International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Hybrid Intrusion Detection using Machine Learning for Wireless Sensor Networks

Revathi G K, Anjana S

Abstract—This Wireless sensor network (WSN) is a network of sensors, which is capable of communicating with each other and sensing some changes in parameters such as temperature, humidity etc. Such networks are beneficial in many fields, such as military industries, health monitoring, environmental tracking, monitoring of traffic. However, WSN's are easy to be attacked because of its properties such as untrusted broadcast transmission media, physical accessibility of sensors. So, protecting networks against attacks is one of most important issues in network and information security domain. As Sensor nodes have limited resources, authentication and encryption cannot be implemented directly to it. Hence, we propose a Hybrid Intrusion Detection System, which consists of Host Based Intrusion Detection system (HBIDS) and Network Intrusion Detection System (NIDS). In NIDS anomaly in network traffic, is detected. In HBIDS, patterns of misuse are detected from information collected at particular host or sensor. The main idea is to collect each sensor node's data and anomaly is detected in network and this detected intrusion is compared with signatures of attack in misuse detection system.

***Keywords:** Sensor data, Wireless Sensor Network, Hybrid IDS, anomaly-based detection, Signature based detection*

I. INTRODUCTION

WSN is made up of group of sensors which are dispersed spatially to monitor and collect the data about physical conditions of environment such as temperature, humidity, sound and so on.WSN contain hundreds of tiny low cost, low power sensor nodes, which perform their function in network. WSN are used in harsh conditions like battle field surveillance which is very difficult for human intervention. WSNs are susceptible to security attacks, as sensor nodes have limited bandwidth, memory and computational property. Attacks on WSN can be achieved by monitoring data between nodes and modifying them either by active or passive way [1] or by compromised nodes which have same capability as sensor being used [2]. Cryptographic methods like encryption and authentication, in which source of data is checked and verified if data was not altered.

But with this approach internal attacks cannot be detected, when attacker knows the key.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Revathi G K*, Department of Computer Science & Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka.
Anjana S, Department of Computer Science & Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka.

As preventive measure to stop attacks on WSN fails, Intrusion Detection System (IDS) to detect and report about the attacks on WSN is implemented. Hybrid IDS is combination of both Anomaly-based IDS and IDS based on Signature. Misuse-based (Signature-based) IDS can find attacks which are known, but it is not possible to detect unknown attacks [3] whereas Anomaly-based IDS detects unknown attack using Machine Learning approach. Normal profile of system is created in training phase and in testing phase it detects deviation from normal profile. As anomaly based system detects only unknown attack it has high false positive rate compared to Signature-based system. In this work Hybrid Intrusion Detection system for WSN with machine learning which consists of NIDS and HBIDS is proposed. First anomaly detection is implemented using Support Vector Machine (SVM)[4] on nodes and detect any deviation of data from normal profile and if there is deviation from normal profile, misuse based IDS is used to locate any known attack on host.

II. LITERATURE SURVEY

In literature various techniques have been proposed for intrusion detection system. This section provides a survey of existing technologies published in literature of anomaly and misuse detection system.

Hidoussi et al, proposed centralized IDS to detect attacks in cluster based WSN. Author has used base station which detects intrusion based on control packets sent from cluster head[5]. Misuse detection algorithm has been used to detect attacks in cluster based WSN. But using only misuse detection IDS is not enough for monitoring and reporting unknown attacks on WSN.

Rassam et al, proposed anomaly detection model which is based on single component classifier for detecting anomalies in sensor data collected by each node [6]. Two real life wireless sensor networks datasets are used and intrusions are detected with relatively lower false alarms. This approach detects anomalies only within training set, but for events that occur outside training set is not detected.

Yan et al, proposed an IDS which is consist of both anomaly detection module and misuse detection module created in cluster head [7].This hybrid IDS is applied to cluster WSN, where cluster head detects intrusion. Decision is done in Cluster-head, which is used to integrate the results and to detect attacks. But Cluster head is vulnerable to attacks by intruders. Shaikh et al, proposed trust management scheme for WSN [8]. It uses clustering to find anomalies in WSN. It requires less memory as trust calculation is based on group of nodes rather than trust values of individual node.

It detects and prevents malicious, selfish and faulty nodes. But power required is more as it depends on broadcasting messages to collect feedback from Cluster Heads.

Wang et al proposed IDS for cluster based WSN, which is an Integrated Intrusion Detection System which resists intrusions and process by analysing the attacks [9]. Three different IDS agents, Intelligent Hybrid IDS, Hybrid IDS and misuse IDS are proposed for sink, cluster head and sensor node respectively. But cluster head calculation becomes difficult since we have to define three IDS agents.

Ozceilk et al proposed an IDS for WSNs by combining the "signature based approach "and "trust based method". Each sensor node computes trust values for its neighbours. Base station detects malicious nodes by combining Signature based rules and trust values [10]. It detects malicious nodes in a base station, without much energy consumption. Proposed IDS detects if there is anomaly in network, but intrusion in host is not detected.

Barbara et al proposed a data mining method to detect intrusions. It uses both classification method and association rules for mining to detect attacks [11]. Anomaly detection detects suspicious data using association rules of mining. This data is passed to misuse detection algorithm to classify it as known attacks or normal data. Misuse detection algorithm must classify attacks as known or normal, but it fails to do so as misuse detection algorithm can detect only known attack.

Depren et al proposed parallel hybrid approach where misuse detection and anomaly detection algorithm runs in parallel. Self-organization map is used in anomaly detection and decision tree is used in misuse detection. Intelligent decision system is used to combine the results from the two detection algorithms [12].

Anderson et al, proposed statistical analysis and rule based model. Rule based model is implemented to detect known attacks and statistical model is implemented to detect outliers from connection that has been established from the data given [13]. With this method known and unknown attack detection rate is enhanced. However for anomaly detection high false positive rate remains as outliers are detected based on previous network connections.

Hwang et al, proposed signature based detection method followed by method to detect anomalies, to develop hybrid IDS [14]. False positive rates are high in anomaly detection. Since both systems are trained independent of each other, it results in raising many false alarms.

Kim and Lee has proposed a new hybrid IDS that integrates both misuse detection and anomaly based detection model. First data was divided into smaller subset using misuse detection model [15]. In second step SVM model was built to have precise behaviour from normal profile data NSL-KDD dataset was used to evaluate model.

Current research focuses on hybrid IDS for WSN data set where intrusion is detected in both network and host and on decreasing high false positive rate for anomaly detection. While in previous results anomaly and misuse detection algorithms were run parallel and combined at the end, proposed Anomaly and misuse detection algorithm are not run in parallel, so we can improve the profiling ability. Details of proposed system are presented in Methodology section.

III. METHODOLOGY

A. *System Architecture*

Proposed IDS model is hybrid in nature as it detects intrusion in both host and network by using anomaly and misuse detection engines as shown in Fig 1.

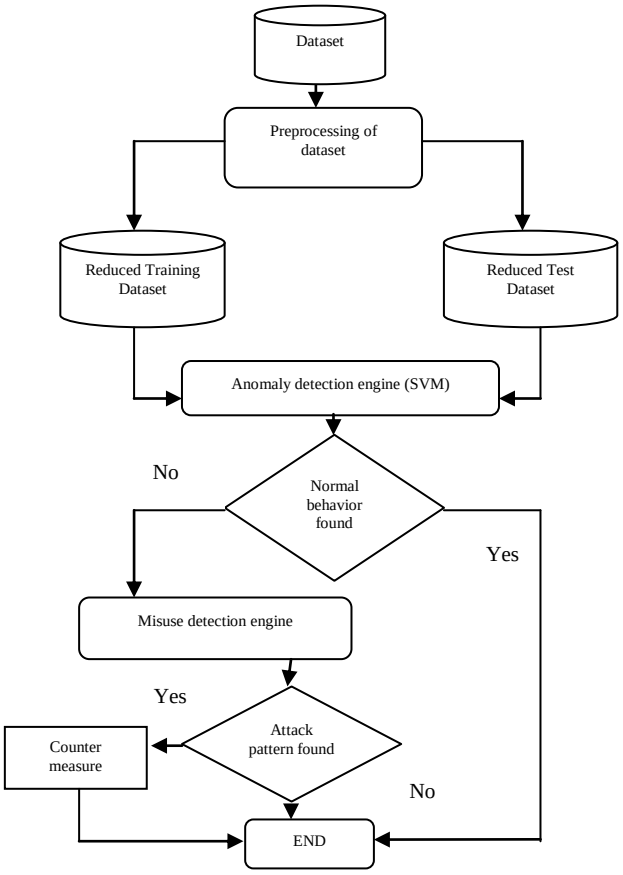

Fig 1. Hybrid intrusion detection model

In proposed Hybrid IDS model dataset is preprocessed and divided into trained dataset and test dataset as described in Fig 1. On this data Machine learning models are used to detect intrusion in the system for both NIDS and HBIDS. In the first step we preprocess the data and later we train the preprocessed data.

B. *Data Pre-processing*

Data gathering methods are loosely controlled and contain irrelevant, redundant information, noisy and unreliable data. Pre-processing for WSN dataset includes integration, cleaning, normalization, feature extraction. Data integration in WSN dataset contains data collected from 10 sensor nodes. Dataset has 18 attributes which is divided into 2 datasets, one for network and second for host. Dataset for network intrusion detection system include data mean value of received signal, standard deviation of received signal strength, mean value of link quality indicator (LQI), mean value of the noise floor, standard deviation of the noise floor standard deviation of link quality indicator, transmission rate, reception rate, mean value of the routing path length, standard deviation of the routing path length, estimated Packet Reception Ratio.

Signal to Noise ratio (SNR) is calculated using received signal and noise floor. Effective SNR can be obtained from LQI [16]. LQI is highly associated with error performance as it is a metric used to measure the error in modulation of packets received. Variation in received signal also increases when the number of hops increases. Distance is directly proportional to signal strength. Re-transmission will occur if received signal and noise floor are close to each other. SNR is expressed in decibels (dB).SNR with value 25 dB to 40 dB is said to be very good signal, SNR with value 15dB to 25dB is said to be Low signal, SNR with value 10dB to 15dB is classified as Very low signal and value 5dB to 10dB is indication of No signal. So, in data preprocessing for network we consider nodes with SNR ratio less than 10dB as outliers.

Dataset for host intrusion detection system include mean value of temperature on the node, standard deviation of temperature on the node, mean value of humidity on the node, standard deviation of humidity on the node, mean value of voltage level on the node, standard deviation of voltage level on the node. Temperature of range 17˚C to 25˚C is considered [17]. Voltage is in range 2 to 3 Volts. As temperature falls, humidity increases. Humidity is in range 0 to 100% which is temperature corrected relative humidity [18].It is inversely proportional to temperature. In the dataset if any value is not in range of the actual values, then it is predicted as an outlier.

On Pre-processed data we run Machine learning algorithm Support Vector Machine (SVM) [19] and naïve Bayes classifier [20]. Naive Bayes classifier is one which apply Baye's theorem with naïve assumptions between the features. Naïve Bayes classifier is multivariate categorical model. Training probabilistic model involves estimating parameters of distributions. Estimation of parameters of a Bernoulli distribution is done by counting number of success 's' in 'n' trials and setting Θ= s/n. Maximum likelihood is predicted as $argmax_z$P(X=x|Z=z).For 2 classes likelihood decision rules can be written as, predict positive if P(X=x| Z=⊕)/ P(X=x| Z= Θ)>$W^{\oplus}/W^{\theta}$, where W_i is the weight to predict $argmax_z$P(X=x| Z=z) which result in less possible loss in misclassified data.

SVM is a supervised learning model, which analyses the data using learning algorithms, data can be used for regression or classification.1 class SVM was motivated by work of [21] in general SVM classifier. Support vector machine defines margin for attack and normal data based on predictor assigned to set of labels. Margin is m/||w|| , where m is decision boundary and nearest training instance. SVM finds decision boundary that maximizes the margin m\||w|| True positive x_i has margin $w.x_i - t > 0$ and True negative x_j has margin $-(w.x_i - t) > 0$. Decision boundaries for intrusions and normal data are called support vectors.

SVM classifier algorithm performs better than Naïve Bayes algorithm as comparison result can be seen in Table 2, hence SVM is chosen as outlier detection algorithm. SVM defines a hyper plane which is used in classification of data among the classes where one of classes represents outlier. 1 class SVM is a binary SVM where training data belongs to first class and original data belong to second class as shown in Fig 2.

Algorithm

Input: Set of data with labels $X_{All} = \{X_1, X_2, \ldots, X_t\}$ collected from Sensor nodes S = $\{S_1, S_2, \ldots, S_n\}$

Output: Detected intrusion in data

Initialize: Integrate data from sensors S = $\{S_1, S_2 \ldots S_n\}$ as training data.

Step 1: Pre-process the input samples $labelled\ x_t \in X_{All}$, where t =1, 2,…. into D_r , where r ≤ t

Step 2: Classify input data as train data and test data with 75:25 ratio

Step 3: Apply anomaly detection (SVM) for the classified data D_r in homogenous coordinates. Select w, where w is weight vector which is used to define y' =w.x, function approximator.

w←0, t←0;
While t<T do
    for j = 1 to $|D_r|$ do
        $y' = w.x$;
        w ←w + $(y_i - y'_i)^2 x_i$;
    end
    t←t+1;
end

Step 4: F1 score is calculated for accuracy using equation,
F1 score = 2*((precision*recall)/ (precision + recall))

C. *Anomaly based Intrusion detection System*

In this model SVM algorithm is used for anomaly detection in both host and network. For SVM algorithm provided dataset is in form 75:25 where training data is 75% and testing data is 25 % of original dataset. With SVM algorithm confusion matrix is generated with which accuracy is calculated. Result generated by SVM algorithm is compared with results generated by WEKA tool. WEKA tool is a tool used for data mining tasks; it contains many machine learning algorithms. Classified data is preprocessed using WEKA tool using SMO (SMO implements algorithm for training support vector classifier using optimization algorithm), it normalizes all attributes and convert numeric to nominal and classify the data and calculate F1 score for generated confusion matrix. Analysis done by WEKA tool is used to compare results generated by Hybrid IDS.

D. *Misuse based Intrusion Detection Algorithm*

In this model, 2 types of Misuse detection techniques namely pattern matching technique[22] and expert rule based systems[23] are used.

Misuse Detection Algorithm in pattern matching technique utilizes knowledge base which contains known attack on sensor data. Remaining data is considered as normal behavior. Log collected from sensor nodes will act as knowledge base. Log contains observed abnormal behaviour in the system. Pattern matching techniques are simple, faster and require knowledge base to classify data.

In this model we run SVM algorithm for misuse detection in both host and network.

For SVM algorithm provided dataset contains log data as training data and testing data is actual data set. Result generated by SVM algorithm is compared with results generated by WEKA tool.

Rule based technique contains rules which are used to classify data. In Expert Systems, knowledge gained about attacks based on rules using if-then implications. Threshold is defined for each rule. Following are the rules defined for Host and Network Intrusion detection System[24].

Integrity Rule: Rule is defined to check data modifications.

Jamming Rule: Rule is defined to check if estimated packet reception ratio and actual number of packet reception ratio are comparable. This is also defined to detect communication noise on sensor nodes. Threshold value $TH_i < a_i / e_i$, where a_i is actual packet reception rate and e_i is estimated packet reception rate.

Interval Rule: Interval time between reception of 2 consecutive packet is considered. If interval time is longer or shorter than predefined time, then failure is introduced. Rules are applied to the data, to detect abnormal behavioral in the sensor nodes data.

E. Counter measures

If data does not satisfy integrity rule, there is possibility that sensor detected wrong data or it has been modified when transmitting. Verify data detected by sensor for average time, if correct value is detected in all other consecutive time, data is corrected else alarm is raised.

If interval time between packets reception is longer than expected (more than threshold defined), sensor may be attacked, replace the sensors.

IV. RESULTS AND ANALYSIS

Results obtained by SVM algorithm are validated by sensitivity, specificity and accuracy of classified algorithm obtained using confusion matrix. Confusion matrix is used to know the performance of classifier on test data. It consists of 4 values true positives, false negatives, false positives and true negatives in matrix as shown in Table 1. For accuracy F1 score is calculated. F1 score is sum of sensitivity and specificity divided by total set of instances taken to perform in the test from dataset. Specificity can be defined as set of negative instances that are true to be classified as negative. Sensitivity (Recall) can be defined as positive instances that are true to be classified as positive.

Precision = true positive value / true positive + false positive

Recall = True positive / true positive + false negative

F1 score = 2*((precision*recall)/ (precision + recall))

Table 1 Confusion matrix

	False (Predicted: NO)	**True (Predicted: YES)**
False (Actual: NO)	True Negative	False Positive
True (Actual: YES)	False Negative	True Positive

Table 2 Comparison of SVM & Naïve Bayes

Classifier	**SVM**	**Naive Bayes**
Accuracy	98.12 %	89.05 %

According to the results obtained we prove that the SVM classifier algorithm is more efficient than Naive Bayes classifier with a classification rate reaching 98.12%.

Table 3 Result of anomaly detection for host

Sensitivity	0.9911
Specificity	0.9752
Positive Pred Value	0.9882
Negative Pred Value	0.9813
Kappa	0.9678
F1 score Accuracy	0.9812

With WEKA tool confusion matrix can be generated, which evaluates the accuracy of the model. It is also called as error matrix. Accuracy for dataset of host was 97.6 % which is comparable with SVM accuracy which is 98.12% shown in Table 3. Kappa calculates level of agreement and compares it with value if 2 were independent.

Table 4 Result of anomaly detection for network

Sensitivity	0.4730
Specificity	0.9667
Positive Pred Value	0.8750
Negative Pred Value	0.7880
Kappa	0.4976
F1 score Accuracy	0.77

Confusion matrix generated by WEKA for dataset of anomaly detection for network accuracy generated is 74.5 % which is comparable with accuracy calculated by SVM algorithm which is 77% shown in table 4.

Table 5 Result of misuse detection for host

Sensitivity	0.99
Specificity	0.9652
Positive Pred Value	0.9872
Negative Pred Value	0.9713
Kappa	0.9878
F1 score Accuracy	0.9906

Confusion matrix generated by WEKA for dataset of misuse detection for host accuracy generated is 99 % which is comparable with accuracy calculated by SVM algorithm which is 98% shown in table 5.

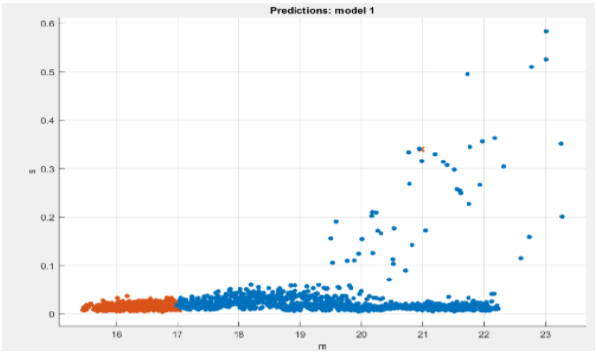

Fig 2. SVM classifier errors

V. CONCLUSION

Hybrid IDS for WSN has been proposed. We propose use of SVM algorithm in Hybrid IDS that is in both Anomaly and misuse-based host IDS and network IDS as it is more efficient compared to Naïve Bayes algorithm with 98% classification rate.

Weka tool has been used to classify and preprocess WSN data and the intrusion is detected in preprocessed data using proposed algorithm. Further research aims to apply the proposed algorithm to IOT data.

ACKNOWLEDGEMENTS

I would like to express my special thanks to my guide Anjana ma'am for her suggestions and guidance. I would also express my thanks to Department of Computer Science and Engineering for providing required resources.

REFERENCES

1. Padmavathi, D. Shanmugapriya, "A survey of attacks security mechanisms and challenges in wireless sensor networks", International J. Computer Science, vol. 4, no. 1, pp. 1-9, 2009.
2. H.K. Patil, S.A. Szygenda, "Security for wireless sensor networks using identity-based cryptography", Auerbach Publications, vol. 18, October 2012.
3. S. Rajasegarar, C. Leckie, and M. Palaniswami, "Detecting Data Anomalies in Wireless Sensor Networks", Security in Ad hoc and Sensor Network, World Scientific Publishing Co, Vol. 3, pp.231-259, 2009.
4. S. Kaplantzis, "Security Models for Wireless Sensor Networks", PhD Conversion Report, Centre of Telecommunications and Information Engineering, Monash University, Australia, 2006.
5. Hidoussi, H. Toral-Cruz, D.E. Boubiche, K.Lakhtaria, A. Mihovska,and M. Voznak, "Centralized IDS based on misuse detection for cluster-based wireless sensor networks", Wireless Personal Communications,vol. 85. No. 1, pp. 207–224, November 2015.
6. M.A Rassam, A. Zainal, M.A Maarof, "One-Class Principal Component Classifier for Anomaly Detection in Wireless Sensor Network", 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), pp. 271-276, 2012.
7. K. Q. Yan, S. C. Wang, S. S. Wang and C. W. Liu "Hybrid Intrusion Detection System for Enhancing the Security of a Cluster-based Wireless Sensor Network" In Proc. 3rd IEEE International Conference on Computer Science and Information Technology Chengdu China, pp.114-118,2010.
8. R.A. Shaikh, H. Jameel, B.J. d'Auriol, H. Lee, S. Lee, "Group-based trust management scheme for clustered wireless sensor networks", IEEE Trans. Parallel Distrib. Syst., vol. 20, no. 11, pp. 1698-1712, Nov. 2009.
9. SS. Wang, KQ. Yan, SC. Wang, CW. Liu, "An integrated intrusion detection system for cluster-based wireless sensor networks", Expert Systems with Applications, vol. 38, no. 12, pp. 15234-15243, 2011.
10. M.M. Ozcelik, E. Irmak., & S. Ozdemir," A hybrid trust-based intrusion detection system for wireless sensor networks". International Symposium on Networks, Computers and Communications (ISNCC), pp. 1-6,2017.
11. D. Barbara, J. Couto,S. Jajodia,L. Popyack & Wu, N, " ADAM: Detecting intrusions by data mining". IEEE Workshop on Information Assurance and Security. 2001.
12. O. Depren, M.Topallar,E. Anarim, & M.K. Ciliz, " An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks". Expert Systems with Applications,vol. 29,no. 4 pp. 713–722, 2005.
13. D. Anderson, T. Frivold, A. Valdes "Next-generation intrusion detection expert system (NIDES)", software users manual, beta-update release SRI International, Computer Science Laboratory(1995).
14. K. Hwang, Y. Chen, M. Qin, "Hybrid intrusion detection with weighted signature generation over anomalous Internet episodes". IEEE Transactions on Dependable and Secure Computing, vol.4 no. 1, pp. 41-55, 2007.
15. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," Expert Systems with Applications, vol. 41, no. 4. pp. 1690-1700, 2014.
16. F. Qin, X. Dai, & J.E Mitchell," Effective-SNR estimation for wireless sensor network using Kalman filter". Ad Hoc Networks, vol 11 no.3. pp. 944-958,2013.
17. S. Bhandari,N. Bergmann, R. Jurdak,B. Kusy," Time series data analysis of wireless sensor network measurements of temperature" Sensors 17, no. 6,pp. 1221, 2017.
18. Wu. Shaomin, D. Clements-Croome "Understanding the indoor environment through mining sensory data-A case study",2006.
19. H.J. Shin, D.H. Eom, S.S. Kim "One-class support vector machines: An application in machine fault detection and classification" Computers & Industrial Engineering, vol. 48, no. 2, pp. 395-408,2005.
20. S. Mukherjee,N. Sharma ," Intrusion detection using naive Bayes classifier with feature reduction", Procedia Technology, vol 4, pp. 119-28,2012.
21. B. Schölkopf, J.C. Platt, J. ShaweTaylor, A.J.Smola, R.C. Williamson" Estimating the support of a high-dimensional distribution**",** Neural Computation, vol. 13, no.7, pp. 1443-1471,2001.
22. Liao, H.J., Lin, C.H.R., Lin, Y.C. and Tung, K.Y " Intrusion detection system: A comprehensive review". Journal of Network and Computer Applications, 36(1), pp.16-24, 2013.
23. Bansal, B. and Singh, K.," Rule Based Intrusion Detection System to Identify Attacking Behaviour and Severity of Attacks." International Journal of Advanced Research in Computer Science and Software Engineering, 5(1), 2015.
24. Deshmukh, R., Deshmukh, R. and Sharma, M, "Rule-based and cluster-based intrusion detection technique for wireless sensor network". Int. J. Comput. Sci. Mobile Comput., 2(6), pp.1-9, 2013.

AUTHORS PROFILE

Revathi G K, Department of Computer Science & Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka.

Anjana S, Department of Computer Science & Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka.