International Journal of Advanced Biotechnology and Research (IJBR)
ISSN 0976-2612, Online ISSN 2278–599X,
Vol-7, Issue-2, 2016, pp760-772
http://www.bipublication.com

Research Article

Classifiers' Accuracy Based on Breast Cancer Medical Data and Data Mining Techniques

Mohammed Abdullah Hassan Al-Hagery
Department of Computer Science,
Computer College, Qassim University, KSA
Email: hajry@qu.edu.sa

[Received-13/06/2016, Accepted-22/06/2016, Published-30/06/2016]

ABSTRACT

Knowledge Data Discovery (KDD) in Breast Cancer data set includes; data collection, data preprocessing, data mining, knowledge extraction, patterns validation, and results visualization. Where, this disease is considered as one of the deadly diseases in the world. The research problem is that there are many machine learning classifiers with different levels of accuracy when applied to Breast Cancer data. The highest accuracy of a classifier model depends upon the nature and the size of the data used to establish this model. To help physicians select the most accurate classifier algorithm to be applied using this kind of data. The paper's objective is to discover the most accurate classifier among all Data Mining classifiers based on Breast Cancer data set and the WEKA software. The results compared with previous results in other researches and show highly accurate results. It was found that the Bayes Net classifier model is the best classifier among the other types.

Keywords:Breast Cancer Disease, Medical Data Sets, Data Mining Models, Classification Accuracy, Decision Making, Quality of Data.

1. INTRODUCTION

The general objective of Data Mining is to extract, advanced information and knowledge from a data set, and then transform it into an understandable structure such as classes, clusters, or rules. These results can be used for further use. There is a necessity to employ the appropriate and recent techniques such as intelligent classifiers that direct doctors or specialists to select better algorithms for diagnoses and early prediction, in the field of Breast Cancer disease.In Data Mining, the strengths and weaknesses of each of the new techniques can be demonstrated on data by discussing applications of these techniques to current problems in all domains.

As a general technology, data mining can be applied to any type of data as long as the data are meaningful for a target application [1]. Data Mining can help in the area of medicine and pharmaceuticals to better determine which patients benefit from a given treatment. Data Mining provides information about cancer, including state of the art information on cancer screening, prevention, treatment, supportive care, and summaries of clinical trials. Many functions of Data Mining are applied in different areas cancer diseases depending on the type of problem and the way that the solution is applied. A number of software technologies are available for researchers to select appropriate techniques for their data analysis and data visualization. The most popular Data Mining analysis and visualization software include WEKA, Rabid

Miner, Tanagra, and Online Analytical Processing (OLAP). WEKA is a great Data Mining tool with a wide range of features that can be employed for several purposes [2]. Data Mining provides information and knowledge about cancer, including state of the art information on cancer screening, prevention, treatment, supportive care, and summaries of clinical trials [3]. Several works were accomplished in this field, especially in Breast Cancer Diagnosis by other methods such as using k-Nearest Neighbor with Different Distances and Classification Rules as in [4]. Various applications of Data Mining relevant to liver and Breast Cancer data sets were studied. The researchers explored that Data Mining techniques offer great promise to uncover patterns hidden in the data using different techniques such as clustering and classification that can help the clinicians in Decision Making [5].

Breast Cancer occurs when malignant tumors develop in the breast cells. These cells can spread by breaking away from the original tumor and entering blood vessels or lymph vessels, which branch into tissues throughout the body. When cancer cells travel to other parts of the body and begin damaging other tissues and organs, the process is called metastasis[6]. The research problem is founding a huge amount of medical data sets are not applied to assist the specialists and professionals in the medical field. The objective of this paper is to identify the more accurate classifier model according to the Breast Cancer data. It is focusing on the application of Data Mining techniques on a big sample of Breast Cancer data set in order to compare six common Data Mining classifiers and to determine the highest rank classifier.

2. Literature Review

Each kind of data reflects a special behavior of any model applied to these data. Therefore, the model that works strictly on the Breast Cancer data may not work the same accuracy on other types of data, such as environmental data, economic, educational or even with other disease data. So, there is a strong twinning relationship between the data and the behavior of the Data Mining models that are applied to these kinds of data. Medical data sets are considered most the famous application to mine. That data provide interesting patterns or rules for the future perspective [7]. A comparative analysis of Support Vector Machine accomplished in[8], itemploysBayesian classifier and other artificial neural network classifiers. Medical Data Mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis and for a better Decision Making. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form.

This collected data can then be integrated to form a hospital information system. Data Mining technology provides a user oriented approach to the novel and hidden patterns in the data [9]. Classification analysis is the organization of data in given classes and it is known as supervised classification. The classification uses given class labels to order the objects in the data collection. Classification is considered as an important task of Data Mining [10],[11]. In addition, Kharya [12] summarized various reviews and technical articles, on the Breast Cancer diagnosis and prognosis. The research focused on several researches that carried out using the Data Mining techniques to enhance Breast Cancer diagnosis and prognosis .

Among the various Data Mining classifiers and soft computing approaches, the decision tree is found to be the best predictor. The Bayesian network is also found to be a popular technique in medical prediction; in particular, it has been successfully utilized for Brest Cancer prognosis and diagnosis .Seven classification methods applied on a data set consists of 340 instances with 7 attributes and two classes. Their research has a similar objective of this research, although they applied a small data set and the results' accuracy in overall was low[13]. There are a number of studies applied machine learning techniques for survivability analysis. These studies have applied different approaches to the given problem and achieved several goals. For instance, [14] provided a solution for the medical practitioners to identify those patients who are in urgent need of chemotherapy before wasting a lot of time in conducting tests and then get the final diagnosis.

There are two stages, in the first stage; SVM and Decision Tree have been used to classify the patients into two classes; Benign and Malignant. And in the second stage the final clustering technique is applied in the pre-processed data sets of two classes Benign and Malignant to get three clusters Poor, Intermediate and good to determine whether the patient is in urgent need of chemotherapy with respect to the survival time of the patient .An evaluation of the accuracy of classification techniques was done based on the selected classifier algorithm. An important challenge in Data Mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications. The performance of SVM shows the high level compare with other classifiers. Hence SVM shows the concrete results with Breast Cancer disease of patient records. Therefore SVM classifier is suggested for diagnosis of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance [15] .An investigation process of the performance of different classification techniques was applied through analyzation of the Breast Cancer data available from the Wisconsin Breast Cancer (WBC) with the aim of developing accurate prediction of Breast Cancer using the Data Mining models. Three popular Data Mining methods were used; Sequential Minimal Optimization (SMO), K-Nearest Neighbor (KNN) classifier, and Bloom Filter Trees (BFT), the comparison carried out in WEKA software. The results show that SMO has higher prediction accuracy than other methods [16]. Two research works applied the same classification mechanism except the number of classifiers applied was different. The data set applied was the WBC database [13],[17].

Data Mining applied for Breast Cancer diagnosis and prediction using the Frequent Pattern algorithm in Association Rule Mining to conclude the patterns frequently found in Benign and Malignant patients. The decision tree algorithm used under the classification to predict the possibility of cancer in the context of the age. Three predictors attribute used, namely Age, Gender, Intensity of symptoms and one goal attributes, namely disease whose values indicate whether the corresponding patient have a certain disease or not [18]. Multivariate linear regression methods were used in breast Cancer diagnosis; logistic regression, KNN method, and discriminant analysis to identify the tumor type in a patient using WBC database [19]. The goal of the classification algorithms is to establish an accurate model amongst several Data Mining Models using a training data set, whose target class labels are known and then this model is used to classify unseen instances. The classification of Breast Cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors.

3. Breast Cancer

Breast Cancer is a dangerous disease, found among females all over the world, this disease being very mischievous to all women. The diagnosis of Breast Cancer disease is an important issue of Data Mining research in the medical field [20]. The Breast Cancer disease starts when the Clump thickness benign cells tend to be grouped in mono-layers, where the cancerous cells are grouped in multi-player. While in the uniformity of cell size/shape, the cancer cell tends to vary in shape and size. That is why these parameters are valuable in determining whether the cells are tumorous or not. In the case of marginal adhesion, the normal cells tend to stick together, where cancer cells tend to lose this ability, as shown in Figure 1. So, the loss of adhesion is a sign of malignancy [21].

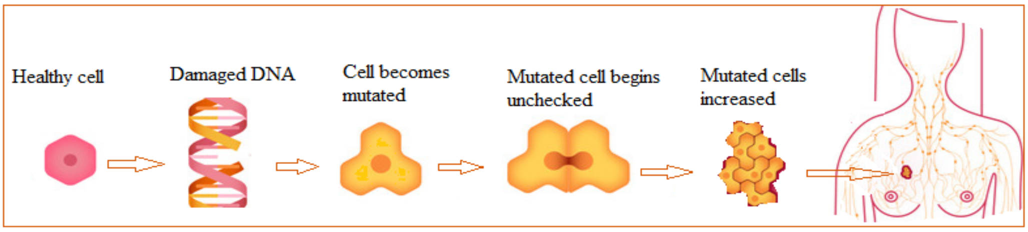


Fig. 1. Steps of breast cancer development, adapted from [6

4. Quality of Data

Real-world data is generally messy, incomplete and inconsistent. Data redundancies may also arise due to integration of data from several sources. The aim of this step is to handle these kinds of problems to improve the data quality. Furthermore, transforming and reducing data can help to improve the accuracy and efficiency of Data Mining function. The basic data preprocessing techniques that are required to apply in this research include three types [22],[23],[2], they are as follows:

(1) Data cleaning involves techniques for filling in missing values, smoothing out noise, handling Outliers, detecting, and removing redundant data.

(2) Data transformation puts the data into appropriate forms for mining when necessary.

(3) Data reduction is applied to reduce the data set to be mined, while dimension reduction techniques eliminate unnecessary attributes, data compression, numerosity reduction techniques that provide other forms of reduced data representations. In this research, for example, the attribute "patient code" is removed as the data reduction process required helping classifiers to learn quickly and to increase the accuracy of learning.

On the other hand, the quality tasks are considering product/process quality description, predicting quality, classification of quality, and parameter optimization [24]. One important issue relevant to the data quality is the data sources, the techniques used for data gathering, the style of data documenting, and archiving method. There are three main elements that affect the data quality; consistency, completeness, and accuracy. Incomplete, inaccurate, and inconsistent data are commonplace properties of large real-world DB and data warehouses. There are various possible reasons for inaccurate data, for example, getting incorrect attributes' values. The instruments used for data collection may be faulty. Another reason is the human or computer errors that occur on data entry. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information.

5. Research Methodology

The research methodology is encompassing a number of processes that should be followed to accomplish the research goals. It includes getting the Breast Cancer data, identify suitable tools, and determine the required techniques, data preprocessing, mining process, knowledge extraction (classifier results), results interpretation, and results evaluation, identify the most accurate model to be used by physicians to discover the type of Breast Cancer whether malignant or benign, the methodology chart is illustrated in Figure 2 (track A and track B).

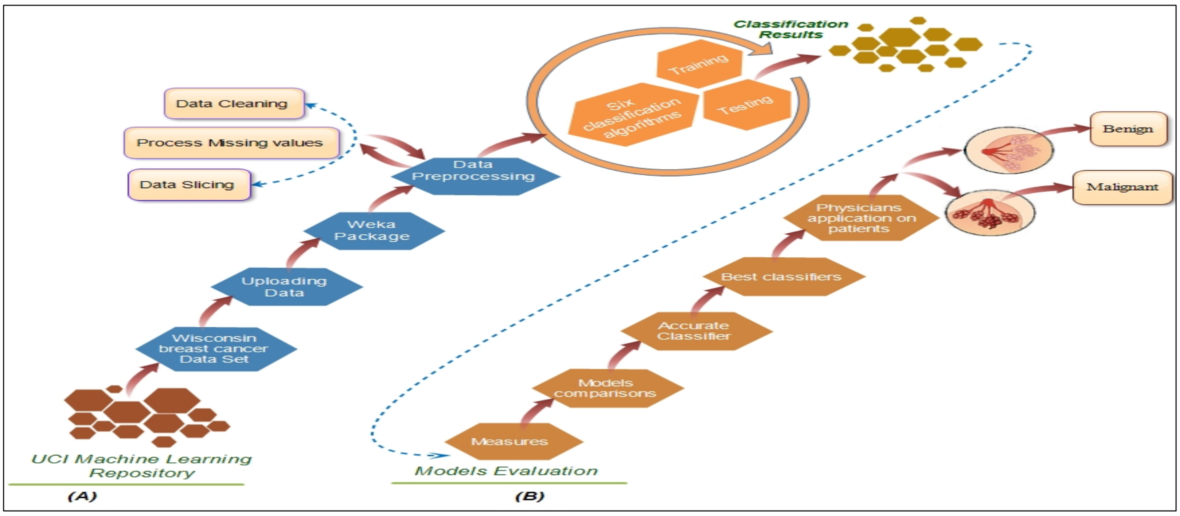
5.1 Proposed Tools and Techniques

Application of Data Mining tool can bring significant benefits. This tool could be applied where there are enough data sets, among which the valuable patterns are hidden [25].

Six algorithms incorporated in the WEKA software package were selected for the application, also using the WEKA in this research because it is supported in several standard Data Mining tasks and the WEKA is a collection of machine learning algorithms for solving real-world Data Mining problems. It was developed in Java as open source software and it can be worked on any platform.

To apply any classifier, there are many steps and tasks that should be accomplished. These steps include; preparation of Breast Cancer data set, data cleaning, data preprocessing (removal of noise/Outliers, or using the best strategies for handling missing values), data slicing, data selection, finding the most useful features to represent the data depending on the goal of the research, data transformation to be easy for the processing, choosing the classification algorithms, interpret the extracted knowledge, and fix the best results. The tasks are shown in Figure 2.

Fig. 2. Methodology steps



The track (A) shows the data pre-processing, maximizing the data quality, and running of 90 experiments on the WEKA, producing the classification results. The track (B) presents the rest of the methodology.

5.2 *Research Data*

The Wisconsin Breast Cancer (WBC) Database was used for analysis and comparison. It consists of 10 attributes plus the class attribute in 699 instances. This type of data was collected from real samples and consist of visually assessed nuclear features of fine needle aspirates (FNAs) taken from patients' breasts. Each sample has been assigned a 9-dimensional vector; each value can be located in the interval 1 to 10, with value 1 corresponding to a normal state and 10 in a most abnormal state. The last attribute is the (class type) designates whether the sample is benign or malignant. Malignancy is determined by taking a sample tissue from the patient's breast and performing a biopsy on it. A benign diagnosis is confirmed either by biopsy or by periodic examination, depending on the patient's choice. Table 1, demonstrate a definition and domain of Wisconsin Breast Cancer Database attributes and Figure 3 shows a sample of this data.

Table 1. Wisconsin Breast Cancer Database attributes

Attribute	Definition	Domain
Sample code number	The number of patient	ID
Clump Thickness	The cancerous cells are often grouped in multi-layer	1 – 10
Uniformity of Cell Size	The cancer cells tend to vary in size	1 – 10
Uniformity of Cell Shape	The cancer cells tend to vary in shape	1 – 10
Marginal Adhesion	The cancer cells tend to lose stick together	1 – 10
Single Epithelial Cell Size	The size relates to the uniformity of cell size	1 – 10
Bare Nuclei	A term used for nuclei that is not surrounded by cytoplasm	1 – 10
Bland Chromatin	Describes a uniform “texture” of the nucleus seen in benign cells	1 – 10

Normal Nucleoli	Small structures seen in the nucleus	1 – 10
Mitoses	Nuclear division plus cytokines and produce two identical daughter cells during prophase	1 – 10
Class	Designates whether the sample is benign or malignant	2 Benign, 4 Malignant

Table 2 displays the first 12 instances in the Wisconsin Breast Cancer Database. Each instance contains 10 attributes. The first is the Clump Thickness and the last is the class type Benign or Malignant.

Table 2: The First 12 Instances in Wisconsin Breast Cancer Database

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
5	1	1	1	2	1	3	1	1	2
5	4	4	5	7	10	3	2	1	2
3	1	1	1	2	2	3	1	1	2
6	8	8	1	3	4	3	7	1	2
4	1	1	3	2	1	3	1	1	2
8	10	10	8	7	10	9	7	1	4
1	1	1	1	2	10	3	1	1	2
2	1	2	1	2	1	3	1	1	2
2	1	1	1	2	1	1	1	5	2
4	2	1	1	2	1	2	1	1	2
1	1	1	1	1	1	3	1	1	2
2	1	1	1	2	1	2	1	1	2

5.3 Data Pre-processing

The objective of this task is to maximize the quality of the data set to get the smaller ratio of the incorrectly classified cases in the final results. Pre-processing is an essential step that is used to transform the raw data into a format that makes it possible to apply the Data Mining techniques and also to improve the quality of data. In this research, the data cleaning tasks are used for pre-processing because there are many missing values in a number of data values, where some instances contain a single missing (i.e., unavailable) value denoted by "?" as a missing value in the attribute "Bare Nuclei" for a 7 patients, as shown in Table 3. The filtering process was applied as an important step before the classification to remove these instances to increase the quality of learning during the classification processes. So, the final size of the data set is 687 instances out of 699.

Table 3: Examples of Missing Values for Some Patients

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
8	4	5	1	2	?	7	3	1	4
6	6	6	9	6	?	7	8	1	2
1	1	1	1	1	?	2	1	1	2
1	1	3	1	2	?	2	1	1	2
1	1	2	1	3	?	1	1	1	2
5	1	1	1	2	?	3	1	1	2
3	1	4	1	2	?	3	1	1	2

Around 90 experiments were conducted in this research on the data set. The data distributed into different sizes of data (nine slices) to get different cases to compare the classifier results. These slices were denoted by (T_r, Ts). TheT_r represents the training sets where, T_r = {76,152, 228, 304, 344, 380, 456, 532, 608} andT_s isthe testing sets also consists of nine slices, T_s = {611, 535, 459, 383, 343, 307, 231, 155, 79}, where (T_r, T_s) ={(76, 611), (152, 535), (228, 459), (304, 383), (344, 343) (380, 307), (456, 231), (532, 155), (608, 79)} see the first column in Table 5.

6. Experimental Results

The selected classifiers including Multi-layer Perceptron (MLP), RBF in Neural network (NN), Naïve Bayes (NB) and Bayesian Net in Bayesian, J48, and Logic Model Tree (LMT) as a decision tree. Each one of these classifiers has its own characteristics. The 10 attributes were considered in the classification with different sizes of data samples, each sample is bigger than those applied by [17]. Classifiers performance evaluating in this research measured according to some important metrics as applied in [1]. Table 4 shows the accuracy of the results of two previous researches. The first accomplished [13], denoted by PR1, and the second research carried out by [15], denoted by PR2.

Table 4: The Results of all Classifiers with Different Training Sets

Training Set	Testing set (PR1)	J48 (PR1)	J48 (PR2)	LMT (PR1)	Bayes Net (PR1)	Naïve Bayes (PR1)	MLP (PR1)	MLP (PR2)	RBF (PR1)
D1: 10% =34	90%	53.28%	79.31%	55.59 %	58.55 %	57.89 %	57.23 %	79.31%	55.92%
D2: 20% =68	80%	57.03%	64.91%	59.62 %	61.11 %	59.62 %	61.85 %	64.91%	60.37 %
D3: 30% =102	70%	54.85%	69.77%	56.54%	59.91 %	59.91%	62.02%	63.95%	67.51 %
D4: 40% =136	60%	57.14 %	71.93%	69.45 %	65.02 %	64.03 %	69.95%	68.42%	67.98 %
D5: 50% =170	50%	64.5 %	70.63%	71.00%	65.08%	65.58 %	65.68 %	67.83%	65.08 %
D6: 60% =204	40%	60.74 %	70.35%	70.37%	61.48 %	61.48 %	66.66 %	69.19%	65.18 %
D7: 70% =238	30%	66.33 %	71.50%	59.40 %	67.32%	67.32 %	69.30 %	71%	68.31 %
D8: 80% =272	20%	63.23 %	73.36%	69.17 %	66.17 %	66.17 %	70.58%	73.80%	54.70%
D9: 90% =306	10%	79.41 %	69.26%	74.47%	76.47 %	76.47%	79.41 %	70.43%	76.47%
Average		62.90%	71.22%	65.07%	64.57%	64.82%	66.96%	69.87%	64.61%

In this research, only six of these classifiers were used and its results are shown in Table 5.The correctly classified results of the first experiment are shown in the first row in this table, as follows: 93.61% for the J48 classifier, 96.23% for the LMT, 97.21% for the Bayes Net, 96.23% for the Naïve Bayes, 94.10% for the MLP, and 92.14% for the RBF. All results got in this research (Table 5) were compared with the previous results (Table 4). The comparison of the results will be discussed in the next section.

Table. 5: Results of all Classifiers with Different Training Sets

Training Set	Testing Set	J48	LMT	Bayes Net	Naïve Bayes	MLP	RBF
D1: 10% = 76	611	93.61%	96.23%	97.21%	96.23%	94.10%	92.14%
D2: 20% = 152	535	94.96%	95.52%	97.20%	96.08%	95.52%	94.58%
D3: 30% = 228	459	94.56%	95.21%	97.60%	96.30%	95.21%	96.52%
D4: 40% = 304	383	95.06%	96.10%	97.40%	95.84%	95.84%	94.28%
D5: 50% = 344	343	93.87%	95.62%	97.37%	95.62%	95.04%	90.96%
D6: 60% = 380	307	93.85%	95.47%	97.08%	95.79%	95.46%	96.44%
D7: 70% = 456	231	93.58%	96.58%	97.43%	96.58%	95.29%	94.44%
D8: 80% = 532	155	95.56%	98.10%	97.46%	95.56%	96.20%	98.10%
D9: 90% = 608	79	93%	98.68%	97.36%	97.36%	98.68%	97.36%
Average		94.01%	96.18%	97.24%	96.05%	95.48%	95.77%

Another method to find out the best results with a high accuracy, the Cross-validation (CF) was applied to the whole data set. The CF method used with two values of folds 10 and 16, which satisfy the best results amongst the other folds. Table 6 shows the comparison of all classifiers based on using 10 and 16 Folds using the whole data set.

Table 6: Accuracy of all Classifiers with Different Folds

Training Set	J48	LMT	Bayes Net	Naïve Bayes	MLP	RBF
CF (10 Folds)	95.77%	96.65%	97.37	96.36	95.77	96.21
CF (16 Folds)	94.17	96.79	97.37	96.21	95.92	96.65

In addition, five experiments accomplished for each classifier based on using different groups of Breast Cancer attribute, as follows:

G1=3, the attributes = {Clump Thickness, Bare Nuclei, Bland Chromatin}. G2= 4, the attributes = {Clump Thickness, Bare Nuclei, Bland Chromatin, Uniformity of Cell Size}. G3= 5, the attributes = {Clump Thickness, Bare Nuclei, Bland Chromatin, Uniformity of Cell Size, Uniformity of Cell Shape}. G4= 6, the attributes = {Clump Thickness, Bare Nuclei, Bland Chromatin, Uniformity of Cell Size, Uniformity of Cell Shape, Uniformity of Cell Size}. G5= 7, the attributes = {Clump Thickness, Bare Nuclei, Bland Chromatin, Uniformity of Cell Size, Uniformity of Cell Shape, Uniformity of Cell Size, Normal Nucleoli}. Table 7 shows the accuracy of all classifiers using different number of attributes in a previous research. Table 7 illustrates the results of previous work when applying different groups of attributes (3, 4, 5, 6, and 7).

Table 7: Accuracy of All Classifiers Using Various Attributes Number (PR1)

No. of Features	J48	LMT	Bayes Net	Naïve Bayes	MLP	RBF
3	76.47%	76.47%	76.47%	76.47%	76.47%	73.52%
4	82.35%	59.40%	82.35%	82.35%	76.47%	79.41%
5	79.41%	82.35%	76.47%	76.47%	88.23%	79.41%
6	79.41%	73.52%	76.47%	76.47%	91.17%	82.35%
7	76.41%	74.47%	76.47%	76.47%	79.41%	79.41%
Average	78.81%	73.24%	77.65%	77.65%	82.35%	78.82%

Table 8 displays the results’ accuracy of this research based on different groups of attributes (3, 4, 5, 6, and 7).

Table 8: Accuracy of All Classifiers Using Different Number of Features

No. of Features	J48	LMT	Bayes Net	Naïve Bayes	MLP	RBF
3	92.30%	94.01%	94.01%	94.87%	94.01%	94.01%
4	94.87%	95.72%	95.72%	96.15%	95.72%	96.15%
5	94.87%	96.15%	97%	96.58%	97.43%	96.15%
6	93.16%	92.30%	95.72%	94.44%	96.15%	93.50%
7	93.16%	94.44%	95.72%	94.44%	95.29%	91.88%
Average	93.67%	94.52%	95.63%	95.30%	95.72%	94.34%

7. RESULTS DISCUSSION

The results will be discussed in the following two sub-sections, depend on the number of features applied and the size of training and testing data set.

7.1 *Using all Features and Different Sizes of Data*

In this section of the results, nine features of Breast Cancer were applied in each experiment. In general, the Research results (R-Results) show a high accuracy for all experiments when compared with that of previous works.

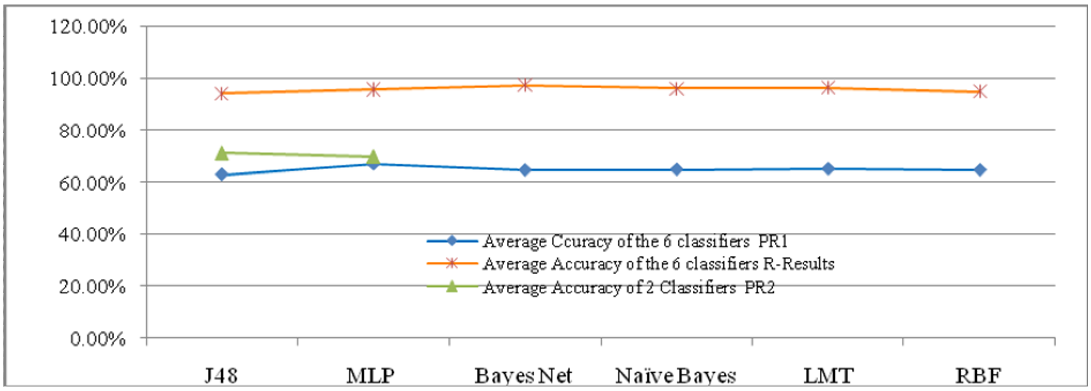


Fig. 3. Results Comparison for J48 Classifier

Figure 3 shows the results of the J48 classifiers in 2 previous works that achieved an accuracy value lower than 80% given by [13],[17], while the accuracy of R-Results are ranging from 93% to 95%.The results of the LMT classifier presented in Figure 4 show better accuracy in R-Results than those generated using the same classifier in other research work [13]. The average accuracy was 73.2% against 94.52%.

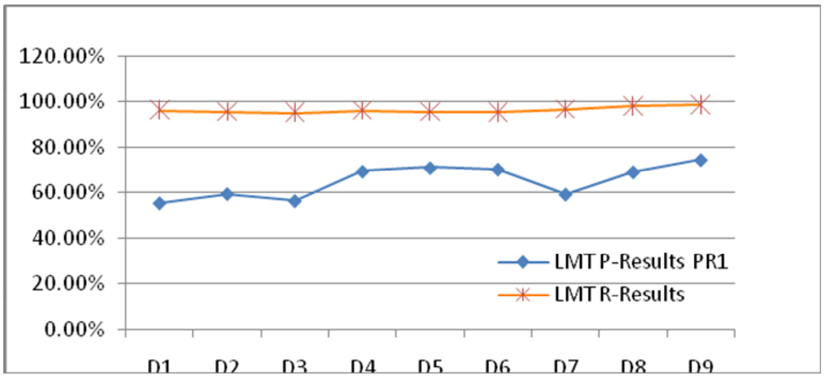


Fig. 4. Results Comparison for LMT Classifier

Although, the results carried out by [13],[15] using the MLP classifier illustrate good accuracy beside the results carried out by this research. The average accuracy in this research is equal to 94% against 62.90% given in[13]& 71.22% in[15] as shown in Table 4 and Figure 5.

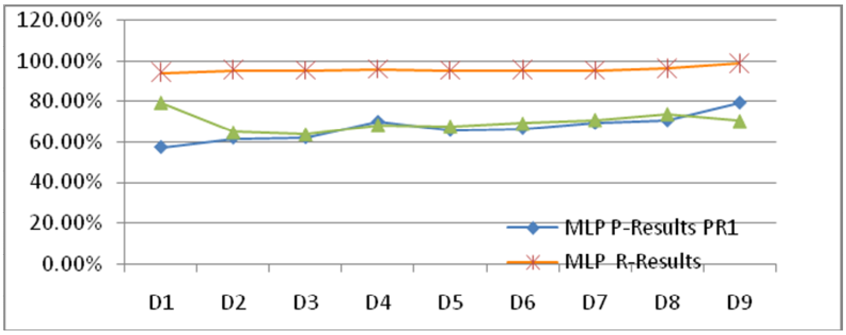


Fig. 5. Results Comparison for MLP Classifier

As seen in Figure 6, the results of the Bayes Net classifier are the highest results ever seen between the results of the other classifiers. These results seem to be the best with an average of 97.24% of the nine samples of data, and based on all data attributes without any exception, furthermore, when using 10folds and 16 folds, the results accuracy of this classifier increased from 97.24% to 97.37%.

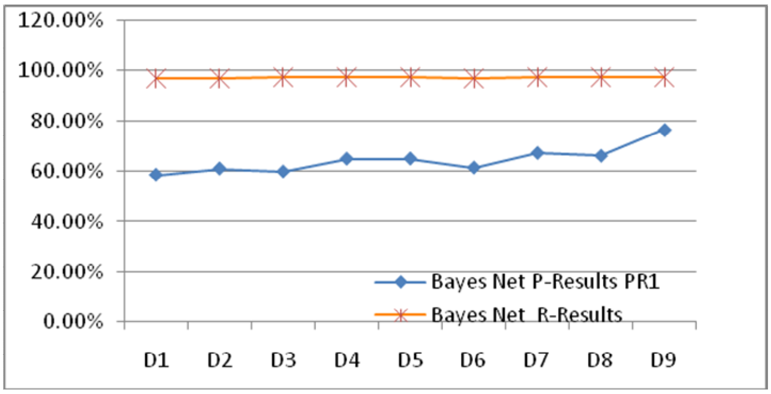


Fig. 6. Results Comparison for Bayes Net Classifier

The Naïve Bayes classifier present highly accurate results in this research when compared with results carried out by [13], as in Figure 7. The established model provides high accurate results whenever applying a big data set during the training process, as in sample D9. The classification accuracy (as an average value of all samples applied by this classifier) in this research is equal to 96.05% against 64.82% given by [13].

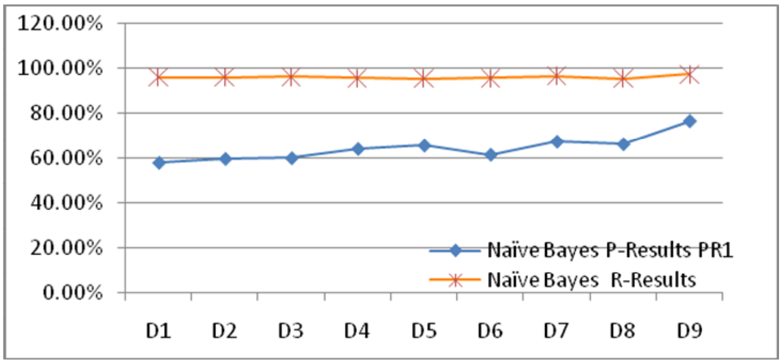


Fig. 7. Results of Naïve Bayes Classifier

The R-Results of the Naïve Bayes classifier present more accurate results, as in Figure 8. The established model provides the most accurate results whenever applying a big data set during the training process, as in sample D9. The classification accuracy average value of all samples applied by this classifier is equal to 96.05% in this research against 64.82% in[13].

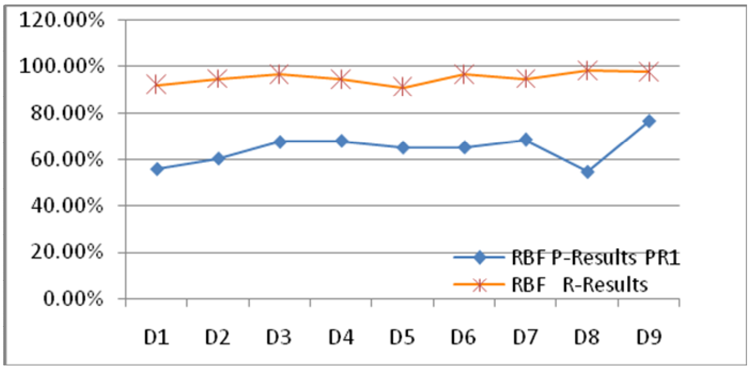


Fig. 8. Results Comparison for RBF Classifier

Figure 9 shows a big difference between the R-Results and the other results based on the average value of the results accuracy of all classifiers. The curve in this figure shows that R-Results are the highest when compared with all classifiers accomplished by [13] and with two classifiers; J48 and MLP measured by [15]. This research gives highly accurate results, and the most accurate classifier amongst all the six types was the Bayes Net, which, achieved 97.24% and 97.37%, as in Table 5 and 6.

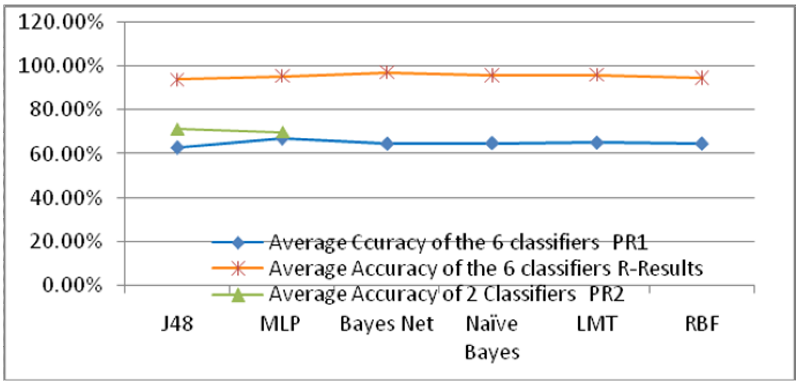


Fig. 9. Average Accuracy of the Six Classifiers

7.2 Using Various Number of Features and all Data

Each experiment carried out above using the six classifiers covered all features of Breast Cancer data without any exception. In this section, the experiments were repeated on the six classifiers, but with a various number of features in each experiment, the selected features were {3, 4, 5, 6, 7} as explained in the Experimental Results. The whole data set divided into two groups; training data (66%) and testing data set (34%). The research results achieved higher accuracy than previous researches. In this type of experiment, it found that the MLP classifier shows the best results amongst other classifiers. It has achieved 97.43% when the number of features =5 against 88.23 in[13] with the same number of features, as shown in Table 7 and 8 and in Figure 10.

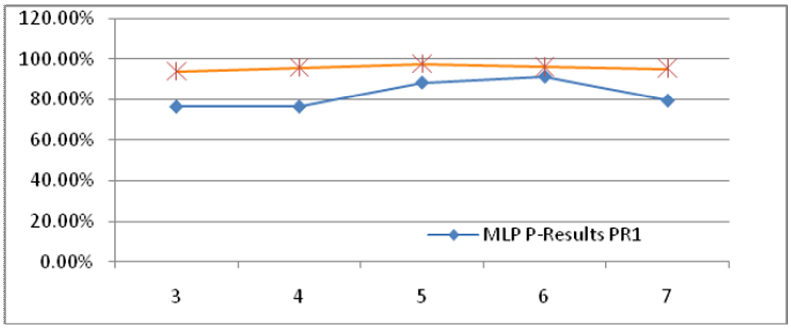


Fig. 10. Results of MLP Classifier with Different Attributes

On the other hand, the average accuracy measured for all classifiers with the whole data set and various numbers of features show in this research higher accuracy of results than those measured in[13]. In this research, all classifiers illustrate accurate results, but the most accurate classifier is the MLP, which achieved an average of 95.72% against 82.35%, as in Table 5 and Figure 11.

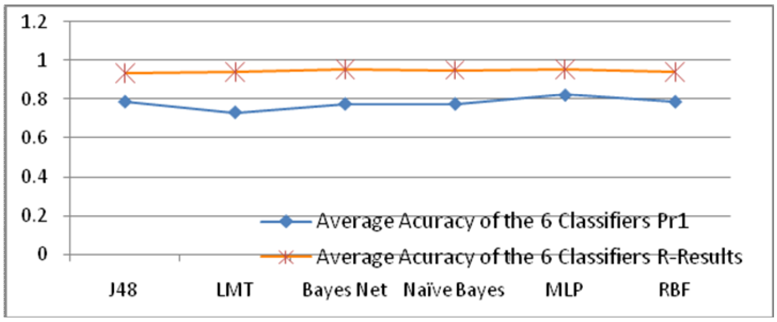


Fig 11. Accuracy of all Classifiers with Different Attributes

8.CONCLUSIONS

In this paper, six types of classification algorithms, including J48, Bayes Net, Naïve Bayes, MLP, RBF, and LMT were applied to the Breast Cancer data set. The results achieved in this research show perfect result with a high degree of accuracy when compared with previous results. Different types of experiments carried out to cover all available cases. In general, the results got in this research achieved the perfect results. All of the results show that the Bayes Net is the best classifier when using different sizes of data and with all Breast Cancer disease features also, it achieved the best results when applying the whole data with different number of attributes. In the last case, the MLP is giving equivalent results to the Bayes Net classifier, although the Bayes Net still according to all cases is the most accurate classifier based on the data sets of breast cancer. The results that were measured according to the data and the methodology of this research were much better than those measured according to the previous researches. Based on the discussed results of this research, the approach allows the

Breast Cancer physicians' committee to support its decision making to diagnose the patient's condition at any time, to determine the type of Breast Cancer (Benign or Malignant). As results of this research, it is highly recommended physicians of this disease to use the Bayes Net classifier model for prediction and diagnosis purpose. This research can be extended by finding the real relationship between individual sets of Breast Cancer features and the infection degree to identify the most features have a high effect on the development of the malignant tumor. Implement the top two accurate classifiers; Bayes Net and MLP in an online system to be available for use by doctors and/or patients.

REFERENCES

1. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques Third Edition. 2011.
2. S. A. Medjahed, T. A. Saadi, and A. Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules," *Int. J. Comput. Appl.*, vol. 62, no. 1, pp. 1–5, 2013.
3. M. R. Koutonin, "The Best Data Mining Tools You Can Use for Free in Your Company," (2013), http://www.siliconafrica.com/the-best-data-minning-tools-you-can-use-for-free-
4. S. A. Medjahed, T. A. Saadi, and A. Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules," Int. J. Comput. Appl., vol. 62, no. 1, pp. 1–5, 2013.
5. M. Thangaraju and R. Mehala, "Novel Classification based approaches over Cancer Diseases," Int. J. Adv. Resh. in Comp. and Commu. Eng. vol.4, no.3, pp. 294–297, 2015. DOI: 10.17148/IJARCCE.2015.4370 294.
6. National Breast Cancer Foundation (NBCF), (2015), http://www.nationalbreastcancer.org/what-is-cancer.
7. D. Delen, G. Walker, and A. Kadam, Predicting Breast Cancer survivability: a comparison of three Data Mining methods. Journal of Artificial Intelligence in Medicine, Elsevier Science Publishers 34(2) (2005) 113-127. DOI: 10.1016/j.artmed.2004.07.002
8. H. You and G. Rumbe, "Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data," Int. J. Interact. Multimed. Artif. Intell., vol. 1, no. 3, p. 5, 2010.
9. J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," Int. J. Comput. Appl., vol. 17, no. 8, pp. 43–48, 2011.
10. A. S. A. Alghamdi, "Efficient Implementation of FP Growth Algorithm-Data Mining on Medical Data," Int. J. Comput. Sci. Netw. Secur., vol. 11, no. 12, pp. 7–16, 2011.
11. M. Ramageri, "Data Mining Techniques and Applications," Indian J. Comput. Sci. Eng., vol. 1, no. 4, pp. 301–305, 2010.
12. S. Kharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease," Int. J. Comput. Sci. Inf. Technol., vol. 2, no. 2, pp. 55–66, 2012.
13. P. M. Barnaghi, V. A. Sahzabi, and A. A. Bakar, "A Comparative Study for Various Methods of Classification," Int. Conf. Inf. Comput. Networks, vol. 27, no. Icicn, pp. 62–66, 2012.
14. R. Yadav, "Chemotherapy Prediction of Cancer Patient by using Data Mining Techniques," Int. J. Comput. Appl., vol. 76, no. 10, pp. 28–31, 2013.
15. G. R. Kumar, G. A. Ramachandra, and K. Nagamani, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques," Ijiet, vol. 2, no. 4, pp. 139–144, 2013.
16. V. Chaurasia and S. Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques," Int. J. Innov. Res. Comput. Commun. Eng., vol. 2, no. 1, pp. 2456–2465, 2014.
17. A. Lebbe, S. Saabith, E. Sundararajan, and A. A. Bakar, "Comparative Study on Different Classification Techniques for Breast Cancer Dataset," Int. J. Comput. Sci. Mob. Comput.,

vol. 3, no. 10, pp. 185–191, 2014.

18. J. Majali, R. Niranjan, V. Phatak, and O. Tadakhe, "Data Mining Techniques For Diagnosis And Prognosis Of Cancer," Ijarcce, vol. 4, no. 3, pp. 613–615, 2015.
19. S. Zarei, M. Aminghafari, and H. Zali, "Application and Comparison of Different Linear Classification Methods for Breast Cancer Diagnosis," Int. J. Analy. Phar. and Biom. Sci., vol. 4, no. 2, pp.123-129, 2015.
20. S. Ghosh, S. Mondal, and B. Ghosh, "A comparative study of Breast Cancer detection based on SVM and MLP BPN classifier," First. Int. Conf Automa. Cont. Energy and Sys (ACES): IEEE, Hooghy, pp. 1-4, 2014, DOI:10.1109/ACES.2014.6808002.
21. G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers," Int. J. Comp. Info. Tech. vol.1, no.1, pp. 36-43, 2012.
22. P. Giudici, Applied Data Mining: Statistical methods for business and industry, (New York: J. Wiley, 2003).
23. D. Pyle, S. Editor, and D. D. Cerra, Data Preparation for Data Mining, vol. 17. 1999.
24. G. Köksal, I. Batmaz, and M. C. Testik, "A review of data mining applications for quality improvement in manufacturing industry," Expert Syst. Appl., vol. 38, no. 10, pp. 13448–13467, 2011.
25. M. Chen and P. S. Yu, "Data Mining: An Overview from Database Perspective," IEEE Trans. Knowledge and Data Eng., vol.3, no.6, pp.866-883, 1996.