A Hybrid High Relevant Low Redundant Feature Selection Method For The Classification of Nodes in Mobile Ad Hoc Networks

T. Dheepak, Research Scholar, Department of Computer Science, Research & Development Centre, Bharathiar University-Coimbatore, Tamil Nadu, India. dheepak.me@gmail.com

Dr.S. Neduncheliyan, Principal, Jaya College of Engineering & Technology, Chennai, Tamilnadu, India. nedun@yahoo.com

Abstract - Intrusion detection systems (IDS) in MANET have to system blocks of packets with many features, which suspend the detecting of anomalies. Sampling and Feature Selection may be employed to depreciate computing time and hence diminishing the time of intrusion detection. A Novel Hybrid High Relevant Low Redundant Feature Selection (HRLR-FS) method mounts on Particle Swarm Optimization (PSO) and Chi-Square (CS) analysis. The execution of the proposed Hybrid HRLR-FS on DARPA dataset to decrease the volume of primary features and accurate by implementing better detection performance in the classification methods relating with other feature selectors. The relevant features and removing redundant features of DARPA dataset is Optimal Dataset. ANN with Multi-Layered Perception classification method was used to classify the nodes of MANET

Keywords — MANET, IDS, Malicious Node, Legitimate Node, High Relevant Low Redundant Feature Selection, Chi-Square analysis, Particle Swarm Optimization.

I. INTRODUCTION

A mobile ad hoc network (MANET) is an auto-configuring network that is designed by an aggregate of mobile nodes without the guidance of a fixed infrastructure or centralized administration [1]. Every node has implemented with a wireless transmitter and receiver, which enable it to interact with another node in its radio communication range[2]. For each node to broadcast a packet to a node that is furthest of its radio range, the collaboration of another node in the network is required; it is called as multi-hop communication [3]. Therefore, every node must serve as both a router and host at the equal time. The network topology periodically varies because of the movement of mobile nodes as they move out, move into or move inside of the network [4]. A MANET with the features defined above was initially developed for military purposes, as nodes have disseminated across an arena and there is no infrastructure to help them form a network [5]. In modern years, MANETs have been improving quickly and are frequently being utilized in various applications, varying from civilian to military and commercial uses, since fixing up such networks can be done without the help of interaction with a human or any infrastructure with a human

[6]. Some cases are data collection, virtual classrooms, conferences and search-and-rescue missions where PDA, laptops or other mobile gadgets experience wireless medium and interact with each other [7]. As MANETs become broadly utilized, the security problem has converted one of the main concerns. For example, a MANETs consider that each node in the network is not malicious and cooperative [8]. Hence, just one compromised node can create the collapse of the whole system.

The ad-hoc wireless networks are intended to be scalable. As the system expands, many routing protocols have proposed. An essential criterion of the protocol scalability is its routing overhead. It was described as the entire amount of routing packets broadcasted over the network, represented in packets/second or bits/seconds. Mobile nodes have coated with power restrictions, and as such, power preservation is a significant factor to examine in the MANET implementation. Furthermore, network size, radio power constraints and channel utilization have investigated. These circumstances restrict the energy of nodes to communicate directly among the destination and source in a MANET to communicate directly. As the amount of nodes developments in the network, the connection between the target and the source more relies on intermediate nodes.

 DOI : 10.18231/2454-9150.2018.0609

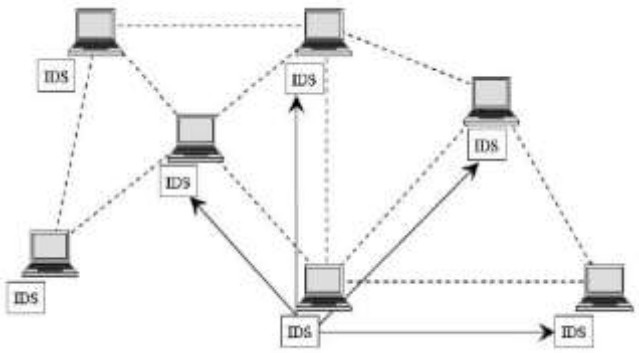

Fig:1 IDS system working in MANET

II. RELATED WORKS

Benjie Chen [9] proposed multi-hop Ad hoc wireless networks which can reduce the level of energy consumption without any specific capability or connective path in the MANET. Each node calculates the number of standby nodes and the capacity of energy level saved for the function. The level of energy consumption is, and therefore network topology is used the saving mode of the span on the top of the MAC power can enhance the packet delivery ratio and throughput.

Krishna Kumar [10] He has proposed power control MAC protocol which helps to calculate power consumptions at various times of transmission. This method shows an excellent result in power consumption of node transmission in multiple times, and the range of communication has also increased. The source node continuous with the strategy by perceiving the neighbor node to find the parameter transmission rate and its capability too. Mesh topology improves the speed of transmission by reducing the distributed power supply to each node. Every time data are transmitted, the power distribution and data rate of each node calculated and monitored in the level by all nodes equally. The main reason is that the nodes are interconnected with mesh topology so that the energy level has maximized.

Binh Hy Dang [11] He describes the cluster technique which helps to evaluate the efficiency of methodologies in MANET. Methods static and dynamic profiles require the most inventive usages of supportive nodes. This method helps to prove whether the disruptive route attack has occurred in the end-to-end communication. This type of scheme supervises unauthorized nodes and track out the reliability of packets in the networks. Many times there happens route invasion attack by incorporating an unauthorized node in the communication network.

Pragya, Mishra, K. V. Arya, and Singh Hardev Pal [12], in this work, IDS algorithm has been introduced that can effectively isolate and detect conspiring nodes of the network so that specific malicious nodes do not influence the execution of the system. Suggested detection method utilizes in-out traffic data and catching statistics of nodes to distinguish conspiring attackers.

Bouhaddi, Myria, Mohammed Saïd Radjef, and Kamel Adi [13], the proposed system enables the reinforcing of network security while protecting the resources of the network nodes (e.g., energy). Further, the suggested explication gives MANETs with a distributed IDS in which activation has performed according to the determined threat. The recommended method in this paper, distinguished by the application of game theory for simulating communications among a possibly malicious sender node and a defense combination node.

III. PROBLEM IDENTIFICATION

Mobile Ad Hoc Network has the following issues due to attacks, intrusion and energy management.

- All system in the network have one to one mapping between the physical and logical identity of the node, but a malicious node violates this and introduce fake logical identity in the network.
- Malicious nodes effect routing of network and creates a false path in a network and disrupt network operation
- Large network nodes are not ready to validate and verify the identity of another node in the system.
- In MANET, energy determines the lifetime of MAC protocol. The node power consumption should minimize as their energy resource is limited MAC protocol should not handle an essential volume of information. The transmission area of the node is defined to maintain energy

IV. PROPOSED FRAMEWORK FOR THE CLASSIFICATION OF MALICIOUS NODE

In this framework, the next two phases are required to build the efficient malicious node identification method. DARPA dataset has utilized to analyze the behavior of the malicious node.

Phase 1: Pre-Processing Phase: The importance of pre-processing is to decrease the dimension of the dataset. It is employed to remove the redundant and irrelevant features and to enhance the classification efficiency of the attacks in the networks. In this stage, a new hybrid feature selection method has proposed. This hybrid method combines the technique of Chi-Square analysis and Particle Swarm Optimization (PSO) technique. These two techniques are hybridized to give the most appropriate features and low redundant for the classification of a node in the network.

Phase 2: Classification phase: In this stage, the ANN classification method is applied to separate the node to their behavior in the network. This method uses a MLP-NNfor classification.

 DOI : 10.18231/2454-9150.2018.0609

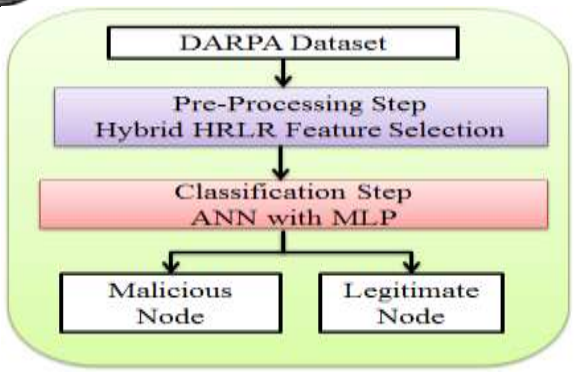

Figure 2: Proposed Framework for Classification of Malicious Node in MANET

1. Pre-Processing Step: Hybrid High Relevant Low Redundant (Hrlr) Feature Selection Method

(a) Chi-Square (CS) Analysis Feature Selection

Feature Selection via chi-square χ^2 test [14] is another, very commonly applied method. CS attribute evaluation estimates the goodness of a feature by measuring the significance of the chi-squared statistic to the class. The first hypothesis H_0 is the premise that the two features are irrelevant, and a chi-squared formula tests it:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{O_{ij} - E_{ij}}{E_{ij}} \right)^2 \quad (4.1)$$

where O_{ij} is the observed frequency, and E_{ij} is the expected (theoretical) frequency, affirmed by the null hypothesis [17]. The higher the value of χ^2, the higher the proof upon the hypothesis H_0.

(b) Particle Swarm Optimization (PSO)

PSO [15][16] was depend on the social behavior connected with bird's assembling for the optimization problem. A social behavior model of organisms that interact and live with big crowds is the motivation for PSO. The PSO is more accessible to put into action than Genetic Algorithm. It is for the motive that PSO does not have a crossover or mutation operators and flow of particles has effected by using velocity function. In PSO, each particle changes its flying memory and its partner's flying inclusion following in mind the top goal of flying in the search space with velocity.

(c) Hybrid High Relevant Low Redundant Feature Selection Method

This method has proposed with the combination PSOand CS analysis.

Then the classification accuracy of the particle is examined with the Decision Feature. If the accuracy of the feature has a value prominent than the Decision Feature, then the feature is collected in T and passed to empty set R. Now, the R has an attribute with higher classification accuracy. Then the velocity is calculated and updated by applying the equation

$$U_{ud} = U_{ud} + C1i1(p_{jd} - y_{jd}) + C2i2(p_{gd} - y_{jd})$$
$$Y_{jd} = Y_{jd} + U_{ud}$$

This procedure iterated until the classification accuracy of the Similarity Uncertainty SU based on decision feature Df is equal to the classification accuracy of conditional feature based on Df. The final output is the optimal data set.

Term	Description
ω	Inertia
DeF	Decision Features
CoF	Conditional Feature
s&c	Social and Cognitive Components
SS	Swarm Size
DFV	Data Fitness Value
p_{best}	Particle Best Position
G_{best}	Global Best Position
R	Null Set
T	Feature Set
U_{ud}	A velocity of the Particle
y_{jd}	Individual Best position
r1 &r2	Random Values

Method The pseudo code for the Hybrid HRLR Feature Selection

Input: DARPA dataset

Output: Optimal Dataset

Step 1: *Initialize Conditional Features(CoF) and Decision Feature (DoF)by using CS analysis.*

Step 2: *Initialize SS = 20, s = 2.0, c=2.0, r1,r2=0.2 and* .ω= 0.33

Step 3:*For each particle r in SS do*

Step 3.1*: Set R ← { }*

Step 3.2*: Set T ← R*

Step 3.3*: Initialize ∀ x ε (CoF – R) after storing the empty data set in T, the conditional features have checked with an empty set*

Table 1: Algorithm terms and its description

Step 3.4*: Calculate Data Fitness Value using Griewangk function*

Step 4:*If DFV > P_{best}*

Step 4.1: *Set P_{best} = DFV*

Step 5:*If P_{best} > G_{best}*

Step 5.1*: Set P_{best} = G_{best}*

Step 6:*if $_{RU}(x) > \gamma_T$(DoF)*

Step 6.1*: Set T ← $R_U(X)$*

Step 6.2*: R ← T*

Step 7: *Calculate and update the velocity by using the equation*

$$U_{ud} = U_{ud} + C1i1(p_{jd} - y_{jd}) + C2i2(p_{gd} - y_{jd})$$
$$Y_{jd} = Y_{jd} + U_{ud}$$

 DOI : 10.18231/2454-9150.2018.0609

Step 8:Repeat the process from step 3 until $\gamma_R(DoF) == \gamma_C(DoF)$ is satisfied.

***Step 8.1**: If $\gamma_R(DoF) == \gamma_C(DoF)$*

***Step 8.2**: Return R.*

2 Classification Step: Artificial Neural Network

ANN is an effective calculating system whose principal theme has acquired from the resemblance of biological NN. ANNs has also described as "Parallel Distributed Processing Systems." ANN obtains a massive number of units that are interrelated in some pattern to enable connection among the units. Those units also mentioned to as neurons or nodes are mere CPUs which operate in parallel. In this work, the ANN has adopted for classification of Intrusion Detection in the system. The following method depicts the steps of the MLP-NN training algorithm.

***Step 1:** Initialize Bias, Learning rate α, weights, to begin the training of Multi-Layered Perceptron Neural Network.For simplicity and calculation, need to set weight =0 and bias $\alpha = 1$.*

***Step 2:** Proceed step 3-8 at the terminating condition is true.*

***Step 3:** Proceed step 4-6 for all training vector a.*

***Step 4:**Initiate each input as follows:*

$$r_j = s_j \ (j = 1 \ to \ m)$$

***Step 5:** Get the net input with the next relations*

$$s_{jn} = b + \sum_{j}^{m} r_j w_{jk}$$

Here bias is given as b, and the whole amount of input neuron is given by 'n'.

***Step 6:**Apply the activation function to obtain the final output for each input*

unit k=1 to n

$$f(s_{jm}) = \begin{cases} 1 & if \ s_{jmk} > \theta \\ 0 \ if - \theta \leq s_{jmk} \leq \theta \\ -1 \ if \ s_{jmk} < -\theta \end{cases}$$

***Step 7:** Adjust the weight and bias for r=1 to m and k=1 to n as follows:*

Step 7.1: Case 1: if $s_k \neq t_k$the m

$$w_{jk}(new) = w_{jk}(old) + \propto t_k r_j$$

$$s_k(new) = s_k(old) + \propto t_k$$

Step 7.2: Case 2: if $s_k = t_k$ then

$$w_{jk}(new) = w_{jk}(old)$$

$$s_k(new) = s_k(old)$$

Here's' is the exact output, and't' is the desired/target output.

***Step 8:** Testing for terminating condition, which will occur while there is no variation in weight.*

3. Implementation Result and Discussion

(a) Dataset Description

The DARPA dataset used for the classification of nodes in this paper [8] [9] [17]. The DARPA dataset is composed of 43 features.

(b) Result and Discussion on Feature Selection Method

The aim of making a feature selection is to reduce the redundant features from the dataset. In this paper, 31 features have selected by using Chi-Square analysis, 29 features are selected by PSO whereas proposed Hybrid HRLR-FS method gives 27 features. Table 2 and figure 1 gave the result obtained from Chi-Square analysis, Particle Swarm Optimization and proposed Hybrid HRLR-FS method.

S .No	Chi-Square analysis	PSO	Proposed Hybrid HRLR-FS method
1	CS_Src_bytes	PSO_logged_in	HRLR_Service
2	CS_Service	PSO_num_5ell	HRLR_srv_Count
3	CS_srv_Count	PSO_dst_host_same_src_port_rate	HRLR_count
4	CS_count	PSO_dst_host_diff_srv_rate	HRLR_dst_host_srv_count
5	CS_dst_host_srv_count	PSO_Protocol_type	HRLR_dst_host_same_src_port_rate
6	CS_dst_host_same_src_port_rate	PSO_dst_host_same_srv_rate	HRLR_dst_host_diff_srv_rate
7	CS_Dst_bytes	PSO_dst_host_srv_serror_rate	HRLR_dst_host_same_srv_rate
8	CS_dst_host_diff_srv_rate	PSO_num_access_fil	HRLR_dst_host_rerror_rate
9	CS_dst_host_same_srv_rate	PSO_srv_Count	HRLR_dst_host_srv_rerror_rate
10	CS_dst_host_rerror_rate	PSO_num_compromised	HRLR_dst_host_srv_diff_host_rate
11	CS_dst_host_srv_rerror_rate	PSO_num_root	HRLR_diff_srv_rate
12	CS_dst_host_count	PSO_srv_serror_rate	HRLR_Flag
13	CS_dst_host_srv_diff_host_rate	PSO_rerror_rate	HRLR_same_srv_rate
14	CS_diff_srv_rate	PSO_srv_diff_host_rate	HRLR_srv_diff_host_rate
15	CS_Flag	PSO_hot	HRLR_rerror_rate
16	CS_same_srv_rate	PSO_dst_host_srv_diff_host_rate	HRLR_srv_rerror_rate
17	CS_srv_diff_host_rate	PSO_is_guest_32	HRLR_dst_host_serror_rate
18	CS_rerror_rate	PSO_dst_host_srv_count	HRLR_logged_in
19	CS_srv_rerror_rate	PSO_Flag	HRLR_dst_host_srv_serror_rate
20	CS_dst_host_serror_rate	PSO_su_attempted	HRLR_serror_rate
21	CS_logged_in	PSO_Root_5ell	HRLR_srv_serror_rate
22	CS_duration	PSO_dst_host_rerror_rate	HRLR_hot
23	CS_dst_host_srv_serror_rate	PSO_Wrong_fragment	HRLR_Wrong_fragment
24	CS_serror_rate	PSO_diff_srv_rate	HRLR_num_failed_32s
25	CS_srv_serror_rate	PSO_serror_rate	HRLR_num_compromised

 DOI : 10.18231/2454-9150.2018.0609

26	CS_hot	PSO_dst_host_serror_rate	HRLR_is_guest_32
27	CS_Wrong_fragment	PSO_srv_rerror_rate	HRLR_land
28	CS_num_failed_32s	PSO_num_file_creations	
29	CS_num_compromised	PSO_dst_host_srv_rerror_rate	
30	CS_is_guest_32		
31	CS_land		

Table 2: Result obtained by using Chi-Square analysis, PSO and Proposed Hybrid HRLR-FS method

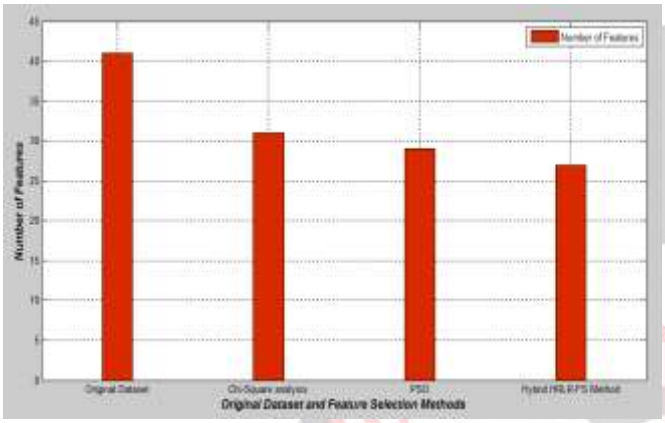

Figure 2: Performance analysis of the Original dataset, Chi-Square analysis, Particle Swarm Optimization and proposed Hybrid HRLR-FS method

(c) Result and Discussion on Classification Methods

Mostly the classification methods are utilized to evaluate the effectiveness of the results obtained from the feature selection techniques. In this paper, the classification methods like ANN and NB have used. By using ANN classification method. Table 3 depicts the proposed Hybrid HRLR-FS method gives better classification accuracy, Kappa Statistic value, reduced error rates like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Root Relative Squared Error (RRSE) and Relative Root Absolute Error (RRAE) than the existing methods like Chi-Square Analysis and Particle Swarm Optimization. Table 3: Performance analysis Chi-Square analysis, PSO and Proposed Hybrid HRLR-FS method by using Artificial Neural Network

Performance Metrics	Original Dataset	*Feature Selection Methods*		
		Chi-Square analysis	Particle Swarm Optimization	Proposed Hybrid HRLR-FS
Classification Accuracy	69.3333 %	94.6667 %	92 %	98 %
Kappa Statistic	0.5539	0.7133	0.7099	0.7594
MAE	0.0411	0.0934	0.0874	0.0788
RMSE	0.1317	0.1671	0.1326	0.1311
RAE	91.7907 %	96.1844 %	90.0778 %	73.7135 %
RRSE	89.9791 %	78.975 %	78.8312 %	60.0193 %
TP Rate	0.693	0.947	0.92	0.98
FP Rate	0.15	0.294	0.307	0.355
Precision	0.553	0.923	0.853	0.98
Recall	0.693	0.947	0.92	0.98
F-measure	0.601	0.933	0.885	0.978
ROC Area	0.847	0.859	0.93	1

From the table 4, it is clear that the proposed Hybrid HRLR-FS method gives the better result than the existing in all the aspects. However, when comparing the classification methods, ANN performs better than Naïve Bayes classification methods for the proposed Hybrid HRLR-FS method. Figure 2 presents the graphical representation of the performance analysis of the classification methods.

Table 3: Performance analysis Chi-Square analysis, PSO and Proposed Hybrid HRLR-FS method by using Artificial Neural Network

Performance Metrics	Original Dataset	*Feature Selection Methods*		
		Chi-Square analysis	Particle Swarm Optimization	Proposed Hybrid HRLR-FS
Classification Accuracy	65.3333 %	75.3333 %	71.3333 %	92.6667 %
Kappa Statistic	0.4664	0.5083	0.4935	0.7432
MAE	0.0198	0.0292	0.0225	0.0128
RMSE	0.1354	0.157	0.1448	0.1077
RAE	44.2956 %	42.273 %	39.4545 %	20.0334 %
RRSE	92.471 %	87.1663 %	88.0308 %	64.0312 %
TP Rate	0.653	0.753	0.713	0.927
FP Rate	0.22	0.26	0.247	0.251
Precision	0.494	0.647	0.549	0.865
Recall	0.653	0.753	0.713	0.927
F-Measure	0.545	0.691	0.619	0.894
ROC Area	0.8	0.822	0.859	0.943

Table 4: Performance analysis Chi-Square analysis, PSO and Proposed Hybrid HRLR-FS method by using Naïve Bayes Classification Method

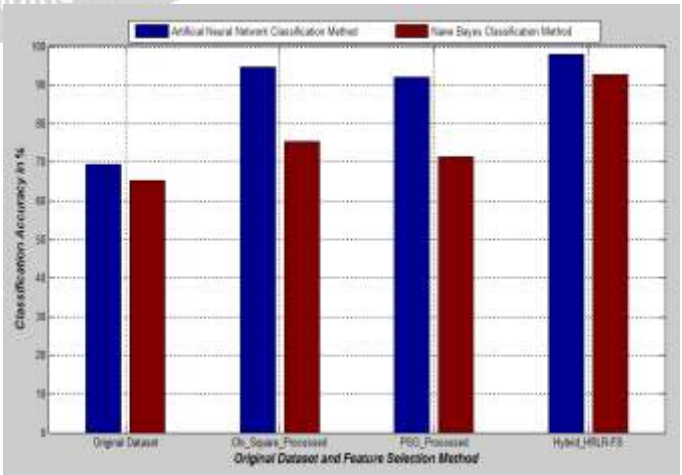

Figure 2: Performance analysis of the ANN and NB classification methods for Original Dataset, Chi-Square analysis, PSO and proposed Hybrid HRLR-FS method

 DOI : 10.18231/2454-9150.2018.0609

V. CONCLUSION

This proposed technique has used to improve the prognostication accuracy. In this work, a novel Hybrid HRLR-FS method has introduced by combining the PSO and CS analysis. This method was presented to eliminate the unnecessary feature for the classification in the IDS dataset. From the obtained results it has determined that the proposed technique worked better than the present feature selection methods in the IDS. Also, it enhances the prognostication accuracy and diminishes the error rates. This diminishing of error rates results in the excellent classification accuracy.

ACKNOWLEDGMENT

Dr. S. Neduncheliyan is known for his excellent guidance and supported this work. The author is so thankful for arranging MAT LAB, Department of Computer Science, Jaya College of Engineering & Technology, Chennai, Tamilnadu. This work could not have been accomplished without the necessary facilities avail-able in the lab. The author would like to thank the technical personnel

REFERENCES

[1] R. Akbani, Korkmaz, T., Raju, G. V. S: Mobile ad hoc network security. In: Lecture Notes in Electrical Engineering, Springer, vol. 127 (2012)

[2] T. Anantvalee and Wu, J.: A survey on intrusion detection in mobile ad hoc networks. In: Wireless/Mobile Security. New York: Springer (2008)

[3] Elhadi, M., Shakshuki, EAACK.: A secure intrusion-detection system for MANETs. In: IEEE Transactions on Industrial Electronics, vol. 60(3) (2013)

[4] Gungor, V.C., Hancke, G.P.: Industrial wireless sensor networks: challenges, design principles, and technical approach. IEEE Trans. Ind. Electron. 56(10), 4258–4265 (2009)

[5] Haldar, N.A.H.: An activity pattern based wireless intrusion detection system. In: Information Technology: pp. 846–847 (2012)

[6] Shen, J.: Network intrusion detection by artificial immune system, IECON, pp. 716–720 (2011)

[7] Khattab, S., Gabriel, S., Melhem, R., Mosse, D.: Live baiting for service-level DoS attackers. Proceeding of the IEEE INFOCOM (2008)

[8] Macia´-Pe´rez, F.: Network intrusion detection system embedded on a smart sensor, industrial electronics. IEEE Trans. 58(3), 722– 732 (2012).

[9] Benjie Chen, Kyle Jamieson, Hari Balakrishnan And Robert Morris" An Energy-Efficient Coordination Algorithm for Topology Maintenance in Ad Hoc Wireless Networks," in Proceedings of the wireless network, 2002.

[10] Krishan Kumar, "power control with transition time for the ad-hoc wireless network," International Journal of Advanced Research in Computer Science, Volume 3, No. 3, May-June 2012.

[11] Binh Hy Dang, Wei Li, "Impact of Baseline Profile on Intrusion Detection in Mobile Ad Hoc Networks," in Proceedings of the IEEE SoutheastCon 2015, April 9 - 12, 2015 – Fort Lauderdale, Florida

[12] Pragya, Mishra, K. V. Arya, and Singh Hardev Pal. "Intrusion Detection System against Colluding Misbehavior in MANETs." Wireless Personal Communications 100.2 (2018): 491-503.

[13] Bouhaddi, Myria, Mohammed Saïd Radjef, and Kamel Adi. "An efficient intrusion detection in resource-constrained mobile ad-hoc networks." Computers & Security 76 (2018): 156-177.

[14] Thaseen I.S., Kumar C.A, "Intrusion Detection Model Using Chi-Square Feature Selection and Modified Naïve Bayes Classifier," Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC – 16').Smart Innovation, Systems and Technologies, Vol. 49.Springer, pp 81-91,2016.

[15] Abualigah, Laith Mohammad, and Ahamad Tajudin Khader. "Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering." *The Journal of Supercomputing* 73.11 (2017): 4773-4795.

[16] Gu, Shenkai, Ran Cheng, and Yaochu Jin. "Feature selection for high-dimensional classification using a competitive swarm optimizer." *Soft Computing* 22.3 (2018): 811-822.

[17] S. Latha, Sinthu Janita Prakash, "HPFSM – A High Pertinent Feature Selection Mechanism for Intrusion Detection System", *International Journal of Pure and Applied Mathematics*118.9 (2018): 77-83.

 DOI : 10.18231/2454-9150.2018.0609