Indonesian Journal of Electrical Engineering and Computer Science
Vol. 18, No. 2, May 2020, pp. 1074~1080
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v18.i2.pp1074-1080

Efficient method for breast cancer classification based on ensemble hoffeding tree and naïve Bayes

Royida A. Ibrahem Alhayali[1], Munef Abdullah Ahmed[2], Yasmin Makki Mohialden[3], Ahmed H. Ali[4]
[1]Department of Computer Engineering, College of Engineering, University of Diyala, Diyala, Iraq
[2]Faculty of Al-Hawija Technical institute, Northern Technical University, Iraq
[3]Department of Computer Science, College Of Science, Mustansiriyah University, Iraq
[4]AL Salam University College Computer Science Department Baghdad, Iraq

Article Info

Article history:

Received Aug 22, 2019
Revised Nov 23, 2019
Accepted Dec 7, 2019

Keywords:

Breast cancer
Classification
Hoeffding tree
Machine Learning
Naïve Bayes

ABSTRACT

The most dangerous type of cancer suffered by women above 35 years of age is breast cancer. Breast Cancer datasets are normally characterized by missing data, high dimensionality, non-normal distribution, class imbalance, noisy, and inconsistency. Classification is a machine learning (ML) process which has a significant role in the prediction of outcomes, and one of the outstanding supervised classification methods in data mining is Naives Bayess Classification (NBC). Naïve Bayes Classifications is good at predicting outcomes and often outperforms other classifications techniques. Ones of the reasons behind this strong performance of NBC is the assumptions of conditional Independences among the initial parameters and the predictors. However, this assumption is not always true and can cause loss of accuracy. Hoeffding trees assume the suitability of using a small sample to select the optimal splitting attribute. This study proposes a new method for improving accuracy of classification of breast cancer datasets. The method proposes the use of Hoeffding trees for normal classification and naïve Bayes for reducing data dimensionality.

Corresponding Author:

Royida A. Ibrahem Alhayali,
Department of Computer Engineering, College of Engineering,
University of Diyala, Diyala, Iraq.
Email: royida.alhayali@engineering.uodiyala.edu.iq

1. INTRODUCTION

Breast cancer is the second leading cancer among women worldwide [1]. The occurrence of breast cancer is increasing yearly due to heredity, increased life expectancy, different lifestyles, and food habits. This research primarily aimed at building a classification model for breast cancer classification, as well as providing an accurate diagnosis to physicians to provide effective treatment to save life. Thus, an efficient classification model can help to reduce cancer-related mortality among women. Classification is one of the best available data mining techniques for the prediction of outcomes from a given dataset. The NBC [2, 3] is a well-known supervised classifier which can be used to predict outcome from a given data set. The NBC generally exhibits good performance when compared to other classifiers; such performances are attributed to the simple nature, less computational difficulty, good prediction accuracy, and less memory-dependence of the NBC. NBC also outperforms other classifiers due to the assumption of independence between the predictors. However, the accuracy of NBC is usually lost due to this assumption of Independence and bad initialized parameters [4]. The presence of inter-related attributes in datasets can also affect the accuracy of NBC. Therefore, it is a tedious task to improve the accuracy of NBC with parameter optimization [5]. Data dimensionality is normally reduced using feature selection (FS) techniques. The FS technique eliminates irrelevant and redundant features from datasets as they have no important part to play during the classification process. Breast cancer dataset has 762691 instances with 134 attributes; however, only 7 attributes are often involved in classification process

using FS technique. In this study, a novel approach of using hoeffding tree to minimize accuracy loss in NBC due to poor parameters initialization is proposed. Hoeffding tree (HT) constructs and analyzes decision trees using the Hoeffding bound. The role of the Hoeffding bounds is to determine the number of required instances to be executed to attain a certain confidence level. The performance of the proposed techniques was evaluated on the Breast Cancer Data set hosted in the UCI Machine Learning Repository.

The organization of this study is as follows: Section-II reviewed the previous works in this domain while section-III presented detailed discussions on NBC and Hoeffding tree. The explanation of the dataset was presented in section-IV while the implementation of the proposed method was presented in section–V. Section-VI presented and discussed the results of the experiments while the conclusion drawn from the study was presented in the last section.

2. RELATED WORK

Breast cancer classification has received several research interests; therefore, the study of data mining techniques and improving the classification of breast cancer is highly required. This section provides a brief review of some of the previous works related to this study. Attarodi et al. [6] presented the combination of Mel-frequency cepstral coefficient (MFCC) and Auto Correlation techniques in which the 1st sound range was separated with a high level of precision. They succeeded in using SVM equipped with RBF and Quadratic kernels to classify 3 groups of newborns into normal sound, murmur sound due to VSD, and murmur due to AS (aortic stenosis). Another study by Kavitha et al. [7] reported the development of a framework with numerous steps such as outlier detection and PCA-guided feature extraction. Wrapper filter was used during the subset features selection to ensure better results. The system presented an improved performance compared to the other scoring functions such as Pearson correlation and Euclidean distance coefficients. Shenfield et al. [8] suggested the consideration of a multi-objective approach to ANNs' evolutionary design using a robust optimizer based on the novel MOEA/DDRA algorithm and incorporation of decision-maker preferences. Amrane et al. [4] classified breast cancer using NBC and K-nearest neighbor (KNN). They implemented the two methods and compared their performance accuracy using cross-validation. Sara et al. [3] presented the categorization and automatic classification of stromal regions with respect to their maturity; they proved that this classification agreed with that of skilled observers, hence, providing a quantitative and repeatable measure for prognostic application. They classified breast cancer stroma regions-of-interest (ROI) using local binary patterns and multiscale basic image features in combination with a random DT classifier.

3. CLASSIFIERS

Classification is a significant process in DM and in the building of learner systems. The learning algorithm [9] builds a classifier based on a set of instances, such as a feature set of values (x1,x2,…. ,xn) in which xi represents the value of feature Xi. Assume c to be the classification feature and c C ℝm as an instance of C. Classification aims at establishing the presence of classes with a given set of observation (for the unsupervised form of learning) or in a situation where there are various classes and the aim is to classify new observation into any of the already existing classes (for the supervised form of learning) [10, 11]. The classification task in this study employed the supervised form of learning.

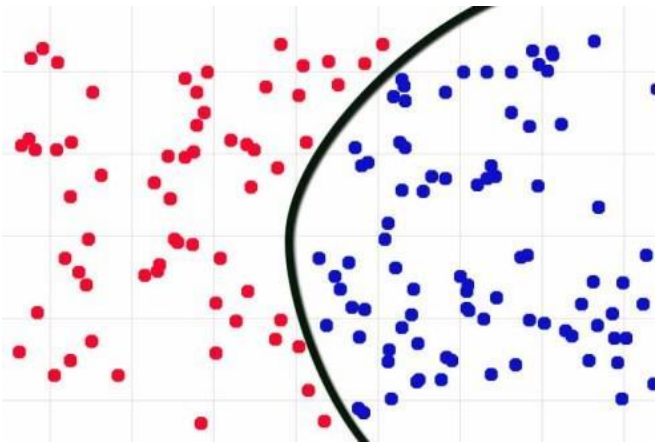

Figure 1. Data Classification

3.1 Naïve Bayes

A classifier mainly aims at performing an accurate prediction of class values considering each instance in a set of data. The NBC [12, 13] is a supervised classification technique which depends on the Bayes' Theorem to predict the class from the attributes of a dataset.

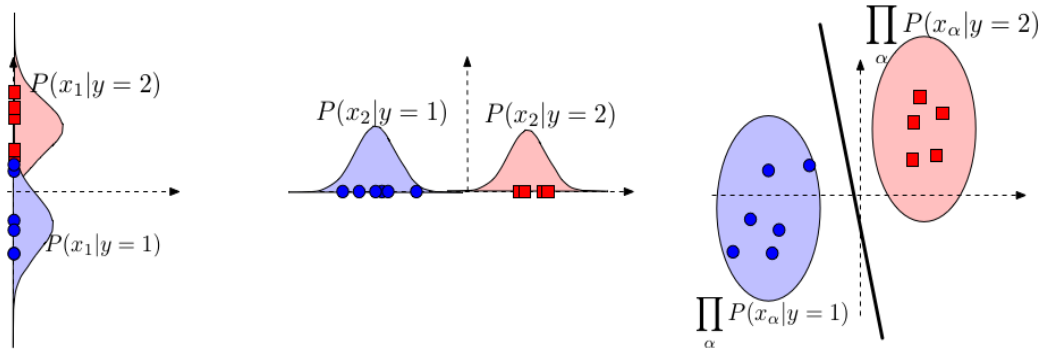

Figure 2. Bayes Theorem

3.2 Hoeffding Tree (HT)

This is a recent development [14-18] in data classification which performs prediction by selecting the majority class at each leaf. The incorporation of Naive Bayes models at the tree leaves can improve the predictive accuracy of HT. However, the naive Bayes method has been outlined previously to initially perform better than the standard HT but is later overtaken. Hence, a hybrid adaptive approach called Hoeffding Naive Bayes Tree (hnbt) which performs better than the component prediction methods for both complex and simple concepts has been proposed. This concept of this method based on executing a naive Bayes prediction on each training feature, then, comparing the prediction performance with the majority class [19-25]. The number of times the naïve Bayes makes a correct prediction of the true class is noted (by taking counts) compared to the majority class. When predicting a test case, the leaf can only output a naïve Bayes prediction when its overall accuracy is more than the majority class, else, it will output a majority class prediction [26-28].

Algorithm (1) Hoeffding tree induction algorithm

1: HT be a tree with a single leaf (the root)

2: for all training examples do

3: Sort example into leaf l using HT

4: Update sufficient statistics in l

5: Increment nil, the number of examples seen at l

6: if NL mod Nmin = 0 and examples seen at l not all of same Class then

7: Compute l (XL) for each attribute

8: Let Ax be attribute with highest l

4. DATA SET

Wisconsin Breast Cancer Database (WBC) was used in this study. This dataset was used because it is widely used in many researches. In a general sense, test results from this paper can be compared with those previous results. Wisconsin Breast Cancer Database (WDC) dataset was collected from the University of Wisconsin Hospitals, Madison by Dr. William H.Walberg in 1991. The dataset includes 699 instances and 10 patient features, which include an instance identifier, tumor information, classes, etc. There are 16 instances

that contain a single missing attribute; so, these records are not considered. After deleting the missing information data, there are 683 instances, 65.01% (444) of them are benign cases, and 34.99% (239) of them are malignant cases. The statistical summary of the 9 input features is given in Table 1.

Table 1. Wisconsin Breast Cancer Database (WBC)

Number of attributes	Description of attributes	Range	Mean	Standard Deviation
1	The thickness of the clump	1.0 -10.0	4.440	2.820
2	Cell size uniformity	1.0 -10.0	3.150	3.070
3	Cell shape uniformity	1.0 -10.0	3.220	2.990
4	Marginal adhesion	1.0 -10.0	2.830	2.860
5	Size of single epithelial cell	1.0 -10.0	3.230	2.220
6	Bare nuclei	1.0 -10.0	3.540	3.640
7	Bland chromatin	1.0 -10.0	3.450	2.450
8	Normal nucleoli	1.0 -10.0	2.870	3.050
9	Mitoses	1.0 -10.0	1.600	1.730

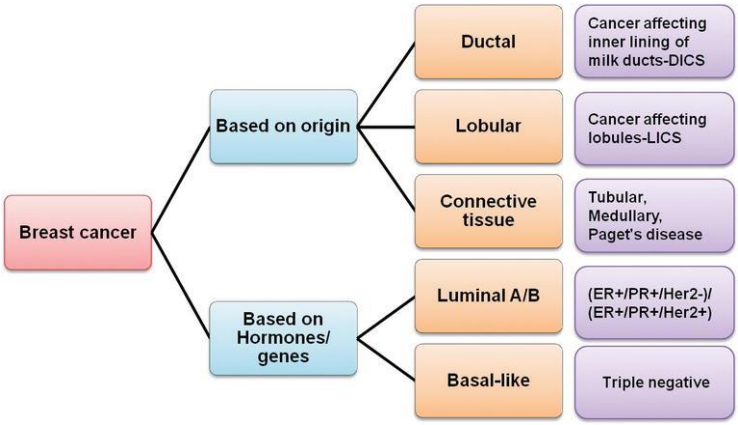

Figure 3. Types of Breast Cancer

5. IMPLEMENTATION

Step-1: The Breast Cancer dataset in CSV is computed as the input.

Step-2: Execute log2 normalization if the size of the dataset is >25 MB, else, resort to Min-Max normalization.

Step-3: Partition the dataset into two (testing and training sets). Cross-validation was used in this study.

Step-4: Differentiate the training dataset according to the class values.

Step-5: Compute the mean and standard values for each data case according to the class values.

Step-6: Choose the Hoeffiding tree for the first phase of classification.

Step-7: First evaluation.

Step-8: Forward the miss-classified sample for naïve Bayes.

Step-9: Second evaluation.

Step-10: Build the final model.

Step-11: Compare the class data of test dataset to determine the prediction accuracy. Evaluate the computed accuracy based on the scale of 0 to 100 %.

Step-12: Generate the predictions using this model.

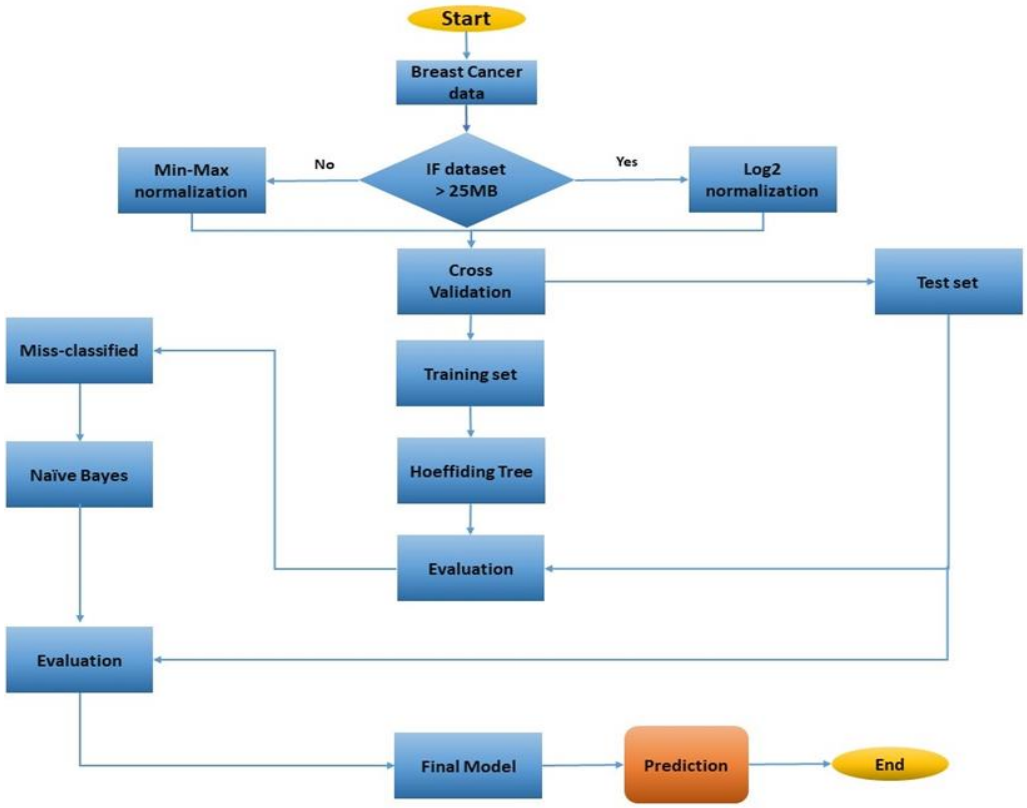

Figure 4. The Proposed Method

6. RESULTS AND COMPARISONS

The proposed model was implemented on the WBC and from the achieved results, there was an increase in the accuracy of the Hoeffiding naïve method to 95.9943% when Step-8 was involved and about 88.33% when Step-8 was not involved. In Step-8, the misclassified sample is forwarded for naïve Bayes and evaluated in the second phase. Using the same dataset, the inbuilt NBC of Matlab recorded an accuracy of 79.09%. The merged analysis of the suggested method in comparison to the other techniques for accuracy is presented in Figure 8. From the results, the suggested method exhibited a lower rate of accuracy loss in the NBC owing to its assumption of conditional independence. Hence, the model presented in this work can improve the performance of NBC. The performance of the proposed approach highlighted the feasibility of using Hoeffiding classifier tree with naïve Bayes on breast cancer dataset. The results shown in Figures 5-8 showed that the suggested method can reduce the rate of accuracy loss in the classification of breast cancer even when conditional independence is assumed. The performance of NBC was also improved by the proposed model in this study.

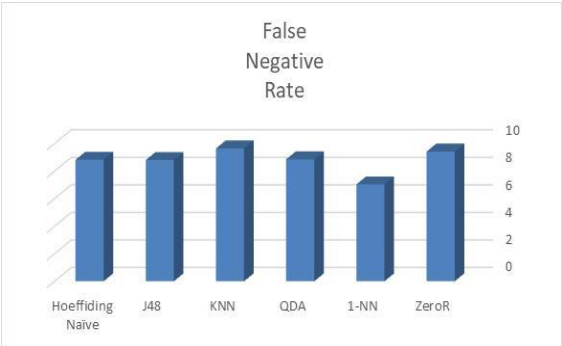

Figure 5. False Negative Rate comparison

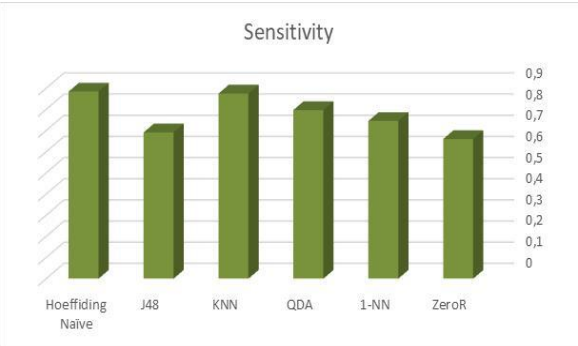

Figure 6. Sensitivity comparison

Figure 7. Specificity comparison

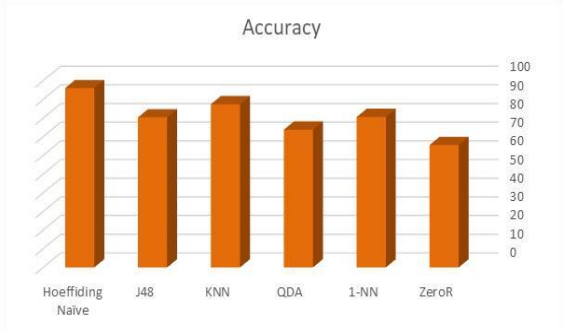

Figure 8. Accuracy comparison

7. CONCLUSION AND RECOMMENDATION

The accuracy of NBC is normally affected by the initial values. However, this assumption makes the probabilities estimation easier. In this study, the adopted separation technique improved the classifiers' accuracy using the Naive Bayes technique. From the achieved results, the employed approach recorded a better prediction accuracy when compared to the traditional NBC in Matlab. Hence, the accuracy of NBC can be improved by the assumption of conditional independence. The proposed approach can classify input breast cancer data into benign, non-benign (malignant), or normal with a good level of specificity, accuracy, sensitivity, and low rate of false negatives. The major derivative from this study is that it can help medical experts in the diagnosis of breast cancer since early cancer detection improves the chances of survival due to the administration of the appropriate treatment.

ABBREVIATIONS

NBC: Naïve Bayes Classifier

REFERENCES

[1] A. B. Ashraf, S. C. Gavenonis, D. Daye, C. Mies, M. A. Rosen, and D. Kontos, "A multichannel markov random field framework for tumor segmentation with an application to classification of gene expression-based breast cancer recurrence risk," *IEEE transactions on medical imaging*, vol. 32, pp. 637-648, 2012.

[2] Y. Tang, W. Pan, X. Qiu, and Y. Xu, "*The identification of fuzzy weighted classification system incorporated with Fuzzy Naive Bayes from data*," in IEEE International Conference on Systems, Man and Cybernetics, 2002, p. 6 pp. vol. 5.

[3] S. Reis, P. Gazinska, J. H. Hipwell, T. Mertzanidou, K. Naidoo, N. Williams, *et al.*, "Automated classification of breast cancer stroma maturity from histological images," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 2344-2352, 2017.

[4] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensarİ, "Breast cancer classification using machine learning," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018, pp. 1-4.

[5] A. Mert, N. Kilic, and A. Akan, "*Breast cancer classification by using support vector machines with reduced dimension*," in Proceedings ELMAR-2011, 2011, pp. 37-40.

[6] G. Attarodi, A. Tareh, N. J. Dabanloo, and A. Adeliansedehi, "Classification of congenital heart disease by SVM-MFCC using phonocardiograph," in *2017 Computing in Cardiology (CinC)*, 2017, pp. 1-4.

[7] R. Kavitha and E. Kannan, "*An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining*," in 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), 2016, pp. 1-5.

[8] A. Shenfield and S. Rostami, "*A multi objective approach to evolving artificial neural networks for coronary heart disease classification*," in 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015, pp. 1-8.

[9] M. D. Hossain, F. Yang, M. Abedin, and A. S. Mohan, "*Time reversal microwave imaging for the localization and classification of early stage breast cancer*," in Asia-Pacific Microwave Conference 2011, 2011, pp. 477-480.

[10] A. H. Ali and M. Z. Abdullah, "*Recent trends in distributed online stream processing platform for big data: Survey*," in 2018 1st Annual International Conference on Information and Sciences (AiCIS), 2018, pp. 140-145.

[11] A.-H. A. Salih, A. H. Ali, and N. Y. Hashim, "Jaya: An Evolutionary Optimization Technique for Obtaining the Optimal Dthr Value of Evolving Clustering Method (ECM)."

[12] M. A. Mohammed, R. A. Hasan, M. A. Ahmed, N. Tapus, M. A. Shanan, M. K. Khaleel, *et al.*, "*A Focal load balancer based algorithm for task assignment in cloud environment*," in 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2018, pp. 1-4.

[13] M. A. Mohammed and N. ŢĂPUŞ, "A novel approach of reducing energy consumption by utilizing enthalpy in mobile cloud computing," *Studies in Informatics and Control*, vol. 26, pp. 425-434, 2017.
[14] X. Song, H. He, S. Niu, and J. Gao, "*A data streams analysis strategy based on hoeffding tree with concept drift on Hadoop system*," in 2016 International Conference on Advanced Cloud and Big Data (CBD), 2016, pp. 45-48.
[15] Hammood, O. A., Nizam, N., Nafaa, M., & Hammood, W. A. (2019). "RESP: Relay Suitability-based Routing Protocol for Video Streaming in Vehicular Ad Hoc Networks". *International Journal of Computers, Communications & Control*, 14(1).
[16] Hasan, R. A., & Mohammed, M. N. (2017). "A krill herd behaviour inspired load balancing of tasks in cloud computing". *Studies in Informatics and Control*, 26(4), 413-424.
[17] Hasan, R. A., Mohammed, M. N., Ameedeen, M. A. B., & Khalaf, E. T. (2018). "Dynamic Load Balancing Model Based on Server Status (DLBS) for Green Computing". *Advanced Science Letters*, 24(10), 7777-7782.
[18] M. A. Mohammed and R. A. Hasan, "*Particle swarm optimization for facility layout problems FLP—A comprehensive study*," in 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), 2017, pp. 93-99.
[19] R. A. Hasan, I. Alhayali, A. Royida, N. D. Zaki, and A. H. Ali, "An adaptive clustering and classification algorithm for Twitter data streaming in Apache Spark," *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 17, 2019.
[20] M. A. Mohammed, Z. H. Salih, N. Ţăpuş, and R. A. K. Hasan, "*Security and accountability for sharing the data stored in the cloud*," in 2016 15th RoEduNet Conference: Networking in Education and Research, 2016, pp. 1-5.
[21] Munef.A.Ahmed , R. A. Hasan .Ahmed.H.A ,and M.A.Mohammed, "Using Machine Learning for the Classification of the Modern Arabic Poetry", *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 17.No.5.
[22] Nada.Q.M ,M.Sh.Ahmed ,and M.A.Mohammed, "*Comparative Analysis Between Solar And Wind Turbine Energy Sources In IoT Based On Economical And Efficiency Considerations*", Paper presented at the 2019 22st 22nd International Conference on Control Systems and Computer Science (CSCS22).
[23] R. A. Hasan, M. A. Mohammed, N. Ţăpuş, and O. A. Hammood, "*A comprehensive study: Ant Colony Optimization (ACO) for facility layout problem*," in 2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet), 2017, pp. 1-8.
[24] R. A. Hasan, M. A. Mohammed, Z. H. Salih, M. A. B. Ameedeen, N. Ţăpuş, and M. N. Mohammed, "HSO: A Hybrid Swarm Optimization Algorithm for Reducing Energy Consumption in the Cloudlets," *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 16, pp. 2144-2154, 2018.
[25] Z. H. Salih, G. T. Hasan, and M. A. Mohammed, "*Investigate and analyze the levels of electromagnetic radiations emitted from underground power cables extended in modern cities*," in 2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2017.
[26] Z. H. Salih, M. A. Mohammed, "*Study the Effect of Integrating the Solar Energy Source on Stability of Electrical Distribution System*", Paper presented at the 2019 22st 22nd International Conference on Control Systems and Computer Science (CSCS22).
[27] Adeeb Salh, Lukman Audah, Nor S. M. Shah, Shipun A. Hamzah, "Pilot reuse sequences for TDD in downlink multi-cells to improve data rates", *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, vol. 17.No.5, pp.2161~2168 2019.
[28] Tejaswini R Murgod, S Meenakshi Sundaram, "Survey on underwater optical wireless communication: perspectives and challenges" *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, Vol. 13,