Application of Machine Learning Approaches in Intrusion Detection System: A Survey

Nutan Farah Haq
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

Musharrat Rafni
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

Abdur Rahman Onik
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

Faisal Muhammad Shah
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

Md. Avishek Khan Hridoy
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

Dewan Md. Farid
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

***Abstract*—Network security is one of the major concerns of the modern era. With the rapid development and massive usage of internet over the past decade, the vulnerabilities of network security have become an important issue. Intrusion detection system is used to identify unauthorized access and unusual attacks over the secured networks. Over the past years, many studies have been conducted on the intrusion detection system. However, in order to understand the current status of implementation of machine learning techniques for solving the intrusion detection problems this survey paper enlisted the 49 related studies in the time frame between 2009 and 2014 focusing on the architecture of the single, hybrid and ensemble classifier design. This survey paper also includes a statistical comparison of classifier algorithms, datasets being used and some other experimental setups as well as consideration of feature selection step.**

***Keywords*—Intrusion detection; Survey; Classifiers; Hybrid; Ensemble; Dataset; Feature Selection**

I. INTRODUCTION

The Internet has become the most essential tool and one of the best sources of information about the current world. Internet can be considered as one of the major components of education and business purpose. Therefore, the data across the Internet must be secure. Internet security is one of the major concerns now-a-days. As Internet is threatened by various attacks it is very essential to design a system to protect those data, as well as the users using those data. Intrusion detection system (IDS) is therefore an invention to fulfill that requirement. Network administrators adapt intrusion detection system in order to prevent malicious attacks. Therefore, intrusion detection system became an essential part of the security management. Intrusion detection system detects and reports any intrusion attempts or misuse on the network. IDS can detect and block malicious attacks on the network, retain the performance normal during any malicious outbreak, perform an experienced security analysis.

Intrusion detection system approaches can be classified in 2 different categories. One of them is anomaly detection and the other one is signature based detection, also known as misuse detection based detection approach [4, 41]. The misuse detection is used to identify attacks in a form of signature or pattern. As misuse detection uses the known pattern to detect attacks the main disadvantage is that it will fail to identify any unknown attacks to the network or system. On the other hand, anomaly detection is used to detect unknown attacks. There are different ways to find out the anomalies. Different machine learning techniques are introduced in order to identify the anomalies.

Over the years, many researchers and scholars have done some significant work on the development of intrusion detection system. This paper reviewed the related studies in intrusion detection system over the past six years. This paper enlisted 49 papers in total from the year 2009 to 2014.This paper enlisted the proposed architecture of the classification techniques, algorithms being used. A Statistical comparison has been added to show classifier design, chosen algorithms, used datasets as well as the consideration of feature selection step.

This paper is organized as follows: Section 2 provides the research topic overview where a number of techniques for intrusion detection have been described. Section 3 represents a statistical overview of articles over the years on the algorithms that were frequently used, the datasets for each experiment and the consideration of feature selection step. Section 4 includes the discussion and conclusion as well as some issues which have been highlighted for future research in intrusion detection system using machine learning approaches.

II. RESEARCH PAPER OVERVIEW

A. Machine Learning Approach

Machine learning is a special branch of artificial intelligence that acquires knowledge from training data based on known facts. Machine learning is defined as a study that allows computers to learn knowledge without being programmed mentioned by Arthur Samuel in 1959.Machine learning mainly focuses on prediction. Machine learning techniques are classified into three broad categories such as – supervised learning, unsupervised learning, and reinforcement learning.

1) Supervised Learning

Supervised learning is also known as classification. In supervised learning data, instances are labeled in the training phase. There are several supervised learning algorithms. Artificial Neural Network, Bayesian Statistics, Gaussian Process Regression, Lazy learning, Nearest Neighbor algorithm, Support Vector Machine, Hidden Markov Model, Bayesian Networks, Decision Trees(C4.5,ID3, CART, Random Forrest), K-nearest neighbor, Boosting, Ensembles classifiers (Bagging, Boosting), Linear Classifiers (Logistic regression, Fisher Linear discriminant, Naive Bayes classifier, Perceptron, SVM), Quadratic classifiers are some of the most popular supervised learning algorithms.

2) Unsupervised Learning

In unsupervised learning data instances are unlabeled. A prominent way for this learning technique is clustering.

Some of the common unsupervised learners are Cluster analysis (K-means clustering, Fuzzy clustering), Hierarchical clustering, Self-organizing map, Apriori algorithm, Eclat algorithm and Outlier detection (Local outlier factor).

3) Reinforcement Learning

Reinforcement learning means computer interacting with an environment to achieve a certain goal. A reinforcement approach can ask a user (e.g., a domain expert) to label an instance, which may be from a set of unlabeled instances.

B. Single Classifiers

One machine learning algorithm or technique for developing an intrusion detection system can be used as a standalone classifier or single classifier. Some of the machine learning techniques have been discussed in this study which have been found as frequently used single classifiers in our studied 49 research papers.

1) Decision Tree

Creating a classifier for predicting the value of a target class for an unseen test instance, based on several already known instances is the task of Decision tree (DT). Through a sequence of decisions, an unseen test instance is being classified by a Decision tree [11]. Decision tree is very much popular as a single classifier because of its simplicity and easier implementation [14]. Decision tree can be expanded in 2 types: (i) Classification tree, with a range of symbolic class labels and (ii) Regression tree, with a range of numerically valued class labels [11].

2) Naive Bayes

On the basis of the class label given Naive Bayes assumes that the attributes are conditionally independent and thus tries to estimate the class-conditional probability[15]. Naive Bayes often produces good results in the classification where there exist simpler relations. Naive Bayes requires only one scan of the training data and thus it eases the task of classification a lot.

3) K-nearest neighbor

Various distance measure techniques are being used in K-nearest neighbor. K-nearest neighbor finds out k number of samples in training data that are nearest to the test sample and then it assigns the most frequent class label among the considered training samples to the test sample. For classifying samples, K-nearest neighbor is known as an approach which is the most simple and nonparametric[8]. K-nearest neighbor can be mentioned as an instance-based learner, not an inductive based [35].

4) Artificial Neural Network

Artificial Neural Network (ANN) is a processing unit for information which was inspired by the functionality of human brains [23]. Typically neural networks are organized in layers which are made up of a number of interconnected nodes which contain a function of activation. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where via a system of weighted connections the actual processing is done. The hidden layers then link to an output layer for producing the detection result as output.

5) Support Vector Machines

Support vector machine (SVM) was introduced in mid-1990's [5]. The concept behind SVM for intrusion detection basically is to use the training data as a description of only the normal class of objects or which is known as non-attack in intrusion detection system, and thus assuming the rest as anomalies [51]. The classifier constructed by support vector machines methodology discriminates the input space in a finite region where the normal objects are contained and all the rest of the space is assumed to contain the anomalies [9].

6) Fuzzy Logic

For reasoning purpose, dual logic's truth values can be either absolutely false (0) or absolutely true (1), but in Fuzzy logic these kinds of restrictions are being relaxed [60]. That means in Fuzzy logic the range of the degree of truth of a statement can hold the value between 0 and 1 along with '0' and '1'[11].

C. Hybrid Classifiers

A hybrid classifier offers combination of more than one machine learning algorithms or techniques for improving the intrusion detection system's performance vastly. Using some clustering-based techniques for preprocessing samples in training data for eliminating non-representative training samples and then, the results of the clustering are used as training samples for pattern recognition in order to design a classifier. Thus, either supervised or unsupervised learning approaches can be the first level of a hybrid classifier [11].

D. Ensemble Classifiers

The classifiers performing slightly better than a random classifier are known as weak learners. When multiple weak learners are combined for the greater purpose of improving the performance of a classifier significantly is known as Ensemble classifier [11].Majority vote, bagging and boosting are some common strategies for combining weak learners [15].Though it is known that the disadvantages of the component classifiers get accumulated in the ensemble classifier, but it has been producing a very efficient performance in some combination. So researchers are becoming more interested in ensemble classifiers day by day.

III. Statistical Comparisons of Related Work

A. Distribution of Papers by Year of Publication

The survey comprises 49 research papers in the time frame between 2009 and 2014. It discussed 8 papers from each of the year 2009, 2010 and 2012.The highest number of papers are studied from the year 2011.The number of papers from that year is 11. 10 papers are enlisted for the year 2013 and 4 papers from 2014.Fig.1 depicts the percentage of distribution of papers by year of publication.

B. Classifier design

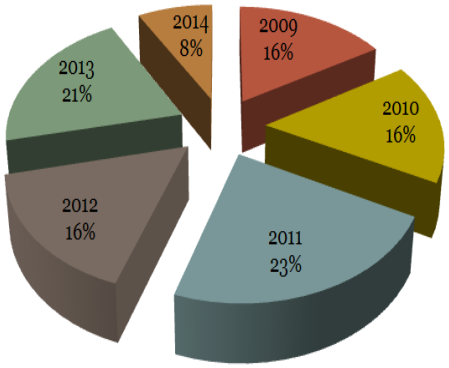

Fig. 1. Year-wise distribution of papers

Intrusion detection method can be categorized in 3 categories namely single, hybrid and ensemble[11] .Fig.2 depicts the number of research papers in terms of single, hybrid and ensemble classifiers used in each year. According

TABLE I. Total Numbers of Research Papers for the Types Of Classifier Design

Classifier design type	No. of research paper	References
Single	20	(D. Sa´nchez, 2009)[12], (Su-Yun Wua, 2009)[50], (Jun Ma, 2009)[27], (Mao Ye, 2009)[31], (Feng Jiang, 2009)[16], (Yung-Tsung Hou, 2010)[58], (Min Seok Mok, 2010)[34], (Han-Ching Wu, 2010)[22], (Chengpo Mua, 2010)[10], (Wang Dawei, 2011)[53], (G. Davanzo, 2011)[17], (Levent Koc, 2012)[29], (Carlos A. Catania, 2012)[9], (Inho Kang, 2012)[26], (Prabhjeet Kaur, 2012)[38], (Yusuf Sahin, 2013)[59], (S. Devaraju, 2013)[42], (Guillermo L. Grinblat, 2013)[21], (Mario Poggiolini, 2013)[32], (Adel Sabry Eesa, 2014)[2].
Hybrid	22	(Kamran Shafi, 2009)[28], (M. Bahrololum, 2009)[30], (Gang Wang, 2010)[18], (Woochul Shim, 2010)[55], (Muna Mhammad T. Jawhar, 2010)[37], (Ilhan Aydin, 2010)[25], (Seung Kim, 2011)[45], (I.T. Christou, 2011)[24], (Mohammad Saniee Abadeh, 2011)[36], (Shun-Sheng Wang, 2011)[47], (Su, 2011)[49], (Seungmin Lee, 2011)[46], (Yinhui Li, 2012)[57], (Bose, 2012)[6], (Prof. D.P. Gaikwad, 2012)[39], (A.M.Chandrashekhar, 2013)[1], (Mazyar Mohammadi Lisehroodi, 2013)[33], (Dahlia Asyiqin Ahmad Zainaddin, 2013)[13], (Seongjun Shin, 2013)[44], (Gisung Kim, A novel hybrid intrusion detection method integrating anomaly detection with misuse detection, 2013)[19], (Wenying Feng, 2014)[54], (Ravi Ranjan, 2014)[40].
Ensemble	7	(Tich Phuoc Tran, 2009)[52], (C.A. Laurentys, 2011)[7], (Dewan Md. Farid M. Z., 2011)[15], (Yang Yi, 2011)[56], (Siva S. Sivatha Sindhu, 2012)[48], (Dewan Md. Farid L. Z., 2013)[14], (Akhilesh Kumar Shrivas, 2014)[3]

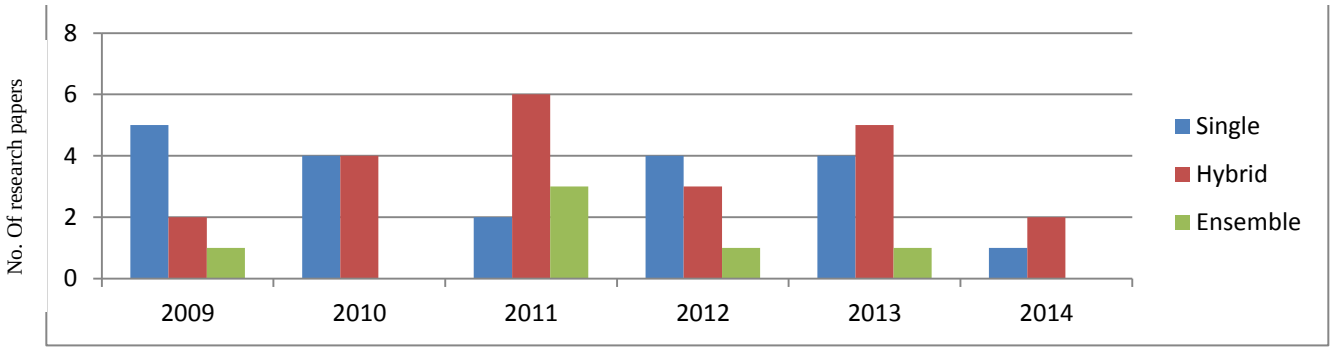

Fig. 2. Year wise distribution of research papers for the types of classifier design

to the statistical comparison between the enlisted papers, hybrid classifiers have the highest number of literatures in the time frame mentioned earlier with a total number of 22. What comes later in terms of study is single classifiers which have been studied in 20 papers.

C. Single classifiers

Fig. 3 depicts the number of single learning algorithms used as classifiers. The number of research papers in the single classifier architecture using different classification techniques, e.g. Bayesian, SVM, DT, ANN, KNN, Fuzzy Logic enlisted in this survey paper is twenty. Table II enlists the proposed algorithms used in all the articles reviewed in this paper. Table IV shows Year wise distribution of single classifiers regarding results and citation of each article.

Support vector machine and Artificial neural network are the most popular approaches for single learning algorithm classifiers. Though we have taken 49 related papers and number of comparative samples is less but the comparison result implies that Support Vector machine is by far the most common and considered single classification technique. On the contrary, Fuzzy logic seems to be less considerable among the single classifiers over the enlisted literatures.

D. Ensemble classifiers

Multiple weak learners are combined in Ensemble classifiers. Table III depicts the articles using ensemble classifiers in intrusion detection system. Statistics shows AdaBoost is the most commonly used learning algorithm along with majority voting. Table III also enlists the detection rate of each of the classifier and the citation of each article throughout the time period.

E. Hybrid classifiers

Table V depicts Year wise distribution of Hybrid classifiers regarding results and citation of each article. Hybrid classifiers in intrusion detection have established in the mainstream study due to the performance accuracy in recent times Statistics shows hybrid classifiers have the highest number of articles in the Year of 2011. The table also shows the used algorithms in each article and their performance in intrusion detection system.

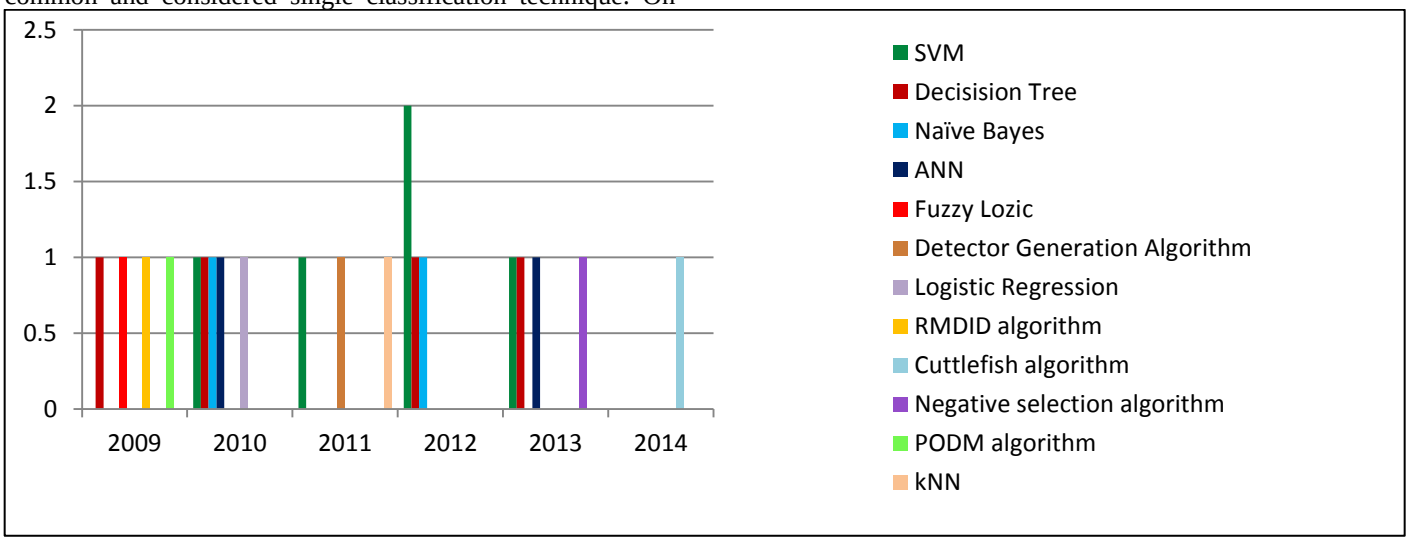

Fig. 3. Distribution of Single classifiers over the Years

TABLE II. ALGORITHMS USED IN SINGLE TYPE OF CLASSIFIER DESIGNED BASED RESEARCH PAPERS

Algorithm	Research paper Title	Reference
Naive Bayes	• A network intrusion detection system based on a hidden naive bayes multiclass classifier. • Malicious web content detection by machine learning.	(Levent Koc, 2012)[29] ; (Yung-Tsung Hou, 2010)[58]
Support Vector Machine	• An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection. • A differentiated one-class classification method with applications to intrusion detection. • Abrupt change detection with One-Class Time Adaptive Support Vector Machines. • Malicious web content detection by machine learning. • Anomaly detection techniques for a web defacement monitoring service.	(Carlos A. Catania, 2012)[9] ; (Inho Kang, 2012)[26] ; (Guillermo L. Grinblat, 2013)[21] ; (Yung-Tsung Hou, 2010)[58]; (G. Davanzo, 2011)[17].
Decision Tree	• Madam id for intrusion detection using data mining. • A cost-sensitive decision tree approach for fraud detection. • Data mining-based intrusion detectors. • Malicious web content detection by machine learning.	(Prabhjeet Kaur, 2012)[38]; (Yusuf Sahin, 2013)[59] ; (Su-Yun Wua, 2009)[50] ; (Yung-Tsung Hou, 2010)[58].
Artificial Neural Network	• Detection of accuracy for intrusion detection system using neural network classifier. • Neural networks-based detection of stepping-stone intrusion.	(S. Devaraju, 2013)[42] ; (Han-Ching Wu, 2010)[22].

Fuzzy Logic	• Data mining-based intrusion detectors.	(Su-Yun Wua, 2009)[50].
Detector Generation Algorithm	• Evolving boundary detector for anomaly detection	(Wang Dawei, 2011)[53].
Negative Selection algorithm	• Application of the feature-detection rule to the Negative Selection Algorithm	(Mario Poggiolini, 2013)[32].
Logistic regression	• Random effects logistic regression model for anomaly detection	(Min Seok Mok, 2010)[34].
RMDID	• Projected outlier detection in high-dimensional mixed-attributes data set.	(Mao Ye, 2009)[31].
PODM	• Information inconsistencies detection using a rule-map technique	(Jun Ma, 2009)[27]
Cuttlefish algorithm	• A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems.	(Adel Sabry Eesa, 2014)[2].
Sequence-based Outlier Detection algorithm	• Some issues about outlier detection in rough set theory.	(Feng Jiang, 2009)[16].
K-nearest neighbour (KNN)	• Anomaly detection techniques for a web defacement monitoring service.	(G. Davanzo, 2011)[17].

TABLE III. YEAR WISE DISTRIBUTION OF ENSEMBLE CLASSIFIERS REGARDING RESULTS AND CITATION OF EACH ARTICLE

Year	Research Paper Title	Reference	Algorithm used	Result (%)	Citation
2009	Novel intrusion detection using probabilistic neural network & adaptive boosting	(Tich Phuoc Tran, 2009)[52]	• NN • AdaBoost • BSPNN	DR : 94.31	14
2011	A novel artificial immune system for fault behavior detection	(C.A. Laurentys, 2011)[7]	• GA • Majority Vote	DR : 97.85	17
	Adaptive intrusion detection based on boosting & naive Bayesian classifier	(Dewan Md. Farid M. Z., 2011)[15]	• NB • AdaBoost	DR : 99.75	14
	Incremental SVM based on reversed set for network intrusion detection	(Yang Yi, 2011)[56]	• SVM • ISVM	DR : 81.377	30
2012	Decision tree based light weight intrusion detection using a wrapper approach	(Siva S. Sivatha Sindhu, 2012)[48]	• Neural ensemble decision tree	DR : 98.38	44
2013	An adaptive ensemble classifier for mining concept drifting data streams	(Dewan Md. Farid L. Z., 2013)[14]	• NB • C4.5 • AdaBoost	DR : 92.65	13
2014	An ensemble model for classification of attacks with feature selection based on KDD-99 & NSL-KDD data set	(Akhilesh Kumar Shrivas, 2014)[3]	• ANN • Bayesian Network • Gain ratio FS	DR : 97.53 (using NSL-KDD) DR: 99.41 (using KDD-99)	[a]

[a]Not cited yet.

TABLE IV. YEAR WISE DISTRIBUTION OF SINGLE CLASSIFIERS REGARDING RESULTS AND CITATION OF EACH ARTICLE

Year	Research Paper Title	Reference	Algorithm used	Result (%)	Citation
2009	Association rules applied to credit card fraud detection	(D. Sa´nchez, 2009)[12]	• Association rule methodology	Certainty factor : 80.08	64
	Data Mining based intrusion detectors	(Su-Yun Wua, 2009)[50]	• C4.5	DR : 70.62 FAR: 1.44	67
	Some issues about outlier detection in rough set theory	(Feng Jiang, 2009)[16]	• Outlier Detection algorithm	DR: SEQ based : 90 DIS based : 92	30
	Projected outlier detection in high dimensional mixed attributes data set	(Mao Ye, 2009)[31]	• PODM algorithm	DR: Credit approval data : 70 Breast Cancer Data : 80 Mushroom Data : 96 Synthetic Data : 97	24
	Information inconsistencies detection using a rule map technique	(Jun Ma, 2009)[27]	• RMDID algorithm	Error scales = 5.0% Inconsistent entries in Train Set = 5, Test Set = 4	1
2010	Malicious web content detection by machine learning	(Yung-Tsung Hou, 2010)[58]	• Naïve Bayes • DT • SVM • AdaBoost	Accuracy : NB : 58.28 DT : 94.74 SVM: 93.51 Boosted DT: 96.14	39
	Random effect logistic regression model for anomaly detection	(Min Seok Mok, 2010)[34]	• Logistic regression model.	Classification accuracy : Training dataset : 79.43 (Normal) 20.57(Attack) Validation dataset: 79.17 (Normal) 20.83 (Attack)	8
	An intrusion response decision making model based on hierarchical task network planning	(Chengpo Mua, 2010)[10]	• Hierarchical task network planning	Roc curve : excellent	20
	Neural Networks based detection of stepping stone intrusion	(Han-Ching Wu, 2010)[22]	• Neural Network	Accuracy : 99.0	13

2011	Evolving boundary detectors for anomaly detection	(Wang Dawei, 2011)[53]	• Detector Generation algorithm	DR : Iris Dataset : 99.28 considering Self radius = 0.08 Boundary threshold = 0.04 KDD dataset : DOS : 94.5 Probing : 93.64 U2R: 78.85 R2L: 50.69 considering Self radius = 0.05 Boundary threshold = 0.025	6
	Anomaly detection techniques for a web defacement monitoring service	(G. Davanzo, 2011)[17]	• K nearest neighbor • Support Vector machine	FPR: K nearest neighbor : 19.43 SVM :6.45	3
2012	A network intrusion detection system based on Hidden Naïve bayes multiclass classifier	(Levent Koc, 2012)[29]	• Hidden Naïve Bayes	Accuracy : 93.73 Error rate: 6.28	45
	An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection	(Carlos A. Catania, 2012)[9]	• Support Vector machine	DR : 88.64 (80% attack) 98.37 (1% attack)	11
	A differentiated one-class classification method with applications to intrusion detection	(Inho Kang, 2012)[26]	• Support Vector machine	DR : M=200* Targeted attack : 96.9 (4.7 % more than ordinary detection)	17
	Madam id for intrusion detection using Data mining	(Prabhjeet Kaur, 2012)[38]	• Decision Tree (J48)	FP rate :75.00 Precession : 1.7 Recall: 66.7	7
2013	A cost sensitive Decision tree approach for fraud detection	(Yusuf Sahin, 2013)[59]	• Decision Tree	TPR: Direct cost : 74.6 Class Probability : 92.1 CS-Gini : 92.8 Cs-IG: 92.6	9
	Detection of accuracy for intrusion detection system using neural network classifier	(S. Devaraju, 2013)[42]	• Neural Network	Accuracy : FFNN : 79.49 ENN: 78.1 GRNN: 58.74 PNN:85.50 RBNN: 83.51	4
	Abrupt change detection with one class time adaptive Support Vector Machine	(Guillermo L. Grinblat, 2013)[21]	• Support Vector Machine	495.9 sequences correctly classified within 500 sequences.	3
	Application of feature –detection rule to the negative selection algorithm	(Mario Poggiolini, 2013)[32]	• Negative Selection algorithm	Feature Detection rule : 0.9375 HD rule : 0.7686 RCHK(No MHC rule):0.8258 RCHK(Global MHC rule) : 0.5155 RCHK(MHC) rule : 0.9482	3
2014	A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection system	(Adel Sabry Eesa, 2014)[2]	• Cuttlefish algorithm	AR : 73.267 DR: 71.067 FPR: 17.685	[b]

[b]Not cited yet.

TABLE V. A DETAILED INFORMATION ON RESEARCH PAPERS DESIGNED WITH HYBRID CLASSIFIER

Year	Research Paper Title	Reference	Algorithm(s) used	Result (%)	Citation
2009	Anomaly intrusion detection design using hybrid of unsupervised and supervised neural network	(M. Bahrololum, 2009)[30]	• NN	TP rate : 97.00(Dos) 71.65(Probe) 26.69(R2L)	11
	An adaptive genetic-based signature learning system for intrusion detection	(Kamran Shafi, 2009)[28]	• GA	Accuracy : 92 FA rate : 0.84	31
2010	A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering	(Gang Wang, 2010)[18]	• ANN. • Fuzzy clustering.	Accuracy : 96.71 Precision : 99.91(Dos) 48.12(Probe) 93.18(R2L) 83.33(U2R)	114
	A distributed sinkhole detection method using cluster analysis	(Woochul Shim, 2010)[55]	• Hierarchical cluster analysis.	DR : 96.61	7
	Design Network Intrusion Detection System using hybrid Fuzzy-Neural Network	(Muna Mhammad T. Jawhar, 2010)[37]	• Fuzzy C-means clustering. • NN	Accuracy : 100(Dos) 100(U2R) 99.8(Probe) 40(R2L) 68.6(Unknown)	21
	Chaotic-based hybrid negative selection algorithm and its applications in fault and anomaly detection	(Ilhan Aydin, 2010)[25]	• Negative selection. • Clonal selection. • KNN.	Accuracy : 97.65	51
2011	Detecting fraud in online games of chance and lotteries	(I.T. Christou, 2011)[24]	• LOF. • K-means clustering. • EXAMCE.	DR : 98	3
	Fast outlier detection for very large log data	(Seung Kim, 2011)[45]	• Kd-tree indexing. • Approximated KNN. • LOF.	Gained time efficiency : 293-8727	11
	Design and analysis of genetic fuzzy systems for intrusion detection in computer networks	(Mohammad Saniee Abadeh, 2011)[36]	• Fuzzy genetic based machine learning methods: (i)Michigan,(ii)Pitsburg,(iii)IRL.	DR : 88.13 (Mitchigan) 99.53 (Pitsburg) 93.2 (IRL)	21

	An Integrated Intrusion Detection System for Cluster-based Wireless Sensor Networks	(Shun-Sheng Wang, 2011)[47]	• BPN. • ART. • Rule based method.	Accuracy: 95.13	24
	Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers	(Su, 2011)[49]	• GA. • KNN.	Accuracy : 97.42 (with known attack) Accuracy : 78 (with unknown attack)	16
	Self-adaptive and dynamic clustering for online anomaly detection	(Seungmin Lee, 2011)[46]	• SOM. • K-means clustering	DR : 83.4 (offline) 86.4 (online)	14
2012	An efficient intrusion detection system based on support vector machines and gradually feature removal method	(Yinhui Li, 2012)[57]	• K-means clustering. • SVM. • Ant colony.	DR : 98.6249	40
	The combined approach for anomaly detection using neural networks & clustering techniques	(Bose, 2012)[6]	• SOM. • K-means clustering.	DR : 98.5 (Dos)	2
	Anomaly Based Intrusion Detection System Using Artificial Neural Network and Fuzzy Clustering	(Prof. D.P. Gaikwad, 2012)[39]	• ANN. • Fuzzy clustering.	*	6
2013	Fortification of hybrid intrusion detection system using variants of neural networks & support vector machines	(A.M.Chandrashekhar, 2013)[1]	• Fuzzy C-means clustering. • Fuzzy neural network. • SVM.	Accuracy : 98.94 (Dos) 97.11 (Probe) 97.80 (U2R) 97.78 (R2L)	2
	Hybrid of fuzzy clustering Neural network over NSL data set for intrusion detection system	(Dahlia Asyiqin Ahmad Zainaddin, 2013)[13]	• Fuzzy clustering.	Recall : 99.1 (Dos) 94.1 (Prob) 78 (U2R) 89 (R2L)	4
	A hybrid framework based on neural network MLP & K-means clustering for intrusion detection system	(Mazyar Mohammadi Lisehroodi, 2013)[33]	• K-means clustering. • MLP	DR : 99.99 (Dos) 99.97 (Probe) 99.99 (U2R) 99.98 (R2L)	[c]
	Advanced probabilistic approach for network intrusion forecasting and detection	(Seongjun Shin, 2013)[44]	• Markov chain. • K-means clustering. • APAN.	DR : 90	9
	A novel hybrid intrusion detection method integrating anomaly detection with misuse detection	(Gisung Kim, A novel hybrid intrusion detection method integrating anomaly detection with misuse detection, 2013)[19]	• C4.5. • 1-class SVM.	DR : 99.98 (with known attack) 97.4 (with unknown attack) Training time : 21.375 sec Testing time : 10.13 sec	9
2014	Mining network data for intrusion detection through combining SVMs with ant colony networks	(Wenying Feng, 2014)[54]	• CSOACN (self organized ant colony network) • SVM • CSVAC(combining support vectors with ant colony)	DR : 94.86 FP : 6.01 FN : 1.00	10
	A new clustering approach for anomaly intrusion detection	(Ravi Ranjan, 2014)[40]	• C4.5. • SVM. • K-means clustering.	DR : 96.12 (Dos) 90.10 (R2L) 70.51 (U2R) 70.13 (R2L). Accuracy : 96.38 False alarm rate : 3.2	4

[c]Not cited yet

F. Used Dataset in Researches

Datasets are assigned for default tasks e.g., Classification, Regression, Function learning, Clustering. Datasets reviewed by this paper is for classification purpose. As Fig.4 depicts, by far the most common dataset being used is KDD cup 1999 dataset. This dataset contains 4,000,000 instances and 42 attributes. The number of papers using KDD cup 1999 data set yields a peak in 2011 and in total 20 research papers has mentioned KDD Cup 1999 as their dataset.

Car evolution dataset [32] contains 1,728 instances with 6 attributes, attribute types are categorical. Wisconsin Breast cancer [16] has multivariate data types, all 10 attributes are integer types and it has 699 instances. Glass [32] dataset with multivariate data types and 214 instances It has 10 real attributes. Mushroom dataset [32] contains 22 categorical attributes and 8,124 instances. Lympography dataset [16] contains 18 categorical attributes and 148 instances. Yeast dataset [24] have 8 real attributes with 1,484 instances. Fisher-Iris dataset [25] contains 4 real attributes with 150 instances.Bicup2006 dataset and CO2 dataset [27] have 1,323 and 296 instances respectively. Public datasets like DARPA 1998, DARPA 2000, Fisher-Iris dataset, NSL KDD datasets are used in many related studies. Study also shows that few private or non-public datasets used over the time frame. Although the study briefly highlights public datasets like KDD cup 99, DARPA 1998, DARPA 2000 being considered as standard datasets for intrusion detection system. DARPA dataset contains around 1.5 million traffic instances [36]. NSL-KDD dataset was proposed by removing all redundant instances from KDD'99. Thus, NSL-KDD dataset is more efficient than KDD'99 in getting more accurate evaluation of different learning techniques [19]. Some of the datasets were randomly used by the researchers. Table VI shows the year-wise distribution of randomly used dataset.

TABLE VI. YEAR-WISE DISTRIBUTION OF RANDOMLY USED DATASET

Data Set	2009	2010	2011	2012	2013	2014	Total
Car Evaluation					1		1
Glass					1		1
DAMADICS			1				1
Yeast			1				1
Ionosphere			1				1
Musk			1				1
Malicious Web pages		1	1				2
Bicup2006	1						1
CO2	1						1
Lympography	1						1

G. Feature Selection

Feature Selection is an important step for the improvement of the system performance. Feature selection is considered before the training phase. Feature selection points out the best features and eliminates the redundant and irrelevant features. Table VII shows the year-wise distribution of feature selection step consideration. Table VII implies that out of 49 studies, 21 used feature selection step for their proposed architecture. It also shows that the number of papers using feature selection yields a peak in the year 2012, where out of 8 papers in that year 7 used feature selection step. On the contrary, in 2009 the scenario was completely opposite. Though we have taken 49 related papers and number of differences in those papers are trivial but the comparison result implies that 21 experiments used feature selection where 28 experiments did not. It implies that feature selection is not a popular procedure in intrusion detection. Table VII and VIII overview the year-wise distribution of feature selection considered in related studies and the count of paper.

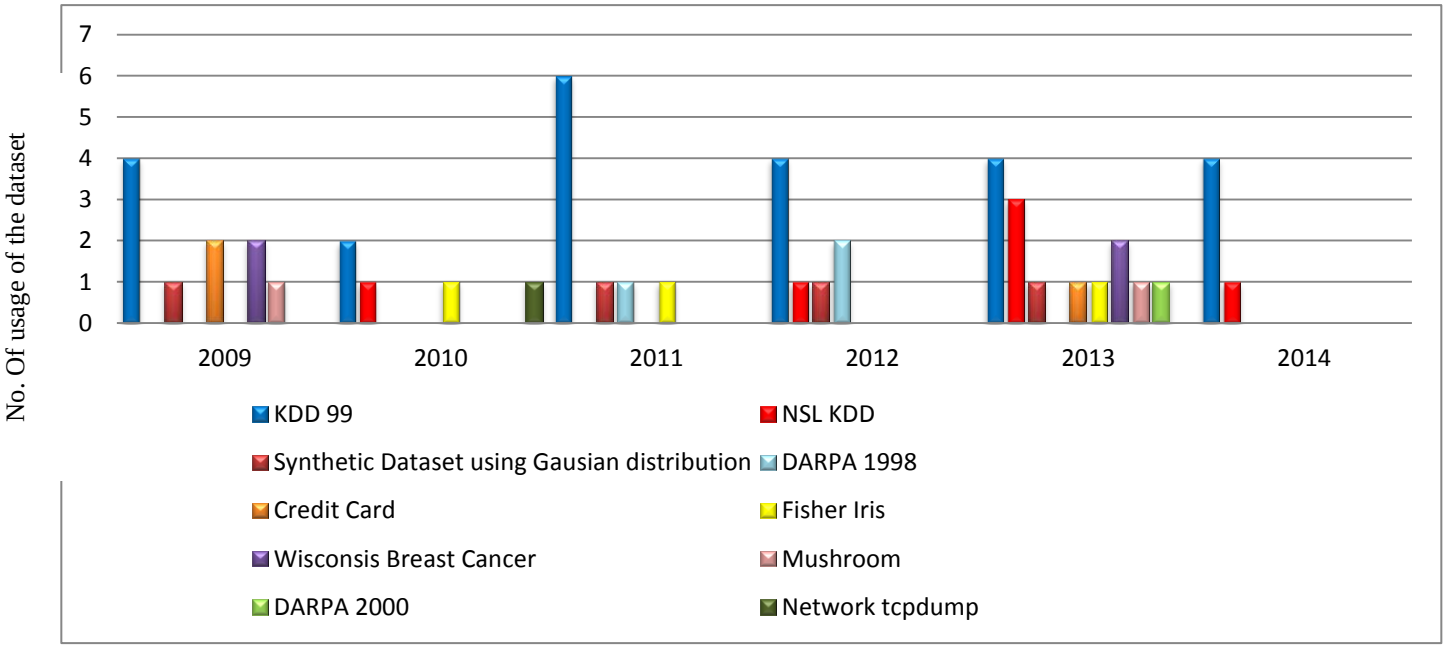

Fig. 4. Distribution of popular datasets over the years

TABLE VII. YEAR-WISE DISTRIBUTION OF FEATURE SELECTION CONSIDERED

Feature Selection Considered	2009	2010	2011	2012	2013	2014	Total
YES	1	3	4	7	4	2	21
NO	7	5	7	1	6	2	28

TABLE VIII. DISTRIBUTION OF RESEARCH PAPERS CONSIDERING THE FEATURE SELECTION STEP

Feature Selection	No. of research papers	Research papers
YES	21	A.m.chandrashekhar, k. (2013)[1]. adel sabryeesa, z. o. (2014)[2]. Akhilesh Kumar Shrivas, A. K. (2014)[3]. Bose, A. A. (2012)[6] Carlos A. Catania, F. B. (2012)[9]. Inho Kang, M. K. (2012)[26]. Levent Koc, T. A. (2012)[29].M. Bahrololum, E. S. (2009)[30]. Mario Poggiolini, A. E. (2013)[32]. Min Seok Mok, S. Y. (2010)[34]. Prabhjeet Kaur, A. K. (2012)[38]. S. Devaraju, S. R. (2013)[42]. Seongjun Shin, S. L. (2013)[44]. Shun-Sheng Wang, K.-Q. Y.-C.-W. (2011)[47]. Siva S. Sivatha Sindhu, S. G. (2012)[48]. Su, M.-Y. (2011)[49]. Woochul Shim, G. K. (2010)[55]. Yang Yi, J. W. (2011)[56]. Yinhui Li, J. X. (2012)[57]. Yung-Tsung Hou, Y. C.-S.-M. (2010)[58]. Yusuf Sahin, S. B. (2013)[59].
NO	28	C.A. Laurentys, R. P. (2011)[7] Chengpo Mua, Y. L. (2010)[10] D. Sa´nchez, M. V. (2009)[12] Dahlia Asyiqin Ahmad Zainaddin, Z. M. (2013)[13]. Dewan Md. Farid, L. Z. (2013)[14] Dewan Md. Farid, M. Z. (2011)[15] Feng Jiang, Y. S. (2009)[16] G. Davanzo, E. M. (2011)[17] Gang Wang, J. H. (2010)[18] Gisung Kim, S. L. (2013)[19](Ravi Ranjan, 2014)[40] Guillermo L. Grinblat, L. C. (2013)[21] Han-Ching Wu, S.-H. S. (2010)[22] I.T. Christou, M. B. (2011)[24] Ilhan Aydin, M. K. (2010)[25]. Jun Ma, J. L. (2009)[27] Kamran Shafi, H. A. (2009)[28]Mao Ye, X. L. (2009)[31]. Mazyar Mohammadi Lisehroodi, Z. M. (2013)[33]. Mohammad Saniee Abadeh, H. M. (2011)[36]. Muna Mhammad T. Jawhar, M. M. (2010)[37]. Prof. D.P. Gaikwad, S. J. (2012)[39] Seung Kim, N. W.-H. (2011)[45]. Seungmin Lee, G. K. (2011)[46]. Su-Yun Wua, E. Y. (2009)[50]. Tich Phuoc Tran, L. C. (2009)[52]. Wang Dawei, Z. F. (2011)[53]. Wenying Feng, Q. Z. (2014)[54].

IV. Discussion, Future Work and Conclusion

Uses of different classifier techniques in intrusion detection system is an emerging study in machine learning and artificial intelligence. It has been the attention of researchers for a long period of time. This paper has identified 49 research papers related to application of using different classifiers for intrusion detection published between 2009 and 2014. Though this survey paper cannot claim to be an in-depth study of those studies, but it presents a reasonable perspective and shows a valid comparison of works in this field over those years. The following issues could be useful for future research:

- Removal of redundant and irrelevant features for the training phase is a key factor for system performance. Consideration of feature selection will play a vital role in the classification techniques in future work.
- Feature selection has many algorithms to work with. Using different feature selection algorithms and working with the best possible one will be helpful for the classification techniques and also increase the consideration of feature selection step in intrusion detection.
- Uses of single classifiers or baseline classifiers in performance measurement can be replaced by hybrid or ensemble classifiers.

References

[1] A.M.Chandrashekhar, K. (2013). Fortification of hybrid intrusion detection system using variants of neural networks & support vector machines. International Journal of Network Security & Its Applications (IJNSA) .

[2] Adel Sabry Eesa, Z. O. (2014). A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Systems with Applications,ELSEVIER .

[3] Akhilesh Kumar Shrivas, A. K. (2014). An Ensemble Model for Classification of Attacks with Feature Selection based on KDD99 and NSL-KDD Data Set. International Journal of Computer Applications .

[4] Anderson, J. (1995). An introduction to neural networks. Cambridge: MIT Press.

[5] Bernhard E Boser, I. M. (1992). A Training Algorithm for Optimal Margin Classiers. Proceedings of the 5th Annual ACM Workshop on Computational , 144-152.

[6] Bose, A. A. (2012). THE COMBINED APPROACH FOR ANOMALY detection using neural networks & clustering techniques. Computer Science & Engineering: An International Journal (CSEIJ) .

[7] C.A. Laurentys, R. P. (2011). A novel Artificial Immune System for fault behavior detection. Expert Systems with Applications,ELSEVIER .

[8] C.M.Bishop. (1995). Neural networks for pattern recognition. England: Oxford University.

[9] Carlos A. Catania, F. B. (2012). An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection. Expert Systems with Applications,ELSEVIER .

[10] Chengpo Mua, Y. L. (2010). An intrusion response decision-making model based on hierarchical. Expert Systems with Applications,ELSEVIER .

[11] Chih-Fong Tsai, Y.-F. H.-Y.-Y. (2009). Intrusion detection by machine learning: A review. expert systems with applications,ELSEVIER .

[12] D. Sa´nchez, M. V. (2009). Association rules applied to credit card fraud detection. Expert Systems with Applications,ELSEVIER .

[13] Dahlia Asyiqin Ahmad Zainaddin, Z. M. (2013). HYBRID OF FUZZY CLUSTERING NEURAL NETWORK OVER NSL DATASET FOR INTRUSION DETECTION SYSTEM. Journal of Computer Science.

[14] Dewan Md. Farid, L. Z. (2013). An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams. Expert systems with Applications,ELSEVIER .

[15] Dewan Md. Farid, M. Z. (2011). Adaptive Intrusion Detection based on Boosting and. International Journal of Computer Applications .

[16] Feng Jiang, Y. S. (2009). Some issues about outlier detection in rough set theory. expert systems with application,ELSEVIER .

[17] G. Davanzo, E. M. (2011). Anomaly detection techniques for a web defacement monitoring service. Expert Systems with Applications,ELSEVIER .

[18] Gang Wang, J. H. (2010). A new approach to intrusion detection using Artificial Neural Networks and. Expert Systems with Applications,ELSEVIER .

[19] Gisung Kim, S. L. (2013). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications,ELSEVIER .

[20] Gisung Kim,J.C,S.K. (2012). A congestion-aware IDS node selection method for wireless sensor networks. IJDSN.

[21] Guillermo L. Grinblat, L. C. (2013). Abrupt change detection with One-Class Time-Adaptive Support Vector Machines. Expert Systems with Applications,ELSEVIER .

[22] Han-Ching Wu, S.-H. S. (2010). Neural networks-based detection of stepping-stone intrusion. Expert Systems with Applications,ELSEVIER .

[23] Haykin, S. (1999). Neural networks: A comprehensive foundation (2nd Edition). New Jersey: Prentice Hall.

[24] I.T. Christou, M. B. (2011). Detecting fraud in online games of chance and lotteries. Expert Systems with Applications,ELSEVIER .

[25] Ilhan Aydin, M. K. (2010). Chaotic-based hybrid negative selection algorithm and its applications in fault. expert systems with applications,ELSEVIER .

[26] Inho Kang, M. K. (2012). A differentiated one-class classification method with applications to intrusion detection. Expert Systems with Applications,ELSEVIER .

[27] Jun Ma, J. L. (2009). Information inconsistencies detection using a rule-map technique. Expert systems with applications,ELSEVIER .

[28] Kamran Shafi, H. A. (2009). An adaptive genetic-based signature learning system for intrusion detection. Expert Systems with Applications, ELSEVIER .

[29] Levent Koc, T. A. (2012). A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier. Expert Systems with Applications,ELSEVIER .

[30] M. Bahrololum, E. S. (2009). ANOMALY INTRUSION DETECTION DESIGN USING. International Journal of Computer Networks & Communications (IJCNC) .

[31] Mao Ye, X. L. (2009). Projected outlier detection in high-dimensional mixed-attributes data set. Expert systems with applications,ELSEVIER .

[32] Mario Poggiolini, A. E. (2013). Application of the feature-detection rule to the Negative Selection Algorithm. Expert Systems with Applications,ELSEVIER .

[33] Mazyar Mohammadi Lisehroodi, Z. M. (2013). A HYBRID FRAMEWORK BASED ON NEURAL NETWORK MLP AND K-MEANS CLUSTERING FOR INTRUSION DETECTION SYSTEM. Proceedings of the 4th International Conference on Computing and Informatics, ICOCI 2013 (p. Paper No. 020). Sarawak, Malaysia: Universiti Utara Malaysia.

[34] Min Seok Mok, S. Y. (2010). Random effects logistic regression model for anomaly detection. Expert Systems with Applications,ELSEVIER .

[35] Mitchell, T. (1997). Machine learning. New york: MacHraw Hill.

[36] Mohammad Saniee Abadeh, H. M. (2011). Design and analysis of genetic fuzzy systems for intrusion detection in. Expert Systems with Applications,ELSEVIER .

[37] Muna Mhammad T. Jawhar, M. M. (2010). Design Network Intrusion Detection System using hybrid. International Journal of Computer Science and Security.

[38] Prabhjeet Kaur, A. K. (2012). MADAM ID FOR INTRUSION DETECTION USING DATA MINING. International Journal of Research in IT & Management,IJRIM .

[39] Prof. D.P. Gaikwad, S. J. (2012). Anomaly Based Intrusion Detection System Using Artificial Neural Network & Fuzzy clustering. International Journal of Engineering Research & Technology (IJERT) .

[40] Ravi Ranjan, G. S. (2014). A NEW CLUSTERING APPROACH FOR ANOMALY INTRUSION DETECTION . International Journal of Data Mining & Knowledge Management Process (IJDKP) .

[41] Rhodes, B. M. (2000). Multiple self-organizing maps for intrusion detection. Baltimore, MD.

[42] S. Devaraju, S. R. (2013). DETECTION OF ACCURACY FOR INTRUSION DETECTION SYSTEM USING NEURAL NETWORK CLASSIFIER. International Journal of Emerging Technology and Advanced Engineering .

[43] Sahoo, R. R. (2014). A NEW CLUSTERING APPROACH FOR ANOMALY INTRUSION DETECTION. International Journal of Data Mining & Knowledge Management Process (IJDKP) .

[44] Seongjun Shin, S. L. (2013). Advanced probabilistic approach for network intrusion forecasting and detection. Expert Systems with Applications,ELSEVIER .

[45] Seung Kim, N. W.-H. (2011). Fast outlier detection for very large log data. Expert Systems with Applications,ELSEVIER .

[46] Seungmin Lee, G. K. (2011). Self-adaptive and dynamic clustering for online anomaly detection. Expert Systems with Applications,ELSEVIER.

[47] Shun-Sheng Wang, K.-Q. Y.-C.-W. (2011). An Integrated Intrusion Detection System for Cluster-based Wireless. Expert Systems with Applications.

[48] Siva S. Sivatha Sindhu, S. G. (2012). Decision tree based light weight intrusion detection using a wrapper approach. Expert Systems with Applications,ELSEVIER .

[49] Su, M.-Y. (2011). Real-time anomaly detection systems for Denial-of-Service attacks by weighted. Expert Systems with Applications,ELSEVIER .

[50] Su-Yun Wua, E. Y. (2009). Data mining-based intrusion detectors. Expert Systems with Applications,ELSEVIER .

[51] Tax, D. &. (1999). Data domain description using support vectors. Proceedings of the european symposium on artificial neural networks, 251-256.

[52] Tich Phuoc Tran, L. C. (2009). Novel Intrusion Detection using Probabilistic Neural. (IJCSIS) International Journal of Computer Science and Information Security.

[53] Wang Dawei, Z. F. (2011). Evolving boundary detector for anomaly detection. Expert Systems with Applications.

[54] Wenying Feng, Q. Z. (2014). Mining network data for intrusion detection through combining SVMs with ant colony networks. Future Generation Computer Systems,ELSEVIER .

[55] Woochul Shim, G. K. (2010). A distributed sinkhole detection method using cluster analysis. Expert Systems with Applications,ELSEVIER .

[56] Yang Yi, J. W. (2011). Incremental SVM based on reserved set for network intrusion detection. Expert Systems with Applications .

[57] Yinhui Li, J. X. (2012). An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert Systems with Applications,ELSEVIER .

[58] Yung-Tsung Hou, Y. C.-S.-M. (2010). Malicious web content detection by machine learning. expert systems with applications,ELSEVIER .

[59] Yusuf Sahin, S. B. (2013). A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications,ELSEVIER.

[60] Zimmermann, H.-J. (2010). Fuzzy set theory. Advanced Review John Wiley & Sons, Inc