Damascus university Journal for engineering sciences
Vol 39 No.2 (2023) :300-305

مجلة جامعة دمشق للعلوم الهندسية
المجلد 39– العدد 2 (2023) 300-305

The Impact of Biomarker Combination on the Breast Cancer Detection Using the Coimbra Dataset and Neural Networks

Eng. Hiba Allah Essa[1] , Prof. Mhd Firas Al-Hinnawi[2].

[1]Researcher, Department of Medical Engineering, Faculty of Mechanical and Electrical Engineering, Damascus University, Damascus, Syria. hiba.essa@damascusuniversity.edu.sy
[2]Professor, Department of Medical Engineering, Faculty of Mechanical and Electrical Engineering, Damascus University, Damascus, Syria. mhd-firas.alhinnawi@damascusuniversity.edu.sy

Abstract

Breast cancer (BC) is the most prevalent type of cancer in women around the globe. Cancer diagnoses have recently included tissue-based biomarkers, protein-based biomarkers, and molecular-based biomarkers. Several machine learning methods have been created and applied to successfully use biomarkers in breast cancer diagnosis and detection. In this study, we investigate the impact of particular biomarker combinations on the final accuracy and performance using the method of normalized features and neural networks. This article provides new findings on how biomarker combination affects cancer detection effectiveness using normalized features and a FNN model. The research found overall performance of detection with Insulin does not show any additional or distinct difference in accuracy, highlighting Insulin's modest role in identifying BC cases compared to Glucose, Resistin, and HOMA with classification accuracy between 83% and 88%.Moreover, HOMA, Glucose, BMI, and Leptin play an essential part in identifying BC when their normalized values are compared to the average of healthy and BC samples with an accuracy of 95% and sensitivity of 92%.

Keywords: Breast Cancer, Feed-Forward Neural Networks, Biomarkers.

Received: 4/5/2023
Accepted: 7/5/2023

تأثير اختلاف المركبات الحيوية على اكتشاف سرطان الثدي باستخدام مجموعة بيانات Coimbra والشبكات العصبونية

م. هبة الله عيسى[1]، أ.د.م. محمد فراس الحناوي[2].

[1]باحثة، قسم الهندسة الطبية، كلية الهندسة الميكانيكية والكهربائية، جامعة دمشق، دمشق، سورية.
hiba.essa@damascusuniversity.edu.sy

[2]أستاذ، قسم الهندسة الطبية، كلية الهندسة الميكانيكية والكهربائية، جامعة دمشق، دمشق، سورية.
mhdfiras.alhinnawi@damascusuniversity.edu.sy

الملخص:

سرطان الثدي (BC) هو أكثر أنواع السرطانات انتشاراً بين النساء حول العالم. تضمنت طرق تشخيص السرطان مؤخراً المؤشرات الحيوية المستخلصة من الأنسجة، والمؤشرات الحيوية القائمة على البروتين، والمؤشرات الحيوية الجزيئية. تم إنشاء العديد من طرق تعلم الآلة وتطبيقاتها لاستخدام المؤشرات الحيوية بنجاح في تشخيص سرطان الثدي واكتشافه. في هذه الدراسة، نعمل على دراسة تأثير مجموعات محددة من المؤشرات الحيوية على الدقة والأداء النهائيين لكشف سرطان الثدي باستخدام طريقة السمات المنسوبة والشبكات العصبونية. تقدم هذه المقالة نتائج جديدة حول كيفية تأثير مجموعة المؤشرات الحيوية على فعالية الكشف عن السرطان باستخدام الطريقة المقترحة. أظهرت النتائج أن الأداء العام للكشف عن الأنسولين لا يظهر أي اختلاف إضافي أو واضح في الدقة، مما يبرز دور الأنسولين المتواضع في تحديد حالات BC مقارنة بالجلوكوز والريزيستين و HOMA مع دقة تصنيف تتراوح بين 83٪ و 88٪. يلعب الجلوكوز ومؤشر كتلة الجسم واللبتين دوراً أساسياً في تحديد BC عندما تتم مقارنة قيمها الطبيعية مع متوسط العينات السليمة وعينات BC بدقة 95٪ وحساسية 92٪.

الكلمات المفتاحية: سرطان الثدي، الشبكات العصبية ذات التغذية الأمامية، المؤشرات الحيوية.

تاريخ الإيداع: 2023/5/4
تاريخ القبول: 2023/5/7

1

Introduction

Breast cancer (BC) is the most prevalent type of cancer in women, the substantial rise in the number of cases over the last decades has made it a serious medical problem [1]. Breast cancer can be detected using a variety of tests, including traditional mammograms, ultrasound, and magnetic resonance imaging (MRI). If doctors identify a suspicious growth, a biopsy should be taken, so it can be examined for signs of cancer. Despite the widespread use of mammogram as the gold standard for identifying and locating cancer, mammography is risky for women, particularly those under the age of thirty [2]. Due the development of spectral analysis techniques and devices for analyzing blood components, science tended to work on the concept of biomarkers and what could help in classifying, diagnosing and predicting the presence of cancers and malignant tumors of all kinds [3].

Over the last several decades, fundamental cancer diagnoses have evolved to include tissue-based biomarkers, protein-biomarkers, and molecular-based biomarkers such as osteopontin (OPN), cancer antigen CA125, and CA15-3 [4].

Patricio *et al.* in 2018 presented the Coimbra dataset in 2018, which includes biomarker samples from 52 healthy and 64 BC patients [5]. The database invloves body mass index (BMI) (kg/m^2), age (years), Glucose (mg/dL), Insulin (µU/mL), homeostatic model assessment (HOMA), leptin (ng/mL), adiponectin (µg/mL), resistin (ng/mL), and monocyte chemoattractant protein-1 (MCP-1) (pg/dL). The Coimbra breast cancer dataset is one of the most important datasets used to investigate and detect BC diagnosis using machine learning tools. This database is available on the UCI Machine Learning Repository. Support vector machine (SVM) models with Resistin, BMI, glucose, and age gave a higher precision (85% to 90%), sensitivity (82% to 88%), and area under the curve (AUC) of (0.87, 0.91) [5]. In 2018, Yixuan Li and Zixuan Chen evaluated five various machine learning algorithms, including SVM, random forest, neural networks, logistic regression (LR) and random forest (RF), and discovered that RF is the best model to use with Coimbra dataset based on assessment criteria [6]. Silva *et al.* used fuzzy neural networks to detect breast cancer in Coimbra dataset in 2019 and achieved a final accuracy of 62% and a sensitivity of 78.3% [7]. The researchers used a fuzzy decision tree (FDT) to categorize the Coimbra dataset and achieved a total accuracy of 70.68% and sensitivity of 78.3% [8].

We recently developed a technique for generating new normalized features to the average values of healthy and BC samples, then used as input for a feed-forward neural network. (FNN) [9,10]. Using the Coimbra dataset, the suggested approach achieved 91.7% classification accuracy and 92.3% sensitivity, compared to 79.3% and 75% for the raw samples. *Patricio et al.* investigated the impact of a specific number of biomarkers on the accuracy of the SVM model and the significance of a specific combination of biomarkers, finding that Resistin, BMI, Glucose, and age work perfectly as combination, with specificity of (85% to 90%) and sensitivity of (82% to 88%).

Whereas the proposed method of normalized features outperformed using all biomarkers in the Coimbra dataset as one combination [9,10], it is worthwhile to investigate the impact of particular biomarker combinations on the final accuracy and performance using the same method. This article provides new findings on how biomarker combination affects cancer detection effectiveness using normalized features and a FNN model. This paper includes a short summary of the Coimbra dataset's structure, a brief explanation of the normalized features with neural networks technique. Finally, the results address the impact of various biomarker combinations on BC detection effectiveness.

Methods and materials

1. Biomarker dataset

In this paper, the Coimbra Dataset biomarkers from the UCI website (University of California, Irvine) are studied to improve breast cancer detection using deep learning techniques. This dataset contains 116 samples (52 healthy subjects and 64 breast cancer patients) [5]. The data consisted of nine biomarkers and categorized as healthy and BC. The nine features are: Age (years), Glucose (mg/dL), Insulin (µU/mL), HOMA, Leptin (ng/mL), Adiponectin (µg/mL), Resistance (ng/mL), and MCP-1(pg/dL) as shown in Table 1.

By calculating the average and standard deviation of the values of each biomarker for healthy and BC samples (Table 1), we found that there is an acceptable variation between healthy and BC samples, especially the average.

Table (1). Biomarkers available in the Coimbra database and samples provided as average and standard deviation.

ID	Biomarker	Healthy samples (average ± standard deviation)	BC samples (average ± standard deviation)
V1	Insulin (µU/mL)	6.93 ± 4.81	12.51 ± 12.22
V2	HOMA	1.55 ± 1.20	3.62 ± 4.55
V3	Glucose (mg/dL)	88.23 ± 10.09	105.56 ± 26.34
V4	BMI (kg/m2)	28.31 ± 5.37	26.98 ± 4.58
V5	Leptin (ng/mL)	26.63 ± 19.14	26.59 ± 19.06
V6	Adiponectin (µg/mL)	10.32 ± 7.55	10.06 ± 6.14
V7	Resistin (ng/mL)	11.61 ± 11.33	17.25 ± 12.53
V8	MCP-1 (pg/dL)	499.73 ± 289.41	563.01 ± 380.98
V9	Age (years)	58.07 ± 18.77	56.67 ± 13.38

2. Proposed methodology

In the proposed method, we have a Coimbra database containing 9 biomarkers, after calculating the average of Biomarker (n) samples (where n is the number of the biomarker in the database) for healthy cases "Average {H_Biomarker (n)}" and BC cases "Average {BC_Biomarker (n)} ", the dataset will be reformulated into 18 features, where each composite sample will be normalized to the average of healthy cases and the average of BC cases, so we have a normalized features about how close each sample is to normal and to BC as it is shown in (Fig. 1) [9].

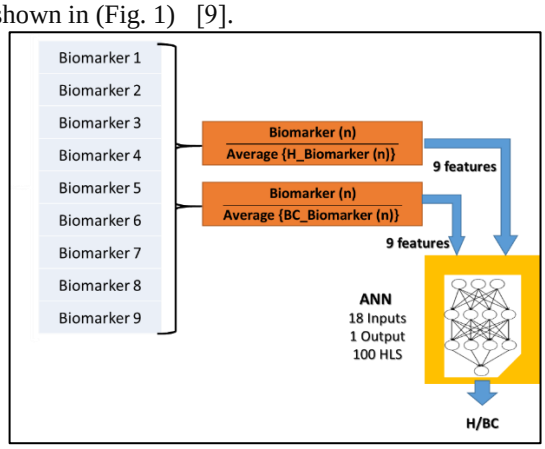

Figure (1). Scheme of the methodology used in extracting normalized features and normalizing features and using them as inputs to the neural network to obtain a case classification [9].

Biomarkers are categorized as healthy or BC in this study using a feedforward neural network. FNNs are artificial neural networks, also known as a multi-layered perceptron. FNN consists of an input layer, several concealed layers, and an output layer [11]. The FNN neural network was implemented to meet the following requirements: an input layer of 9 nodes, an output layer of one node (healthy/BC), and N-neurons in the hidden layer (HLS) were determined using Eq. 1 [12]:

$$N = \sqrt{0.43 \times m + 0.1n^2 + 2.54m + 0.77n + 0.35} + 0.5 \quad (1),$$

where m is the input size, and n is the number of output nodes.

The training process depended on the value of the desired final performance (10^{-8}), the number of epochs 500, and the training continued using the scaled conjugate gradient backpropagation, which updates the values of weights and bias according to the method of the imputed companion gradient.

3. Performance metrics

Output accuracy and efficiency were determined by accuracy (AC) and sensitivity (SE), and specificity (SP) using the formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

Where:

True Positive (TP): The number of correctly classified data indicating the presence of cancer (BC).

False Positive (FP): The number of data that are misclassified as cancer (BC) and considered as healthy (H).

True Negative (TN): The number of data classified as healthy (H) denoting the absence of cancer.

False Negative (FN): The number of data classified as healthy (H) and they are cancerous (BC).

Results and discussion

In this section we address and discuss the findings of using different combinations of the Coimbra biomarkers with the newly computed normalized features. The Coimbra dataset is initially split into 80% for training and 20% for detection. The study is predicated on researching various biomarker combinations in order to use them as input to the classification algorithm. Accordingly, the neural network topology must be updated in terms of hidden layer size for each biomarker combination (Eq. 1).

The following findings demonstrate the classification performance after initializing and training the FNN with 80% training data. The studied biomarkers were labeled with "V1-V9" as shown in Table (1).

The first form of combination uses Insulin (V1) as the fundamental biomarker, and then the V2-V9 biomarkers are introduced accumulatively. The accuracy and sensitivity of the testing results using various Insulin-based combinations (Figure 2) are 91% and 92%, respectively, which are similar to the prior findings [9,10]. Furthermore, classification accuracy varies between 83% and 88% when only Insulin, HOMA, Glucose, and BMI are used. The overall performance of detection with Insulin does not show any additional or distinct difference in accuracy, highlighting Insulin's modest role in identifying BC cases compared to Glucose, Resistin, and HOMA. The age was able to give a significant enhancement in BC detection where the age of patients with other biomarkers gave an accuracy of 92%, which meets the previous finding of [5].

Figure (2). Classification performance using Insulin as a basic biomarker.

The biomarker combinations based on HOMA are used in the second step of performance analysis. As shown in (Figure 3), the detection performance using the V2-V4 combination of HOMA, Glucose, and BMI has a reasonable accuracy and sensitivity of 85% and 89%, respectively. While V2-V5 provides the highest testing results, with an accuracy of 95% and sensitivity of 92%. In addition to the V2-V6 efficiency, with accuracy and sensitivity of 93% and 91%, respectively. The findings show that HOMA, Glucose, BMI, and Leptin play an essential part in identifying BC when their normalized values are compared to the average of healthy and BC samples. Adiponectin (V6) has a detrimental effect on classification performance because it reduces overall performance when is used.

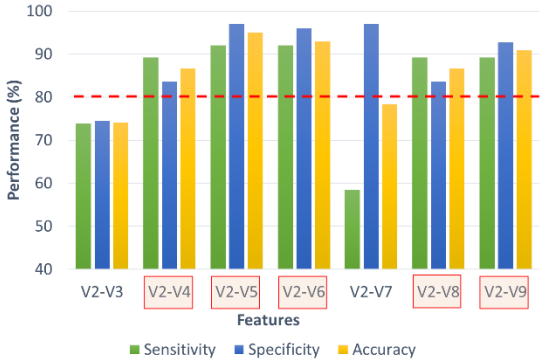

Figure (3). Classification performance using HOMA as a basic biomarker.

By excluding the V1-V2 biomarkers, there is no distinct increase in total performance as shown in Figure 4. Moreover, the absence of V2 decreases the accuracy of V3-V6 to 88% which proves the importance of HOMA and Insulin in any biomarker combination. In addition, the performance of V3-V9 around 80%, shows the importance of V1-V2 in any biomarker combination.

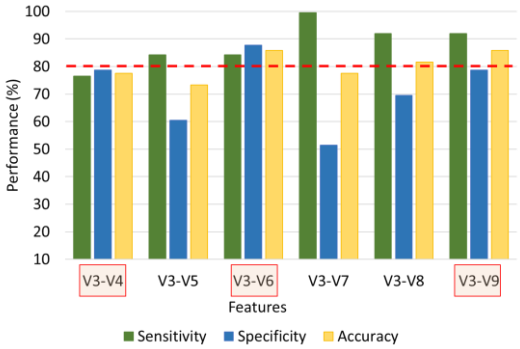

Figure (4). Classification performance using Glucose as a basic biomarker.

The last stage contains only V4-V9 as shown in Figure 5 and Figure 6. There is no clear evidence about the ability of any combination involves BMI and Leptin with Adiponectin and Resistin alone in enhancing the classification performance using normalized features. Any combination involves biomarkers between V6 and V9 cannot exceed an accuracy of 72% expect the fact that V6-V9 (Figure 6) was able to reach an accuracy of 80% and sensitivity of 100% which highlight their ability to classify BC samples but misclassify the healthy subjects.

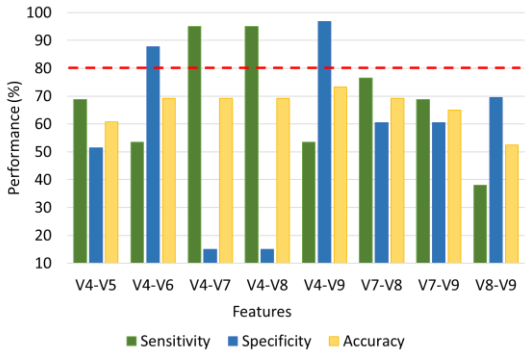

Figure (5). Classification performance using BMI and Resistin as basic biomarkers.

The absence of Insulin, HOMA and Glucose features from the dataset clearly reduces the accuracy and sensitivity of BC detection where the detection model shows a good sensitivity of (50-90%) while the specificity of healthy subjects ranges within 10-90% with an average of 40% which led to low average accuracy using different combinations.

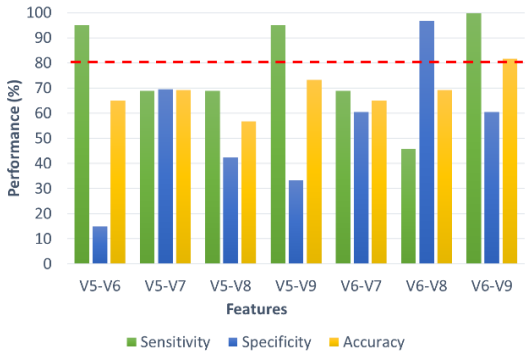

Figure (6). Classification performance using Leptin and Adiponectin as basic biomarkers.

Conclusion

The impact of biomarker combination on the breast cancer detection using the normalized features technique and neural network model based on Coimbra dataset, were investigated in this study. This paper provided a short summary of the Coimbra dataset's structure, the structure of the neural network used for classification with normalized features. The research found that HOMA, Glucose, BMI, and Leptin play an essential part in identifying BC when their normalized values are compared to the average of healthy and BC samples. The overall performance of detection with Insulin does not show any additional or distinct difference in accuracy, highlighting Insulin's modest role in identifying BC cases compared to Glucose, Resistin, and HOMA.

3

References

1. Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. International journal of cancer, 149(4), pp. 778-789.
2. Nover, A. B., Jagtap, S., Anjum, W., Yegingil, H., Shih, W. Y., Shih, W. H., & Brooks, A. D. (2009). Modern breast cancer detection: a technological review. Journal of Biomedical Imaging, 2009, pp. 1-14.
3. Sarhadi, V. K., & Armengol, G. (2022). Molecular biomarkers in cancer. Biomolecules, 12(8), pp. 1021.
4. Chang, J. C., & Kundranda, M. (2017). Novel diagnostic and predictive biomarkers in pancreatic adenocarcinoma. International journal of molecular sciences, 18(3), pp. 667.
5. Patrício M., Pereira J., Crisóstomo J., Matafome P., Gomes M., Seiça R., Caramelo F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC cancer, 18(1), pp.1-8.
6. Li Y., Chen Z. (2018). Performance evaluation of machine learning methods for breast cancer prediction. Appl Comput Math, 7(4), pp.212-214.
7. Silva Araújo V.J., Guimarães A.J., de Campos Souza P.V., Rezende T.S., Araújo V.S. (2019). Using resistin, glucose, age and BMI and pruning fuzzy neural network for the construction of expert systems in the prediction of breast cancer. Machine Learning and Knowledge Extraction, 1(1), pp.479-481.
8. Idris N.F., Ismail M.A. (2021). Breast cancer disease classification using fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition. PeerJ Computer Science, 7, pp. 11-13.
9. عيسى هـ. ا. (2023). تشخيص سرطان الثدي بالاعتماد على تحليل الدم وتقنيات الذكاء الاصطناعي. أطروحة ماجستير. جامعة دمشق.
10. الحناوي م. ف.، عيسى هـ. ا. (2023). كشف سرطان الثدي باستخدام القيم النسبية للمؤشرات الحيوية مع الشبكات العصبونية. مجلة جامعة دمشق للعلوم الهندسية.
11. Laudani A., Lozito G.M., Riganti Fulginei F., Salvini A. (2015). On training efficiency and computational costs of a feed forward neural network: a review. Computational intelligence and neuroscience, 2015(13), pp.1-4.
12. Huang M.L., Hung Y.H., Chen W.Y. (2010). Neural network classifier with entropy based feature selection on breast cancer diagnosis. J Med Syst, 34:pp. 865–73

1