A Combined Decision for Secure Cloud Computing Based on Machine Learning and Past Information

Zina Chkirbene, Aiman Erbad and Ridha Hamila
College of Engineering, Qatar University

***Abstract*—Cloud computing has been presented as one of the most efficient techniques for hosting and delivering services over the internet. However, even with its wide areas of application, cloud security is still a major concern of cloud computing. In order to protect the communication in such environment, many secure systems have been proposed and most of them are based on attack signatures. These systems are often not very efficient for detecting all the types of attacks. Recently, machine learning technique has been proposed. This means that if the training set does not include enough examples in a particular class, the decision may not be accurate. In this paper, we propose a new firewall scheme named Enhanced Intrusion Detection and Classification (EIDC) system for secure cloud computing environment. EIDC detects and classifies the received traffic packets using a new combination technique called most frequent decision where the nodes' [1] past decisions are combined with the current decision of the machine learning algorithm to estimate the final attack category classification. This strategy increases the learning performance and the system accuracy. To generate our results, a public available dataset UNSW-NB-15 is used. Our results show that EICD improves the anomalies detection by 24% compared to complex tree.**

Index— Cloud security, firewalls, attack signatures, machine learning technique, past performance.

I. Introduction

Cloud security is considered as an important issue in cloud computing (CC) environment which attracted a lot of research in past few years [1], [2], [3]. For instance, an attacker can explore the vulnerabilities in CC to introduce multiple security threats such as denial of services (DoS), Domain Name Server spoofing (DNS) and Address Resolution Protocol (ARP) [4].

Many security systems have been proposed [5], [6] to protect the CC environment and most of them are based on attack signatures. These systems may seem effective and efficient; however, they suffer from multiple issues [7]. For instance in case of failure, the attacker can exploit the time needed for system maintenance and access to the network. Moreover, researchers have shown that if the signatures are not sufficiently described then the attacker can succeed to introduce many malicious threats [8]. Furthermore, these systems need the human intervention to create, test, and deploy the signatures and it may take hours or days to generate a new signature for an attack. To overcome these problems, systems that do not rely on human intervention have been proposed based on machine learning and well-trained model to improve significantly the anomaly detection performance with a reasonable cost and complexity [9]. However, maximizing the learning efficiency is still one of the major goals of secure systems that can be only achieved through a big amount of data that might be challenging to collect, required to generate a robust model.

In this paper, we propose a new firewall called **Enhanced Intrusion Detection and Classification** (EIDC) system for secure CC environment. EIDC combines the performance of supervised learning and past information about the network users. Basically, by taking the history of each node into account, the proposed scheme improves the estimation of the node reliability when making the final packet classification [6] [10].

The main contributions of this paper can be summarized as follows:

1) A novel firewall for cloud computing anomaly detection and classification model namely Enhanced Intrusion Detection and Classification (EIDC) is proposed.
2) The proposed supervised learning model helps the system evaluate the trustworthiness of every node and significantly increase the classification accuracy compared to traditional models.
3) Extensive simulation results have been conducted to confirm the efficiency of the proposed model in detecting and identifying most of the attack categories.

The rest of the paper is organized as follows: In Section II, a brief overview of related work carried out is presented. We describe the proposed model of EICD in section III. Section IV presents the simulation results and section V concludes the paper.

II. Overview

Several techniques have been proposed to protect the CC environment from multiple security threats [11], [12]. In [13], the authors proposed Support Vector Machine (SVM) based approach to secure cloud computing while reducing the over-fitting problem. However, the kernel models can be quite sensitive to over-fitting the model selection criterion. In [14], Artificial Neural Network (ANN) based approach has been proposed to build the security model. Although, ANN has the ability of storing information on the entire network with high parallel processing capability, it depends on the hardware performance since it requires processors with parallel processing power. In [15], the authors proposed a decision tree based approach for network anomalies detection. Moreover, a hybrid machine learning model is proposed in [16]. This technique is able to mix the advantages of two learning schemes; however it

[1] In this document we will use the words "node" and "user" interchangeably.

978-1-5386-7646-2/19/$31.00 ©2019 IEEE

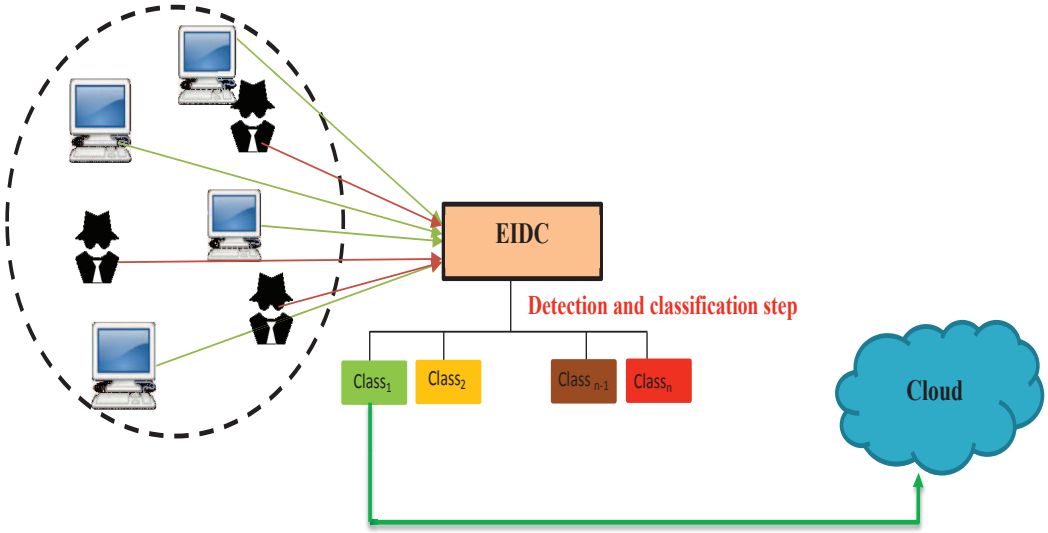

Fig. 1. The proposed model.

increases the system complexity. In [17], the authors proposed Bayesian network based model to detect the network threats. But, Bayesian network is extremely computationally expensive due to sigma functions and cross-corpora calculations.

III. Enhanced intrusion detection and classification

A. Network Traffic Model

We assume that the proposed network is composed of N nodes for which the communication categories (normal users or attackers) are proportionally distributed to the categories distribution in the considered dataset [2]. We assume also that each node i can be identified and differentiated by a set of features denoted by x^i that reflect the characteristics of transmitted data packets (for example IP destination, transmission protocol). Let y^i be the output class of each feature set. We denote by TS the input data (a training dataset) modeled as follows:

$$\forall i \in \{1..N\} \qquad (1)$$
$$TS = \left\{ \left(x_1^i, y_1^i\right), \left(x_2^i, y_2^i\right), ..., \left(x_n^i, y_n^i\right) \right\}.$$

We assume that each node keeps the same status for an enough period of time that allows the firewall to detect it. In particular, this strategy is applied to protect the network from any unknown node that keeps trying to inject malicious packets in the network. We assume that the system receives a new packet from each node in each time unit TU.

[2]Note the distribution of the nodes categories does not affect the performance of proposed model and is only used to exploit the available dataset to model a multinode continuous communication CC network

B. Overall Framework

The overall structure of the proposed framework is composed of three major steps, as depicted in Figure 2. In the first step, a **learning phase** is performed to train the machine learning algorithm and create the learning model which will be used to make the first classification for any received packet. The second step includes **decision history storing** where EICD scheme stores in a separate database all the previous decisions made for each node in the network. The final step where the final classification decision is made namely, **combined decision phase**. For each received packet from any node in the network, EICD combines the decision of the machine learning model with the saved past information (nodes attack types classification decision histories) and the most frequent decision will be taken (.i.e. for a received packet in time t, the decision is made based on all the previous decisions from instant 0 to t).

C. Leaning Phase

The **Leaning phase** step is composed of the machine training and model creation. In the machine training, an input data that includes packets for normal and abnormal behaviors is needed. This process focuses on modeling the input features for each node to an output class. The learning algorithm tries to find a function $f(x)$ of the input features that best predicts the output class y for any node i so that:

$$y^i = f(x^i). \qquad (2)$$

Algorithm 1 creates the model $Model$ using a selected learning algorithm LG. The predict function $predict$ is used to predict the classification of new received packets. At the end of this step, the created model should be able to classify any received packet.

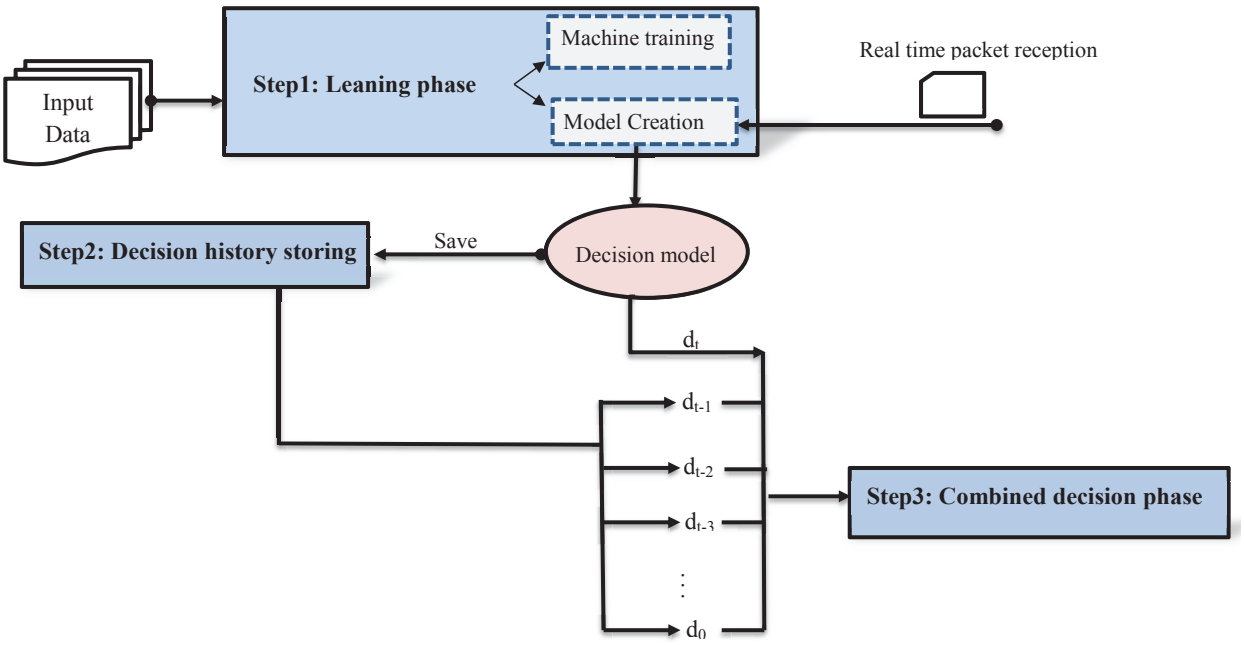

Fig. 2. The general Framework.

Algorithm 1 Creation of models

Input:
LG: The selected learning algorithm
TS: The input data
p_t^i: The received real time packet from node i
Output:
d_t^i: The classification d_t of node i using LG

TS
Create classifier:
$Model \leftarrow ML\_Model\_create(LG, TS)$.
Generate the predicted decision :
$d_t^i \leftarrow predict(Model, p_t^i)$
Return(d_t^i)

D. Decision History Storing

For any received packet at time t, the created model generates a new decision classification d_t^i and the system adds it to the database. The saved decisions will be updated so that each node i has a behaviors history D_i in the dataset that can be written as

$$D^i = \left\{d_0^i, d_1^i, d_{t-1}^i, d_t^i\right\}. \quad (3)$$

E. Combined Decision Phase

Combined decision phase is the last step in EIDC. The decision combination is performed in accordance with the following steps:

1) Step 1: Determine node i using the set of features x_i.
2) Step 2: Access to the decision history database and extract D^i of the node i.
3) Step 3: Combine D^i with the current decision d_t^i .
4) Step 4: Compute the most frequent decision and select it as the best decision.

Figure 3 shows an example of the most frequent decision for a node i. After 5 time slots, EIDC takes into consideration all the decisions made for this node, and the most frequent (0) is taken even if the last decision of the machine learning is 1.

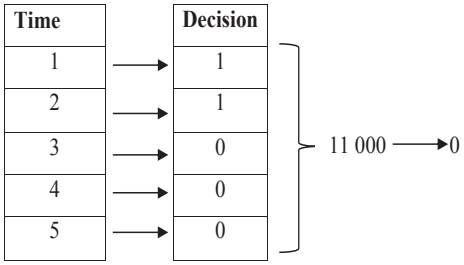

Fig. 3. An example of the most frequent decision.

IV. Simulation Results

A. Simulation Setup

In 2015, UNSW dataset has been released, it includes 10 different types of traffic packets and it is more suitable to be used in the contemporary anomaly detection models. It includes normal packets as well as 9 types of attacks, which are Analysis, Backdoor, DoS, Exploits, Fuzzers, Reconnaissance, Shellcode and Worms. Table I shows the notation of these

classes in the paper. UNSW-NB-15 is composed of two parts: a training set *UNSW-NB-15 training-set.csv* which has been used for model creation and a testing set, *UNSW-NB-15-testing-set.csv* used for the testing step and modeling the received real time packets. The number of records in the training and testing sets are 175,341 and 82,332, respectively.

TABLE I
CLASSES NOTATION

Number	Class
1	Normal
2	Analysis
3	Backdoor
4	DoS
5	Exploits
6	Fuzzers
7	Generic
8	Reconnaissance
9	Sellcode
10	Worms

Complex tree from decision tree is used. The total number of nodes is fixed to 50 nodes proportionally distributed to the number of classes in the dataset.

B. Definitions

In this section, the performance of the proposed model is evaluated and compared with traditional model in terms of different criteria, namely:

- **False Positive (FP)** and **False Negative (FN)** indicate false alarms and miss detection, respectively.
- **True Positive (TP)** presents the packets that are correctly identified (as normal) and **True Negative (TN)** presents the portion of correctly rejected packets.
- **The accuracy** is the ratio of correct predictions over the total number of the packet in the testing set.

$$Accuracy = \frac{TN + TP}{number\ of\ packets} \tag{4}$$

C. Simulation results

Figure 4 shows the accuracy of the EIDC model compared to complex tree as a function of time. We can see that by increasing the time, the accuracy of EICD increases too. However, the accuracy in complex tree is fixed to 69% which proves that the EICD is sensitive to the time. In fact, EICD takes into consideration the nodes past performance when making the final classification which increases the accuracy especially for long time where the system has enough information about nodes history. We can remark also that using the EICD increases the accuracy by 24% compared to complex tree. This proves that the EICD detects better the traffic anomaly than complex tree.

Figure 5 presents the detection performance after 1, 50, and 150 time unit (TU) for EICD. It can be seen that the proposed algorithm successfully distinguishes the malicious users from the normal nodes by giving them low reputation scores. We remark also that 150 TU are enough to detect the malicious users in the network.

Figure 6 presents the detection performance after 1, 30, 50, and 150 TU for EICD. Note that when the time index is equal to 1, the EICB is equivalent to the classic complex tree. Note also that the normal node average decision is equal to 0 and the average for malicious node is 1. It can be seen that complex tree was not able to classify the packets of the classes (2,3,4,10) since they doesn't have sufficient number of packet in the data set that allows the system to train itself for better detection which has an impact on the performance detection of EICD on these classes. We remark also that whatever is the time, the difference between complex tree and EICD can be seen starting from the first iteration and increases in time.

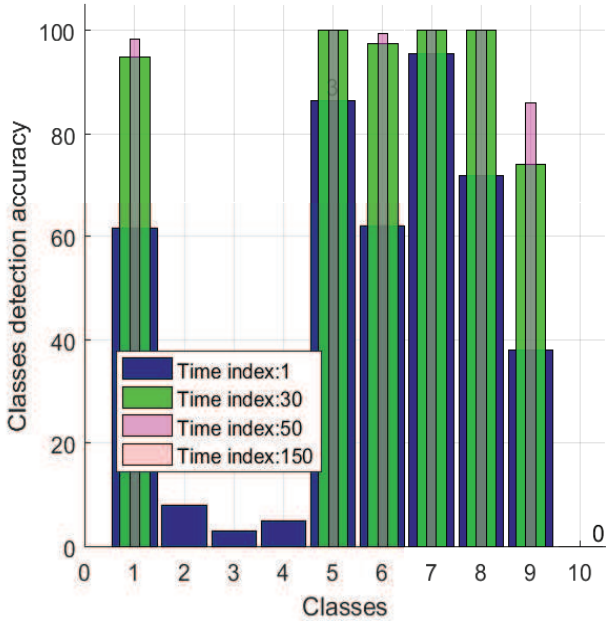

Fig. 6. The detection performance of the proposed model per classes in function of time.

Figure 6 presents the confusion matrices of EICD as a function of time. Confusion matrix is a specific table that shows the performance of supervised learning. Each column of the matrix presents the predicted classes and each row presents the true classes. The diagonal of this matrix shows the true detection performance for each class of the system. First, we can remark that complex tree has a weak detection performance in the detection of classes (2, 3, 4, and 10). In particular, class 10 has 0 packets detected out of 100 received packets. We can remark also that by increasing the time index the matrix becomes almost diagonal and all the classes converge. The detection rate reaches approximately 90% for each node. So, the bigger the number of packets and time iterations (time index) are, the better and faster the malicious users can be distinguished and classified.

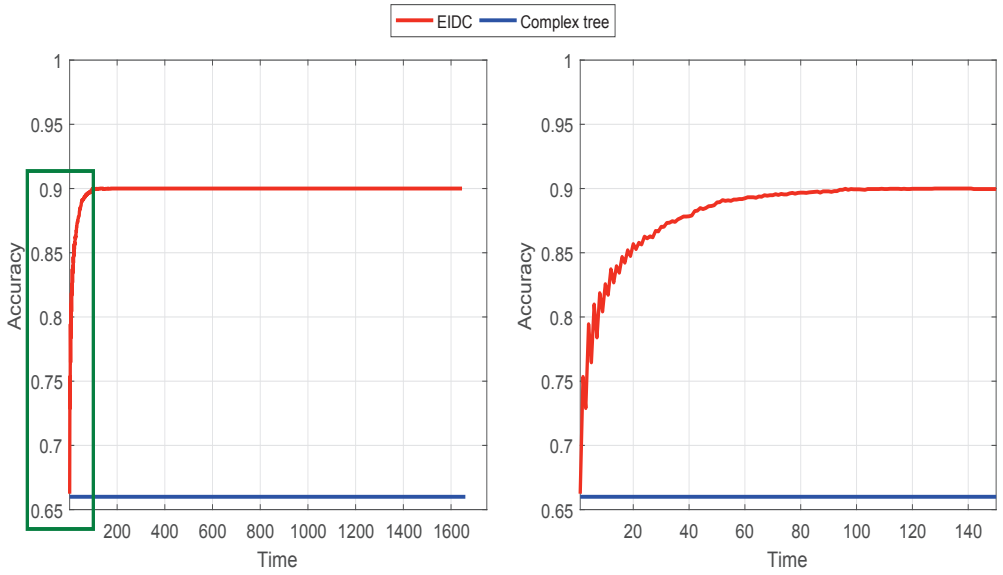

Fig. 4. The accuracy of the proposed model compared to complex tree as a function of time

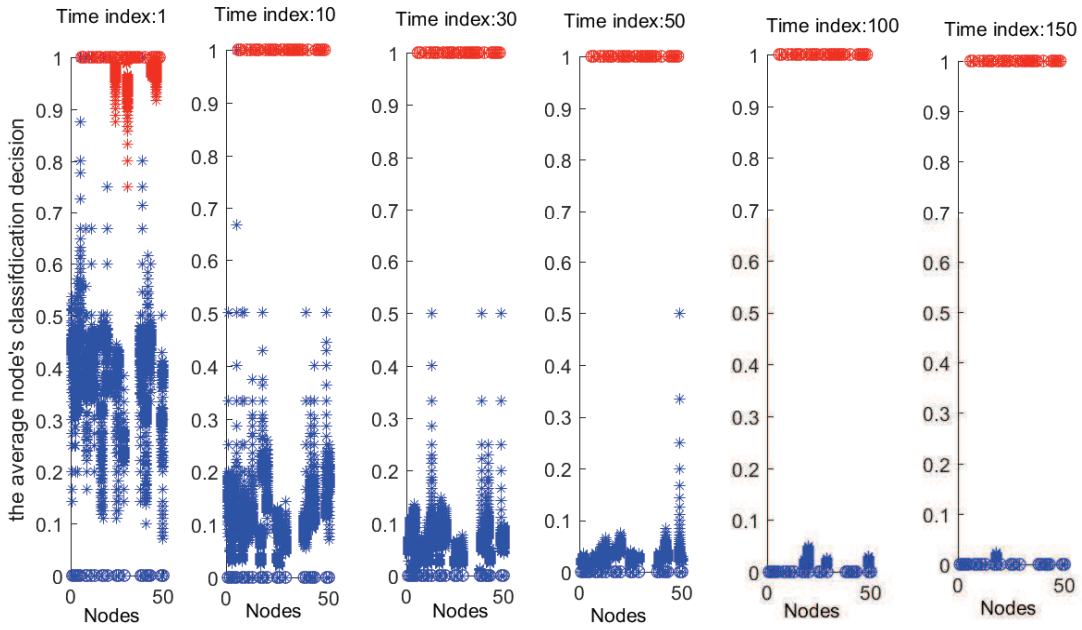

Fig. 5. Detection performance of the proposed model

V. CONCLUSION

In this paper, a new firewall called **Enhanced Intrusion Detection and Classification** (EIDC) system has been proposed to realize a secure cloud environment. EIDC improves the detection and classification of malicious users using a novel combination technique that exploits the decision histories as well as current decisions. Simulation results showed that this scheme is more efficient for attack detection compared to the classic learning algorithm and can increase the classification accuracy from 66 % to up to 90%. Future work will address more advanced techniques for profiling normal packets while improving the classification performance using methods such as data mining and data clustering.

ACKNOWLEDGMENT

This publication was made possible by the NPRP award [NPRP 8-634-1-131] from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the author[s].

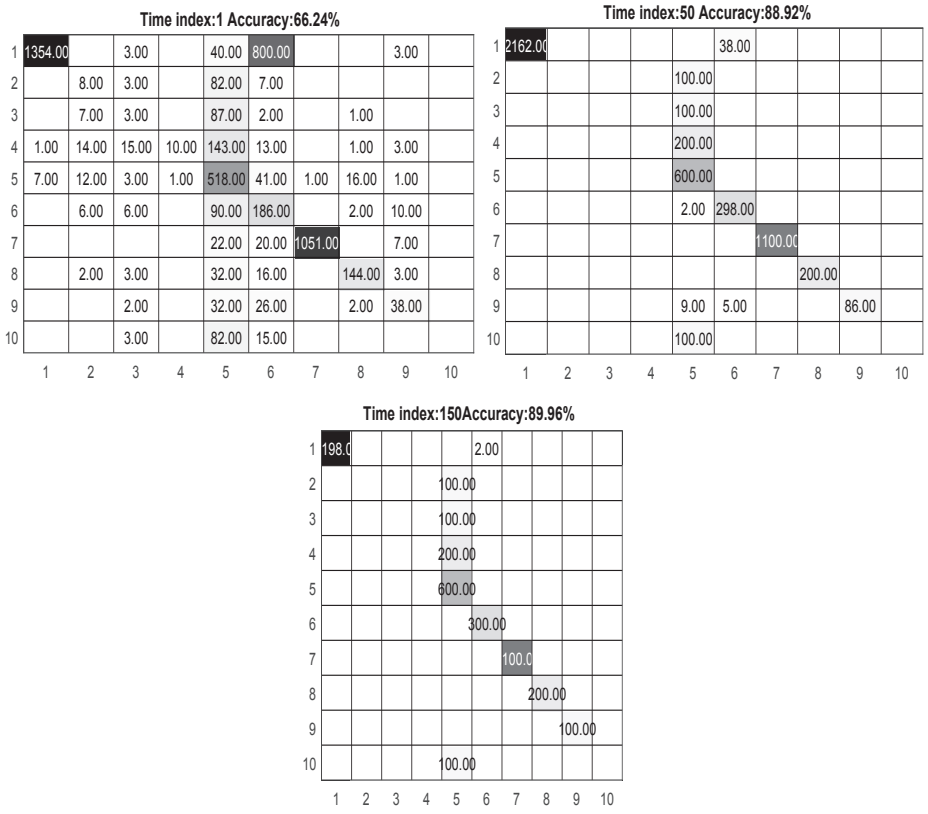

Fig. 7. Confusion matrices of EICD as a function of time

REFERENCES

[1] Z. Chkirbene, S. Foufou, M. Hamdi, and R. Hamila, "Scalnet: A novel network architecture for data centers," in *2015 IEEE Globecom Workshops (GC Wkshps)*, Dec 2015, pp. 1–6.

[2] Z. Chkirbene, S. Foufou, and R. Hamila, "Vaconet: Variable and connected architecture for data center networks," in *2016 IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–6.

[3] Zina Chkirbene, Ala Gouissem, Rachid Hadjidj, Sebti Foufou, and Ridha Hamila, "Efficient techniques for energy saving in data center networks," *Computer Communications*, vol. 129, pp. 111 – 124, 2018.

[4] Saurabh Singh, Young-Sik Jeong, and Jong Hyuk Park, "A survey on cloud computing security: Issues, threats, and solutions," *Journal of Network and Computer Applications*, vol. 75, pp. 200 – 222, 2016.

[5] Muhammad Abedin, Syeda Nessa, Latifur Khan, and Bhavani Thuraisingham, "Detection and resolution of anomalies in firewall policy rules," in *Data and Applications Security XX*, Ernesto Damiani and Peng Liu, Eds., Berlin, Heidelberg, 2006, pp. 15–29, Springer Berlin Heidelberg.

[6] Chkirbene Zina, Mazen Hasna, Ridha Hamila, and Noureddine Hamdi, "Location privacy preservation in secure crowdsourcing-based cooperative spectrum sensing," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, pp. 85, Mar 2016.

[7] Giovanni Vigna, William Robertson, and Davide Balzarotti, "Testing network-based intrusion detection signatures using mutant exploits," in *Proceedings of the 11th ACM Conference on Computer and Communications Security*, New York, NY, USA, 2004, CCS '04, pp. 21–30, ACM.

[8] Giovanni Vigna, William Robertson, and Davide Balzarotti, "Testing network-based intrusion detection signatures using mutant exploits," in *Proceedings of the 11th ACM Conference on Computer and Communications Security*, New York, NY, USA, 2004, CCS '04, pp. 21–30, ACM.

[9] S. Kiranyaz, I. Turker, M. Gabbouj, and R. Hamila, "Method and apparatus for performing feature classification on electrocardiogram data," vol. 858, 02 2008.

[10] L. Khalid and A. Anpalagan, "A weighted fusion scheme for cooperative spectrum sensing based on past decisions," in *2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept 2011, pp. 354–358.

[11] S. A. Almulla and Chan Yeob Yeun, "Cloud computing security management," in *2010 Second International Conference on Engineering System Management and Applications*, March 2010, pp. 1–7.

[12] B. R. Kandukuri, R. P. V., and A. Rakshit, "Cloud security issues," in *2009 IEEE International Conference on Services Computing*, Sept 2009, pp. 517–520.

[13] Laura Auria and Rouslan A. Moro, "Support vector machines (svm) as a technique for solvency analysis," vol. 1, 02 2008.

[14] S Ryszard Michalski, G Jaime Carbonell, and M Tom Mitchell, Eds., *Machine Learning an Artificial Intelligence Approach Volume II*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1986.

[15] Gary Stein, Bing Chen, Annie S. Wu, and Kien A. Hua, "Decision tree classifier for network intrusion detection with ga-based feature selection," in *Proceedings of the 43rd Annual Southeast Regional Conference - Volume 2*, New York, NY, USA, 2005, ACM-SE 43, pp. 136–141, ACM.

[16] Sandhya Peddabachigari, Ajith Abraham, and Johnson Thomas, "Intrusion detection systems using decision trees and support vector machines," vol. 11, 01 2004.

[17] Bo Jin, Yong Wang, Zhenyan Liu, and Jingfeng Xue, "A trust model based on cloud model and bayesian networks," *Procedia Environmental Sciences*, vol. 11, pp. 452 – 459, 2011, 2011 2nd International Conference on Challenges in Environmental Science and Computer Engineering (CESCE 2011).