Proposed Intrusion Detection Model using Decision Tree Classifier with Feature Engineering

Abstract

Due to the recent advancements in the Internet of things (IoT) and Cloud computing technologies and growing number of devices connected to the internet, the security and privacy issues are important to be resolved and protect the data and computer network. To provide security, a real time monitoring of the network data and resources is needed. Intrusion detection systems have been used to monitor, detect and alert an intrusion event in real time. Recently, the intrusion detection systems (IDS) incorporate several machine learning (ML) techniques. One of the techniques is decision tree, which can take reliable network measures and make good decisions by increasing the detection rate and accuracy. In this paper, we propose a reliable network intrusion detection approach using decision tree with enhanced data quality. Specifically, network data pre-processing and entropy decision feature selection is done for Enhancing the data quality and relevant training; then, a decision tree classifier is built for reliable intrusion detection. Experimental study on two datasets shows that the proposed model can reach robust results. Actually, our model achieves 99.42% and 98.80% accuracy with NSL-KDD and CICIDS2017 datasets, respectively. The novel approach gives many advantages compared to the other models in term of accuracy (ACC), detection rate (DR) and false alarm rate (FAR).

Keywords— Network security; IDS; ML; Decision tree; Data Quality; Entropy decision; Feature selection;.

Introduction

The computer security threats are becoming quite challenging with the growing capabilities of the adversaries, influencing the reliability of data communication and networks. The recent advancements in cloud computing and IoT technologies enabled new attack vectors for the adversary and even more prone to attacks [1-3]. The IoT applications enables the attacks not only focusing stealing the data but can also impact human lives. For example, a hacked home utility smart heater can be used to automatically increase the temperature and indirectly impact the human beings living in the home [4, 5]. Therefore, the main objective of the security is satisfy all the properties integrity, availability, confidentiality and availability by utilizing the various security tools and policies aiming at protecting the data and also able to detect the attacks targeting the IoT [4, 6]. An intrusion tries to violate one of security objectives and infects systems. Hence, many tools and methods, such as IDS, are developed to secure networks and systems from intrusions [7-9]. Thereby, Intrusion detection is a set of techniques implemented to detect undesirable activities by classifying data activity into normal or intrusion [6, 8].The intrusion detection techniques detect and stop intrusions from outside or within a monitored network.

For this reason, two fundamental detection approaches can be used. The first one is called misuse detection; it is based on a known attack signature to detect intrusion. The second one is named anomaly detection or behavioral detection, based on a deviation from a normal model [1, 8, 11]. The hybrid detection approaches combine advantages of both misuse and anomaly detection and aim to increase detection rate and accuracy of IDS [9, 12, 13]. A considerable distinction is made between network IDS (NIDS) and host IDS (HIDS) [1, 8]. Formally, an IDS can be software or hardware which detects malicious traffic, makes accurately automatic decisions and interrupt intrusions quickly in real-time with an automatic response [6, 8].

Despite their efficiency, the IDS suffer from a number of limitations, such as real-time analysis and detection, generated alarm and data quality that can decrease detection rate and accuracy performances [6, 8]. Therefore, intrusion detection is still an effective and dynamic research field.

Recently, ML methods have been integrated to enhance intrusion detection and reinforce computer security. Numerous research contributions explore how to incorporate ML techniques in intrusion detection to obtain reliable IDS with accurate performances by enhancing data quality and training [14-21]. The decision tree is an induction algorithm which has been used for classification in many issues. It is based on splitting features and testing the value of each one. The splitting process continues until each branch can be labelled with just one classification [22, 23]. The decision tree is more than equivalent representation to the training set. Hence, it can be used to predict the values of other instances not in the training set. Decision tree is widely used as a mean of generating classification rules because of the existence of a simple but very powerful algorithm called Top Down Induction of Decision Trees (TDIDT). It is guaranteed to give a decision tree that correctly corresponds to the data provided Two of the best known being ID3 and C4.5 [23].

On the other side, the data is not always obtained in a structured form. For relevant analysis, the unstructured data have to be pre-processed. This operation is an essential stage which performed to enhance data quality and make accurate decisions. Data quality techniques are implemented before training and classification process [18, 24, 25]. Besides, feature selection is a desirable process aiming to select the useful features to both reduce the computational cost of modelling and to improve the performance of the predictive model [14, 25].

In this paper, we propose a novel network intrusion detection approach based on decision tree method to train and build a binary classifier model and make accurate decisions. The features engineering techniques were used to improve the data quality. Experimental results on NSL-KDD dataset and CICIDS2017 dataset demonstrates that our proposed approach gives good performances in terms of accuracy DR and FAR. Two main contributions have been validated in this research work. Firstly, we implement feature selection using entropy decision technique to improve data quality. Secondly, we build a classifier model based on decision tree algorithm to achieve effective network intrusion detection approach.

The remainder of this paper is organized as follows. The section 2 presents related work on intrusion detection, especially which integrated ML techniques to improve IDS performances. Section 3 describes in detail the proposed solutions for the novel approach. In section 4, we discuss experimental results, performance of proposed model, and its comparison with other models. Finally, the conclusion and future works presented in the paper.

Our Proposed model

As depicted in Figure 1, the proposed model consists of three main components including data quality component, building of classifier component and intrusion detection deployment component. The details of those three components are given in the following.

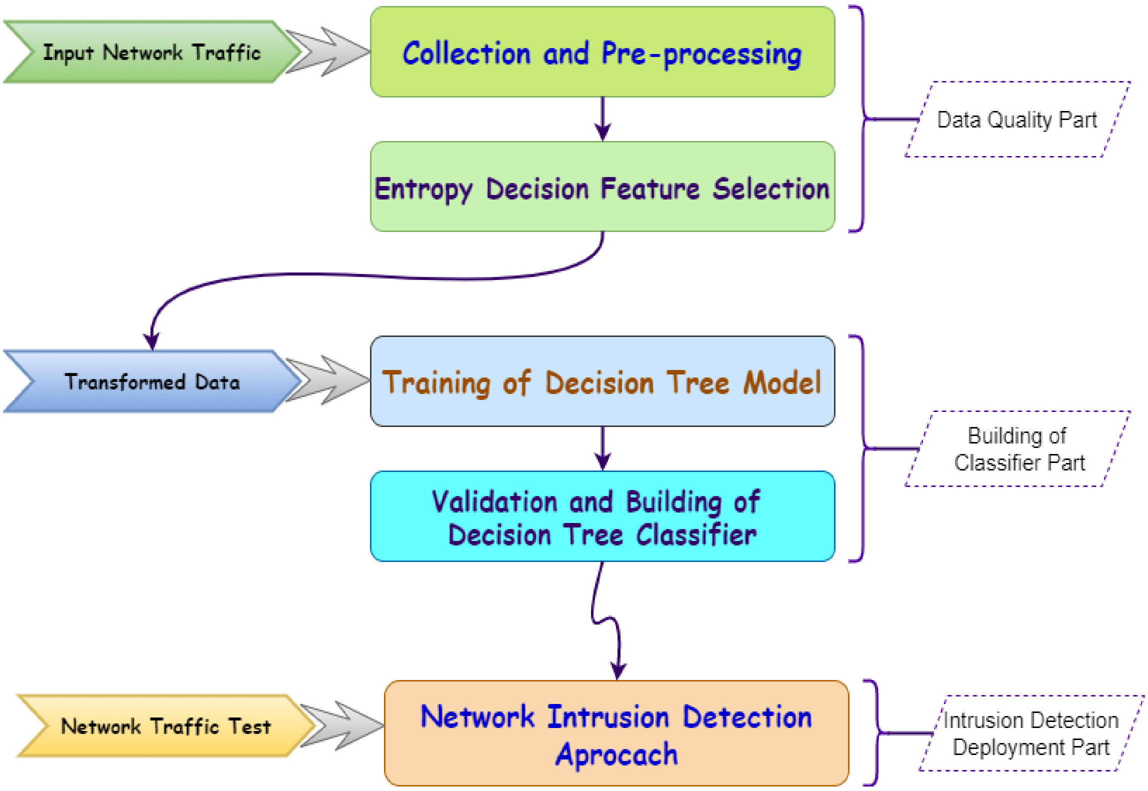

Input Network Traffic
Collection and Pre-processing
Entropy Decision Feature Selection
Data Quality Part
Transformed Data
Training of Decision Tree Model
Validation and Building of Decision Tree Classifier
Building of Classifier Part
Network Traffic Test
Network Intrusion Detection Aprocach
Intrusion Detection Deployment Part

Figure 1: Proposed network intrusion detection model.

Part 1: Data quality process

The main goal of this component is collecting and per-processing the data. Hence, the system executes the process that can gather and accumulate necessary data from networks. Once the data are collected, a specific data pre-processing is performed on gathered network traffic. The data pre-processing portion evaluate the data and ignore the incompatible data types. Further, the data is sanitized and save the resulting data. In addition, the data is transformed and finalize the features of network dataset features. We used the entropy decision technique to select the features.

Part 2: Building of classifier

Once the first part is completed, the second one is started. Generally, the objective of second part, as it is clear in its name, is to build a classifier model. The input here is the transformed data obtained in data quality process part. In the classifier building part, we can distinguish between two main phases: model training phase and model validation phase. In first phase, three portions of data are used for training a decision tree classifier implemented in our proposed approach. Then, in the second phase the rest of data are used to validate our model.

Part 3: Network intrusion detection deployment

After building of classifier model, the third part comes for deploying the network intrusion detection. At this point, actual tests are necessary to improve the performance of reliable IDS. Hence, we are in aptitude to check its capacity to classify activities in normal or abnormal. So, based on the classification results the IDS can make accurate.

Description of proposed solutions

As we mentioned above, the first step which is made by our approach is to collect and transform data with feature selection according to needs of analysis and detection. The data quality is an important and essential task to train and build an accurate intrusion detection model. Hence, this step aims to prepare data for analysis and making accurate decision. We start first with data transformation by applying feature selection using entropy decision on original traffic collected within network traffic to obtain a good training set. In fact, it is a critical step aiming to improve accuracy of our approach. It aims also to overcome training complexity by reducing analysed data and obtain a great model with best performances in terms of accuracy, detection rate and real-time detection. A particular pre-processing is applied on collected network traffic before analysis step. Data normalization is performed. For this, we suggest an implement a particular coding to enumerate feature values and establish a pattern of activities facilitating the distinction between the activities. The goal of the feature extraction is to reduce the number of features in collected data from networks. It aims to summarize most the information contained in this original data by creating new features. The feature selection aims instead to choose the important existing features in the original data and discard less important ones. For this reason, we use entropy decision technique for feature selection. The implementation of components that constitute our approach is described in Figure 2.

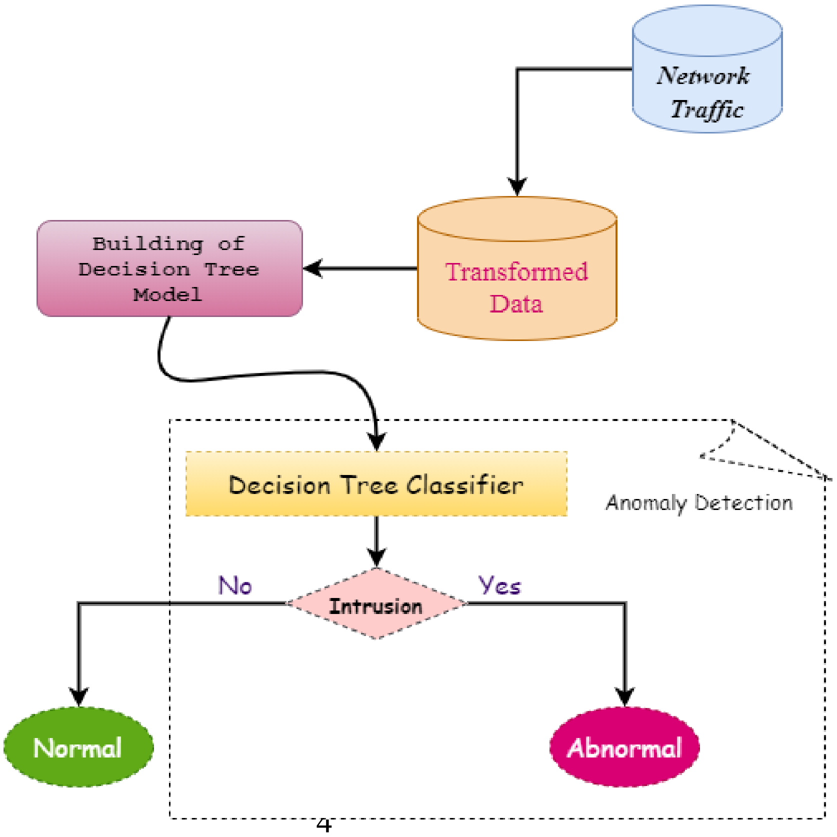


Figure 2: Procedure of validation and building classifier.

We obtain a transformed data, by implementing proposed data quality techniques, aiming to increase our approach accuracy. This allows training and validating of an effective intrusion detection model based on decision tree to make relevant decisions in real time. Moreover, intrusion detection is considered as a classification task aiming to classify incoming traffic in normal activity or intrusion. Hence, the main objective of this part is to predict a binary value to validate classifier able to answer question with a yes or a no. Thus, we encoded both classes in numerical variable: +1 for normal activity and – 1 for intrusion. We remember that the number of features must be fixed in advance. For validation step of our model, there are various strategies used to split the data into a training and test set. In this case, we use the efficient and recommended one, k-fold [1].

According to standard components of an IDS mentioned in [8] [28], our approach is constituted by four parts: data collection part, pre-processing part, making decision part and response part. The proposed approach focuses on pre-processing part by improving data quality technique used to train and build an accurate classifier able to discover intrusions within traffic network. It focuses also on enhancing making decision part by integrating decision tree classifier. A set of research works have been made in [6, 14, 25] to improve others parts of IDS, such as data collection, dimensionality reducing and real-time response which are not taken into account in this research work.

Experimental results and Discussion

Dataset description

The assessment of datasets plays a vital role in validation of intrusion detection approaches. Therefore, for evaluate any IDS using ML techniques one can select desired dataset among a large number of appropriate and available datasets. For instance, numerous public datasets are available [9, 10, 33, 34] and can be used freely for evaluation proposed methods capability. In our case, we have selected two types of datasets including NSL-KDD and CICIDS2017, which are used for training, performances evaluation and validation of proposed approach.

NSL-KDD dataset was created from KDD cup 99 dataset [9, 30]. It contains 125,973 records of training set and 22,544 for test set. It has 22 training instances attacks and 41 features which 21 of them describe connection itself and 19 for nature of connection of the same host [33, 38]. The novelty and instances volume of NSL-KDD dataset make it very practical. On the other hand, CICIDS2017 dataset was created from Canadian Institute for Cyber Security. It aims to overcome the limitations of actual dataset and present an effective dataset for intrusion detection. It is a labelled data set that comprises behaviour and new malware attacks and is consisted of 8 files containing 2,830,743 instances. The CICIDS2017 dataset integrates 80 features network flow captured at July 2017 from network traffic using CICFlowMeter tool [9].

Those two used datasets in this research work, NSL-KDD dataset and CICIDS2017 dataset, are available at [39][40]respectively.

Experiments environment

The experimental setting of our research work is performed and evaluated on a computer with a Core-i7 2700K CPU@ 2.50 GHz and 32 GB of DDR3 running windows 7 professional 64 bits. The entropy feature selection and decision tree model training are implemented using python version 3.8.0.

To validate our proposed intrusion detection model, we use the 10-fold cross validation technique to obtain training and test set. Hence, we split randomly full dataset into ten parts with same size. Nine parts are used in the training and the last part in the test step. Finally, the performances of model are presented by repeating this procedure ten times.

Data transformation

In the implementation step, we propose to extract samples of dataset to avoid some drawbacks such as, processing and big volume of data. The data extraction from each used dataset is given in Table 1.

Table 1: Data extraction from NSL-KDD and CICIDS2017 datasets

	Category	Original size	Extracted size
NSL-KDD dataset	Training	125,973	25,195
	Test	22,544	4,509
	Total	148,517	29,704
CICIDS2017 dataset	Benign	2,273,097	113,655
	Attack	557,646	27,883
	Total	2,830,743	141,538

Feature selection is a relevant technique included by our network intrusion detection approach. It is implemented and incorporated to select useful features for reliable detection and decision making. For this, we implement entropy decision technique.

The encoding step is performed to assign numeric values to categorical features for making relevant processing**.** To avoid undesirable influence problem of high weights, we normalize continuous features values. The Equation 1 is used to find the new value. Hence, we make the values of each feature run from 0 to 1. If the lowest value of a given feature x is min and the highest value is max, we convert each value of x to:

$$(value(x) - \min)/(max - \min) \quad (1)$$

Furthermore, all continuous features are in range [0, 1].

Metrics evaluation and discussion

The most obvious criterion to use for estimating the performances of a classifier is predictive accuracy. The proportion of a set of unseen instances that it correctly classifies. For numerical performances evaluation of the proposed model, the following metrics are used.

These metric performances are not dependent on the size of the training and test set and can be really helpful in assessing the performance of the full model. Based on confusion matrix (table 2), the performances metrics are calculated:

ACC is obtained from equation 2. It is the ratio of instances that are correctly predicted as normal or attack to the overall number of instances in test set.

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \qquad (2)$$

DR is calculated using equation 3 and indicates the ratio of the number of instances that are correctly classified as attack to the total number of attack instances present in test set.

$$\mathrm{DR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (3)$$

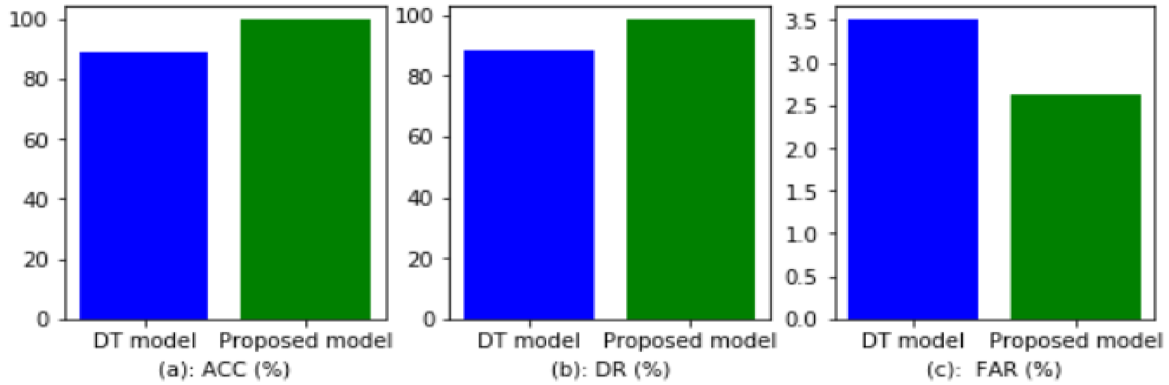
FAR is obtained from equation 4 and represents the ratio of instances which is categorized as attack to the overall number of instances of normal behaviour.

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \qquad (4)$$

Table 2: Confusion matrix

Actual class	Predicted class	
	Attack	Normal
Attack	TP	FN
Normal	FP	TN

In this research work, we start with comparing detection assessment of our proposed model for novel approach and decision tree model only. The results showed in Figure 3 and Figure 4 demonstrate this comparison according to ACC, DR and FAR on NSL-KDD dataset and CICIDS2017 dataset.



(a): ACC (%) (b): DR (%) (c): FAR (%)

Figure 3: (a) ACC results of DT model and our proposed model on NSL-KDD dataset (b) DR results (c) FAR results

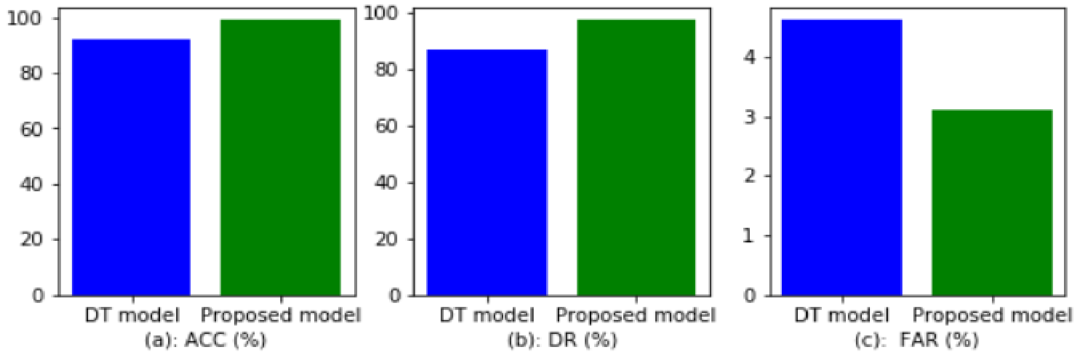


Figure 4: ACC results of DT model and our proposed model on CICIDS2017 dataset (b) DR results (c) FAR results

The figures 3 (a) and 4 (a) show that accuracy of the proposed model is specifically better than model based on decision tree only, figures 3 (b) and 4 (b) demonstrates the DR of both IDS. It validates that the DR of proposed IDS model is higher than the IDS based on decision tree only on NSL-KDD dataset and CICIDS2017 dataset.

The results demonstrated above are summarized in Table 3 and Table 4. They show that our proposed model can reach significant performances than decision tree only. For the NSL-KDD dataset, the ACC of our proposed model achieves 99.42% while decision tree only exceeds 89%. In terms of DR and FAR, our proposed model obtains 98.2% and 2.64% respectively, while decision tree only presents DR 88.5% and FAR 3.5%. For CICIDS2017 dataset, our proposed model indicates high performances in terms of ACC 98.8%, DR 97.3% and FAR 3.10%. Besides, the decision tree only gives ACC 92%, DR 86.7% and FAR 4.6%.

Table 3: performances metrics of decision tree and proposed model using NSL-KDD dataset

	ACC (%)	DR (%)	FAR (%)
Decision Tree	89.00	88.50	3.50
Proposed approach	99.42	98.20	2.64

Table 4: performances metrics of decision tree and proposed model using CICIDS2017 dataset

	ACC (%)	DR (%)	FAR (%)
Decision Tree	92.00	86.70	4.60
Proposed approach	98.80	97.30	3.10

The results obtained validate that our approach gives great detection capability in terms of ACC, DR and FAR. Specifically, they demonstrate that the performances metrics of our proposed model are higher o NSL-KDD dataset but low CICIDS2017 dataset. According to the evaluation performances; our proposed IDS model can reach great performances. The comparison with model which uses decision tree only indicates the effectiveness of our network intrusion detection approach.

Conclusion and future works

Intrusion detection is a set of enhanced techniques implemented to monitor systems and data to be more secure. In this paper, we present a reliable network intrusion detection approach based on decision tree classifier and engineering feature techniques. According to heterogeneity of data, a pre-processing phase is setting up to increase detection rate and accuracy of IDS. Also, a feature selection technique based on entropy decision tree method is handled before building model for high data quality. The validation of novel approach is achieved by proposed solutions that guarantee an efficient accuracy. The performances are evaluated on two datasets: NSL-KDD and CICIDS2017. Hence, the novel proposed network intrusion detection approach presents many advantages and provides high accuracy compared with other models. The future works will integrate other efficient ML techniques such as deep learning in various parts to empower detection rate and accuracy of our approach.

References

[1] Chiba Z., Abghour N., Moussaid K., Elomri A., and Rida M., "Intelligent Approach to Build A Deep Neural Network Based IDS for Cloud Environment Using Combination of Machine Learning Algorithms," Computers & Security, vol. 86, pp. 291-317, 2019.

[2] A. Irshad, S. A. Chaudhry, O. A. Alomari, K. Yahya and N. Kumar, "A Novel Pairing-Free Lightweight Authentication Protocol for Mobile Cloud Computing Framework," in *IEEE Systems Journal*, 2020.

[3] Chaudhry, S.A., Kim, I.L., Rho, S. et al. An improved anonymous authentication scheme for distributed mobile cloud computing services. *Cluster Comput* 22, 1595–1609, 2019.

[4] A. Irshad, M. Usman, S. A. Chaudhry, H. Naqvi and M. Shafiq, "A Provably Secure and Efficient Authenticated Key Agreement Scheme for Energy Internet-Based Vehicle-to-Grid Technology Framework," in *IEEE Transactions on Industry Applications*, vol. 56, no. 4, pp. 4425-4435, 2020.

[5] S. A. Chaudhry, M. S. Farash, N. Kumar and M. H. Alsharif, "PFLUA-DIoT: A Pairing Free Lightweight and Unlinkable User Access Control Scheme for Distributed IoT Environments," in *IEEE Systems Journal*, 2020.

[6] Fernandes, G., Rodrigues, J.J.P.C., and Carvalho, L.F., "A comprehensive survey on network anomaly detection. Telecommun Syst 70, pp. 447-489 2019.

[7] Khraisat A., Gondal I., Vamplew P., and Kamruzzaman J., "Survey of intrusion detection systems: techniques, datasets and challenges," Cybersecurity, 2019.

[8] Ferrag M. A., Maglaras L., Moschoyiannis S., and Janicke H., "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," Journal of Information Security and Applications, vol. 50, 2020.

[9] Ji S.Y., Jeong B.K., Choi S., Jeong D.H., "A multi-level intrusion detection method for abnormal network behaviors," Journal of Network and Computer Applications, vol 62, pp. 9-17, 2016.

[10] Çavuşoğlu Ü., "A new hybrid approach for intrusion detection using machine learning methods," Appl Intell 49, pp. 2735–2761, 2019.

[11] Masdari M., and Khezri H., A survey and taxonomy of the fuzzy signature-based Intrusion Detection Systems, Applied Soft Computing, vol 92, 2020.

[12] Alazzam H., Sharieh A., Sabri K.E., "A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer," Expert Systems with Applications, vol.148, 2020.

[13] Aldweesh A., Derhab A., and Ahmed Z. E, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," Knowledge-Based Systems, vol. 189, 2020.

[14] Amini M., Rezaeenour J., and Hadavandi E., "A Neural Network Ensemble Classifier for Effective Intrusion Detection Using Fuzzy Clustering and Radial Basis Function Networks," International Journal on Artificial Intelligence Tools, Vol. 25, No. 02, 2016.

[15] Fang W., Tan X., and Wilbur D., "Application of intrusion detection technology in network safety based on machine learning," Safety Science, vol. 124, 2020.

[16] Gu I., Wang L., Wang H., Wang S., "A novel approach to intrusion detection using SVM ensemble with feature augmentation," Computers & Security, vol. 86, pp. 53-62, 2019.

[17] Hassan M.H., Gumaei A., Alsanad A., Alrubaian M., and Fortino G., A hybrid deep learning model for efficient intrusion detection in big data environment, Information Sciences, vol. 513, pp. 386-396, 2020.

[18] Sethi, K., Sai Rupesh, E., Kumar, R. et al. "A context-aware robust intrusion detection system: a reinforcement learning-based approach," International Journal of Information Security, pp. 657–678, 2020.