VOL 3 (2019) NO 4

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Breast Cancer Prediction Using a Hybrid Data Mining Model

Elham Bahmani#, Mojtaba Jamshidi*, Abdusalam Abdulla Shaltooki**

# *Department of Computer Engineering, Malayer Branch, Islamic Azad University, Malayer, Iran*
**Department of Electrical, Computer and IT Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*
***Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq*

E-mail: bahmani.elham66@gmail.com, jamshidi.mojtaba@gmail.com, salam.abdulla@uhd.edu.iq

Abstract— **Today, with the emergence of data mining technology and access to useful data, valuable information in different areas can be explored. Data mining uses machine learning algorithms to extract useful relationships and knowledge from a large amount of data and offers an automatic tool for various predictions and classifications. One of the most common applications of data mining in medicine and health-care is to predict different types of breast cancer which has attracted the attention of many scientists. In this paper, a hybrid model employing three algorithms of Naive Bayes Network, RBF Network, and K-means clustering is presented to predict breast cancer type. In the proposed model, the voting approach is used to combine the results obtained from the above three algorithms. Dataset used in this study is called Breast Cancer Wisconsin taken from data sources of UCI. The proposed model is implemented in MATLAB and its efficiency in predicting breast cancer type is evaluated on Breast Cancer Wisconsin dataset. Results show that the proposed hybrid model achieves an accuracy of 99% and mean absolute error of 0.019 which is superior over other models.**

Keywords— Data mining, breast cancer prediction, hybrid model, RBF Network, Naive Bayes, K-means.

I. Introduction

Cancer is a type of disease in which body cell grow and proliferate uncontrollably. Cancer is usually named with the name of the involved part of the body. For instance, abnormal growth of cells in breast tissue is called breast cancer. Breast cancer symptoms include a lump in the breast, change in the shape of the breast, discharge from the nipple, skinning on some part of the skin. Patients in whom the disease metastases to other tissues, symptoms might include bone pain, intumescent lymph nodes, breath shortness or jaundice [1, 2].

Breast cancer is the leading cause of death in women between 40 to 55 years old and the second cause of death after pulmonary cancer. According to statistics of WHO, breast cancer was one of the most common cancers in 2012. More than 1.2 million women around the world are diagnosed with breast cancer, annually. Fortunately, in recent years, the death rate caused by breast cancer has reduced due to emphasis on diagnosis and treatment techniques. The main factor in this process is fast and correct diagnosis [1-3].

Today, using classification systems for medical diagnosis in increasing gradually. Classification systems can help reduce the error which might be caused by low experience experts and provides the possibility to investigate medical data in a shorter time with more details [3].

In this paper, three algorithms including Naive Bayes network, RBF network and K-mean clustering algorithm are combined to present an efficient predictor model for diagnosing the type of breast cancer. In the proposed model, raw data is loaded first and then they are pre-processed. Next, all data is divided into two training set and test set. Training data set is given to all the three algorithms in parallel so that three independent predictor models are created. Then, the test dataset is given to each model and the results are combined using the voting approach to obtain the final result.

The rest of this paper is organized as follows. Section II presents related work, breast cancer dataset, and the proposed model. Section III presents the simulation results. Finally, the paper is concluded in Section IV.

II. Material And Method

In this section, some existing works are studied first. Then, the breast cancer dataset used in this study is introduced. Finally, the proposed model is presented.

A. Related Work

In [4], an analysis has been presented on the survival of patients suffering from breast cancer using data mining methods. The pre-processing dataset includes 151886 records and 16 attributes. Three simple techniques including Bayes, feed-forward neural network, and C4.5 decision tree have been used. Results showed that C4.5 outperforms the other two techniques.

In [5], the performance of some common data mining algorithm in classifying breast cancer has been investigated. Results of experiments on two datasets show that among various data mining algorithms and software calculation methods, decision tree gives better results with an accuracy of 93.62%.

In [6], neural networks have been used to classify medical data sets. Backpropagation error method with variable learning rate and acceleration has been used to train the network. In order to analyze the performance of the network, various training data have been used as input of the network. In order to speed up the learning process, parallelization is performed in each neuron at all output and hidden layers. Results showed that the multi-layer neural network is trained faster than a single-layer neural network with high classification efficiency.

In [7], a model based on the J48 algorithm has been presented to predict recurring breast cancers. In this study, data of 908 patients suffering from breast cancer and 89 features from each patient has been used. Since there is a large data loss in this dataset, only information of 666 could have been used. Since there are missing values in the remaining records, these values are estimated through EM algorithm using SPSS.V20 as one of the pre-processing and data preparation phases.

In [8], it has been claimed that for more accurate analysis of breast cancer, all features of the dataset should be studied. In this study, a dataset of an institute in Portugal including a high percentage of unknown classified data (most clinical data of the patient is incomplete) has been investigated which is challenging in terms of complexity. In this study, KNN, decision tree, logistic regression, and SVM have been used for prediction. Results showed that KNN with an accuracy of 81% has offered better efficiency compared to the other algorithms.

In [9], a breast cancer diagnosis system has been presented based on simple logistic, RBF network, and RepTree. The data used in this study are provided by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. The data set has 10 attributes and 286 rows. Results showed that Simple Logistic with an accuracy of 74.47% offers better results compared to the other two algorithms.

In [10], the application of decision trees in predicting breast cancer has been investigated. It has also analyzed the performance of conventional supervised learning algorithms viz. Random tree, ID3, CART, C4.5, and Naive Bayes. Then, data is transferred to Rapid Miner data mining tool and breast cancer diagnosis for each sample in the test set is predicted with seven different algorithms which are Discriminant Analysis, Artificial Neural Networks, Decision Trees, Logistic Regression, Support Vector Machines, Naïve Bayes, and KNN. Results showed that Random tree achieves higher accuracy in cancer prediction.

In [11], seven different algorithms including Discriminant Analysis, Artificial Neural Networks, Decision Trees, Logistic Regression, Support Vector Machines, Naïve Bayes, and KNN have been evaluated in terms of the breast cancer diagnosis. The data about the patients have been taken from the UCI Machine Learning Repository thanks to Dr. William H. Wolberg from the University of Wisconsin Hospitals, Madison [13]. Results showed that Discriminant Analysis with an accuracy of 98.4% and Logistic Regression with an accuracy of 97.33% outperforms other algorithms.

B. *Breast Cancer Dataset*

The studied dataset is called Breast Cancer Wisconsin [13] taken from UCI data repositories. This dataset is collected by Dr. William at the University of Wisconsin. This dataset includes 699 samples and 11 attributes as presented in Table 1, values of all features are an integer. Output field of this dataset is class. All samples of this dataset are classified as benign and malignant. 458 cases are benign and 241 cases are malignant.

TABLE 1.
DATASET OVERVIEW

Attribute No.	Attribute name	Domain
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
11	Class	2 for benign, 4 for malignant

C. *The Proposed Method*

The main idea of the proposed method is to combine the k-means clustering algorithm with Naive Bayes and RBF to design a system for diagnosing breast cancer. The proposed system operates as follows:

I. First, data is loaded and pre-processing is performed to eliminate missing data.
II. The whole dataset is divided into a training set (70%) and test set (30%).
III. K-means, RBF, and Naïve Bayes algorithms are executed independently on the training dataset to create the predictive models. In this step, three predictor models are developed based on K-means, RBF, and Naïve Bayes algorithms.
IV. Each predictor model is evaluated through the test dataset and the results are stored in temporary memory.
V. Finally, the results obtained from these three models are combined and the final prediction results are presented to the user.

The flowchart of the proposed method is shown in Fig. 1. In the following, details of each step of the flowchart of the proposed system are described.

First, the data are extracted from UCI repositories and stores as a .csv file. Then pre-processing is performed on data. In the studied dataset, there are 16 missing values which should be initialized to execute clustering and classification algorithms. Here, a random initialization (from 1 to 10) are used to fill the empty fields of the dataset. In addition, the first feature of the dataset (sample code number) containing the ID of each record of the dataset cannot present any useful information for disease diagnosis. Thus, this feature is eliminated from the dataset. Then, the resulting dataset is divided into two training and test datasets. In this step of the proposed method, 70% of data (489 samples) is selected randomly for training and the remaining 30% (210 samples) is selected for the test.

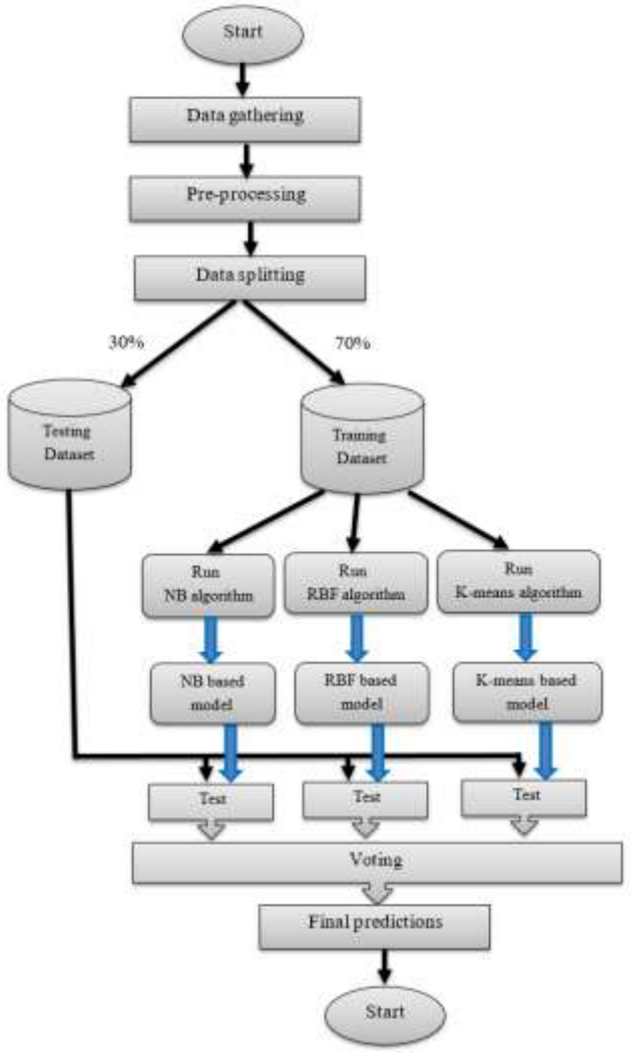

Fig. 1. Flowchart of the proposed method

In the next step, Naïve Bayes, RBF, and K-means are executed on the training dataset to create their predictor models. Here. It is required to review these three algorithms in brief:

Naïve Bayes: Bayes classification represents the membership of a sample to each class with a probability. Statistical concepts like mean, standard deviation, and histogram of features are used to generate rules. Bayes network is a graphical model describing the probable relationship between a set of variables. Structure of a Bayes network is an acyclic directed graph in which nodes represent random variables and edges represent a one-to-one relationship among variables. It can be easily constructed without requiring complicated iterative parameter estimation programs. That is, it can be used for a wide range of data and it performs astonishing. It might not be the best classifier for a specific application but its robustness can be relied on in most cases. There are a variety of algorithms from this family including Naive Bayes, AODE, and ByaseNet. In general, Bayesian algorithm operates under these two assumptions:

- Classification feature (or output field) should be nominal.
- There should be no missing values in the dataset.

In the Bayesian method, there are four various algorithms for estimating probability tables. The search operation is performed using K2 or TAN algorithm and several complicated methods based on hill-climbing, Tabu search method, and genetic algorithm. In this study, Simple Estimator is used as the estimator algorithm and the K2 algorithm is used for search [12].

RBF Network: This algorithm is an artificial neural network which employs radial basis functions as activation functions. The output of this network is a linear combination of radial basis functions for input parameters and neurons. This type of network is used in approximation function, time-series prediction, system control, and classification. This algorithm is called the radial function interpolation [12].

K-Means clustering: is the most common and simplest clustering method. Clustering is a class of unsupervised learning. Clustering is an automatic process in which samples are divided classes with similar members and each class is called a cluster. Therefore, a cluster contains a set of similar objects which are not similar to objects of other clusters. Different measures can be considered for similarity; for instance, distance measure can be used for clustering and consider the objects which are closer to each other as a cluster. This type of clustering is called distance-based clustering [12].

Despite the simplicity, K-means is a basic method for many other clustering methods (like fuzzy clustering). K-Means algorithm has an iterative process which estimates the following for a constant number of clusters:

- Obtaining points as the center of clusters. These points are the average points belonging to each cluster.
- Assigning each sample to a cluster in which the sample has a minimum distance to the center of the cluster.

In general, the K-Means algorithm operates as follows:

I. First, k points are selected as central points of the clusters.

II. Each sample is attached to the cluster which has a minimum distance from the center.

III. After attaching all samples to the clusters, a new point is calculated as the center of the cluster (average points belonging to each cluster).

IV. Steps 2 and 3 are repeated until no other change is made in the center of clusters.

In the proposed method, in order to construct a model based on the K-means algorithm, the training dataset is clustered into two “benign” and “malignant” clusters. When clustering is finished, centers of these two clusters are calculated and extracted for use in the test phase.

In the next step of the proposed method, predictor models constructed in the previous step are tested. In order to test each model, the test dataset is used. In this step, the test operation is performed separately for each model. To this end, all existing data of the test dataset is given to the predictor model to determine the type of cancer. Test operations of Naïve Bayes and RBF classification algorithms are identical and simple. That is, the test data is given to the model and output of the model is returned as the cancer type. But, for the K-means clustering algorithm, the following steps are performed:

I. calculating Euclidean distance of each test sample from the center of two benign and malignant clusters
II. classifying each test sample as benign or malignant based on minimum Euclidean distance from the center of the cluster.

In the last step, predictions obtained from these three models are combined with each other. In the combination step, voting and majority law are used. Thus, the prediction which has maximum votes is accepted as the final prediction of the proposed model. That is, if at least two models of the three mentioned models have predicted the same class for a testing sample, the voting result would be benign class.

III. Discussion And Simulation Results

In this section, the model proposed for predicting cancer type is evaluated. In order to construct and evaluate and the proposed models, MATLAB is used.

One of the common tools used for evaluating classification algorithms is to employ the confusion matrix. As can be seen in Table 2, the confusion matrix includes results of predictions of the classifier algorithm in 4 different classes including True Positive, False Negative, False Positive and True Negative.

TABLE II
CONFUSION MATRIX

		Predicted	
Observed		True	False
	True	TP	FN
	False	FP	TN

Considering the confusion matrix, the following measures can be defined and evaluated:

- **True Positive** refer to the positive samples that were correctly labelled by the classifier.
- **True Negative** refer to the negative samples that were correctly labelled by the classifier.
- **False Positive** is an error in data reporting in which a test result improperly indicates presence of a condition, such as a disease (the result is *positive*), when in reality it is not present.
- **False Negative** is an error in which a test result improperly indicates no presence of a condition (the result is *negative*), when in reality it is present.
- **Precision** is the fraction of retrieved instances that are relevant:

$$\frac{TP}{TP+FP} \tag{1}$$

- **Accuracy** is the proportion of true results (both true positives and true negatives) among the total number of cases examined:

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

- **Recall** is the fraction of relevant instances that are retrieved:

$$\frac{TP}{TP+FN} \tag{3}$$

- **F-Measure** combines precision and recall (harmonic mean):

$$\frac{2\times Recall\times Precision}{Recall+Precision}=\frac{2\times TP}{2\times TP+FP+FN} \tag{4}$$

- **Root Mean Squared Error (RMSE)** is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed. The RMSD represents the square root of the second sample moment of the differences between predicted values (p_i) and observed values (a_i) or the quadratic mean of these differences.

$$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}} \tag{5}$$

- **Mean Absolute Error (MAE)** measures how far predicted values (p_i) are away from observed values (a_i).

$$\frac{|p_1-a_1|+\cdots+|p_n-a_n|}{n} \tag{6}$$

The confusion matrix obtained from the proposed model is shown in Table 3. The test results showed that the proposed model has detected all test samples as benign correctly except one sample. In addition, it has failed in detecting malignant cancers just for one sample.

TABLE III
CONFUSION MATRIX OF THE PROPOSED METHOD

		Predicted	
Observed		**Benign**	**Malignant**
	Benign	162	1
	Malignant	1	46

Table 4 and Table 5 shows the results obtained from executing the proposed model for all evaluation measures. The results showed that the precision, recall, and F-measure of the proposed method for the benign class are 0.993 and their results for the malignant class are 0.978, 0.978, and 0.993, respectively. In addition, experiment results showed that the accuracy of the proposed model in the classification of samples is 0.99. Also, the results showed that the MAE and RMSE of the proposed model are 0.019 and 0.195, respectively.

TABLE IV
THE RESULTS OBTAINED FROM THE PROPOSED METHOD IN TERMS OF PRECISION, RECALL, AND F-MEASURE.

Measure	Benign	Malignant
Precision	0.993	0.978
Recall	0.993	0.978
F-Measure	0.993	0.993

TABLE V.
THE RESULTS OBTAINED FROM THE PROPOSED METHOD IN TERMS OF ACCURACY, RMSE, AND MAE.

Measure	Result
Accuracy	0.99
RMSE	0.195
MAE	0.019

Fig. 2 compared the proposed hybrid model and J48, Random Forest, RBF, Naïve Bayes, and Multilayer Perceptron (MLP) Neural Network in terms of the accuracy metric. As can be seen from the results, the accuracy of the proposed model is 0.99 while the accuracy of other compared algorithms is 0.97.

In addition, MAE and RMSE of the proposed model are compared with those of J48, Random Forest, RBF, Naïve Bayes, and MLP. Fig. 3 showed that the proposed model offers better results in terms of MAE with an error of 0.019 compared to other algorithms but in terms of RMSE, Fig. 4 showed that the RBF and Naïve Bayes with RMSE of 0.1837 and 0.1822, respectively, offer better results compared to the proposed model with RMSE of 0.195.

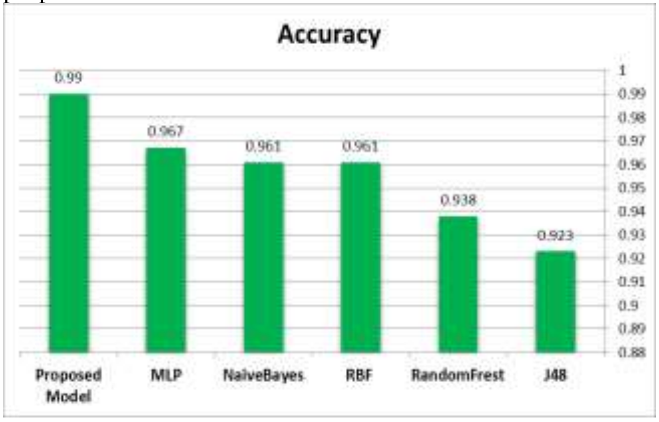

Fig. 2. Comparing the accuracy of the proposed model and some other common methods

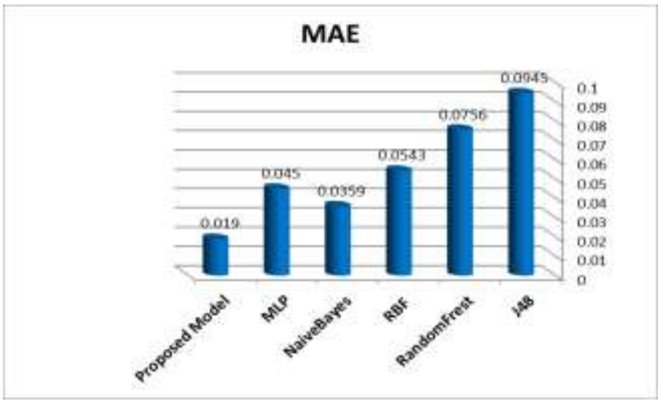

Fig. 3. Comparing the MAE of the proposed model and some other common methods

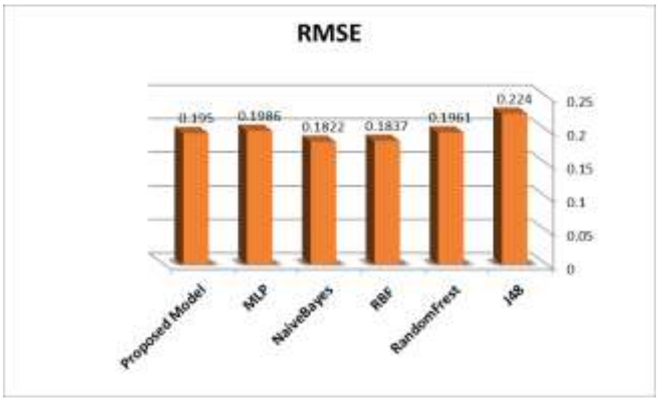

Fig. 4. Comparing the RMSE of the proposed model and some other common methods

IV. Conclusions

In this paper, a hybrid model is presented using K-means clustering, RBF and Naïve Bayes for diagnosing the type of breast cancer. In the proposed model, after extracting and preprocessing data, training and test datasets are created. K-means, RBF and Naïve Bayes algorithms are executed on the training dataset to create predictor models. Then, the test dataset is given as input to each model and obtained results are combined through voting. Evaluation results showed that the proposed method is more efficient than other models in terms of accuracy (0.99). In addition, it has offered better results compared to other algorithms in terms of MAE (0.019) but in terms of RMSE, RBF and Naïve Bayes algorithms offer better results with values of 0.1837 and 0.1822 compared to the proposed model with RMSE of 0.195.

References

[1] DeSantis, C., Ma, J., Bryan, L. and Jemal, A., 2014. Breast cancer statistics, 2013. CA: a cancer journal for clinicians, 64(1), pp.52-62.

[2] Harirchi, I., Karbakhsh, M., Kashefi, A. and Momtahen, A.J., 2004. Breast cancer in Iran: results of a multi-center study. Asian pacific journal of cancer prevention, 5(1), pp.24-27.

[3] Delen, D., Walker, G. and Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), pp.113-127.

[4] Gupta, S., Kumar, D. and Sharma, A., 2011. Data mining classification techniques applied for breast cancer diagnosis and prognosis. Indian Journal of Computer Science and Engineering (IJCSE), 2(2), pp.188-195.

[5] Kharya, S., 2012. Using data mining techniques for diagnosis and prognosis of cancer disease. arXiv preprint arXiv:1205.1923.

[6] Rani, K.U., 2010. Parallel approach for diagnosis of breast cancer using neural network technique. International Journal of Computer Applications, 10(3), pp.1-5.

[7] Kiani, B. and Atashi, A., 2014. A prognostic model based on data mining techniques to predict breast cancer recurrence. Journal of Health and Biomedical Informatics, 1(1), pp.26-31.

[8] García-Laencina, P.J., Abreu, P.H., Abreu, M.H. and Afonoso, N., 2015. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Computers in biology and medicine, 59, pp.125-133.

[9] Chaurasia, V. and Pal, S., 2017. Data mining techniques: to predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC, 3(1), pp. 10 – 22.

[10] Shajahaan, S.S., Shanthi, S. and ManoChitra, V., 2013. Application of data mining techniques to model breast cancer data. International Journal of Emerging Technology and Advanced Engineering, 3(11), pp.362-369.

[11] Senturk, Z.K. and Kara, R., 2014. Breast cancer diagnosis via data mining: performance analysis of seven different algorithms. Computer Science & Engineering, 4(1), p.35.

[12] Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

[13] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).