UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Comparative Study on Feature Selection Techniques in Intrusion Detection Systems using Ensemble Classifiers

Bello Nazifi Kagara & Maheyzah Md Siraj
Faculty of Computing,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
Email: nazefballo@gmail.com, maheyzah@utm.my

Submitted: 22/12/2020. Revised edition: 4/4/2021. Accepted: 6/4/2021. Published online: 24/4/2021
DOI: https://doi.org/ 10.11113/ijic.v11n1.286

***Abstract*—Network usage has become a paramount aspect of life, therefore, securing our networks is crucial. The world is experiencing a rapid breakthrough of internet usage, most especially with the concept of internet of things (IoT), now internet of everything (IoE.). Real network data is rowdy, noisy and inconsistent. These issues with the data influences the performance of intrusion detection systems (IDS) and develop manifold of false alarms. Feature selection technique is used to remove the inconsistent and rowdy data from a large data set and presents a refined set of data. This research work adopts the use of two distinct feature selection technique in parallel: ReliefF ranking and particle swarm optimization, using linear discriminant analysis (LDA) and logistic regression (LR) as the machine learners, to first clean the data, train the classifiers, and subsequently classify new instances. The results showed that, the combination of the ReliefF with the ensemble machine learning (Linear Discriminant Analysis and Logistic Regression) has a higher classification accuracy of 99.7% compared to the Particle swarm optimization (PSO) which attained an accuracy of 98.6%.**

***Keywords*—Machine learning, particle swamp optimization, relief ranking, linear discriminant analysis, logistic regression**

I. INTRODUCTION

Network security has been one of the underlying issues for decades and various types of built systems are being introduced. Maintenance of network security is one of the leading security concerns to neutralize any undesired activities. It is not only intended to protect information and privacy issues on the network, but also to avoid dangerous situations. Microsoft Security Intelligence Report from January to June 2010 indicates that infection rates continue to rise at a higher rate around the world on average [1]. A network is said to be intruded, when there is an unauthorized access to the network, which could lead to the loss of sensitive information and could lead to the unavailability of the entire network. The security system must therefore be consistent and well-configured [2]

Intrusion can be considered as a series of activities that try to challenge resources accessibility, discretion, or integrity. Using intrusion detection usually requires monitoring of significant incidents taking place in a system and then evaluating them for potential device intrusions. A meticulous intrusion detection description can be overviewed as a team of error detection mechanisms, procedures and practices that could potentially lead to security failure by detecting and diagnosing anomalies and signature based attacks and intrusions [3]. It can also be added that, an IDS is a realistic application of frameworks and theory for intrusion detection in the network [4]. This method consists of combining software and hardware components running on a host machine tracking the actions of users and programs to the host for outsider threats. An IDS's goal is to send alerts for suspicious events to administrators and attempt to curb the attacks. The principles used in IDS vary from other methods of security, including access control, encryption, or firewalls to secure a system. Having emphasized on this, nevertheless, such techniques of safety are highly recommended for the simultaneous stabilization of the defense of a network and the protection of a broader range [5].

This paper consists of five major sections, its starts with an introduction followed by related works, then succinctly lays the

proposed methodology, then we look at the experiment results and discussions, and lastly conclusion, future works and references.

II. RELATED WORKS

IDS is an integral grown spectrum in dealing with computer systems and networks. Hence, various researchers have developed variety of IDS depending on the intended goal. Some systems are developed by combining weaker machine learners to create a stronger one, while others entail combining multiple feature selection techniques, in order to reduce computational complexity. Feature selection techniques have been and are still adopted widely for the refinement of data sets, which help machine learners and classifiers results more efficient and accurate. For the purpose of this paper, we looked at related works peculiar to feature selection and machine learning.

A. Works on Feature Selection and Ensemble Classification

In an attempt to develop an IDS that is efficient and has low false positives, [6], developed and IDS whose target was to improve the efficiency, with the use of the NSL-KDD dataset, KDD-CUP, and CIC-IDS 2017. The methodology adopted by this work, was a combination of three feature selection techniques in order to have a robust and improved classification efficiency, and the ensemble classifier includes also three techniques based on AOP combination rule. The conclusion of this project shows that Accuracy and precision has greatly improved. Another ensemble model was proposed by Paulaukas and Aukalnis (2017) with the employed the use of four distinct classifiers, J48, C5.0, Naïve Bayes and PART, with focus on putting together poorer classifiers to compose richer ones. Their result of the ensemble model achieved more accurate results in the Intrusion detection system. The scientists proposed in [7] an SVM-based intrusion detection program that includes the hierarchical clustering algorithm, simple selection process for features, and a SVM technique. There have been fewer, abstract and higher qualified training instances of the hierarchical clustering algorithm, extracted from the KDD-Cup 1999 training set. It could considerably reduce the practice time, and also enhance the success of the resulting SVM. The straightforward feature selection technique was administered to exclude irrelevant features in the training set in order to enable the obtained SVM model to interpret the network traffic information more precisely. The famous data set was used to test the proposed system for the 1999 KDD Cup. This program showed better performance in detecting DoS and Probe attacks and the best overall accuracy performance compared to other intrusion detection systems centred on the same dataset [7]

III. PROPOSED METHODOLOGY

In order to succinctly draw explicit facts and conclusions about the two widely used feature selection techniques; Filter and wrapper approach. We selected an algorithm of each; Relief ranking and particle swarm optimization respectively, to develop an intrusion detection system using Linear discriminant analysis and Logistic regression as the ensemble machine learners and classifiers. The Figure 1 below shows the framework of the proposed hybridized machine learning based ids.

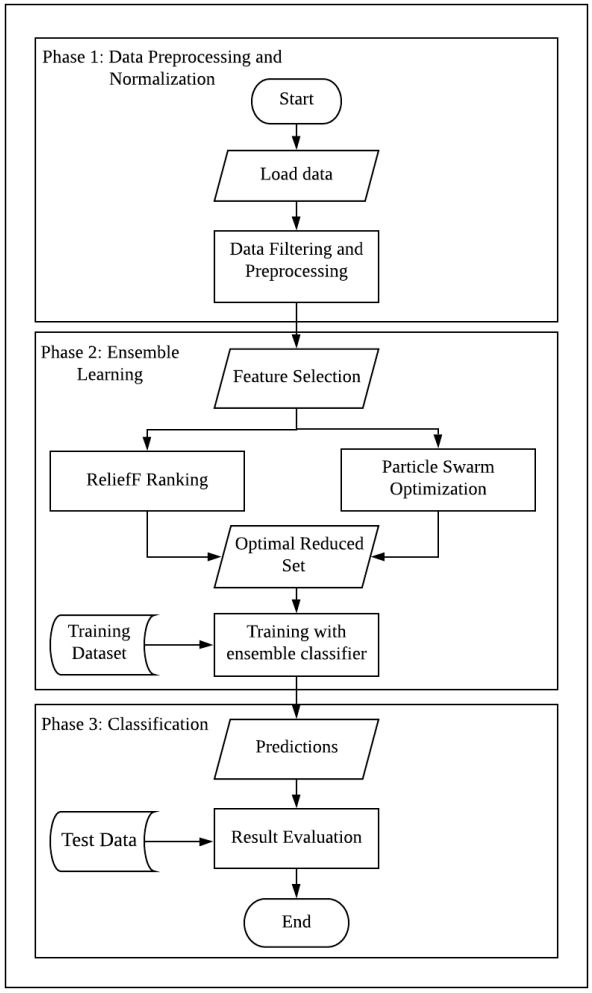

Fig. 1. Framework of the proposed methodology

The framework has the following main phases:

1. Dataset Normalization: This is useful in classification. Generally, for classifying algorithms. It involves the process of scaling the attributes of data, so that it falls withing a smaller acceptable range. Normalization is generally required when dealing with attributes on different scales. This is an integral part considering the real-life network data being used which contain noisy inconsistent and rowdy data. The normalization method adopted for this research work is decimal scaling.

2. Variable subset selection: The feature selection phase was carried out by both algorithms; ReliefF ranking and particle swarm optimization. In order to obtain the refined subset data. The refined data obtained from each of the techniques will be passed to the ensemble classifier for training.
3. Classification: This stage employs the implementation of bagging ensemble by sci-kit learning. The ensemble algorithm is a combination of linear discriminant analysis and logistic regression in order to improve the classifying accuracy.

IV. RESULTS AND DISCUSSIONS

The feature selection stage was both carried out by the ReliefF ranking Algorithm and the Particle Swamp Optimization, Filter and Wrapper approach respectively. The selected features are given below for both cases.

A. Particle Swamp Optimization Selection

The particle swamp optimization worked by the collection of individuals called particles, moving in steps through a region. At each step, the algorithm evaluates the objective function at each particle. After the evaluation, the algorithm decides on the new velocity of each particle and picks the optimal feature. The mean of the data was obtained and used as the objective function of the PSO.

Objective function = EFX/EF, where X = Dataset

The PSO algorithm was able to minimize and selects the best subset features to 18, as shown in Table 1.

TABLE 1. Selected features of the PSO

No.	Feature name
21	count
14	root_shell
7	land
18	num_shells
39	dst_host_srv_rerror_rate
28	diff_srv_rate
35	dst_host_rsv_diff_host_rate
13	num_compromised
29	srv_diff_host_rate
26	srv_rerror_rate
8	wrong fragment
11	num_failed_logins
15	su_attempted
20	is_guest_login
6	dst_bytes
19	num_access_files
17	num_file_creations
10	hot

B. ReliefF Ranking

The reliefF computes ranks and weights of attributes for the input data matrix and response vector, the ReliefF filter selection method was able to rank the predicting variables with respect to the class label in accordance with their respective weight score. The features subsets, which was totals to 22 features.

TABLE 2. Selected features of the ReliefF ranking

Selected Features	Ranking Scores
3	0.0074
30	0.0147
31	0.0602
34	0.0056
33	6.55E-04
36	0.0148
32	0.0021
35	0.0038
38	0.0039
14	0.0125
39	0.0066
28	0.0122
37	9.47E-04
29	0.0214
6	0.0034
2	9.47E-04
25	0.0037
10	0.0034
12	0.0036
27	0.0078
20	-0.003
1	0.0019

C. Model Evaluation

The different statistical criteria will be used to measure the model's effectiveness in terms of power and precision of prediction. An example is the true positive rate, the false-negative rate, the false positive rate, the accuracy of the classifier, the precision, and the recall. Find the equations below:

i. True Negative (TN): Number of correctly forecasted cases as non-attacks.
ii. False Negative (FN): Number of cases wrongly forecasted as non-attacks.
iii. False Positive (FP): Number of cases wrongly forecasted as attacks.
iv. True Positive (TP): Number of correctly forecasted cases as attacks.

1. **Sensitivity/TPR** = TP/(TP+FN)
2. **Precision** = TP/(TP+FP)
3. **Accuracy** = (TP+TN)/ total number of classified item = (TP+TN) / (TP+TN+FP+FN)
4. **False positive rate** = 1 – Specificity
5. **Recall/sensitivity** = TP/(TP=FN)

Fig 2. Equations for model evaluation

D. Experimental Results Evaluation.

The experimental results are listed based on the classification algorithm. The evaluation parameter shows the result of the Ensemble classifier obtained for both cases. The testing (probing) evaluation was achieved using the True Positive rate (TP), False Positive (FP), True Negative (TN) and, False Negative (FN), accuracy and error rate as well. The evaluation parameters for classification rate that were achieved are, classification Accuracy, sensitivity, Specificity and Error Rate.

E. Ensemble Classification of the ReleifF selected features

The table 3 below shows the analysis per each class based on the class label from the Normal, Dos, Probe, U2R and R2L attack group. The table highlights the true positive value, the true negative value, false positive value and false negative value of each of the class groups.

a) Analysis Per class

TABLE 3. Analysis per class

Analysis per class.	True Positive	True Negative	False Positive	False Negative
Class 1	3346	2919	16	17
Class 2	2303	3974	15	6
Class 3	562	5720	6	10
Class 4	1	6296	0	1
Class 5	47	6244	2	5

b) Confusion Matrix

Confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class based on the testing set of data. The class 1 represents the normal class which gives a total of 3363 from the test observation set, a total of 3346 was classified correctly and 17 was misclassified, class DOS represented by label 2 gives a total of 2309 from the test observation set, a total of was 2303 classified correctly and 6 was misclassified, class PROBE represented by label 3 gives a total of 571 from the test observation set, a total of was 562 classified correctly and 10 was misclassified, class U2R represented by label 4 gives a total of 2 from the test observation set, a total of was 1classified correctly and 1 was misclassified, lastly class R2L represented by label 5 gives a total of 52 from the test observation set, a total of was 47 classified correctly and 5 was misclassified

TABLE 4. Confusion Matrix

Confusion Matrix					
	1	2	3	4	5
1	3346	12	4	0	1
2	4	2303	2	0	0
3	6	3	562	0	1
4	1	0	0	1	0
5	5	0	0	0	47

c) Evaluation Parameters for Classification Phase

The table 5 shows the evaluation parameters of the Ensemble+ReleifF based on the f-score, specificity, sensitivity, accuracy and error rate.

TABLE 5. Evaluation parameter for classification phase

Technique	F-score	SPECIFICITY	SENSITIVITY	ACCUARACY (%)	ERROR RATE (%)
Ensemble Classifier +ReleifF	0.914776	0.997884	0.875742	99.7523	0.00247698

d) Result of System Computational Time

The actual computational time used in processing the Ensemble Classifier for training the dataset is taken, which is

measured in terms of the total seconds use time for executing the training process. The result is shown below.

TABLE 6. Training time

Technique	**Training Time(Secs)**
Ensemble+ReliefF	43.1904

F. Ensemble Classification of the PSO selected features

The Table 7 shows the analysis per each class based on the class label from the Normal, Dos, Probe, U2R and R2L attack group. The table highlights the true positive value, the true negative value, false positive value and false negative value of each of the class groups.

a) Analysis per class

TABLE 7. Analysis per class

Analysis per class.	**True Positive**	**True Negative**	**False Positive**	**False Negative**
Class 1	3312	2818	117	51
Class 2	2272	3914	76	36
Class 3	462	5708	17	111
Class 4	1	6296	0	1
Class 5	34	6239	7	18

b) Confusion Matrix

Confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class based on the testing set of data. The class 1 represents the normal class which gives a total of 3363 from the test observation set, a total of 3312 was classified correctly and 49 was misclassified, class DOS represented by label 2 gives a total of 2408 from the test observation set, a total of 2372 was classified correctly and 36 was misclassified, class PROBE represented by label 3 gives a total of 574 from the test observation set, a total of 462 was classified correctly and 112 was misclassified, class U2R represented by label 4 gives a total of 2 from the test observation set, a total of was 1classified correctly and 1 was misclassified, lastly class R2L represented by label 5 gives a total of 52 from the test observation set, a total of was 34 classified correctly and 18 was misclassified.

TABLE 8. Confusion Matrix

Confusion Matrix					
	1	2	3	4	5
1	3312	30	14	0	7
2	33	2372	3	0	0
3	65	46	462	0	1
4	1	0	0	1	0
5	18	0	0	0	34

c) Evaluation Parameters for Classification Phase

The Table 9 shows the evaluation parameters of the Ensemble classifier + PSO based on the f-score, specificity, sensitivity, accuracy and error rate.

TABLE 9. Evaluation Parameters for Classification Phase

Technique	**F-score**	**SPECIFICITY**	**SENSITIVITY**	**ACCUARACY (%)**	**ERROR RATE (%)**
Ensemble Classifier+PSO	0.845477	0.9874	0.787873	98.6218	0.00137822

d) Result of System Computational Time

The actual computational time used in processing the Ensemble Classifier for training the dataset is taken, which is measured in terms of the total seconds use time for executing the training process. The result is shown below.

TABLE 10. Training Time

Technique	**Training Time(Secs)**
Ensemble+PSO	31.2383

G. Graphical Analysis

The graphical analysis shows a comparative result of the training time, classification accuracy, specificity, sensitivity, error rate and f-score of both the filter and wrapper selection

technique when passed into the ensemble classification algorithm.

1. Result Analysis for Training Time

The training time shows the time taken by the model to create knowledge retention of the data supplied to the Ensemble Classifier for both feature selection cases. Fig. 3 shows that the PSO+ Ensemble Classification has a better optimal time than its counterpart case.

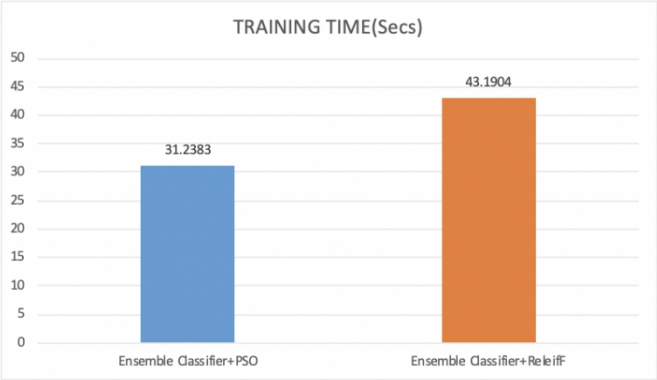

Fig. 3. Training time

2. Result for Classification Accuracy

The classification accuracy shows the correct classification rate attained by the Ensemble Classifier for both cases. The classification accuracy in percentage shows the percentage of instances that were classified correctly. The classification accuracy results shows the Ensemble Classifier+ReleifF Ranking as more accurate than the Ensemble+PSO has an accuracy of 99.7523% and 98.6218% respectively.

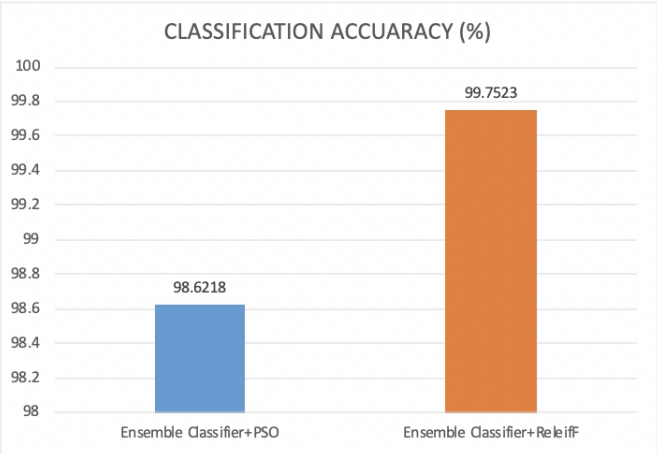

Fig. 4. Classification Accuracy

3. Result of Comparative Analysis for Error rate

The error rate shows the lowest possible error rate for any classifier in a random outcome during the classification. The Ensemble Classifier shows the lowest error rate of 0.00247698 at Ensemble Classification +ReliefF, which is pointing to the fact that the Ensemble classifier shows a very high positive rate detection at Ensemble Classification +ReliefF case.

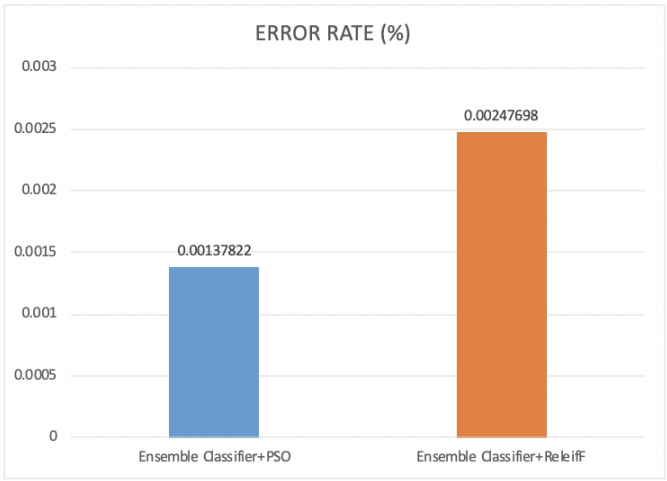

Fig. 5. Error Rate

4. Specificity and Sensitivity

The Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives, Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. The best sensitivity and specificity falls at 1. From the obtained results shows the sensitivity and the specificity rate has value close 1 indicating a good predictive rate. The Ensemble Classifier+ReleifF Selection proved more better than its counterpart as its specificity and sensitivity rate are closer to 1 than Ensemble Classifier + PSO.

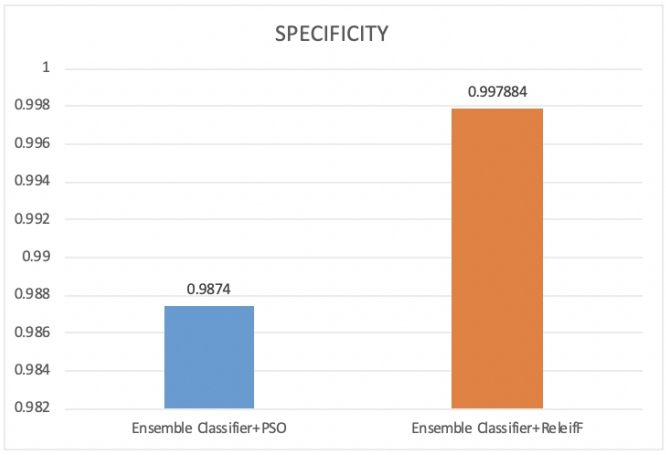

Fig.6. Specificity

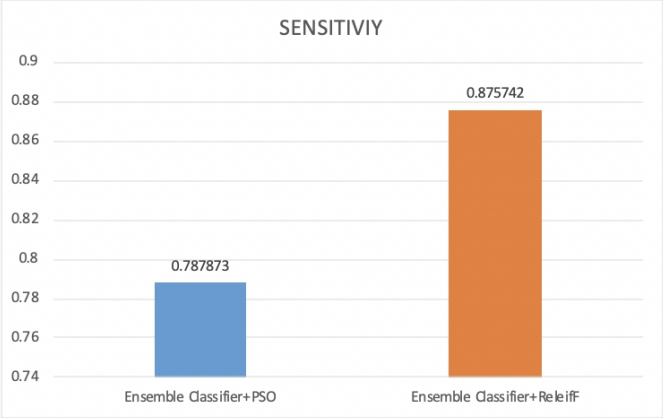

Fig.7. Sensitivity

V. CONCLUSION

In this research work, a hybrid method and comparative approach was used to achieve high success prediction for intrusion detection, the model followed a filtering, feature selection and classification technique of a data mining process, examining both filter and wrapper techniques for feature selection. The performance rate of the Ensemble Classifier + ReliefF gives a higher classification accuracy of 99.7523% as compared with the Ensemble Classifier + PSO, which was able to attain an accuracy of 98.6218%. For this case study, the ReliefF ranking feature selection technique is ascertained to be better than the PSO wrapper approach and can therefore be recommended for intrusion detection problems in determining the most predominant factor that helps in predicting Normal and Attacks in Intrusion detection systems.

ACKNOWLEDGMENT

I would first like to thank my thesis supervisor Dr. Maheyzah Md Siraj, of school of computing, faculty of engineering, Universiti Teknologi Malaysia who has helped me immensely by always steering me in the right direction.

I would also like to thank the experts, my examiners who were involved in the validation for this research project: Dr. Mohd. Foad and Dr. Ismail Fauzi, for their passionate participation and input.

Finally, I must express my very profound gratitude to my parents for providing me with an unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have ever been possible without them. Thank you.

REFERENCES

[1] G. M. D. Batchelder, J. Blackbird, P. Henry. (2014). Microsoft Security Intelligence Report-Volume 17, Microsoft Secure. *Intell. Rep.,* 16, 2014.

[2] M. Kagara, Nazifi & Md Siraj. (2020). A Review on Network Intrusion Detection System Using Machine Learning. *International Journal of Innovative Computing.*

[3] S. A. Abhaya, K. Kumar, R. Jha and S. Afroz. (2014). Data Mining Techniques for Intrusion Detection: A Review. *Int. J. Adv. Res. Comput. Commun. Eng.*, 3, 6938-6941.

[4] D. A. Prasanna P., RaghavRamana A. V. T, Kumar R. K. (2012). Network Programming and Mining Classifier for Intrusion Detection Using Probability Classification. *Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering.*

[5] M. A. A. K. Chabathula, K., C. D. Jaidhar. (2015). Comparative Study of Principal Component Analysis based Intrusion Detection approach using Machine Learning Algorithms. *Signal Processing Communication and Networking, 3rd Int. Conf.*

[6] D. K. Li Y., Xia J., Zhang S., Yan J., Ai X. (2012). An Efficient Intrusion Detection System Based on Support Vector Machines and Gradually Feature Removal Method. *Expert Syst. with Appl.* 39: 424-430.

[7] C. D. Horng, S. J., Su, M. Y., Chen, Y. H., Kao, T. W., Chen, R. J., Lai, J. L., Perkasa. (2011). Anovel Intrusion Detection System Based on Hierarchical Clustering and Support Vector Machines. *Expert Syst. Appl.,* 38: 306-313.