JOURNAL OF EPIGENETICS

Journal homepage: http://jep.usb.ac.ir/

Document Type: Original Article

Early Detection of Breast Cancer in Women Using a Cost-effective Procedure

Yahya Kord Tamandani [a,*]

[a]*Department of Computer Science, University of Sistan and Baluchestan ,Zahedan, Iran; Email: yahya.kord@cs.usb.ac.ir*

[*]*Corresponding author at: Email Address: yahya.kord@cs.usb.ac.ir*

ARTICLE INFO

Article history:

Received 20 March 2020

Accepted 18 April 2020

Available online 18 April 2020

DOI: 10.22111/jep.2020.34052.1022

KEYWORDS:

Breast cancer, Prediction model, Anthropometric data, SVM, K-NN.

A B S T R A C T

Breast cancer is considered to be the second most common type of cancer affecting the female population worldwide. It is estimated that more than 508 000 women died in 2011 as a result of breast cancer. The survival rates of breast cancer are lower in less developed countries mainly due to the absence of early detection methods resulting in a great percentage of women showing with late-stage disease. Early detection and medical diagnosis are known to be the most effective solution to minimize the risk of tumor development and progression. There are different methods for Early detection of breast cancer which include screening tests and clinical breast exams performed by a well-trained health professional. Due to a lack of facilities and cost, many women in less developed countries may not be able to use the mentioned methods. The objective associated with this research was to achieve an affordable and cost-effective prediction model of breast cancer based on anthropometric data and parameters that can easily be collected in a routine and regular blood test. For every one of the 166 individuals number of clinical features such as age, Body Mass Index (MBI), serum glucose levels, plasma levels of insulin, etc. were measured and observed. Various learning algorithms including Support Vector Machines (SVM), K-Nearest Neighbors (K-NN) and logistic regression(LR), etc. have been applied and compared with one another. The result shows that SVM and K-NN models perform well and allow prediction of breast cancer in women with accuracy more than 78%, the sensitivity of 78% and 79%, and Specificity value is 77% and 79% respectively.

Introduction

Cancer occurs due to unnatural changes or mutations, in the genes accountable for managing and controlling the growth of cells and maintaining them healthy. Typically, the cells replace themselves via an organized procedure. healthy new cells in our body are formed and the old ones die out. However, in some cases, this process might go wrong. Particular genes are turned on and others are turned off by mutations in a cell. That modified cell gains the capacity to keep dividing and not having control or order, resulting in a lot more cells identical to it and producing a tumor. Breast cancer that is more common in women indicates a malignant tumor having formed from cells in the breast. It is known to be the second most cause of death in the female population worldwide. survival rates of Breast cancer vary tremendously worldwide, over 80% in North America, Sweden, and Japan, approximately 60% in middle-income countries and less than 40% in low-income countries

* Corresponding author: Tel +9831136396.
E-mail address: yahya.kord@cs.usb.ac.ir

(Coleman, Quaresma et al.2008). The high death rates in low-income and less developed countries are mostly due to the absence of facilities for early detection, resulting in a huge percentage of females having late-stage cancer. Hence the major health issue here is to find these breast cancers early in individual women with the hope of decreasing their chances of dying from breast cancer and to provide them with better treatment options. Focusing on early detection is essential in order to prevent the growth of breast cancer. Hence a number of researches using different strategies have been done on it (Guan Z, Yu H et al.2019). There are also several medical imaging techniques used for the early detection of breast cancer. One Popular example of such techniques is mammography. Mammography is great at identifying little tiny white dots on a mammogram which represents calcification of the earliest stage of breast cancer. Mammography has been shown to reduce mortality from breast cancer. Other medical imaging techniques for breast cancer include breast ultrasound and magnetic resonance imaging (MRI). Table 1 shows the comparison of most frequently used image screening techniques and their limitations for the detection of breast cancer (Wang L.2017). Indeed, breast cancer screening strategy is a vital method allowing for early detection and diagnosis. However effective models of prediction that are based on gathered records from routine blood tests and consultation are required to offer an essential contribution through additional tools for screening. The objective associated with this research was to achieve a cost-effective prediction model for the detection of breast cancer based on anthropometric data and parameters that can easily be collected in a routine and regular blood test.

Table 1 - Different Breast Screening Methods aand Their Limitations

Type	Usage	Sensitivity*	Specificity*	limitations	Time Required
Mammography	Screening tool for early detection of breast cancer in women	67.8%	75.0%	Low sensitivity and specificity, increase in tissue density drops the sensitivity level	Less than a minute
Ultrasound	Evaluation of lumps detected in mammography; Not appropriate for bony structures	83.0%	34.0%	Low sensitivity; well-trained operator is needed throughout examination; resolution of image is low;	10–20 minutes
Magnetic Resonance Imaging (MRI)	Using a large magnet and radio waves in order to look at small details of soft tissues.	94.4%	26.4%	Unable to detect all type of cancer such as lobular carcinoma; costly;	40–60 minutes
Computed Tomography (CT)	Determining and imaging distant cancer	91%	93%	Risk of radiation; costly scanner;	5 minutes
Positron Emission Tomography (PET)	Imaging test to show how tissues and organs in body are functioning	61.0%	80.0%	Radioactive tracer injection	90–240 minutes

* Sensitivity and specificity are related to the types of cancer and breast composition.

Related work

A number of recent researches using different cancer biomarkers have been done on early detection of breast cancer (Bartosik, Jirakova et al .2019). In (Crisóstomo J, et al.2016) classical biomarkers for breast cancer have been reviewed, the tactics for the development of new biomarkers have been highlighted. In (Patrício, M., Pereira et al.2018) based on only four biomarkers that are glucose, resistin, BMI, and age the prediction of breast cancer has been done. The existence of breast cancer in females has been predicted with a sensitivity of 88% and specificity of 90%. In (Santillan-Benitez JG, et al.2013) also four variables namely BMI, leptin, L/A ratio, and cancer antigen (CA) 15-3 were analyzed in order to predict which women are at high risk to developing breast cancer. For the prediction model, a sensitivity of 83.3% and specificity of 80% was reported. In (Assiri AM, Kamel et al.2015) CA15-3, hsCRP, resistin, visfatin, and leptin were used as biomarkers for

early detection of breast cancer in women. They measured by enzyme-linked immunosorbent assay (ELISA). ROC (receiver operating characteristic curve) analysis for serum values of leptin shown AUC=0.795; 95% CI, 0.724-0.866. Resistin revealed AUC(area under the curve)=0.875; 95% CI, 0.821-0.928. The study stated that visfatin more than 12.2ng/mL confirmed a sensitivity and specificity of 97.6% and 92.6%, respectively, and AUC=0.724; 95% CI, 0.643-0.804. Different machine learning techniques have been implemented to databases that are publicly available in the UCI Machine Learning Repository (Toprak, A. 2018). In (Chaurasia, Pal .2014) the performance of various supervised learning classifiers, including Naive Bayes, Decision Tree, and SVM-RBF were evaluated in an effort to discover the most effective classifier in breast cancer datasets. The research revealed that the SVM-RBF kernel has higher accuracy compared with other classifiers; according to the study, the level of accuracy was 96.84% on the Wisconsin Breast Cancer Dataset (WBCD). The Model offered in this research is based on a population of females early-diagnosed with breast cancer. The collected data used in this work are described in the Dataset section. In the Methodology section contains characteristics of different classification algorithms that were implemented in the article. In the result section, the overall performance of each model is discussed.

Dataset

The dataset used in this paper is publically available (Patrício, M., Pereira et al.2018). To create the dataset samples are taken from Females lately recognized as having breast cancer from the Gynecology Department of the University Hospital Centre of Coimbra (CHUC). For any individual, the positive test result was obtained from positive mammography and was verified histologically. All of the samples were collected prior to the treatment. A total of 52 healthy volunteers in addition to 62 females having breast cancer were included in the dataset. Blood samples for all the individuals were collected following overnight fasting at the same time. Data that are used as a biomarker in this study are given in Table 2.

Table 2 - List of Biomarkers aand Method of Analyzing Used iin This Study

Breast Cancer biomarkers	Description	Method of analyzing	values
Age	Age of patient	-	numeric
BMI	The BMI, expressed in kg/m2	Determined by dividing the weight by the squared height	numeric
Glucose	Serum Glucose levels were deter- mined by an automatic analyser	Olympusn kit - Diagnóstica Portugal, Produtos de Diagnós- tico SA, Portugal	numeric
Insulin	Plasma levels of Insulin	ELISA kit using Mercodia Insulin ELISA, Mercodia AB, Sweden	numeric
HOMA	Homeostasis Model Assessment (HOMA) index was calculated to evaluate in- sulin resistance	HOMA = logarithm ((If) x (Gf)) / 22.5, where (If) is the fasting insulin level (μU/mL) and (Gf) is the fasting Glucose level (mmol/L)	numeric
Leptin	Serum values of Leptin	Kit of Duo Set ELISA Development System Human Leptin R&D System, UK	numeric
Adiponectin	Serum values of Adiponectin	Kit of Duo Set ELISA Development System Human Adiponectin R&D System, UK	numeric
Resistin	Serum values of Resistin	Kit of Duo Set ELISA Development System Human Resistin R&D System, UK	numeric
MCP_1	Chemokine Monocyte Chemoattractant Protein 1 (MCP-1)	Kit of Human MCP-1 ELISA Set, BD Biosciences Pharmingen, CA, EUA	numeric

Methodology

Several classification algorithms were implemented in this study namely; Logistics Regression, Naïve Bayes, K- Nearest Neighbor (KNN), Discriminant Analysis, Support Vector Machines (SVM), and Random Tree algorithms. The characteristics of different classification algorithms applied in this work are given in Table 3 (Mathworks) .

DOI: 10.22111/jep.2020.34052.1022

Table 3 - Characteristics of Different Classification Algorithms

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Medium Gaussian SVM	Binary: Fast Multiclass: Slow	Binary: Medium Multiclass: Large	Hard	Medium
Weighted KNN	Medium	Medium	Hard	Medium
Linear Discriminant	Fast	Low	Easy	Low
Logistic Regression	Fast	Medium	Easy	Low
Medium Tree	Fast	Low	Easy	Medium
Kernel Naive Bayes	Slow	Medium	Easy	Medium
RUSBoost Trees	Fast	Low	Hard	Medium

Result

Considering Table 4. and Figure 1. Two classification algorithms namely Gaussian SVM and Weighted KNN have the best performance in comparison to others. Weighted KNN has the highest rate in terms of sensitivity that was near 80% and the specificity rate was about 79%. The algorithm also had the best performance in case of precision that was close to 82%. However, taking into consideration the value of AUC (area under the curve) the Gaussian SVM has the highest value which was 0.85.

Table 4 - Performance of Different Classification Algorithms

Methods	Accuracy	Sensitivity	Specificity	Error rate	Precision	AUC
Medium Gaussian SVM	0.7844	0.7812	0.7692	0.2155	0.8095	0.85
Weighted KNN	0.7844	0.7968	0.7884	0.2155	0.8196	0.82
Logistic Regression	0.7672	0.75	0.7884	0.2327	0.8135	0.78
Ensemble RUSBoosted Trees	0.75	0.7656	0.7307	0.25	0.7777	0.82
Linear Discriminant	0.7413	0.7031	0.7884	0.2586	0.8035	0.78
Medium Tree	0.7155	0.7343	0.6923	0.2844	0.7460	0.77
Kernel Naive Bayes	0.6724	0.6562	0.6923	0.3275	0.7241	0.77

Figure 1. indicates the performance of different classification algorithms in case of accuracy, sensitivity, specificity, error rate, precision, and the value of AUC.

DOI: 10.22111/jep.2020.34052.1022

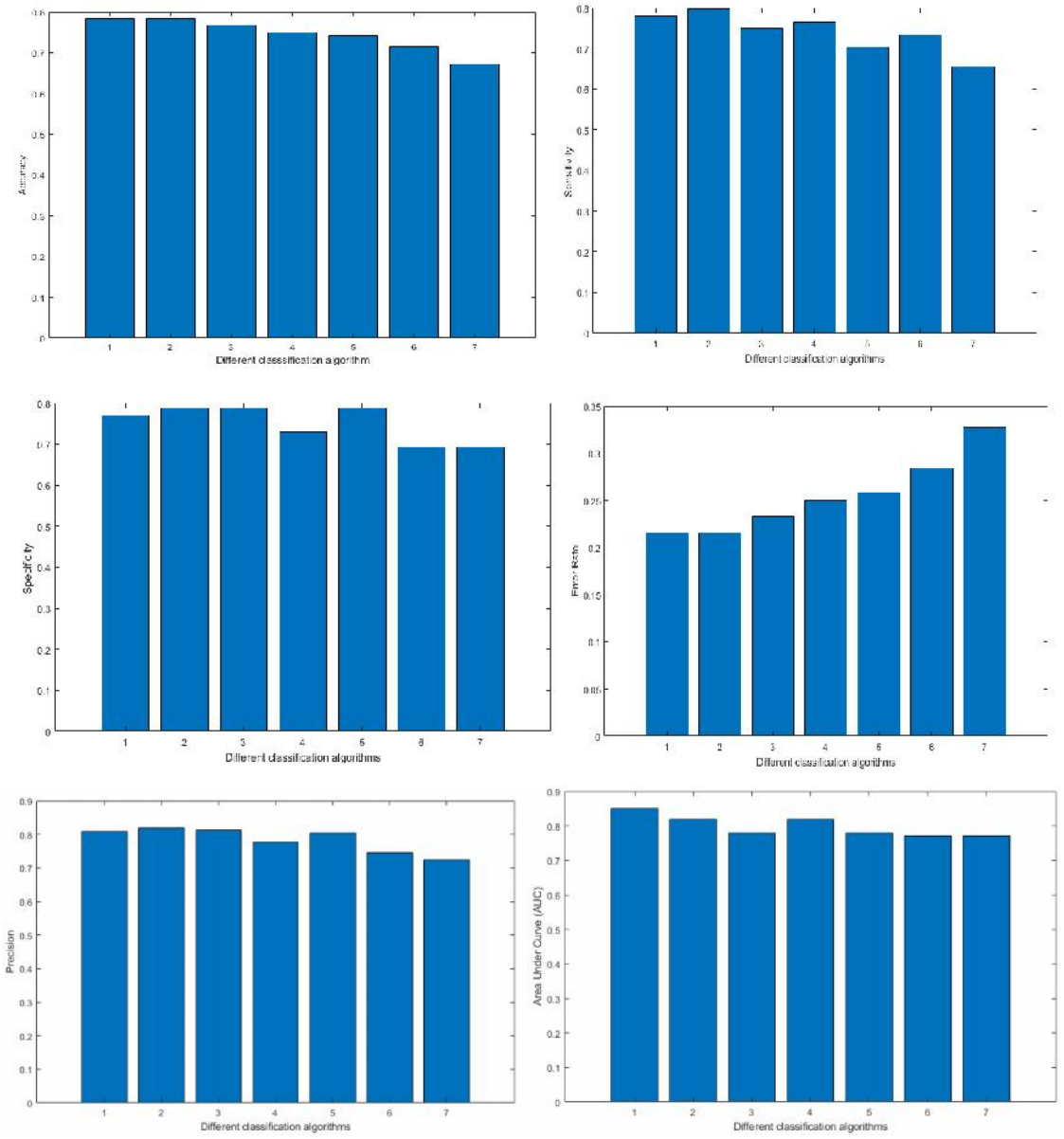

Fig.1 - performance of Different Classification Algorithms (1=Medium Gaussian SVM;2=Weighted KNN;3=Logistic Regression;4=Ensemble RUSBoosted Trees;5=Linear Discriminant;6=Medium Tree;7=Kernel Naive Bayes)

Figure 2. displays the performance of different classification algorithms in the case of ROC curves generated with nine predictors as mentioned earlier in the cross-validation procedure.

DOI: 10.22111/jep.2020.34052.1022

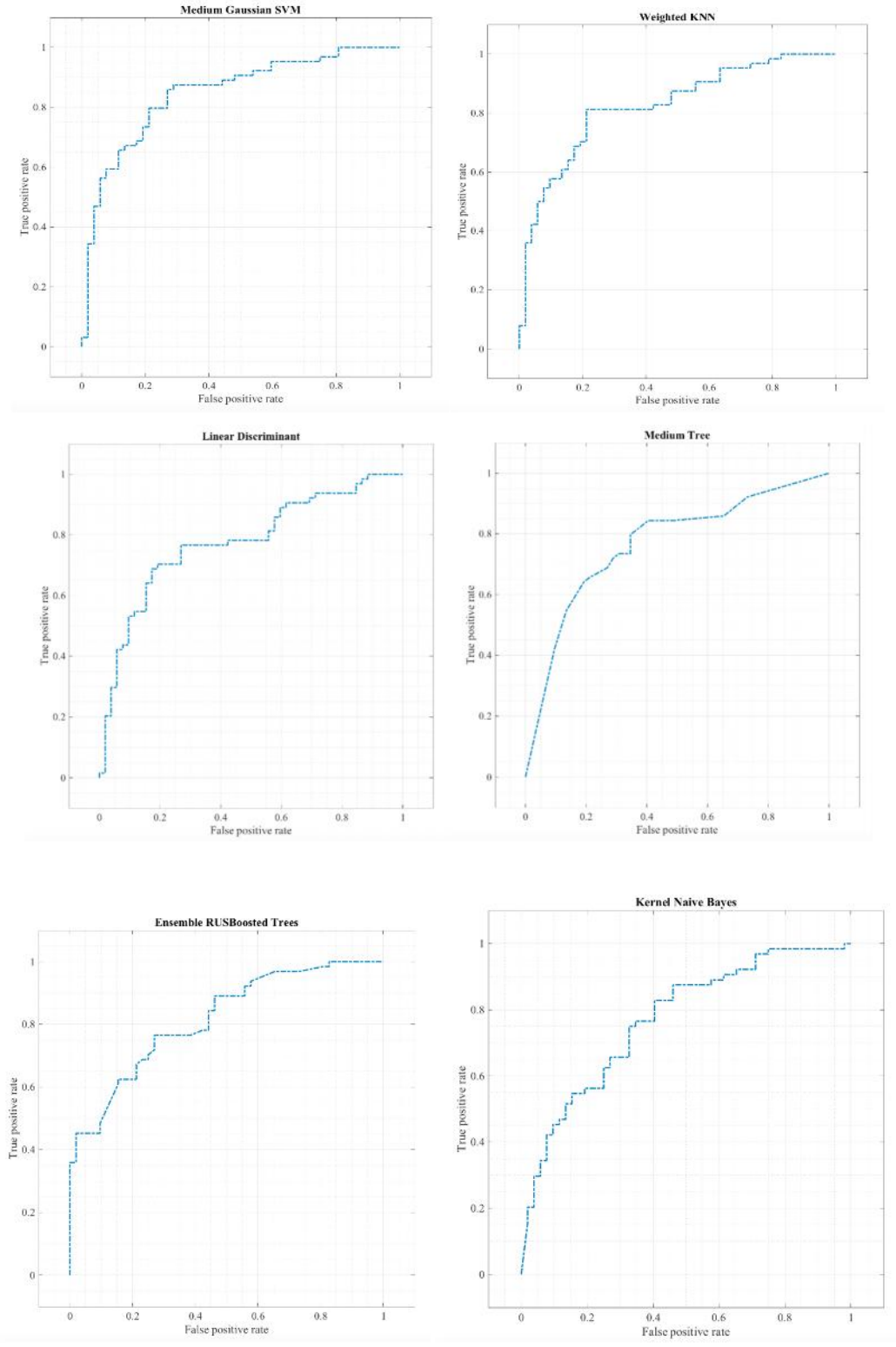

Fig.2 - ROC Curves Generated with Nine Predictors in the Cross-validation

DOI: 10.22111/jep.2020.34052.1022

Conclusion

Based on anthropometric data and parameters that can easily be collected in a routine and regular blood test the prediction of breast cancer in women can be achieved with a sensitivity value of 80% and specificity value of 79%. This cost-effective procedure can be easily done with a simple blood test that could potentially improve the health of many women especially residing in underdeveloped countries with a limited number of resources and facilities for early detection of breast cancer.

References

Assiri, A. M., & Kamel, H. F. (2015). (2015). Evaluation of diagnostic and predictive value of serum adipokines: Leptin, resistin and visfatin in postmenopausal breast cancer. *Obes Res Clin Pract.*, 10(4):442-53.

Bartosik, M., & Jirakova, L. (2019). Electrochemical analysis of nucleic acids as potential cancer biomarkers. *Current Opinion in Electrochemistry*, *14*, 96–103.

Coleman, M., Quaresma, M., Berrino, F., Lutz, J.M., Angelis, R., Capocaccia, R., & et al. (2008). *Cancer survival in five continents: a worldwide population-based study (CONCORD) Lancet Oncol.* 9:730–756.

Chaurasia, V., & Pal, S. (2014). (2014). Data mining techniques: to predict and resolve breast cancer survivability. *Int J Comput Sci Mobile Comput.*, 3: 10–22.

Crisóstomo, J., & et al. (2016). Hyperresistinemia and metabolic dysregulation: the close crosstalk in obese breast cancer. *Endocrine*, 53(2):433-42.

Guan, Z., Yu, H., Cuk, K., Zhang, Y., & Brenner, H. (2019). Whole-Blood DNA methylation markers in early detection of breast cancer: A systematic literature review. *Cancer Epidemiol Biomarkers Prev. 28*, 496–505.

"Mathworks.com." *Supervised Learning Workflow and Algorithms - MATLAB & Simulink - MathWorks* , it.mathworks .com/ help/ stats/supervised-learning-machine-learning-workflow-and-algorithms.html.

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).*

Toprak, A. (2018). Extreme Learning Machine (ELM)-Based Classification of Benign and Malignant Cells in Breast Cancer. *Medical Science Monitor, 24*, 6537–6543.

Wang, L. (2017). (2017). Early diagnosis of breast cancer. *Sensors. doi: 10.3390/s17071572. 17(7)*: 1572.

DOI: 10.22111/jep.2020.34052.1022