HETEROGENEOUS ENSEMBLE MODELS FOR GENERIC CLASSIFICATION

Balogun A. O., Balogun A. M., Sadiku P. O., Adeyemo V. E.

Department of Computer Science, University of Ilorin, Ilorin, Nigeria

Corresponding author: Balogun A. O., bharlow058@gmail.com

ABSTRACT: This paper presents the application of some data mining techniques in the field of health care and computer network security. The selected classifiers were used individually and also, they were ensemble methods using four different combinations for the purpose of classification. Naïve Bayes, Radial Basis Function and Ripper algorithms were selected and the ensemble methods were majority voting, multi-scheme, stacking and Minimum Probability. The KDDCup'99 dataset was used as the benchmark for computer network security, while for the health care, breast cancer and diabetes dataset from the WEKA repository were used. All experiments and simulations were carried out, analyzed and evaluated using the WEKA tool. The Multi-scheme ensemble method gave the best accuracy result for the KDD dataset (99.81%) and the breast cancer dataset (73.08%) but its value of (75.65%) on breast cancer is the least of them all. Ripper algorithm gave the best result accuracy (99.76%) on KDD dataset amongst the base classifier but it was slightly behind in the breast cancer and diabetes dataset.

KEYWORDS: Ensemble, Data mining, Classification, KDD Dataset, Machine Learning.

1.0 INTRODUCTION

With our diversified community, culture, industry, and perspective, the application and usage of an important and significant idea or product is eminent; and this application and usage should also be diversified. Over the years data mining had evolved and recently it has been largely promoted but mostly the domain of computer network security had received much attention from this field. Other sectors of human diversity such as health care – which is an important part of our society, should receive even more attention.

This paper in the wise of the application of machine learning algorithms to computer network security developed a network intrusion detection system using the aforementioned base classifiers and further ensemble the classifiers having selected a suitable dataset for the purpose. The issue of developing a model for intrusion detection is of great concern and importance regarding the massive usage of computers and its network by various small, medium and large enterprises and even by the government. The enormous flow of information is possible via the usage of computers and its network; collection and analysis of data and information is also possible via is prudent means. Computers and its network had significantly change and help improve the way of human life in so many ways. Jobs are done faster with accurate, efficiency and effectiveness, flow of information is now flexible and fast, the whole world is a global village etc., nevertheless the security of these means is of great importance as a breach of it may lead to a great havoc. This problem is the basis of the development of intrusion detection system – not just a rigid intrusion detection system but one that is flexible and adaptive, that has the ability to learn. Also, the improvement of such intrusion detection system is paramount as knowledge is being accumulated on daily basis.

More so, the health care sector of human life is more importance of than the sector discussed above. There are several deadly diseases, many are not been diagnosed in time and even so that are been diagnosed takes longer time before the test result is being known – the result of the test that must have been conducted for days or weeks back. The application of machine learning algorithms in this sector will improve the way activities can be carried out. A sure predictive standard can be developed to help fasten the treatment of a patient, for instance a patient showing some symptoms of diabetes will be subjected to several tests before he or she can be diagnosed for the disease. If there is in place, a relatively sure predictive model that predict the whether or not a patient has the disease, treatment can be administered even before the result of the test comes out as this might save the patient life in time before the result is being known.

Succinctly, this paper present model for intrusion detection system and both diabetes and breast cancer predictive models. These models were all developed using three types of classifiers: Naïve Bayes, Radial Basis Function Network, and Ripper (JRip) which are selected from different methods of classifiers ; these classifiers are furthered ensemble using four different combinations: Multi-scheme, majority voting, stacking and minimum probability. The

selected classifiers are heterogeneous and are combined in four different methods using three types of datasets.

The reminder of this paper is segmented thus: Section 2 talked about related works, section 3 explains the methodology of the study while section 4 discussed the results and finally section concludes and summarizes the paper.

2.0 RELATED WORKS

The proposed paper by Santana et al. ([S+10]) compared the use of two optimization techniques in heterogeneous ensembles to find out whether an ensemble built with 3, 6 and 12 individual classifiers together with or without feature selection algorithm will perform better. Results showed that an ACO optimization outperformed GA in ensembles built with fewer individual classifiers while GA outperformed ACO in ensembles constructed with more individual classifiers.

In paper by Neto and Canuto [(NC14]), the use of Meta-learning technique and multi-objective optimization was proposed to make the ensemble system to investigate how the initial configuration of an ensemble would affect the outcome of NSGA II optimization algorithm. Results indicate that when Meta-learning is used, a more accurate ensemble system is obtained in more than 50% of the cases analyzed. An empirical investigation of Rand, Meta, and Equal was conducted, in which the lowest error rate and statistically significant result was obtained with Meta.

The author Wang ([Wan10]) developed a heterogeneous ensemble and a framework for constructing different kinds of ensemble for classifying spam emails. Results indicate that the heterogeneous ensemble can increase diversity as well as performance when compared to individual classifiers and other ensemble models. Meanwhile, Wang ([Wan08]) focuses on examining how the ensemble accuracy was impacted by what components and the degree of their effects. Three results were found, where they found that as the number of base classifiers increases, so does the diversity. Secondly, in terms of the accuracy, as the diversity increases, the accuracy also increases when voting strategy is used. Their final finding indicates that as the number of members' increase, so does accuracy and diversity but the accuracy increases even further when odd numbers are used in building the ensembles instead of even numbers.

Gupta and Thakkar ([GT14]) compared two evolutionary algorithms based on stacking ensembles optimization techniques. Result showed that ACO algorithm is more flexible in terms of meta-classifier selection, has a larger search and GA ensemble can only find the best classifier for some dataset if it is either majority voting scheme or model tree. They concluded that GA is superior in terms of accuracy than ACO while ACO is more proficient than GA.

Anwar, Qamar and Muzaffar Qureshi ([AQM14]) proposed an ensemble approach, which was used to develop a global optimization method for classification models that aims to improve the accuracy of many classifiers on any given dataset. Results showed an overall improvement in all the classifiers, some high and others are very low, varying between 1% to 3% depending upon the complexity of the algorithm and how it handles bias and variance.

A research work carried out by Borji ([Bor07]), proposed the combination of classification approach for intrusion detection. The author fused the outputs of four base classifier using three combination strategies: Bayesian averaging, majority voting and a belief measure. The later output support the superiority of his proposed approach compare with single classifier for detecting intrusion.

In Zhang, Jiang, and Kamel ([ZJK04]), two hierarchical neural network frameworks; serial hierarchical IDS (SHIDS) and parallel hierarchical IDS (PHIDS), were proposed. Backward Propagation Learner (BPL) and Radial Basis Function (RBF) were the two important learning algorithms used in these neural networks. Authors have shown that BPL has a slightly better performance than RBF in the case of misuse detection, while the RBF takes less training time. On the other hand RBF shows a better performance in the case of anomaly detection.

Wua-Hua, Sheng-Hsun, & Hwang-Pin ([WSH05]) proposed Artificial Neural Networks (ANN) and support vector machine (SVM) algorithms for intrusion detection with frequency-based encoding method. On the chosen DARPA dataset, they used 250 attacks and 41,426 normal sessions. The percentage of detection rate (%DR) they archived were between 43.6% and 100% while percentage of false positive rate (%FPR) varied from 0.27% to 8.53% using different thresholds.

Pavel, Patrick, Christin, and Reick ([P+05]) proposed an experimental framework for comparative analysis of both supervised and unsupervised learning techniques including C.45, multi-layer perceptron (MLP), K-nearest neighbor (KNN), etc. The best result they attained was 95% DR and 1% FPR using C.45 algorithm. Mukkamala, Sung, and Abraham, ([MSA05]) also presented an ensemble method for intrusion detection. They considered two types of classifiers; Artificial Neural Network (ANN) and Support Vector Machine (SVM).

However, this paper looks into the ensemble of heterogeneous machine learning algorithms on different datasets cutting across the scope of computer network security and health care. The datasets are the famous KDDCup'99, Diabetes (from the WEKA dataset) and breast cancer (from the WEKA dataset) respectively. For the domain of intrusion detection, the experiment will be done by evaluating the classifiers performance individually and ensemble methods on the KDDCup'99 dataset thereby exposing its potency for categorizing any kind of attack, while the experiment will be repeated for the health care field with different dataset from diabetes and breast cancer to classify whether or not it is positive or negative in the case of diabetes and recurrence or not in the case of breast cancer respectively.

3.0 METHODOLOGY

For this study, the chosen classifiers belong to the classification category of the machine learning techniques. These algorithms are predictive classifiers model, they study the attributes of a given dataset for a particular label with several instances and then predict the outcome of an instance without a label. The selected classifiers (Naïve Bayes, Radial Basis Function Network and RIPPER) for this study were trained and tested individually and also they were all ensemble as base learner for ensemble processes. More so, the data preprocessing stage was carried out using "Resample technique". This techniques was used on huge datasets, it filters the dataset by randomly resampling the dataset into a predefine percentage of its original size.

As this study is aimed at diversifying the application of machine learning techniques – in both single and meta-classification processes, the datasets used correlate with the domain of computer network security and health care services. Three datasets were selected, KDDCup'99, Diabetes, and Breast Cancer; and the algorithms were trained and tested with the three datasets individually and also being ensemble. The ensemble methods are of four various form vis-à-vis multi-scheme, majority voting, stacking and minimum probability.

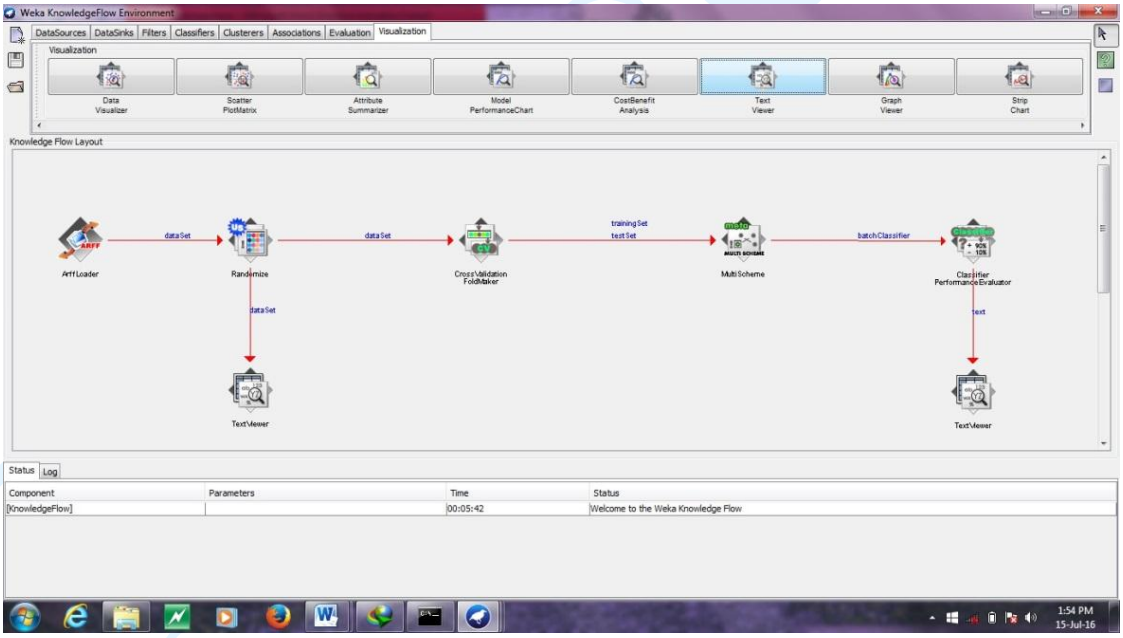

Figure 1: Proposed System Architecture (WEKA)

3.1 Proposed System Activity Architecture

Arff Loader: This module loads the datasets which will be used to train and test the algorithm.

Attribute Selection: This module see to data preprocessing. The Resample technique was used to resize the original dataset of any huge dataset

Cross Validation Maker: This is the process of breaking the dataset in 10 fold (which is the default fold) and passing the fold one after the other to the classifier.

Meta-Classification: This is where the base learners are selected and combined to form an ensemble method.

Classifier Performance Evaluator: This module evaluates and measures the performance of the selected algorithm, either a single or ensemble classifier.

Text Viewer: This module receives and display outputs in text format from any module.

3.2 Performance Evaluation

The results of the classifiers used will be evaluated and measure using the following parameters: correctly classified instances (%), incorrectly classified instances (%),TP (True Positive) rate, FP (False Positive) rate, and Kappa Statistics (is a chance-corrected measure of agreement between the classifications and the true classes. A value greater than 0 means that the classifier is doing better than chance).

3.3 Evaluation Setup

The experiments were carried out on a HP probook 6470b laptop with the following configurations Intel(R) Core(TM)i5-3230M, CPU 2.60GHz, 6GB RAM (5.55 GB usable), 64-bit operating system whose platform is Microsoft Windows7 Professional (Service Pack 1) . The latest Weka – an open source machine learning package was used for setting up the experimental and evaluation environment (Weka 3.6.11).

3.4 Data Preprocessing

Resample technique: is a supervised instance filter that can be used to randomly select instances from a larger set of instances. It produces a random subsample of a given dataset either with replacement or without replacement.
This technique was only applied on the KDDCup'99 dataset as it is huge in nature.

3.5 Classifier Algorithm

The classifiers that were used are heterogeneous which makes this research work a benchmark.
Naïve Bayes: is based on Bayes' rule and "naively" assumes independence. For this experiment, it was selected and all its parameters were set to false.
Radial Basis Function: implements a normalized Gaussian Radial basis function network. We ran RBF with the following parameters: clustering Seed = 1, debug = False, maxIts = -1, minStdDev = 0.1, numClusters = 2, and ridge = 1.0E-8.
Ripper (JRip): implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction, is an optimized version of IREP. It is based on association rules with reduced error pruning (REP). It parameters was set to: checkErrorRate = True; debug = False, folds = 3, minNo = 2.0, optimzations =2, seed = 1, and usePruning = True.
Multi-Scheme: is an ensemble class for selecting classifier from among several using cross validation on the training data or the performance on the training data. Its parameters after selecting the base classifiers were: debug = False, numFolds = 0; seed = 1.
Majority Voting: is an ensemble method for combining classifiers that simply combine the predictions of base learners via voting.
Stacking: combines the predictions of base classifiers together and make predictions. It learns how to combine and make predictions through the usage of a meta-learner.
Minimum Probability: is another combination technique for ensemble that makes predictions based on minimum probability of the outcomes of the base learners.

4.0 RESULTS AND DISCUSSION

The data preprocessing module was applied to the KDDCup'99 datasets, as the dataset contains more than 400,000 connections. The application of the technique on the dataset was to the produce another random subsampled dataset with lower connections with exactly 19,490 connections having instances with all form of labels present in the original dataset.

Table 1: Details of the datasets used for classifiers evaluation

Dataset	No of attributes (with label)	Total No of Actual Connections	No of Connections Used for the Experiment
KDD Cup	42	487,271	19,490
Diabetes	9	768	768
Breast Cancer	10	286	286

Table 2: Performance evaluation of the classifier algorithms – correctly classified instances, incorrectly instances, TP rate, FP rate, and Training Time (TT)

Classifiers	Datasets / Evaluation Parameters	KDD	Diabetes	Breast Cancer
Naïve Bayes	Correctly Classified Instances (%)	94.1508	76.3021	71.6783
	Incorrectly Classified Instances (%)	5.8492	23.6979	28.3217
	True Positive Rate	0.942	0.763	0.717
	False Positive Rate	0	0.307	0.446
	Kappa Statistics	0.9033	0.4664	0.2857
Radial Basis Function	Correctly Classified Instances (%)	99.3792	75.3906	70.979
	Incorrectly Classified Instances (%)	0.6208	24.6094	29.021
	True Positive Rate	0.994	0.754	0.71
	False Positive Rate	0.001	0.345	0.517
	Kappa Statistics	0.9895	0.4303	0.2177
JRip (Ripper)	Correctly Classified Instances (%)	99.764	76.0417	70.979
	Incorrectly Classified Instances (%)	0.236	23.9583	29.021
	True Positive Rate	0.998	0.76	0.71
	False Positive Rate	0	0.322	0.489
	Kappa Statistics	0.996	0.4538	0.2409
Multi-Scheme	Correctly Classified Instances (%)	99.8102	75.651	73.0769
	Incorrectly Classified Instances (%)	0.1898	24.349	26.9231
	True Positive Rate	0.988	0.757	0.731
	False Positive Rate	0.001	0.316	0.494
	Kappa Statistics	0.9968	0.4513	0.2686
Majority Voting	Correctly Classified Instances (%)	99.6716	77.2135	72.3776
	Incorrectly Classified Instances (%)	0.3284	22.7865	27.6224
	True Positive Rate	0.997	0.772	0.724
	False Positive Rate	0	0.311	0.483
	Kappa Statistics	0.9945	0.4791	0.266
Stacking	Correctly Classified Instances (%)	95.844	76.1719	72.3776
	Incorrectly Classified Instances (%)	4.156	23.8281	27.6224
	True Positive Rate	0.958	0.762	0.724
	False Positive Rate	0.029	0.335	0.456
	Kappa Statistics	0.9294	0.4484	0.288
Minimum Probability	Correctly Classified Instances (%)	99.1842	77.3438	71.6783
	Incorrectly Classified Instances (%)	0.7132	22.6563	28.3217
	True Positive Rate	0.993	0.773	0.717
	False Positive Rate	0	0.327	0.52
	Kappa Statistics	0.988	0.4721	0.2246

The Multi-scheme ensemble method gave the best accuracy result for the KDD '99 dataset (99.81%) amongst other ensemble method and the base classifiers. The Ripper classifier gave a good accuracy result of 99.76% on the KDD '99 dataset as compared to other ensemble methods which gave less (stacking: 95.84%; Majority Vote: 99.67%; Minimum probability: 99.18%), which kind of go against the normal assumptions that all ensemble methods are better than individual classifiers. Results from the diabetes dataset are very similar in terms of the performance of the ensemble methods and the base classifiers when compared.

The Minimum probability ensemble method gave the best (77.34%) followed by the Majority voting (77.21%) and then the Naïve Bayes (76.30%) which had a better accuracy compared to stacking and multi-scheme ensemble methods. In the case of the breast cancer dataset, the ensemble methods gave a better result over the base classifiers with the Multi-scheme ensemble method having 73.08% with the Radial Basis Function and the Ripper having the lowest accuracy of 70.98% of them all.

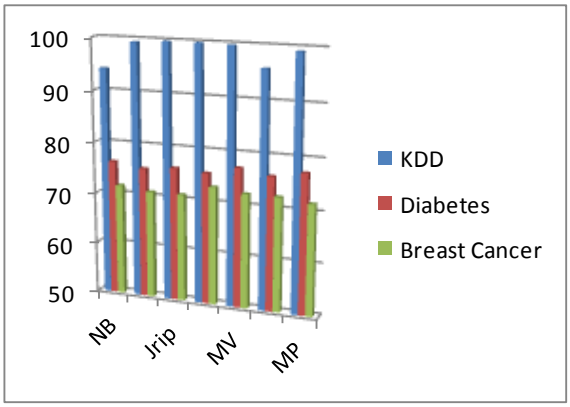

Figure 2: Accuracy of the Algorithms

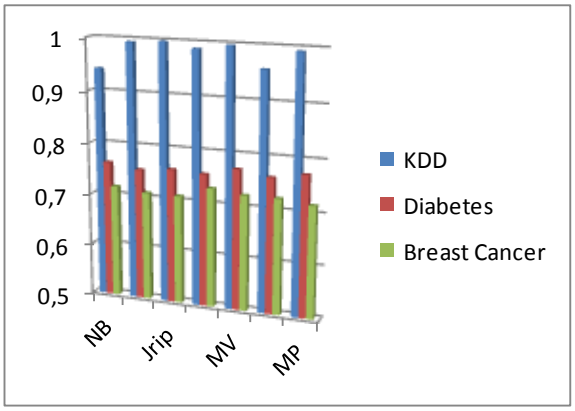

Figure 3: True Positive of the Algorithms

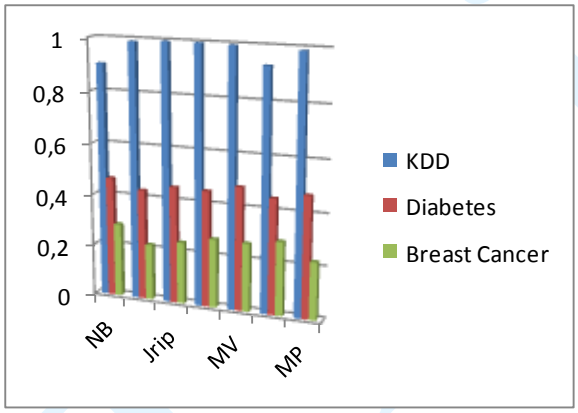

Figure 4: Kappa Statistics of the Algorithm

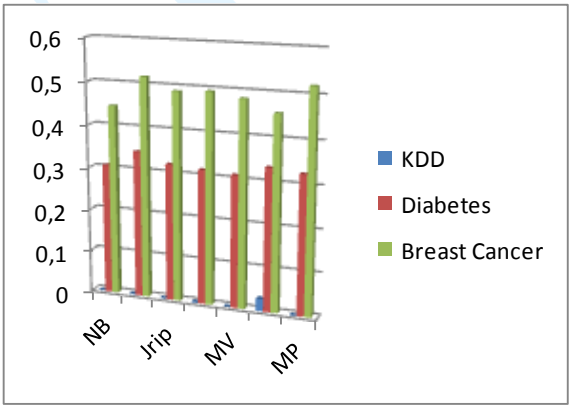

Figure 5: False Positive the Algorithms

5.0 CONCLUSION

From the various results, the choice of data preprocessing technique used on the KDDCup'99 dataset was chosen in order to see the effect of instance resampling instead of the popular attribute or feature selection process. The performances of the algorithms were well above average in terms of accuracy (correctly classified instances). The results from the experiments revealed that the predictive accuracy of the various machine learning algorithms are reliable as the performance of the classifiers on the KDDCup'99 dataset valued well above 90% while the corresponding values for both diabetes and breast cancer datasets ranges above 70%. As some classifiers outperformed some certain ensemble methods, it is safe to say the assumptions of ensemble methods are better than individual classifier is subject to the type of ensemble method and the base classifier. It is also evident that the choice of ensemble method for a particular classification problem depends on the kind of dataset both in volume and value. However, the researchers strongly recommended further study to be carried out on other field of studies using other types of classifiers, instance selection or attribute reduction technique and also the ensemble methods in order to shed more light on the disparity between ensemble methods and base classifiers.

REFERENCES

[AQM14] **H. Anwar, U. Qamar, A. W. Muzaffar Qureshi** – *Global Optimization Ensemble Model for Classification Methods*. The Scientific World Journal, Hindawi Publishing Corporation. 2014.

[Bor07] **A. Borji** - *Combining Heterogeneous Classifier for Network Intrusion Detection*. ASIAN 2007, LNCS 4846, pp. 254 – 260. 2007.

[GT14] **A. Gupta, A. R. Thakkar** - *Optimization of Stacking Ensemble Configuration Based on Various Metaheuristic Algorithms*. In Advance Computing Conference (IACC), IEEE International pp. 444-451, IEEE. 2014.

[MSA05] **S. Mukkamala, A. H. Sung, A. Abraham** - *Intrusion Detection Using Ensemble of Soft Computing Paradigms*. Journal of Network and Computer Applications 28, 167–182 2005.

[NC14] **A. A. F. Neto, A. M. Canuto** - *Meta-Learning and Multi-Objective Optimization to Design Ensemble of Classifiers*, 2014 Brazilian Conference on Intelligent Systems. pp. 91 – 96, IEEE. 2014.

[P+05] **L. Pavel, D. Patrick, S. Christin, K. Rieck** - *Learning Intrusion Detection: Supervised or Unsupervised.* In: Roli, F., Vitulano, S. (eds.) ICIAP 2005. LNCS, vol. 3617, pp. 50–57. Springer, Heidelberg. 2005.

[S+10] **L. E. Santana, L. Silva, A. M. Canuto, F. Pintro, K. O. Vale** – *A Comparative An Analysis of Genetic Algorithm and Ant Colony Optimization to Select Attributes for a Heterogeneous Ensemble of Classifiers*, In Evolutionary Computation (CEC), 2010 IEEE Congress on pp. 1-8, IEEE**.**

[Wan08] **W. Wang** - *Some Fundamental Issues in Ensemble Methods.* In Neural Networks, IJCNN. IEEE World Congress on Computational Intelligence. IEEE International Joint Conference on pp. 2243-2250. IEEE. 2008.

[Wan10] **W. Wang.** - *Heterogeneous Bayesian Ensembles for Classifying Spam Emails*, The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, IEEE. (2010)..

[WSH05] **C. Wun-Hua, H. Sheng-Hsun, S. Hwang-Pin** - *Application of SVM and ANN for intrusion detection*. Computer. Operation. Research. 32(10), 2617–2634. 2005.

[ZJK04] **C. Zhang, J. Jiang, M. Kamel** - *Intrusion detection using hierarchical neural networks*. Pattern Analysis and Machine Intelligence Research Group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. 2004.