International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Intrusion Detection System for Large Scale Data using Machine Learning Algorithms

Sayali R. Kshirsagar, P.B.Kumbharkar

Abstract*: *To provide security to internet assets, Intrusion Detection System (IDS) is most essential constituent. Due to various network attacks it is very hard to detect malicious activities from remote user as well as remote machines. In such a manner it is mandatory to analyze such activities which are normal or malicious. Due to insufficient background knowledge of system it is hard to detect malicious activities of system. In this work we proposed intrusion detection system using various soft computing algorithms, the system has categorized into three different sections, in first section we execute the data preprocessing as well as generate background knowledge of system according to two training data set as well as combination genetic algorithm. Once the background knowledge has generated system executes for prevention mode. In prevention mode basically it works for defense mechanism from various networks and host attacks. System uses two data sets which contain around 42 attributes. The system is able to support for NIDS as well as HIDS respectively. The result section will show how proposed system is better than classical machine learning algorithms. With the help of various comparative graphs as well as detection rate of systems we conclude proposed system provides the drastic supervision in vulnerable network environment. The average accuracy of proposed system is 100% for DOS attacks as well as around more than 90% plus accuracy for other as well as unknown attacks respectively.

Index Terms*: *Genetic Algorithm, HIDS Machine Learning Algorithm, NIDS, Ensemble method.

I. INTRODUCTION

Intrusion Detection Systems (IDS) focuses on identifying possible incidents or threats, logging information, attempting to stop intrusion or malicious activities, and report it to the management station. Additionally, it record info associated with ascertained actions, inform security directors of considerably ascertained actions and generate reports. Several Intrusion detection systems also react to a detected hazard by making an attempt to forestall it from following. They have used varied response techniques like fixing the protection surroundings for instance, reconfiguration of a firewall or fixing of the contents of attack for stopping attack itself. So IDS helps in applied math analysis for malicious behavior. Our goal is to spot novel attacks by unauthorized users in an exceedingly specific network. If the vulnerability is unknown to the target's administrator or user, we have a tendency to think about an attack to be novel although the attack or signature pattern is usually illustrious. We have a tendency to square measure in the main taking note in four forms of remotely launched attacks: denial of service (DOS), probe, U2R and R2L. A DoS attack may be a sort of attack within which the hacker or assaulter makes a memory resources or computing resources thus busy or full to serve rightful networking requests and deny users to access to a system. The samples of Dos attacks square measure Neptune, apache, ping of death, mail bomb, smurf, UDP storm etc. A far off to user (U2R) attack is an attack within which assaulter or hacker sends packets to an ADP system over a selected network, so as to reveal the machines weakness and vulnerabilities and abuse rights that a neighborhood user would wear the machine that he/she doesn't have access rights. The samples of U2R attacks square measure sendmail lexicon, xnsnoop, xlock, guest, phf, etc. A R2L attack is an attack within which attackers exploits a system by beginning or accessing a system with traditional approved user account and gain user privileges. The samples of R2L attacks square measure xterm, perl etc. A probe is an attack within which the hacker scans a networking device or a system for crucial weaknesses or vulnerabilities thus by compromising the system. This method is usually employed in data processing.

II. LITERATURE SURVEY

In this section we illustrates the complete literature review background of intrusion detection system the various existing systems has done different security mechanisms to provide the security for vulnerable environments. DARPA organization has already introduced KDDCUP99 data set in 1999. Similarly NSLKKD as proposed in 2003, the basic difference of both data set KDD contains around 23 sub attacks for all four classes rather than NSLKDD contains 38 sub attacks for four classes respectively. The data set having numerous flexible attribute like numeric as well as string, the first 6 attribute in entire data set might be effective for generating the dynamic rule from machine learning algorithm. Below are the various existing systems where many authors have already done some intrusion detection work. We had also found some gaps in all those given survey and given the oven contribution to eliminate such problems in IDS. In [1] authors implemented IDS for detection of attacks in the Android mobile devices using flow anomaly detection technique. This system uses ANN (Artificial Neural Network) on Android Operating System (AOS) for discovery of abnormal action in android mobiles.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Sayali R. Kshirsagar, M.E. Computer Engineering from JSPM's Rajarshi Shahu College of Engineering.
P.B.Kumbharkar, Professor in Computer Engineering, Dean (Planning and Development) and IQAC CO-ordinator, Rajarshi Shahu College of Engineering Tathawade Pune

Accuracy and detection rate stated by this method are 85% and 81% respectively. Limitation of this method is treated respecting CPU, memory and battery power .This effort aims to identify a lightweight, extensible and efficient IDS for an android domain. For forwarding public assaults various utilities are arranged. Efficient machine learning algorithms are used for evaluation of data surge.

There are many opportunities to improve efficiency and detection rate.

In [2] system is implemented for a network intrusion detection in cyber security using neuromorphic computation. We know that cyber security is serious concern in the cyberspace. Authors admitted the presentation of a neuromorphic intellectual measuring reach towards network IDS for computerized insurance using deep learning. The scheme used is Discrete Vector Factorization. The NSL-KDD dataset is used to gain accuracy and classification. Result stated are accuracy as 90.12% and classification rate as 81.31% respectively. Deep learning attains individual-level achievement in particular for respective tasks. Deep learning method merges the features of extraction classification. Here new challenge is to resolve the depiction of data in spiking pattern for the better usage in the True-North-System.

In [3] a Hidden Markow Model for IDS is refined for software-defined networking (SDN). SDN associate can benefit to audit the global insurance of a system by analyzing the web as a crack and formulating decisions to fight the network based on the knowledge from the full network. It adds use of ANN IDS. This technology grants greater influential control of a network surrounding. This method have advantages like increased in the scope of activities and security utilization. It has shown that machine learning application influences the capability to be used to fix the risk in networking environment. The future scope is to broaden the feature vector use by HMM in determining the maliciousness of a set data which need to be added.

In [4], authors described Intrusion Detection System for PS-Poll DOS attack in 802.11. It is a networks corruption duration separate event system. For sleuthing DOS intrusion this methodology uses RTDES on time period separate event system. One in all the vital benefits are high accuracy and detection rate. A loss of frames could be one in all the foremost deficiency. It is noticed that the PS-DOS attack need cryptography amendment in protocol or installation of proprietary hardware.

Authors developed Secure Intrusion Detection System for MANETS using hybrid cryptography which is described in [5]. EAACK uses the ideas of hybrid cryptography techniques to scale back system burden caused by digital signature. By providing Hybrid Cryptography technique to EAACK theme, it'll become tough for assaulter to interrupt the network still as retrieved the data.

Mehdi Ezzarii [6] proposed a well-known system on sequential algorithmic rule which relies on gene copy and mutation. Current analysis has seen that extra information embedded aboard individual chromosomes transmits information into future offspring. This extra transmission of data into kid generations outside deoxyribonucleic acid is understood as epigenetic. Extra information is taken into account because the epigenetic issue that helps United States of America to outline randomness crossover and mutation utilized in classical genetic algorithmic rule. This paper conjointly presents a state of art wherever we tend to attempt or to explore epigenetic algorithms inside the context of Intrusion Detection System. Here we tend to review the methodology utilized in genetic algorithmic rule and the way our proposed methodology will achieve detection of intrusions for a competent security.

In [7] authors proposed an advanced method for detection of botnet traffic using Internal Intrusion Detection. It is verified that this is the advanced method for detection and to improve the security by identifying and tracking the attacker using machine learning, ranking and Voronoi clustering. Machine learning, ranking and Voronoi clustering ensures reduction in the size of data set and high detection accuracy. A data set called ISOT has been used for detection of botnet traffic. The processing lag in the large scale network UDP and TCP are audited to observe attained disciplined growth in network traffic. It is considered that machine learning components act like deep neural network. Different botnet approaches are provided and DNA based method is developed for the system benefit. This paper also adopts characteristics of the network stream to observe the botnet intrusion against packet payload size, which supports in encryption of packet.

In [8] authors stated that use of common path mining helps hybrid IDS for better detection. Data mining is used for the building of a power system which operates on data logs. This is an automated way to frame the hybrid IDS. Important benefit of this method is disclosure efficiency which is near about 73%. Abduction of a large data logs is very time consuming and complex hence this method is not much helpful for big data problems. The structure drags attributes of signature-based and specification based IDS. To review the standard path the data mining method that combines investigated records from numerous system accessories are used. The automated methodology wipe outs the demands of manual investigation and manual code pattern.

In [9] authors have used ADS-B techniques for the detection of intrusion. In this method authors have used HMAC data set. The reason behind use of this dataset is to boost the performance of air traffic control. This arrangement works with minimum or essential burden. When we keep ADS-B position valid, system gives better performance. Its distance at the time of task must be within the secure territory. ADS-B come up as a substitute to modern radio, radar standards in aircraft signaling. Exceptional location reliability is contributed by GPS utilizing the cyber-physical environment which helps in attack detection confirmation. A methodology is put forward to interchange the keys used for the HMAC innovation securely. ATC Centre commences strong harmonies with ATC's that regulate other zone in the flight pathway to pass on the private key over public key infrastructure (PKI) arrangements.

Based on genetic fuzzy classification Intrusion Detection System has been developed in [10] with various machine learning algorithms. Fuzzy systems are accustomed to solve many classification issues.

Genetic-fuzzy systems crossbreed the approximate reasoning technique of fuzzy systems with the educational capability of organic process algorithms. During this paper a completely unique intrusion detection technique is bestowed, capable of police investigation.

Traditional and intrusive behavior that extracts each correct and explicable fuzzy IF-THEN rules from network dataset for classification. This technique uses the fuzzy association rule based classification technique. Top dimensional issues are supported by stages to get associate with correct degree and compact fuzzy rule based classifier with a less computational cost.

III. RESEARCH METHODOLOGY

A long-term issue in network traffic classification is generated by unwanted and unrelated attributes in dataset. These attributes not only weaken the procedure of classification but also blocks the classifier from accomplishing authenticate findings, when third party produces the flash actions on vulnerable network.

In this research work, system defines GA based rule creation system according to their feature selection method that worked on NIDS as well as HIDS. Genetic algorithm is an optimization algorithm, which is used for finding optimal solution. Ensemble approach with various classification algorithms will provide a best detection with NIDS in all type of sub attacks with master class. With the proposed research work our aim to generate strong rules and increase the detection rate for DOS, PROBE, U2R and R2L for NIDS and HIDS. Fig.1 shows overall system architecture.

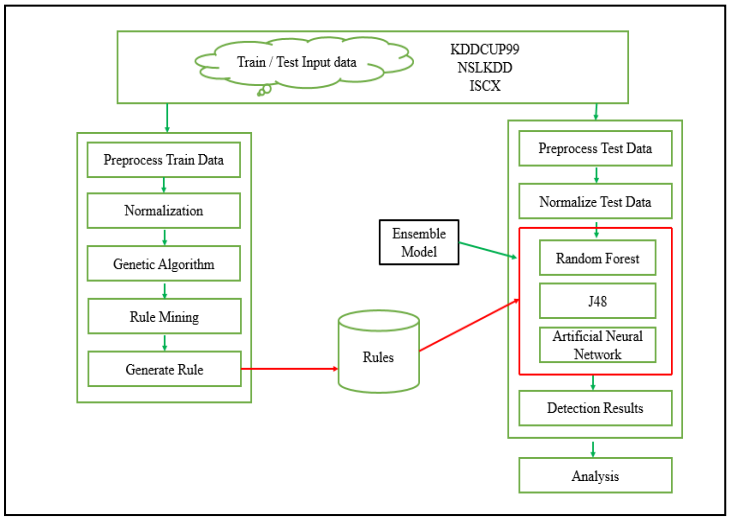

Fig 1: System Architecture

Overall system consists of two modules as follows:

A. Training module :

In this Phase, Genetic algorithm is used where, we first initialize the chromosomes and group of chromosomes we say as population is created. Once the population is created crossover is applied to obtain new generation of chromosomes. Mutation is applied for updating bit value of attributes of chromosomes randomly. The fitness function will define the fitness value of each chromosome and a selection criterion is applied for selected optimal rules. When variation is completed then Genetic algorithm will get terminated. The outputs of genetic algorithm are genetic rules. The output of genetic algorithm that is genetic rules is given as an input to fuzzy logic. In this phase probability of each attribute is calculated which is used for classification of data as attack or normal

Step 1: System first collect network traffic from network audit data using remove environment or some synthetic dataset like KDDCup99, NSLKDD etc.
Step 2: Select features of each connection and apply Genetic Algorithm (GA) for rule creation.
Step 3: Once rule created store it into local database directory called as BK rules.

B. Testing module :

In this Phase, the rules which we are getting from association rule mining are considered as final rules and these rules are given as an input to the ensemble method for the classification of sub attack. Here system collect the network traffic data using PacketXLib and Wincap Driver. On each instance neural network algorithm will be applied. Transfer function will be used for calculating each node weight .Using defined threshold, sub attacks can be classified.
Step 1: System collect the network traffic data using remote driver or NSLKDD
Step 2: Read each instance and apply ensemble (J48, ANN, NB) algorithm.
Step 3: Calculate the weight using given functions for each connection.
Step 4: Finally classifies each attack with sub attack type using define threshold (e.g. DoS, PROBE, U2R, R2L, Network attacks, Active Attack, Passive Attack, Advance attack etc)
The whole system consist three different phases as
A. Intrusion Detection System (IDS):-This phase execute the first GA and Fuzzy for features extraction for creating background rules, once background rules has created, testing phase has done with the help of Decision Tree (DT) for master and sub attack classification.
B. Intrusion Prevention System (IPS):-The prevention system work for prevent the known attacks which is already generated by remote sources. The system automatically block when any attack has generated. Here some pattern matching algorithms work for find the same network flow as well as packet signatures.
C. Intrusion Response System (IRS):-Basically it work for provide the security from different type of unknown attacks. The system holds ensemble modules for detecting malicious activity. Different classification approaches work for find such kind of unknown attacks. This module also provides the forensic features for creating the attacks log reports.
Algorithm:
Algorithm 1: GA (Genetic Algorithm) For Rules Creation [11]
Input: Set of network packet which consist 41 attributes with class label
Output: Set of normal as well intrusion rules
Step 1: Initialize randomly population with 41 Chromosomes.
Step 2: Initialize N (In the training set total number of records).
Step 3: The new population for each chromosome.
Step 4: Apply Crossover to best selected chromosome.
Step 5: Apply Mutation for each chromosome to new population.

Step 6: Calculate fitness= F(x) /sum (F(x)).
Step 7: Select best fit chromosome as 50 % and delete worse fit chromosome.
Step 8: End
Pseudo code for Fuzzy Algorithm [12]
Input: Test record with attribute value as from intrusion pool (data)
Output: Return record with attack type

```
for each record
{
        for each attribute
        {
                prob = fuzzy(attribute x);
                totalprob = totalprob + prob;
        }
        If (totalprob > threshold)
                class is attack;
         else
                class is normal;
  }
```

IV. DATASET DESCRIPTION

KDD cup 99 dataset [4] consist of essential records of the complete KDD data set. There are number of downloadable files at the disposal for the researchers.

Table 1: Dataset Description

Sr. No.	Name of the file	Description	Instances
1	KDDCUP 99 Train	The training dataset which contains 41 attributes including class labels	10000
2	KDDCUP 99 Test	The testing dataset which is around 23 attacks data excluding class labels	25000
3	NSKKDD Train	The training dataset which contains 41 attributes including class labels with multi variety attribute types	10000
4	NSKKDD Test	The testing dataset which is around 23 attacks data excluding class labels	25000

V. SET DEPENDENCY

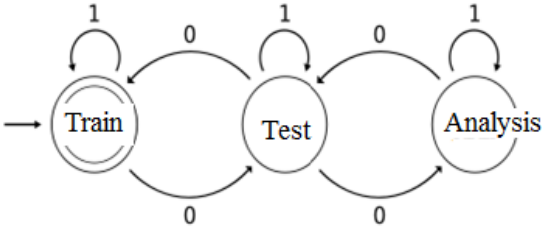

Fig. 2: State diagram for the system

System has define with 3 stages

Train: if training has successfully done then it returns 0 and system forward to test.
Test: Once test has done it will forward for analysis.
Analysis: Shows the result state
State => 1: Under execution state
0: Process successfully done state

System= {Train, Test, Analysis}
Train = {GA, Fuzzy, ARM}
GA = {Cross → Mutate → Fitness → Selection}
Fuzzy = {Probability, {0, 1}}
{GA → Fuzzy → ARM} → {0, 1}

Test = {PatternMatch, Th, Weight, Subclass}
Class = {Input → BkRules →Weigh} →{Noraml, Attack} →{subattacks}

Analysis = {DoS, probe, U2R, R2L, Normal, unknown}

VI. RESULTS AND DISCUSSIONS

To evaluate the proposed system performance analysis we have used various data sets for system testing which is already define in table 1. Each data set contains different features as well as different kind of attacks. Once the system has train according to specific data set, it generates training rule accordingly. The average accuracy for entire system with all data set is around 90%.

The below Fig. 3 carried out attack detection accuracy of proposed system in the testing phase. This experiment has done with KDDCUP99 and NSLKDD dataset. According to this graph DOS attack has highest accuracy than probe, U2R and R2L respectively. Various Thresholds have used to measure the accuracy level with incoming packet, each rule generate the similarity weight during the testing phase once any new packet has received into the victim port. After the validation of this experiment we conclude proposed system provides highest accuracy than other classical machine learning algorithms.

Table 2: Detection Rate based on rule desired rules

Attack	Accuracy	Error Rate
DOS	98.2	1.5
Probe	91.6	6.55
U2R	93.25	5.3
R2L	93.1	8.1

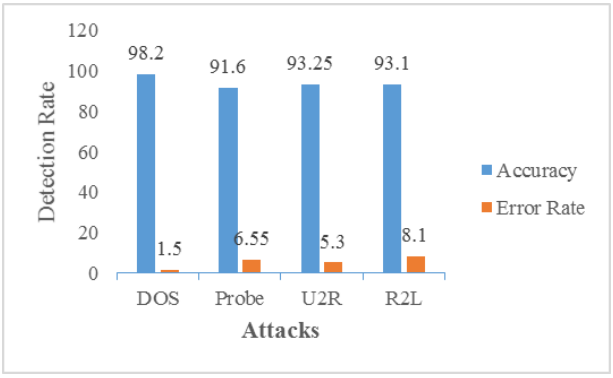

Fig. 3: Detection Rate based on rule desired rules

In Fig. 4 we have done the same experiment with different threshold values. The below figure shows how classification accuracy can be changed when different threshold values are resettled. According to this figure we conclude 0.50 is the optimal threshold to detect different known as well as unknown attacks in network environment.

Table 3: No. of unknown attack detection by system with various thresholds

Threshold	Accuracy	Error Rate
0.5	99.2	0.8
0.6	97.5	2.5
0.7	90.6	9.4
0.8	84.1	15.9

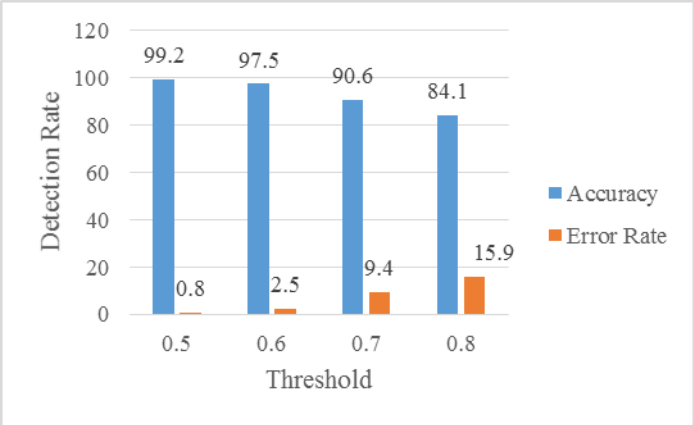

Fig. 4: No. of unknown attack detection by system with various thresholds

The below Fig.5 shows the number of unknowns packets successfully detected by IRS system during the testing phase of NIDS as well as HIDS respectively. IRS module continues updates the rule pool and generates new signatures.

Table 4: Average detection accuracy of proposed system

Samples	Detection Rate
5000	89
10000	87
15000	85
20000	84

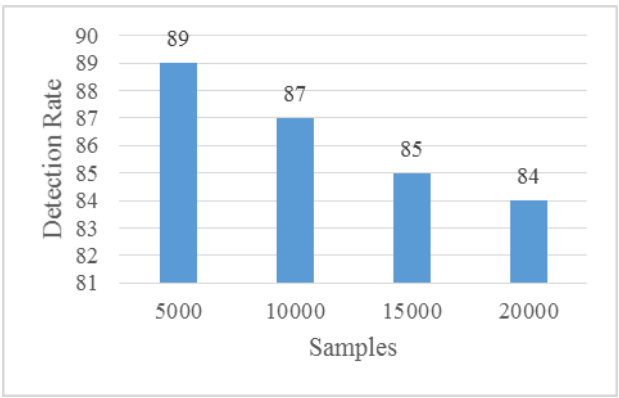

Fig. 5: Average detection accuracy of proposed system

According to overall experimental analysis we conclude that this work provides the security to the system from different type of unknown attacks. The system holds ensemble modules for detecting malicious activity. Different classification approaches work for finding such kind of unknown attacks. This system also provides the forensic features for creating the log report of an attack.

VII. CONCLUSION

Many experts are working on intrusion detection to gain strength in the security community roughly from last ten to twelve years. They have proposed number of distinct methodologies for intrusion detection to encounter this problem. Every developed Intrusion detection uses different sources to obtain data as well as uses specific techniques to analyze this data. Today after long research most systems separates information either by misuse detection or anomaly detection. Each approach has some advantages and relatively some set of restrictions. It is likely not practical to expect that an intrusion detection system be capable of correctly classifying every event that occurs on a given system. By considering all advantages and limitations it is difficult to gain perfect detection, like perfect security. It is very difficult to attain all security goals by considering intricacy and rapid appraisal or evolution of modern systems. After the completion of this survey we can conclude there are different techniques that can used for detection, some soft computing as well as some classification approaches are effective for detect the different attacks. Some systems have worked on signature base anomaly detection with creation of different rules. KDD cup dataset has used for training and testing purposed. Finally every system shows the maximum accuracy for attack detection, but none of these are has focused on unknown attack detection or misuse detection.

By considering all experimental analysis our system provides the security from different type of unknown attacks. Ensemble method uses majority technique for attack detection. Ensemble method with different classification algorithm performs better detection than individual algorithms.

REFERENCES

1. Panagiotis I. Radogloa-Grammatikis; Panagiotis G. Sargannidis, "Flow Anomaly Based Intrusion Detection System for Android Mobile Devices", 2017 6th International Conference on MOCAST, May 4-6, 2017, Kazani, Greece.
2. Md Zahangir Alom, Tarek m. Taha, "Network Intrusion Detection for Cyber Security on Neuromorphic Computing System", 2017 International Joint Conference on Neural Networks (IJCNN), May 14-15, 2017, USA.
3. Parisa Alaei, Fakhroddin Noorbehbahani," Incremental Anomaly-based IntrusionDetection System Using Limited Labeled Data", 2017 3th International Conference on Web Research (ICWR), IEEE.
4. Saad Mohamed Ali Mohamed Gadal, Rania A. Mokhtar, "Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique", 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE), Khartoum,Sudan , 2017 IEEE.

5. Trae Hurley, Jorge E. Perdomo, Alexander Perez-pons, "HMMBased Intrusion Detection System for Software-Defined Networking", 2016 15th IEEE Conference on Machine Learning and Application, Dec 18-20, 2016, Miami, Florida.
6. Mayank Agarwal, Sanketh Purwar, Santosh Biswas, Sukumar Nandi, "Internal Detection System for PS-Poll DOS attack in 802.11 networks using real-time discrete event system",IEEE, Vol.4, Issue4,2017.
7. Sharad Awatade, Shweta Joshi. "Improved EAACK: Develop Secure Intrusion Detection System for MANETS using Hybrid Cryptography", 2016 International Conference on computing communication control and automation (ICCUBEA), Aug 12-13, 2016, Maharashtra, India.
8. Alireza Shameli-Sendi, Habib Louafi, Wenbo He, "Dynamic Optimal CountermeasureSelection for Intrusion Response System", 2016 IEEE.
9. Mehdi Ezzarii, Hamid Elghazi, Hassan El Ghazi, Tayeb Sadiki, "Epigenetic Algorithm for Performing Intrusion Detection System", 2016 International Conference on ACOSIS, Oct17- 19, 2016, Rabat, Morocco.
10. Manoj s. Koli, Manik K. Chavan, "An Advanced Method for Detection of Botnet Traffic using Interhnal Intrusion Detection", 2017 International Conference on (ICICCT), March 10-11, 2017, Sangli, India.
11. Shengyi Pan, Thomas Morris, Uttam Adhikari, "Developing a Hybrid Intrusion Detection System using Data Mining for Power System", IEEE Transactions on, Vol. 6, Issues. 6, Nov. 2015.
12. Uzair Bashir, Manzoor Chachoo," Intrusion Detection and Prevention System:Challenges and Opportunities", 2014 IEEE.
13. Thabet Kacem, Duminda Wijesekera, Paulo Costa, Alexander Barreto, "An ADS-B Intrusion Detection System", 2016 IEEE on ISPA, 2016, Fairfax, Virginia.
14. Mariem Belhor, Farah Jemili, "Intrusion Detection Based on Genetic Fuzzy Classification System", 2016 IEEE 13th International Conference on Computer Systems and Application
15. Geethapriya Thamilarasu, Genetic Algorithm based Intrusion Detection System for Wireless Body Area Networks, IEEE 2015
16. P. Jongsuebsuk, N. Wattanapongsakorn, C. Real- Time Intrusion Detection with Fuzzy Genetic Algorithm, 978-1-4799-0545-4/13/ ©2013 IEEE

AUTHORS PROFILE

Ms.Sayali Kshirsagar is research student currently pursuing M.E. Computer Engineering from JSPM's Rajarshi Shahu College of Engineering. Author has published papers in international and national conferences.

Dr. P.B. Kumbharkar is presently working as Professor in Computer Engineering, Dean (Planning and Development) and IQAC CO-ordinator, Rajarshi Shahu College of Engineering Tathawade Pune. He also works as Adjunct faculty for BITS Pilani WILP program in Wipro Ltd. Hinjewadi Pune. He Published 25+ papers in various National, International conferences and journals. He works as a reviewer for various National and International journals.