Damascus University Journal for engineering sciences
Vo 41 number (3) -2025: 268-276

مجلة جامعة دمشق للعلوم الهندسية
المجلد41 العدد (3) – 2025: 268-276

Breast Cancer Detection Using a Syrian Biomarkers Dataset and Machine Learning Algorithms

Hiba Allah Essa[*1] Mhd Firas Alhinawwi[2]

*[1]. Engineer, Department of Biomedical Engineering, Faculty of Mechanical and Electrical Engineering, Damascus University.
hiba.essa@damascusuniversity.edu.sy

[2]. Professor, Department of Biomedical Engineering, Faculty of Mechanical and Electrical Engineering, Damascus University.
hiba.essa@damascusuniversity.edu.sy

Abstract:

Breast cancer is the most common type of cancer that affects women worldwide. The development of spectral analysis techniques and devices for blood analysis enhanced the use of biomarkers and what they can help in classifying, diagnosing and predicting the presence of cancers and malignant tumors. Recently, machine learning algorithms have been used with biomarkers in the diagnosis and detection of breast cancer. This research aims to create a Syrian dataset of biomarkers from Syrian hospitals and laboratories. The Syrian dataset is used to detect breast cancer using feed-forward neural networks (FNN), support vector machine (SVM) and the k-nearest neighbours algorithm (K-NN). The test performance showed that the FNN gives the best accuracy with a value of 95%, a sensitivity of 95.2% and a specificity of 94.7%. The SVM model gave an accuracy of 92.5% and a specificity of 95.2%. Compared to related work, the Syrian dataset shows the efficiency of obtained biomarkers in detecting breast cancer.

Keywords: Breast cancer; feed−forward neural networks; biomarkers; machine learning.

Received: 10/3/2023
Accepted: 29/5/2023

ISSN (online): 2789-6854
http://journal.damascusuniversity.edu.sy/index.php/index/index

كشف سرطان الثدي باستخداممجموعة بيانات سورية من المؤشرات الحيوية وخوارزميات تعلم الآلة

هبة الله عيسى*[1] محمد فراس الحناوي[2]

*[1]. مهندسة، قسم الهندسة الطبية، كلية الهندسة الميكانيكية والكهربائية، جامعة دمشق.
hiba.essa@damascusuniversity.edu.sy
[2]. أستاذ في قسم الهندسة الطبية، كلية الهندسة الميكانيكية والكهربائية، جامعة دمشق.
mhdfiras.alhinnawi@damascusuniversity.edu.Sy

الملخص:

سرطان الثدي هو أكثر أنواع السرطانات شيوعاً التي تصيب النساء في جميع أنحاء العالم. تطور تقنيات التحليل الطيفي وأجهزة تحليل مركبات الدم اتجه بالعلم إلى العمل على مفهوم المؤشرات الحيوية وما يمكن أن تساعد في تصنيف وتشخيص والتنبؤ بوجود السرطانات والأورام الخبيثة بأنوعها.مؤخراً تم الاعتماد على خوارزميات تعلم الآلة لاستخدام المؤشرات الحيوية في تشخيص سرطان الثدي واكتشافه. يهدف هذا البحث إلى إنشاء مجموعة بيانات سورية من المستشفيات والمخابر السورية والتي تعتمد على المؤشرات الحيوية المتعارف عليها من قبل الأطباء المختصين.

ثانياً، التأكد من قدرة هذه المؤشرات على كشف سرطان الثدي من خلال استخدام الشبكات العصبونية ذات الانتشار الأماميFNN ونموذجي تعلم آلة هما آلة متجه الدعمSVM وخوارزمية الجيران الأقرب K-NN في تصنيف الحالات السليمة والسرطانية منها. أظهرت نتائج الاختبار أن الشبكة العصبونية FNN تعطي أفضل دقة بقيمة 95%وحساسية95.2% ونوعية94.7%. بينما أظهر نموذج SVM دقة 92.5% ونوعية 95.2%. بالمقارنة مع الأبحاث الأخرى، تظهر مجموعة البيانات السورية مدى كفاءة المؤشرات الحيوية المستحصلة من التحاليل الطبية في كشف سرطان الثدي.

الكلمات المفتاحية: سرطان الثدي، الشبكات العصبونية ذات التغذية الأمامية، المؤشرات الحيوية، تعلم الآلة.

تاريخ الإيداع: 2023/3/10
تاريخ القبول: 2023/5/29

(ng/mL), adiponectin (µg/mL), resistin (ng/mL), and monocyte chemoattractant protein-1 (MCP-1) (pg/dL). This database is available on the UCI Machine Learning Repository. In the same research, SVM models using Resistin, BMI, glucose, and age show the best performance with specificity (85% to 90%), sensitivity (82% to 88%) and area under the curve (AUC) of (0.87 to 0.91) (Patricio et al., 2018, 1-8).

The study (Li et al., 2018, 212-21) found that RF is the best model to use with the Coimbra dataset with accuracy of 74.3%compared to different machine learning algorithms such as SVM, RF, neural networks and LR. Moreover, (Silva et al., 2019, 479-481) used the fuzzy neural networks approach to detect breast cancer based on Coimbra data and reached a final accuracy of 62% and a sensitivity of 78.3% for cancer detection.

Later, a fuzzy decision tree (FDT) model was used to classify the Coimbra dataset, and the results reached a final accuracy of 70.68% and a sensitivity of 78.3% (Idris et al., 2021, 11-13). Recently, (Alnowamiet al., 2022, 104-110) used sequential backward selection (SBS) as a prior step of the SVM model, where features are sequentially removed from the dataset until the best performance is obtained. This research shows an accuracy of 92% and a sensitivity of 94% in detecting breast cancer.

The study (Wong et al., 2019, 332-341) improved the detection of cancers based on the Cohen dataset (Cohen et al., 2018, 926-930)by designing the CancerA1DE algorithm, but the sensitivity of detecting breast cancer did not exceed 77% while the sensitivity of ovarian, liver, stomach, pancreas, and esophagus cancer detection ranged between 69% and 99%. Cohen dataset was obtained from 1817 samples (blood test), 626 samples belong to BC. Cohen dataset includes 40 compounds and biomarkers. Some of these compounds: OmegaScore, CA-125 (U/ml), carcinoembryonic antigen (pg/ml) (CEA), CA19-9 (U/ml), Prolactin (pg/ml), Interleukin-8 hepatocyte growth factor concentrate Concentration (HGF) (pg/ml), osteopontin (OPN) (pg/ml), myeloperoxidase (ng/ml), Tissue Inhibitor of Metalloproteinases (pg/ml) (TIMP-1).

Researcher Gao and his colleagues (2020) presented a survey of seven important biomarkers in the early detection of breast cancer: micro ribonucleic acid (miRNA), circulating cell free deoxyribonucleic acid

Introduction:

Breast cancer (BC) is the most common type of cancer among women in general, and is a disease in which breast cells get out of control while dividing and growing. There were more than 2.26 million new cases of breast cancer in women in 2020 (Ferlay et al., 2021, 778-789). Breast cancer can be diagnosed through a mammogram, ultrasound, magnetic resonance imaging (MRI), and biopsy. Although mammography is the best method for detecting and locating breast cancer, its extensive and frequent use makes it dangerous for women, especially under 30 years old (Nover et al., 2009, 1-14). On the other hand, through the development of spectroscopy techniques and devices for blood analysis, science has been oriented to work on the concept of biomarkers and what can help in classifying, diagnosing and predicting the presence of cancers and malignant tumors of all kinds (Sarhadi et al., 2022, 1021).

The National Cancer Institute defined the biomarker as a molecule in the blood that is a marker of whether the body is functioning normally or abnormally. Thus the concept of a vital marker includes the presence of a physical or biological sign of a condition or disease, such as breast cancer. Blood-based biomarkers are non-invasive, measurable features that can be used in the detection, pre-diagnosis, and staging of cancer, such as osteopontin (OPN), cancer antigen CA125, and CA15-3 (Chang et al., 2017, 667).

Biomarker data is enormous and interrelated and samples of biomarkers cannot sometimes be distinguished between healthy and cancerous, so it was crucial to invest in machine learning (ML) techniques in classifying biomarker data. The most popular ML algorithms for breast cancer detection and diagnosis based on available features are support vector machines (SVM), random forests (RF), logistic regression (LR), and K-nearest neighbors (KNN). Traditionally, classification outputs are evaluated using performance metrics like sensitivity, specificity, and accuracy (Dlamini et al., 2020, 2300-2302).

The paper (Patricio et al., 2018, 1-8) presented the Coimbra dataset containing samples of 52 healthy cases and 64 breast cancer patients. The dataset contains body mass index (BMI) (kg/m2), age (years), Glucose (mg/dL), Insulin (µU/mL), homeostatic model assessment (HOMA), leptin

In this paper, biomarkers will be classified into healthy or BC case using a feedforward neural network (FNN). FNN is a type of artificial neural network. It is often referred to as a multi-layered perceptron. FNN consists of an input layer, a number of hidden layers, and an output layer (Laudani et al., 2015, 1-4). The FNN neural network is built according to the following specifications: the input layer contains 7 nodes, the output layer contains one node (healthy or BC), the size of the hidden layer (HLS) is equal to 5 neurons, where the number of neurons N for the hidden layer is calculated from the equation (1) according to (Huang et al., 2010, 865–873):

$$N = \sqrt{0.43 \times m + 0.1n^2 + 2.54m + 0.77n + 0.35} + 0.51 \quad (1)$$

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

The training process depends on the value of the desired final performance (10^{-8}), the number of iterations 500, and the training continued using the scaled conjugate gradient backpropagation, which updates the values of weights and bias.

3. K-nearest neighbor algorithm:

The k-NN algorithm is a simple ML technique that uses the distribution of labelled data points on a space to classify unlabeled data using the nearest distance from their neighbors. The KNN model consists of three parameters, the first is the "k" value which indicates the number of closest data points to the reference label, the second is the distance or "D" value between adjacent data points and the reference label, and the third parameter indicates the distance measurement function or method. The principle of KNN is similar to the voting procedure. To classify an untrained data, KNN calculates the distance to the labelled reference point and finally classifies the point according to the nearest neighboring reference label. In this paper, the K-NN model is used with k=3 and the Jaccard-type distance function (Dlamini et al., 2020, 2300-2302).

4. Support Vector Machine:

The SVM is an ML technique that can be used for both classification and regression. It mainly has two variables to support linear and nonlinear problems. A linear SVM has no kernel and finds a small-margin linear solution to the problem. SVM is used with kernels when the solution is not linearly separable. SVM is a supervised learning technique widely used in the classification of labelled data (DNA), adenomatous polyposis coli (APC) gene promoter methylation, 14-3-3σ promoter methylation, CA153, CEA, and prostate-specific antigen (PSA) (Gao et al., 2020, 97-108).

The subject of biomarkers-based cancer detection is a recent methodology that does not require surgical intervention and considers a promising diagnostic tool. This research aims to create a Syrian dataset from Syrian hospitals and laboratories, which relies on crucial biomarkers recognized by specialized doctors and clinicians. Then, ensure the ability of these biomarkers to detect breast cancer using a neural network and two machine learning models (SVM and K-NN) and classify the normal and BC cases. Finally, the research will highlight the efficiency of the proposed dataset and compare it to Coimbra and Cohen dataset using the performance of employed ML approaches.

Methods and materials

1. Syrian Dataset of biomarkers:

Syria Breast Cancer Dataset (SBCD) contains 160 samples collected from hospitals and laboratories in the Syrian Arab Republic for a number of women who were diagnosedwith breast cancer. Despite the sanctions imposed on Syria, which led to a decrease in the amount of raw materials needed for such analyzes, we were able to collect these samples. The samples included 80 breast cancer patients and 80 healthy samples. Samples were collected from Al-Bayrouni Hospital in Damascus and Tishreen University Hospital in Lattakia for a community sample of women between the ages of 45 and 65 years. The study included patients who were under treatment or before starting treatment, and it was previously confirmed that they had BC. Patients who had mastectomy but continued treatment were excluded because there are possibilities of being treated but still has the biomarker values of BC or after the lumpectomy, the tumor moved to the second breast or to another place in the body. Samples were collected during the years 2021-2023. The presence of cancer was confirmed through mammographic x-rays and biopsy for those over 40 years old, and by echocardiography and biopsy for those under 40 years old. The dataset contains seven biomarkers: glucose (mg/dL), alanine transaminase (ALT), aspartate aminotransferase, cancer antigen CA15-3, carcinoembryonic antigen (CEA), and phosphatases. Alkaline phosphatase (ALP), Calcium (Ca).

2. Neural network:

Table (1) The Syrian dataset and data structure in the terms of the average and standard deviation.

Biomarker	healthy samples mean ±) standard (deviation	BC samples mean ±) standard (deviation
Glucose (mg/dl)	82.66±9.88	112.2±27.24
ALT (U/l)	21.36±5.59	20.95±11.24
AST (U/l)	20.7±4.77	22.52±8.98
CA15-3 (U/ml)	15.76±4.65	60.55±65.78
CEA (ng/ml)	2.53± 1.11	13.42±12.28
ALP (U/l)	91.4± 28.16	140.2±57.7
Ca (mg/dl)	9.71±0.58	8.8±1.34

Glucose, CA15-3, CEA and ALP show a clear differences in the mean values of healthy and cancerous but with some overlaps in the values in terms of standard deviation.Data were pre-processed and prepared to be classifiable using FNN, K-NN and SVM. The data were divided into 75% (120 samples) as training samples and 25% (40 samples) as testing samples equally between normal and BC.

In the beginning, the training process of the FNN was carried out according to the parameters previously mentioned in Paragraph (2) of the Methods and Materials section, and K-NN and SVM models were built based on the training samples.

The training results (Table 2) show a final classification accuracy of 97.5%, a sensitivity of 96.6%, and a specificity of 98.4% using FNN.

Table (2) Performance of the training process using the proposed ML tools and the Syrian dataset.

ML approach	AC	SE	SP
FNN	97.5%	96.6%	98.4%
SVM	95%	94.9%	95.1%
K-NN	92.5%	91.7%	93.8%

(Dlamini et al., 2020, 2300-2302).In this study, the used kernel is linear with scale of 1.

5. Performance metrics:

The accuracy and efficiency of detection performanceare determined by accuracy (AC), sensitivity (SE), and specificity (SP) and are calculated using the following formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{(2)}$$

$$Sensitivity = \frac{TP}{TP + FN} \quad \textbf{(3)}$$

$$Specificity = \frac{TN}{TN + FP} \quad \textbf{(4)}$$

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) but are healthy (H).

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

False Negative (FN): The number of data classified as healthy (H) but the data belong to (BC).

Results and discussion:

By collecting data from the Syrian hospitals and getting a final dataset of 7 biomarkers and 160 samples (80 healthy and 80 diagnosed with BC). The biomarkers show a clear distinction between the samples of healthy and BC by calculating the average and standard deviation as shown in table (1).

While the process of initializing SVM and KNN models using the training samples shows an accuracy of 95% and 92.2%, respectively. These results show the ability of biomarkers and their samples to diagnose healthy and BC cases.

Test samples are used to ensure the effectiveness of using biomarkers in detecting breast cancer using untrained samples. The test results in Table (3) show that the FNN model gives the best accuracy with a value of 95%, a sensitivity of 95.2%, and a specificity of 94.7%. While the SVM model shows an accuracy of 92.5% and a sensitivity of 95.2%, which indicates that most of the BC samples are detected, but some healthy samples are classified as BC. While K-NN failed to detect several BC and healthy samples and misclassified them with a detection accuracy of 85%, a sensitivity of 87.5%, and a specificity of 81.2%.

Table (3) Performance of the testing procedure using the proposed ML tools and the Syrian dataset.

ML approach	AC	SE	SP
FNN	95%	95.2%	94.7%
SVM	92.5%	95.2%	89.5%
K-NN	85%	87.5%	81.2%

Table (4) Comparison of performance metrics between the different datasets and the Syrian data.

Ref.	Dataset	ML approach	AC	SE
(Silva et al., 2019, 479-481)	Coimbra	fuzzy neural networks	62%	78%
(Idriset al., 2021, 11-13)	Coimbra	fuzzy decision tree (FDT)	70.69 %	78.3%
(Alnowami et al., 2022, 104-110)	Coimbra	SBS+SVM	٪92	٪94
)Cohen et al., 2018, 926-930(	Cohen	Cancer SEEK	-	- ٪75 ٪65
(Wong et al., 2019, 332-341)	Cohen	Cancer A1DE	-	٪75
Proposed study	Syrian	FNN	95%	95.2%

Since there are no common biomarkers between the Syrian dataset and the two Cohen and Coimbra datasets, therefore, it is not possible to set criteria for a comprehensive comparison approach between the results of breast cancer detection. Since each dataset was created based on samples from the hospitals of the studied community, as well as the type of blood analysis used in those hospitals, which are considered to have a role in detecting breast cancer, according to the considerations of the researchers and authors. But the importance of biomarkers can be emphasized by performing classification using machine learning algorithms and the used dataset. By examining the difference in the classification performance using Coimbra and Syrian datasets, the SVM performance after optimizing the selection of features used in the classification of the Coimbra dataset shows a breast cancer detection accuracy of 92% and a sensitivity of 94% (Alnowami et al., 2022, 104-110). While the Syrian dataset using SVM directly with the raw features shows higher performance with a sensitivity of 95.2%.

results of enhancing breast cancer detection in the related work using Coimbra and Cohen datasets and machine learning tools are shown in Table (4). The comparison between the Syrian dataset and related work can clearly note to what extent the Syrian dataset shows the ability to detect breast cancer in terms of accuracy and sensitivity. Regarding Coimbra data, (Idris et al., 2021, 11-13) obtained a final accuracy of 70.69% and (Alnowami et al., 2022, 104-110) had an accuracy of 92% and a sensitivity of 94%. While the results of (Cohen et al., 2018, 926-930) and (Wong et al., 2019, 332-341) studies did not exceed 75% of sensitivity in detecting breast cancer. In the end, the Syrian dataset achieved an accuracy of 95% and a sensitivity of 95.2%, thus showing the efficiency of biomarkers obtained from Syrian hospitals in detecting breast cancer.

experience of doctors working in them. These data showed a significant superiority in detecting breast cancer based on three machine learning techniques with a maximum accuracy of 95% and a sensitivity of 95.2% using the FNN model by using 160 samples of biomarkers and splitting them into 120 samples for training and 40 samples to test the detection of breast cancer. In comparison with the related work regarding the use of the Coimbra and Cohen datasets, we have shown the extent of the efficiency of biomarkers using the Syrian dataset in detecting breast cancer.

Funding: This Research Is funded by DamascusUniversity – Funder No (501100020595).

Thus the current research proves that using biomarkers CA15-3, Glucose, ALT, CEA, ALP, Ca and AST as an integrated dataset of samples has an efficient impact on detecting breast cancer with machine learning algorithms. Future work may focus on studying and collecting more important biomarkers that have the ability to detect several types of cancer.

Conclusion:

Medical decision support using machine learning tools for biomarker-based cancer diagnosis has become an important and promising area of clinical studies. In this research, a Syrian dataset is built, obtained from Syrian hospitals and supported by the

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. Dlamini Z., Francies F.Z., Hull R. Marima R. (2020). Artificial intelligence (AI) and big data in cancer and precision oncology. Computational and Structural Biotechnology Journal, 18, pp.2300-2302.
6. 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.
7. Li Y., Chen Z. (2018). Performance evaluation of machine learning methods for breast cancer prediction. ApplComput Math, 7(4), pp.212-214.
8. 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.
9. 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.
10. Alnowami, M. R., Abolaban, F. A., &Taha, E. (2022). A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer. Journal of Radiation Research and Applied Sciences, 15(1), pp. 104-110.
11. Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., ... & Papadopoulos, N. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science, 359(6378), pp. 926-930.
12. Wong, K. C., Chen, J., Zhang, J., Lin, J., Yan, S., Zhang, S., ... & Yu, J. (2019). Early cancer detection from multianalyte blood test results. Iscience, 15, pp. 332-341.
13. Gao, Y., Liu, M., Shi, S., Sun, Y., Li, M., Zhang, M., ... & Cancer Biomarker Assessment Working Group Jing Wang XinyueLuoRunjinCai. (2020). Diagnostic value of seven biomarkers for breast cancer: an overview with evidence mapping and indirect comparisons of diagnostic test accuracy. Clinical and Experimental Medicine, 20, 97-108.
14. Laudani A., Lozito G.M., RigantiFulginei 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.
15. 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.