Machine learning approach for intrusion detection systems as a cyber security strategy for Small and Medium Enterprises

NEVILA BACI, KRESHNIK VUKATANA, MARIUS BACI
Department of Statistics and Applied Informatics
University of Tirana
Nënë Tereza Square, 4
ALBANIA

Abstract: Small and medium enterprises (SMEs) are businesses that account for a large percentage of the economy in many countries, but they lack cyber security. The present study examines different supervised machine learning methods with a focus on intrusion detection systems (IDSs) that will help in improving SMEs' security. The algorithms that are tested through a real dataset, are Naïve Bayes, Sequential minimal optimization (SMO), C4.5 decision tree, and Random Forest. The experiments are run using the Waikato Environment for Knowledge Analyses (WEKA) 3.8.4 tools and the metrics used to evaluate the results were: accuracy, false-positive rate (FPR), and total time to train and build a classification model. The results obtained from the original dataset with 130 features show a high value of accuracy, but the computation time to build the classification model was notably high for the cases of C4.5 (1 hr. and 20 mins) and SMO algorithm (4 hrs. and 20 mins). the Information Gain (IG) method was used and the result was impressive. The time needed to train the model was reduced in the order of a few minutes and the accuracy was high (above 95%). In the end, challenges that SMEs can have for choosing an IDS such as lack of scalability and autonomic self-adaptation, can be solved by using a correct methodology with machine learning techniques.

Key-Words: Intrusion detection systems, Machine learning, Small and medium enterprises, Cyber-security.

Received: August 9, 2021. Revised: January 7, 2022. Accepted: January 19, 2022. Published: January 20, 2022.

1 Introduction

This paper presents an overview of machine learning techniques applied in intrusion detection systems with a focus on Small and Medium Enterprises (SMEs). Recently, data breaches and cyber-attacks continue to increase not only in SMEs but in every business in the market. The growing number of attacks derives an increasing cost of dealing with them, that is why security should be a priority for the businesses. Most of the Intrusion Detection Systems (IDSs) in the market are signature-based and for this reason, the process of discovering new intrusions becomes a big challenge. The selection of appropriate classification algorithms for IDSs is a challenging task and has become a priority in the security field. A lot of techniques of machine learning have been proposed as a solution to improve the accuracy of signature-based methods[1]. These techniques have enormous potential and can be used to build robust models for the classification of malicious activities on the SME information systems. The IDSs must be capable to identify the existing malware or discover new ones.

Different factors should be considered when applying these techniques such as the dataset size and the presented features in the dataset, because they have a big impact on the system performance[2]. There are cases when irrelevant features present in the dataset, can lead machine learning techniques to different issues such as classification misleading, overfitting, generality reduction, model run-time enhancement, and processing complexity. When it comes to SMEs, one of the challenges to face with the IDSs is the presence of a false-positive rate (FPR) resulting in a high workload for analyzing the logs. SMEs having in place IDSs can reduce the streamline and improve system accuracy. Applying machine learning techniques can be a solution to the intrusion detection process. The classification of the attacks in different classes is the most important task performed by an IDS and can be performed using different machine learning techniques. These techniques must be properly tuned and not blindly applied to reduce complexity by not affecting the performance of the system.

Nowadays, SMEs are using different means of communication such as Cloud services, social media, mobile devices, etc. This leads to more breaches, rendering the SME systems more vulnerable. SMEs are more exposed to cyber-criminals than other big enterprises. The vulnerability of SMEs is shown by the number of breaches on their systems that for the year 2018 is increased by 424%[3]. Hackers are increasingly targeting more small businesses rather than big

ones. The main reason that SMEs are becoming a target for cybercriminals is the assumption that their security systems are less strong compared to the big companies. In SMEs, the vulnerabilities often arise because of not taking adequate cyber security measurements, mainly due to the lack of financial and human resources. By doing so, they increase the risk to guarantee the data confidentiality and integrity of their clients. For example, in the year 2019, around 58% of SMEs have been a victim of a cyberattack, resulting on average downtime for every breach in more than 8 hours[4]. In terms of money, these attacks are estimated to cost around $3 million, resulting in losing profits, but most important losing clients because of trustiness. On the other hand, big enterprises unlike SMEs have human resources, technical expertise, and finance to protect their information assets as explained in the Kshetri[5]. So, the solution is to increase the cyber security investment.

Machine learning techniques are wildly used in IDSs to achieve effectiveness with datasets that are not suffering from irrelevant, and redundant feature sets. The aim is to analyze the impact and consequences of cyber-attacks in an information system with a focus on SMEs, and to show the effectiveness of applying machine learning techniques in intrusion detection systems. For example, in cases when an attacker tends to gain access or interrupt normal operations of an information system, almost always he is trying to cause damage and malfunctions. Different supervised and unsupervised machine learning techniques are used to address the major challenges faced by IDSs such as Decision Tree algorithm (DTA) and Support Vector Machine (SVM) as shown by Ektefa research[6]. Some methods outperform others in terms of classification accuracy, but less interest is shown in computational time that is an important factor in choosing the right algorithm and is addressed in this work.

With the new General Data Protection Regulation (GDPR)[7], which came into force in May 2018, new regulations must be followed by enterprises during a data breach. If the company systems incur any data breach, it should be documented no later than 72 hours after having become aware of it. In these circumstances, implementing strong IDS can guarantee the enterprises to monitor the network or the systems for malicious activity and policy violations, and have the possibility to document it, for example through logs.

In this paper, the focus is to investigate the different machine learning techniques used in the context of IDS to ascertain the potential presence of any technique through experimental exploration which can be used for SME scenarios by showing the power of feature selection methods in improving the classification of different attacks into classes. The purpose is to show the effectiveness of using the right machine learning techniques for the IDS to solve the most significant challenges faced such as high computational time and low accuracy. To evaluate these two parameters on the IDSs, several experiments were conducted with real data, the Aegean Wi-Fi Intrusion Dataset (AWID) dataset[8]. Initially, the data were pre-processed, and then the relevant features were extracted to reduce the dimensionality of the dataset. These two steps were important for improving the classification accuracy and reducing the computation time. In the end, different machine learning methods were applied, and the results were compared through the metrics of accuracy, FPR, and total time to build the classification model.

2 Materials and Methods

2.1 Intrusion Detection Systems

Cyber security experts implement different methods to defend from malicious attacks like firewalls, Intrusion Prevention System (IPS), or IDS. The latter is one of the most essential components of computer security used to detect attacks before they are widely spread. An intrusion is classified as the set of actions aimed to compromise the security goals that are integrity, confidentiality, and availability of computer resources[9]. An IDS is a device or software that detects any malicious activity or attack on protected assets. It can analyze the collected data in a given network to identify malicious behavior or policy violations and then prepare a report for the system administrator to handle the intrusion, summarizing the functions of IDS such as:

- to monitor user and system activity;
- to detect attacks as soon as possible;
- to enforce the network traffic;
- to analyse statistical patterns;
- to audit of operating system.

There is also, a classification on types of IDS that are Network-based IDS (NIDS) or Host-based IDS (HIDS), depending on weather the system monitors a single host or a network[10].

2.1.1 HIDS

A HIDS relies heavily on audit trials, becoming limited in finding new attacks. It monitors and analyses the input/output packets from a single device performing log analyses, file integrity checking, policy monitoring, etc. In any case, HIDS tends to be desirable for some reasons. For example, because it can

Table 1: Difference between HIDS and NIDS.

Types	Advantages	Disadvantages
NIDS	Monitors multiple hosts at a time. Attacks of different hosts can be correlated. It does not decrease the host performance.	Problem with encrypted network traffic. Keep up with the network speed.
HIDS	Can analyse data for specific use in the host by using audit trials. Operates in environments that are encrypted.	Cross platform based. Can't see network traffic. Large cost in setting up.

Table 2: Comparison of detection methods in IDS.

Types	Advantages	Disadvantages
AIDS	Able to detect new attacks. Signature database can be updated based on new attacks.	Cannot handle encrypted packets. False-positive alarms are high. Difficult to classify alerts in different categories. Training phase is complex.
SIDS	False-positives are low. Better for detecting the known attacks. Simple design.	Cross platform based. Can't see network traffic. Large cost in setting up.

Table 3: AWID Dataset Characteristics[8].

Dataset	Purpose	No. of Records	Target variables
AWID-CLS-R-Trn	Training phase	1,795,575	4 classes
AWID-CLS-R-Tst	Testing phase	530,643	

monitor access to information in terms of "who accessed what", this system can trace the activities of a specific user and determine whether an attack has occurred or not. Moreover, this system is capable to operate in an encrypted computer environment. Since HIDS comes with the system, there are also cost advantages to using it among other systems.
On the other hand, there are some disadvantages. The main is that it cannot monitor the network traffic being heavily dependent on the operating system that is hosting it. More in detail, a typical HIDS must raise a flag or report the information about any malicious activity that occurred. This can be a downside on the performance of the hosting machine as HIDS uses the same resources and does not have a standalone operating system like other types of IDS[11].

2.1.2 NIDS

The NIDS offers a different approach. The data are collected from the network rather than from a single host. NIDS checks for misbehavior by inspecting Internet Protocol (IP) protocol-level activities and network packet structure to detect many IP-based Denial Of Service (DOS) attacks such as Transmission Control Protocol (TCP) Synchronized attacks.
The disadvantage with NIDS is that it has limited visibility within the host machine and there is no effective way to analyze the encrypted network traffic to defend the system. There exists available software or tools with different solutions such as Network Intrusion Detection & Prevention System (SNORT) or NetSTAT, a command-line network utility on Unix-like operating systems that monitor the network traffic in real-time.
Table 1. summarizes shortly the advantages and disadvantages of both, NIDS and HIDS. The first focuses more on vulnerability abuse while the second focuses on privilege abuse. From a financial perspective, NIDS costs less and is faster in time response than HIDS, because it monitors the traffic in real-time or close to real-time.

2.2 Intrusion Detection Approaches

Based on the detection method, IDSs can be principally classified into two main categories, signature-based and anomaly-based, but in general, some systems operate as a hybrid system.
Signature-based Intrusion Detection Systems (SIDS) detect attacks based on the most used method, the pattern matching technique. Patterns detected in IDSs are known as signatures. The system tries to match a new intrusion with the existing ones stored in the signature database, and when a match occurs an alarm is triggered. SIDS among experts is known also as Knowledge-Based Detection[12]. This system can detect already known attacks, whose signature already exists in the system, but they are incapable of detecting new attacks because their signatures are not stored in the database. The problem here is that the signature database must be updated frequently, otherwise, attacks whose signatures do not exist in the catalog are unlikely to be detected. In practice, SIDS gives a good classification accuracy for the detection of previously known attacks.
Anomaly-based Intrusion Detection Systems (AIDS) have drawn interest because they overcome the lim-

itations of SIDS. In AIDS a reliable behavior model is developed using different approaches such as Machine Learning, statistical methods, or knowledge-based methods. The observed behavior is compared with the data model and every significant deviation is an anomaly. These anomalies can be classified as intrusions. The statistical method deals with anomaly detection from randomness, while the knowledge-based method includes capturing the alleged behavior by network traffic instances and other relevant system data[13]. AIDS goes throw two processes: the training phase when a model is developed filled with the normal behavior data, and then in the testing phase, a new data set is used to test the capability of the system to detect the "not normal behavior" classified as an intrusion. In comparison to SIDS, AIDS is better when it comes to the chance to identify zero-day attacks because it does not depend on matching the data with patterns in the signature database. The main differences between these two types of systems are shown in Table 2.
Taking into consideration the advantages of both systems, a hybrid one can be implemented with both methods. This system can detect zero-day attacks and reduce the number of false alarms. In their research[13] found that no system was just signature or anomaly-based, IDSs are usually deployed as a hybrid system.

2.3 Datasets

The experiments conducted in this research are evaluated on the AWID Dataset[8]. This dataset is available and public. It is focused on 802.11 networks and was introduced in 2015. There are different datasets available for the intrusion detection systems, but this dataset is recommended to be used as it contains real data, captured through Wireless Local Area Network(WLAN) traffic in a packet-based format.
The data collected are around 37 million packets, captured in one hour. Originally this dataset has 155 attributes and one target variable. There are two types of datasets available, the "CLS" and the "ATK". The first type named "CLS" has four target classes: normal, impersonation, overflow, and injection. On the other dataset, the attacks are classified into 15 different categories. The dataset creators have produced for research purposes two different datasets for the first type "CLS", a complete and a reduced one. To simplify our work, and because the experiment runs on one personal computer (PC), the reduced version of the data is used. The properties of the dataset are shown in Table 3.

2.4 Data pre-processing

To enhance the classification accuracy, the data are pre-processed. First, all the string attributes are converted into numeric ones. Some missing values in different attributes were discovered in the dataset. There were not applied imputation methods for the missing values as the focus of this work is not on those techniques. The approach followed was to replace all the missing values with zero. Some machine learning techniques do not work with missing data, for this reason, the transformation of the data was obligatory to be done. Also, some attributes have the same constant values for all the instances, but they do not contain any relevant information to interfere in creating new classes. In this case, the attributes were simply discarded. Next, followed a process of normalizing the data to avert the feature influence of the measurement scale, transforming the raw data feature values between zero and one.

2.5 Algorithms for Machine Learning Classification

Thiis section describes some of the machine learning methods that were applied in the experiments. There are a lot of methods that might be applied in the dataset selected, but onlys some of them were chosen, evidenced during the literature review. The classification algorithms used in machine learning models for classifying a given dataset were as follows: Bayesian Network, SVM, C4.5 decision tree, and Random Forest.
The Bayesian Network method provides a graphical representation of different probabilistic relationships between variables[14]. Compared to other statistical methods, this method has some advantages such as:

- Graphical representation of the relationship between all the variables.
- Can be used to learn about causal relationships.
- Identifies random relationships.
- Addresses multidimensional statistical problems.

The dependencies between variables are presented graphically by a Directed Acyclic Graph (DAG) and a probability table. Bayesian networks are widely used because they provide an efficient method for preventing overfitting.
SVM is a linear model used for classification or regression problems where each data item is plotted as a point in n-dimensional space[15]. N is the number of attributes that a dataset has. This method is recommended in small and medium datasets. SVM uses a high dimensional feature space, a kernel function, and the training phase is based on the optimization theory. Sequential minimal optimization (SMO) is a typical algorithm for solving problems that arise during the training of SVMs[16].

The C4.5 decision tree is a model based on decision trees. It is used when dealing with supervised classification problems. C4.5 comes as an extension of Iterative Dichotomiser 3 (ID3), which has two limitations[17]. The first occurs when two or more cases with identical values belong to different classes. The second is related to the risk of overfitting. The decision tree is built using the concept of entropy. According to this algorithm, the attribute is selected as the tree node separates objects more productively compared to other attributes, thanks to the gain of information. This algorithm is based on probability concepts to create a complete data table. The most important benefit of using decision trees is comprehensibility, because of their visualization in the form of a tree. The tree is created by a recursive-separation algorithm, which in each non-terminal node, determines a value for a variable. In this manner, the remaining branches or classes have a better differentiation. The root of the tree is a question or a qualitative variable, which has several categories. The main disadvantage is that Decision trees suffer from overfitting when the dataset is small and from testing only one attribute at a time.
Random Forest is one of the most popular assembling supervised machine learning algorithms that are capable of unpruned classification or regression. The random forest creates decision trees on data samples and is very efficient in large datasets. Random forest models are robust to overfitting and are based on the bagging technique to combine the decision trees[18]. Compared to decision trees, random forest models are more difficult to be interpreted.

2.6 Experimental environment and Evaluation metrics

The experiments are run using WEKA 3.8.4 (Waikato Environment for Knowledge Analyses) on a PC with 4 Core CPU Intel i7-4900MQ. Steps involved in the experiments and the logical flow of the process are shown in Fig. 1.

Different metrics[19] can be used to evaluate the results of machine learning methods like accuracy or recognition rate, confusion matrix, recall, FPR, sensitivity or true positive rate (TPR), specificity, learning time, precision, and Receiver Operating Characteristic (ROC) curve. When it comes to IDS, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are important metrics to measure system reliability. The concepts of accuracy and FPR are defined by the metrics as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} (1)$$

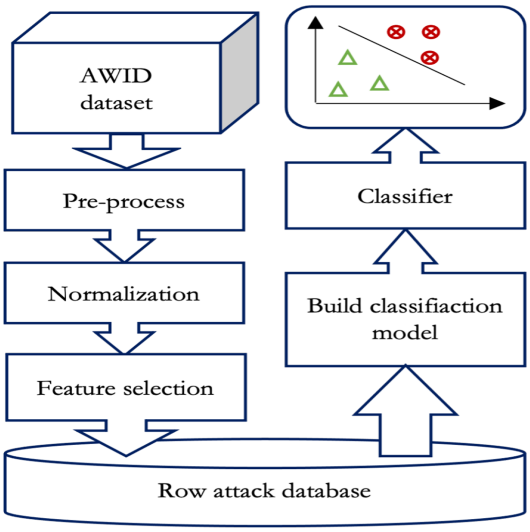

Figure 1: Steps Involved in the Experiments.

$$FPR = \frac{FP}{FP + TN} (2)$$

where accuracy (1) is the percentage of group records that are correctly classified and FPR (2) is the report of normal events classified as attacks and training time is the total time to build a classification model.

2.7 Feature selection process

To show the differences between the original dataset and a reduced one, we applied a feature selection method known as Information Gain (IG). In this technique, the significance of each feature is calculated as the relationship between the information gain factor following the class. This method is a ranked-based technique that does provide the final list of selected features and it is part of filter-based feature selection techniques.
Selection of the features is an important task during the training data set as it influences the correct classification. In literature is stressed the fact that reducing the feature space can often contribute to increasing system accuracy[20]. Table 4 shows the feature selection results based on the AWID datasets where is applied the IG feature selection method. All the selected machine learning methods are running in both datasets, the original one and the reduced dataset using the IG methods, to have a comparison.

3 Results and Discussion

Table 4 shows the results obtained from the original dataset with 130 variables, taking into consideration the accuracy and the FPR. The accuracy is high,

Table 4: Classifier Evaluation with/without applying feature selection method.

Method	Accuracy	FPR	Model Training Time	Feature Selection[No. of features]
Naïve Bayes	83.02%	0.138	00:04:47	Original dataset[130]
SMO	95.21%	0.512	04:19:23	
C4.5	93.99%	0.114	01:20:29	
Random Forest	95.98%	0.231	00:10:32	
Naïve Bayes	95.02%	0,143	00:01:01	IG[39]
SMO	96.69%	0,021	00:04:22	
C4.5	99.76%	0,072	00:05:29	
Random Forest	97.98%	0,029	00:01:67	

where the best result is obtained from the application of the Random Tree algorithm, which achieved a 95.98% accuracy, and the lowest is 83.02% obtained by the Naïve Bayes algorithm. The accuracy achieved is quite satisfactory, but the computation time is notably high, especially in the cases of C4.5 (1 hour and 20 minutes) and SMO algorithm (4 hours and 20 minutes). To reduce this time required for the training phase is used the IG method.

After the feature selection method is applied, the original dataset is reduced to 39 features. Using the IG method, the irrelevant features are removed. In the experiment, all the machine learning methods are re-executed. Results of the reduced dataset are shown in Table 4. The table shows that all the methods have a high accuracy rate above 95%. More in detail, there is respectively for Naïve Bayes, SMO, C4.5, and Random Forest an increment of 12%, 1,48%, 5,77%, and 2%.

The experimental results demonstrate also an improvement for all the algorithms regarding the FPR, except the Naïve that remain in the same order. Most important, the time needed to train the model is reduced drastically in order of 5 minutes also for the C4.5 and SMO method. This is a satisfactory result and shows that the feature selection algorithm has selected the right components, improving the performance of intrusion classification in terms of accuracy and computing time.

4 Conclusions and Further Work

SMEs are considered the backbone of the economy, as they have great potential for job creation, growth, and innovation. Cloud services, social media, and mobile devices bring a range of challenges and demands to SMEs affecting how business is done via different means of communication. Nowadays, it is common for SMEs to face cyber security incidents where personal data is stolen. Mostly, these security incidents go undetected or unreported. While big enterprises have long started employing cyber security strategies including IDS and IPS to defend against attacks, SMEs are still standing on the edge.

This paper aimed to address the challenges that SMEs must face related to IDS as the most important defense tool against network attacks. Several experiments were conducted applying different supervised machine learning methods to improve the detection of attacks in these systems. The algorithms applied were SMO, C4.5, Random Forest, and Naïve Bayes. All the experiments were conducted in the AWID dataset which has been widely used in the research for simulating the IDS. For all the experiments were used Weka 3.8 tools.

In our findings, once the feature selection method is applied, the dataset complexity and dimensionality are drastically reduced improving the system accuracy and reducing the computation time. This result is based on the experiments ran on the comparison between the original dataset and the one reduced with the IG feature selection method.

The dataset used in this paper is an unbalanced dataset where there is a huge difference between the number of records of the normal class (majority class) and the malicious class (minor class). In literature different balancing approaches are known and, in the future, these findings can be enriched with analysis on experiments on the effect of balancing in the system performance.

References:

[1] Feizollah, A., Anuar, N. B., Salleh, R. and Wahab, A. W. A., A review on feature selection in mobile malware detection, *Digital investigation*, Vol. 13, 2015, pp. 22–37.

[2] Faris, H., Hassonah, M. A., Ala'M, A.Z., Mirjalili, S. and Aljarah, I., A multi-verse optimizer approach for feature selection and optimizing parameters based on a robust system architecture, *Neural Computing and Applications*, Vol. 30, No. 8, 2018, pp. 2355–2369.

[3] Identity Breach Report 2019, [Online] Available: https://4iq.com/2019-identity-breach-report/ (Last accessed September 15, 2021).

[4] Yeng, P., Nimbe, P., Weyori, B., Solvoll, T. and Yang, B., Web Vulnerability Measures for SMEs, *NISK*, Vol 12., 2019, pp. 1-16.

[5] Kshetri, N., The Economics of Cyber-Insurance, *IT Professional*, Vol. 20, No. 6, 2018, pp. 9-14.

[6] Ektefa, M., Memar, S., Sidi, F., and Affendey, L. S., Intrusion Detection Using Data Mining Techniques, *In the proceedings of IEEE International Conference on Information Retrieval & Knowledge Management*, Exploring Invisible World, CAMP10, 2010, pp. 200-203.

[7] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), *Official Journal of the European Union* Vol. 59: L. 119, 2016, pp. 1-89.

[8] Kolias C., Kambourakis, G., Stavrou, A. and Gritzalis, S., Intrusion Detection in 802.11 Networks: Empirical Evaluation of Threats and a Public Dataset, *IEEE Communications Surveys & Tutorials*, Vol. 18, 2016, pp. 184-208.

[9] Heady, R., Luger, G., Maccabe, A. and Servilla, M., The architecture of a network level intrusion detection system, 1990.

[10] Othman, S., Alsohybe, N., Ba-Alëi, F., Zahary, A., Survey on Intrusion Detection Types, *International Journal of Cyber-Security and Digital Forensics*, Vol. 7, No. 4, 2018, pp. 444-462.

[11] Singh, A. P. and Singh, M., Analysis of Host-Based and Network-Based Intrusion Detection System, *International Journal of Computer Network and Information Security*, Vol. 8, 2014, pp. 41-47.

[12] Modi C., Patel D., Borisaniya B., Patel H., Patel A., and Rajarajan, M., A survey of intrusion detection techniques in cloud, *The Journal of Network and Computer Applications*, Vol. 36, No. 1, 2013, pp. 42–57.

[13] Butun, I., Morgera, SD. and Sankar, R., A survey of intrusion detection systems in wireless sensor networks. *IEEE Communications Surveys & Tutorials*, Vol. 16, No.1, 2014, pp. 266–282.

[14] Holmes, D.E. and Jain, L.C., *Innovations in Bayesian Networks. Studies in Computational Intelligence, Vol 156*, Springer, Berlin, Heidelberg, 2008.

[15] Cortes, C., Vapnik, V., Support-vector networks, *Machine Learning*, Vol. 20, 1995, pp. 273–297.

[16] Platt, J., *Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Machines*, Microsoft Research Technical Report, 1998.

[17] Salzberg, S.L., C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993, *Machine Learning*, Vol. 16, 1994, pp. 235–240.

[18] Breiman, L., Bagging predictors, *Machine Learning*, Vol. 24, 1996, pp. 123–140.

[19] Sokolova M., Japkowicz N., Szpakowicz S., Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. *In: Sattar A., Kang B. (eds) AI 2006: Advances in Artificial Intelligence. AI 2006. Lecture Notes in Computer Science, Vol. 4304*, Springer, Berlin, Heidelberg, 2006.

[20] Liu, H. and Motoda, H., *Computational methods of feature selection. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, Taylor & Francis, New York, 2007.