Original
ISSN 2413-0516

Breast cancer decisive parameters for Iraqi women via data mining techniques

Suhad Faisal Behadili,[a*] Mustafa S. Abd,[a] Iyden Kamil Mohammed,[b] and Maha Mohammed Al-Sayyid[c]

[a]Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq.
[b]Biomedical Department, Al-Khawarzmy Engineering College, University of Baghdad, Baghdad, Iraq.
[c]Oncology Teaching Hospital, Medical City, Baghdad, Iraq.
*Correspondence to Suhad Faisal Behadili (email: suhad.behadili@scbaghdad.edu.iq).
(Submitted: 12 December 2018 – Revised version received: 05 January 2019 – Accepted: 29 January 2019 – Published online: 26 April 2019)

Objective This research investigates Breast Cancer real data for Iraqi women, these data are acquired manually from several Iraqi Hospitals of early detection for Breast Cancer. Data mining techniques are used to discover the hidden knowledge, unexpected patterns, and new rules from the dataset, which implies a large number of attributes.
Methods Data mining techniques manipulate the redundant or simply irrelevant attributes to discover interesting patterns. However, the dataset is processed via The Waikato Environment for Knowledge Analysis platform. The OneR technique is used as a machine-learning classifier to evaluate the attribute worthy according to the class value.
Results The evaluation is performed using a training data rather than cross validation. The decision tree algorithm J48 is applied to detect and generate the pattern of attributes, which have the real effect on the class value. Furthermore, the experiments are performed with three machine-learning algorithms J48 decision tree, simple logistic, and multilayer perceptron using tenfold cross-validation as a test option, and the percentage of correctly classified instances as a measure to determine the best one from them. As well as, this investigation used the iteration control to check the accuracy gained from the three mentioned above algorithms. Hence, it explores whether the error ratio is decreasing after several iterations of algorithm execution or not.
Conclusion It is noticed that the error ratio of classified instances are decreasing after 5–10 iterations, exactly in the case of multilayer perceptron algorithm rather than simple logistic, and decision tree algorithms. This study realized that the TPS_pre is the most common effective attribute among three main classes of examined dataset. This attribute highly indicates the BC inflammation.
Keywords CA 15-3, CEA, breast cancer, saliva, MLP, SLR, J48, data mining, OneR, Iraq

Introduction

Breast Cancer (BC) is the leading cause of death in women in developing countries, and a second cause in developed countries as per the statistics of national cancer institute. The BC may occur in both male and female. However, it has high occurrence in female throughout the world. In addition, BC is most frequently discovered as an asymptomatic nodule on a mammogram.[1,2]

BC is the most frequent cancer in women worldwide. It is the most mutual cause of cancer death among women (522,000 deaths in 2012), and the most widespread diagnosed cancer among women in 140 out of 184 countries worldwide, that constitutes 1:4 of all cancers in women. It represents the most commonly occurring cancer in women, and the second most common cancer overall. There were over 2 million new cases in 2018. Whereas, it has elevated occurrences in Iraqi women (all ages), which became one of the major menace to Iraqi female health. However, it is essential for clinical researchers to look at several body fluids to identify biomarkers. This study attempt to investigate multiple biomarkers as shown in Table 1, which are detectable in blood and saliva of 181 Iraqi women in different sites and hospitals for early detection of BC and gynecology during the period from July 2013 to October 2014.[3–5] In accordance with the Iraqi Cancer Registry data during period 2000–2009, the 23,792 total incidents registered BC cases among females aged ≥15 years. They stand for 33.8% of all cancers. The prevalence ratio of all female BC in Iraq (all ages) incremented from 26.6/100,000 in 2000 to 31.5/100,000 in 2009, which make it one of the major threats to Iraqi female health.[3]

As well as, BC incidence rates in Arab women increased. Since, improved life expectancy, urbanization growth, adopt the western lifestyles, and retarded and diminished fertility play a role to this augmentation. The lack of breast health programs is obvious in many Middle Eastern countries. As well as, a deficiency in service provision, in addition to the conservative culture, may be correlated with the increased rate of advanced stages of BC in Arab countries.[6]

Although, BC can be diagnosed by classifying tumors. There are two general types of tumors, the malignant and benign tumors. Even the physicians need a reliable diagnosis procedure to differentiate between them. Nevertheless, it is difficult to distinguish tumors even by the experts. Thereby, computerizing the diagnostic system is required to help in tumors diagnosis. The researches endeavored to apply machine-learning algorithms to detect survivability of cancers in human beings.[7] However, the researchers realized that these algorithms are done well in cancer diagnosis.[2,4] BC survivability prediction is challenging, also it is a complex research assignment. Therefore, the existing approaches incorporate statistical methods or supervised machine learning to predict the survival chances of patients.[2,7] The groundwork results gifted to the function of the data mining (DM) methods into the survivability forecast problem in BC. In spite of all challenges, DM algorithms have good results in different methodologies of BC, but more work is have to be done in managing different studies, and reviews of DM technologies.[2,7,8]

The biomarkers are very important not for reliable disease diagnostic only, but also to have a good choice from multiple

available therapeutic alternatives, that is likely to benefit the patients.[3] Whoever, it is crucial for clinical researchers to look at multiple body fluids, and different molecular techniques to identify biomarkers. Saliva one of body fluids, which is simply and non-invasively gained, and contains several types of potential protein biomarkers. Therefore, finding of CA 15-3 protein for breast cancer in saliva proposed renewed interest in the potential use of saliva as a diagnostic fluid.[3] It has been stated that person with BC secretes a different profile of proteins compared with the healthy individual.[3] As well as, the levels of vascular endothelial growth factor, epidermal growth factor (EGF), and carcinoembryonic antigen (CEA) in the saliva are considerably raised in BC patients. The salivary biomarkers appearance reflect precisely the normal and disease states. This make saliva an attractive diagnostic fluid beside the sampling benefits compared with blood sampling.[3] Women with early-stage BC have excellent survival rates. Therefore, it is critical to identify factors that anticipate diagnosis of early-stage BC.[2,4,9] Then, concluding the proportion of BCs that were identified at an early stage (stage I) in different racial/ethnic groups, and whether the ethnic differences well explained by early detection, or by intrinsic biological differences in tumor aggressiveness.[2,9] Medical organizations generate, and gather large quantity of data. The medical domain is considered as one of the leading areas for applying DM to recognize some significant properties of data.[2,10] In medical field, there are various problems, such as in medical imaging like classification, segmentation, extraction, and selection. Medical datasets categorized always by huge amount of disease measurements, and comparatively small amount of patient records. These measurements (feature selection) are irrelevant. This irrelevant and redundancy features are difficult to evaluate. Meanwhile, to represent the data set, then a large number of features causes memory storage problem. Therefore, different DM techniques can convenient with imprecision and uncertainty in data analysis, and can efficiently remove noisy and redundant information.[2,11]

Through the means of big data analytics and machine-learning techniques of patient's data of precise clinical manifestations. Hence, it became possible to identify precise biomarkers in blood or urine, which can be utilized for early diagnosis of BC upon early diagnosis, administration of precise personalized nanomedicine is quite possible, that can reduce the time and cost of drug regimen. By the next decade because of big data analytics, time, and cost of bench to bedside drug discovery lifecycle can be drastically reduced from 9 to 12 years, and current value of around US$ 1.1 billion.[12]

Convergence technology is a revolution, it is not an interdisciplinary collaboration, however taking science, research, and technology development into next revolution by interdisciplinary integration. Thereby, foremost barriers faced can be overpowered. The construction of largest patient database (1 million patients), incorporating genetic, behavioral (societal), and clinical information are under progression nowadays. Therefore, using big data analytics and machine-learning techniques on huge database, and their outcomes have to be generalized to researchers in engineering, physical, biological, and clinical science, thus to be explored by them.[12]

This research applied machine-learning algorithms in detecting BC for Iraqi women. In this article, the remaining of this paper is organized as follows. Section 2 specifies literature review about B.C analysis via DM techniques. Section 3 gives information about DM techniques, and its learning rules. Section 4 specifies related works on BC using DM. Section 5 implies other machine-learning algorithms and its types, with related work on those algorithms. Finally, Section 6 concludes the results and perspectives.

Data Mining Techniques for Breast Cancer Analysis

Data mining is an interdisciplinary field and is fast reputation because of exploring database technology, information science, machine learning, and neural networks (NN)[4,7,10] along with the statistical techniques. However, DM algorithms not applied on the medical data by common people, but the knowledge obtained can be very useful for them if shared within a comprehensible form.[2,8] In addition, the machine learning is a branch of artificial intelligence, it is a scientific discipline interested with the design, and development of algorithms, which evolve the computers behavior with regard to empirical data, from either sensor data or databases.[4]

Priyanga and Prakasam[13] proposed a cancer prediction system based on DM technology. The user's genetic and non-genetic factors are collected. Thus, that helps to predict the BC at early stage. However, it is cost effective to the user. Thereafter, Waikato Environment for Knowledge Analysis (Weka) system analyzed the medical information. Whenever, the attributes are finalized, hence the risk range could be determined via the prediction system. This system applied successfully on BC data sets, it achieves better accuracy level comparing to other existing systems. As well as, it gives earlier stage warning to the users, cost and time benefits to the user.[2,8,11] Kharya[14] proposed several efficient DM techniques to classify the BC. They were the soft computing approaches, and decision tree. They provide best predictor with 93.62% accuracy on benchmark and SEER data set.[2,7] This predictor was used to design the web page application.[11]

Delen et al.[15] used common DM algorithms, such as artificial neural network (ANN) and decision tree, and along with logistic regression[4] to formulate the prediction model for the BC by investigating large database. Then, compared the prediction models, which gives 93.6% accuracy with decision tree, 91.2% accuracy with ANN, and worst accuracy 89.2% with logistic regression.[11]

The algorithms such as decision tree, ANN, regression,[4] support vector machine (SVM), Naïve bays,[16] and backpropagation[2] are frequently considered. They introduced various results based on speed, accuracy, performance, and cost. As well as, the effective classification data aids to obtain the patient treatment.[11] DM has become a substantial methodology for computing applications in the medicine domain. The progression of DM applications and its implications are indicated in data management of healthcare administrations, epidemiology, patient care, intensive care systems, significant image analysis to information extraction, and automatic identification of unknown subjects. DM includes multiple techniques such as classification, clustering,[2] prediction, association rules, decisions tress, and NNs. Among the diverse classification algorithms, the well-known algorithms ID3 and C4.5 play an essential role in BC analysis. Plenty of researchers attempt to use machine-learning algorithms for detecting cancers survivability in human beings.[7]

Gad[17] implemented diagnosis method of BC on Wisconsin diagnosis BC (WDBC) dataset and Wisconsin prognosis BC (WPBC) dataset,[2] which combined unsupervised learning method *K*-means with SVM supervised learning method. Consequently, eliminates the inapplicable attributes using feature selection method Chi-square.[18] Shiv Shakti et al.[19] reviewed the use of DM in BC, they observed that, many researchers are mainly used the NN and decision approach to create a predictive model, and decision rules from the BC data. Most of them performed a comparative study of algorithm to take BC data. Then, to conduct an experimental work, and find various if….then rules from decision tree, which represented and used J48 classifier of WEKA.[2,8,11] Ravi Kumar et al.[20] explored a comparison among different DM classifiers on the database of BC WBC,[2,8] using classification accuracy to establish an accurate classification model for BC prediction. For full usage of invaluable information in clinical data, particularly that often neglected by most of the existing methods, whenever aimed to predict in high accuracies.[8] The dataset was divided into training set with 499, and test set with 200 patients. It compares six classification techniques in Weka software. So that, comparison results show that SVM has higher prediction accuracy than those methods. In addition, the SVM are more suitable in handling the classification problem of BC prediction.[8] Chaurasia and Pal[21] suggested a diagnosis system for detecting BC based on RepTree, RBF Network, and Simple Logistic. In test stage, tenfold cross-validation method was applied to the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia database to evaluate the proposed system performance. The classification rate of the system is 74.5%. The Simple Logistic used for reducing the feature space dimension, and proposed Rep Tree and RBF Network model used to obtain fast automatic diagnostic systems for other diseases.

Zand[2] proposed the classification of BC data could be useful to predict the result of some diseases, or discover the genetic behavior of tumors. Hence, presented a comparative study on DM techniques in diagnosis and prediction of BC, and prediction analysis for survivability rate of BC patients. Ghosh et al.[22] applied different classification techniques namely, multilayer perceptron (MLP) using backpropagation NN,[2] and SVM on BC Wisconsin dataset[2] from the UCI machine-learning repository to BC detection. The conclusion said that SVM classifier has the potential to improve significantly the conventional classification methods, hence to be used in medical or general Bioinformatics field.[10]

As well as, Delshi Howsalya and Indra[23] used machine-learning algorithm in for automatic examination of hazardous illness, BC has been considered. It identifies cancer occurrence during its beginning, and its reoccurrence, which has three stages. The first stage enclosing the data to number related entities by applying Farthest First clustering algorithm. Computation time took less time, due to decrease in the size of dataset. The second stage, deviations from the normality (outliers) are detected from BC dataset (BCD) using outlier detection algorithm. Thereafter, Final stage, J48 classification algorithm identifies whether the cancer is benign or malignant from the pre-processed data set. Wisconsin BCD (WBCD) and WDBC show an accuracy of 99.9%, which serves the doctors to diagnose the BC.[2,18] Various studies indicated factors regarding BC prognosis based on different factors. These studies recently focused on predicting BC through SVM, and on survival since the time of first diagnosis.[2,7] It is a challenging task.

Methodology

The key intention to this study is to construct data analytical model. This can provide more comprehension of BC, by forming patients' cohorts (Healthy, Benign, and Malignant), where the most common attributes extracted from 42 attributes of the mentioned three classes, which share several attributes. Moreover, identify the explicitly effectiveness of the attributes on these classes.[7]

Data Set Attributes

Breast Cancer is a heterogeneous disease with varieties in the biological profile, and subsequent clinical prognosis.[24] Prognostic information of individual patient based on the biological analysis for markers in the primary tumor, that including estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor (HER2) and Ki67, Family history, it is a powerful risk attribute for women <40 years old. Its effect increasing when first-degree relatives with breast or ovarian cancer, and if the relatives are young at diagnosis, Previous chest radiotherapy, where women with previous radiotherapy at a young age (10–30 years) have about six times as high a risk to develop BC. BRCA1/2 mutations are rare, only 0.2% of women are estimated generally to be BRCA1/2 mutation carriers. Breast density is a strong independent risk factor for the development of BC. Obesity increases the risk of postmenopausal BC, but it is a protective factor for premenopausal women. Alcohol and smoking are not risk factors for BC in women aged,[25] together with age, tumor size, histological grade, and lymph node involvement. However, there are 42 attributes examined in this study as appears in Table 1.

Table 1. **Data set attributes before and after space reduction (features selection)**

Before			After		
No.	**Attributes**	**Rank**	**No.**	**Attributes**	**Rank**
1	Ca_B_pre	93.889	1	Ca_B_pre	93.889
2	Ca_S_pre	92.222	2	Ca_S_pre	92.222
3	E2_B_Pr	91.111	3	E2_B_Pr	91.111
4	E2_S_pr	91.111	4	E2_S_pr	91.111
5	TP_S_pre	90.556	5	TP_S_pre	90.556
6	Pg_B_ps	89.444	6	Pg_B_ps	89.444
7	Type	89.444	7	Type	89.444
8	ER	89.444	8	ER	89.444
9	E2_S_ps	89.444	9	E2_S_ps	89.444
10	E2_B_ps	89.444	10	E2_B_ps	89.444
11	Grade	89.444	11	Grade	89.444
12	Pg_S_ps	89.444	12	Pg_S_ps	89.444
13	PR	89.444	13	PR	89.444
14	HER2	89.444	14	HER2	89.444
15	PH_S_pos	89.444	15	PH_S_pos	89.444
16	Ca_B_pos	89.444	16	Ca_B_pos	89.444

(*Continued*)

Table 1. **Data set attributes before and after space reduction (features selection) —*Continued***

Before			After		
No.	**Attributes**	**Rank**	**No.**	**Attributes**	**Rank**
17	CEA_B_ps	89.444	17	CEA_B_ps	89.444
18	TP_B_pos	89.444	18	TP_B_pos	89.444
19	Ca_S_pos	89.444	19	Ca_S_pos	89.444
20	CEA_S_ps	89.444	20	CEA_S_ps	89.444
21	TP_S_pos	89.444	21	TP_S_pos	89.444
22	CEA_B_pre	87.778	22	CEA_B_ pre	87.778
23	TP_B_pre	87.222	23	TP_B_pre	87.222
24	PH_S_pre	85.556	24	PH_S_pre	85.556
25	CEA_S_pre	85	25	CEA_S_ pre	85
26	Stage	85	26	Stage	85
27	BMI	77.778			
28	TBF	75.556			
29	Pg_B_pr	72.222			
30	Age	72.222			
31	Pg_S_pr	70.556			
32	W_R	65.556			
33	Blood_G	61.667			
34	Rh	61.667			
35	Lew	61.667			
36	Family_H	61.667			
37	Lact	61.667			
38	S	61.667			
39	Mns	61.667			
40	Kind_of_Co	61.667			
41	PP_meno	61.667			

CEA, carcinoembryonic antigen.

However, these data in primarily stage are classified into obese cancer, healthy obese, healthy non-obese, benign obese, and benign non-obese. Thereafter, in advanced preprocessing, they are classified into three main classes namely Healthy, Malignant, and Benign.

Data Analysis Approach

The descriptive statistics are identified for 181 women with invasive BC. The diagnosed data are captured during July 2013 to October 2014. This investigation accomplished using OneR classifier for attributes reduction to extract the most influential features, and J48 for patterns recognition. Finally, J48, simple logistic regression (SLR), and MLP algorithms were used for classifiers evaluation, and compared their performance with each other according to the correct classified instances. The predominant objective of this research is to explore the DM techniques to enhance the BC diagnosis. Peculiarly, this investigation discusses the use of the classification algorithms J48, SLR, and MLP in BC analysis. This examination used Weka platform[2,8,11] to perform the study experiments. Figure 1 presents the proposed DM architecture of this study. The preprocessing phase accomplished on the data set to prepare them for manipulation in DM techniques. The raw data set in this phase are extracted, cleaned up, and transfigured into .arff format, which was accepted by Weka platform. Thereafter, the resulted data would be classified into three main classes. These classes are Healthy (H), Malignant (M), and Benign (B). Later on, the feature selection method performed, which measures the weight of attribute according to the class. Applying an attribute selection method is to reduce the feature space, and keep only the ones that have highest affect according to a threshold, which is determined by the data analyst.

Furthermore, implying OneR machine-learning classifier to evaluate the worth of an attribute according to the class value. The attribute rank 85 is determined as a threshold of an accepted attributes, and prune the ones that is less than it. It is issued under data analyst control of this study. Then, the evaluation performed using training data rather than cross validation. Thus, it extracts 26 attributes plus class attribute instead of the initial 42 attributes with their class. The classified classes based on attributes are M, B, and H to indicate cancer, benign, and healthy instances respectively, as demonstrated in Table 1.

Moreover, the decision tree J48 algorithm applied to detect and generate the attributes patterns. These patterns have the real effect on the class value. Therefore, it is founded that, the TP_S_pre attribute has a decisive effect on the appearance of BC when it is >0.28. Otherwise, its benign or healthy case. As well as, when Ca_S_pre attribute > 1.39, and PH_S_pre > 7, then it is healthy case. Otherwise, it is benign case unless CEA_B_pre ≤ 1.51, then it will be healthy case. Hence, when Ca_S_pre ≤ 1.07, and PH_S_pre > 7 then it is healthy case. Meanwhile, when PH_S_pre ≤ 7 it is benign case. Consequently, the Ca_S_pre, PH_S_pre, and CEA_B_pre are the major attributes to distinguish between the healthy and benign cases. However, the tree in Figure 2 demonstrates all the mentioned results. Finally, the previous phases discovered hidden knowledge from a raw data. The major three methods examined in this study are explored in Table 2. The examined results are produced using tenfold cross-validation as test option, and the percentage of correctly classified instances as a measure to determine the best one from them. Furthermore, the iteration control is used to check the accuracy gained from the three mentioned algorithms in the case of 1, 5, and 10 repetitions, and to realize if the error ratio descends during iterations ascending of algorithm execution. Accordingly, the error ratio of the correctly classified cases is descending whenever number of repetitions applied increasing, exactly in the case of MLP algorithm rather than SLR. Table 2 demonstrates the percentage ratios for the correctly classified instances. Finally, the classification accuracy is estimated for the used techniques. Hence, stated that the highest efficiency achieved by MLP in detection new supposed instances, either they are malignant, benign, or healthy cases. Figures 3–5 represent the accuracy classification for used methods J48, SLR, MLP respectively.

Conclusion

This study introduced an approach for addressing BC analysis for Iraqi women, it is a comparative data mining examination. The data set analysis aims to uncover the ambiguity of the attributes relationships and their patterns. They include

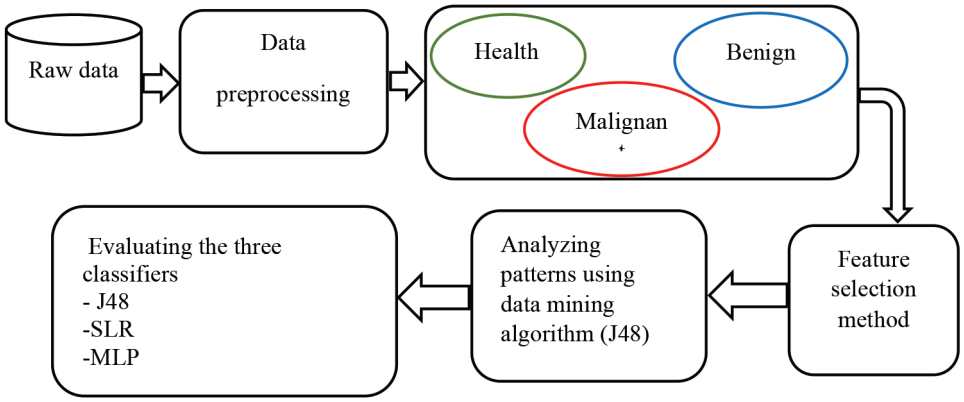

Fig. 1 **Architecture of data mining for BC investigation of Iraqi women.**

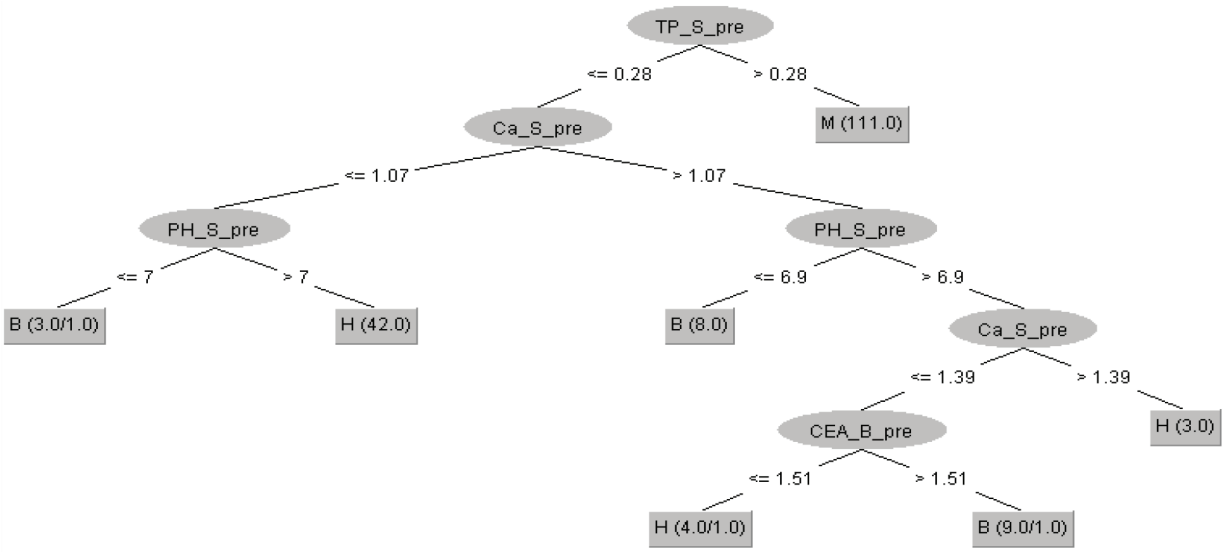

Fig. 2 **Pattern visualization of J48 decision tree.**

Table 2. **Data mining techniques with 1, 5, and 10 iterations**

Algorithm / Tenfold	1-Repetition	5-Repetition	10-Repetition
Trees J-48	93.33	91.33	91.72
SLR	98.89	98.33	97.89
MLP	97.78	98.56	98.94

MLP, multilayer perceptron; SLR, simple logistic regression.

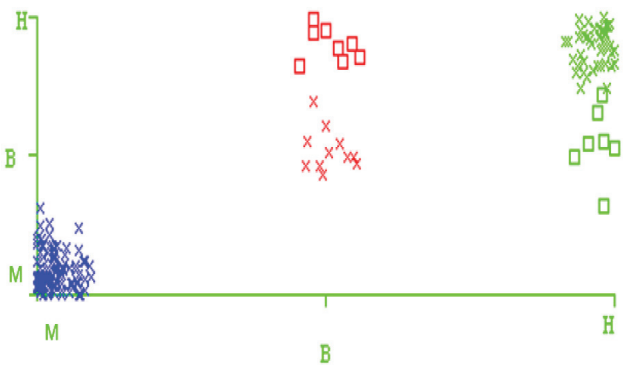

Fig. 3 **Accuracy classification J48 technique.**

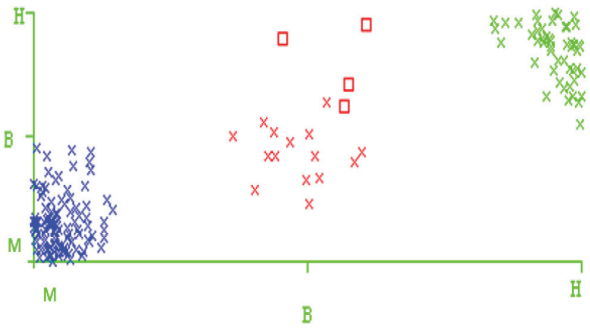

Fig. 4 **Accuracy of classification SLR technique. SLR, simple logistic regression.**

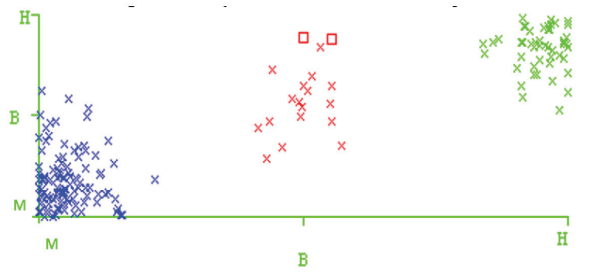

Fig. 5 **Accuracy of classification MLP technique. MLP, multilayer perceptron.**

42 attributes, which are manipulated, hence reduced to just 26 significant attributes. The acquired attributes are explored as the great influential on the BC inflammation.

This study used the information gain or data-driven method (OneR classifier) for variable selection instead of manual variable selection methods. Furthermore, used machine-learning methods decision tree (J48), MLP, and SLR. They can handle numeric and nominal attributes and map raw patient records into subsets of patient cohorts (classes), which share commonalities in attributes. Hence, the proposed method formed multiple patient subgroups with different attributes; it helps in improving the baseline accuracy of mentioned classifiers.

As a result, data mining techniques could be considered as a powerful research tool for medical researchers, and it is recommended to identify and exploit the patterns and relationships among large number of diseases attributes. Therefore, they are able to predict outcome of a disease using the historical datasets. It can also assist in realizing the most effective biomarkers. Furthermore, this investigation results supports the results in Salih et al.'s[3] study, which considered the TPS as an effective biomarkers to BC indication, this is explored using decision tree (J48) classifier. As well as, the gained results accuracy is evaluated for each of the mentioned methods. However, it is appeared that MLP is the most accurate classifier with 98.94% of correctness precision, bypassing multiple iterations of execution.

As a future perspective, it is counseled to examine big data of BC for Iraqi patients, for the sake of deriving more hidden relations, and examine larger attributes. Especially, investigates the historical family records of patients, this could help in predicting approach for BC survivability. However, the great challenge of this kind of studies is the lack of data; huge number of disease attributes, electronic registering rareness in Iraq, and the lose communication between the medical domain and informatics.

Conflicts of Interest

None.

References

1. Anupama YK, Amutha S, Ramesh Babu DR. Breast cancer prediction using data mining techniques. Int J Adv Res Sci Eng. 2018;7:41–44.
2. Zand HK. A comparative survey on data mining techniques for breast cancer diagnosis and prediction. Ind J Fundam Appl Life Sci. 2015;5:4330–4339.
3. Salih KM, Mohemmed AK, Al-Shaikh M, Al-Sayyid MM. Salivary fraction of CA 15-3 and CEA as tumor markers for breast cancer in Iraqi women. Int J Eng Technol. 2015;4:5114–5117.
4. Estanislao GL, Campos RA. Breast Cancer, Primary Peritoneal Malignant Mixed Mullerian Tumor and Fallopian Tube Carcinoma: Incidental Concomitant Malignancies or Evidence for a New Genetic Cancer Predisposition Syndrome? Conference 2018 of the European Society of Gynaecological Oncology, Lyon, France, October 4–6, 2018.
5. Al-Akeedi JM, Mahmod AS, Ali MA. Potential prognostic roles for IL-6 and CRP in Iraqi women with breast cancer. Int J Adv Biol Res. 2013;3:530–534.
6. Madkhali NA, Santin O, Noble H, Reid J. Understanding breast health awareness in an Arabic culture: qualitative study protocol. J Adv Nurs. 2016;72:2226–2237.
7. Shukla N, Hagenbuchner M, Win KT, Yang J. Breast cancer data analysis for survivability studies and prediction. Comput Methods Programs Biomed. 2018;155:199–208.
8. Yuvarani S, Jothi V. Breast cancer detection in data mining: a review. Int J Comput Appl. 2015;7:45–48.
9. Iqbal J, Ginsburg O, Rochon PA, Sun P, Narod SA. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. JAMA. 2015;313:165–173.
10. Barot V, Brahmbhatt N. A survey on breast cancer diagnosis using data mining technique. Int J Innov Res Sci Eng Technol. 2017;6:147–150.
11. Saranya P, Satheeskumar B. A survey on feature selection of cancer disease using data mining techniques. Int J Comput Sci Mobile Comput. 2016;5:713–719.
12. Khargonekar P, Sinskey A, Miller C, Ranganathan B. Convergence revolution – piloting the third scientific revolution through start-ups for breast cancer cure. Cancer Sci Res Open Access. 2017;4:1–6.
13. Priyanga A, Prakasam S. Effectiveness of data mining - based cancer prediction system (DMBCPS). Int J Comput Appl. 2013;83:11–17.
14. Kharya S. Using data mining techniques for diagnosis and prognosis of cancer disease. Int J Comput Sci Eng Inform Technol. 2012;2:55–66.
15. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34:113–127.
16. Mustafa TK, Abd MS. Proposed approach for analysing general hygiene information using various data mining algorithms. Iraqi J Sci. 2017;58:337–344.
17. Gad W. SVM-Kmeans: support vector machine based on Kmeans clustering for breast cancer diagnosis. Int J Comput Inform Technol. 2016;5:252–257.
18. Anupama YK, Amutha S, Ramesh Babu DR. Survey on data mining techniques for diagnosis and prognosis of breast cancer. Int J Recent Innov Trends Comput Commun. 2017;5:252–257.
19. Shiv Shakti S, Sant A, Aharwal RP. An overview on data mining approach on breast cancer data. Int J Adv Comput Res. 2013;3:256–262.
20. Ravi Kumar G, Ramachandra GA, Nagamani K. An efficient prediction of breast cancer data using data mining techniques. Int J Innov Eng Technol. 2013;2:139–144.
21. Chaurasia V, Pal S. Data mining techniques: to predict and resolve breast cancer survivability. Int J Comput Sci Mobile Comput. 2014;3:10–22.
22. Ghosh S, Mondal S, Ghosh B. A Comparative Study of Breast Cancer Detection based on SVM and MLP BPN Classifier. 2014 First International Conference on Automation, Control, Energy and Systems (ACES), India, IEEE, Hooghy, India, 2014.
23. Delshi Howsalya R, Indra M. Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer. Int J Adv Eng Technol. 2016;VII:93–98.
24. Fouda MA, Sherif FZ, Ghannam AA, Al-shorbagy SH. Prognostic value of breast cancer subtypes based On ER/ PR, Her2 expression and Ki-67 index in women received adjuvant therapy after conservative surgery for early stages breast cancer a retrospective clinical study. JSM Clin Oncol Res. 2017.
25. Fredholm H. Breast cancer in young women - aspects on mortality and local recurrence, Ph.D., University Hospital Solna, 2017.