An Assessment of Predictive Power of Biomarkers in Breast Cancer Diagnosis

Saulo Mendes de Melo
Applied Computational Intelligence
PPGETI
Fortaleza, Brazil
saulomelo96@hotmail.com

***Abstract*—Every year breast cancer claims the lives of a great number of women worldwide. The clinical research on the area has been long providing insights on causes and correlated attributes and indicators, and the existence of strong biomarkers is known in the scientific community. This work presents an statistical analysis and evaluation of the predictive power of a few clinical indicators extracted from typical blood sample analysis, in both patients of breast cancer and a control group, provided by the publicly available Breast Cancer Coimbra Data Set. It was concluded that the Glucose and Resistin present an abnormal profile for breast cancer patients and using these indicators in classification models such as Logistic Regression, SVM and Artificial Neural Networks has provided accurate predictions of presence of Breast Cancer.**

***Index Terms*—Breast cancer, Resistin, SVM, Neural Networks**

I. Introduction

Cancer is among the leading causes of death worldwide and every year its various types claim a great number of lives. Amongst women victims, breast cancer (BC) is the leading type, with as much as 322.000 deaths in the year 1990 [1]. A study published in 2019 analyzing the US population has found that approximately 13% of women will be diagnosed with BC in their lifetime, and as much as 1 in 39 women will eventually succumb to it [2].

The incidence of BC among women increases with age. The probability of a diagnose for woman in the age of 20 is of 0.1%, and goes to 3.0% by the age of 80. The rates of incidence and mortality also hold some relationship with ethnicity. The incidence rate is higher among whites (130.8 per 100.000), but for blacks, while incidence is lower the mortality is up to 40% higher (28.4 per 100.00) [2]. Other risk factors associated are late first birth (post 30), nulliparity, use of oral contraceptives and having first and/or second-degree relatives diagnosed [3].

Although non-clinical factors provide a valuable insight, other studies have focused on the analysis of biological indicators more tipically related to medical assessment. Dalamaga [4] has analyzed Resistin as biomarker, and it's links to obesity and cancer. Crisóstomo et al. [5] provided a study of biomarkers in the context of BC, where groups were separated for both obese and non-obese control and patients, and an extensive statistical analysis was made with indicators such as Glucose, Insulin, Resistin, among others. A set metabolic characteristics was found in obese women with BC, which includes glucose, insulin disorders and other anomalies.

The present work aims to provide an analysis of biomarkers on a clinical dataset, evaluate how correlated they might be with BC detection in patients. We also have evaluated the predictive performance of various models when classifying between Healthy and Patient, between two scenarios of dimensionality reduction, principal components classification and fitting into a subset of features.

II. Methodology

A. Data

The dataset used for this work is the publically available Breast Cancer Coimbra Data Set, and it was collected from women diagnosed with BC before any surgery or treatment. All patients had not had any cancer treatment before, neither had any sort of infection, acute diseases or comorbities. A total of 64 BC patients and 52 control group healthy volunteers had their data collected (Age, weight, height...) as well as blood samples after an overnight fasting and these samples had their biomarker levels determined by a variety of clinical tests [6].

The variables itself are composed of 2 anthropometric measurements, Age (years) and BMI (kg/m²), and 7 biomarkers, Glucose(mg/dL), Insulin (µU/mL), Homeostasis Model Assessment (HOMA), Leptin (ng/mL), Adiponectin (µg/mL), Resistin (ng/mL), MCP-1 (pg/dL). HOMA is an indicator of insulin resistance and its value is given by the formula:

$$\text{HOMA} = \frac{log(\text{If} \times \text{Gf})}{22.5} \tag{1}$$

where (If) is the fasting insulin level (µU/mL) and (Gf) is the fasting Glucose level (mmol/L).

B. Univariate and Bivariate analysis

For every variable of the data, it was computed a set of descriptive statistics, such as class-wise mean and standard deviation, skewness and those are listed in Table I.

The table also includes a field which is the ratio between the minimum and the maximum value of the distribution (Max-min Ratio). The skewness is a parameter which indicates how much the values of a distribution are concentrated in one side in relation to the center.

TABLE I

Descriptive statistics of the Dataset

	Mean		Std. Dev.		Skew	Max-Min Ratio
	Healthy	Patient	Healthy	Patient		
Age	58.08	56.67	18.96	13.49	0.02	3.71
BMI	28.32	26.98	5.43	4.62	0.17	2.1
Glucose	88.23	105.56	10.19	26.56	2.59	3.35
Insulin	6.93	12.51	4.86	12.32	2.58	24.04
HOMA	1.55	3.62	1.22	4.59	3.81	53.59
Leptin	26.64	26.6	19.33	19.21	1.31	20.94
Adipon.	10.33	10.06	7.63	6.19	1.82	22.97
Resistin	11.61	17.25	11.45	12.64	2.58	25.58
MCP.1	499.73	563.02	292.24	384.0	1.42	37.05

The skewness of each variable can be clearly seen in the histogram plot in Figure 1. A general rule of thumb that allows us to determine whether a distribution is skewed or not is checking if the Max-min Ratio is greater than 20 [7]. Making use of this rule, it is possible to determine that every biomarker is right-skewed, with the exception of Glucose, which is not above the threshold.

Age and BMI are fairly homogenous in their distribution. The right-skew present in the biomarkers denotes a general trend of having mostly lower values and fewer ones of higher magnitude. Some of the biomarkers present a higher mean and standard deviation for patients, as well as a wider interquartile range. In Figure 2 this pattern is the most evident for Glucose, Insulin, HOMA and Resistin. This hints at a possible class separability between these variables.

The boxplots also point out the existence of potential outliers, but in this case of patients, which have a trend of having higher values, its is unlikely that these are true outliers. The trend is also mostly normal in the case of healthy samples, with the exception of a few couple abnormaly high Resistin and Adiponectin samples in Healthy patients.

C. Data Transformations

While skewness indicates important attributes of the data itself, it has the downside of confering some instability to statistical methods, specially those that assume a normal distribution of the input data. The Box-Cox transformation [8] is an important step in resolving skewness in a distribution. It is composed by a family of transformations made for a predictor variable x controlled by a parameter λ

$$f(x) = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & if \lambda \neq 0 \\ log(x) & if \lambda = 0 \end{cases}$$

The transformation is only valid for positive predictors, which is the case of the BC dataset, and the parameter λ can be estimated through maximum likelihood. The result is an unskewed distribution.

Other relevant transformations are centering, which subtracts the mean of a predictor from every value, resulting in a zero mean distribution, and scaling, which divides each value by the standard deviation, resulting in a distribution with standard deviation equals to one [7]. These transformations lead to loss of interpretability, but are required for further analysis methods, and more specifically, for Principal Component Analysis. Besides that, all of the predictive models used in this study were shown to work better in data that is centered and scaled

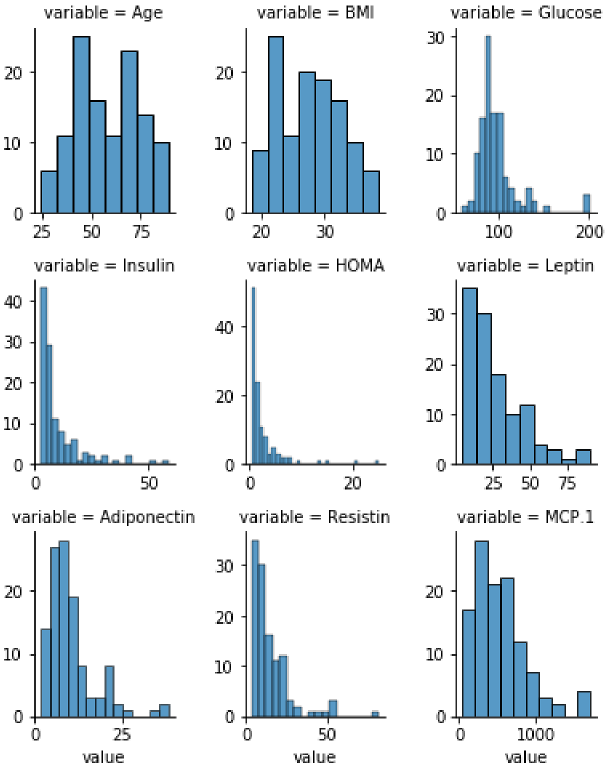

Fig. 1. Histograms of the variables

An evaluation of the relationships between the variables was done by plotting the correlations between the variables in a correlation matrix, indicated in Figure 3. The correlation matrix plots the pair-wise correlation values between the predictors. There is a substantial correlation in the pairs Insulin vs. HOMA and Glucose vs. HOMA, but this is because they hold a relationship given by (1). Other meaningful correlations found are Leptin vs. BMI, Resistin vs. MCP.1 and Leptin vs. HOMA.

D. Classification

In the classification problem we seek to model a qualitative response instead of a quantitative one, as in the case of regression. This usually means that the model must return a value, for instance 0 or 1, corresponding to a specific *label* or *class*, such as ”healthy” or ”ill”. In the context of this work, the objective is to predict whether a subject has breast cancer (Patient) or not (Healthy).

1) Logistic Regression: A possible way of solving the classification problem would be to attempt to model the probability of a sample belonging to a certain class. For a 0 or 1 class problem, we could consider the sample to be class 1 if the output were $y > 0.5$.

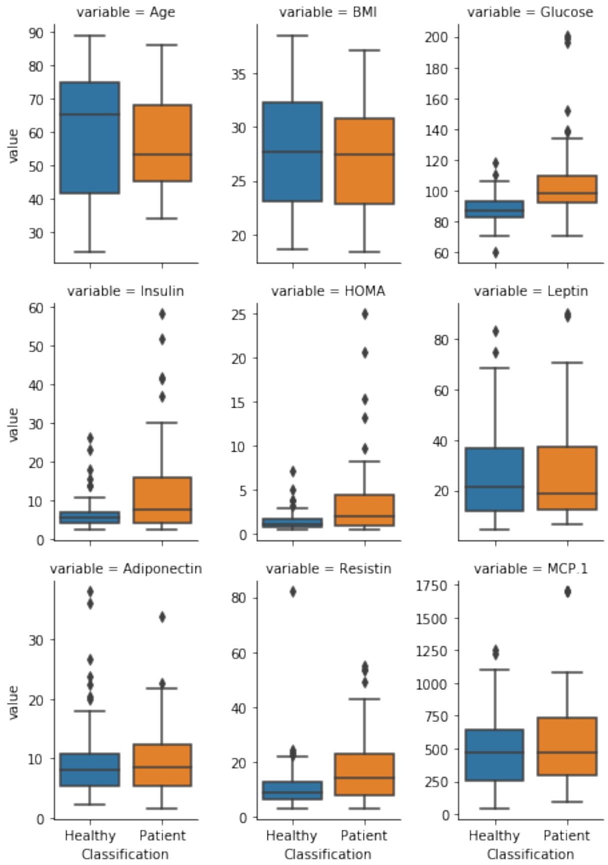

Fig. 2. Class-wise boxplots of the variables

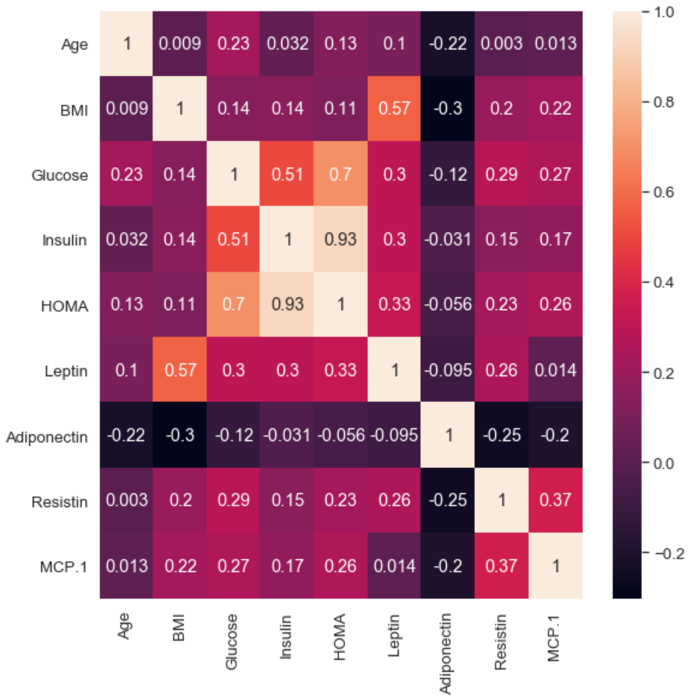

Fig. 3. Correlation matrix of the predictors

Logistic regression seeks to solve this problem by modeling the posterior probability for a given class through a linear function.

$$\mathrm{P}(y=1|X)=p(x)=\frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)} \qquad (2)$$

The function represented by (2) is called the Logistic function and it is characterized by an S shape with values in the [0,1] range. This is an adequate response to model a probability. We can obtain (3) by making a simple algebraic manipulation of (2) which is linear in X. This is called the *logit*. The model is optimized via Maximum Likelihood parameter Estimation.

$$\log\left(\frac{p(x)}{1-p(x)}\right)=\beta_0+\beta_1 X \qquad (3)$$

2) Support Vector Machine (SVM): The SVM classifier consists of a method to find a hyperplane that separates optimally the classes. For a data matrix X with p predictors, i samples, and label $y \in \{-1,1\}$, such a hyperplane could be defined by the function (4).

$$y_i(\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip})>0 \qquad (4)$$

Naturally, there might be an infinite amount of hyperplanes that fit this definition for a given data matrix, so we must present the more useful concept of the maximal margin hyperplane, which is the separating hyperplane that is farthest from the training samples. In other words, this line has the largest margin from the training samples.

It can be shown that the perpendicular distance from a given observation i is in fact given by the left side of (4) [9], so the problem then consists of finding optimal parameters β such that $y_i(\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip}) \geq M$ and M is as large as possible. This is usually called *hard-margin* classifier because it requires that no sample violate the margin.

A *soft-margin* version of this algorithm is obtained by considering a cost constraint. For every observation i consider a slack variable $\{\epsilon_1,\ldots,\epsilon_i\}$ such that $\epsilon_i \geq 0$ subject to $\sum_{i=1}^{n}\epsilon_i \leq C$, where C is a total global optimization parameter. The C corresponds to a total amount of margin violation error tolerated by the classifier. This version of the problem allows some margin or border violations in exchange for finding a good enough fit in datasets where there is no clear separating hyperplane.

The margin classifier can also be further extended to model non-linear decision boundaries through the use of kernel functions. This is usually called the ”kernel trick”, and consists of using non-linear functions such as polynomial, hyperbolic tangent and radial basis functions to obtain a more complex decision boundary [10].

3) Feedforward Neural Networks (FFNN): The basic building block of an Artificial Neural Network (ANN) is the neuron. A neuron is characterized having a set of input signals X and respective weights W that are summed and then passed into an activation function φ into an output signal y, as in (5). Different activation functions may be used, such as threshold, sigmoid or hyperbolic tangent.

$$v = \sum_{j=0}^{m} w_j x_j + b \qquad y = \varphi(v) \tag{5}$$

Neural Networks per se are built by organizing neurons in specific arrangements. The FFNN is a Network composed of layers of Neurons that have as inputs the output signals of the previous layer, and feed their resulting output signals into the next layer, moving the information forwardly to the output layer. The middle layers contain information that is usually not observed and so are called hidden layers or hidden units.

The learning process in FFNNs consists of adjusting the weights and threshold values until a certain performance criterion is satisfied. These conditions are usually defined by a continuous differentiable cost function $E(w)$, where w is a weight vector. The objective is to find and optimal solution w^* such that $E(w) \leq E(w^*)$, or in other terms, $\nabla E(w^*) = 0$, where ∇ is the gradient operator.

Starting with an initial weight vector $w(0)$, the optimal weights can be found via the steepest descent method. The weights are successively adjusted in the direction opposite of the gradient vector $\nabla E(w)$, regulated by a positive constant η called the learning rate. The learning rate regulates the steps taken by the weights at each iteration.

$$w(n+1) = w(n) - \eta \nabla E(w)(n) \tag{6}$$

A fundamental part of the learning process is the Back-propagation. In general, a FFNN has an input signal that moves forward through the network to the output, and an error signal $E(w)$ computed with the output and the desired response, that propagates backward through the network. The error signal is used to compute the gradients with respect with every hidden unit parameter of the network with the steepest descent method [11].

III. Results

The models used in this study were fit after the variables were centered and scaled and had the skewness adjusted via Box-Cox method. All hyperparameters were validated with 10-fold cross validation. The kernel for the SVM model was chosen according to the best performance and a Radial Basis Function, along with a cost $C = 1.9$ were found to be the best fit. The FFNN architecture that had the best fit consisted of 2 hidden layers, respectively with 50 and 30 hidden units, with an activation Rectified Linear Unit (ReLU) function.

TABLE II
10-fold CV Accuracy on different sets of features

	Full	PCA	Reduced
Log. Regression	0.5341	0.525	0.578
SVM (RBF)	0.481	0.489	0.741
FFNN	0.568	0.595	0.750

Three different sets of pre-processing steps were validated via the cross validation process as well, as seen in Table II. Fitting a model on all of the columns has yielded poor results in terms of classification accuracy. A PCA was conducted on the dataset and 5 features were retained, which resulted in slightly better results for the non-linear models. In his study, [4] took an approach of modeling a reduced set of features; Age, BMI, Glucose and Resistin; with good results. This approach was attempted in this study and has resulted in much higher accuracy in non-linear models.

The results seen in table II may be explained by the existance of collinearity in the dataset. Glucose and HOMA are directly correlated. Another possible factor is a low number of samples, which is known to affect more severely non-linear models.

TABLE III
Performance results of the models

	Accuracy	Sensitivity	Specificity
Log. Regression	0.83	0.86	0.80
SVM (RBF)	0.86	1.00	0.73
FFNN	0.90	1.00	0.80

The reduced feature dataset was then divided in 75% for the train set, and 25% for the test set. The performance results for each model on the test set can be seen in Table III. It is clear that the FFNN has a much better accuracy performance, with up to 0.9 accuracy rate. The non-linear models also presented a much higher true positive rate (Sensitivity) of 1.0, meaning that every person with breast cancer was correctly predicted, while maintaining a worse or equal true negative rate (Specificity).

The ROC curves for every model were computed and can be seen in figure 4. The area under curve (AUC) for Logistic regression obtained was 0.943, and for the non-linear models were 0.948 and 0.905 for SVM and FFNN respectively. The non-linear models are capable of having a very high sensitivity. The Logistic regression model has an interesting characteristic, it is capable of having a reasonably good true positive rate with a zero false positive rate on this test set.

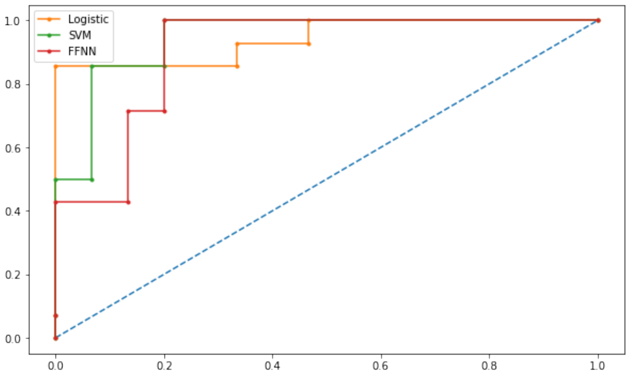

Fig. 4. ROC curve of the models

The Neural Network model presented a lower AUC. For a given threshold it can have a perfect sensitivity at the cost of having some false positives. This might be an indication of overfitting, or in other terms, a possible excess in variance in the bias variance balancing of the problem.

IV. Conclusion

This study allows us to affirm that Age, BMI, Glucose and Resistin had a strong predictive power for Breast Cancer diagnose. The models trained have had a very good accuracy and we have shown that while SVM attains a very good performance, even better results may be obtained with a neural network approach. While a low number of Principal Components may hold most of the variance of the dataset, they have not provided considerably better predictive results.

References

[1] C. J. Murray and A. D. Lopez, "Mortality by cause for eight regions of the world: Global burden of disease study," *The lancet*, vol. 349, no. 9061, pp. 1269–1276, 1997.

[2] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 6, pp. 438–451, 2019.

[3] V. G. Vogel, "Epidemiology of breast cancer," in *The breast*. Elsevier, 2018, pp. 207–218.

[4] M. Dalamaga, "Resistin as a biomarker linking obesity and inflammation to cancer: potential clinical perspectives," *Biomarkers in medicine*, vol. 8, no. 1, pp. 107–118, 2014.

[5] J. Crisostomo, P. Matafome, D. Santos-Silva, A. L. Gomes, M. Gomes, M. Patrício, L. Letra, A. B. Sarmento-Ribeiro, L. Santos, and R. Seiça, "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer," *Endocrine*, vol. 53, no. 2, pp. 433–442, 2016.

[6] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, "Using resistin, glucose, age and bmi to predict the presence of breast cancer," *BMC cancer*, vol. 18, no. 1, p. 29, 2018.

[7] M. Mulas, "Data analysis and pre-processing," November 2020.

[8] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[9] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[11] M. Mulas, "Neural networks," November 2020.