ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 9, No. 4, December 2021, pp.1102-1109

Comparison and analysis of supervised machine learning algorithms

Alaa Abdulhussein Daleh Al-magsoosi [1] , Ghassan Nashat Mohammed[2], Zamen Abood Ramadhan[3]
[1]ITMO University and College For Pure Science, University of Wasit, Iraq
[2]Department of planning and Studies, Ministry of Higher Education, Baghdad Iraq
[3]College for pure science University of Wasit, Iraq

ABSTRACT

When investigating a network for signs of infiltration, intrusion detection is used. An intrusion detection system is designed to prevent unwanted access to the system. Data mining techniques have been employed by a number of researchers to detect infiltrations in this field. Based on distance measurements, this study proposes algorithms for supervised machine learning. In terms of detection rate, accuracy, false alarm rate, and Matthews correlation coefficient, supervised machine learning techniques surpass other algorithms. When it comes to serial execution time, the supervised machine learning algorithms surpassed all other Actions in terms of serial execution performance.

Keywords: Classification of normal anomalies, Parallel feature selection, rule weighting algorithm

Corresponding Author:

Alaa Abdulhussein Daleh Al-magsoosi
ITMO University and College for Pure Science
University of Wasit, University
Wasit, Iraq
aal-magsoosi@itmo.ru, adleah@uowasit.edu.iq

1. Introduction

More than two quintillion data bytes have produced and traded every day. Conventional detection systems are unable to sense the intrusion in a way that is compatible with high data volume and speed. Big data technologies are being used to deal with this threat efficiently. The use of supervised machine learning algorithms can be a special device or software system that spontaneously screens and identifies attacks or infiltrations and issues alerts to the computer or network. It is a system that monitors the traffic of data within the network with the intention of detecting any suspicious activity or any potential threats [1, 2]. The alert report helps administrators or users to identify security holes in the system or network and thus solve them. Intrusion detection technologies such as host and network detection are the subject of anomalous approaches to data analysis. Network Intrusion Sensor can detect network traffic and control multiple network hosts to identify any errors. SVM can be used as an enhanced, non-programmable machine learning technology that demonstrates how different SVM, LR, LDA, RF and CART algorithms can be implemented [2],[3, 4] such as modified logistic regression, decision tree address, artificial network and use of machine learning. Unusual behaviours may be detected using software. And network traffic equipment.

2. Intrusion detection ML algorithm

ML refers to (Machine Learning) that stands for Artificial Intelligence (AI) branch. Where machine learning allows the system to learn and predict and improves its automatic ability to experiment without being programmed in detail depending on a set of algorithms. In addition, ML algorithms work more precisely in sensing attacks of an enormous quantity of data in the shortest time possible [5-8]. And ML algorithms have been categorized into three classes:

- Supervised
- Unsupervised
- Semi-supervised

2.1. Supervised machine learning algorithms

The monitoring algorithm deals with completely classified information and identifies the connection among data and its class [9]. This can be achieved by either regression or classification. There are dual stages in the classification: preparation and assessment. The training data was carried out using the reaction vector. Supporting Vector Machine (SVM), discrimination, Nauve Bayes, Nearest Neighbor's Network and logistic regression are the usual algorithms under classification group [10, 11]. Such algorithms are Linear Regression SVR Ensemble Methods, Decision Tree is shown in Figure 1, and Random Forest while other algorithms are Linear Regression. In this article, support logistical regression of vector machines The discussion is on the Linear Discriminant Analysis SVM, LR, LDA, RF and CART.

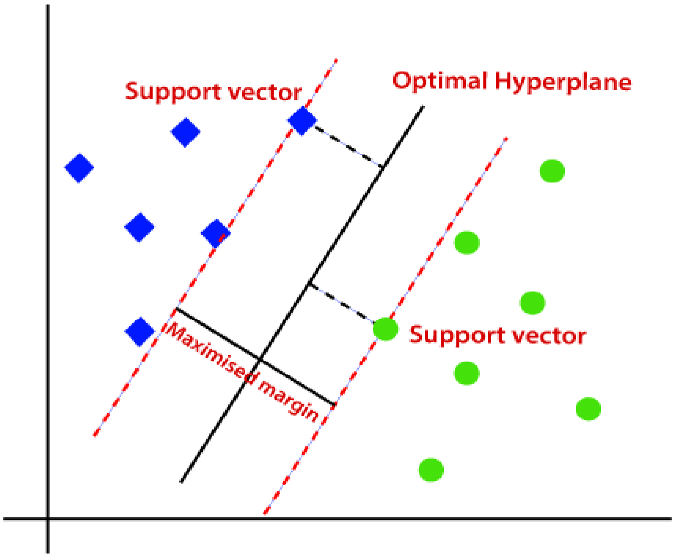

Figure 1. Linear Regression SVR

Controlled anomaly detection learning methods can be classified into four methods:
1) Data training consisting only of regular events;
2) data training consisting only of anomalous events;
3) data training consisting of classified ordinary and anomalous events;
4) data training consisting of several labeled incident groups.
Single classification formulations are Methods 1 and 2 Method 3 has a difficulty with two-class classification, and Method 4 has a multi-class classification. Note that Methods 3 and 4 have here shown that training data must be allocated to labels[12]. This distinguishes between unattended methods of learning that don't need marks to be attached to the training results. This statement is often made in order to differentiate between methods 1 and 2, where all training information is collected by a single class [13], and the marking is therefore negligible.

2.2. Support vector machine (SVM)

SVM is one of the most popular machine learning (ML) algorithms out there. Regression and classification can both be accomplished with SVM. The algorithm may be trained with labeled data and the hyper plane can be used to divide the data into classes, maximizing the range of all the attacking classes [14, 15]. Cascaded multi-class classification can be achieved using SVM, according to Mehmood et al. [16]. In Figure 2, the types and parameters of the kernels utilized in SVM are clearly illustrated.

The AE model employs an unsupervised learning method with one or more hidden layers. Trait learning and dimensional reduction can benefit from nonlinear generalization of trait learning [17]. Units of input and output are the same in AE, and the function vector parts are also equal in number. There are exactly as many units in the hidden layer as there were in the bottleneck layer, which was determined before to training.

A vector-based machine learning approach, SVM is generally a supervised model [18]. Two classes and two people are shown to be making use of categorization learning methods in Figure 1. You can categorize a new text because each classification on the SVM system has its own unique data set number [19].

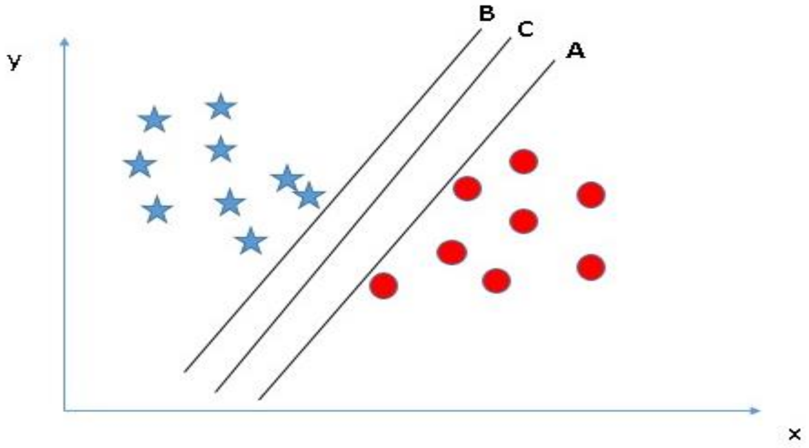

Figure 2. Svm algorithm

2.3. Logistic regression (LR)

Logistic regression is named for the function used at the core of the method, the logistic function. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1. It was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out is shown in Figure 3.

It is another way of borrowing from the mathematical data profession through machine learning. It is also the aim of the process of binary classification issues (difficulties with more than just two class moral values). Logistical regression is used to form a group results such as real passes/completely failures, optimistic and constructive/no or neutral again, and then we use the probability distribution class as fraudulent and not fraud in case of credit card fraud identification [20].

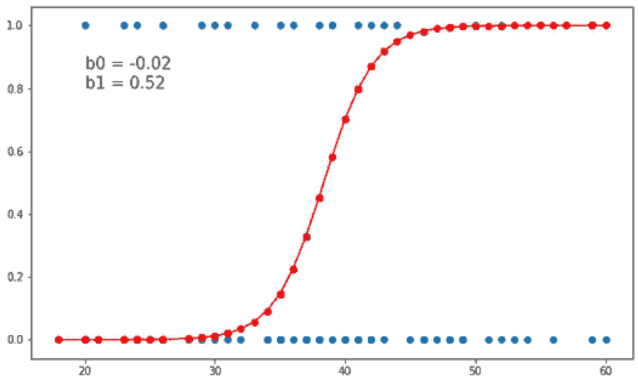

Figure 3. Logistic Regression algorithm

Linear models are intended for regressions in which a linear combination of the input variables is expected to be a target value. LR is a linear classification instead of regression model, despite its name. The probabilities that describe the potential result of a single test are modeled with a logistic function in this model. Scikit-learn is using the Coordinate Descent (CD) algorithm [21] Logistic Regression, the default resolver of which is.

Logistic regression is often referred to as the regression of Binomial logistics. The sigmoid function is dependent on which the output is likely and the input will range from -infinity -+infinity. Let us address some Linear Regression benefits and drawbacks.

2.4. Linear discriminant analysis (LDA)

For dimensionality reduction and prediction, LDA is a linear supervised linear ML technique. Bayesian inference is used to determine the likelihood that a new input belongs to a particular class.

Data sets and test vectors can be analyzed using two alternative approaches in the converted space. Based on class, a person is transformed. The ratio of class variance to class variance is maximized in this strategy. In order to obtain a high enough level of class separability, it is critical that this ratio be made even better. Data sets must be transformed independently using two optimization parameters in the class approach. Class-independent transformation: This solution aims to minimize the gap between total differences and the differences between classes. Figure 4 illustrates a strategy that applies only one optimization criterion for changing data sets and consequently discards all data points.

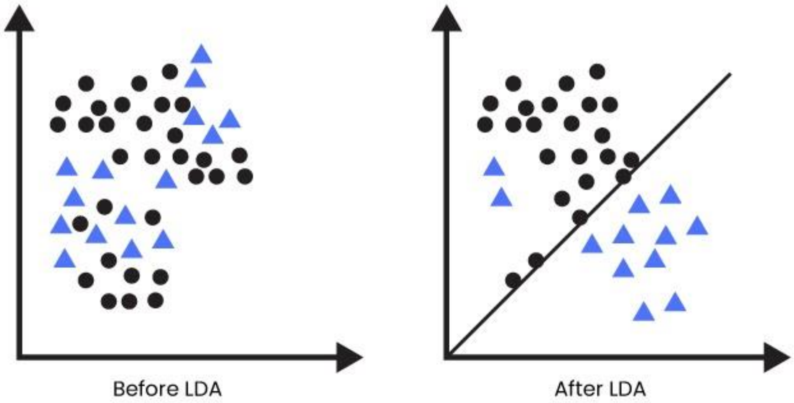

Figure 4. Linear discriminant analysis algorithm

The LDA technique's aim is to project a lower dimension of the original data matrix. Three steps needed to be taken to accomplish this objective. The first step is to quantify the segregation between groups, which is called the interclass or interclass matrix. The second step is to measure the distance between the mean and the samples of a class known as the internal or internal variance [22]. The third step is to design the lower dimensional space that maximizes the variance within classes and reduces the variance within class. These three are explained in this segment. Linear Discriminant Analysis (LDA) and Variable Inference in Near Real Time are used to track deviations in internet traffic [23]. A technical solution, which uses Natural Language Processing (NLP) methodology to identify potential malicious attacks and network configuration issues, is explained, and results are presented demonstrating the implementation of the concept. There are potential use cases for this technology in the areas of anomalous data detection

2.5. Classification and regression tree (CART)

No online supervision is required for CART, which uses a simple ML algorithm for classification. When using CART, the target variable should be categorical, however when using regression trees, the target variable should be continuous. In CART, the Gini index is a measure used to describe the data.

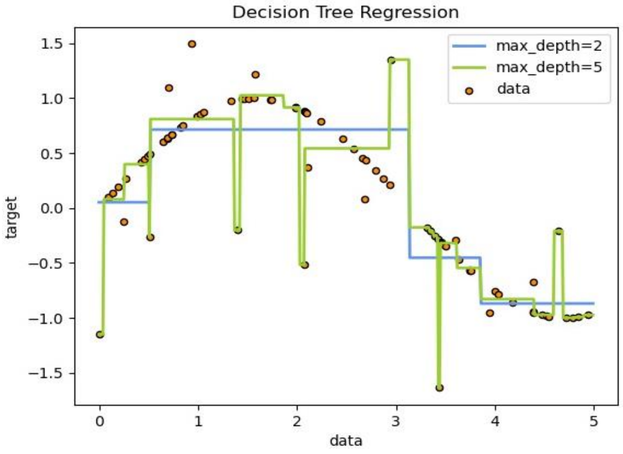

Figure 5. Classification and Regression Tree algorithm

Being able to apply classification and regression techniques simultaneously is a significant benefit of this methodology. As a result, a binary tree is formed, with each internal node having two outbound edges. Cost Complexity Pruning and IG, GI and twoing parameters can all be used in the splitting process. CART is an algorithm that we employed in our work with the scikit-learn library [24].

2.6. Random forest (RF)

RF stands for a dynamic non-linear algorithm that is employed for regression and classification. This will build several decision-making bodies for model education, with the results of predictions collected from all the trees producing a response, as Ensemble techniques are mentioned. The RF classification system operates the following: the more trees the model contains, the better the precision and the more the model is not over fitted is shown in Figure 6.

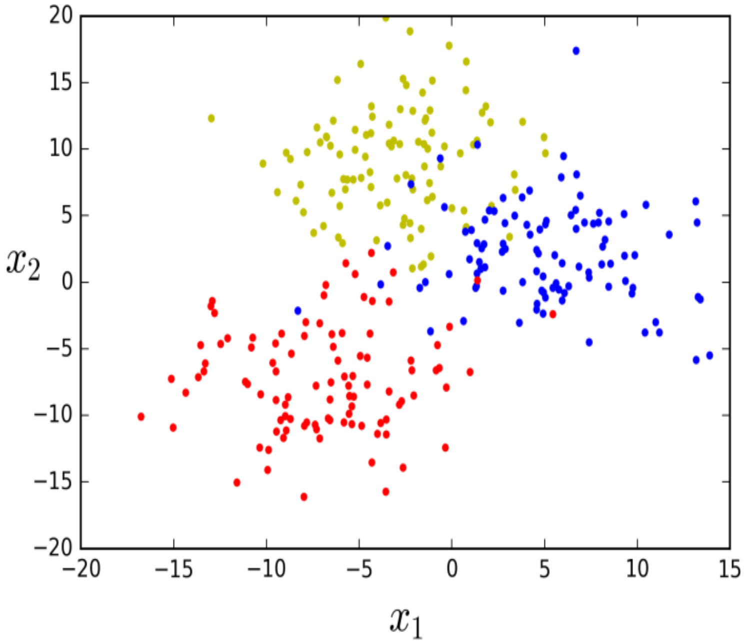

Figure 6. Random Forest algorithm

Classifier is a classification ensemble that integrates multiple classifiers of decision-making-trees to forecast the class[25]. Each tree is sampled individually and uniformly using the majority rule. Every new input data point is forwarded to each of its trees by the RF classifier to select the class class classified by the most trees.
One of the ensemble classifier approaches is Random Forest. If an ensemble classifier is a decision-making classifier, the classifier set is a 'land.' The random collection of attributes for each separation node [26] is used to build each decision tree. In 2001, Breich suggested the random forest algorithm. Study performed by [27] included several anomaly detection experiments using random forests.

2.6.1. Ensemble methods

This ML technique integrates many simulations to create the desired predictive model. The core concept behind ensemble approach is to group all weak students into a strong learner, thus increasing the model's accuracy. Bagging, boosting and stacking are some typical forms of ensemble approaches. In approach to this method, Gautam et al .[7] have developed a path with ML algorithms , in the recent new papers that use machine learning algorithm which used for anomaly detection the algorithm partial decision. It showed that the approach of the ensemble is higher than SVM, LR, LDA, RF and CART algorithms.

2.6.2. Performance evaluation

All the preprocessing strategies have been validated with Supervised learning algorithms and we present in Table 1 the findings of the best methods of All SVM ,LR,LDA, RF and CART algorithm. The below table is shown the advantages of each algorithm and the drawback with performance analysis for each machine learning algorithm to compare between them.

Table 1. Supervised learning algorithms analysis

Author and Publication	Methodology Employed	Dataset	Advantage	Limitation	Performance Analysis
[16]	Comparison of different moderated algorithms for a deviation-based detection technique (Svm).	NSL-KDD	Has high detection rate	The training and testing speed is slow	It has cannot detect novel attacks
[17]	The Fuzzy clustering and Svm used for classification	KDD CUP	The process was carried out by dividing the heterogeneous training group into subgroups	Showed vulnerability to handling complex data in a large data set	Higher IDS detection accuracy, fewer attacks, and stronger detection stability
[18]	Most strategies were implemented throughout the identification of card fraud.by using (svm)	Cc data set	It achieves its own feedback process by enhancing classifier detection rate as well as effectiveness.	The method was successful with cc dataset	This model was very accurate, with a false alarm rate of 1.87%.
[21]	Show better results in long distances vs. Attacks	KDD- -99	Better results appear in long distances versus during attacks.	Because of its strong rules, it is good at detecting anomalous data	Good at detecting anomalies
[20]	Anomaly detection techniques based on a single classification	KDD’99	Mechanisms for detecting anomalies within a single classifier.	Force of class stability versus lack of data	To improve the performance of intrusion detection systems.
[24]	Make the neural network classification work as effectively as possible	KDD 99	This system provides less time for online learning.	This system was characterized by an increased false alarm rate.	The quality of the rating and the error rates were used as a parameter to evaluate the performance

The anticipated work in this paper can be developed more and more employing genetic algorithm [28], internet of thing (IoT) [29], cloud computing and Arduino as viable future trends [30, 31].

3. Conclusion

Anomalous data detection algorithms are classified into two types in terms of detecting misuse and defects. As we adopted in the detection of anomalies, the generation of predictive patterns, sequence matching, statistics, and supervision. After we analysis the methodologies which has been used for anomaly detection, we found that Support Vector Machine algorithm is the preferred one and the high accuracy in anomaly detection.

References

[1] A. A. Sodemann, M. P. Ross, and B. J. Borghetti, "A review of anomaly detection in automated surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* vol. 42, no. 6, pp. 1257-1272, 2012.

[2] H. ALRikabi, H. Tauma, "Enhanced Data Security of Communication System using Combined Encryption and Steganography," *International Journal of Interactive Mobile Technologies,* vol. 15, no. 16, pp. 144-157, 2021.

[3] H. T. Salim, N. A. Jasim, "Design and Implementation of Smart City Applications Based on the Internet of Things," *International Journal of Interactive Mobile Technologies (iJIM),* vol. 15, no. 13, pp. 4-15, 2021.

[4] A. S. Abdalrada, O. H. Yahya, A. H. M. Alaidi, N. A. Hussein, H. T. Alrikabi, and T. Al-Quraishi, "A predictive model for liver disease progression based on logistic regression algorithm," *Periodicals of Engineering and Natural Sciences,* Article vol. 7, no. 3, pp. 1255-1264, 2019.

[5] S. Ganapathy, K. Kulothungan, S. Muthurajkumar, M. Vijayalakshmi, P. Yogesh, and A. Kannan, "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey," *EURASIP Journal on Wireless Communications and Networking,* vol. 2013, no. 1, pp. 1-16, 2013.

[6] A. Al-zubidi, R. K. Hasoun, S. Hashim, H. Salim, "Mobile Application to Detect Covid-19 pandemic by using Classification Techniques: Proposed System," *International Journal of Interactive Mobile Technologies,* vol. 15, no. 16, pp. 34-51, 2021.

[7] N. A. Hussien, A. A. Daleh Al-Magsoosi, H. TH, and F. T. Abed, "Monitoring the Consumption of Electrical Energy Based on the Internet of Things Applications," *International Journal of Interactive Mobile Technologies,* vol. 15, no. 7, pp. 17-29, 2021.

[8] I, A. Aljazaery, H. Salim, "Encryption of Color Image Based on DNA Strand and Exponential Factor," *International Journal of Interactive Mobile Technologies (iJIM),* 2021.

[9] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing,* vol. 18, no. 1998, pp. 1-8, 1998.

[10] V. Cabannes, F. Bach, and A. Rudi, "Disambiguation of weak supervision with exponential convergence rates," *arXiv preprint arXiv:2102.02789,* 2021.

[11] A. S. Hussein, R. S. Khairy, S. M. M. Najeeb, and H. T. ALRikabi, "Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 5, 2021.

[12] Q. Liu, X. Wang, X. Huang, and X. Yin, "Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data," *Tunnelling and Underground Space Technology,* vol. 106, p. 103595, 2020.

[13] D. Snow, "Machine learning in asset management—Part 1: Portfolio construction—Trading strategies," *The Journal of Financial Data Science*, vol. 2, no. 1, pp. 10-23, 2020.

[14] D. Snow, "Machine Learning in Asset Management—Part 2: Portfolio Construction—Weight Optimization," *The Journal of Financial Data Science*, vol. 2, no. 2, pp. 17-24, 2020.

[15] R. M. Al_airaji, I. A. Aljazaery, S. K. Al_dulaimi, and H. T. S. Alrikabi, "Generation of high dynamic range for enhancing the panorama environment," *Bulletin of Electrical Engineering and Informatics,* Article vol. 10, no. 1, pp. 138-147, 2021.

[16] T. Mehmood and H. B. M. Rais, "Machine learning algorithms in context of intrusion detection," in *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*, 2016, pp. 369-373: IEEE.

[17] S. Kim, W. Jo, and T. Shon, "APAD: autoencoder-based payload anomaly detection for industrial IoE," *Applied Soft Computing,* vol. 88, p. 106017, 2020.

[18] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "An efficient credit card fraud detection model based on machine learning methods," *International Journal of Advanced Science and Technology,* vol. 29, no. 5, pp. 3414-3424, 2020.

[19] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. A. Khan, "Performance analysis of machine learning algorithms in intrusion detection system: A review," *Procedia Computer Science,* vol. 171, pp. 1251-1260, 2020.

[20] K. Demestichas, N. Peppes, T. Alexakis, and E. Adamopoulou, "An Advanced Abnormal Behavior Detection Engine Embedding Autoencoders for the Investigation of Financial Transactions," *Information,* vol. 12, no. 1, p. 34, 2021.

[21] Á. M. Guerrero-Higueras, N. DeCastro-Garcia, and V. Matellan, "Detection of Cyber-attacks to indoor real time localization systems for autonomous robots," *Robotics and Autonomous Systems,* vol. 99, pp. 75-83, 2018.

[22] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI communications,* vol. 30, no. 2, pp. 169-190, 2017.

[23] A. Thornton, B. Meiners, and D. Poole, "Latent Dirichlet Allocation (LDA) for Anomaly Detection in Avionics Networks," in *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC),* 2020, pp. 1-5: IEEE.

[24] P. I. Radoglou-Grammatikis and P. G. Sarigiannidis, "An anomaly-based intrusion detection system for the smart grid based on cart decision tree," in *2018 Global Information Infrastructure and Networking Symposium (GIIS),* 2018, pp. 1-5: IEEE.

[25] M. Injadat, F. Salo, and A. B. Nassif, "Data mining techniques in social media: A survey," *Neurocomputing,* vol. 214, pp. 654-670, 2016.

[26] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," *The Morgan Kaufmann Series in Data Management Systems,* vol. 5, no. 4, pp. 83-124, 2011.

[27] R. K. Singh, S. Dalal, V. K. Chauhan, and D. Kumar, "Optimization of FAR in intrusion detection system by using random forest algorithm," in *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE),* 2019.

[28] Y. S. Mezaal, S. F. Abdulkareem, “Affine Cipher Cryptanalysis Using Genetic Algorithms,” JP Journal of Algebra, Number Theory and Applications, vol. 39, no. 5, pp. 785-802, 2017.

[29] Y. S. Mezaal, L. N. Yousif, Z. J. Abdulkareem, H. A. Hussein, S. K. Khaleel, “Review about effects of IOT and Nano-technology techniques in the development of IONT in wireless systems,” International Journal of Engineering and Technology (UAE), vol. 7, no. 4, 2018.

[30] Y. S. Mezaal, H. H. Madhi, T. Abd, S. K. Khaleel, “Cloud computing investigation for cloud computer networks using cloudanalyst,” Journal of Theoretical and Applied Information Technology, vol. 96, no. 20, 2018.

[31] Z.K. Hussein, H.J. Hadi, M.R. Abdul-Mutaleb, Y.S. Mezaal, "Low cost smart weather station using Arduino and ZigBee." *Telkomnika* , vol.18, no. 1, pp.282-288, 2020.