ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

A Hybrid Intrusion Detection System Based on C5.0 Decision Tree Algorithm and One-Class SVM with CFA

Meesala Shobha Rani, S. Basil Xavier

PG Scholar, Dept. of Computer Science and Engineering, Karunya University, Coimbatore, India

Assistant Professor, Dept. of Computer Science and Engineering, Karunya University, Coimbatore, India

ABSTRACT: Cyber security threats have become increasingly sophisticated and complex. Intrusion detection which is one of the major problems in computer security has the main goal to detect infrequent access or attacks and to protect internal networks. A new hybrid intrusion detection method combining multiple classifiers for classifying anomalous and normal activities in the computer network is presented. The misuse detection model is built based on the C5.0 Decision tree algorithm and using the information collected anomaly detection model is built which is implemented by one class Support Vector Machine (SVM).The key idea is to take advantage of cuttlefish algorithm (CFA).In the proposed algorithm, Cuttlefish can find best selected features to remove the redundant and irrelevant features to evaluate the accuracy of classification. Integration of multiple algorithms helps to get better performance. The Experimental results are performed on NSL-KDD Dataset, and it is shown that overall performance of the proposed approach is improved in terms of detection rate and low false alarms rate in comparison to the existing techniques.

KEYWORDS: Misuse detection; Anomaly detection; Hybrid approach; C5.0 Decision tree Algorithm; One Class Support Vector Machine; Cuttlefish Algorithm; Feature Selection; Chromatophores; Iridopheres; leucophores.

I. INTRODUCTION

The survey on 'Information Security' in India (2015) reveals that security breaches are increasing year by year. The security attack incidents is in the range of 1 million attacks every year which is in turn about 2800 attacks every day. The global estimated financial loss is about 2.7 million USD, which 34% more than in 2013[1].

Cyber Security is one of the major business risks. The US Federal Bureau of Investigation (FBI) has notified 3000 companies who have been victims of cyber security breach. The survey of stock exchanges conducted by International Organization of Securities Commissions (IOSCO) and World Federations of Exchange Office have found that 53% of the exchanges have been affected by cyber attacks [Global State of Information Security survey 2015].Interconnected devices are more vulnerable to attacks. HP viewed commonly used connected devices and found 70% of serious vulnerability. Google has launched Project Zero initiative, in identifying and stopping threats (unknown code) before any of hackers can exploit by using the attacks.

A proper intrusion detection system when deployed in an organization can avoid threats and vulnerabilities. Intrusion detection is the art of detecting inappropriate, incorrect, or anomalous activity both internally and externally. Generally intrusion detection algorithms are categorized as misuse detection and anomaly detection [2]. The misuse detection algorithm detects attacks based on the known attack signature. It is effective in detecting known attack with low errors. It cannot detect newly created attacks that do not have similar behaviour to the known attacks. In contrast anomaly detection algorithm confirms the normal behaviour profiles. It analyses the current activities with the normal profiles and reporting significant deviations as intrusions. Anomaly detection algorithms can be useful for identifying new attack patterns; it is not effective as compared to the misuse detection model in terms of detection rate and low false alarm rate.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

In general IDS deals with large amount of data which contain redundant and irrelevant features causing excessive training and predicting time. Dimensionality reduction is a commonly used step in machine learning, especially when dealing with a high dimensional feature space. Feature selection (FS) is a part of dimensional reduction which is known as the process of choosing an optimal subset of features that represents the whole dataset. FS has been used in many fields, such as classification, data mining, object recognition and so forth, and has proven to be effective in removing irrelevant and redundant features from the original dataset.

In order to solve the limitations of these conventional intrusion detection methods, hybrid intrusion detection method that combines misuse detection method and anomaly detection method with Cuttlefish feature selection has been proposed.CFA selects best optimal subset selection which give higher accuracy rate. The hybrid intrusion detection system uses both combination of misuse detection and anomaly detection in order to achieve high detection rate and low false alarm. Both normal attacks and anomaly attacks can be detected by using these two models.

The paper is organized as follows: Section 2 presents the related hybrid intrusion detection methods are studied. Section 3 describes the detailed description of introduction and overview of Cuttlefish Optimization Algorithm and new feature selection based on CFA ,C5.0 Decision Tree Algorithm and One-Class SVM section 4 describes experimental setup and result analysis finally conclusion of the paper is given in section 5.

II. RELATED WORK

Extensive research is being carried out for detection of hybrid intrusion detection and feature selection. Gisung Kim et al. [2] presents a new hybrid intrusion detection method hierarchically integrates a misuse detection and anomaly detection in a decomposed structure. The misuse detection model is built based on C4.5 decision tree algorithm and is used to decompose the normal training data into smaller subsets. The one-class SVM is used to create anomaly detection for the decomposed region.C4.5 decision tree does not form a cluster, which can degrade the profiling ability. Adel Sabry Eesa et al.[3] proposed A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Cuttlefish. The model uses the cuttlefish algorithm (CFA) is used to find optimal subset of features and ID3 classifier as a judgment on the selected features that are produced by the CFA. This model has been evaluated on KDD99.Adel Sabry Eesa et al.[4] proposed an Cuttlefish Algorithm – A Novel Bio-Inspired Optimization Algorithm. a new meta-heuristic bio-inspired optimization algorithm, Cuttlefish Algorithm (CFA) is presented. This method achieves better performance in comparison to the Genetic Algorithm (GA), Particle Swarm Optimization (PSO) , Ant Colony Optimization (ACO), Bees Algorithm (BA).Amuthan Prabakar Muniyandi et al. [5] presents an anomaly detection method using K-Means+C4.5 , a method to cascade k-means clustering and the C4.5 decision tree methods. This method achieves better performance in comparison to the K-Means, ID3, Naïve Bayes, K-NN, and SVM.Basant Agarwal et al. [6] proposed an anomaly traffic detection system based on the Entropy of network features and Support Vector Machine (SVM) are compared, then hybrid method is a combination of both Entropy and SVM is compared with individual methods. The Hybrid method outperforms the single method in terms of accuracy but it is not dynamic to decide whether it has attack or not it causes high false alarms. Gang Wang et al. [7] proposed a new approach called FC-ANN, based on ANN (Artificial Neural Network) and fuzzy clustering, to solve the problems in the IDS. This approach achieves better detection precision rate and detection stability in comparison to the back propagation neural network, Decision tree and Naïve Bayes.Hyun Joon Shin et al. [8] proposed a novel test technique for machine fault detection and classification in electro-mechanical machinery from vibrating measurements using one-class Support Vector Machines (SVM).This method gives better performance in detecting outliers in comparison of multi-layered perception it is one of the artificial neural technique.Sankar Mahadevan et al. [9] proposed a new approach for fault detection and diagnosis using 1-class SVM and SVM-recursive feature elimination. This approach is based on non-linear distance metric distance metric measured in feature space. This method achieves better performances in terms of false alarm rates, detection latency and fault detection rates in comparison to the conventional techniques such as PCA and DPCA.Shih-Wei Lin et al. [10] proposed an intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection using Support Vector Machine (SVM), Decision Tree (DT) and simulated annealing (SA).The goal is to find best selected features to elevate the accuracy for only anomaly detection. This method achieves better accuracy in comparison to the hybrid processes of DT, SA, and feature selection, the hybrid process of particle swarm optimization (PSO), SVM and feature selection, only DT, only SVM are used to simulate the results.Shi-Jinn Horng et al. [11] proposed an SVM-based intrusion detection system, which combines a hierarchical clustering algorithm, a simple feature selection procedure, and the SVM technique. This

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

approach provides better performance in terms of accuracy in comparison to the other NIDS. It only detects Dos and Probe attacks not U2L and R2L attacks. Siva S. Sivatha Sindhu et al. [12] proposed a light weight Intrusion Detection System to detect anomalies in the network using a wrapper based feature selection algorithm, the main goal is removing redundant instances, identifying suitable subset of features that maximizes the specificity and sensitivity, adding neural ensemble decision tree to evolve better optimal features. This method increases the detection rate in comparison various six decision tree classifiers are Decision Stump, C4.5, Naïve Bayes Tree, Random Forest, Random Tree and Respective tree model.Vahid Golmah. [13] proposed an efficient hybrid intrusion detection method based on C5.0 and SVM .This method achieves a better performance compared to the individual SVM. Evaluate the proposed method using DARPA dataset. Yinhui Li et al. [14] proposed an efficient Intrusion detection system based on Support vector machines and gradually feature removal method, combination of clustering method, ant colony algorithm and support vector machine.

III. PROPOSED ALGORITHM

A. *Cuttlefish algorithm (CFA)*

A new meta-heuristic bio-inspired optimization algorithm which is called Cuttlefish Algorithm (CFA)[3][4]. The algorithm mimics the mechanism of colour changing behaviour of the cuttlefish to solve numerical global optimization problems. The colours and patterns of the cuttlefish are produced by reflected light from three different layers of cells. The proposed algorithm considers mainly two processes: reflection and visibility. Reflection process simulates light reflection mechanism used by these layers, while visibility process simulates visibility of matching patterns of the cuttlefish. These two processes are used as a search strategy to find the global optimal solution. The diagram in the fig 1: cuttlefish skin detailing the three main skin structures (chromatophores, iridophores and leucophores),two example states (a, b) and three distinct ray traces (1, 2, 3), shows the sophisticated means by which cuttlefish can change reflective colour.CFA reorders these six cases shown in Fig.1 to be as shown in Fig 2: The formulas for finding the new solution (newP) using reflection and visibility is described in eq.(1)

$$newP = reflection + visibility \quad \text{eq. (1)}$$

Global search using the value of each point to find a new area around the best solution with a specific interval

$$Reflection = R * G_1[i].Points[j] \quad \text{eq. (2)}$$
$$Visibilityj = V * (Best.\ Points[j] - G_1[i].Points[j]) \quad \text{eq. (3)}$$

Where G_1 is a group of cells , i is the i^{th} cell in G_1.Points[j] represents the j^{th} point of the i^{th} cell, Best. Points represents the best solution points, R represents the degree of reflection, and V represents the visibility degree of the final view of the pattern's and V are found as follows:

$$R = random\ () * (r_1 - r_2) + r_2 \quad \text{eq. (4)}$$
$$V = random\ () * (v_1 - v_2) + v_2 \quad \text{eq. (5)}$$

Where, random () function is used to generate random numbers between (0, 1) andr1, r2, v1, v2 are four constant values specified by the user. As a local search, CFA uses Cases 3 and 4 to find the difference between the best solution and the current solution to produce an interval around the best solution as a new search area. The formula for finding the reflection is as follows

$$Reflection_j = R * Best.\ Point[j] \quad \text{eq. (6)}$$

While the formulation for finding the visibility remains as in Cases 1 and 2.For Case 5, the algorithm also uses this case as a local search, but this time the difference between the best solution points and the average value of the Best points is used to produce a small area around the best solution as a new search area. The formulas for finding reflection and visibility in this case are as follows

$$Reflection_j = R * Best.\ Point[j] \quad \text{eq. (7)}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

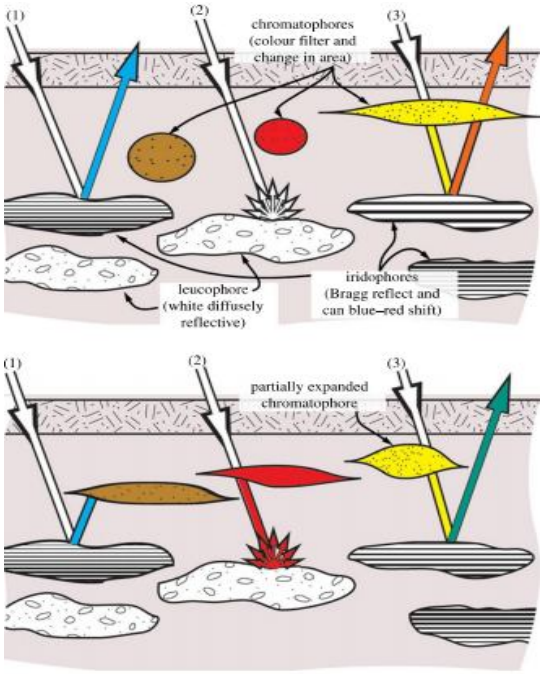
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

$Visibility_j = V*(Best.\ Points[j] - AV_{best})$ eq. (8)

Where, AV_{Best} is the average value of the Best points. Finally, the CFA uses Case 6 as the random solutions. The general principle of CFA is shown in Fig.3.



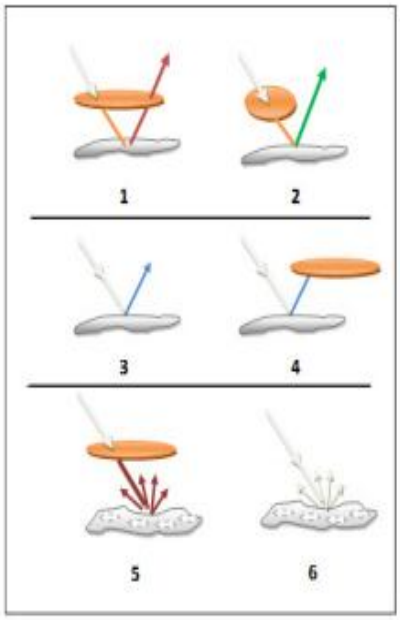


Fig.1. Diagram of cuttlefish skin detailing the three main skin structures (Chromatophores, iridopheres and leucophores) [25]

Fig.2. Rearrange of the six cases in fig.1.[3]

B. *A New Feature Selection approach based on General CFA in proposed scheme*

Step 1: Simulation of cases 1 and 2:

The simulation of these two cases is started by sorting the population P in descending order according to the fitness values. After that the new Subset will be generated from p_i, where i = 1, 2, . . . , k. k is an integer number generated randomly between (0, and N/2). In the original algorithm, R represents the reflection degree used to find the stretch interval of the saccule when the muscles of the cell are contracted or relaxed, while V represents the degree of visibility of the final view of the matched pattern. Here, the main equation and the operator of reflection and visibility in the original algorithm for Cases 1 and 2 are modified as follows:

$newSubset_i = Reflection_i \ \ Visibility_i$ eq. (9)
$Reflectioni = random\ subset[R] \quad pi.selectedFeatures$ eq. (10)
$Visibility_i = random\ subset\ [V] \quad p_i.unselectedFeatures$ eq. (11)

Where $Reflection_i$ and $Visibility_i$, are two subsets with size R and V their elements are produced randomly from selected Features and unselected Features, respectively. The value of R and V can be calculated as follows:

R = random (0, selectedFeatures: Size)
V = selectedFeatures.Size _ R

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

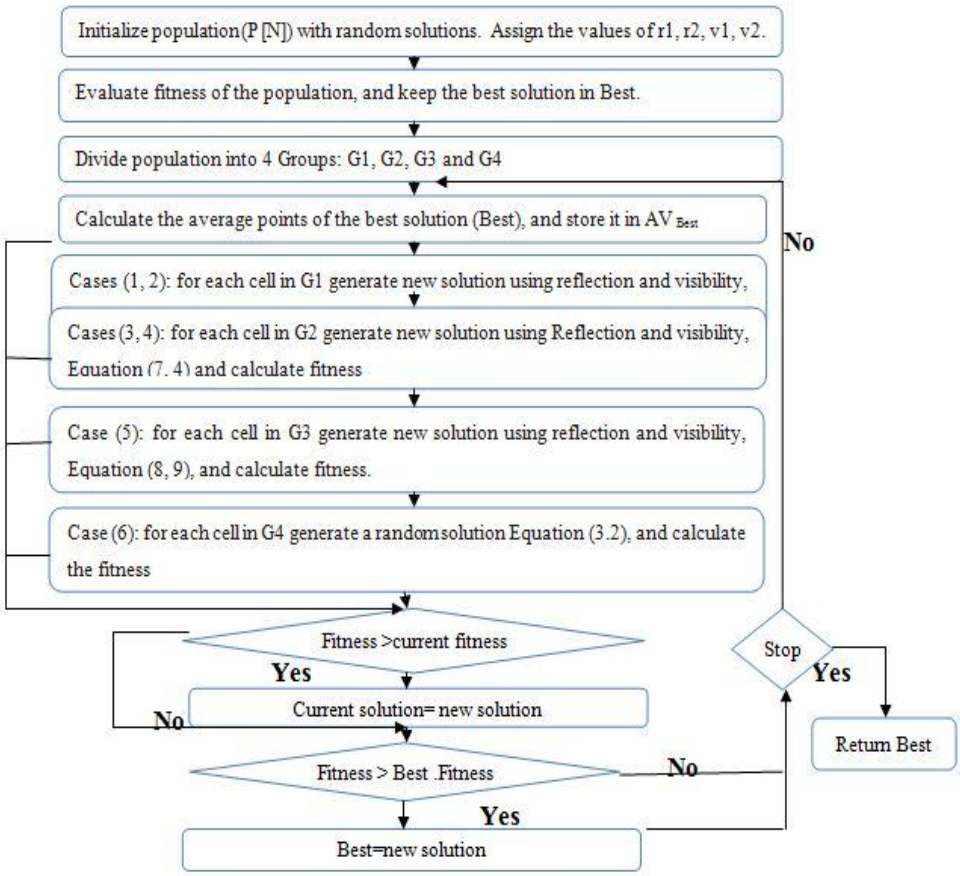


Fig.3. General Principal of CFA

The operator represents the combination (union) of these two subsets to find the new Subset(i).The size of selected Features is bigger than unselected Features it may cause a problem, because the number of elements in unselected Features is not enough to generate a new Subset, in this case we can use original features to produce a new Subset without repetition of any element. In short, Cases 1 and 2 use the best k solutions from the half best solution of population P to be a part of the new solution (new subset), by keeping some elements from the best solutions and completing them with some new elements that have a chance to be a part of the new solution.

Step 2: Simulation of Cases 3 and 4:

Iridophores cells are light reflecting cells which are assisting in organs concealment. That means the final reflected colour from Iridophores cells around the organs is very similar to the organs original colours. Therefore, we assumed that the organs are represented by the best solution (best Subset), and the final reflected colour is represented by the new solution. The new solution colour should be very similar to the colour of the organs. As a simulation to the similarity of the incoming colour and the reflected colour, we considered that the reflected colour is a subset produced by removing only one random feature from the incoming colour (selected Features) whereas the visibility is considered as a single feature which is selected randomly from the unselected Features. The combination between the reflection and the visibility will produce the new solution (new Subset). So the formulation of finding the new solution, reflection and the visibility are reformulated as follows:

Reflection = bestSubset_SelectedFeatures- best Subset Features[R] eq. (12)
Visibility = best Subset unselected Features [V] eq. (13)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Where R represents the index of the feature that should be removed from selected Features. V represents the index of the feature that should be selected from unselected Features. The calculation of finding R and V are as follows:

R= random (0, bestSubset_selected Features. Size)
V = random (0, best Subset unselected Features. Size)

The new Subset is then calculated by using Eq. (9), by adding the feature represented by the visibility to the subset represented by the reflection. In another words, the new Subset is a set produced by exchanging the feature R with the feature V.

Step 3: Simulation of Case 5:

In this case, the leucophore cells work as a mirror. The cells will reflect the predominant wavelength of light in the environment. In this case, the light is coming through chromatophore cells with specific colour. The outgoing light is very similar to the light coming from the chromatophore cells. In order to cover the similarity between the incoming colour and the outgoing colour, we proposed that the values of the incoming colour and the reflection be equal to selected Features in $AV_{Bestsubset}$ and the outgoing colour be a new sub-set produced from the selected Features in $AV_{Bestsubset}$ by removing one feature from it. While the visibility represents the feature I that should be removed from selected Features. These operations make the matched pattern very similar to the original pattern that appears in the environment. In this way, we can produce R new sub-sets, the value of R is equal to the size of selected Features each subset representing the matched pattern by removing one feature from selected Features at each time. The two equations for finding the reflection and the visibility and the main equation are modified as follows.

$newSubset_i$ = Reflection - $Visibility_i$ eq. (14)
Reflection = $AV_{Best\ subset}$ – selected Features eq. (15)
$Visibility_i$ = $AV_{Best\ subset}$ – selected Features [i] eq. (16)

Where, I represents the index of the features that should be removed from selected Features: I = {1,2, . . . , R}, and R is the size of selected Features.

Step 4: Simulation of Case 6:

In this case, the leucophore cells will just reflect the incoming light from the environment. This operator allows the cuttlefish to blend itself into its environment. As a simulation, one can assume that any incoming colour from the environment will be reflected as it can be represented by any random solution. In the initial algorithm, this case is used to generate random solutions. Also, we use this case as a random generator process to generate random solutions. The number of generations is equal to m, where m = N _ k. k is a random number which was previously generated in Cases 1 and 2. The new generation will be started at the location k after sorting the population P in descending order. If the new generated solution is better than the current solution, then the current solution is replaced with the new solution. The process of random solutions is the same as that used with the initialization process which is described in Step1.Fig 4.shows that the Block Diagram of new Cuttlefish Feature Selection based on General CFA

C. *Fitness Function*

C5.0 Decision Tree classifier evaluates the quality of the each subset of features. Decision tree decide that which subset of feature is better according to the Eq. (18) [3]

$$\text{Fitness} = \alpha * DR + \beta * (1 - FPR) \quad (18)$$

Equation.(18) shows that DR and FPR have different importance based on α and β, where $\alpha \in [0, 1]$ and β=1-α are two parameters referring to the importance of the DR quality and FPR, In the experiment ,the quality of DR is more important than FPR and α=0.7, β=0.3

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

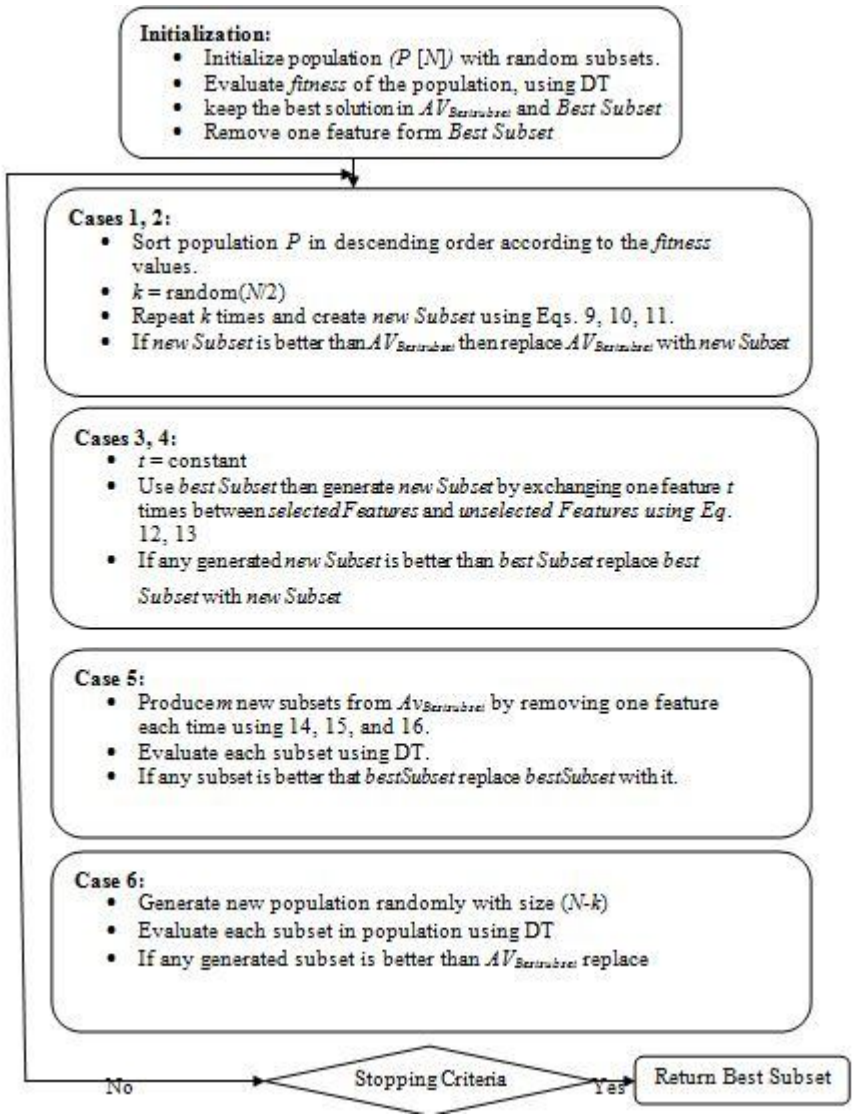


Fig.4. Block Diagram of new Cuttlefish Feature Selection based on General CFA

D. *C5.0 Algorithm for building misuse detection in proposed scheme*

The C5.0 algorithm is the latest version of machine learning algorithms (MLAs) developed by Quinlan, based on decision tree [15]. The decision trees are built based on list of possible attributes and set of training instances, and then the tree can be classified by using subsequent set of test instances. It is a modified version of well-known and widely used C4.5 Classifier and it has several important advantages over its ancestors [16]. C5.0 supports boosting of decision trees. Boosting is a technique for generating and combining multiple classifiers to give improved final predictive accuracy. C5.0 incorporates variable misclassification costs. It allows separate cost for each predicted / actual class pairs. C5.0 constructs classifiers to minimize estimated misclassification costs rather than the error rates. New attributes are dates, times, timestamps, ordered discrete attributes. The values can be marked as missing or not applicable for particular cases. It supports sampling and cross-validation.C5.0 models are quite robust in the presence of problems such as missing data and large numbers of input fields. It does not require long training times to estimate. In addition, it is easier to understand than some other model types, since the rules derived from the model have a very straightforward interpretation. C5.0 has option to convert the tree to rules .C5.0 tree or rule sets are usually smaller than C4.5 [17].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

E. *Information Gain and Entropy*

Information gain is used to decide how well an attribute separates the training data according to the target model. It is based on a measure commonly used in information theory known as entropy. The units of entropy are bits.

Let T is the training sample set.

C_i is Class I; i= 1,2,…….n

$I(T_1,T_2\ldots,Tn) = -\sum pi\ log2(pi)$ eq. (19)

T_i is the number of samples in class i

$Pi=T_i/T$

$Log_{2\ is}$ the binary Logarithm

Let attribute F have v distinct value

Entropy= E(F) is $\sum\{(T1j + T2j + \cdots + Tnj)/T\} * I(T1j \ldots..Tn)$ j=1 eq. (20)

Where T_{ij} is Samples in Class i and subset j of attribute F

$I(T1j, T2j, \ldots.Tnj) = -\sum pijlog2(pij)$ eq. (22)

$Gain(F) = I(T1, T2, \ldots Tn) - E(F)$ eq. (22)

F. *Decision Tree Based On C5.0 Classification Algorithm*

Step 1: The C5.0 node generates either decision tree or a rule set [18].

Step 2: A C5.0 works by splitting the sample into subsample based on the field that provides maximum information gain by using eq. (22)

Step 3: The target field must be categorical .Multiple Splits into more than two subgroups are allowed.

Step 4: Each subsample defined by the first split is then split again, based on a different field, and the process Iterated until the subsamples cannot be split any more or the partitioning tree has reached the threshold.

Step 5: Finally, the lowest-level splits are re-examined, and those that do not contribute significantly to the value of the model are removed or pruned.

G. *One Class SVM for anomaly detection in proposed scheme*

The One-class SVM was proposed by Scholkopf et al. was inspired by general SVM. One-class SVM is a famous outlier (or) novelty (or) anomaly detection algorithm in various application like machine fault detection and document classification [8]. It identifies outliers among positive instances and uses them as negative instances. It is used to classify anomalous packets as outliers.

Let $x1, x2 \ldots\ldots xl \pounds X$ be the training data instances belonging to original space X *and* l be the number of instances. The 1-classSVM may be viewed as a regular binary SVM where all training data lies in the first class and the origin belongs to the second class. It discovers the maximal margin hyper plane that best separates the training data from the origin (Scholkopf et al., 2001). It is difficult to locate a hyper plane that creates training data patterns separable from the origin in the original space X, the SVM uses a feature map ($\phi: X \rightarrow F$) , which non-linearly transforms the data from the original space to the feature space in order to locate the hyper plane in the feature space. The 1-class SVM is formulated as the following quadratic programming.

$$\min_{w,\varepsilon,\rho} \frac{1}{2}||w||^2 + \frac{1}{vl}\sum_{i=1}^{l} \varepsilon_i - \rho \text{ Subject to } (w.\phi(xi)) \geq \rho - \xi i \ \ \xi i \geq 0, i = 1, \ldots. l \quad eq.(23)$$

Where w is the weight vector orthogonal to the hyper plane,$\varepsilon = (\varepsilon 1, \ldots, \varepsilon l)$ is the vector of slack variableused to panalize the rejected instances,and ρ represents the margin (the distance of hyperplane from the origin), v is the parameter that controls the trade-off between maximizing the distance of hyperplane from the origin and fraction data containing in the separate region.Due to curse of dimensionality[8][19], the SVM utilizes the kernal theory,the inner dot product in the feature space is calculated using a simple kernel function $k(x,y) = \phi(x).\phi(y)$, such as Gaussian kernel, $k(x,y) = e^{-}\boldsymbol{\gamma}||x\text{-}y||^2$.Using the kernel function and Lagrangian multiplie to the original quadratic programming ,the solution of eq.(23) creates a decision function. The generic test instance (x) is

$$f(x) = sgn((w.\phi(x) + b) - \rho) \quad \text{eq. (24)}$$

The test instance (x) is accepted when f(x) is positive and it is rejected when f(x) is negative. Positive instances indicate that test instance (x) is similar to the training data and the Negative instances indicate that it departs from the training data and is considered as anomaly.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

H. **System Architecture**

Fig.6. shows the architecture of proposed hybrid intrusion detection system with C5.0 and One-class SVM with Cuttlefish feature selection. The NSL-KDD dataset is divided into two training data and testing data, then the training data is prepared and pre-processing .The Cuttlefish takes the input as pre-processed training data; train the features that are selected from the training data. Calculate the fitness function and select the best subset. The best subset selection gives the input to the C5.0. C5.0 generates the node information along with the original set of information. The node information is passed through the One-Class SVM to obtain the final output whether it is normal or anomaly.

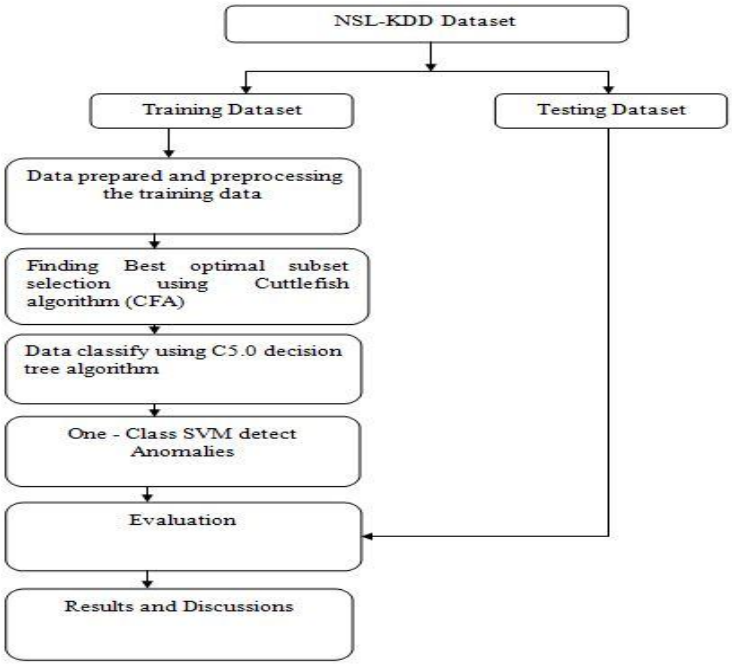


Fig.6.Frame Work of Proposed Architecture

IV. EXPERIMENTAL SETUP

The NSL-KDD[20[21] dataset are taken to evaluate the proposed C5.0 & One-class SVM with Cuttlefish feature selection Algorithm(CFA) in terms of Accuracy, True positive (TP),True negative (TN), False positive(FP), False negative(FN), Recall, Precision, Specificity, F-Measure, Roc. The experiment has been performed using Intel core 5 Processor with 4 GB of RAM and LIBSVM (MATLAB)[22]. LIBSVM is an integrated software tool for support vector classification, regression and distribution estimation, which can handle One-class SVMs. The proposed method is compared with C4.5 and one-class SVM, C5.0 and One-class SVM. To evaluate the performance of proposed technique Confusion matrix is used, it contains data about actual and predicted classifications [23]. The simulation results shows that by implementing the proposed methodology is better when compared to the existing Techniques. The accuracy rate of proposed algorithm is 98.2%, it is very high compared to the existing approaches i.e., C4.5 and One-Class SVM is 88.5% and C5.0 and One-Class SVM is 93%.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

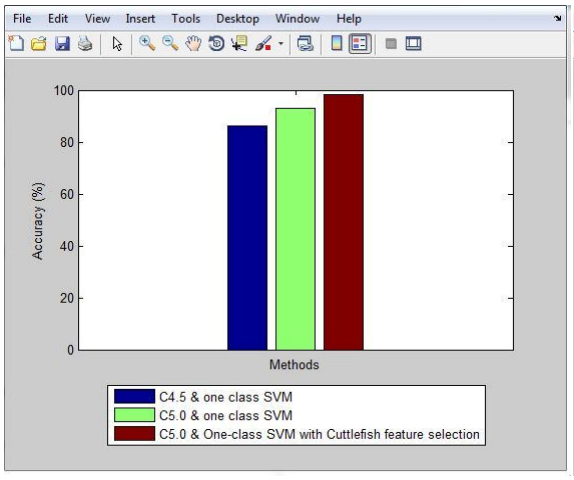


Fig .7.Accuracy Comparison

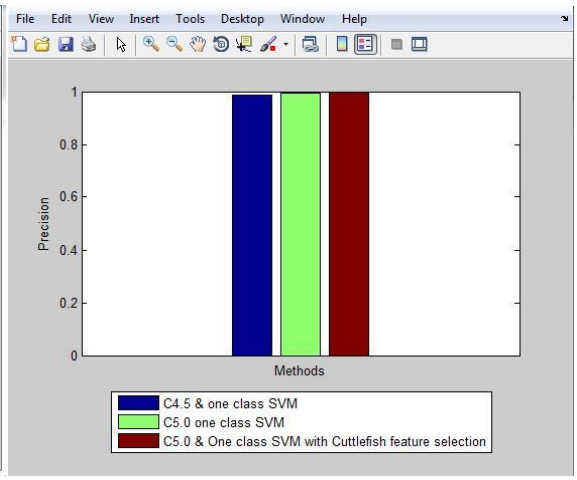


Fig.8.Precision Comparison

Fig.7. shows that the comparison of the accuracy of the attack prediction. The comparison is made between the existing method called C4.5 with one-class SVM, C5.0 and one-class SVM and the proposed C5.0 & One-class SVM with Cuttlefish Feature Selection (CFA). Accuracy is defined as the proportion total number of predictions that are correct. The X- axis in the graph represents the input data set whereas the Y- axis represents the Accuracy rate of attack prediction. The results of the graph show that by implementing the proposed methodology the accuracy rate is increased when compared to the existing method. Fig.8. shows that the occurrences of Precision. Precision is defined as the proportion of the predicted positive cases that were correct.

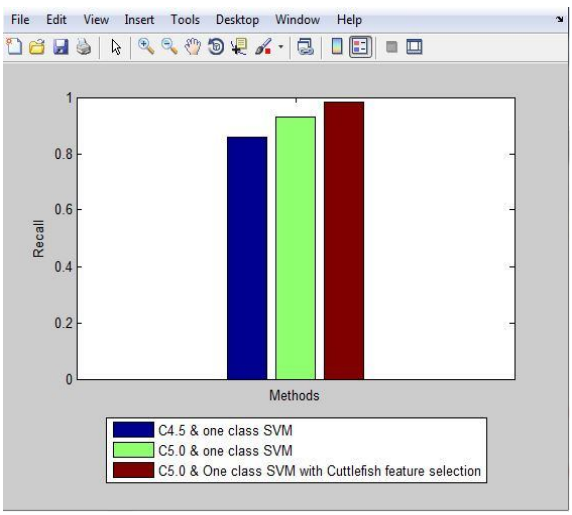


Fig.9. Recall Comparison

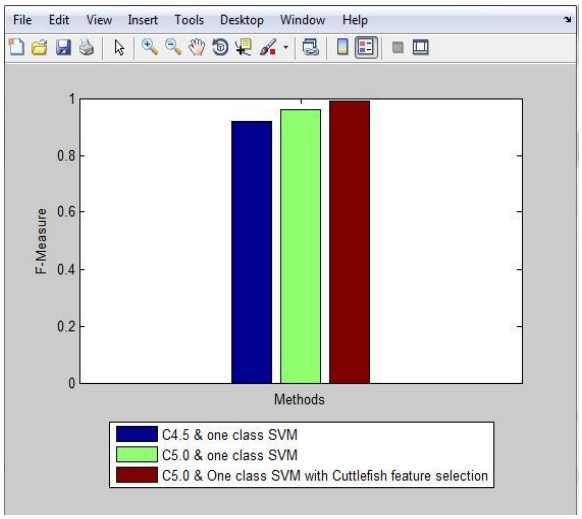


Fig .10.F-Measure Comparision

Fig.9. shows that the occurrences of recall rate. The recall or true positive rate (TP) is defined as the proportion of positive cases that were correctly identified.Fig.10.shows that the occurrences of F-Measure. The F-Score consider both precision and recall of the procedure to compute the score The X- axis represents the input data set whereas the Y – axis represents the F-Measure rate.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

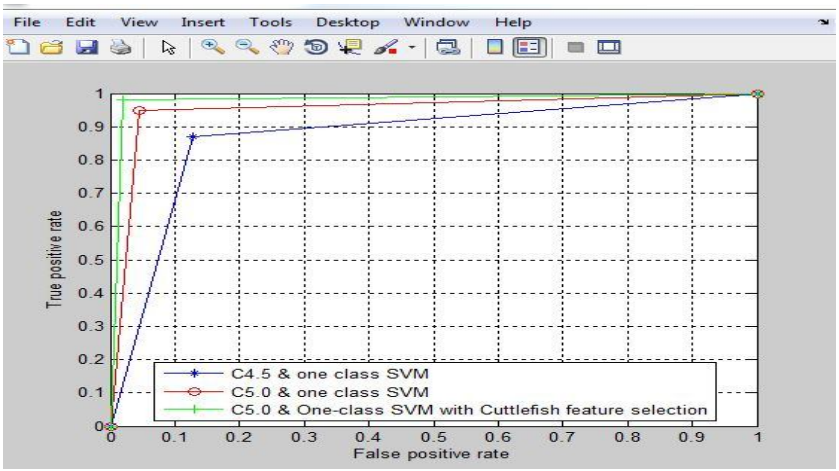


Fig.11. Roc Curve

Fig.11. shows that the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the detection probability in the y-axis and the false-alarm probability in x-axis.

V. CONCLUSION

Intrusion detection is one major research problem in network security; main goal is to detect infrequent access or attacks to protect internal networks. In this study, a new proposed feature selection algorithm based on CFA is used to find the best optimal subset selection, C5.0 decision tree classifier is used to evaluate the best selected features and One-class SVM is used for finding Normal and Anomaly attacks. The hybrid intrusion detection system is a combination of misuse detection and anomaly detection. The Combination of hybrid detection model gives better performance, minimizes the time complexity, achieves high detection rate and low false alarm rate, reduces irrelevant features and improves the system performance. According to the experiment on the NSLKDD dataset, the proposed system could reach a high accuracy 98.2% with a false alarm rate 1.7%.Compared with other techniques that also applied NSLKDD as dataset. The proposed algorithm C5.0 AND One-Class SVM with Cuttlefish algorithm (CFA) outperforms other existing techniques. Simulation results demonstrate that the proposed algorithm is successful in detecting Normal and anomaly intrusion detection system.

REFERENCES

1. Managing cyber risks in interconnected world Key findings from The Global State of Information Security® Survey 2015 is Available at http:// www.pwc.com/gsiss2015
2. Gisung Kim and Seungmin Lee., "A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection With Misuse Detection", ELSEVIER, Expert Systems with Applications,Vol.41, pp.1690 – 1700, 2014.
3. Adel Sabry Eesa., Zeynep Orman and Adnan Mohsin Abdulazeez Brifcani., "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems", Expert Systems with Applications, Vol.42, pp.2670–2679, 2015.
4. Adel Sabry Eesa., Adnan Mohsin Abdulazeez Brifcani and Zeynep Orman., "Cuttlefish Algorithm – A Novel Bio-Inspired Optimization Algorithm", International Journal of Scientific & Engineering Research, Vol. 4, Issue 9, September 2013
5. Amuthan Prabakar Muniyandi., R.Rajeswari and R. Rajaram., "Network Anomaly Detection By Cascading K-Means Clustering And C4.5 Decision Tree Algorithm", ELSEVIER, In Procedia Engineering,Vol.30 pp.174 – 182, 2012.
6. Basant Agarwal and Namita Mittal., "Hybrid Approach For Detection Of Anomaly Network Traffic Using Data Mining Technique", ELSEVIER, In Procedia Engineering, Vol. 6 , pp.1996 – 1003, 2012.
7. Gang Wang., Jinxing Hao., Jian Ma and Lihua Huang., "A New Approach To Intrusion Detection Using Artificial Neural Networks And Fuzzy Clustering", ELSEVIER, Expert System with Applications Vol.37 pp.6225 – 6232, 2010.
8. Hyun Joon Shin., Dong-Hwan Eom and Sung-Shick Kim.,"One-class support vector machine-an application in machine fault detection and classification", ELSEVIER, ,Computer & industrial engineering,Vol.48, pp.395-408, 2005.
9. Sankar Mahadevan and Sirish L. Shah., "Fault detection and diagnosis in process data using one-class support vector machines", ELSEVIER , Journal of Process Control Vol.19, pp. 1627-1639, 2009.
10. Shih-Wei Lin., Kuo-Ching Ying., Chou-Yuan Lee and Zne-Jung Lee., "An Intelligent Algorithm With Feature Selection And Decision Rules Applied To Anomaly Intrusion Detection", ELSEVIER, Applied Soft Computing,Vol.12 pp.3285-3290, 2012.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

11. Shi-Jinn Horng and Ming-Yang Su., "Novel Intrusion Detection System Based On Hierarchical Clustering And Support Vector Machines", ELSEVIER, Expert Systems with Applications,Vol.38, pp.306-313, 2011.
12. Siva S. Sivatha Sindhu., S. Geetha and A. Kannan., "Decision Tree Based Light Weight Intrusion Detection Using A Wrapper Approach", ELSEVIER, Expert Systems with Applications Vol.39, pp.129-141, 2011.
13. Vahid Golmah., "An Efficient Hybrid Intrusion Detection System Based On C5.0 And SVM", International Journal of Database Theory and Applications vol.7 No.2 , pp.59 – 70. http://dx.doi.org/10.14257/ijdta.2014.7.2.06, 2014.
14. Yinhui Li and Jingbo Xia., "An Efficient Intrusion Detection System Based On Support Vector Machines And Gradually Feature Removal Method", ELSEVIER, Expert Systems with Applications Vol.39, pp.424-430, 2012.
15. Information on See5/C5.0-RuleQuest Research Data.See5/Mining Tools, 2011 .[Online].Available: http://www.rulequest.com/see5-info.htm
16. Is See5/C5.0 Better Than C4.5?,[Online].Available: http://www.rulequest.com/see5 -comparison.html,2009.
17. See/C5.0 updated record [Online]. Available: https://www.rulequest.com/see5-previous.html
18. Prof Manasi Kulkarni and Ms Rashmi R. Tundalwar., "Web Spam Detection Using C5.0 Classification Algorithm", IJARCSSE .vol.3 issues 2, 2013.
19. Manevitz and Yousef, M ., "One-class SVMs for document classification", ELSEVIER, Journal of Machine Learning Research Vol.2, pp.139-154, 2002
20. M. Tavallaee., E. Bagheri., W. Lu, and A. Ghorbani., "A Detailed Analysis of the KDD CUP 99 Data Set", Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.
21. Hee-su Chae., Byung-oh Jo., Sang-Hyun Choi., Twae-kyung Park., Feature Selection for Intrusion Detection using NSL-KDD, Recent Advances in Computer Science.
22. LIBSVM 2.5 is available at http://www.csie.ntu.tw/~cjlin/libsvm
23. Kohavi and Provost, Confusion Matrix. [Online].Available:http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html,1998.
24. Swets, ROC Graph is Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/ROC/ROC.html,1988.
25. Eric Kreit, Lydia M. Mathger., Roger T. Hanlon., Patrick B. Dennis., Rajesh R.Naik., Eric Forsythe and Jason Heikenfeld., "Biological verses electronic adoptive coloration: how can one inform the other", http://dx.doi.org/10.1098/rsif.2012.601.

BIOGRAPHY

Meesala Shobha Rani ,received the B.Tech degree in Computer Science and Engineering from JNTU Ananthapur 2012, and M.Tech degree in Computer Science and Engineering with a specialization of Computer and Communication Engineering from the Karunya University 2015, Coimbatore respectively**.**