Al-Khwarizmi Engineering Journal, Vol. 16, No. 3, Sptember, (2020)
P. P. 1-10

Al-Khwarizmi Engineering Journal

Two-Stage Classification of Breast Tumor Biomarkers for Iraqi Women

Iyden Kamil Mohammed* **Ali Hussein Al-Timemy****
Javier Escudero***

*,** *Department of Biomedical Engineering / Alkhwarizmi College of Engineering/ University of Baghdad/ Baghdad/ Iraq*
*** *School of Engineering/ Institute for Digital Communications/ The University of Edinburgh/ Alexander Graham Bell Building/ EH9 3FG/ UK*
*Email: aydenel_1969@yahoo.com
**Email: ali.altimemy@kecbu.uobaghdad.edu.iq
***Email: javier.escudero@ed.ac.uk

(Received 18 November 2019; accepted 26 April 2020)
https://doi.org/10.22153/kej.2020.04.003

Abstract

Objective: Breast cancer is regarded as a deadly disease in women causing lots of mortalities. Early diagnosis of breast cancer with appropriate tumor biomarkers may facilitate early treatment of the disease, thus reducing the mortality rate. The purpose of the current study is to improve early diagnosis of breast by proposing a two-stage classification of breast tumor biomarkers fora sample of Iraqi women.
Methods: In this study, a two-stage classification system is proposed and tested with four machine learning classifiers. In the first stage, breast features (demographic, blood and salivary-based attributes) are classified into *normal* or *abnormal* cases, while in the second stage the abnormal breast cases are further classified into either *malignant* or *benign*. The collected 20 breast cancer features are utilized to test the performance of the proposed classification system with Leave-One-Out (LOO) cross validation and Synthetic Minority Over-Sampling Technique (SMOTE) to balance the classes. Furthermore, correlation-based feature selection (CFS) was employed in an exploratory analysis to find the best features for the 2-stage classification system.
Results: Classification accuracy of 9٤% for stage-1 and 100% for stage-2was achieved with a Naïve Bayesclassifier which outperformed other three methods. In addition, CFS selected small subset of features as being the best five features out of the all 20 features for both stage-1 and stage-2.
Conclusion: We achieved a high classification accuracy which is promising to help improve the early diagnosis of breast tumor. The outcome of this study also shows the importance of CA15-3protein in saliva and blood as well as carcinoembryonic antigen level and total protein in blood, and Estrogen hormone level in saliva, for predicting breast tumors.

Keywords: *Breast cancer, correlation-based feature selection, decision tree, machine learning, oner algorithm, two-stage classification.*

1. Introduction

In Iraq, breast cancer is regarded the most common type of malignancy [1], [2]. Among all the malignant diseases, breast cancer is assessed as one of the main causes of death in post-menopausal women, accounting for 23% of all cancer deaths in 2017 [3]. In 2010, breast cancer is almost recognized as the deadliest cancer in women since it is regarded as number one cause of cancer mortality among women [4].
Biomarkers have many potential applications in oncology, including screening, risk assessment, determination of prognosis, prediction of

response to treatment, differential diagnosis, and monitoring of progression of disease. Due to the major role that biomarkers may play at all stages of disease, they should undergo rigorous evaluation, including clinical validation, analytical validation, and assessment of clinical utility prior to incorporation into regular clinical care [5].

A tumor biomarker is a molecular or process-based change that discloses the status of an underlying malignancy. A tumor biomarker may be diagnosed and assessed via one or more biomarker assays or tests. Patient management is progressively being derived by tumor biomarker tests. This can be done by recognizing patients who do not require any, or recognizing other patients whose tumors are so unlikely to respond to a given type of treatment that it will drive to more harm than good. Thus, patient management should be guided by a tumor biomarker test, to inspect if it has analytical validity, which means it is accurate, reliable and reproducible [6].

Machine learning techniques have been utilized to classify cancer attributes and biomarkers aiming at improving the diagnosis rate. For instance, a classifier- based expert system was proposed in [7] for early diagnosis of prostate cancer with Artificial Neural Network (ANN) and Support Vector Machines (SVM). Thirteen attributes were acquired from 300 men to classify benign and malignant tumors. Classification accuracy of 79.3% and 80.1% was obtained with ANN and polynomial SVM, respectively. Other researchers employed machine learning and data mining techniques to investigate breast tumors. Three popular machine learning classifiers (Naive Bayes, Radial Basis Function Neural Network (RBFNN), J48 decision tree) were used [8] to develop prediction models for 683 breast cancer cases. Classification accuracy of 97.36%, 96.77 and 93.41% was obtained for the Naive Bayes, RBFNN, J48 classifiers, respectively. In another study, Behadili et al. [9] analyzed 42 attributes of the Iraqi women and selected 26 attributes for the classification of three classes with the decision tree J48 algorithm with 98% accuracy.

Other researchers tried to reduce the size of the feature set to detect breast cancer with Independent Component Analysis (ICA) [10]. A publicly available data set, Wisconsin diagnostic breast cancer (WDBC) dataset, was utilized to test the proposed algorithm where the 30 attributes have been reduced to only one feature (IC). Then, reduced feature was utilized to evaluate diagnostic accuracy with multiple classifiers: k-nearest neighbor (k-NN), ANN, RBFNN, and SVM.

In some occasions, three- class classification problems have been tackled such as the work in[11] by proposing two-stage classification where Computer-aided diagnosis (CAD) system have been proposed to classify brain tumors. The system classified brain tumor MRI images into normal and abnormal images in the first stage and if the output was abnormal, then it was additionally classified into malignant or benign tumor.

In this paper, we propose a two-stage classification for breast cancer classification with four machine learning classifiers. In the first stage, 20 breast attributes are classified into normal or abnormal cases whereas in the second stage the abnormal breast cases are further classified into malignant or benign. Furthermore, correlation-based feature selection will be employed to find the best attributes out of the 20 attributes.

The main contribution of this paper is that it presents a two-stage classification system for the classification of breast markers with machine learning classifiers. The most important attributes that are influential in the prediction of breast tumor will be investigated with correlation-based feature selection on a data set of 181 samples.

2. Materials and Methods

A. Details of Breast Cancer Data Collection

The data set utilized in this study was acquired from the center for early breast cancer detection and Elwiya Oncology teaching hospital in 2013 and 2014. Ethical approval to conduct the data collection was obtained from the Ministry of Health. In addition, data collection was performed in accordance with the *Declaration of Helsinki 1964*, and its later amendments. The data set consisted of 181subjects (111 malignant, 50 normal control and 20 benign cases). The oncologist examined subjects, with suspected breast tumor, and approved their inclusion in the study, confirmed with the following: 1) clinical examination, 2) breast biopsy, 3) mammogram and 4) Ultrasound (US) scan. Before the data collection, subjects were debriefed and consented to participate in the study.

The data set has 20 attributes which can be categorized in 3 main categories, i) demographic information (9 attributes),ii) attributes derived

from blood samples (5 attributes) and iii) attributes derived from saliva (6 attributes).
The nine demographic attributes include age, body mass index (BMI), total body fat index (TBF), Waist Hip Ratio (WHR), number of menstruation (Mens.) cycles per year, duration of contraceptive intake per year, Menstruation (Mens.) cycle status (normal or abnormal), type of Fucosyl transferase 2 (FUT2) gene (secretor or non-secretor), type of Lewis blood group.

The 5biomarkers that are derived from blood samples included the level of Estrogen hormone (Es-B), the level of progesterone hormone (Pg-B), the level of CA15-3 protein (CA15-3 B), carcinoembryonic antigen level (CEA-B) and Total Protein (TP-B).

The biomarker that were obtained from analyzing salivary samples included the PH level (PH-S), salivary Total Protein (TP-S), Estrogen hormone level (Es-S), the level of CA15-3 protein (CA15-3 S), saliva progesterone hormone level (Pg-S) and salivary carcinoembryonic antigen level (CEA-S).

B. Details of the Pattern Classifcation

In this study, a two-stage classification system is proposed where the breast attributes are classified into either *normal* or a*bnormal* in the first stage. In the second stage of the classification, the abnormal instances can be further classified into *malignant* and *benign* cases. The general block diagram of the two-stage classification system is shown in Fig. 1.

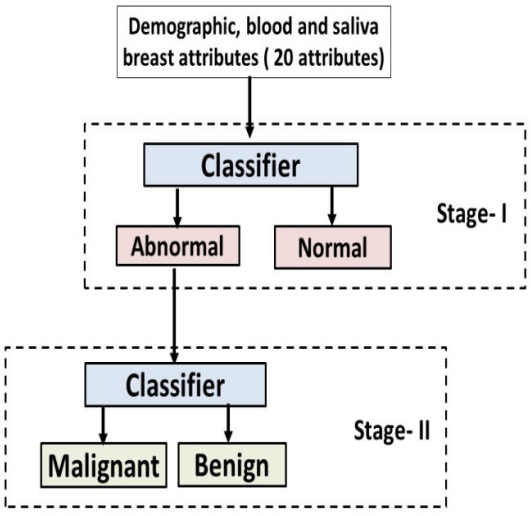

Fig. 1. Block diagram of the proposed breast attribute classification method.

How to select the best machine learning classifier for the breast cancer detection can be considered as an open research question. Therefore in this study we investigate the performance of the proposed two-stage classification system with four different classifiers: logistic regression [12], Naïve Bayes [13], decision tree [12] and OneR [14]. The rationale behind choosing these machine learning classifiers to perform the analysis in this study was that these classifiers are relatively straightforward and simple to implement, and they have been utilised in the previous literature. Weka, an open source software package [12] was used to perform classification with the four classifies in the current study.

Logistic regression is a traditional classification method where class probabilities are estimated by means of applying the logit transformation to a linear regression model [3]. It provides a mechanism for applying linear regression for performing classification [15]. For more details about the mathematical derivation, the reader is referred to [15]. *Logistic* function in Weka has been utilised for logistic regression [12].

Naïve Bayes classifiers is a probabilistic classifier which is based on the Bayes theorem and being considered as a simple classifier [16]. In addition, it assumes independence in a naïve way [12]. The hypothesis is that the probabilities of each feature multiply is only valid if the events are independent. Despite the simplistic assumption of independent attributes in real life, Naïve Bayes works very effectively when utilized to classify real life datasets [12]. The pseudocode for Naïve Bayes classifier [17] is displayed in Fig.2.

DecisionStump, a fast decision tree learner which uses reduced-error pruning, was used. It builds one level binary decision tree with categorical or numeric class to perform the classification. *DecisionStump* in Weka was utilized to do decision tree. For more details about the implementation of *Decision Stump*, the reader is referred to [18].

OneR classifier is simple but accurate classifier. OneR algorithm generates one rule for each predictor in the data by forming a frequency table for each predictor against the target [19]. Afterwards, the rule with minimal total error is selected. Fig. 3 illustrates the pseudocode for OneR algorithm [12].

```
Input:
Training dataset T,
F= (f1, f2, f3,.., fn) // value of the predictor
variable in testing dataset.
Output:
A class of testing dataset.
Step:
1. Read the training dataset T;
2. Calculate the mean and standard deviation of the
predictor variables in each class;
3. Repeat
Calculate the probability of fi using the gauss
density equation in each class;
Until the probability of all predictor variables
(f1, f2, f3,.., fn) has been calculated.
4. Calculate the likelihood for each class;
5. Get the greatest likelihood;
```

Fig. 2. Algorithm for Naïve Bayes classifier.

```
OneR Pseudocode:

for each attribute:
 for each value of that attribute make a rule as follows:
    count how often each class appears
    find the most frequent class
    make the rule assign that class to this attribute-value.
 Calculate the error rate of the rules
Choose the rules with the smallest error rate
```

Fig. 3.The pseudocode for OneR algorithm.

The dataset used in this study has an imbalanced number of the three classes. Class imbalances may limit the performance of machine learning classifiers [20]. In addition, different classification techniques are sensitive to the imbalanced data when the samples of one class in a dataset outnumber the samples of the other class. This may lead to biased models due to overfitting. To tackle the issue of class imbalance, we utilized Synthetic Minority Over-Sampling Technique (SMOTE), proposed by Chawla et al. [21] to equalize the number of classes in both classification stages. In SMOTE, *k*-nearest neighbor (kNN) is used to generate synthetic instances to oversample the minority classes while the size of the majority class is kept the same [21].

The Weka filter *'SMOTE' was used* for both classification stages where the number of *k*-nearest neighbor was set to the default value of 5. In stage 1, the total number of instances became 256 for stage 1 and 221 for stage 2, after applying SMOTE.

To evaluate the performance of the breast classification, the exhaustive Leave-One-Out (LOO) Cross-Validation (CV) was utilised in each stage of the classification. LOO CV will prevent overfitting and bias in the evaluation of classification performance despite that it requires lot of computations compared to the 10-fold CV since it requires to go through all dataset.

Classification accuracy, precision, recall and were calculated given true positive (TP), true negative (TN), false positive (FP), and false negative (FN), as follows

$$Precision = \frac{TP}{TP+FP} \quad \ldots(1)$$

$$Recall = \frac{TP}{FP+FN} \quad \ldots(2)$$

$$Accuracy(Ac) = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots(3)$$

Furthermore, we calculated Matthews correlation coefficient (*MCC*) [22], given in eq. 4, which is an indicator used to evaluate the performance of classification quality where the output value is between the range -1 to 1; high value of MCC of 1 indicates an excellent classification.

$$MC = \frac{TP \times TN - FP \times FN}{\sqrt{P(1-P)S(1-S)}} \quad \ldots(4)$$

Where the values of P and S are given below

$$P = (TP + FP)/ (TP + FP + TN + FN) \quad \ldots(5)$$

$$S = (TP + FN)/ (TP + FP + TN + FN) \quad \ldots(6)$$

C. Selecting the Best Attributes for Breast Cancer Classification

In this study, Correlation-based Feature Subset Selection (CFS) [23] was utilised, for exploratory and interpretability purposes, to select the best attributes for each stage of the breast cancer classification. The main theory is that features which have high correlation with the class label but are uncorrelated with each other may represent the base for a good feature set. A feature evaluation formula is developed from the aforementioned theory. CFS then combines the developed evaluation formula with a heuristic search strategy and suitable correlation measure.
The best classifier from the previous analysis (Section II.B) will be utilised alongside Weka function *AttributeSelectedClassifier*, CFS (Weka evaluator '*CfsSubsetEval*') with the search method selected to be 'Best first'. It should be noted that '*SMOTE'* filter in Weka was utilised in this part of the analysis to oversample the minority class for both stage-1 and stage-2.

3. Results and Discussion

The Minimum (Min), Maximum (Max), mean and standard deviation (STD) for the 20 attributes in this study are shown Table 1 for the normal subjects, Table 2 for the malignant patients and Table 3 for the benign patients, respectively. There are large differences between some attributes for the three groups (such as *CA15-3 B, CEA-B and* CA15-3 S) while other attributes have smaller differences (such as TBF and PH-S).

Table 1,
The Min, Max. Mean and Std values of the 20 attributes for normal subjects (n=50).

Attribute	Min	Max	Mean	STD
Age	25	67	44.26	11.44
BMI	21.2	26.4	23.70	1.35
TBF	25.91	46.9	33.67	4.48
WHR	0.63	0.95	0.78	0.08
Mens. cycles/year	10	14	12.52	0.74
Dur. of contracept./year	0	9	0.50	1.66
Mens. cycle status	0	1	0.06	0.24
FUT2 gene type	0	1	0.72	0.45
Lewis blood type	0	2	1.24	0.52
Es-B	11	97	50.48	25.70
Pg-B	0.73	4	1.81	0.99
CA15-3 B	3.11	15.3	6.61	2.74
CEA-B	0.59	2.94	1.78	0.57
TP-B	6.02	7.3	6.47	0.35
Es-S	3.18	28.1	14.62	7.43
Pg-S	0.24	1.8	0.68	0.41
CA15-3 S	0.5	2.68	0.79	0.41
CEA-S	0.5	0.5	0.50	0.00
TP-S	0.05	0.28	0.14	0.04
PH-S	6.9	7.4	7.23	0.12

Table 2,
The min, max. mean and std values of malignant patients (N=111).

Attributes	Min	Max	Mean	STD
Age	26	73	54.21	8.14
BMI	21.4	34.4	26.90	2.86
TBF	27.34	53.5	39.53	4.16
WHR	0.67	1.1	0.88	0.11
Mens. cycles/year	10	15	12.66	0.79
Dur. of contracept. /year	0	18	3.52	5.08
Mens. cycle status	0	1	0.31	0.46
FUT2 gene type	0	1	0.39	0.49
Lewis blood type	0	2	1.30	0.75
Es-B	182	384	270.86	50.27
Pg-B	0.62	4.9	1.42	1.16
CA15-3 B	18	79	52.44	18.79
CEA-B	2.7	14.3	8.71	2.82
TP-B	7	10.8	8.76	1.22
Es-S	52	111.2	77.68	14.78
Pg-S	0.2	1.8	0.54	0.46
CA15-3 S	2.57	13.57	8.98	3.27
CEA-S	0.5	2.7	1.21	0.56
TP-S	0.61	1.52	0.91	0.27
PH-S	5.4	7.2	6.06	0.58

Table 3,
The min, max. mean and std values of the 20 attributes for benign patients (N=20).

Attributes	Min	Max	Mean	STD
Age	28	58	45.15	7.86
BMI	21.1	32	25.63	3.18
TBF	26.36	44.91	35.17	5.01
WHR	0.69	0.92	0.81	0.07
Mens. cycles/year	11	13	12.10	0.45
Dur. of contracept. /year	0	5	0.55	1.39
Mens. cycle status	0	1	0.20	0.41
FUT2 gene type	0	1	0.60	0.50
Lewis blood type	0	2	1.25	0.55
Es-B	18	96	59.75	30.89
Pg-B	0.65	4.2	2.29	1.32
CA15-3 B	6.59	13.6	10.49	1.80
CEA-B	1.02	5.71	2.46	1.15
TP-B	6.1	7.6	6.78	0.55
Es-S	5.3	27.8	17.24	8.93
Pg-S	0.2	1.6	0.84	0.54
CA15-3 S	0.69	2.35	1.51	0.47
CEA-S	0.5	0.68	0.51	0.04
TP-S	0.12	0.28	0.17	0.04
PH-S	6.4	7.3	7.00	0.27

In order to investigate the performance of the four machine learning classifiers, Table 4 and 5 shows the results of the classification of stage-1 and stage-2 in terms of precision, recall, accuracy and MCC for logistic regression, Naïve Bayes, decision tree and OneR classifiers. Naïve Bayes classifier is outperforming other classifiers for stage-1 when we are classifying *normal* versus *abnormal* who had tumors as well as achieving similar performance for stage 2 for the case of classifying *malignant* and *benign* patients.

Table 4,
The results of stage-1 classification using different classifiers with LOO cross validation. The best performer is shown in bold

Classifier	Precision	Recall	Ac	MCC
OneR	0.914	0.914	91.4	0.828
Logistic	0.934	0.934	93.4	0.867
Naïve Bayes	**0.948**	**0.941**	**94.1**	**0.889**
Decision Tree	0.930	0.930	93	0.860

Table 5,
The results of stage-2 classification using different classifiers with LOO cross validation.

Classifier	Precision	Recall	Ac	MCC
OneR	1	1	100	1
Logistic	0.996	0.995	99.5	0.991
Naïve Bayes	1	1	100	1
Decision Tree	1	1	100	1

The confusion matrix (CM) for the classification of breast cancer attributes is plotted with Naïve Bayes (best performer) classifier in Table 6 and 7. The results in the diagonal of CM show the correct classification rates while the misclassifications are shown off-diagonal. There were only 15 cases (table 6) out of the 256 (with SMOTE) cases that were misclassified for stage 1 while all the 221 (with SMOTE) malignant and benign cases were classified correctly in stage 2 as shown in table 7.

Table 6,
Confusion matrix of the Naïve Bayes classifier for stage 1 after using SMOTE

	Predicted Class	
Actual group	**Abnormal**	**Normal**
Abnormal	116	15
Normal	0	125
Overall accuracy= 94.1 %		

Table 7,
Confusion matrix of the Naïve Bayes classifier for stage 2 of the classification after using SMOTE

	Predicted Class	
Actual group	**Malignant**	**Benign**
Malignant	111	0
Benign	0	110
Overall accuracy= 100%		

When comparing the results obtained in this study with that in [9] who investigated 42 attributes and selected 26 attributes for the classification of three classes of breast tumors, we utilized SMOTE to balance the classes unlike [9] who used imbalanced classes, which may cause overfitting, despite the high accuracy obtained on their work 98%.

To find the best breast attributes that have the most influence for the classification of breast cancer in the stage 1 and 2, we utilizedCFS [23]. Table 8 shows the best *ranked* selected attributes for stage-1 where the CFS selected 10 attributes while for stage-2, the best selected attributes were equal to 11. It is worth noting that the classification accuracy with the best selected attributes with LOO was equal to 94.9 %, slightly higher than that of the full set of attributes (94.1%).

In stage-1 for the classification of normal or abnormal cases, BMI, menstrual cycle status, and FUT2 gene type were the 3 most important ranked attributes, selected by the CFS. As for stage-2 for the classification of benign and malignant cases, the selected features were age, WHR and number of menstrual cycles per year.

It can be noted also that the common five attributes that are shared between stage-1 and stage-2 are CEA-B, CA15-3 B, CA15-3 S, TP-B and Es-S. CEA-B is tumor biomarker that is derived from blood; it is not specific for breast tumor. However, CA15-3 B is regarded as tumor biomarker and it is specific protein biomarker for breast cancer. Moreover, CA15-3 S and Es-S are

newly derived breast tumor biomarker that is derived from saliva. It is highly promising since it is non-invasive based on easy to acquire salivary sample.

Table 6,
The results of the best ranked selected attributed with CFS with Naïve Bayes classifier. The attributes that are common the two classification stages are shown in bold.

	Stage-1	Stage-2
1	BMI	Age
2	Mens. cycle status	WHR
3	FUT2 gene type	Mens. cycles/year
4	**CA15-3 B**	Lewis blood type
5	**CEA-B**	Es-B
6	**TP-B**	**CA15-3 B**
7	**Es-S**	**CEA-B**
8	Pg-S	**TP-B**
9	**CA15-3 S**	**Es-S**
10	PH-S	**CA15-3 S**
11		TP-S

4. Conclusion

In this paper, 20 breast cancer attributes have been collected for the Iraqi women and utilized to test the performance of two-stage classification withfourclassifiers where the attributes are classified into *normal* and *abnormal* cases in the first stage. If the case was abnormal, then the second stage of the classification is performed to predict either the patient has a *malignant* or *benign* tumor. Synthetic Minority Over-Sampling Technique (SMOTE) was utilized to deal with the problem of class imbalance. Classification accuracy of 94% for stage-1 and 100 % was achieved with Naïve Bayes classifier and LOO cross-validation. The level of CA15-3 protein in blood (CA15-3 B) and saliva (CA15-3 S), and carcinoembryonic antigen level (CEA-B) and total protein (TP-B) in blood and also Estrogen level (Es-S) in saliva, were the best selected features with CFS for both classification stages which show the importance of those parameters in predicting malignant breast tumors.

Acknowledgments

The first author would like to thank the contribution of Dr Khalid Mahdi Salah, University of Mustansiriyah for his comments on the data collection. The authors are grateful for the reviewers for their insightful comments.

Conflicts of Interest Disclosure

Authors declare that there are no conflicts of interest.

5. References

[1] N. A. S. Alwan, “Breast cancer: demographic characteristics and clinico-pathological presentation of patients in Iraq,” 2010.

[2] M. S. Dawood and A. A. Mohammed, “Breast Tumor Diagnosis Using Diode Laser in NearInfrared Region,” Al-Khwarizmi Eng. J., vol. 5, no. 2, pp. 20–31, 2009.

[3] M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan, “Awareness and current knowledge of breast cancer,” Biol. Res., vol. 50, no. 1, p. 33, 2017.

[4] G. N. Sharma, R. Dave, J. Sanadya, P. Sharma, and K. K. Sharma, “Various types and management of breast cancer: an overview,” J. Adv. Pharm. Technol. Res., vol. 1, no. 2, p. 109, 2010.

[5] N. L. Henry and D. F. Hayes, “Cancer biomarkers,” Mol. Oncol., vol. 6, no. 2, pp. 140–146, 2012.

[6] D. F. Hayes, “Biomarker validation and testing,” Mol. Oncol., vol. 9, no. 5, pp. 960–966, 2015.

[7] M. Çınar, M. Engin, E. Z. Engin, and Y. Ziya Ateşçi, “Early prostate cancer diagnosis by using artificial neural networks and support vector machines,” Expert Syst. Appl., vol. 36, no. 3, Part 2, pp. 6357–6361, 2009.

[8] V. Chaurasia, S. Pal, and B. B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques,” J. Algorithm. Comput. Technol., vol. 12, no. 2,

pp. 119–126, 2018.

[9] S. F. Behadili, M. S. Abd, I. K. Mohammed, and M. M. Al-Sayyid, "Analyzing Breast Cancer Data for Iraqi Women using Data Mining Techniques," in 3rd International Medical Education CONGRESS, 2018.

[10] A. Mert, N. Kılıç, E. Bilgili, and A. Akan, "Breast cancer detection with reduced feature set," Comput. Math. Methods Med., vol. 2015, 2015.

[11] M. K. Abd-Ellah, A. I. Awad, A. A. M. Khalaf, and H. F. A. Hamed, "Design and implementation of a computer-aided diagnosis system for brain tumor classification," in 2016 28th International Conference on Microelectronics (ICM), 2016, pp. 73–76.

[12] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[13] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," in Eleventh Conference on Uncertainty in Artificial Intelligence, 1995, pp. 338–345.

[14] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," Mach. Learn., vol. 11, pp. 63–91, 1993.

[15] C. Sammut and G. I. Webb, Eds., "Logistic Regression BT - Encyclopedia of Machine Learning," Boston, MA: Springer US, 2010, p. 631.

[16] M. Karabatak, "A new classifier for breast cancer detection based on Naïve Bayesian," Measurement, vol. 72, pp. 32–36, 2015.

[17] M. F. A. Saputra, T. Widiyaningtyas, and A. P. Wibawa, "Illiteracy Classification Using K Means-Naïve Bayes Algorithm," JOIV Int. J. Informatics Vis., vol. 2, no. 3, pp. 153–158, 2018.

[18] C. Sammut and G. I. Webb, Eds., "Decision Stump BT - Encyclopedia of Machine Learning," Boston, MA: Springer US, 2010, pp. 262–263.

[19] S. Sayad, "Tutorial on OneR classifier." [Online]. Available: http://www.saedsayad.com/oner.htm. [Accessed: 28-Jan-2019].

[20] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intell. data Anal., vol. 6, no. 5, pp. 429–449, 2002.

[21] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[22] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," Bioinformatics, vol. 16, no. 5, pp. 412–424, 2000.

[23] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

تصنيف على مرحلتين للمؤشرات الحيوية لأورام الثدي للنساء العراقيات

ايدن كامل محمد* علي حسين التميمي خافير اسكوديروا*****

*،**قسم هندسة الطب الحياتي/ كلية الهندسة الخوارزمي/ جامعة بغداد

*** كلية الهندسة/ معهد الاتصالات الرقمي/ جامعة ادنبره/ المملكة المتحدة

البريد الالكتروني: aydenel_1969@yahoo.com

**البريد الالكتروني: ali.altimemy@kecbu.uobaghdad.edu.iq

***البريد الالكتروني: javier.escudero@ed.ac.uk

الخلاصة

الهدف: يعتبر سرطان الثدي سرطان مميت في النساء ويسبب الكثير من الوفيات. إن التشخيص المبكر لسرطان الثدي باستخدام العلامات الحيوية المناسبة للورم قد يسهل العلاج المبكر للمرض ، مما يقلل من معدل الوفيات. الغرض من الدراسة الحالية هو تحسين التشخيص المبكر للثدي من خلال اقتراح تصنيف على مرحلتين للمؤشرات الحيوية لأورام الثدي لعينة من النساء العراقيات.

الطرق: في هذه الدراسة ، تم اقتراح نظام تصنيف على مرحلتين واختباره مع أربعة مصنفات للتعلم الآلي. في المرحلة الأولى ، تُصنَّف ملامح الثدي (الخصائص الديموغرافية والسمات المستخلصة من عينات الدموالخصائص المستخلصة من عينات اللعاب) لحالات طبيعية واخرىمرضية ، في حين يتم تصنيف حالات الثدي غير الطبيعية في المرحلة الثانية إما بشكل ورم خبيث أو حميد. يتم استخدام السمات العشرون لسرطان الثدي التي تم جمعها لاختبار أداء نظام التصنيف المقترح . علاوة على ذلك ، تم استخدام اختيار الميزات المستندة إلى الارتباط (CFS) في تحليل استكشافي للعثور على أفضل الميزات لنظام التصنيف على مرحلتين.

النتائج: تم تحقيق دقة التصنيف بنسبة ٩٤٪ للمرحلة الأولى و ١٠٠٪ للمرحلة ٢ مع مصنف Naïve Bayesالذي تفوق على الطرق الثلاثة الأخرى. بالإضافة إلى ذلك ، حددت CFS مجموعة فرعية صغيرة من السمات باعتبارها أفضل خمسسمات من بين كل ٢٠ سمة لكل من المرحلتين الأولى والثانية.

الخلاصة: لقد تم تحقيق دقة تصنيف عالية والتي تعد بالمساعدة في تحسين التشخيص المبكر لورم الثدي. تظهر نتائج هذه الدراسة أيضًا أهمية البروتين CA15-3في اللعاب والدم وكذلك مستوى المستضد السرطاني في الدم للتنبؤ بأورام الثدي.