Original Research Article **Biotechnology**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

ANALYSIS OF EXPRESSION LEVEL OF BREAST CANCER GENE USING MACHINE LEARNING ALGORITHMS FOR DIAGNOSIS OF BREAST CANCER

J. SUMITHA[1] AND T. DEVI[2]

*[1]Ph.D Research Scholar *, Department of Computer Applications, Bharathiar University*
[2]Professor and Head, Department of Computer Applications, Bharathiar University

ABSTRACT

The breast cancer is the second leading type of cancer which leads to death among women all over the world. Breast cancer exists due to the mutation occurs in the normal growth of BRCA gene under certain circumstances. In this paper, we used a novel approach for finding the disease – causing gene using computer-assisted algorithms. The existing algorithms are compared with each other to determine the efficiency in detecting the diseases from gene expression value. The results proved that the effectiveness of Hybrid Radial Bias Neural Network (HRBFNN) algorithm performs better than sequential and Divide and Conquer Kernel Solving Support Vector Machines (DCKSVM) algorithm in finding the diseased gene.

KEYWORDS: DCKSVM, HRBFNN, Identification of diseased gene, Sequential model

* **J. SUMITHA**
Ph.D Research Scholar *, Department of Computer Applications, Bharathiar University

***Corresponding author**

Received on : 30-09-2016
Revised and Accepted on : 18-11-2016
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.1.b79-85

INTRODUCTION

Worldwide, cancer is the second leading cause of death and more than 1,500 people per day are affected by this disease and approximately 1,479,350 new cancer cases are diagnosed all over the world. Breast cancer is the most common and rigorous cancer among women and continuing to be a significant public health problem in the world. Approximately 182,000 new cases of breast cancer are diagnosed and 46,000 women die of breast cancer each year. Nearly 192,370 new cases of invasive breast cancer are diagnosed among women and thus, the incidence and mortality of breast cancer are very high. Hence breast cancer is the second leading cause for cancer death in women.[1]The normal BRCA gene exists in the nature of proto-onco gene in the human body. But when the mutation occurs, this proto-onco gene is translated into onco gene which leads to cancer. Breast cancer has very high frequency of death rate, however the reason for breast cancer is still mysterious and there is no efficient way to prevent the existence of breast cancer. So, early detection is the first significant step towards treating breast cancer. Breast cancer screening is imperative to detect breast cancer and the most common screening methods are mammography and sonography. Of these methods, mammography is the most important tool that doctors use to detect, diagnose, and evaluate breast cancer and this technique has been in use for about forty years.[1,2] However , mammography has some disadvantages for detecting breast cancer. Although it is very sensitive, it is not accurate in detecting breast cancer as mammography gives false positive results.[3, 4] Hence it is essential to develop software which could give reliable diagnostic results prevent these drawbacks. The objective of this paper is to detect the disease-causing gene with the help of gene expression value using computer-assisted algorithms. The algorithm which used for this research is the sequential algorithm, DCKSVM and HRBFNN and the prediction is done on the basis of confusion matrix method. The breast cancer dataset is taken from UCI repository URL. Mat lab is the software that has been used for implementing this work. Particularly, the prognostic type of breast cancer dataset is selected for this research.This paper is organized as follows: Section 2 presents proposed methods to solve the task of identifying breast cancer. Section 3 contains results obtained and discussions. Finally Section 4 for conclusion.

METHODS

The methods used for the research is catagorized as sequential algorithm, DCKSVM and HRBFNN.

Dataset Description

The breast cancer dataset used in this research is taken from the URI repository having 567 data with gene expression value is dipicted in Figure-1. From this, 380 data are taken as training data and remaining data are used as testing data. The parameters used for predicting the performance are accuracy, precision, recall and f-measure.

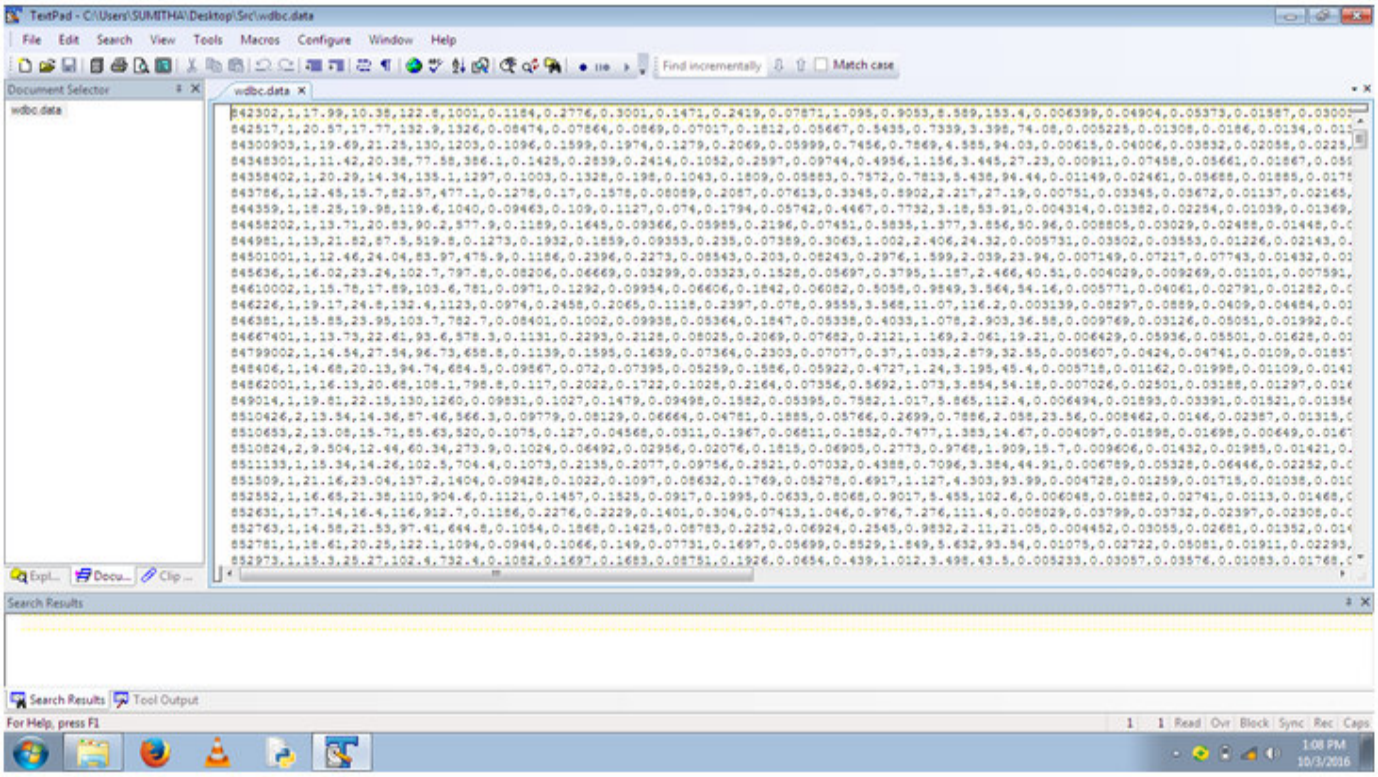


Figure 1
Breast Cancer Dataset

The performance criterion for these classifiers in disease detection is based on confusion matrix method.[5]

Algorithms

Sequential Model

It is used to differentiate the differences between the pair of gene expression values in the dataset. For example, gene1, gene2…..gene10 are genes in the dataset in which the gene pair (g4g8g2) and the another gene pair (g8g4g2) are found to be identical.[6] Then it implies that this two gene pair stimulates the same disease in the human body. This sequences of each gene in breast cancer dataset are taken as an input for predicting results.[7]

DCKSVM

The most commonly used classification algorithm is the DCKSVM algorithm which applied to the breast cancer dataset for predicting the accuracy.[9-10] The purpose of DCKSVM algorithm is to split the main clusters of data into sub-clusters and makes its predictivity which becomes consistent on that clusters.[8]

Algorithm 1: Divide and Conquer SVM

Input: Training Data

Output: The SVM solution A.

Step 1: Partitioning the variables into k subsets {v1….vk}

Step 2: Time complexity for solving sub-problems reduced to O(k*n/k)2=O(n2/k) with space complexity, where n is the variable and k is the cluster subset.

Step 3: After compution of all subproblem solutions, concatenate them to form solution for whole problem a^b=[a1….ak] Step 4: A bound is derived on $||a^b-a^*||_2$ where ab is the optimal solution by adding cluster-kernel values.

Step 5: Minimizing the off-diagonal values of the kernel matrix with a balancing normalization.

Step 6: For each cluster k, go to Step 2 for partitioning data and computing Step 4 for absolute scale.

HRBFNN

To increase the capabilities of data granulation in the dataset, HRBFNN is applied to this dataset which inhibits the characteristics of data granulation and Principal Component Analysis [PCA] for preprocessing the data.[11,12,13] Algorithm 2: HRBFNN

Step 1: Preprocess the data set using PCA. To obtain dimensionality reduction, principal component analysis is used to preprocess data sets for feature extraction and reduction of data.

Step 2: Training and testing data sets are formed.

Step 3: The generic parameters used inthis research are decided.

Step 4: Selected inputs are determined.

Step 5: PFNs are designed. For the selecting r inputs, the number of nodes (PFNs) generated in each layer becomes equal to

$$k = \frac{n!}{n!(n-r)!\,r!}$$

where, n is the number of total inputs and r stands for the number of the chosen input variables and k is the clusters.[11]

Step 6: Check the termination criterion

Step 7: Select the best predictive capability nodes and construct their corresponding layer. It takes only a few seconds to complete 1000 iterations and generates better results than other algorithms in terms of accuracy, precision, recall and f-measure.

RESULTS AND DISCUSSION

The machine learning algorithms applied in this research to analyze the breast cancer dataset for identifying diseases is shown in the Table-1. HRBFNN algorithm gives better results when compared to the sequential and DCKSVM algorithm. The performance can be calculated in terms of accuracy, precision, recall and F-Measure. The criterion for these classifiers in disease detection is based on Confusion matrix. The accuracy percentage of the sequential model, DCKSVM and HRBFNN is 78, 80 and 85 respectively as depicted in Table 1. The results show that the HRBFNN gives 85.18 pecentage of accuracy, 0.79 percentage of precision and 0.83 percentage of f-measure which is higher than the sequential model and the DCKSVM algorithms. However DCKSVM algorithm shows higher recall percentage than sequential model and HRBFNN. The graph plotted for these results are shown in Figure-2.

Table 1

Results of sequential, DCKSVM and HRBFNN

Algorithms / Parameters	Sequential Model(%)	DCKSVM(%)	HRBFNN (%)
Accuracy	77.7778	80.4233	85.1852
Precision	0.7103	0.7504	0.7937
Recall	0.7659	0.8923	0.8713
F-measure	0.7371	0.7892	0.8307

Accuracy is the percentage of correct predictions from the dataset. The Accuracy is the Proportion of the total number of predictions that were correct. It can be calculated using the equation,

$$Accuracy = p + s / (p+q+r+s) \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The recall is the proposition of positive cases that are correctly identified, calculated using the equation,

$$Recall = s / (r+s) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

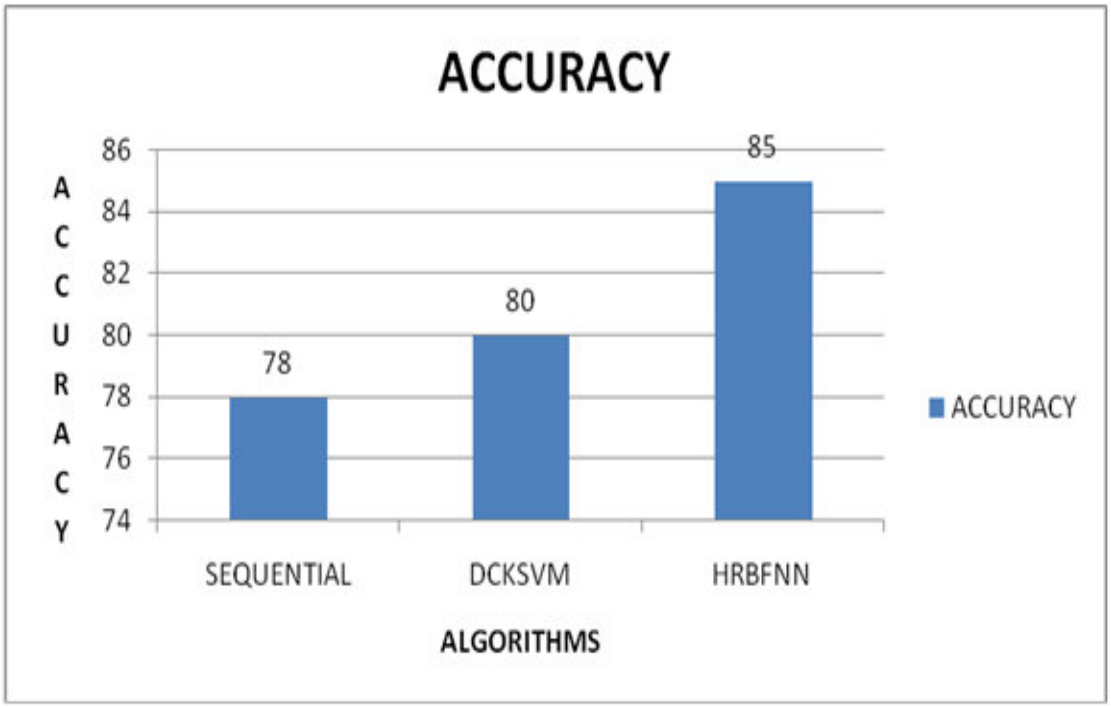
Precision (P) is the proposition of the predicted positive cases that were correct, calculated using the equation,

$$Precision = s/(q+s) \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

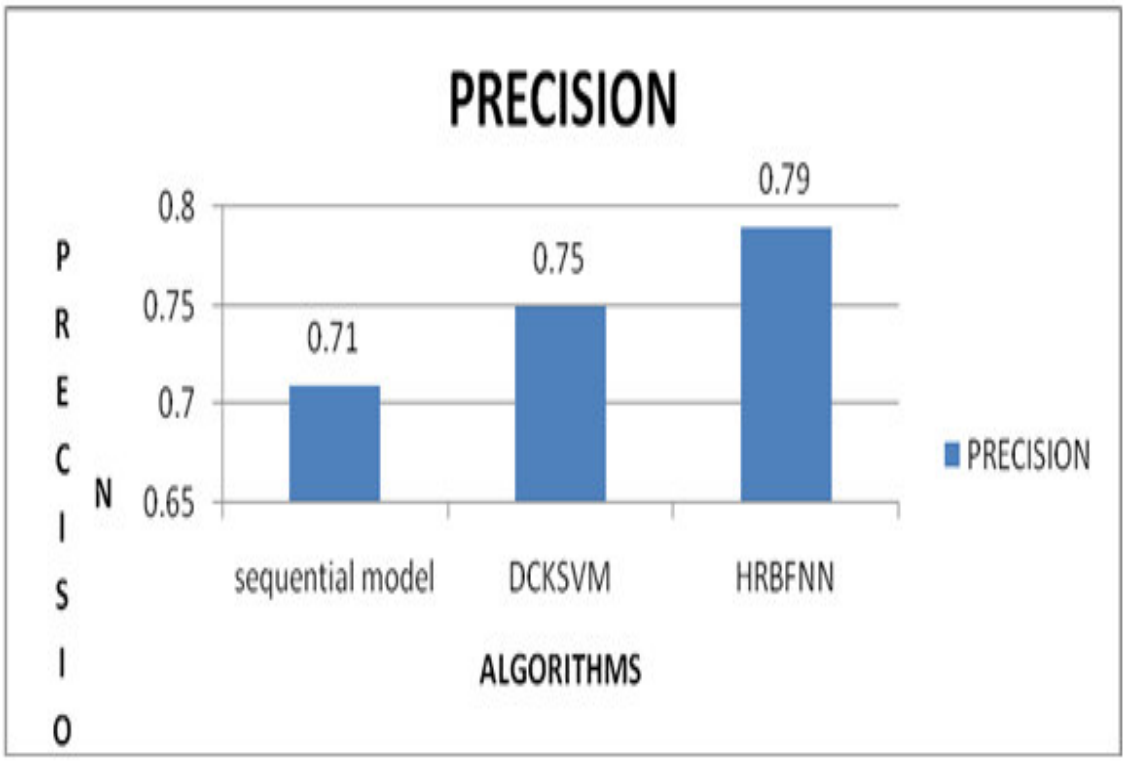
F-measure (F) can be calculated using the formula,

$$F = 2 * (Precision*recall) / (precision + recall) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

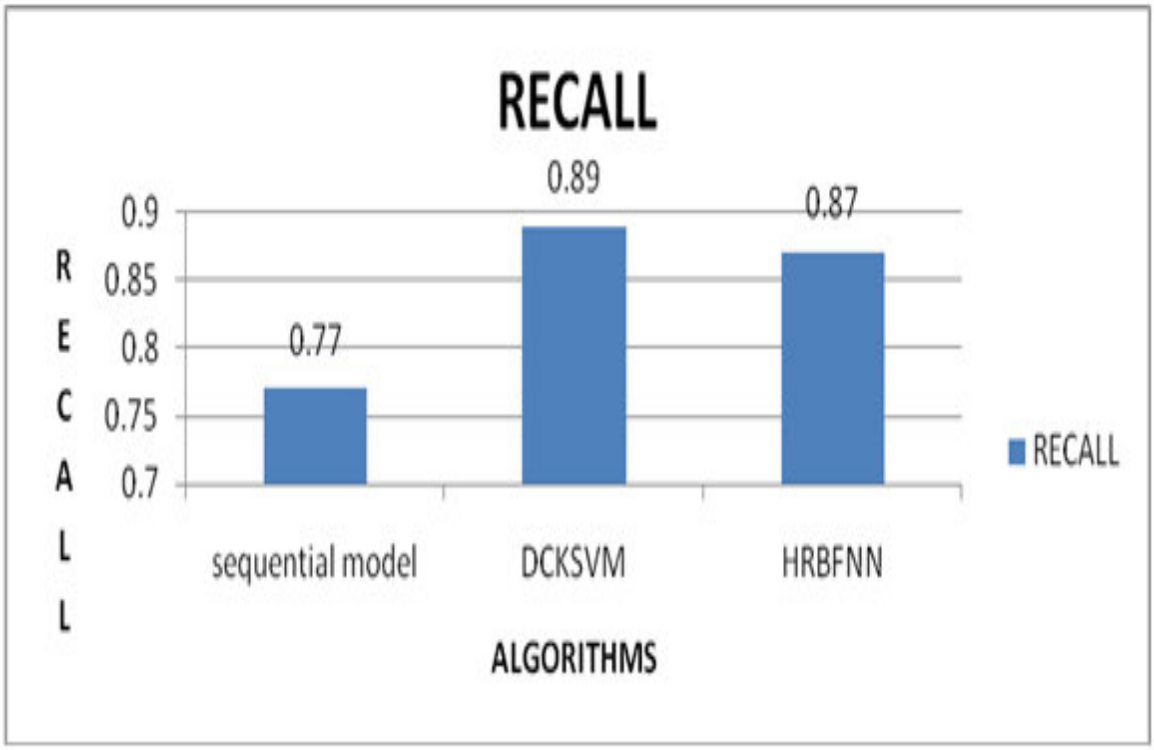
Where p is the number of correct predictions that an instance is negative i.e. correctly predicted genes which are not disease-causing gene, q is the number of incorrect predictions that an instance is positive i.e. incorrectly predicted which are diseases causing gene, r is the number of incorrect predictions that an instance negative i.e. incorrectly prediction of undiseased gene and s is the number of correct predictions that an instance is positive i.e. correctly predictions of disease-causing genes in the dataset. Machine learning algorithm is prevalently used algorithm in related to this research work. Table 1 concludes that HRBFNN predicts higher efficiency than other classifiers. Based on the confusion matrix calculation method, the parameters used for predicting the results are shown in (1), (2), (3) and(4).



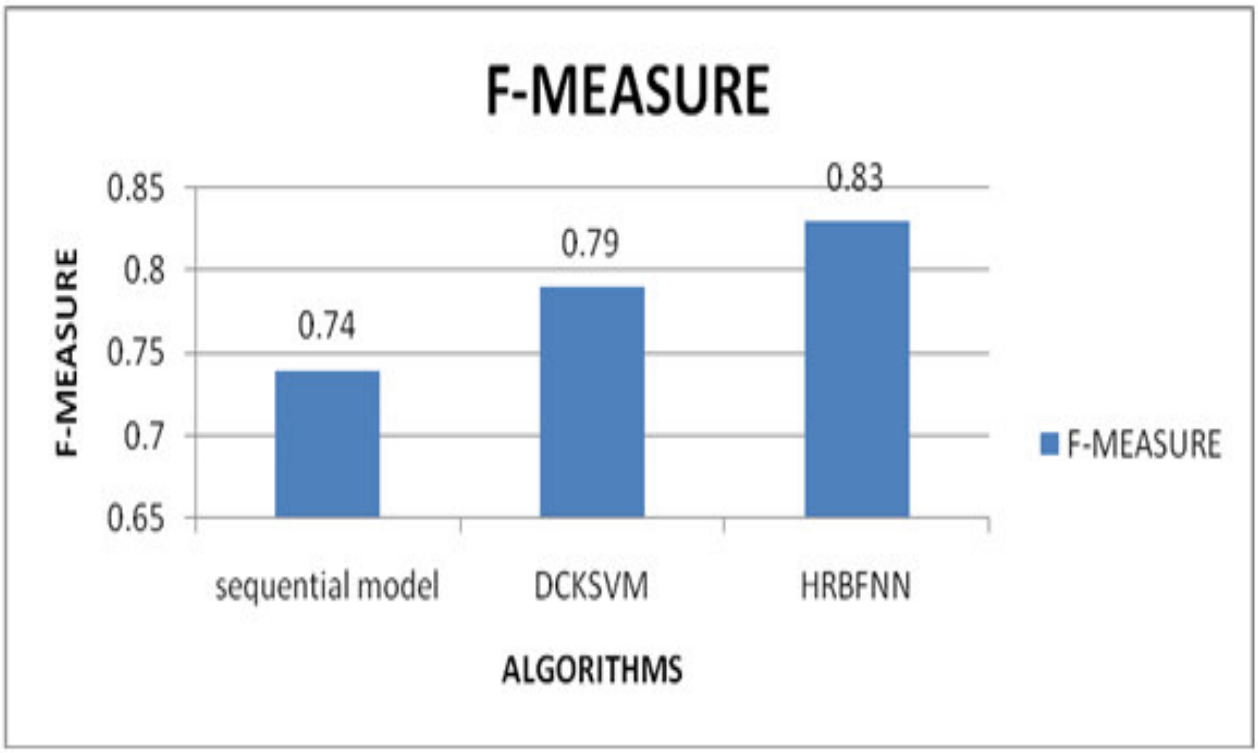
(a)


PRECISION
0.8
0.75
0.7
0.65
0.71
0.75
0.79
PRECISION
sequential model
DCKSVM
HRBFNN
ALGORITHMS

(b)


RECALL
0.9
0.85
0.8
0.75
0.7
0.77
0.89
0.87
RECALL
sequential model
DCKSVM
HRBFNN
ALGORITHMS

(c)



(d)

Figure 2

(a) Accuracy graph of Sequential, DCKSVM and HRBFNN Algorithms, (b) Precision graph of Sequential, DCKSVM and HRBFNN Algorithms (c) Recall graph of Sequential, DCKSVM and HRBFNN Algorithms (d) F-measure graph of Sequential, DCKSVM and HRBFNN Algorithm

CONCLUSION

In this research, HRBFNN proved that it is effective in analyzing the diseased gene from the breast cancer dataset. The performance analysis between the sequential model, DCKSVM and HRBFNN shows that the HRBFNN gives better performance than the DCKSVM and sequential model. In future, it can also be enhanced with some other new computational algorithms to predict new results.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Erin Linnenbringer L. Social Constructions, Biological Implications: A Structural Examination of Racial Disparities In Breast Cancer Subtype [dissertation]. University of Michigan; 2014.
2. Laura Van J, Mao M, Hans L, Matthew Marton J, Peter S. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature*. 2012 Mar 12; 41(5):96-102.
3. Guha S, Meyerson A, Mishra N, Motwani R. Clustering Data Streams: Theory and Practice. IEEE Trans. Knowl. Data Eng.2003 Jan 20; 15(3):515-28.
4. Ingrid Hedenfalk A, Markus Ringner, Jerey Trent M, Ake Borg. Gene Expression in Inherited Breast Cancer. Adv Cancer Res *.2002 June 13;* 8(4):1-34
5. Isabelle G, Jason W, Stephen B, Vladimir V. Gene Selection for Cancer Classification using Support Vector Machines. J Mach Learn Res. 2014 May 24; 46(1):389-422.
6. Ken Kaneiva. A sequential pattern mining algorithm using rough set theory. INT J APPROX REASON. 2011 May 15; 5(2):881–93.
7. Carl Mooney H, John Roddick F. Sequential Pattern Mining – Approaches and Algorithms. ACM Comput. Surv. 2013 June 6;9(1):512-20.
8. Cho Jui H, Si S, Inderjit Dhillon S. A Divide-and-Conquer Solver for Kernel Support Vector Machines. J Mach Learn Res. 2014 May 24; 46(1):300-12.

9. Minseung K, Sung-Hou K. Empirical prediction of genomic susceptibilities for multiple cancer classes . *PNAS*. 2015 May 24;111(5):1921–6.
10. Oona F, Diana I, Thomas T. A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts. IEEE Trans. Knowl. Data Eng. 2016 Jan 25; 3(6):152-60.
11. Wei H, Sung-Kwun. Design of hybrid radial basis function neural networks (HRBFNNs) realized with the aid of hybridization of fuzzy clustering method (FCM) and polynomial neural networks (PNNs). Neural Netw.2015 June 5; 6(3):166–81.
12. Hilmi Berk Celikoglu. Application of radial basis functions and generalized regression neural networks in non-linear utility function specification for travel mode choice modeling. Math Comput Model. 2006 May 24; 44(5): 640–58.
13. Seema Singh, Sushmitha H. An Efficient Neural Network Based System for Diagnosis of Breast Cancer. Math Comput Model. 2014 June 14; 5(3):354-60.