World Applied Sciences Journal 36 (6): 742-750, 2018
ISSN 1818-4952
© IDOSI Publications, 2018
DOI: 10.5829/idosi.wasj.2018.742.750

Improving Breast Cancer Diagnosis by Applying Data Mining Classification Techniques

Nancy Awadallah Awad

Sadat Academy for Management Sciences, Department of Computer and Information Systems, Egypt

Abstract: Breast cancer has become the main reason of death in women in the world. The use of data mining techniques has a vital role in the breast cancer diagnosis and prediction. This study aimed to compare between four data mining classification techniques (Decision Tree, Decision Table, Naïve Bayes and SMO) on breast cancer datasets to predict benign / malignant and to reduce the ratio of false-positives (FP). To carry out this task, researcher used data mining algorithms, on WEKA tool. Findings explored that SMO was the best classifier achieved accuracy with 97% and 0.037 (FP) in 0.11 second.

Key words: Data Mining Techniques · Breast Cancer Datasets · Decision Tree · Decision Table · Naïve Bayes

INTRODUCTION

Cancer is defined as abnormal cells divide without control and is able to spread to other tissues. Cancer cells can spread to other parts of the body through the blood and lymph systems. Breast cancer is a common cancer among women, a vital role of information technology is in the early detection and diagnosis of breast cancer [1]. Successful treatment of patients is based on early prediction and diagnosis of breast cancer - on the other hand, to distinguish between benign / malignant breast tumors, it requires a reliable and accurate procedure [2].

Data Mining is set of techniques used in several fields to give meaning to the available data, its objective is to fit data to a model such as predictive and descriptive. The difference between the two models that the predictive model is the process creating a classification model from a set of examples, called the training set, which belongs to a set of classes. While, descriptive model is to describe the general or special features of a set of data in a concise manner [2].

Researcher in this paper focused on applying data mining classification techniques on breast cancer dataset to predict if the tumor is benign or malignant. These techniques are Decision Tree, Decision Table, Naïve Bayes and SMO which applied the process of classification on complete dataset, replace missing values and remove tuples filters in WEKA tool.

Several breast cancer studies are needed to apply data mining methods with different objectives and data sources [3]. Lu *et al*. [3] transformed patient data into useful information by using data mining techniques. Five models have been evaluated to get valuable patterns to help understand cancer outcomes.

Delen *et al*.[2] presented several prediction models based on a large cancer database in the USA by using Artificial Neural Networks, logistic regression and decision trees [2]. Chang and Liou [4] used data set of breast cancer patients of Wisconsin University. They focused on the artificial neural network model which achieved 0.9502 (sensitivity 0.9628, specificity 0.9273), decision tree model which achieved 0.9434 (sensitivity 0.9615, specificity 0.9105), logistic regression model which achieved 0.9434 (sensitivity 0.9716 and specificity 0.9482 and genetic algorithm model which achieved 0.9878 (sensitivity 1, specificity 0.9802). They explored that the accuracy of the genetic algorithm was highest model [5].

Jurca *et al*. [6] identified new potential biomarkers for breast cancer by integrating the social network analysis and text mining.

Diz *et al.* presented comparison between two breast cancer datasets to detect the best breast density classification in predicting benign/malignant tumor [7].

Ruilan and Zhixin [8] set experimental results divided into two categories, namely benign cluster and malignant cluster by setting Wisconsin breast cancer data

Corresponding Author: Nancy Awadallah Awad, Department of Computer and Information Systems, Egypt.

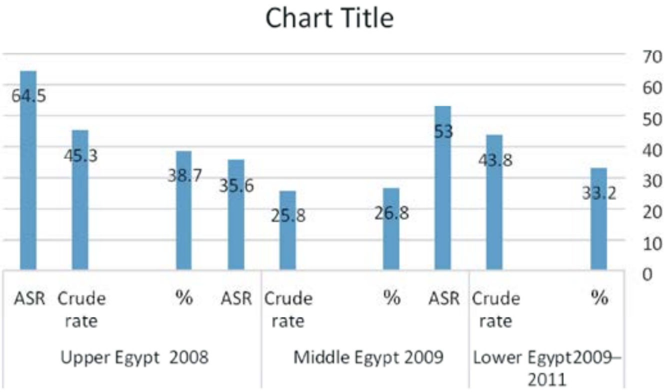


Fig. 1: Proportions and incidence rates of breast cancers in the 3 regions of Egypt [11]

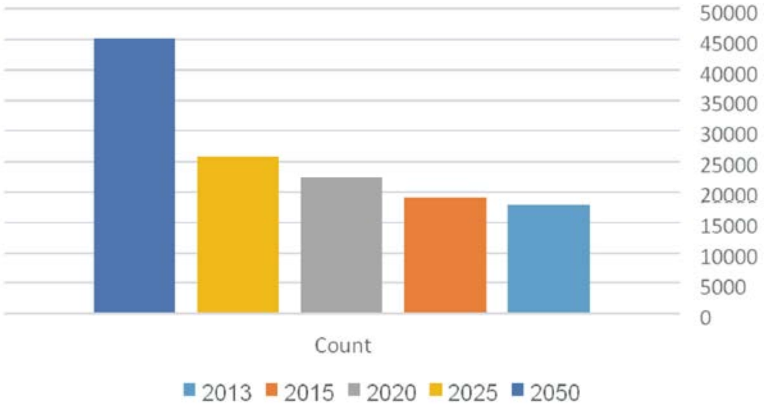


Fig. 2: Estimated number of breast cancer, Egypt 2013-2050 [11]

for analysis and using K-means to do SMOe clustering operation. Compare the clustering results with the known experimental data, to arrive at the results.

Lee *et al.* [9] get three prognostic groups: Good, Poor and Intermediate from clustered 253 breast cancer patients. Their results suggested that patients with chemotherapy (represented by the Intermediate group) is better than the same group whose without chemotherapy in but the patients should not receive chemotherapy represented by the Good group.

Michael *et al*.[10] examined by using a classification tree the surgical treatment factors from breast cancer. They explored that for predicting patient choice the classification trees perform well as logistic regression and the selected tree can inform clinicians' advice to patients.

Tran and Le [11] used indications of patients to build predictive models which can classify patients' breast cancer categories if (benign or malignant).

Ibrahim *et al.* [12] applied the age-specific incidence rates of each registry to the population of the 3 regions (Lower, Middle and Upper Egypt). This step is represented to get incident cases in different age groups by using data of Egypt last census. ASR: Age standardized rate.

MATERIALS AND METHODS

Data Mining Classification Techniques: One of the most important task of data mining is classification technique. It is a supervised learning as targets are predefined via applying this technique prediction assign patients to non- cancerous ”*benign*” group or cancerous ”*malignant*” group. The data mining classification tasks which commonly used can be classified into the decision tree, Naïve Bayes, neural networks, Support Vector Machine, Decision table, Fuzzy sets, Genetic algorithms [13 -15].

Decision Tree: In decision tree starting at the root node and follow down until we reach a leaf and this to classify a data item. A decision is made when a terminal node is approached.

Decision Tree in WEKA J48 (C4.5) J48 can handle numeric attributes it divided and conquer algorithm, convert tree to classification rules. Attribute Selection which has the highest value of Information gain will be used [15].

Decision Table: It likes decision trees as it is used for prediction in classification models [16]. A decision table consists of a hierarchical table in which each entry in a higher level table gets broken down by the values of a pair of additional attributes to form another table [17].

Naïve Bayes: A Naive Bayes classifier is based on applying Bayes' theorem with strong (naive) independence assumptions [18, 19]. Naive Bayes classifier supposed that the presence (or absence) of an attribute of a class is unrelated to the presence (or absence) of any other feature [20, 21].

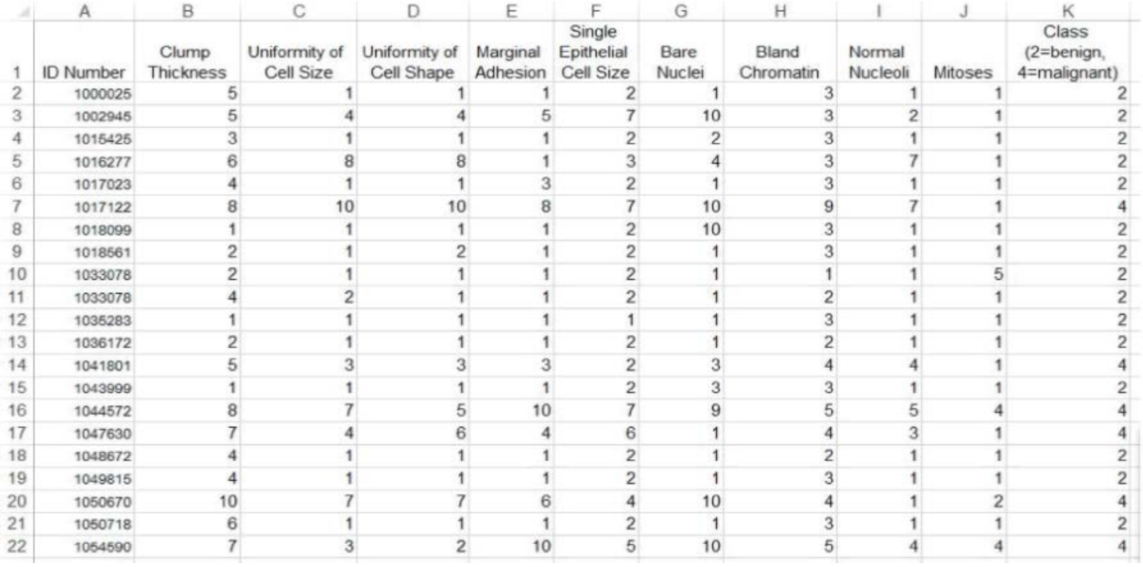
Support Vector Machine (SVM): It is the most used technique for machine learning tasks. Through a sequential optimization algorithm (SMO) function on WEKA tool, (SVM) can be used easily [22, 23].

Data Set: Breast cancer prediction in this paper was performed by using Wisconsin Breast Cancer Database. This dataset consists of 11 variables and 699 observations, the 11 variables are Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and Class.

Table 1: Data attributes [11]

#	Variable	Attribute	Domain
1	ID	Sample code number	ID number
2	Clump_Thickness	Clump thickness	1–10
3	Cell_Size	Uniformity of cell size	1–10
4	Cell_Shape	Uniformity of cell shape	1–10
5	Adhesion	Marginal adhesion	1–10
6	Epi_Cell_Size	Single epithelial cell size	1–10
7	Nuclei	Bare nuclei	1–10
8	Chromatin	Bland chromatin	1–10
9	Nucleoli	Normal nucleoli	1–10
10	Mitoses	Mitoses	1–10
11	Class	Target variable	2-Benign, 4-Malignant

Table 2: Sample of the dataset



	A	B	C	D	E	F	G	H	I	J	K
1	ID Number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class (2=benign, 4=malignant)
2	1000025	5	1	1	1	2	1	3	1	1	2
3	1002945	5	4	4	5	7	10	3	2	1	2
4	1015425	3	1	1	1	2	2	3	1	1	2
5	1016277	6	8	8	1	3	4	3	7	1	2
6	1017023	4	1	1	3	2	1	3	1	1	2
7	1017122	8	10	10	8	7	10	9	7	1	4
8	1018099	1	1	1	1	2	10	3	1	1	2
9	1018561	2	1	2	1	2	1	3	1	1	2
10	1033078	2	1	1	1	2	1	1	1	5	2
11	1033078	4	2	1	1	2	1	2	1	1	2
12	1035283	1	1	1	1	1	1	3	1	1	2
13	1036172	2	1	1	1	2	1	2	1	1	2
14	1041801	5	3	3	3	2	3	4	4	1	4
15	1043999	1	1	1	1	2	3	3	1	1	2
16	1044572	8	7	5	10	7	9	5	5	4	4
17	1047630	7	4	6	4	6	1	4	3	1	4
18	1048672	4	1	1	1	2	1	2	1	1	2
19	1049815	4	1	1	1	2	1	3	1	1	2
20	1050670	10	7	7	6	4	10	4	1	2	4
21	1050718	6	1	1	1	2	1	3	1	1	2
22	1054590	7	3	2	10	5	10	5	4	4	4

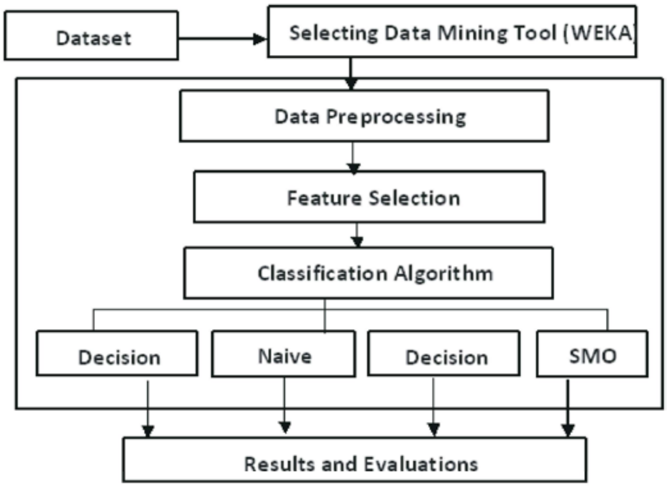


Fig. 3: The framework of using data mining classification techniques for breast cancer detection

These attributes are as follows:

Table 2 describes sample of the dataset which its attributes are described in Table 1.

Research Design: Knowledge discovery in databases (KDD) represented data mining process in discovering hidden information from a huge datasets [13]. The characteristic of this process is to be capable of prediction a series of supposed risks, form descriptive data [14].

The next figure illustrates the framework of this research which describes the performance evaluation cycle and consists of several phases, firstly: Extracting dataset, secondly: selecting data mining tool, thirdly: data pre-processing, fourthly: feature selection, fifthly: classification algorithm (Decision Table, Naïve Bayes, Decision tree and SMO) and finally get results.

Data Preprocessing: Data in the real world is dropped into 3 categories: Incomplete – Noisy - Inconsistent In this study, researcher used filters (Remove tuple with missing values) and (Replace missing values) to avoid outlier's data.

Quality decisions must be based on quality data measures which are accuracy, completeness, consistency, timeliness, believability, value added and accessibility.

Feature Selection: The objective of this step is:

- Improving the prediction performance.
- Providing a faster and more cost-effective predictors.

Transforming a dataset by removing some of its columns to detect quality of a classifier.

Classification Techniques: In this study, researcher applied several data mining classifiers such as decision table, Naïve Bayes, decision tree and SMO algorithms. The reasons to use classification are:

- Building accurate and efficient classifiers.
- Classification improve predictions compared to unclassified data.

Evaluation Methods: WEKA is a collection of machine learning algorithms for data mining tasks.

- The algorithms can either be applied directly to a dataset or called from your own Java code.
- WEKA contains tools for data preprocessing, classification, regression, clustering, association rules, visualization and feature selection.

RESULTS AND DISCUSSION

The study applied 10-fold cross validation. This test mode work as following:

- Split data into 10 equal sized pieces
- Train on 9 pieces and test on remainder [5, 16]

Evaluation: The sample data were divided into the next 4 categories:

- True positive (TP) = No. of positive samples which predicted correctly.
- False negative (FN) = No. of positive samples which predicted wrongly.

Table 3: Using Decision table algorithm for breast cancer detection

Test Set		Performance Evaluation			
Classifier	Class	TP Rate	FP Rate	Precision	F-Measure
Decision Table					
Complete Data Set	Benign	0.963	0.066	0.965	0.964
	Malignant	0.934	0.037	0.930	0.932
	Weighted Avg.	0.953	0.056	0.953	0.953
Replace missing va lues	Benign	0.950	0.075	0.960	0.955
	Malignant	0.925	0.050	0.907	0.916
Weighted Avg		0.941	0.066	0.942	0.941
Remove the tuple with values missing	Benign	0.963	0.066	0.965	0.964
	Malignant	0.934	0.037	0.930	0.932
Weighted Avg		0.953	0.056	0.953	0.953

Table 4: Using Naive Bayes algorithm for breast cancer detection

Test Set		Performance Evaluation			
Classifier	Class	TP Rate	FP Rate	Precision	F-Measure
Naive Bayes					
Complete Data Set	Benign	0.952	0.025	0.986	0.969
	Malignan	0.975	0.048	0.914	0.944
	Weighted Avg	0.960	0.033	0.962	0.960
Replace missing values	Benign	0.952	0.025	0.986	0.969
	Malignan	0.975	0.048	0.914	0.944
Weighted Avg		0.960	0.033	0.962	0.960
Remove the tuple with missing values	Benign	0.952	0.025	0.986	0.969
	Malignan	0.975	0.048	0.914	0.944
	Weighted Avg	0.960	0.033	0.962	0.960

- False positive (FP) = No. of negative samples which predicted as positive wrongly.
- True negative (TN) = No. of negative samples which predicted correctly.

So the sensitivity or the true positive rate (TPR) is defined by TP / (TP + FN), the specificity or the true negative rate (TNR) is defined by TN / (TN + FP), the accuracy is defined by (TP + TN) / (TP + FP + TN + FN).

Table 3 describes the results of applying 3 datasets (Complete, replace missing values filter and remove the tuple with missing values) by using Decision table algorithm. Classification techniques presents two classes Benign and Malignant. Performance evaluation used for this algorithm are (TP rate, FP rate, Precision, F-Measure). The complete dataset is similar to remove the tuple with missing values filter (dataset) which are better than replace missing values (dataset) as the precision is 95.3% and the error ratio (FP) 0.056.

Table 4 describes the results of applying 3 datasets (Complete, replace missing values filter and remove the tuple with missing values) by using Naive Bayes algorithm. It presented two classes Benign and Malignant and performance evaluation used for this algorithm are (TP rate, FP rate, Precision, F-Measure).

The 3 datasets were similar to each other, which describes that the precision is 96.2% and the error ratio (FP) is 0.033.

Table 5 describes the results of applying 4 datasets (Complete, replace missing values filter, remove the tuple with missing values and Info Gain Attribute) by using decision tree algorithm. It presented two classes Benign and Malignant and performance evaluation used for this algorithm are (TP rate, FP rate, Precision, F-Measure). Replace missing values filter is better than others datasets as it achieves the precision 95.2% and error ratio (FP) is 0.053.

Fig. 4 illustrates the results from Decision Tree J48 via WEKA tool.

Table 6 describes the results of applying 3 datasets (Complete, replace missing values filter and remove the tuple with missing values) by using SMO algorithm. It presented two classes Benign and Malignant and performance evaluation used for this algorithm are (TP rate, FP rate, Precision, F-Measure).

Complete dataset is better than others datasets as it achieves the precision 97% and error ratio (FP) is 0.034.

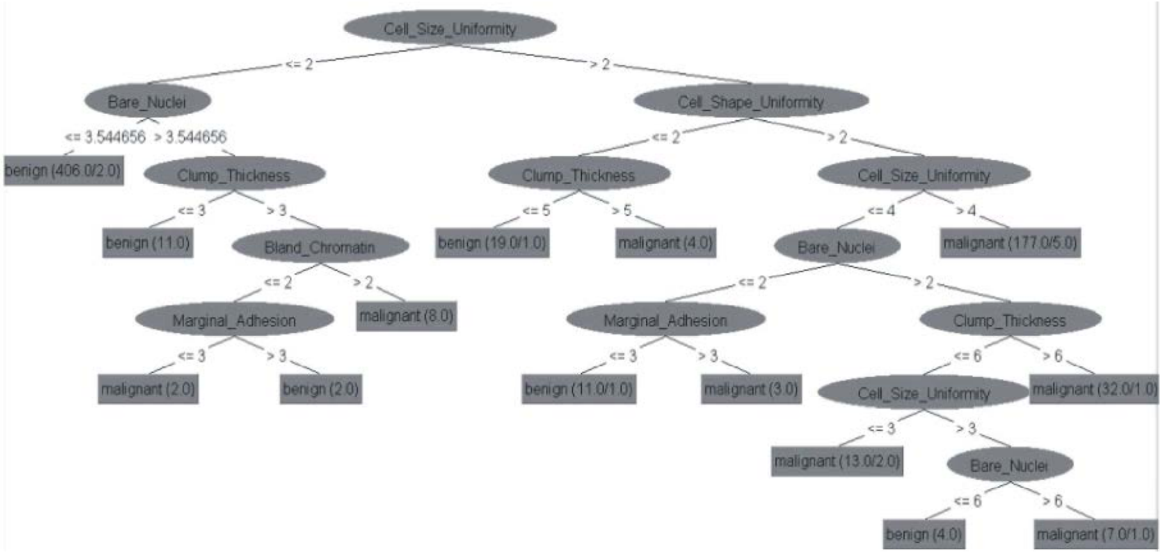


Fig. 4: Applying Decision tree (J48) After Using Replace missing values Filter

Table 5: Using Decision tree algorithm for breast cancer detection

Test Set		Performance Evaluation			
Classifier	Class	TP Rate	FP Rate	Precision	F-Measure
Decision Tree (J48)					
Complete Data Set	Benign	0.956	0.075	0.961	0.958
	Malignant	0.925	0.044	0.918	0.921
	Weighted Avg.	0.946	0.064	0.946	0.946
Replace missing values	Benign	0.956	0.058	0.969	0.963
	Malignant	0.942	0.044	0.919	0.930
	Weighted Avg.	0.951	0.053	0.952	0.951
Remove the tuple with missing values	Benign	0.956	0.075	0.961	0.958
	Malignant	0.925	0.044	0.918	0.921
	Weighted Avg.	0.946	0.064	0.946	0.946
Info Gain Attribute	Benign	0.956	0.075	0.961	0.958
	Malignant	0.925	0.044	0.918	0.921
	Weighted Avg.	0.946	0.064	0.946	0.946

Table 6: Using SMO algorithm for breast cancer detection

Test Set		Performance Evaluation			
Classifier	Class	TP Rate	FP Rate	Precision	F-Measure
SMO					
Complete Data Set	Benign	0.974	0.037	0.980	0.977
	Malignant	0.963	0.026	0.951	0.957
	Weighted Avg	0.970	0.034	0.970	0.970
Replace missing values	Benign	0.974	0.041	0.978	0.976
	Malignant	0.959	0.026	0.951	0.955
Weighted Avg		0.969	0.036	0.969	0.969
Remove the tuple with missing values	Benign	0.974	0.041	0.978	0.976
	Malignant	0.959	0.026	0.951	0.955
	Weighted Avg	0.969	0.036	0.969	0.969

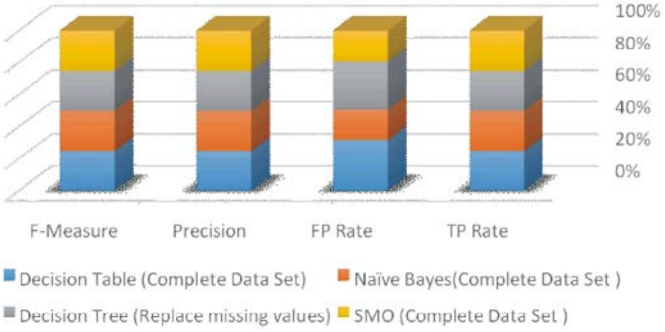


Fig. 5: Classification Techniques Results

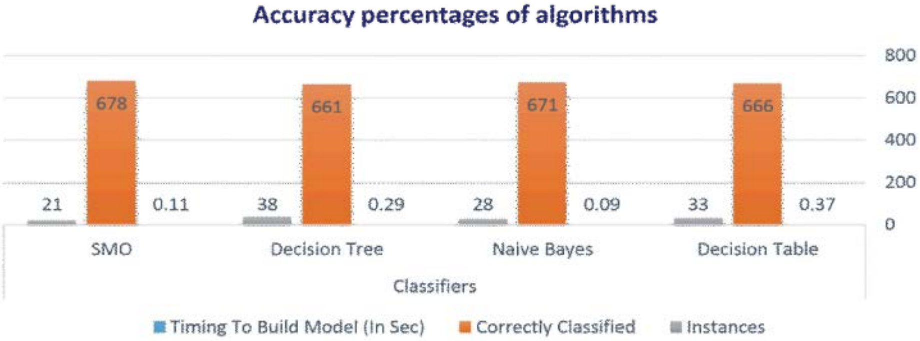


Fig. 6: Accuracy percentages of algorithms

Table 7: Evaluation criteria for Decision Table, Naïve Bayes, Decision tree and SMO algorithms for breast cancer detection

	Classifiers			
Evaluation Criteria	Decision Table	Naive Bayes	Decision Tree	SMO
Timing To Build Model (In Sec)	0.37	0.09	0.29	0.11
Correctly Classified Instances	666	671	661	678
Incorrectly Classified Instances	33	28	38	21
Accuracy (%)	95.3 %	96.2 %	94.6 %	97 %

Table 8: Confusion Matrix for Decision Table, Naïve Bayes and Decision tree algorithm

Classifiers	Benign	Malignant	Class
Decision Table	441	17	Benign
	16	225	Malignant
Naive Bayes	436	22	Benign
	6	235	Malignant
Decision Tree	438	20	Benign
	18	223	Malignant
SMO	446	12	Benign
	9	232	Malignant

Fig. 5 illustrates the performance evaluation (F-Measure, Precision, FP rate, TP rate) for four classification techniques (Decision table, Naïve Bayes, Decision Tree and SMO).

Table 7 describes the results of applying four classification techniques (Decision table, Naïve Bayes, Decision Tree and SMO) on complete dataset and presents evaluation criteria such as (Timing to Build Model (In Sec), Correctly Classified Instances, Incorrectly Classified Instances and Accuracy).

The results presents that the SMO algorithm is the best one than decision table decision tree and Naïve Bayes algorithm, as it achieve the accuracy ratio 97%, the time to build the model is 0.11 second and the incorrectly classified instances is 21 (less than other classifiers).

Fig. 6 illustrates the accuracy percentages of (Decision table, Naïve Bayes, Decision Tree and SMO) algorithms.

Table 8 illustrates the matrix for each algorithm toward each class (Benign, Malignant).

CONCLUSIONS

In this paper, researcher applied data mining techniques on breast cancer dataset to classify, predict benign / malignant tumor to reduce the ratio of false-positives. Performance evaluation cycle of this research consists of several phases, firstly: Extracting dataset, secondly: selecting data mining tool, thirdly: data pre-processing, fourthly: feature selection, fifthly: classification algorithm (Decision Table, Naïve Bayes, Decision tree and SMO) and finally get results.

The results of this study showed that SMO algorithm is the best classifier used as achieved accuracy 97% with 0.037 false positive and take 0.11 second to build a model.

REFERENCES

1. Lu, W., Z. Li and J. Chu, 2017. A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning" , Computers in Biology and Medicine, 83: 157-165.
2. Delen, D., G. Walker and A. Kadam, 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med., 34(2): 113-127.
3. Lu, J., A. Hales, D. Rew and M. Keech, 2015. Data Mining Techniques in Health Informatics: A Case Study from Breast Cancer Research, Springer International Publishing Switzerland 2015, pp: 56-70, DOI: 10.1007/978-3-319-22741-2_6
4. Chang W.P. and D.M. Liou, 2008. Comparison of three data mining techniques with genetic algorithm in the analysis of breast cancer data. J. Telemed Telecare., 9: 1-26.
5. Ming. Liou, D. and W. Pin Chang, 2015. Applying Data Mining for the Analysis of Breast Cancer Data, Methods in molecular biology, Springer Science+Business Media New York, DOI 10.1007/978-1-4939-1985-7_12.
6. Jurca, G., O. Addam, A. Aksac, S. Gao, T. Özyer, D. Demetrick and R. Alhajj, 2016. Integrating text mining, data mining and network analysis for identifying genetic breast cancer trends, BMC Research Notes , DOI 10.1186/s13104-016-2023-5.
7. Diz, J., Goreti Marreiros and Alberto Freitas, Applying, 2016. Data Mining Techniques to Improve Breast Cancer Diagnosis, J. Med. Syst., 40: 203. doi: 10.1007/s10916-016-0561-y .
8. Zhang Ruilan and Feng Zhixin, 2014. The Data Mining of Breast Cancer Based-on K-Means, Advances in Computer Science and Its Applications, DOI: 10.1007/978-3-642-416743_155
9. Lee, Y.J., O.L. Mangasarian and W.H. Wolberg, 2003. Survival-time classification of breast cancer patients, Comput. Optim. Appl., 25(1-3): 151-166.
10. Michael A. Martin, Ramona Meyricke, Terry O'Neill and Steven Roberts, 2006. Mastectomy or breast conserving surgery? factors affecting type of surgical treatment for breast cancer: a classification tree approach. BMC Cancer 6, 98.
11. Tran, T. and U. Le, 2018. Predicting Breast Cancer Risk: A Data Mining Approach. In the Proceedings of the 6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6), IFMBE Proceedings, vol 63, Springer, Singapore, pp 223-228, https://doi.org/10.1007/978-981-10-4361-1_37.
12. Ibrahim, A.S., H.M. Khaled, N.N.H. Mikhail, H. Baraka and H. Kamel, 2014. Cancer Incidence in Egypt: Results of the National Population-Based Cancer Registry Program, J. Cancer Epidemiol., doi: 10.1155/2014/437971 .
13. Tseng, W.T., W.F. Chiang, S.Y. Liu, J. Roan and C.N. Lin, 2015. The application of data mining techniques to oral cancer prognosis. J. Med. Syst. 39(5): 59, doi:10.1007/s10916-015-0241-3.
14. Malucelli, A., A. Stein Junior, L. Bastos, D. Carvalho, M.R. Cubas and E.C. Paraíso2010. Classification of risk micro-areas using data mining. Rev Saude Publica, 44(2): 292-300, doi:10.1590/S0034-89102010000200009.
15. Gupta, S., D. Kumar and A. Sharma, 2011. Data Mining Classification Techniques for Breast Cancer Daignosis and Prognosis, Indian Journal of Computer Science and Engineering (IJCSE), 2(2) Apr-May.
16. Kohavi, R., 1995. The Power of Decision Tables, Proceedings of the European Conference on Machine Learning, Lecture Notes in Artificial Intelligence 914, Springer Verlag, Berlin, Heidelberg, NY, pp: 174-189.
17. Becker, B.G., 1998. Visualizing Decision Table Classifiers, Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258), DOI: 10.1109/INFVIS.1998.729565.

18. Senturk, Z.K. and R. Kara, 2014. Breast Cancer Diagnosis Via Data Mining: Performance Analysis Of Seven Different Algorithms" Computer Science & Engineering: An International Journal (CSEIJ), Vol. 4, No. 1, February DOI : 10.5121/cseij.2014.4104.
19. Bueno, G., N. Vállez, O. Déniz, P. Esteve, M.A. Rienda, M. Arias and C. Pastor, 2011. Automatic breast parenchymal density classification integrated into a CADe system. Int J Comput Assist Radiol Surg., 6(3): 309-318, doi:10.1007/s11548-010-0510-z.
20. Witten, I.H. and E. Frank , 2005. Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Francisco. ISBN: 0120884070.
21. Castella, C., K. Kinkel, M.P. Eckstein, P.E. Sottas, F.R. Verdun and F.O. Bochud, 2007. Semiautomatic mammographic parenchymal patterns classification using multiple statistical features. Acad Radiol. 14(12): 1486-1499, doi:10.1016/j.acra.2007.07.014.
22. Lesniak, J., R. Hupse, R. Blanc, N. Karssemeijer and G. Székely, 2012. Comparative evaluation of support vector machine classification for computer aided detection of breast masses in mammography. Phys. Med. Biol., 57(16): 5295-5307.
23. Pérez, N., M.A. Guevara, A. Silva, I. Ramos and J. Loureiro, 2014. Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection. In: Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on. IEEE, pp: 209-217.