Detection of Network Intrusions using Hybrid Intelligent Systems

Afreen Bhumgara
Department of Computer Science
Nowrosjee Wadia College
Pune, India
afreenbhumgara97@gmail.com

Anand Pitale
Department of Computer Science
Nowrosjee Wadia College
Pune, India
pitalevanand@gmail.com

***Abstract*- With the advancement in technology, our society has become so dependent on the internet and the number of internet users keeps rising each day. However, with the increased users comes the risk of hacking and malicious activities. One of the major concerns facing the technology sector today is the risk of intrusion. Thus, in the domain of security and computer networks, research in intrusion detection is essential. To combat the threats and malicious activities of the internet, the computer industry has gone a mile by creating new software and hardware products such as the Firewalls, Intrusion Prevention Systems and Detection Systems. Recently researchers created a Network-IDS to prevent such intrusions. However, these systems were prone to manipulations and had defects that were based on classification techniques. These systems failed to provide the necessary protections as they used a single classifier system or the individual classification technique. A single classifier classifies all of the data as normal or not, however due to the evolution of new attack patterns these systems failed to provide optimal attack detection rates with poor false alarm rates. The rise of different attack patterns meant that these systems cannot offer complete protection hence researchers came up with more sophisticated classification techniques that uses blends of several classification algorithms known as a hybrid intelligent system, leading to more detection accuracy. The aim of this study is to contrast various classifiers for network intrusions while combining these classifiers to direct the study towards hybrid intelligent systems. The study is carried out by performing an empirical and literature review while simultaneously providing a base for future studies.**

Keywords—Network Intrusion Detection, Hybrid Intelligent Systems, Network Attacks, Clustering, Classification, NSL-KDD intrusion dataset

I. INTRODUCTION

Over the past few decades, the usage of the internet is on a rise, and computer networks now encompass a broad domain that holds significance. With this rise in computer networks, the use of devices like computers, mobile phones etc have been commonplace. It is obvious that secure and important information often passes through these devices. [1],[2]

When any confidential information is at the risk of being compromised, or there is a loss of such data - these set of actions can formally define an intrusion as stated in [3]. Intrusion is when particular steps are aimed at compromising business plans - that is the confidentialities and integrity or the losses of critical data and the tampering with integral resources within the system. Intrusion detection is the last defensive mechanism in the security system security that detects intrusions. They are of two types: Host based IDS and Network based IDS. HIDS works by taking in data from person's individual computer whereas Network based IDS work by performing real-time analysis of the network traffic.

A. Background Information

The urgent need for new hybrid intelligent systems in this interconnected era, keeps researchers at global level, working on the development of new solutions that could be customized to the customer´s needs. The integration of different techniques and technologies to overcome each specific limitation, has provided the solution to intelligent systems design, helping the development of smarter systems and capabilities. Therefore, every day innovative approaches are introduced to construct intelligent hybrid systems solutions that are even appealing to artificial neural networks, genetic algorithms, etc. Different techniques have already been applied in network-IDS. One of the approaches is data mining. Form models are obtained from data. Data mining techniques includes clustering, regression, classification as well as association rule analysis. For the activities of classifying and predicting, decision trees are tools that are in use. The decision tool technique is characterized by components which are arcs, nodes, and leaves. Bayesian networks is the other technique applied in network intrusion detection. It is based on use of graphical models, capable of showing the various connections. Also, it is characterized by capabilities of representing the causal relationships. The clustering approach is also important as its algorithm has the capability of helping the current

IDS in various ways. Clustering technique is valuable as it help in organizing data collection in clusters. Most IDS work on the basis of the single classification technique. However, they are unable to provide accurate results. Hence, the need for hybrid intelligent systems which use a combination of classifiers to improve the overall performance has been adopted.

As demonstrated by [4] and [5] there are two types of IDS: (1) misuse detection and (2) anomaly detection. Intrusion Detection Systems that use misuse detection learn from labeled data, having an advantage of no false alarms. However, as misuse detection only learns from labeled data, it allows new signatures to pass through it i.e. it cannot determine information that is not in the training dataset. If IDS employs anomaly detection, it can identifying new attacks i.e. it is capable of comparing real time network traffics, and hence this is widely used.

Because of the multiple discrepancies in the KDD cup99 dataset, the NSL KDD dataset has been used for more accuracy. A significant issue in the KDD cup99 dataset is the redundancy, for about 78% of training sets and 75% of test sets being duplicated as seen in [6]. This therefore makes the learning algorithm inherently one-sided, that makes User to Root types of attacks more dangerous. Hence, the NSL-KDD dataset is used - which is the cleaner dataset compared to the KDD cup99 dataset and is available to research. [7] Even though the NSL-KDD dataset may suffer from some inefficiencies and may not represent present networks, it is still widely used to compare and contrast varying intrusion detection methods [8]. In [9] a comprehensive review is available on the NSL-KDD which uses a number of ML classifiers. Various research has been previously carried out on the NSL-KDD which employ a number of different methods - the sole purpose being to develop an IDS. Hence, this paper consists details of the various classifiers. This makes it easy to run all of the required experiments to achieve optimal and consistent evaluation results.

B. Research Aims and Objectives

Each technology is susceptible to a certain degree of risks, with the main risks in computer systems being an intrusion. According to Juez et al. (2018) [18], intrusion refers to unauthorized access to a system to cause harm or to steal vital data to gain undue advantage over the owner of the system. Computer software experts have continually fought to develop systems that have enhanced security features to minimize the level of risk associated with the intrusion. Hybrid intelligent systems are systems development using a combination of artificial intelligence knowledge such as genetic fuzzy systems, neuro-fuzzy systems among other sub-branches of Artificial Intelligence (AI). The coming together of several branches of AI ensures that hybrid intelligence systems are well prepared to thwart unauthorized access to computer systems due to a combination of several security features. The aim of this study is to put forth the relevance of the dataset used, to provide information on the various Machine Learning classifiers and to pre-process and analyze the dataset to use a combination of various algorithms and classifiers so as to use hybrid intelligent systems, rather than using a single classifier so as to improve accurate rates.

II. LITERATURE REVIEW

In the literature review section, background information that is useful to understanding the cup99 dataset is provided. This consists of the various classifiers and techniques that have been used and this part gives a brief overview of how the earlier cup99 dataset is extremely beneficial for contrasting the different types of ML algorithms. As specified, all the major problems of the earlier KDD cup99 dataset that lead to the new version dataset are mentioned in [7] and [9]. [6] performed a comprehensive analysis on the cup99 and found two significant problems which result in extremely poor accuracy for anomaly detection and thereby affect the overall optimality of the given system. Therefore a new dataset was proposed, namely NSL-KDD, which only consisted of selected records from the earlier cup99 dataset, and this did not suffer with any of the above mentioned inefficiencies. Even though the NSL-KDD does not provide a reflection of existing or present networks, it is still widely used as a common ground for various researchers to study and contrast various intrusion detection methods. Classifiers study and tag normal behavior and thus correctly indentify traffic through the network. [15] In general, a number of ML algorithms and other Data Mining techniques have been used specifically in the prevention and detection of intrusions mostly by studying various behavior patterns from traffic and data through the network. Therefore, most ML algorithms include techniques that help establishing models to categorize or classify patterns correctly. Developing such an algorithm to obtain consistent evaluations requires examining the performance of many such algorithms which has been carried out vastly in research. Vipin Kumar et all [9] made use of classification algorithms on the new dataset to figure out the rate of accurate detections. M. Shyu et al [10] proposed a Classifier Selection Model which included research of IDS and the cup99 dataset. Here, they took over 49000+ instances of the cup99 datasets which

were further used to test various ML algorithms viz. Naive Bayes and MLP algorithms. They therefore were able to propose two models to detect various intrusions from the given and underlying KDD dataset. In [11] M. K. Lahre et al tested the SVM algorithm to detect network intrusions using the KDD dataset. Research showed that the Support Vector Machine Algorithm requires a long amount of time to train, and hence, this resulted in a limited usage of the SVM. In [12] the authors Shilpa et al performed Principal Component Analysis for selection of various features and for methods to reduce the dimensions in specific on anomaly detection performed on the newer dataset.

Various attacks that take place in an IDS which are probe attacks, Denial Of Service attacks, Root to Local attacks and User to Root attacks. G. Meera Gandi [16] tested the performance of ML algorithms to specify the attack in one of the four categories. The outputs showed that the decision tree classifier provides accurate performance. According to F. Haddadi et al [13], the authors carried out the various preprocess phases on the KDD dataset viz. normalizing the attributes range to [-1,1] and conversion of a number of symbolic attributes. The result was that Neural Networks are suitable for R2L and U2R attacks but do not provide consistent accuracy rates for DoS or Probe attacks [14]. Similary, Zhang et al [17] tested neural networks on KDD datasets and impleneted Perceptron Back propogation-hybrid (PBH), Radial-Based Functions, Back Propagation (BP) and ARTMAP algorithms. BP and PBH algorithms were the most accurate. Hence, feature selection can lead to a reduce in the time required for computation. However, this study focuses on determining the best combination leading to hybrid intelligent systems for accurate network intrusion detections. For this, we will thus focus on using an ML classifier which will have an optimal rate of detection of attacks.

III. DATASET DESCRIPTION

Due to the various inefficiencies of the KDD cup 99 [19] which have been revealed [7] by a number of statistical analysis, it is evident that these issues affect the capabilities of the IDS and lead to poor detection accuracy. This also lead to a poor evaluation of anomaly detections. Thus, a new and refined dataset, the NSL-KDD [9] has been proposed, wherein only select records of the previous KDD cup99 dataset are present. The advantages are manifold, firstly, the reduction in redundancy in the training sets and testing sets, thereby leading to unbiased results and providing more accuracy. Second, no duplication and thus better reduction rates. The known attacks are present, whereas the additional and unknown datasets are not present and not available in the training sets. In this, every record consists of 41 different attributes and classification it as a normal or not.

TABLE I: FOUR TYPES OF ATTACK CLASSES SPECIFIED

Denial of Servial	Back, Land, Neptune,
Probe-attack	Satan, Mscan, Saint (6)
Root To Local	Guess_Password, Ftp_write, Imap, Warezclient, Spy, Httptunnel
User To Root	Loadmod

TABLE II: NSL-KDD DATASET ON NORMAL OR ATTACK TYPES

Type	**Number**					
	Record	**Normal**	**Denial of Service**	**Probe**	**User To Root**	**Root To Local**
KDDTrain+ 20%	25192	13449	9234	2289	11	209
		53.39 %	36. 65 %	9.09%	0.04 %	0.83 %
KDD Train+	125973	67343	45927	11656	52	995
		53.46 %	36. 46 %	9.25 %	0.04 %	0.79 %
KDD Test+	22544	9711	7458	2421	200	2754
		43.08 %	33.08 %	10.74 %	0.89 %	12.22 %

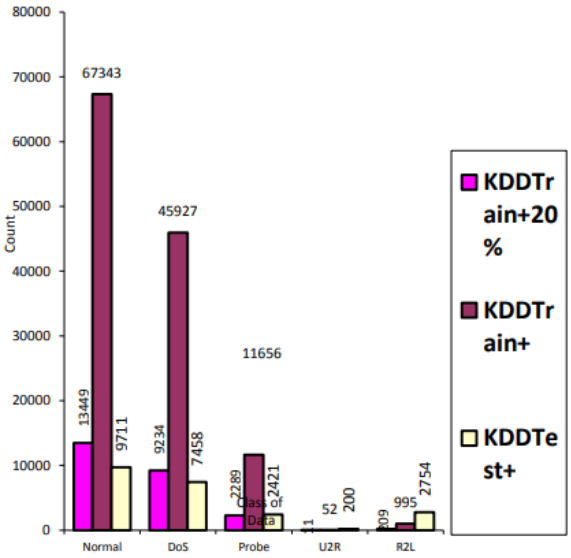

Fig 1: Cup99 data for different attack behavior

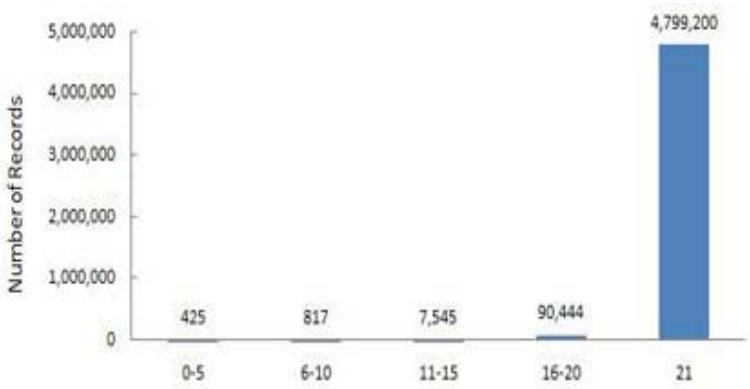

Fig 2: No. of entries in the training-data

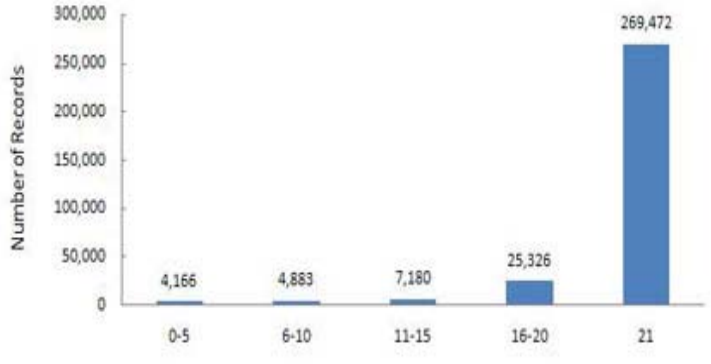

Fig 3. No. of entries in the test-data

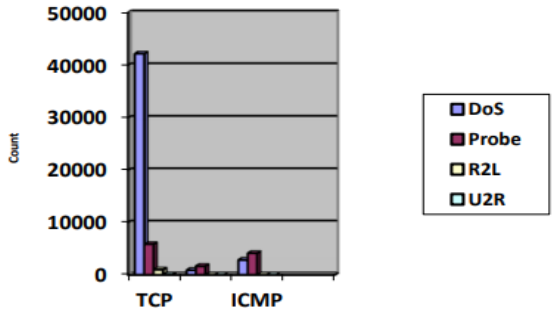

Fig 4. Cup99 data for different attack behavior based on protocol - whether TCP or ICMP

Thus, with further review of the KDDTrain+ dataset, we understand significant information about the various types of attacks. Similarly, the study shows us that a large number of attacks have been made by using the TCP protocol. TCP is widely exploited because of its easy usage and is used to attack victim computers.

TABLE III : DETAILED ANALYSIS OF VARIOUS ATTACKS

	Denial Of Service	Probe-attack	Root To Local	User to Root
Transmission Control Protocol	41723	5785	982	51
User Datagram Protocol	873	1772	-	2
ICMP	2789	4220	-	-

IV. CLASSIFICATION TECHNIQUES

The following section provides an overview of the various ML algorithms and gives an understanding of the need of implementing these algorithms, especially in IDS. This is necessary to extract data as well as analyze the data from a large number of datasets.

Data mining consists of multiple methods, one of them is classification, which is a process of assigning various instances of data to one from all the types. There are a number of classification techniques which exist, all designed to function efficiently than the last. These classification algorithms are built on the basis of mathematical methodologies, neural networks and programming. These classification techniques perform analysis on the present data to make evaluations and make predictions. Most ML algorithms are divided into supervised and unsupervised

categories [21]. Supervised algorithms consist of trained data with pre-labeled object through which they learn and predict. Unsupervised algorithms consist of the unlabeled data and finds grouping of objects. Here, the study focuses on supervised algorithms as the dataset includes pre-defined and pre-labeled classes.

Decision Tree: Here, the problem is further divided into sub-problem. It builds a tree that develops a model to classify. Time is required to build the tree.
Neural Networks: Various statistical learning models are combined which are largely inspired by biological neural networks. These depend on a huge amount of training data, and are used to approximate. This works best with numerical data and so textual data may have to be converted to numerical data, which may require a lot of time.
Nearest-Neighbor: Here, new class are built similar to all the previously saved classes supplied on the basis on training data. This method is extremely time-consuming if the dataset grows.

As each method has specific drawbacks, hence this study involves hybridization leading to optimization. Here, in the hybridized approach, only the pros of the existing technique that could work with current domain and for the specific problem are considered. Algorithms such as SVM, Apriori Algorithm, Decision Forest Algorithm and Naive Bayes are known for their accurate results. [22]

V. Result and Analysis

The following part deals in understanding the methodology, how the process was carried out and what output was obtained.

A. Features and Classification

Data Mining processes viz. cleaning of the data, has multiple features already implemented in Waikato Environment for Knowledge Analysis . WEKA is a tool used to perform classification on the 20percent newer NSL-KDD dataset. Steps include as mentioned in [20] - First, selection and preprocessing of the dataset. Second, running the classifier algorithm. Third, comparison of results.

B. Pre-Processing of Data, Selection of Features and Classification

We now perform normalization on the dataset which is the step wherein dataset is preprocessed and normalized to values 0-1. We have performed normalization as some classifiers produce better results on normalized data [20]. We have filtered 41 features to 5 features using the CFS subset technique for training and testing all the datasets. We use Decision Tree-J48, Support Vector Machine and Naive Bayes algorithms for classification.

C. Discussion

All experiments are carried out in Waikato Environment for Knowledge Analysis and classification algorithms on the NSL-KDD dataset is analyzed. Rates of accurate measurements are shown. We used Correlation-Based Feature Selection to reduce dimensions. Here, J48-Decision Tree has the highest accuracy.

TABLE IV: DETECTION OF ACCURACY WITH DIFFERENT CLASSIFIERS

J48-Decision Tree	Normal-Attack	98.7
	Denial of Service	98.1
	Probe-attack	97.6
	User To Root	97 .5
	Root To Local	97.7
Support Vector Machine	Normal-Attack	96.6
	Denial of Service	97.5
	Probe-attack	97.1
	User To Root	93.4
	Root To Local	93.3
Naïve Bayesian	Normal-Attack	75.7
	Denial of Service	74.2
	Probe-attack	73.9
	User To Root	71.1
	Root To Local	69.9

Results for the normal and attack records for each protocol is given in the table below -

TABLE V: DATA IN THE CUP99 BASED

Data	Transmission Control Protocol	User Datagram Protocol	ICMP
KDDTrain+20 %	20522	3010	1657
KDDTrain+	102679	14991	8299
KDDTest+	18820	2624	1040

VII. CONCLUSIONS & FUTURE WORK

Hybrid Intelligence systems are highly sophisticated that uses a blend of techniques and approaches which are built on different AI fields like fuzzy and symbolic reasoning. They are called hybrid systems because they combine two or more intelligent technologies. There are so many areas where these systems can be used such as network security. The study proved that the NSL-KDD dataset currently is the best dataset to approach IDS. Here, various tests were performed to evaluate a number of different Machine Learning classifiers - J48, SVM and Naive Bayes. All tests took place on the NSL-KDD dataset. We also understand that no single classifier is useful and rather a hybrid intelligence system is more preferred. The Correlation-based Feature Selection method reduces the time required for detection and improves results. The researcher now has a clear idea and understanding of the NSL-KDD dataset. We also know of how using the TCP protocol may lead to more attacks because of its vulnerability. The future scope of the research lies in optimizing algorithms to develop IDS with higher accuracy rates.

REFERENCES -

[1] Upadhyaya, D., & Jain, S. (2013). Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification. International Journal of Computer Science Issues (IJCSI), 10(3), 231 - 236.

[2] "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSLKDD.html, March 2009.

[3] Adetunmbi AO, Zhiwei S, Zhongzhi S, Adewale OS. Network anomalous intrusion detection using fuzzy-Bayes. In IFIP International federation for information processing, 3rd edn, Vol. Volume 228. Intelligent Information Processing, Shi Z, Shi-mohara K, Feng D eds. Springer: Berlin, 2006; pp.525-530.

[4] Lee W and Stolfo S., "Data Mining techniques for intrusion detection", In: Proc. of the 7th USENIX security symposium, San Antonio, TX, 1998.

[5] Dokas P, Ertoz L, Kumar V, Lazarevie A, Srivastava J, and Tan P., "Data Mining for intrusion detection", In: Proc. of NSF workshop on next generation data mining, 2002.

[6] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddc up 99/kddcup99.html, October 2007.

[7] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set", Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009)

[8] J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory," ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294, 2000.

[9] Vipin Kumar, Himadri Chauhan, Dheeraj Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume3, Issue-4, September 2013.

[10] M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172– 179, 2003

[11] M. K. Lahre, M. T. Dhar, D. Suresh, K. Kashyap, and P. Agrawal, "Analyze different approaches for ids using kdd 99 data set," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 1, no. 8, pp. 645–651, 2013

[12] Shilpa lakhina, Sini Joseph and Bhupendra verma, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD", International Journal of Engineering Science and Technology, Vol. 2(6), 2010, 1790- 1799.

[13] Almseidin, Mohammad & Alzubi, Maen & Szilveszter, Kovács & Alkasassbeh, Mouhammd. (2018). Evaluation of Machine Learning Algorithms for Intrusion Detection System retrieved from https://arxiv.org/abs/1801.02330

[14] Almseidin, Mohammad & Alzubi, Maen & Szilveszter, Kovács & Alkasassbeh, Mouhammd. (2018). Evaluation of Machine Learning Algorithms for Intrusion Detection System retrieved from https://arxiv.org/abs/1801.02330

[15] Noureldien A. Noureldien, Izzedin M. Yousif - Accuracy of Machine Learning Algorithms in Detecting DoS Attacks Types - Science and Technology 2016, 6(4): 89-92 DOI: 10.5923/j.scit.20160604.01

[16] G. Meera Gandhi, Machine learning approach for attack prediction and classification using supervised learning algorithms, International Journal of Computer Science & Communication, Vol. 1, No. 2, pp. 247-250. July-December 2010.

[17] Z. Zhang, J. Li, C. Manikopoulos, J. Jorgenson, and J. Ucles, "Hide: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification," in Proc. IEEE Workshop on Information Assurance and Security, 2001, pp. 85–90

[18] Juez, F., Villar, J., Cal, E., Herrero, A., Quintián, H., Saez, J. & Corchado, E. (2018). Hybrid artificial intelligent systems : 13th International Conference, HAIS 2018, Oviedo, Spain, June 20-22, 2018, Proceedings. Cham, Switzerland: Springer.

[19] Sapna S. Kaushik, Dr. Prof.P.R.Deshmukh," Detection of Attacks in an Intrusion Detection System",

International Journal of Computer Science and Information Technologies, Vol. 2 (3), 2011, 982-986

[20] S. Revathi, Dr. A. Malathi "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December - 2013 ISSN: 2278-0181

[21] M. Al-Kasassbeh, "Network intrusion detection with wiener filter-based agent," World Appl. Sci. J, vol. 13, no. 11, pp. 2372–2384, 2011.

[22] XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, "Top 10 algorithms in data mining", Knowledge and Information Systems Journal,SpringerVerlag London, vol. 14, Issue 1, pp. 1-37, 2007.

[23] Medsker, L. R. (2012). Hybrid Intelligent Systems. Boston, MA: Springer US.

[24] Panda, M., Abraham, & Patra, M. (2012). Hybrid intelligent systems for detecting network intrusions. Security And Communication Networks, 8(16), 2741-2749. doi: 10.1002/sec.592

[25] Ravi, V., Naveen, N., & Pandey, M. (2013). Hybrid classification and regression models via particle swarm optimization auto-associative neural network based nonlinear PCA. International Journal Of Hybrid Intelligent Systems, 10(3), 137-149. doi: 10.3233/his-130173