Data-driven Intrusion Detection System for Small and Medium Enterprises

Ogerta Elezaj, Sule Yildirim Yayilgan, Mohamed Abomhara, Prosper Yeng and Javed Ahmed
Department of Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway
Email: {ogerta.elezaj, sule.yildirim, mohamed.abomhara, prosper.yeng, javed.ahmed}@ntnu.no

***Abstract*—Small and Medium Enterprises (SMEs) have become targets of attack by cyber criminals in resent times. This paper therefore aim to address awareness and challenges of SMEs related to IDSs as the most important defense tool against sophisticated and ever-growing network attacks.An IDSs framework was actually introduced for efficient network anomaly detection for SMEs and provided experimental results to illustrate the benefits of the proposed framework. The proposed framework deals with one of the main challenges that IDSs of SMEs are facing, the lack of scalability and autonomic self-adaptation. Training, testing and evaluation of IDSs applying different machine learning (ML) techniques are presented. Results of experiments show that using feature selection approaches can lead to better classification accuracy and improved computational speed.**

***Index Terms*—intrusion detection, attacks, anomaly detection, cyber security, SME, machine learning**

I. Introduction

The European Commission (EC) [1] have adopted a definition of SMEs which is the only definition widely accepted. According to the EC, medium-sized enterprises are defined as enterprises with less than 250 employees, annual turnover of less than 50 million and a balance sheet total of less than 43 million. Small enterprises are defined as enterprises with less than 50 employees, a turnover less than 10 million and a balance sheet total less than 10 million. According to the Organization for Economic Cooperation and Development (OECD) [2] SMEs accounts for 99% of all enterprises in the European Union providing millions of jobs.

Due to the use of internet as a medium for business operations, SMEs are now exposed to new threats of cybercrime. They are becoming among the main targets of cyber-criminals because of their affiliation with bigger enterprises. Most SMEs do not consider themselves as having crucial data that is of interest to cyber-criminals. To guarantee success on the market, SEMs must guarantee data confidentiality and integrity, while securing and making available all communication channels during business operations.

According to [3], 58% of SMEs have suffered a cyberattack in the past 12 months. The average cost of an attack is estimated to be around $3 million with average downtime of more than 8 hours. About 20% of system breaches occur due to human errors. Attacks on SMEs result in loss of profit, clients, partners, trustworthiness, mitigation cost etc.

The main reasons why SMEs do not adapt appropriate cybersecurity practices are lack of investment, human resource and budget restrictions, which originate from enterprise owners and employees' low levels of security awareness [3]. As the threat landscape evolves, SMEs are facing the same security threats as large enterprises. However, they do not have sufficient IT security resources and expertise to continually monitor, detect and prevent security incidents, rendering them more and more prone to revenue losses and reputation damage.

Moreover, the new General Data Protection Regulation (GDPR) [4] enhances the need for awareness of personal data breaches, compliance and accountability. This new regulation puts pressure on SMEs to reevaluate data security policies and measures used for personal data protection. To comply with the regulation, enterprises need to re-evaluate their security systems for their effectiveness.

Data breaches caused by unauthorized users, accessing the network are considered as the main type of incidents that compromise personal data. Strong IDSs and IPSs that can monitor network traffic and trigger alarms for unusual or suspicious activity are among the best safeguards to avoid these risks. A properly tuned IDS guarantees enterprises deeper insight into network problems by providing visibility and control measures, which contribute to minimizing of threats and consequences of security breaches [5].Machine Learning (ML) techniques can be of great benefit to SMEs for the daily monitoring of security events, but these techniques should not be blindly applied so as to avoid increased in IDS complexity, and system performance decrease [6].

This paper identifies the main gaps and challenges that exist for attack detection systems and infrastructure related to SME systems. Also, promising research directions that should be pursued to address these gaps are proposed.A framework that combines anomaly and signature detection is proposed to meet the research objectives for SMEs. Different anomaly classifiers are applied, and the results are compared with findings from literature. Most existing IDSs are signature-based, which means they are capable of detecting only known attacks. The solution proposed in this paper combines signature-based IDS with anomaly detection to achieve a self-learning IDS. One

of the challenges that IDSs face is the high false alarm rate, which is an even bigger problem when it comes to SMEs due to the high workload for employees

The remaining of the paper is structured as follows: Section II displays related work on IDSs in SMEs. Section III presents the materials with a classification of methods depending on the techniques used in these systems as well as provides a hybrid IDS framework. In section IV, practical aspects of the experiments and analyses of data-driven approaches of different ML techniques are provided. Section VI comprises the conclusions and an outlook on possible future work.

II. Background

A. Related work

As this study is aimed at discovering cyber security defense systems (CSDS) for SMEs that are efficient, effective and practical to implement, various literature are explored. A CSDS for SMEs ought to be less expensive, since SMEs are known to lack the financial muscle to invest into other areas either than their core business area. A CSDS for SMEs should also be easy to manage by these organizations, since they often do not have the power to engage high-level IT security experts. Aside, threat neutralization, the cyber-security defense system for SMEs should also be able to detect and prevent intrusions with an acceptable false positive rate. An efficient and effective CSDS for SMEs should basically support the core business objectives and be capable of improving the SMEs risk status [7]. It should be able to provide a secure business environment for the SMEs to operate and deliver on their core objectives in alignment with their strategic direction and vision.

In this vein, Kent et al. [7] explored SMEs' perceptions of cyber-security and the factors that influence SME cyber-security implementation. A case study with an interpretive approach of inquiry was adopted to achieve the study objective. This method was quite useful for the study to focus on the complexity of human sense-making. A qualitative approach and interviews served as data collection means. Three case studies were interviewed, observed and the study arrived at the conclusion that budget, organizational complexity, attitude towards security and expertise for implementing security mechanisms were the main factors contributing to cyber security implementation problems. Respondents also perceived that web server logging could be performed and used for anomaly detection in SMEs. Open source software was also suggested for intrusion detection for SMEs. However, anomaly detection was perceived as more expensive to acquire, implement and maintain. Moreover, the respondents were doubtful about the usefulness of the ID technique. All respondents thought that it was not viable to perform intrusion detection based on web server logs alone, as web server logs are huge and do not contain enough useful information to perform proper intrusion detection. Kent et al. focused precisely on CSDS for SMEs and provided explicit knowledge on the challenges of SMEs that require research directions to develop CSDS to meet the needs of SMEs. However, the study did not result in experimentation to assess the proposed ideas.

Chamiekara et al. [8] analyzed and proposed an efficient and low-cost scenario-base security operations center for SMEs, having envisaged the high cost of existing security operations centers (AutoSOCs) for most SMEs. AutoSOC is a complete automated security operations center, except for the forensic investigation system aspect that requires the generation of a ticket and user's approval. The study [8] focused on cyber-security defense systems for SMEs and provided a framework for comprehensive CSDS development for SMEs, but the idea was not practically implemented or evaluated.

Recently, many ML-based methods both supervised and unsupervised have been used to address classification problems in IDSs. In [9], C4.5 decision tree algorithm and Support Vector Machine (SVM) methods were used for detecting network intrusion and the results indicated that the accuracy of C4.5 was better than SVM. A hybrid model was presented in [10], where the k-means unsupervised method was combined with the supervised SVM method. The contribution of the proposed methods is the increased detection accuracy, but less interest was directed to the computational time required for the learning and classification process. In [11], the authors studied the anomaly detection performance of a new hybrid model combining a genetic algorithm with latent Dirichlet allocation. The solution enhanced the detection accuracy to 98.1%, but the precision was low (87%), the computational time was high (1.42 hours) and the false positive rate was high as well (14%).

Kolias et al. [12] applied different conventional supervised ML techniques to the AWID dataset. The feature selection process was done manually based on the observations and theoretical analysis of the attack patterns. Eight supervised ML techniques were applied to a dataset with 20 features and the overall accuracy obtained varied between 89.43% and the highest of 96.2% with C4.5. More recently, the literature indicates some techniques proposed for detecting different attack classes using deep learning models [16].

B. Challenges for SMEs

In this section, findings of the literature reviewed such as the information security challenges for SMEs are presented.

1) **Lack of public datasets for IDSs in SMEs:** There are some datasets used by researchers for experiments, but the data are outdated and mostly comprises synthetic data. Activity found in laboratory networks has fundamental differences from real activity networks. The reason for the lack of publicly available data is the fact that data is highly sensitive as it contains confidential information about network users, making researchers face legal barriers in accessing the data.
2) **Semantic gap between intrusion detection results and their interpretation:** IDSs must transform their output into understandable reports for network operators to interpret. The majority of the systems are capable of identifying only deviations from the normal profile.
3) **Lack of self-adaptation:** One of the main characteristics of an IDSs is the self-adaptiveness, which means the nature of detection changes when it is necessary. If there

are any changes in the network configuration hardware, the system should adapt to these changes and work properly. The major drawback of existing IDSs is their inability to detect new classes rather than predefined classes of attacks.

III. MATERIALS AND METHOD

A. ML for Anomaly Detection

The goal of this study was to determine suitable ML methods that can be used for IDS for SMEs through experimental analysis. A literature review was conducted (Section II) to determine the challenges often faced by SMEs in relation to their security related countermeasures. The review also sorted for efficient ML algorithms that can be employed in SMEs for efficient IDS. Literature sources including IEEE-Xplore, Science Direct, Google Scholar and Digital ACM were searched for conference and journal articles of security solutions and IDEs implemented for SMEs. Various challenges associated with SMEs security solutions were also explored. A conceptual model of IDE for SMEs was then developed and evaluated with experimentation. The dataset used to evaluate the proposed framework is AWID Dataset [12], which is one of the largest benchmark Wi-Fi network datasets collected from real network introduced in 2015. Even though ML methods are used in order to classify intrusions behaviour, it is difficult to decide on a single method to apply on a given data set. In the experiments, we used peer reviewed ML techniques found in literature that are used for intrusion detection to enable a comparison of our proposed solution. The following classifiers were employed: C4.5 - a decision tree that uses the pruning process to mitigate over-fitting; Bayes Network - a probabilistic graphical model representing a set of variables and their conditional dependencies; Random Forest - a bagged decision tree model that split on a subset of features on each split; SVM - a classifier defined by a separating hyperplane; ANN - a feedforward artificial neural network that utilizes backpropagation for training and can classify data that is not linearly separable. These traditional techniques depend on feature selection and extraction, and an attempt was made to adapt these methods to detect attacks in real time with high accuracy and low false positive rate. Unfortunately, IDS face the problem of highly imbalanced data distribution and most machine learning algorithms have poor detection on minority classes of malicious attacks. A dataset is considered imbalanced if the class distribution is not uniform among the normal and malicious traffic classes. Standard ML techniques are more sensitive to detect the majority classes because they did not have enough data to learn about the minority classes during the training phase. There are two techniques of random resampling for dealing with imbalanced datasets: over-sampling that add synthetic data point belonging to the minority class and down-sampling that removes majority class data points. To tackle the imbalance problem, in this paper we used the Synthetic Minority Over-sampling Technique (SMOTE), an intelligent oversampling method that adds samples belonging to the minority classes to achieve equal distribution among classes. SMOTE works by joining the points of the minority classes with line segments and then places artificial points on these lines applying K-Nearest Neighbor (KNN) algorithm [13]. The reason of using SMOTE to handle the imbalance challenge is its compatibility with machine learning methods. All the results obtained with different techniques are compared with recent approaches as mentioned in literature and conclusions are drawn.

B. Feature selection

Feature selection is extremely important during the data preprocessing phase because the entire classification process depends on the inputted features. It is defined as the process of selecting a subset of features that contribute most towards the classification task. This process in machine learning brings many advantages such as: reduction in the data dimensionality, reduction in computation time, improvement of classifier predictive accuracy and enhancement of model understanding. Data features can be divided into relevant (informative) and irrelevant (uninformative). Applying ML-based algorithms to datasets with irrelevant features reduces the overall accuracy of classifiers. A lower number of features in a dataset translates in a shorter training times, an enhancement of generalization by reducing over-fitting and an improvement in execution performance. The key to building an intelligent IDS is to select the most informative features, which increases the systems predictive accuracy and reduces its complexity. There are three methods of feature selection: filter, wrapper and embedded. In the filter method, the features are selected based on statistical analysis of the feature set and the selection of features is done independently of any machine learning algorithm. During data preprocessing, when filter methods are applied, a score is calculated per each feature based on univariate statistics. The features are ranked by the score and either selected to be kept or not. The main advantages of this method are: independence of the classifier algorithm, lower computation cost and model generalization. Wrapper methods are based on greedy search algorithms and train a new model for each subset and fort this reason are computationally expensive. Embedded methods perform variable selection during the learning procedure and are depended on the applied classifier. In our experiments common feature selection methods implemented in Weka were applied because they are peer-reviewed and implemented in WEKA [14]. In order to show the effectiveness of the proposed framework, the following three filter feature selection methods are applied during data preprocessing and their results are compared:

- *CfsSubsetEval*: Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
- *Infogain*: Calculates the information gain of the attributes isolated by each other to determine which of them are most useful for discriminating between classes during the learning phase.

- *Correlation-based Feature Selection* : used to select a subset of features that increases the feature to class correlation [15] and reduces the feature to feature correlation.

C. Experimental environment and Evaluation metrics

The proposed solution was implemented using Waikato Environment for Knowledge Analysis (WEKA 3.8) [14] and was executed on a PC with Intel Core i7 processor, 2.1GHz speed and 8 GB RAM. False Negative (FN) and False Positive (FP) are two important metrics to measure IDSs reliability. For a given network event, the performance of a classifier is summarized in the confusion matrix, illustrated in Table I.

TABLE I
CONFUSION MATRIX

Actual Class	Table Column Head		
		Attack	**Normal**
	Attack	TP	FN
	Normal	FP	TN

The ML algorithms were evaluated by comparing the following evaluation metrics:

- **Accuracy or recognition rate**: Percentage of test set records that are correctly classified. Formally: $Accuracy = \frac{TP+TN}{TP+TN+FN+FP}$
- **Recall**: measures the accuracy of positive instances detected. Formally: $Recall = \frac{TP}{TP+FP}$
- **False Positive Rate (FPR)**: the ratio of normal events that are classified as attacks. Formally: $FPR = \frac{FP}{FP+TN}$
- **Learning time**: Total time taken to build a classifier model using the training dataset.

D. Proposed IDS Framework

The proposed model illustrated in Figure 1 it is divided into two main phases: data preprocessing and intrusion detection. The first critical process of an IDS includes data collection and preprocessing. The data sources and the locations from where the data were collected must be clearly defined. Starting from the collected network package, the flow data is used to define and label different flow records representing specific network activities. After the data is retrieved from network sources at regular time intervals, it is stored in a raw attack database. In the second stage, the data is decoded as a vector of numbers. To enhance the reliability and quality of the collected data, data cleaning activities are performed to reduce the noise in the data, to fill in missing values, to identify and remove outliers and to resolve data inconsistency. Moreover, the data is normalized in order to avoid the bias of features with greater values. Following normalization, the input data falls into the same range of the interval [0-1]. After normalization, three different filter feature selection methods are applied and their performance are compared in order to determine a subset of features.

The proposed framework is a hybrid IDS combining signature-based and anomaly-based detection. First, specific patterns such as byte sequences or known malicious intrusion

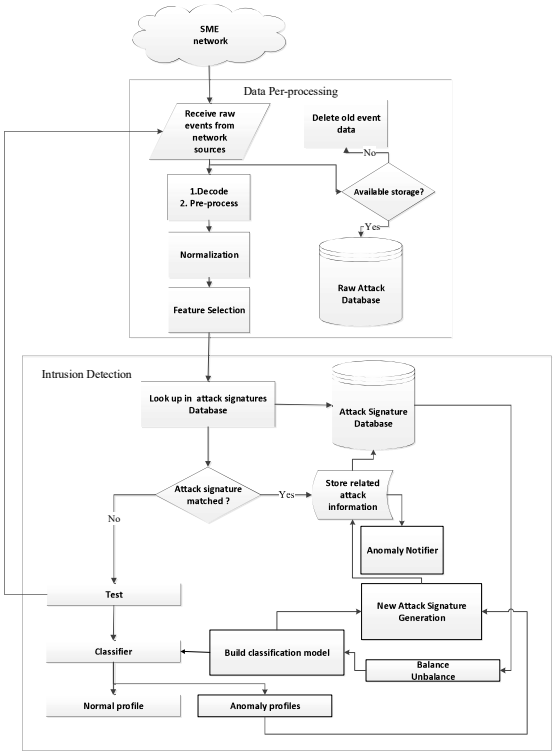

Fig. 1. Proposed IDS.

sequences are looked up in the attack signature database. If the signature matches, it is classified as an anomaly profile and the system produces an accurate notification of malicious activity. On the other hand, if the attack signature does not match, the input data is tested based on a classifier model in order to decide whether it belongs to specific types of attacks or benign traffic. First relevant features are selected by applying different filter feature selection methods. After the optimal feature subset is obtained, the dataset is balanced using random sampling balancing methods such as SMOTE. Then, the balanced and unbalanced test datasets are used to train the model. Based on the trained model, each new flow in the system is assigned to a certain class applying different ML methods. The goal of this framework is not only to increase the overall classification accuracy, but to keep a higher accuracy for the attack profile classes. The results of the anomaly profiles are fed back to the signature database to update it with new attacks in order to tune the ML algorithms.

IV. RESULTS AND DISCUSSIONS

V. DATA SOURCE AND DATA PREPROCESSING

The main reason for using AWID dataset in the present work is the real life like characteristics and the large number of records it contains. Based on the number of target classes, there are two datasets: “CLS” and “ATK.” “CLS” groups the different attacks into 4 main classes: normal, impersonation, flooding and injection. On the other hand, the “ATK” dataset has 16 target classes belonging to the 4 main classes. Also, the

datasets are divided into two different types (full and reduced) based on the number of data instances included. This dataset contains 155 numerical attributes and one class attribute. For simplicity in this study, it is used the reduced "CLS" dataset and its details are presented in Table II.

TABLE II
CLASSIFIERS EVALUATION OF AWID-CLS_R WITH 34 FEATURES

Classes	Dataset	Number of instances	% of attack class
Flooding Impersonation Injection Normal	AWID-CLS-R-Trn	1,795,575	5.04%
	AWID-CLS-R-Tst	530,643	4.89%

During the data preprocessing phase, the data were cleaned and transformed in order to be used for the classification process. This phase includes, data cleaning, data transformation and feature selection. First the attributes with string and hexadecimal values was encoded as numbers to be used as input features in ML methods. Also, the dataset contains missing values denoted by "?" which are replaced with zeroes. The attributes with constant values are dropped out, as they do not contribute to the class distinction. After removing irrelevant features, the dataset has 124 attributes from 154 it had originally. The features were extracted according to the three filter methods described in section 3, and the selected features are normalized. According to the applied feature selection methods, it was observed that the number of selected features ranged from 6 to 34, illustrated in Tables III . By comparing these numbers against the total number of features (n = 124), a decrease in computational performance was expected. As a result, all the selected ML methods were performed with four different datasets containing respectively 124, 6, 16 and 34 attributes and one class label (3 attack classes and 1 normal class) for each.

TABLE III
FEATURE SELECTION METHODS

Method	Search Method	No. of features
$CfsSubsetEval$	GreadyStepwise	6
$Infogain$	Ranker	16
$CorrelationAttributeEval$	Ranker	34

A. Experimental results

This section presents the principal findings of the experiments. The results obtained from analyzing the performance of traditional classifiers with 124, 6, 16 and 36 features are summarized in Tables IV, V, VI and VII respectively.

Table IV shows the resulted obtained from the experiment using the dataset with 124 attributes. We have measured and reported the time complexity of the training phase in the datasets with highest and lowest number of features, respectively the dataset with 124 features and the datset with 6 features. This is done to conclude related to the computational challenges that high dimensional data poses. What stands out in these tables is that the highest accuracy in the dataset with 124 features was 95.34% obtained by applying the Random Forest algorithm, whereas, the lowest was 90.63% applying the Bayesian Network classifier. The obtained accuracy is not really satisfactory, the classifiers failed to classify some of attacks in AWID. Also, the computation time is high for all the applied ML methods. For example, SVM method takes 3 hours and 45 minutes to train the model. So, in order to increase the detection accuracy and to shorten the computational time of learning phase, feature selection methods are applied. Three different filter based feature selection methods were applied, and all ML methods were used to test the proposed framework.

TABLE IV
CLASSIFIERS EVALUATION OF AWID-CLS_R WITH 124 FEATURES

Learning method	Nr. of attributes	Accuracy	FP	Time to train the model
C.5	**124**	95.25	0.128	11:59
Bayesian Network		90.63	0.48	05:20
Random Forest		95.34	0.149	38:55
SVM		90.83	0.055	03:45:32
ANN		95.02	0.088	01:17:04

Table V shows the performance for the dataset with 6 attributes. The accuracy of all algorithms are similar, and their overall accuracy is degraded compared to the accuracy of the dataset with 124 attributes. The main reason for the decreased of accuracy is related to the fact that some features that are significant to predict classes of small areas of instance space are eliminated during feature selection. C4.5 performs better with an accuracy of 91.96%. As the complexity of the dataset is reduced, the worst computation time is obtained by applying SVM, one hour and nine minutes to train the classifier. This is approximately 4 times faster than the full feature set. C4.5 with 6 features takes 41 seconds for the training of the data, which is 18 times faster than the full feature set.

TABLE V
CLASSIFIERS EVALUATION OF AWID-CLS_R WITH 6 FEATURES

Learning method	Nr. of attributes	Accuracy	FP	Time to train the model
C.5	**6**	91.69	0.63	00:41
Bayesian Network		91.90	0.25	00:01
Random Forest		91.90	0.24	00:10
SVM		91.78	0.53	01:09:58
ANN		90.99	0.22	05:02

Table VI, represents the results obtained with infogain method where 16 feature were selected and best accuracy was obtained by using C4.5 (95.39%). Based on this feature

selection method, the number of selected features is low (13% of features selected over 124 features), and 87 % of the original data are not taken into account, the accuracy obtained is higher.

TABLE VI
CLASSIFIERS EVALUATION OF AWID-CLS_R WITH 16 FEATURES

Learning method	Nr. of attributes	Accuracy	FP
C.5	16	95.39	0.128
Bayesian Network		92.26	0.576
Random Forest		95.11	0.18
SVM		94.12	0.22
ANN		91.3542	0.169

In Table VII are presented the resulted obtained by using the dataset with 34 features selected using CorrelationAttributeEval feature selection methods. Using this dataset, we obtained the highest accuracy by applying C4.5 (98.4508%). SVM produced the worst results with 88.21% accuracy, which can be explained by the purely linear nature of this method.

Another significant problem in IDSs is the management of alerts. If the system has a high FP alarms rate, it makes it very challenging for SMEs when human operators must understand the problem and take action. Thus, it is necessary to develop an IDS with a high detection rate and that can be tuned accurately to ensure a low FP rate and to enable adjustment to environment changes with less human intervention. In terms of the FP metric, C4.5 classifier that was applied to the dataset with 34 attributes had the best FP rate of 0.029% compared to other ML methods.

What stands out from the comparison between all feature selection methods and the original dataset, is the significant improvement in accuracy for the dataset with 34 attributes (the highest accuracy of 98.4508% and the lowest FP rate 0.029%) obtained using C4.5. The result confirm that feature selection is a very important factor to be considered in IDSs, since IDSs deal with real-world high-dimensional feature spaces. IDSs collect high volume of audit data for small or medium networks. The accuracy and performance of a classifier is reduced if the numbers of attributes in a data set are reduced.To accurately develop IDS detection methods, the data dimensionality should be reduced. As the intrusion classification is much related to the data set characteristics, different ML methods should be considered for finding the method that better fits the needs. Usually, real-world data sets in the intrusion domain are imbalanced and the classifiers attempt to maximize only the accuracy of major class. Considering this problem in the AWID-CLS dataset, we adopted the SMOTE technique (k=5) to balance the dataset. The training dataset has about 1.8 million records and adding synthetic cases for balancing would bring computational challenges. Consequently, we applied SMOTE methods to the test dataset with 16 attributes to train C4.5. SMOTE was used to balance the 3 minor classes, so the training dataset had 50% normal flow and 50% attack flow belonging to 3 attack classes, as illustrated in Figure 2. Also, 10% of records were randomly selected from the

TABLE VII
CLASSIFIERS EVALUATION OF AWID-CLS_R WITH 34 FEATURES

Learning method	Nr. of attributes	Accuracy	FP
C.5	34	98.4508	0.029
Bayesian Network		92.7434	0.530
Random Forest		95.3596	0.154
SVM		88.2149	0.055
ANN		98.32	0.031

training dataset and were utilized to test the model. The overall experimental procedure is shown in Figure 3.

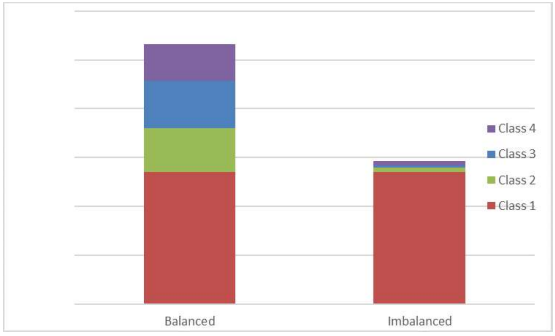

Fig. 2. Balanced dataset using SMOTE vs imbalanced dataset.

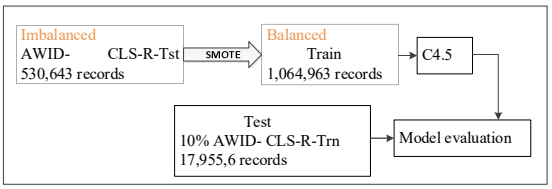

Fig. 3. Experimental procedure

After balancing the data, the accuracy obtained by C4.5 was 99.9872%, which is the highest among all methods applied. Moreover, according to the confusion matrix in Table VIII, a true positive rate (TPR) of over 99% was obtained for all attack categories and normal flows. Our experiments conclude that a 50:50 balance ratio between the normal and attack classes in the training dataset improved the classifier performance. For SMEs where the data set is not to large, the process of balancing the data is beneficial because the size of samples in the attack classes is not big enough to contain significant information for the learning phase.

The following paragraph compares our experiments with previous related works where the same dataset is tested. The comparison results are summarized in Table IX. Thanthrige et al. [16] achieved the highest accuracy (94.33%) using Random Forest with 111 attributes and 92.44% using C4.5 with 10 attributes. Aminanto et al. [17] achieved 99.88% accuracy using the Deep-Feature Extraction and selection earning method with a dataset with 35 features. Our results indicate better performance with an overall detection accuracy of 99.9872%. The proposed framework is relevant for SMEs

because the methods with high accuracy found in literature are based on deep learning, which are used for large-scale classification problems.

TABLE VIII
CONFUSION MATRIX OF THE TESTING DATA SET

		Predicted class			
		Normal	Impersonation	Flooding	Injection
Threat agent	Normal	163308	2	9	0
	Impersonation	9	4843	0	0
	Flooding	3	0	4845	0
	Injection	0	0	0	6537

TABLE IX
COMPARISON OF THE PROPOSED METHOD WITH OTHER ML TECHNIQUES

Study	**Learning method**	**Number of attributes**	**Accuracy**
Thanthrige [16]	Random Forest	111	94.33
Thanthrige [16]	C4.5	10	92.44
Kolias [12]	C4.5	20	96.2
Kaleem [18]	Neural Network	6	99.3
Thing [19]	Deep Learning	154	98.67
Aminanto [17]	Deep-Feature Extraction	35	99.88
Proposed method	SMOTE—C4.5	34	**99.9872**

VI. CONCLUSIONS

The aim of this study was to explore for the main gaps and challenges that exist for attack detection systems and infrastructure related to SME systems. A framework that combines anomaly and signature detection was proposed to meet the research objectives for SMEs. Different anomaly classifiers were applied, and the results were compared with findings from literature. Many existing IDS are rule-based systems and their performance is mostly based on predefined rules in the system. This approach is not scalable for analyzing a large volume of traffic data so to overcome this limitation, ML techniques were explored. A conceptual framework was therefore developed and assessed for its suitability for SMEs. The ML algorithms which were evaluated for their suitability were C4.5, Random Forest, Artificial Neural Network, Bayesian Neural Networks and SVM.

The experimental results showed that by selecting the most significant attributes, and by applying an intelligent over-sampling technique such as SMOTE to balance the dataset, the IDS overall accuracy is increases. The overall accuracy of the proposed solution is 99.9872% and the accuracy of each malign classes is higher than 99%. A comparison with recent approaches cited in literature showed an improvement in accuracy and in the FP rate for all attack profiles.

In the future, we would like to test our proposed framework with other real datasets covering a broader range of attacks, including mobile and Internet of Things security platforms. The experiments should also be performed using ML methods that are capable to handle real time intrusion detection.

REFERENCES

[1] EU. (2005) Commission adopts a new definition of micro, small and medium sized enterprises in europe. Accessed 30th May 2019. [Online]. Available: http://europa.eu/rapid/press-release_IP-03-652_en.pdf

[2] Organisation for Economic Co-operation and Development . (2019) Oecd sme and entrepreneurship outlook 2019. Accessed 5th June 2019. [Online]. Available: https://www.oecd.org/industry/oecd-sme-and-entrepreneurship-outlook-2019-34907e9c-en.htm

[3] M. Zec and M. Kajtazi, "Examining how it professionals in smes take decisions about implementing cyber security strategy," in *ECIME2015-9th European Conference on IS Management and Evaluation: ECIME 2015*. Academic Conferences and publishing limited, 2015, p. 231.

[4] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[5] Y. B. Choi and G. D. Allison, "Intrusion prevention and detection in small to medium-sized enterprises," 2017.

[6] R. Zuech and T. M. Khoshgoftaar, "A survey on feature selection for intrusion detection," in *Proceedings of the 21st ISSAT International Conference on Reliability and Quality in Design*, 2015, pp. 150–155.

[7] C. Kent, M. Tanner, and S. Kabanda, "How south african smes address cyber security: The case of web server logs and intrusion detection," in *2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech)*. IEEE, 2016, pp. 100–105.

[8] G. Chamiekara, M. Cooray, L. Wickramasinghe, Y. Koshila, K. Abeywardhana, and A. Senarathna, "Autosoc: A low budget flexible security operations platform for enterprises and organizations," in *2017 National Information Technology Conference (NITC)*. IEEE, 2017, pp. 100–105.

[9] M. Ektefa, S. Memar, F. Sidi, and L. S. Affendey, "Intrusion detection using data mining techniques," in *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*. IEEE, 2010, pp. 200–203.

[10] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified k-means for intrusion detection system," *Expert Systems with Applications*, vol. 67, pp. 296–303, 2017.

[11] B. Kasliwal, S. Bhatia, S. Saini, I. S. Thaseen, and C. A. Kumar, "A hybrid anomaly detection model using g-lda," in *2014 IEEE International Advance Computing Conference (IACC)*. IEEE, 2014, pp. 288–293.

[12] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, "Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 184–208, 2015.

[13] A. Sen, M. M. Islam, and K. Murase, "An algorithmic framework based on the binarization approach for supervised and semi-supervised multiclass problems," in *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2014, pp. 175–182.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[15] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[16] U. S. K. P. M. Thanthrige, J. Samarabandu, and X. Wang, "Machine learning techniques for intrusion detection on public dataset," in *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2016, pp. 1–4.

[17] M. E. Aminanto, R. Choi, H. C. Tanuwidjaja, P. D. Yoo, and K. Kim, "Deep abstraction and weighted feature selection for wi-fi impersonation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 621–636, 2017.

[18] D. Kaleem and K. Ferens, "A cognitive multi-agent model to detect malicious threats," in *Proceedings of the 2017 International Conference on Applied Cognitive Computing (ACC17)*, 2017.

[19] V. L. Thing, "Ieee 802.11 network anomaly detection and attack classification: A deep learning approach," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2017, pp. 1–6.