IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 27, No. 1, February, 2022, pp. 151-156
Journal homepage: http://iieta.org/journals/isi

Effective Intrusion Detection System Using Classifier Ensembles

Muthukumarasamy Govindarajan

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar 608002, Tamil Nadu, India

Corresponding Author Email: govind_aucse@yahoo.com

https://doi.org/10.18280/isi.270118

Received: 16 December 2021
Accepted: 12 January 2022

Keywords:
accuracy, arcing, bagging, ensemble, heterogeneous, homogeneous, support vector machine, radial basis function

ABSTRACT

The problem of network intrusion detection poses innumerable challenges to the research community, industry, and commercial sectors. Moreover, the persistent attacks occurring on the cyber-threat landscape compel researchers to devise robust approaches in order to address the recurring problem. Given the presence of huge web traffic, standard machine learning approaches are rather inefficient if adapted in network intrusion detection areas. Instead, a hybrid multiple classifier model when attempted enhances the performance henceforth leading to valid predictions. Thus, novel ensemble approaches are presented in this research work that involves bagged homogeneous classifier ensembles and arcing of heterogeneous ensembles. Then the classification performances of classifier models are assessed using accuracy. Here, classifier ensemble is built using base classifiers such as RBF and SVM. The feasibleness and the advantages of the proposed approaches are illustrated with the help of existing intrusion detection dataset. pre-processing phase, classification phase and combining phase are the three major phases of this proposed method. A broad series of analogous experiments are done for standard dataset of intrusion detection. Furthermore, comparisons with previous work on standard dataset of intrusion detection are also exhibited. The experimental outcomes demonstrate that this proposed ensemble approaches are competitive.

1. INTRODUCTION

With the growing reliance on the Internet, Network Intrusion Detection system (NIDs) becomes a vital part of cyber safety system. NIDs focus at discriminating the web traffic as normal and abnormal cases. Designing an intellectual and effectual intrusion detection system with good detection rates and low false-alarm rates is essential to face the variety of network practices and the fast increase of intrusion approaches. The main advancement in machine learning in recent years is the ensemble method that develops both accurate as well as diverse classifiers combining their results so that the resulting classifier surpasses the other single base classifiers.

Many researchers have built models to determine machine learning classifiers and to classify intrusion data set using BBN, ANN, SVM etc. [1-3]. Most intrusion detection systems (IDSs) mostly use a single classifier algorithm to classify the network traffic data as normal behaviour or anomalous. However, these single classifier systems fail to provide the best possible attack detection rate with low false alarm rate. In this paper, ensemble approaches are proposed using combination of classifiers in order to make the decision intelligently, so that the overall performance of the resultant model is enhanced [4-6]. The contributions of the paper are as follows:

(i) Homogeneous ensemble classifier and heterogeneous ensemble classifier are built with bagging and arcing respectively and their classification accuracy are estimated.
(ii) SVM and RBF are used as base classifiers to design classifier ensemble model. The main innovation of this proposed technique covers three major phases: pre-processing phase, classification phase and combining phase.
(iii) The classification performance of homogeneous and heterogeneous classifier models are compared with that of base classifiers upon intrusion detection data.
(iv) In comparison with the single classifiers, the proposed ensemble classifiers exhibit remarkable accuracy enhancement. Moreover, heterogeneous classifiers are found to perform better than homogeneous models.
(v) Furthermore, comparisons with previous research on existing intrusion detection dataset are also enlisted and are found that the proposed ensemble methods are competitive.

The rest of the article is framed as follows. Section 2 discusses the previous works done. Section 3 describes the novel technique proposed in this work. Section 4 depicts the classification performance as well as evaluation measures. Section 5 and 6 deals with the results and conclusion.

2. RELATED WORK

A lot of research is done in the field of intrusion detection where many techniques are covered and still many remains to be covered.

Fossaceca et al. [7] present the innovative Extreme Learning Machine with Multiple Adaptive Reduced Kernel. Multiple Classification Reduced Kernel ELM and Multiple Kernel Boosting are combined in this work. Experiments on intrusion detection data show that MARK-ELM outruns other

approaches exhibiting higher detection accuracy.

Aburomman et al. [8] proposes a new ensemble construction approach which applies PSO generated weights to generate classifier ensemble with better network intrusions detection performance. Better behavioral parameters are discovered for PSO by adapting local unimodal sampling as meta-optimizer. For this empirical study, five random subsets of the popular KDD99 intrusion detection training data is used. Novel technique and weighted majority algorithm are applied to generate the classifier ensembles. It is found that superior classification accuracy is achieved with the proposed method than weighted majority algorithm in creating ensemble classifiers.

Based on well-known machine learning approaches, Aburomman et al. [9] give a brief overview of intrusion detection algorithms. In particular, several ensemble techniques and hybrid approaches were studied along with the homogeneous and heterogeneous ensemble approaches. Furthermore, voting based ensemble techniques that can be easily applied and will generate beneficial results were also considered. A recent literature survey reveals that hybrid techniques, in which a single classifier is fused with feature selection or reduction component has been quite common. Hence, the scope of this study has been expanded to encompass hybrid classifiers.

Divyasree & Sherly [10] proposed an efficient IDS using Ensemble Core Vector Machine (CVM) approach that detects Probe attack, DoS attack, U2R attack and R2L attack. A core vector machine classifier is built for each kind of attack. The classifiers are trained and tested using KDD-99 dataset. In this approach the appropriate features are selected for each attack using Chi-square test and dimensionality reduction is performed by applying a weighted function to these selected features.

Salo et al. [11] presented a novel hybrid approach for dimensionality reduction in which the Principal Component Analysis and Information Gain approaches are combined for intrusion detection with instance based learning classification algorithms, multilayer perceptron and a support vector machine based classifier ensemble. In this work, the detection rate of the proposed ensemble model is determined using the most popular NSL-KDD, ISCX 2012 and Kyoto 2006+ datasets were used to estimate.

Kunal & Dua [12] constructed an ensemble model using Random Tree, REP Tree, IBk (K-NN), j48graft and Random Forest classifiers where number of attributes were reduced and was evaluated by ranker-based attribute evaluation technique.

Several methods have been put forth by researchers to perform network intrusion detection using a combination of algorithms. This paper presents combination of SVM and RBF as base classifiers to build a hybrid system in order to improve the overall performance and produce classifiers with better accuracy compared to prior research work. Various experiments were performed on NSLKDD data to estimate robustness of proposed bagged classifiers and hybrid system. It is found that the heterogeneous models outrun homogeneous models for NSL-KDD dataset [13].

3. PROPOSED METHODOLOGY

3.1 Preprocessing

In preprocessing of dataset, cleaning and transformation are performed. Cleaning process means removing the redundant labels and filling missing value in the dataset. Transformation means translate full data set into the desired form (it means convert numeric value to the string type data).

3.2 Existing classification methods

3.2.1 Radial basis function neural network

This is an artificial neural network formulated by Broomhead & Lowe [14]. RBF uses radial basis functions for activation to change along the distance from a location. For functional approximation, it uses time-series prediction, classification, and system control. A multi-layer feed forward neural network, RBF is used to classify data in a non-linear mode and compare input data with training data. The production of the RBF neural network is weighted linear superposition of all basis functions. The frequently used basis function in the RBF model is the Gaussian basis function.

3.2.2 Support vector machine

This is widely used for training SVMs and was formulated by Platt [15]. SMO is one way to solve a quadratic programming (QP) issue that arises during SVM training. SMO divides the large QP problem into a series of very tiny sub-problems. These small sub-problems are solved analytically, preventing the use of time-consuming numerical QP optimization as an inner loop. It is the fastest for linear SVMs and sparse datasets and can be more than 1000 times faster than the chunking algorithm. The amount of memory needed for SMO is linear in the training dataset size, allowing SMO to handle very large training sets. It scales somewhere between linear and quadratic in the training set size for several test problems.

3.3 Homogeneous ensemble classifiers

3.3.1 Proposed bagged RBF and SVM classifiers

Given a set D, of d tuples, bagging [16] works as follows. For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. The bootstrap sample, D_i, created by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeatedly or not at all in any particular replicate training data set D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The bagged RBF and SVM, M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: RBF and SVM ensemble classifiers using bagging

Input:

- D, a set of d tuples.
- $k = 2$, the number of models in the ensemble.
- Base Classifiers (Radial Basis Function, Support Vector Machine).

Output: Bagged RBF and SVM, M^*

Method:

(1) for $i = 1$ to k do // create k models;

(2) Create a bootstrap sample, D_i, by sampling D with replacement, from the given training data set D repeatedly. Each example in the given

training set D may appear repeated times or not at all in any particular replicate training data set D_i;

(3) Use D_i to derive a model, M_i;

(4) Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i, based on the accuracies of percentage of correctly classified example in training data D_i.

(5) endfor

To use the bagged RBF and SVM models on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value.

3.4 Heterogeneous ensemble classifiers

3.4.1 Proposed RBF-SVM hybrid system

Given a set D, of d tuples, arcing [17] works as follows; For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. some of the examples from the dataset *D* will occur more than once in the training dataset D_i. The examples that did not make it into the training dataset end up forming the test dataset. Then a classifier model, M_i, is learned for each training examples *d* from training dataset D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The hybrid classifier (RBF-SVM), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Hybrid RBF-SVM using Arcing Classifier
Input:

- D, a set of d tuples.
- $k = 2$, the number of models in the ensemble.
- Base Classifiers (Radial Basis Function, Support Vector Machine).

Output: Hybrid RBF-SVM model, M^*.

Procedure:

1. For i = 1 to k do // Create k models,
2. Create a new training dataset, D_i, by sampling D with replacement. Same example from given dataset *D* may occur more than once in the training dataset D_i.
3. Use D_i to derive a model, M_i,
4. Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i, based on the accuracies of percentage of correctly classified example in training data D_i.
5. endfor

To use the hybrid model on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value;

The basic idea in Arcing is like bagging, but some of the original tuples of D may not be included in Di, whereas others may occur more than once.

4. CRITERIA FOR EVALUATION

4.1 Cross validation technique

Cross-validation is a statistical technique which involves partitioning the data into subsets, training the data on a subset and use the other subset to evaluate the model's performance.

4.2 Criteria for evaluation

Accuracy is one of the essential measures for describing the performance of any algorithm. It is the degree to which an algorithm can properly predict positive and negative instances, and it can be determined by the following formula: Accuracy = TP + TN/TP + FN + FP + TN.

5. EXPERIMENTS AND ANALYSIS

5.1 Properties of NSL-KDD dataset

The data used in classification is NSL-KDD, which is a new dataset for the evaluation of researches in network intrusion detection system as given in Table 1. NSL-KDD consists of selected records of the complete KDD'99 dataset [18]. NSL-KDD dataset solve the issues of KDD'99 benchmark [KDD'99 dataset]. Each NSL-KDD connection record contains 41 features (e.g., protocol type, service, and ag) and is labeled as either normal or an attack, with one specific attack type.

Table 1. Properties of intrusion detection dataset

Datasets	Instances	Attributes
NSL-KDD	11850	42

5.2 Performance comparison of the homogeneous and heterogeneous ensembles

Accuracy of the proposed ensembles is assessed to analyze the performance of the homogeneous and heterogeneous models.

The homogeneous and heterogeneous models are compared with base classifiers in terms of accuracy for NSL-KDD dataset as given in Table 2. According to Figure 1, high improvement of accuracy is observed for proposed hybrid methods than single base classifiers and heterogeneous ensembles exhibit higher performance compared to homogeneous ones.

Table 2. Performance comparison of homogeneous and heterogeneous ensembles

Datasets	Classifier models	Accuracy
NSL-KDD	RBF	84.74%
	Proposed Bagged RBF	86.40%
	SVM	91.81%
	Proposed Bagged SVM	93.92%
	Proposed Hybrid RBF-SVM	98.46%

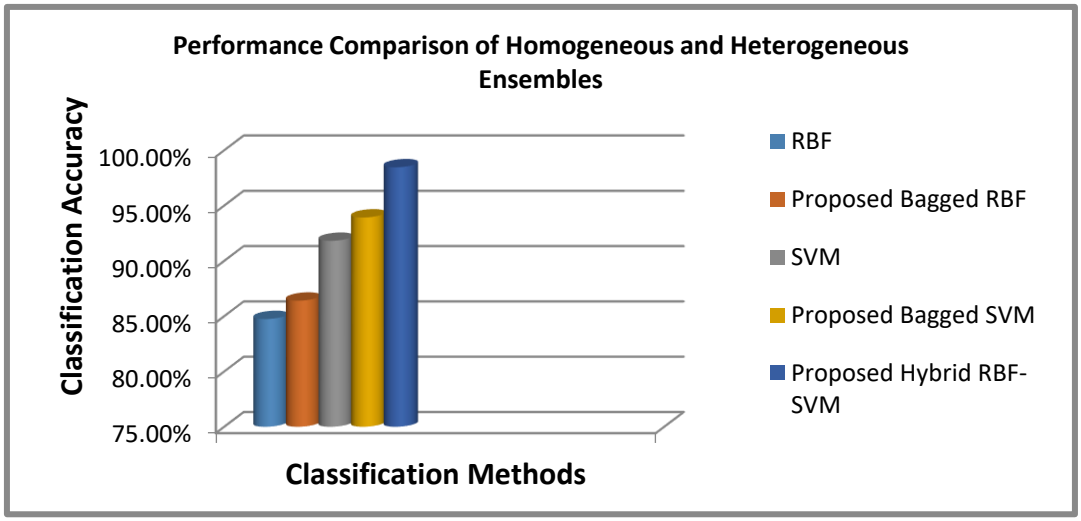

Figure 1. Accuracy for homogeneous and heterogeneous ensembles

5.3 Performance comparison with prior research work

It is found in Table 3 that higher accuracy is accomplished with homogeneous and heterogeneous models in comparison with prior work on the intrusion detection. Also, the proposed classifier proves to show statistically significant performance than state of the art techniques.

Table 3. Experimental results for homogeneous and heterogeneous ensembles

Techniques	Accuracy Claimed
RBF	84.74%
SVM	91.81%
Homogeneous Ensemble Classifiers	
Proposed Bagged RBF	**86.40%**
Reference [19]	80.00%
Reference [20]	82.60%
Reference [21]	84.43%
Reference [22]	84.54%
Reference [23]	86.00%
Proposed Bagged SVM	**93.92%**
Reference [24]	91.14%
Reference [25]	92.13%
Reference [26]	89.40%
Reference [27]	89.02%
Reference [28]	90.53%
Reference [23]	89.00%
Heterogeneous Ensemble Classifiers	
Proposed Hybrid RBF-SVM	**98.46%**
Reference [29]	97.75%
Reference [30]	95.76%
Reference [31]	96.28%
Reference [32]	97.20%
Reference [33]	97.65%
Reference [28]	93.60%
Reference [23]	98.00%

6. CONCLUSIONS

In this work, a novel technique of combining the classification models involving homogeneous ensembles with bagging are implemented using NSL-KDD data and the classifier performance is depicted with accuracy. Here, the proposed ensembles integrate features of corresponding single classifiers. In the same way, new hybrid RBF-SVM heterogeneous ensemble model is constructed and the accuracy is evaluated.

The following observations are revealed from the results.

- ❖ Among the individual classifiers used, SVM depicts significantly higher performance in key aspect of accuracy.
- ❖ The bagged models have been found to achieve remarkable enhancement of classification accuracy when compared to the corresponding individual classifiers.
- ❖ RBF-SVM model exhibits higher accuracy percentage in comparison with standalone classifiers.
- ❖ The hybrid models show significantly high accuracy results than combined models on NSL-KDD dataset.
- ❖ The statistical significance is also found to be high for the proposed classifiers than base classifiers.
- ❖ Results also indicate the homogeneous and heterogeneous models outperforming previous work on the intrusion detection dataset.
- ❖ The limitation for the ensemble is hard to learn and any wrong selection can lead to lower predictive accuracy than an individual model.

Developing and implementing highly accurate classifiers specifically for the NSL-KDD dataset will be the future work.

ACKNOWLEDGMENT

Author acknowledges Annamalai University authorities for their support to do this work.

REFERENCES

[1] Panda, M., Abraham, A., Patra, M.R. (2015). Hybrid intelligent systems for detecting network intrusions. Security and Communication Networks, 8(16): 2741-2749. https://doi.org/10.1002/sec.592

[2] Ingre, B., Yadav, A. (2015). Performance analysis of NSL-KDD dataset using ANN. In Proceedings of the IEEE International Conference on Signal Processing and

Communication Engineering Systems, Guntur, India, pp. 92-96. https://doi.org/10.1109/SPACES.2015.7058223
[3] Farahani, G. (2020). Feature selection based on cross-correlation for the intrusion detection system. Security and Communication Networks, 2020: 1-17. https://doi.org/10.1155/2020/8875404
[4] Albayati, M., Issac, B. (2015). Analysis of intelligent classifiers and enhancing the detection accuracy for intrusion detection system. Int. J. Comput. Intell. Syst., 8(5): 841-853. https://doi.org/10.1080/18756891.2015.1084705
[5] Tsai, C.F., Hsu, Y.F., Lin, C.Y., Lin, W.Y. (2009). Intrusion detection by machine learning: A review. Expert Syst. Appl., 36(10): 11994-12000. https://doi.org/10.1016/j.eswa.2009.05.029
[6] Panda, M., Abraham, A., Patra, M.R. (2012). A hybrid intelligent approach for network intrusion detection. Procedia Eng., 30: 1-9. https://doi.org/10.1016/j.proeng.2012.01.827
[7] Fossaceca, J.M., Mazzuchi, T.A., Sarkani, S. (2015). MARK-ELM: Application of a novel multiple kernel learning framework for improving the robustness of network intrusion detection. Expert Systems with Applications, 42(8): 4062-4080. https://doi.org/10.1016/j.eswa.2014.12.040
[8] Aburomman, A.A., Reaz, M.B.I. (2016). A novel SVM-kNN-PSO ensemble method for intrusion detection system. Applied Soft Computing, 38: 360-372. https://doi.org/10.1016/j.asoc.2015.10.011
[9] Aburomman, A.A., Reaz, M.B.I. (2017). A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers & Security, 65: 135-152. https://doi.org/10.1016/j.cose.2016.11.004
[10] Divyasree, T.H., Sherly, K.K. (2018). A network intrusion detection system based on ensemble CVM using efficient feature selection approach. Procedia Computer Science, 143: 442-449. https://doi.org/10.1016/j.procs.2018.10.416
[11] Salo, F., Nassif, A.B., Essex, A. (2019). Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. Computer Networks, 148: 164-175. https://doi.org/10.1016/j.comnet.2018.11.010
[12] Kunal, Dua, M. (2020). Attribute selection and ensemble classifier based novel approach to intrusion detection system. Procedia Computer Science, 167: 2191-2199. https://doi.org/10.1016/j.procs.2020.03.271
[13] Govindarajan, M. (2014). Hybrid intrusion detection using ensemble of classification methods. Int. J. Comput. Netw. Inf. Secur., 6(2): 45-53. https://doi.org/10.5815/ijcnis.2014.02.07
[14] Broomhead, D.S., Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. Complex Syst., 2: 321-355. Accession Number: ADA196234.
[15] Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines; Technical Report; MSR-TR-98-14; Microsoft Research: Redmond, WA, USA.
[16] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2): 123-140. https://doi.org/10.1007/BF00058655
[17] Breiman, L, (1996). Bias, Variance, and Arcing Classifiers, Technical Report 460, Department of Statistics, University of California, Berkeley, CA.
[18] Lu, Y.J., Cohen, I., Tian, Q., Zhou, X.S. (2007). Feature selection using principal feature analysis. Proceedings of the 15th international conference on Multimedia, Augsburg, Germany, pp. 301-304. https://doi.org/10.1145/1291233.1291297
[19] AbdElrahman, S.M., Abraham, A. (2014). Intrusion detection using error correcting output code based ensemble. Proceedings of the International Conference on Hybrid Intelligent Systems, Kuwait, pp. 181-186. https://doi.org/10.1109/HIS.2014.7086194
[20] Amini, M., Rezaeenour, J., Hadavandi, E. (2016). A neural network ensemble classifier for effective intrusion detection using fuzzy clustering and radial basis function networks. International Journal on Artificial Intelligence Tools, 25(02): 1-32. https://doi.org/10.1142/S0218213015500335
[21] Dubey, R., Rathore, D., Kushwaha, D., Maurya, J.P. (2017). An empirical study of intrusion detection system using feature reduction based on evolutionary algorithms and swarm intelligence methods. International Journal of Applied Engineering Research, 12(19): 8884-8889.
[22] Gao, Y., Liu, Y., Jin, Y.Q., Chen, J.Q., Wu, H.R. (2018). A novel semi-supervised learning approach for network intrusion detection on cloud-based robotic system. IEEE Access: Special Section on Cloud-based Robotic Systems for Intelligent Services, 6: 50927-50938. https://doi.org/10.1109/ACCESS.2018.2868171
[23] Khonde, S.R., Ulagamuthalvi, V. (2020). Hybrid framework for intrusion detection system using ensemble approach. International Journal of Advanced Trends in Computer Science and Engineering, 9(4): 4881-4890. https://doi.org/10.30534/ijatcse/2020/99942020
[24] Zhao, H. (2013). Intrusion detection ensemble algorithm based on bagging and neighborhood rough set. International Journal of Security and Its Applications, 7(5): 193-204. http://dx.doi.org/10.14257/ijsia.2013.7.5.18
[25] Tesfahun, A., Bhaskari, D.L. (2015). Effective hybrid intrusion detection system: A layered approach. International Journal of Computer Network and Information Security, 7(3): 35-41. http://dx.doi.org/10.5815/ijcnis.2015.03.05
[26] Sethia, T.S., Kantardzic, M. (2017). On the reliable detection of concept drift from streaming unlabeled data. Expert Systems with Applications, 82: 77-99. https://doi.org/10.1016/j.eswa.2017.04.008
[27] Teng, S.H., Wu, N.Q., Zhu, H.B., Teng, L.Y., Zhang, W. (2018). SVM-DT-Based adaptive and collaborative intrusion detection. IEEE/CAA Journal of Automatica Sinica, 5(1): 108-118. https://doi.org/10.1109/JAS.2017.7510730
[28] Mbugua, J., Thiga, M., Siror, J. (2019). A comparative analysis of standard and ensemble classifiers on intrusion detection system. International Journal of Computer Applications Technology and Research, 8(4): 107-115. https://doi.org/10.7753/IJCATR0804.1005
[29] Wang, H., Zhang, G.L., Mingjie, E., Sun, N. (2011). A novel intrusion detection method based on improved SVM by combining PCA and PSO. Wuhan University Journal of Natural Sciences, 16(5): 409-413. http://dx.doi.org/10.1007/s11859-011-0771-6
[30] Enache, A.C., Sgârciu, V. (2014). Anomaly intrusions detection based on support vector machines with bat algorithm. Proceedings of the 18th International

Conference on System Theory, Control and Computing, Sinaia, Romania, pp. 856-861. https://doi.org/10.1109/CSCS.2015.12

[31] Dhanabal, L., Shantharajahm, S.P. (2015). Intrusion detection and classification using hybrid support vector machine and dynamic ant colony algorithm. Australian Journal of Basic and Applied Sciences, 9(23): 328-335.

[32] Moustafa, N., Creech, G., Slay, J. (2017). Big data analytics for intrusion detection system: Statistical decision-making using finite Dirichlet mixture models. Data Analytics and Decision Support for Cyber Security, pp. 127-156. https://doi.org/10.1007/978-3-319-59439-2_5

[33] Alareqi, E., Abed, K. (2018). Predictive hybrid machine learning model for network intrusion detection. Proceedings of the 2018 International Conference on Data Science, Las Vegas, Nevada, USA, pp. 258-262.