The Use of the Binary Bat Algorithm in Improving the Accuracy of Breast Cancer Diagnosis

Reyhaneh Yaghoubzadeh[1,*], Seyed Reza Kamel[1], Hossein Barzegar[2], Bahareh Moshajeri San'ati[2]

[1] Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran
[2] Department of Computer Sciences, Islamic Azad University, Neyshabur Branch, Neyshabur, Iran
*Corresponding author: Reyhaneh Yaghoubzadeh, Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran. Tel: +989154133477; Fax: +985137251857; E-mail: yaghobzadehr@gmail.com

Introduction: The early diagnosis of breast cancer as prevalent cancer among women, is a necessity in the research on cancers since it could simplify the clinical management of other patients. The importance of the classification of breast cancer patients into high- or low-risk groups has led research groups in the biomedical and informatics departments to evaluate and use computer techniques such as data mining. To date, various methods have been used for breast cancer diagnosis which has shown unfavorable accuracy due to issues such as computational complexities and prolonged implementation.

Methods: The present study aimed to apply the feature selection method based on the binary bat algorithm (BBA) to increase the accuracy of the breast cancer diagnosis. Feature selection is carried out to select the most important features from a dataset. We applied the naïve bayes (NB), support vector machine (SVM), and J48 algorithms in MATLAB software; based on the dataset obtained from Wisconsin to evaluate the accuracy, sensitivity, and diagnostic criteria of the proposed model.

Results: The BBA had 99.28%, 96.43%, and 92.86% accuracy in SVM, NB and J48 algorithms, respectively.

Conclusions: According to the results, the feature selection technique, along with the BBA and SVM, yielded the most accurate results regarding breast cancer detection.

INTRODUCTION

One of the most common cancers in the world is breast cancer as a result of the uncontrollable growth of tissues [1]. The disease is more common in high-income countries, but it is also increasing in low-income regions such as in Africa, Asia, and Latin America [2]. In the medical industry, the early diagnosis of breast cancer at the early stage, preventing its development to the other parts of the body, decreasing the treatment costs, and increasing the treatment quality has turned into a major goal. Recently, the development of data-based methods has led to the introduction of new and effective ways of diagnosing breast cancer. Data mining and expert systems have been not only employed in solving medical issues but have been extensively used for other industrial applications [3]. Some of the strong specialized and data-based methods are the artificial neural network, fuzzy systems, decision tree, support vector machine (SVM), Bayesian Network (BN), etc. Actually, by employing data mining techniques and data modeling, cancer patients with dangerous

conditions might be diagnosed [2]. For example, in recent studies, the precision and accuracy of sequential minimal optimization (SMO) and SVM algorithms in diagnosing breast cancer were evaluated [4-8]. The studies indicated that the SVM and SMO algorithms were superior to other algorithms in increasing the precision level of diagnosis. In a study carried out by Aljawad et al., two BN and SVM algorithms were evaluated with regards to the prediction of conservation status in patients with breast cancer [4]. The investigated criteria were precision, accuracy, and recall. The criterion of receiver operating characteristics (ROC) was also used. The experimental results of the simulations indicated that SVM outperformed BN and had more precision. Moreover, in the research conducted by Lakshmi et al., three data mining algorithms including naïve bayes (NB), RBF, and J48 were used for the early diagnosis of breast cancer, and the superiority of these three algorithms were proved [9]. The neural network algorithm was used to diagnose breast cancer, and the superiority of the algorithm was proved [1, 10, 11]. Furthermore, in the research carried out by Chaurasia et al., it was attempted to improve the precision of diagnosing breast cancer, using the statistical characteristics [12]. K-NN with three different distance metrics including euclidean, cosine, city-block, and their combination were used for classification. The extracted characteristics led to the classification of return on investment (ROI) into each of the classes of the breast tissues including density, fat, and glands. The results showed that the combined classification of K-NN indicated more precision and might be effectively used in categorizing the breast tissues. Conventional data mining techniques (SVM and ANN) have provided variable accuracy in the diagnosis of breast cancer due to complications such as the lengthy run time and computational complexity. Therefore, the combination or improvement of these techniques and the use of feature selection techniques could increase diagnostic accuracy and eliminate the defects of these methods. The present study aimed to propose a feature selection method inspired by bats' behaviors in nature, and the article includes a materials and methods, introduction of the proposed method, preferences and criteria for comparison, discussion and conclusion.

METHODS

Figure 1 depicts the algorithm proposed in the present study. We applied the bar binary algorithm [14] to select the valuable features for the detection of breast cancer. In addition, the SVM, J48 and NB algorithms were used to classify the feature selection data. As depicted in Figure 1, the method encompassed three main phases, which have been elaborated on in the following sections.
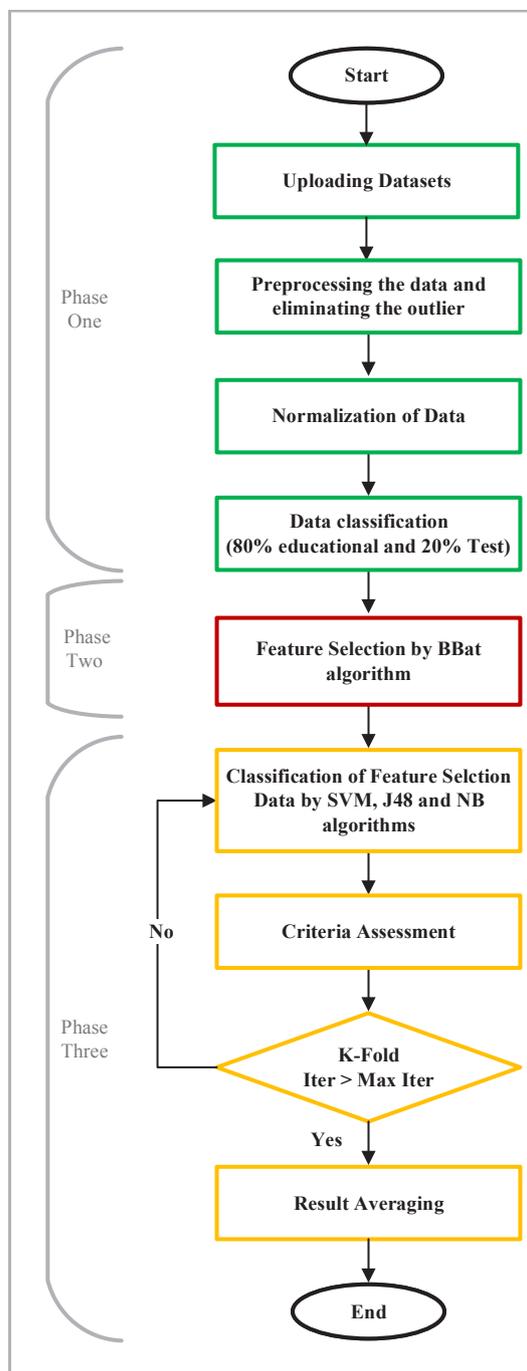

Figure 2: Three Phases of Proposed Method

Phase One

This phase consisted of three stages, including data uploading, preprocessing, and normalization.

Data uploading

At this stage, the breast cancer datasets were extracted from the repository university of California Irvine (UCI) [13] to perform the preprocessing and normalization of operations. In total, 699 records, 10 features, and one cancer class were extracted (class zero was related to healthy Individuals and class one was related to breast cancer patients) (Table 1) [5, 14].

Table 1: Input Features of Datasets

	Features
1	Sample code number: id number
2	Clump Thickness: 1 - 10
3	Uniformity of Cell Size: 1 - 10
4	Uniformity of Cell Shape: 1 - 10
5	Marginal Adhesion: 1 - 10
6	Single Epithelial Cell Size: 1 - 10
7	Bare Nuclei: 1 - 10
8	Bland Chromatin: 1 - 10
9	Normal Nucleoli: 1 - 10
10	Mitoses: 1 - 10

Preprocessing to eliminate the missing data

In the present study, the preprocessing stage and elimination of the missing data were carried out; using the nearest neighbor technique, which replaces the NaN values with the values of the nearest neighbor's column [15].

Normalization

In this phase, the repeated records were cleaned using the Min-Max normalization method. Afterward, the data were divided into the educational and experimental categories, and normalization was performed for each feature. In general, normalization refers to the classification of data within a specific range (e.g., 1, -1).

Phase Two

This phase consisted of two stages, including data selection and data classification.

Feature selection based on the binary bat algorithm

Bat algorithm

The bat optimization algorithm is inspired by the tracking of small bats in search of prey so that the small bats could hunt their prey in absolute darkness through the emission and reception of sound. The following ideal rules are applied to develop the algorithm [16]:

All the bats used sound reflection to detect the distance and spot the difference between food and the leading obstacles. Bats fly at the random speed of vi in the xi location, with the fixed frequency of f_{min}, various wavelengths of λ, and sound loudness of A to hunt the prey. In addition, they automatically regulate the distributed waves and their pulse rate based on the proximity of the prey.

$\in$ *[r (0, 1)]* depending on the proximity of their target

Since sound loudness may differ in various aspects, it is often considered to be within the range of Rmax (highest value)-Rmin (lowest value).

For each bat, X^t_i and the velocity of V^{t-1}_i were considered based on the mentioned rules, and the *ith* location in *t* repetition and f_i frequency were measured; using the following Equations:

Equation 1: $f_i = f_{min}(f_{max} - f_{min})B$

Equation 2: $V^t_i{}_{v_i{}^t=} V^{t-1}_i + (X^t_i - X_*)f_i$

Equation 3: $X^t_i{}_{X_i{}^t=} X^{t-1}_i + V^t_i$

Equation 4: $x_{new} = X_{old} + \epsilon A^t$

Where $\in$ *[β 1 ,0]* is a random vector with uniform distribution, and X shows the optimal current location (solution), which is found in each replication after the comparison of all the solutions among *n* bats. In addition, the *f* frequency is often considered to be within the range of *fmin*=(0), *fmax*=(100). In each replication, an answer is selected as optimal to find a location, and the new location of each bat is updated locally with a random step; using Equation 4, where $\varepsilon \in$ *[0,1]* is a random number, and $A^t = <A_i{}^t>$ represents the mean sound loudness of all the bats in this phase.

Binary bat algorithm for feature selection

A standard approach to feature selection is that each solution could be in two modes of zero and one, with zero indicating the unselected feature and one; showing the selected feature. At this stage, we applied

the binary bat algorithm (BBA) to select the important features in breast cancer diagnosis. In this algorithm, the decision variables could only be attributed to one of the zero or one modes. In the binary bat algorithm, each bat (feature) moves to the continuous values in the search space, which is modeled in the form of an n-dimensional Boolean network in case of the selection of the feature, where the bats (features) move across the corners of a hypercube. The location of the bats is exhibited by a binary vector since the issue is the selection or lack of selection of the features. The binary version of the bat algorithm uses a sigmoid function to limit the position of the new bat to the binary values only, as follows [16]:

Equation 5: $\delta(V_i^j) = \frac{1}{1+e^{v_i j}}$

Therefore, Equation 3 in the bat algorithm could be replaced by Equation 5.

Equation 6: $x_i^j = \begin{cases} 1 & if \quad S(V_i^i) > \sigma \\ 0 & othewise \end{cases}$

In these equations, $\sigma \sim (0, 1)$; therefore, Equation 6 could only provide the binary values for the specifications of each bat in the Boolean network; regarding the presence or absence of the features.

Data classification

At this stage, the training data were taught; using the SVM-BBA, NB-BBA, and J48-BBA algorithms and assessed based on the experimental data that were not used in the training phase.

Phase Three

This phase consisted of two steps, including determining the accuracy and using the K-fold technique.

Determining the accuracy of the test method

At this stage, the accuracy determined by the three algorithms used in the current research was calculated, assessed, and compared.

Validation of the K-fold cross technique

The efficiency of a model could be more accurately assessed by dividing the dataset into the educational and experimental subcategories. After the division, the prediction model was established with the education data, and its accuracy was evaluated; using the experimental data. In this technique, the main dataset was randomly placed in the k data subgroup, and the prediction model was developed each time with a k-1 subgroup and evaluated with the remaining subgroup. The education and test processes were also repeated k times, followed by averaging the results obtained from the procedure; therefore, the obtained results were considered to be favorable.

Preferences and Criteria for Comparison

In the previous section, we presented a method to increase the accuracy of breast cancer prognosis. In this section, the evaluation criteria of the proposed method have been assessed, followed by the simulation of the proposed method in MATLAB software [17]. The results obtained from the simulation were assessed based on the evaluation criteria and using tables.

Confusion matrix

In Table 2, the matrix shows the function of the classification algorithm based on the input dataset and various categories of the classification problem to evaluate the performance of the proposed method. Presented in Equations 7 to 12, the set of criteria is used as the most important evaluation criteria.

Equation 7: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Equation 8: $Sensitivity = \frac{TP}{TP+FN}$

Equation 9: $Specificity = \frac{TN}{TN+FP}$

Equation 10: $F - measure = \frac{2*Recall*Precision}{Precision+Recall}$

Table 2: Confusion Matrix

		Predicted Amounts	
		Patients	**Healthy Individuals**
Actual Amounts	**Patients**	True Positive (TP)	False Negative (FN)
	Healthy Individuals	False Positive (FP)	True Negative (TN)

Equation 11: $MSE = \sum_{t=1}^{n}(y-y)^2/n$

Equation 12: $RMSE = \sqrt{\sum_{t=1}^{n}(y-y)^2/n}$

RESULTS

In the present study, the BBA was applied to select a valuable subset of features, and the selected features were evaluated for the classification and assessment of the criteria. The simulation results regarding the accuracy, sensitivity, and specificity criteria were shown with feature selection; using the bat algorithm, as well as the SVM [9], J48 [18], and NB [18] classifiers (Table 3). According to the information in Table 3, the SVM algorithm, along with the feature selection; using the BBA had a better function with the accuracy of 99.28%, which could be attributed to the superior feature of the SVM in its functioning with large datasets. After the SVM, the NB algorithm, along with the feature selection; using the binary bat algorithm played a key role in the accurate detection of breast cancer with 96.43% accuracy. The second column of Table 3 indicates the sensitivity of the proposed method, which measures the positive cases that have been correctly identified. As observed, the SVM, along with the BBA had the optimal performance. The third column in Table 3 shows the diagnostic criterion regarding the negative cases that were correctly identified. The SVM algorithm, along with the feature selection by the BBA had the optimal function in this respect (97.77%). Table 3 also shows the other assessed criteria. Accordingly, the use of the data mining algorithms, along with the feature selection approach by the BBA had higher accuracy in the identification of breast cancer patients owing to the higher and faster convergence compared to other algorithms.

The mean square error (MSE) is a statistical tool used to determine the predictive accuracy in modeling [14]. The value of this criterion is shown in Table 3 for the proposed method and three algorithms. The root-mean-square error (RMSE) is a measure that was used for the proposed scenarios as presented in Table 3. RMSE is mostly used to calculate the differences between the predicted values by the model and the observed value [14]. The accuracy of the proposed model is higher if the MSE of one particular model is lower than the other model.

Selected Features

In this study, feature selection was carried out; using the BBA to detect breast cancer. The method was proposed to increase the accuracy of breast cancer diagnosis by reducing the features. The applied dataset was extracted from the UCI and analyzed and compared using the SVM, J48, and NB algorithms. According to the results, the performance of the SVM algorithm, along with the use of the BBA algorithm provided higher accuracy in the diagnosis of the disease compared to the other algorithms. The features selected by BBA, which were evaluated in three algorithms are presented in Table 4.

Table 4: Selected Features [a]

	Sample Code Number: Id Number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
SVM	0	1	1	1	1	1	0	0	0	1
Tree	0	1	0	0	0	0	1	1	0	0
NB	0	1	0	1	0	0	1	0	0	0

[a] Abbreviations: NB, naïve bayes; SVM, support vector machine

Also the comparison of the proposed method with several examined articles is presented in Table 5. As shown in Table 5, the proposed method has gained better results in terms of accurate diagnosis in comparison to the results obtained in other studies [19-22]. In the meantime, llangovan et al., have

Table 3: Results of Criteria Evaluated in the Proposed Method [a]

	Algorithm	F-measure, %	Specificity, %	Sensitivity, %	Accuracy, %	MSE	RMSE
1	SVM	99.48	97.77	100	99.28	2.86	1.69
2	J48	94.38	86.27	66.15	92.86	7.14	2.67
3	NB	97.44	95.24	70.37	96.43	3.57	1.89

[a] Abbreviations: MSE, mean square error; NB, naïve bayes; RMSE, root-mean-square error; SVM, support vector machine

selected the attributes; using the relief algorithm [20]. This algorithm can perform well but it doesn't consider the attributes that have an extension, so the results of using this algorithm will be a non-optimal series that will reduce the accuracy of the diagnoses. In the study of Yang et al., the attributes are chosen; using the PCA method [22]. Using this fast and simple method has several defects. In this technique, the attributes are not selected automatically; as a result, there is always the possibility of making a mistake by the user. But in the proposed method of this study, the attributes are chosen in an optimal way and without any human interference which in turn has increased the efficiency of this algorithm. One of the key features of the BBA is its fast convergence capability in the primary stage through switching from exploration to extraction. This feature allows this algorithm to be effective in applications, such as classification, and generally when a quick solution is required. The results from this research showed that using the BBA, along with the SVM algorithm, yields good results in diagnosing diseases.

Table 5: Comparison of the Results for the Proposed Method

Research	Accuracy, %	Sensitivity, %	Specify, %
2018 [21]	82	45	−
2019 [20]	85.89	76	86
2019 [22]	97.71	98.42	99.50
2019 [19]	98.24	−	−
Proposed Method	99.28	100	97.77

DISCUSSION

The main purpose of this study was to recognize the best attributes in diagnosing breast cancer; using the BBA. Extracting useful attributes from a massive amount of data related to people's disease and medical records was done; using data mining processes. This technique can lead to early diagnoses and prevent death. Based on the results of this study, the important attributes found by the BBA and SVM categorization are clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, and single epithelial cell size. The definite diagnoses of this disease need multiple tests and is a time-consuming process, but the proposed method has been able to prove the diagnoses of this disease with the precision of 99.28%. These attributes greatly help the early clinical and empirical diagnoses of breast cancer and increase their accuracy. This can reduce the number of medical tests that are time-consuming and put a lot of expenses on the health care system and patients. The process of feature selection with the BBA and three data mining algorithms are evaluated in this research. Proper feature selection is an important step in pattern detection, while noisy, fake, inappropriate, and additional features in the main datasets not only increase the computational costs, but they also reduce the quality of the classifier function. The reduction of the required information leads to the reduced duration of implementation, simplified structure, and improved classification performance. Therefore, feature selection by reducing the dimensions of the data plays a key role in the accurate detection of the classification algorithm. As a result, finding the optimal subset of features could effectively differentiate various models.

CONFLICT OF INTEREST

The authors declared no conflict of interest.

ETHICS APPROVAL

Not applicable.

REFERENCES

1. Ramachandran P, Girija N, Bhuvaneswari T. Early detection and prevention of cancer using data mining techniques. nt J Comput Appl. 2014;97(13):48-53. DOI: 10.5120/17069-7492.
2. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. J Algorithm Comput Technol. 2018;12(2):119-26. DOI: 10.1177/1748301818756225.
3. Higa A. Diagnosis of Breast Cancer using Decision Tree and Artificial Neural Network Algorithms. Int J Comput Appl Technol Res (IJCATR). 2018;7(1):23 - 7.

4. Aljawad DA, Alqahtani E, Al- Kuhaili G, Qamhan N, Alghamdi N, Alrashed S, et al., editors. Breast cancer surgery survivability prediction using bayesian network and support vector machines. 2017 International Conference on Informatics, Health & Technology (ICIHT); 2017 Feb 21-23; Riyadh, Saudi Arabia: IEEE. DOI: 10.1109/ICIHT.2017.7899000.

5. Heidari M, Zargari Khuzani A, Danala G, Mirniaharikandehei S, Qian W, Zheng B, editors. Applying a machine learning model using a locally preserving projection based feature regeneration algorithm to predict breast cancer risk. SPIE Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications; 2018 Mar 6; Houston, Texas, United States: SPIE. DOI: 10.1117/12.2291524.

6. Osman AH. An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. Int J Adv Comput Sci Appl. 2017;8(4):158-65. DOI: 10.14569/IJACSA.2017.080423.

7. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM Ensembles in Breast Cancer Prediction. PLoS One. 2017;12(1):e0161501. DOI: 10.1371/journal.pone.0161501 PMID: 28060807.

8. Tuo Y, An N, Zhang M. Feature genes in metastatic breast cancer identified by MetaDE and SVM classifier methods. Mol Med Rep. 2018;17(3):4281-90. DOI: 10.3892/mmr.2018.8398 PMID: 29328377.

9. Lakshmi IV V, Krishnaveni G. Performance assessment by using SVM and ANN for breast cancer mammography image classification. Int J Eng Technol Sci Res. 2017;4(9):620-6.

10. Alickovic E, Subasi A, editors. Normalized Neural Networks for Breast Cancer Classification. International Conference on Medical and Biological Engineering (CMBEBIH) 2019. IFMBE Proceedings; 2020; Cham: Springer International Publishing. DOI: 10.1007/978-3-030-17971-7_77.

11. Vaidehi K, Subashini TS. Breast tissue characterization using combined K-NN classifier. Indian J Sci Technol. 2015;8(1):23–6. DOI: 10.17485/ijst/2015/v8i1/52818.

12. Chaurasia V, Pal S. Performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer disease. Rev Res. 2014;3(8).

13. Wolberg WH. UCI Machine Learning Repository Wisconsin, USA: University of Wisconsin Hospitals Madison; 1992 [updated 1992 Jul 15; cited 2020 Nov 28]. Available from: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).

14. Kamel SR, YaghoubZadeh R, Kheirabadi M. Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer. J Big Data. 2019;6(1):90. DOI: 10.1186/s40537-019-0247-7.

15. Speed T. Statistical analysis of gene expression microarray data. Washington, DC, USA: CRC Press; 2003. 240 p. DOI: 10.1201/9780203011232.

16. Nakamura RYM, Pereira LAM, Costa KA, Rodrigues D, Papa JP, Yang X, editors. BBA: A Binary Bat Algorithm for Feature Selection. 25th SIBGRAPI Conference on Graphics, Patterns and Images; 2012 Aug 22-25; Ouro Preto, Brazil: IEEE. DOI: 10.1109/SIBGRAPI.2012.47.

17. MathWorks. MATLAB for Artificial Intelligence. MathWorks. USA: The MathWorks, Inc.

18. Sa'di S, Maleki A, Hashemi R, Panbechi Z, Chalabi K. Comparison of data mining algorithms in the diagnosis of type II diabetes. Int J Comput Sci Appl. 2015;5(5):1-12. DOI: 10.5121/ijcsa.2015.5501.

19. Al- Muhaimin K, Mahmud T, Acharya S, Islam A. Breast cancer prediction using different machine learning models. [Dhaka, Bangladesh]: Brac University 2019. 33 p.

20. Sangaiah I, Vincent Antony Kumar A. Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (RF-EGA) approach: application to breast cancer prediction. Cluster Comput. 2019;22(3):6899-906. DOI: 10.1007/s10586-018-1702-5.

21. Sun D, Wang M, Li A. A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. IEEE/ACM Trans Comput Biol Bioinform. 2019;16(3):841-50. DOI: 10.1109/TCBB.2018.2806438.

22. Yang L, Xu Z. Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning. Int J Mach Learn Cybern. 2019;10(3):591-601. DOI: 10.1007/s13042-017-0741-1.