GA-ANN Frame Work for Breast Cancer Classification using NSGA-II

Mallikarjuna Rao Gundavarapu[1], M.Divya Satya Padma[2],
Ch. Mallikarjuna Rao[3], D. V Lalitha Parameswari[4], and Saaketh Koundinya G[5]

[1,2,3]Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
[4] G, Narayanamma Institute of Technology and Science, Hyderabad, India
[5]National Institute of Technology Wrangal, Hyderabad. India
[1] gmr_333@yahoo.com, [2] divyasatyapadma@gmail.com, [3]chmksharma@yahoo.com,
[4]lalla_mks@yahoo.com, [5]gsaaketh@studentnitw.ac.in

Abstract. The Indian Medical Research Council records 1.5 lakh new cases of breast cancer in India, 70,000 of which occur each year. One of the main aspects of this problem is proper diagnosis at proper time In spite of various approaches suggested by researchers in this direction but their dependability on dataset, outliers, technology and transitory behaviour restricts their efficiency. The multi-object optimization approaches are extremely useful in dealing with this sporadic behaviour of cancer cells. In this paper, we use multi-target evolutionary hybrid system, combination of Genetic algorithm and Artificial Neural Network, to build a smart technology for the context neural propagation in the direction of breast cancer classification. The first stage focusses on development un-dominated genetic sorting algorithm (NSGA-II) and second stage hybridize the enhanced algorithm to improve the conversion speed into the non-dominated front. The key idea of the proposed algorithm was to introduce a new technique to solve artificial neural architecture problems automatically. The scientific findings obtained by the proposed intelligent technology tested using the breast cancer dataset demonstrated the potential for enhancing the efficiency of the algorithm proposed with superior performance with 97.68% accuracy compared to individual artificial neural network approaches.

Keywords: NSGA-II**,** least successive improvement**,** crowding distance, multi objective optimization**.**

1 INTRODUCTION

Cancer is the world's leading cause of death [7][8], making early detection of cancer significant. It is stated that microRNAs (miRNAs) are associated with different cancers and vary between normal tissues and tumour expressions [3][4][5]. Most miRNA trials are, however, biological [1]. Based on their expression, miRNAs can be classified into normal and tumours types. There is, therefore, a pressing problem in

automatically classifying a miRNA sample as either of these classes. For the classification of data, many regulated master learning algorithms are created, which inputs marked data and generate inference for unknown data mapping. These algorithms consist of many parameters that can be altered by the issue at hand and increase the classification performance. Gaspar-Cunha etal, used SVM-supporting vector machine (MSM) data for the automatic classification of single-proton computation (SPECT) emission data while developing Reduced Pareto Set Elitist Genetic Algorithm. Peng et al. use MRNAs and MIRNAs to select a set of mRNAs and miRNAs for the classification of cancer effective tissue gradating, while Mukhopadhya etal, use SVM-nRFE wrapper method to select the necessary miRNAs for the normal and tumour tissue classification. All such studies have the same disadvantage, i.e. using only one classifier such as SVM it alone but may not sufficient to overcome various classification problems. Some classifiers are good for certain regions and others are good for others. It is therefore important that the classifier for a specific classification problem is automatically selected from a collection of classifiers. In addition, automatic function and parameter combination selection is also required for the selected classifier. None of the available methods therefore offers a means of combining the number of solutions obtained following the implementation of any multi-target optimization technique.

Artificial Neural Network (ANN) has recently grown into the substratum of soft computing methods used in the efficient resolution of various issues [1] [2]. ANN is the machine model that imitates the way the human brain neuron system functions. Therefore, various approaches and methods for helping the computer systems have been used. In particular, due to their high predictability and adaptability [7], ANNs are most frequently used classificatory. More intensive work is therefore needed for the design and production of the ANN classification for problems of classification. The use of a machine system has increasingly increased. Continuous research efforts have been ongoing over the last decade to use soft machine techniques for classification problems. Healthy candidates for multi-target optimization problems (MOOP)[2] are the Evolutionary Algorithm (eAs). Due to their ability to find many optimal solutions in Pareto and their success in the global search field. One of the hottest fields in the field of evolutionary computation is multi-objective evolutionary algorithms (MOEAs).They were ideal for creation and design of adequate and exact ANNs in order to optimize two opposing goals, namely the minimization of ANNs structural complexity and network capacity maximization. Therefore, MOEAs were recently introduced with great success to simultaneously optimize structure, link weights and network training [2][4[[5].

Any peculiar findings are found in almost all datasets. The data analysis method may be influenced by such outliers. The numerous methods and techniques for external identification show the significance of outliers in data collection. Distance and density-based approaches are distinct from linear models. Statistical methods are generally used for the analysis of quantitative, actual, continuous values data sets, or at least qualitative, ordinal values. Currently, the analysis of non-ordinal data is very important. The multidimensionality of information (variable) must be discussed. The literature includes distance-based approaches. The nearest neighboring method is one of the most commonly used and updated methods. This is a non-parametric approximation approach that uses measures like Euclid, Manhatan, Shebyshev, or

other behavior to separate our world from our neighbor. For the detection of contours, the k-NN algorithm is also improved. See function, e.g. [3]. The fundamental issue that may occur in this situation is that the consumer does not have the requisite details about data distribution and so the model is incorrectly calculated and the outliers are wrongly identified.
In summary, tumor classification approaches suffers from some single classifier difficulty, insufficient data distribution, ANN over fitting and under fitting. In this regard we proposed GA-ANN framework which automatically provides optimal ANN classifier based on the problem behaviour. The genetic algorithm model was developed using NSGA II to provide optimal artificial neuron architecture. The approach proposed emphasizes with two fitness functions to test the efficiency of the BP algorithm and classification precision based on the MSE This optimized ANN architecture is capable of adapting insufficient data scenario, deal with outliers, avoid over fitting and improves classification accuracy.

2. LITERATURE REVIEW

Siti Mariyam Shamsuddin,(2018): One of the common causes for cancer of the breast is breast cancer. This condition is diagnosed on a human basis. It requires some investment and has a human mistake factor in the outcomes. In a solitary run for boundary parametric streamlining of counterfeit nerve systems (ANNs), Pareto ideal developmental multi-target advancement is utilized to accomplish numerous conclusive outcomes. In this paper, a robotized classifier framework for the determination of bosom malignant growth is presented in a mechanized manner. So as to acquire right arrangement consequences of analyzed bosom malignancy issues, the proposed system utilized a multi-layer perceptron organize (MLP) in view of improved not-commanded arranging hereditary calculation (NSGA-II). Also, it is utilized to improve the system structure and all the while decrease the mistake pace of the MLP neural system. The proposed approach is effective for breast cancer diagnosis in contrast to other approaches found in the literature.

Xiaoke Ma,(2017): The advances in biotechnology allow data to be produced simultaneously for several conditions. In multiple networks, discovery of the condition-specific modules is essential for Understanding of the molecular processes underlying the cells. The available algorithms allow for a single, low-precision objective optimization problem for the different networks. A genetically engineered multi-target algorithm for the discovery of the condition modules in multiple networks (MOGA-CSM).We exhibit that the MOGA-CSMout utilizes counterfeit systems to achieve cutting edge strategies with extraordinary accuracy. What's more, MOGA-CSM distinguishes stage-explicit modules in bosom malignant growth systems dependent on information produced from the TCGA. The proposed model and algorithm provide for efficient multi-network analysis.

Lukasz Chomatek, (2019): Any abnormal findings are found in almost all datasets. The data analysis method may be influenced by such outliers. While there are many outlier detection methods, a modern, more efficient approach must still be sought. In

this article we suggest a series of goals to classify outliers effectively using multi-target genetic algorithms. Research has shown that the most common genetic algorithms intended for multi-objective optimization can successfully employ this procedure. The results of tests on the medical data set in the repository show that our approach can be applied successfully to the medical problem.

Ashraf Osman Ibrahim,(2014): Evolutionary algorithms (EAs) were population based algorithms that permit simultaneous exploration of various components in optimum collection of Pareto. This paper introduces a 3-term back-propagation Network for Classification Problems Memetic Elitist Pareto Evolutionary Algorithm. The memetic elitist evolutionary Pareto algorithm called METBP is used for the development of the 3-term back propagation (TBP) network that is suitable for weight, error rates and architectural complexity at the same time. METPB is based on the NSGA-II search algorithm that enhances individuals within the algorithm population. The numerical results of METPB show the benefit of the combination of local search algorithm and can achieve a TBP network with a higher accurateness of classification and simpler structure than the multi-objective TBP (MOGATBP) genetic algorithm network and some of the methods described in the literature.

Sriparna Saha,(2017): MicroRNA (miRNA) assumes a basic job in natural cycles including RNA grafting and guideline of quality articulation. Examination has indicated the conceivable connection among's oncogenesis and the articulation profiles of certain miRNAs, because of their particular articulation of the typical tissue and tumor tissue. However, it was rarely considered to immediately assign miRNAs into various groups by looking at the similarities between their expressive values. This article provides a method to solve problems with cancer, miRNA, and mRNA expression data sets for real-life classification. In the first object, a multiple goal optimization system based on a non-dominated genetic sorting algorithm II is proposed, in addition to its suitable parameter and feature combinations that are applicable to the classification of a given dataset, to automatically determine the suitable classification type. A stack-based ensemble technology is used on the second page to achieve a single integrated solution from the first stage solutions.

3. MATERIAL AND METHODS

The technique proposed incorporates data on breast cancer by LS method to boost the final solutions. The NSGA-II has been developed to correctly improve and optimize the structure of ANN. The proposed algorithm is therefore able to find the right amount of nodes in middle layer with the low rates of error.

3.1 *Non-dominated Sorting Genetic Algorithm (NSGA-II):*

The enhanced version in the basic NSGA-II has improved the BP algorithm. The technique proposed fixes the accuracy and network architecture at the same time as each BP neural network is completely defined. Stepwise procedure is given below:

Step1: Copying, standardizing and interpreting data.

Step 2: Break date into data for training and research**.**

Step 3: Set the highest, least number of secret nodes and a highest number of iterations**.**

Step 4: Comparison of person length.

Step 6: Generate and launch ENHANCED nsga-II community**.**

Step 7: Assess every individual according to fitness functions in each iteration**.**

Step 8: Stops and releases a collection of BP networks that are not dominated, if the full iteration reaches.

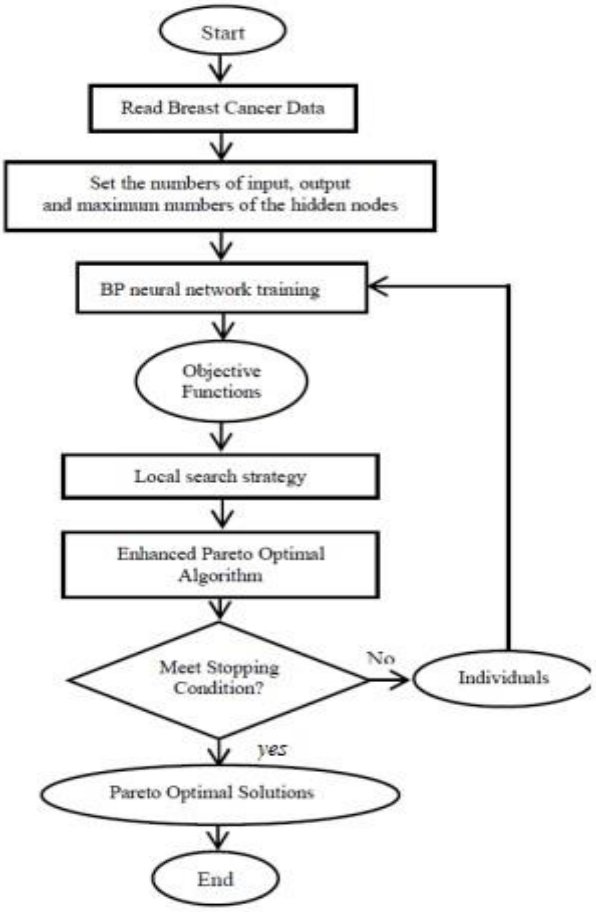

Fig. 1. Process flow model

Under NSGA-II Algorithm (non-dominated), Deb, Pratap, Agarwal, and Meyarivan (2002); a new version of NSGA (Srinivas & Deb, 1994) was proposed (in

2006). Two algorithms were GA operators based. Furthermore, NSGA-II was a fast and elitist multi-target genetic algorithm used to generate the optimal frontal solutions from Pareto [9][10]. The good performance of the global quest for a multi-target genetic optimization algorithm of non-dominated sortation has been the favourite method of optimisation. The new method and arithmetic operator are proposed, a simple, un-dominated sorting approach with a crowded comparator. NSGA-II is known as an optimal solution algorithm of the most popular in Pareto. It needs to minimize or optimize two or more objective functions simultaneously. All these experiments have shown the possibility of optimizing the genetic algorithm and its improved variants. A random population of chromosomes or solutions of size N starts with the NSGA-II algorithm. Firstly, all parents and descendants are merged into a single 2N population rather than simply considering the non-dominated frontiers of their descendants. Then, the whole population undergoes the non-dominated sorting process. This method permits a global non-dominance.

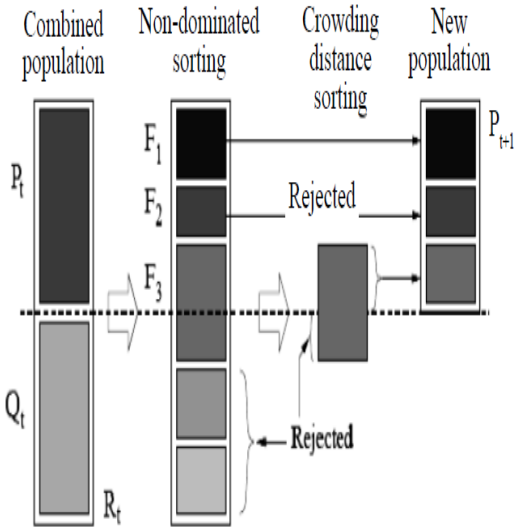

Fig. 2. Schematic of NSGA-II algorithm

It controls between parent and offspring solutions and helps NSGA-II to converge more quickly. The NSGA-II scenario is shown in Figure 2. The crowding distance is used to choose parents for a new person and to choose a new population based on the solution's comparison of congestion. In order to preserve the versatility of the solutions, a greater crowd gap is desired.The approach proposed emphasizes two fitness functions to test the efficiency of the BP algorithm. The first target is precision, which is based on the MSE. Although network complexity is the second goal, the number of neurons in the midfield of the network is defined.

3.2 Formulation of the proposed approach using multi-objective optimization

A multi-target optimization technology has been developed for selecting a group of classifiers, optimizing precise performance, retrieving and the number of features selected at the same time. In order to achieve the result in the data sets, the selected classifiers will then use a stacked ensemble technique. There are two step methods, as discussed below, to the proposed technique.

First stage: In order to determine the classification problem for the particular classifier type, parameter blends and functional combination, we used the standard multi-target optimisation method for genetic sortation non-dominated (NSGA-II) research capability. The basic steps of NSGA-II are shown in Figure 1.

String representation: As inputs of NSGA-II, individuals or strings are used. The solution to the problem is encoded with a string. The string is described in three sections, i.e. classification form, parameters and function combinations, since the problem has three sub-components (Figure 3). The principal section shows the sort of order. Four arbitrary tree (RF), irregular tree (RT), least successive improvement (SMO) and calculated relapse (LR) have been utilized in the current examination. All the arrangements spoke to by values 1, 2, 3 and 4 of RF, RT, SMO and LR are accessible in a specific arrangement. The second bit of the string/arrangement contains boundaries for a particular order. The boundaries utilized for RF are the quantity of arbolics (potential qualities: 10, 20 and 30) and the quantity of capacities (potential qualities: 0, 5 and 6). Case loads of the leaf are remembered for the boundaries utilized in RT (potential qualities: 1.0, 1.05, and 1.25) and irregular arrangements of capacities (potential qualities: 0, 3 and 7). In our calculation only one SMO boundary, unpredictability (possible values: 1, 3 and 8) is considered. There have been no unique LR parameter values; Instead of default value. A binary string function is the third component of a solution / string, where the '0 'and'1' indicates a certain element is absent or present.

Population initialization: Both strings will be initialized automatically. Random selection of values 1, 2, 3 and 4 for RF, RT, SMO and LR respectively can be done in the first section. A random set value is selected to assign parameter values to a specific classifier to initialize the second part of the chainstring... For instance, if you have the SMO value in the first component, then you can select a parameter 1, 3, or 8 randomly from the set. Finally, the third element has "0" or "1" values. If there are total features of N in the data set, then the location of and feature is 0 or 1. This produces a binary N-string.

Objective function Calculation: Let (S) indicate a set of features whose values in the string/solution function are "1' (third section). You will obtain the encoded classifier and its parameter and combination of functions (S). The chosen classifier is performed in the existing data set with a cross-validation leave one-out (LOOCV), along with the selected parameters and function combinations. The first two objective functions whose value is to be maximized (higher accuracy values and retrieval are equal to the high quality) are determined by using two classification metrics, average accuracy and retrieval values. The third feature is to be reduced to a minimum, that is, the number of functions (S).

A

Classifier	Parameters	0 1 1 1 0..... 0 1

Feature combination

B

	F_1	F_2	.....	F_k	P_1	P_2	.....	P_k
S_1	a_1	a_2	.....	a_k	1	1	.....	0
S_2	b_1	b_2	.....	b_k	0	1	.....	0
S_3	c_1	c_2	.....	c_k	0	0	.....	1
.	.	.	.....	.	.	.	.....	.
.	.	.	.....	.	.	.	.....	.
S_n	d_1	d_2	.....	d_k	1	0	.....	0

Fig. 3. (A) Structure of solution (B) Solution Stack

A proposed. First stage string / solution representation. Three sections include the classifier sort, the parameters for the selected classifier and the combination of features.

B. The second step of the suggested solution is a stack-based ensemble. F1, F2, Fk represent the corresponding characteristics; P1, P2... Pk is the expected class labelling that corresponds to a specific classification. P1, F2: ..., Pk represents samples that are in the dataset. Therefore, ' '0' and "1' signify the lack and presence of a particular feature. Genetic algorithm-II of NSGA-II, non-dominated sortation.

3.3 Mathematical Model of Proposed Approach

If X is a random variable with a Pareto (Type I) distribution, then the probability that X is greater than some number x, i.e. the survival function (also called tail function), is given by

$$\overline{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_{\mathrm{m}}}{x}\right)^{\alpha} & x \geq x_{\mathrm{m}}, \\ 0 & x < x_{\mathrm{m}}, \end{cases}$$

From the definition, the cumulative distribution function of a Pareto random variable with parameters α and x_{m} is

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_{\mathrm{m}}}{x}\right)^{\alpha} & x \geq x_{\mathrm{m}}, \\ 0 & x < x_{\mathrm{m}}. \end{cases}$$

It follows (by differentiation) that the probability density function is

$$f_X(x) = \begin{cases} \frac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}} & x \geq x_{\mathrm{m}}, \\ 0 & x < x_{\mathrm{m}}. \end{cases}$$

The expected value of a random variable following a Pareto distribution is

$$\mathrm{E}(X) = \begin{cases} \infty & \alpha \leq 1, \\ \frac{\alpha x_{\mathrm{m}}}{\alpha - 1} & \alpha > 1. \end{cases}$$

4. RESULTS AND ANALYSIS

Experimental results

This segment presents the mathematical aftereffects of the proposed technique which incorporate an improved NSGA-II with the LS system for bosom malignant growth. For this situation, the strategy for cross-approval used to assess the method proposed is 10fold. This paper utilizes different measurements, for example, **affectability, qualities and exactness**, as shown in conditions 1-3. The affectability of similar positive examples by the quantity of valid and bogus positives. The accuracy is utilized to help and gauge the specific negative cases by the quantity of valid and bogus negatives. Exactness is a test for right results. True positive (TP), True negative (TN), False positive (FP) and False negative (FN) parameters are used to find accuracy, Sensitivity and Specificity. Figure 5 and figure 6 reflects the same.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FP} \quad (2)$$

$$Specificity = \frac{TN}{TN + FN} \quad (3)$$

We can see the training and test error rates from Table 1. The findings also synthesize the algorithm's generalization error. Table 1 indicates a more effective method for the treatment of breast cancer. However, the average error rates that were obtained with the LS and BP algorithm for one single series of the multi-target evolutionary improved NSGAI hybrid.

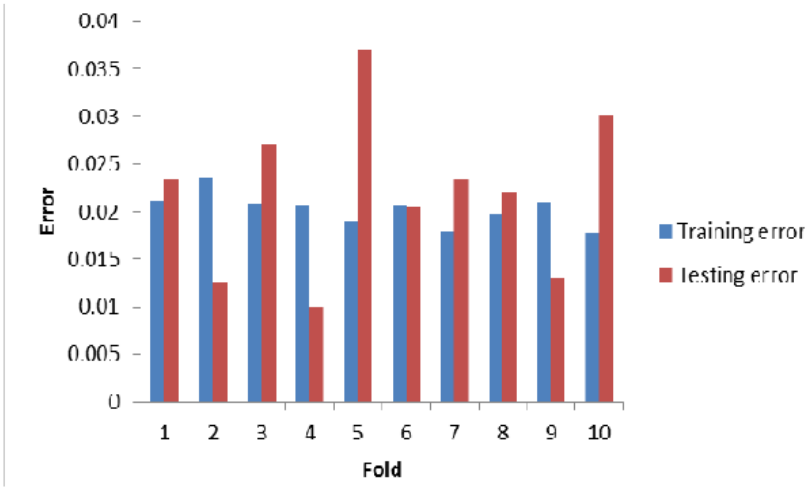

Fig. 4. The training and testing errors

Fig. 5. Training and testing specificity results

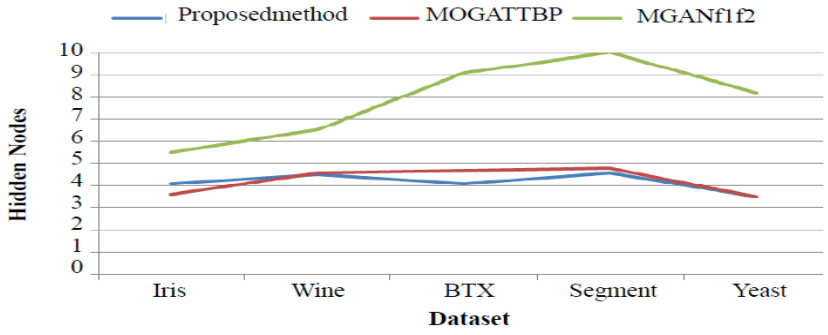

Fig. 6. . Comparison of the hidden nodes of the proposed method and other Methods with different datasets.

Table 1. Training and Testing Error

#Fold	Train error	Test error
1	0.0211	0.0233
2	0.0236	0.0125
3	0.0209	0.0271
4	0.0207	0.0100
5	0.0188	0.0371
6	0.0206	0.0204
7	0.0180	0.0234
8	0.01986	0.0125
9	0.0210	0.0220
10	0.0177	0.0301
Average	0.0202	0.0218
Standard Deviation	0.0017	0.0084

Table 2. Performance comparison of proposed approach with others

Method	Hidden Nodes	Sensitivity Specificity	Accuracy
ANN	4,10	96.67	96.57
SVM	4,6	98.77	97.01
Genetic Algorithm	4,7	96.01	96.97
Proposed Approach: NSGA II GA+ANN	1,3	97.60	97.68

5. CONCLUSION

Our proposed frame work, as shown figure 6 performs well across the various cancer data sets compared to other approaches. In comparison to the non- optimizing neural network, the model built with the genetic network algorithm results in a greater accuracy, as shown in table 2. The change can be seen by the precision rate for the Neural Network model algorithm, which after optimization amounted to 96.57%; the value of the Genetic-based Neural Network algorithm has increased to 97.68%, with a difference of 1.11%. The ROC curve is used for the assessment. While improved results have been achieved in the implementation of the GA optimized model of the neural network algorithm, there are some drawbacks, such as the need for a longer calculation time from 6 to 20.33 minutes. Table 2 further reveals that better performance is achieved with less number of hidden nodes in the ANN architecture.

ANN convergence time increases with the number of hidden nodes. Thus the goal is to achieve a better accuracy of prediction for breast cancer diagnosis with optimal classifier is obtained using the hybrid GA-NN technology than NN alone. Our proposed GA-ANN frame work shown consistency in its performance across different datasets.

REFERENCES

[1.] Ashraf Osman Ibrahim,2018," Intelligent breast cancer diagnosis based on enhanced Pareto optimal and multilayer perceptron neural network".

[2.] Ashraf Osman Ibrahim, 2019, Back propagation Neural Network Based on Local Search Strategy and Enhanced Multi-objective Evolutionary Algorithm for Breast Cancer Diagnosis.

[3.] Derisma, Meza Silvana, Imelda (2018), Optimization of Neural Network with Genetic Algorithm for Breast Cancer Classification, International Conference on Information Technology Systems and Innovation, 978-1-5386-5692-1, 22-25.

[4.] Deja ., W. Froelich, G. Deja, A. Wakulicz-Deja, Hybrid approach to the generation of medical guidelines for insulin therapy for children. Information Sciences. 384, 157–173 (2017).

[5.] Duraj, A., Chomatek, L.: Supporting breast cancer diagnosis with multi-objective genetic algorithm for outlier detection. In: International Conference on Diagnostics of Processes and Systems, pp. 304–315. Springer (2017)

[6.] Duraj, A., Szczepaniak., P.S.: Information outliers and their detection. In: Information Studies and the Quest for Trans disciplinary, pp. 413–437. World Scientific Publishing Company (2017)

[7.] Gevaert, O.; Tibshirani, R.; Plevritis, S.K. Pan Cancer analysis of DNA methylation-driven genes using MethyMix. Genome Biol.2015, 16,17,doi:10.1186/s13059-014-0579-8.

[8.] Rodina, A.;Wang, T.; Yan, P.; Gomes, E.D.; Dunphy, M.P.; Pillarsetty, N.; Koren, J.; Gerecitano, J.F.; Taldone, T.;Zong, H.; et al. The epichaperome is an integrated chaperome network that facilitates tumour survival. Nature 2016, 538, 397–401.

[9.] Xiaoke Ma,2017," Multi-Objective Optimization Algorithm to Discover Condition-Specific Modules in Multiple Networks".

[10.] Lukasz Chomatek, 2019," Efficient Genetic Algorithm for Breast Cancer Diagnosis". DOI: 10.1007/978-3-319-91211-0_6 In book: Information Technology in Biomedicine (pp.64-76)