TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 3, June 2020, pp. 1688~1694
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i3.13764

A new model for large dataset dimensionality reduction based on teaching learning-based optimization and logistic regression

Hind Raad Ibraheem[1], Zahraa Faiz Hussain[2], Sura Mazin Ali[3], Mohammad Aljanabi[4], Mostafa Abdulghafoor Mohammed[5], Tole Sutikno[6]
[1,2]Department of Computer Science, Al Salam University College, Iraq
[3]College of Political Science, Mustansiriyah University, Iraq
[4]Department of Computer Science, College of Education, Aliraqia University, Iraq
[5]Faculty of Automatic Control and Computers, University Polytechnic of Bucharest, Romania
[6]Department of Electrical Engineering, Universitas Ahmad Dahlan, Indonesia
[6]Embedded System and Power Electronics Research Group, Indonesia

Article Info

Article history:

Received Jul 30, 2019
Revised Feb 28, 2020
Accepted Mar 24, 2020

Keywords:

Feature selection
FSS
IDS
NTLBO
Subset
TLBO

ABSTRACT

One of the human diseases with a high rate of mortality each year is breast cancer (BC). Among all the forms of cancer, BC is the commonest cause of death among women globally. Some of the effective ways of data classification are data mining and classification methods. These methods are particularly efficient in the medical field due to the presence of irrelevant and redundant attributes in medical datasets. Such redundant attributes are not needed to obtain an accurate estimation of disease diagnosis. Teaching learning-based optimization (TLBO) is a new metaheuristic that has been successfully applied to several intractable optimization problems in recent years. This paper presents the use of a multi-objective TLBO algorithm for the selection of feature subsets in automatic BC diagnosis. For the classification task in this work, the logistic regression (LR) method was deployed. From the results, the projected method produced better BC dataset classification accuracy (classified into malignant and benign). This result showed that the projected TLBO is an efficient features optimization technique for sustaining data-based decision-making systems.

Corresponding Author:

Mohammad Aljanabi,
Department of Computer Science,
Aliraqia Univeristy, Alsaydya, Baghdad, Iraq
Email: mohammad.cs88@gmail.com

1. INTRODUCTION

Breast cancer (BC) is the most common cancer among women around the world. About 25% of all new cancer cases are diagnosed as BC in women as stated by the American Cancer Society (ACS). One woman dies from BC every minute and more than 1400 women die every day from BC [1]. Cancer results from the rapid and uncontrollable division of cells which results in the formation of extra tissue mass called tumors in the body [2]. Such tumors are either cancerous or non-cancerous. The spread of the malignant tumors is faster as they spread rapidly to cause harm to the neighboring tissues. Cancer is commonly named based on the affected body part or where it started. Hence, BC is a form of malignant tumor which results from the uncontrolled division of breast cells. Among the major signs of BC are changed and increase in the increase normal size and shape of the breast, pain in the breast, inflammation of the affected or all parts of the breast, varying breast skin colors, presence of a lump beneath the arm area, etc. [2-4].

The World Health Organization (WHO) [5] reported that there are around 1.2 million cases of BC diagnosed in women every year. One out of 8 women in the USA is living with BC. The identification of BC can be done manually by the physician, but it is a difficult task due to the need to remember all the required information for each case, giving rise to low identification accuracy. Breast cancer-related mortality rate can be reduced through disease detection at the early stage [6-9]. Several conventional BC detection methods exist but the higher accuracy of machine learning (ML) classifiers is making them more useful recently. As such, there are several ML methods for early cancer detection and also checking for its relapse. Among these ML methods are artificial neural network (ANN), support vector machine (SVM), Naive Bayes, relevance vector machine, decision trees, K-means, K-nearest neighbor, random forests, etc.

The use of ML-based classification schemes is gaining attention in the medical field as they can help both skilled and unskilled experts in reducing possible errors and accurately providing the required medical data for diagnosis within a short time. However, the high dimensionality of the dataset represents one of the major limitations for effective use of ML classifiers. The important criteria which must be considered for effective ML-based classification are the data quality and a careful feature selection. Feature selection (FS) is the process of extracting a subset of relevant features from the original dataset [10-12]. It involves the use of FS algorithms to filter out irrelevant and redundant data features from the original dataset to prevent over-fitting [6, 13] and improve the classification accuracy of the model. Feature selection also reduces the classification models' complexity in time and space domains [14-18]. The main idea of this paper is to employ the TLBO-based algorithm for features subset selection in BC diagnosis. A recent metaheuristic, teaching-learning-based optimization (TLBO), has been reported to be an efficient optimization tool that is inspired by the knowledge passing mechanisms of teachers and learners in a classroom [19-22]. It has been applied to several well-known combinatorial optimization problems, producing good results [23-26]. The following sections discussed the novel multi-objective TLBO optimization algorithm for attaining better feature selection accuracy.

2. RELATED WORKS

Chuang et al. [27] proposed the catfish binary PSO (CatfishBPSO) algorithm. In this algorithm, few features are selected via the introduction of new catfish (particles) into the solution space to achieve 2 major advantages: i) reduced computation time, and ii) higher classification accuracy using the k-NN algorithm. It was applied and compared to 10 classification problems taken from the literature. Experimental results show that CatfishBPSO simplifies the feature selection process effectively, and either obtains higher classification accuracy or uses fewer features than other feature selection methods.

Bahassine et al. [28] proposed a novel feature selection method for Arabic text classification. The method uses an enhanced Chi-square method to improve the classification accuracy. The combination of the proposed Arabic text classification model with SVM classifier significantly enhanced the performance of the model as it achieved the best F-measure value of 90.50% using 900 features.

Sridev and Murugan [29] developed a feature selection technique for medical analysis of BC and compared with several classification algorithms. The objective of the presented algorithm is to select a minimum number of features to provide high classification accuracy. They reduced the feature vectors to 222 for both diagnosis and prognosis BC data sets using rough sets and correlation techniques.

Agrawal et al. [30] proposed a feature selection system for classification of cervical cancer CT images using artificial bee colony algorithm (ABC) and k-NN classifier; and artificial bee colony algorithm with SVM classifier. The result shows that the combination of ABC with SVM gave better performance compared to the combination of ABC with K-NN classifier.

Allam et al. [31] mentioned the importance of automatic medical disease diagnosis to handle the problems efficiently in the early stages. The study also discussed various imaging modalities for capturing the images, feature extraction methods for collecting the required attributes, and feature selection techniques for necessary features like texture, and color.

Chen et al. [32] proposed a coarse-grained parallel genetic algorithm (CGPGA) for optimizing the features in the dataset and constraints for SVM. They also proposed a new fitness function which is composed of classification accuracy, number of selected features, and the number of support vectors to optimize generalization errors. The results showed that the performance was ten times for the accuracy, size of subset features, number of support vectors, and the practice time.

Shahbeig et al. [33] proposed a mutated fuzzy adaptive PSO combined with TLBO algorithm for finding the most relevant and least set of genes in BC microarray data. The need to reduce the number of genes and increase the performance led to the use of a multi-objective for optimization problems. The result showed the model to achieve an accuracy of 91.88% with SVM classifier.

Jung et al. [34] presented a method to obtain additional numerical parameters from BC image data analysis using many neural network algorithms to explains how to get the highest number of numerical parameters from data of BC image and made a comparison between these algorithms to find the best classification between benign and malignant.

Thein et al. [35] proposed the training of ANN using the island-based model for distinguishing different types of BC with better accuracy and reduced training time on Wisconsin diagnostic and prognostic breast cancer. They proposed 2 different migration topologies with random-random policy and later compared their results. From the results, the torus topology needed more training time compared to the random topology although it presented similar solution performance to the random topology.

Thawkar et al. [36] explored the use of Firefly algorithm to select a subset of features. Artificial neural network and support vector machine classifiers are employed to evaluate fitness of the selected features. Features selected by Firefly algorithm are used to classify masses into benign and malignant, using artificial neural network and support vector machine classifiers. Results show that Firefly algorithm with artificial neural network is superior to Firefly algorithm with support vector machine.

Sasikala et al. [37] proposed a novel shapely value embedded genetic algorithm (SVEGA). The method selects the genes that can maximize the capability to discriminate between different classes. Thus, the dimensionality of data features is reduced and the classification accuracy rate is improved. The number of features reduced from 24,481 to minimum of 6 features.

Sridev and Murugan [29] presented a modified correlation rough set feature selection (MCRSFS). It is composed of two feature reduction algorithms. Rough set quick reduct algorithm is applied at first to obtain the minimal feature subset. Then the second algorithm correlation feature selection (CFS) is used to do further reduction in minimal feature subset. The MCRSFS achieved highest classification accuracy compared to other feature selection methods.

Durgalakshmi and Vijayakumar [38] proposed an efficient method for breast cancer detection based on Wisconsin prognostic breast cancer (WPBC) data set. The correlation matrix method is used for feature selection which remove the insignificant features from the massive amount of dataset, followed with the classification algorithms such as support vector classification, logistic regression and random forest was deployed. The proposed method improves the accuracy [3, 4, 7-12, 15-18, 20-22, 24-26, 39, 40].

3. THE PROPOSED ALGORITHM AND MACHINE LEARNING TEQNIQUES

Feature selection is a process of optimizing individuals (records) by extracting the best subset of attributes from such records. During feature selection, fitness is assessed for each record in every generation while new records are generated to establish the population of the subsequent generations. After many generations, the components of the successive generations are better compared to the initial population. The construction of a dataset of optimal features requires a novel algorithm. The technique proposed in this study has two phases. The first phase involves the use of an optimization scheme to select the best set of features for the classification process. Then, in the second phase, the classification models are generated for the evaluation of the proposed intended scheme on BC dataset in terms of its performance. In this study, a multi-objective TLBO optimization algorithm was modeled with two major objectives which are results accuracy maximization and minimization of the number of features such that there will be a set of solutions instead of just one solution. The projected model is comprised of the TLBO and LR for subset feature selection. LR-based classification involves the estimation of an events' occurrence probability based on the similarity of given data points. The LR uses sigmoid function to determine the events' occurrence probability. An event with an occurrence probability of >0.5 is predicted as 'occurred', else, as 'not occurred'. Figure 1 depicts the proposed model while Figure 2 depicts the TLBO algorithm. Three parameters inherit from genetic algorithm to represent and update the set of features, each set of features represented in a chromosome.

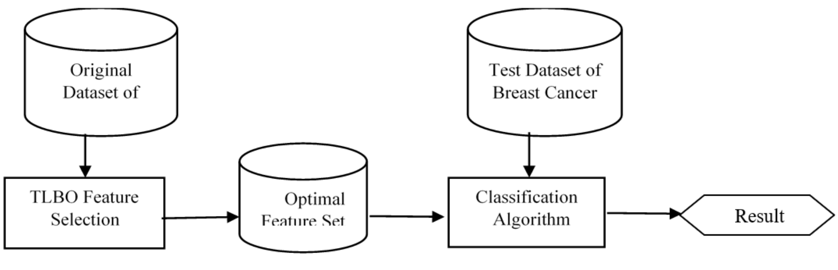

Figure 1. The projected feature selection model based on TLBO

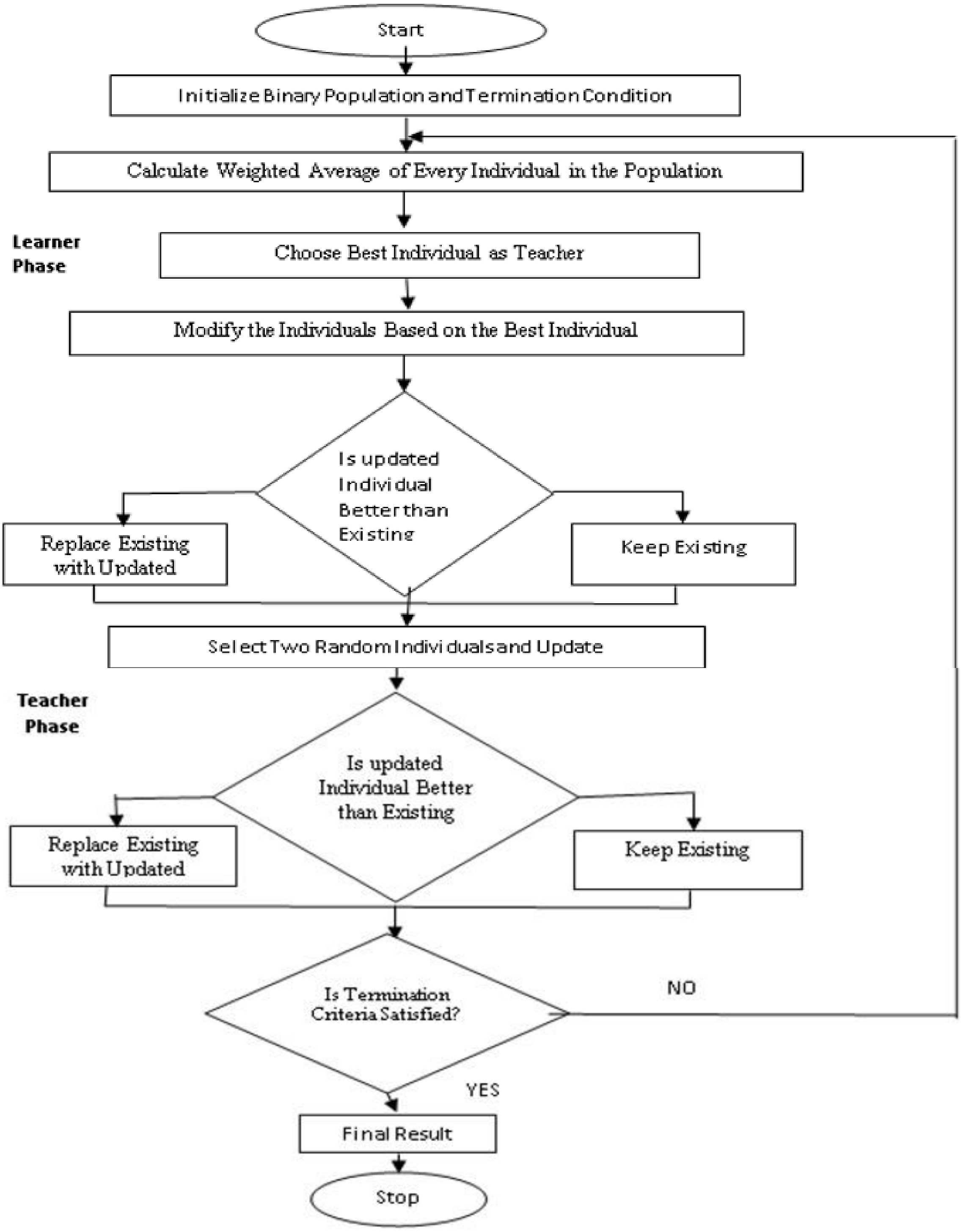

Figure 2. The TLBO Algorithm

Each gen in the chromos represent one feature, if the value of the gen 1 the feature is selected and 0 represent unselected feature, as shown in Figure 3. To updates the value of each gen (features) from selected to unselected or Vis versa, crossover between two chromosome and then mutation is used to get new subset of features (chromosome) as shown in Figure 4. The parameters used in this algorithm shown in Table 1.

1	0	1	0	0	0	0	0	1	0

Figure 3. Structure of the chromosome

Chromosome 1	1	1	0	0	0	1	0	1
Chromosome 2	1	1	1	0	0	0	0	1
Crossover								
New Chromosome	1	1	1	0	0	1	0	1
Mutation								
New Chromosome	1	1	1	0	1	0	0	1

Figure 4. Crossover and mutation operations

Table 1. The Parameters of algorithm

Parameter	Value
Population size	20
Number of generation	40
Croddover type	Half-uniform
Mutation type	Bit-flip

4. RESULTS AND DISCUSSION

The results of the proposed algorithm are shown in Table 2. The results showed each set of features and the accuracy. Statistical tests were performed on the result to verify the results. Table 3 showed the confusion matrix. Based on the confusion matrix, we can calculate the detection rate using (1).

$$\text{Detection rate} = \frac{TP}{TP+FP} \quad (1)$$

$$= \frac{444}{444+0} = 1.00$$

The comparison of the proposed algorithm in this study with PCA and Fine tree built in MATLAB as shown in Table 4 revealed the superiority of the proposed model over the benchmarked schemes.

Table 2. Results for each set of features

No. Features	Accuracy
1	0.97
2	0.99
3	1.00

Table 3. Confusion matrix

	Predicated malignant	Predicated Benign
Actual malignant	444	0
Actual Benign	0	239

Table 4. Comparision table

No. features	Our Model	PCA+ Fine Tree
1	0.97	62.1
2	0.99	95.6
3	1.00	96.5

5. CONCLUSION

This study presented a multi-objective TLBO algorithm for solving optimal feature selection problems. Accuracy is most important in the field of medical diagnosis to diagnose the patient's disease. The objective of this algorithm is to select minimum number of features providing high classification accuracy. The results from the investigations showed that the proposed TLBO-based feature optimization system can improve the accuracy of classifiers compared to previous methods on BC dataset. The classification model was trained with a reduced number of features. Our future studies will focus on working with a new hybrid algorithm.

REFERENCES

[1] S. Wang, R. V. Rao, P. Chen, Y. Zhang, A. Liu, and L. Wei, "Abnormal breast detection in mammogram images by feed-forward neural network trained by Jaya algorithm," *Fundamenta Informaticae,* vol. 151, No.1-4, pp. 191-211, 2017.

[2] F. Gao, T. Wu, J. Li, B. Zheng, L. Ruan, D. Shang, et al., "SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis," *Computerized Medical Imaging and Graphics*, vol. 70, pp. 53-62, December 2018.

[3] M. A. Ahmed, R. A. Hasan, A. H. Ali, and M. A. Mohammed, "The classification of the modern arabic poetry using machine learning," *TELKOMNIKA Telecommunication Computing Electronics and Control,* vol. 17, No.5, pp. 2667-2674, 2019.

[4] A. H. Ali and Mahmood Zaki Abdullah, "A Survey on Vertical and Horizontal Scaling Platforms for Big Data Analytics," *International Journal of Integrated Engineering,* vol. 11, No. 6, pp. 138-150, 2019.

[5] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. A. Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," *Computers & Electrical Engineering,* vol. 70, pp. 871-882, August 2018.

[6] J. H. Wang, J. H. Jiang, and R. Q. Yu, "Robust back propagation algorithm as a chemometric tool to prevent the overfitting to outliers," *Chemometrics and intelligent laboratory systems,* vol. 34, No. 1, pp. 109-115, Aug. 1996.

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

[8] A. H. Ali and M. Z. Abdullah, "A novel approach for big data classification based on hybrid parallel dimensionality reduction using spark cluster," *Computer Science,* vol. 20, no.4, pp. 413-431, 2019.

[9] M. A. H. Ali, "An Efficient Model for Data Classification Based on SVM Grid Parameter Optimization and PSO Feature Weight Selection," *International Journal of Integrated Engineering,* vol. 12, no. 1, pp.1-12, 2018.

[10] N. Q. Mohammed, M. S. Ahmed, M. A. Mohammed, O. A. Hammood, H. A. N. Alshara, and A. A. Kamil, "Comparative Analysis between Solar and Wind Turbine Energy Sources in IoT Based on Economical and Efficiency Considerations," *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*, pp. 448-452, 2019.

[11] 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," *2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp.1-4, 2017.

[12] Z. H. Salih, G. T. Hasan, M. A. Mohammed, M. A. S. Klib, A. H. Ali, and R. A. Ibrahim, "Study the Effect of Integrating the Solar Energy Source on Stability of Electrical Distribution System," *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*, pp. 443-447, 2019.

[13] K. Huber-Keener, "Beyond BRCA: Cancer Risk Assessment in the Era of Panel Genetic Testing," *Michigan Medicine-University of Michigan,* pp. 1-53, 2018.

[14] S. Vanaja and K. Ramesh Kumar, "Analysis of feature selection algorithms on classification: a survey," *International Journal of Computer Applications,* vol. 96, No. 17, pp. 28-35, June 2014.

[15] 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, No. 6, pp. 3086-3099, December 2019.

[16] 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, No. 5, pp. 2144-2154, October 2018.

[17] R. A. Hasan, M. A. Mohammed, N. Ţăpuş, and O. A. Hammood, "A comprehensive study: Ant Colony Optimization (ACO) for facility layout problem," *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, pp. 1-8, 2017.

[18] Z. F. Hussain, H. R. Ibraheem, M. Alsajri, A. Hussein Ali, M. A. Ismail, S. Kasim, *et al.*, "A new model for iris data set classification based on linear support vector machine parameter's optimization," *International Journal of Electrical & Computer Engineering (2088-8708),* vol. 10, No. 1, pp. 1079-1084, February 2020.

[19] R. V. Rao, V. J. Savsani, and D. Vakharia, "Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems," *Computer-Aided Design,* vol. 43, No. 3, pp. 303-315, March 2011.

[20] 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),* pp. 93-99, 2017.

[21] 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),* pp. 1-4, 2018.

[22] M. A. Mohammed, A. A. Kamil, R. A. Hasan, and N. Tapus, "An Effective Context Sensitive Offloading System for Mobile Cloud Environments using Support Value-based Classification," *Scalable Computing: Practice and Experience,* vol. 20, No. 4, pp. 687-698, December 2019.

[23] T. Dokeroglu, "Hybrid teaching–learning-based optimization algorithms for the quadratic assignment problem," *Computers & Industrial Engineering,* vol. 85, pp. 86-101, July 2015.

[24] M. A. Mohammed, I. A. Mohammed, R. A. Hasan, N. Ţăpuş, A. H. Ali, and O. A. Hammood, "Green Energy Sources: Issues and Challenges," in *2019 18th RoEduNet Conference: Networking in Education and Research (RoEduNet),* pp. 1-8, 2019.

[25] 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,* pp. 1-5, 2016.

[26] 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, no. 4, pp. 425-434, December 2017.

[27] L.-Y. Chuang, S.-W. Tsai, and C.-H. Yang, "Improved binary particle swarm optimization using catfish effect for feature selection," *Expert Systems with Applications,* vol. 38, No. 10, pp. 12699-12707, September 2011.

[28] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University-Computer and Information Sciences,* vol. 32, no. 2, pp. 225-231, February 2018.

[29] T. Sridevi and A. Murugan, "A novel feature selection method for effective breast cancer diagnosis and prognosis," *International Journal of Computer Applications,* vol. 88, No. 11, pp. 28-33, January 2014.

[30] V. Agrawal and S. Chandra, "Feature selection using Artificial Bee Colony algorithm for medical image classification," in *2015 Eighth International Conference on Contemporary Computing (IC3)*, pp. 171-176, 2015.

[31] M. Allam and M. Nandhini, "A Study on Optimization Techniques in Feature Selection for Medical Image Analysis," *International Journal on Computer Science and Engineering (IJCSE),* Vo. 9, No. 3, pp. 75-82, March 2017.

[32] Z. Chen, T. Lin, N. Tang, and X. Xia, "A parallel genetic algorithm based feature selection and parameter optimization for support vector machine," *Scientific Programming,* vol. 2016, No. 2, pp. 1-10, January 2016.

[33] S. Shahbeig, M. S. Helfroush, and A. Rahideh, "A fuzzy multi-objective hybrid Tlbo-Pso approach to select the associated genes with breast cancer," *Signal Processing,* vol. 131, pp. 58-65, February 2017.

[34] I.-S. Jung, D. Thapa, and G.-N. Wang, "Neural network based algorithms for diagnosis and classification of breast cancer tumor," in *International Conference on Computational and Information Science*, pp. 107-114, 2005.

[35] H. T. T. Thein and K. M. M. Tun, "An approach for breast cancer diagnosis classification using neural network," *Advanced Computing: An International Journal (ACIJ),* Vol. 6, no. 1, pp. 1-11, January 2015.

[36] S. Thawkar and R. Ingolikar, "Classification of Masses in Digital Mammograms Using Firefly based Optimization," *International Journal of Image, Graphics & Signal Processing,* vol. 10, no. 2, pp. 25-33, February 2018.

[37] S. Sasikala, S. A. alias Balamurugan, and S. Geetha, "A novel feature selection technique for improved survivability diagnosis of breast cancer," *Procedia Computer Science,* vol. 50, pp. 16-23, 2015.

[38] Durgalakshmi B, Vijayakumar V, "Impact of Dimensionality Reduction and Classification in Breast Cancer," *International Journal of Innovative Technology and Exploring Engineering*. vol. 8, no. 8, pp. 2599-2603, June 2019.

[39] S. B. Meskina, "On the effect of data reduction on classification accuracy," *2013 3rd International Conference on Information Technology and e-Services (ICITeS)*, pp. 1-7, 2013.

[40] S. A.-B. Salman, A.-H. A. Salih, A. H. Ali, M. K. Khaleel, and M. A. Mohammed, "A New Model for Iris Classification Based on Naïve Bayes Grid Parameters Optimization," *International Journal of Sciences: Basic and Applied Research (IJSBAR),* vol. 40, No. 2, pp. 150-155, August 2018.