Biomarkers in Breast Cancer Diagnosis: an Exploratory Data Analysis

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 Principal Components of several 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 biomarkers present an abnormal profile for breast cancer patients and the use of these indicators can be coupled with other external information for diagnostics and monitoring.**

***Index Terms*—Breast cancer, PCA, biomarkers, Principal Components**

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 glucoes, 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 and the possibilities of dimensionality reduction through the use of Principal Component Analysis.

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 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. For a predictor variable x with mean $\bar{x}$ and n number of values, the skewness γ_x is defined as [7]

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

$$\gamma_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)v^{3/2}} \text{ where } v = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

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, which is important to maintain a control on these variables, as the purpose of the data is most focused on biomarkers. The right-skew present in the biomarkers denotes a general trend of having mostly lower values and fewer ones of higher magnitude.

A class-wise analysis shows a possible explanation on the nature of these higher values. 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

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

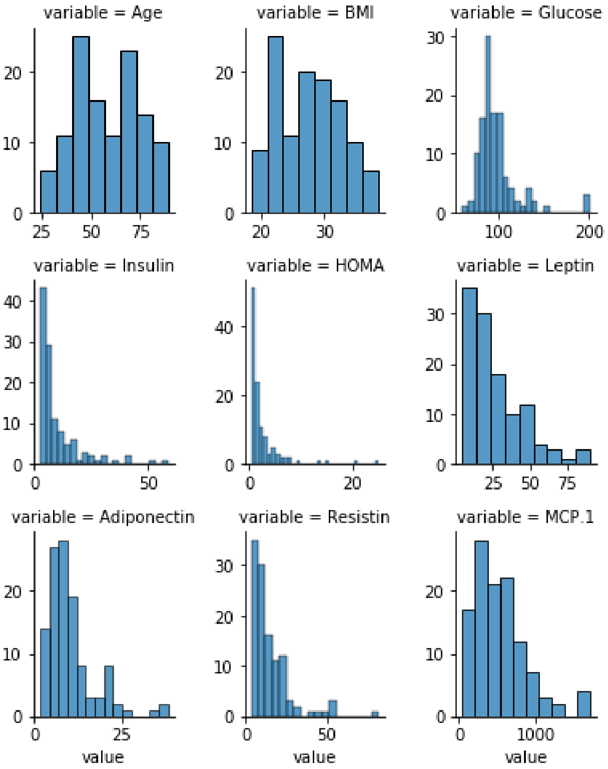

Fig. 1. Histograms of the variables

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.

D. Bi-variate Analysis

An evaluation of the relationships between the variables was done through the use of pair-wise scatter plots and by plotting the correlations between the variables in a correlation matrix, indicated in Figures 4 and 3. Most of the scatter plots denote a scenario of prevalent overlap between both classes, but with a subset of Patients off the general trend, usually with higher values.

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.

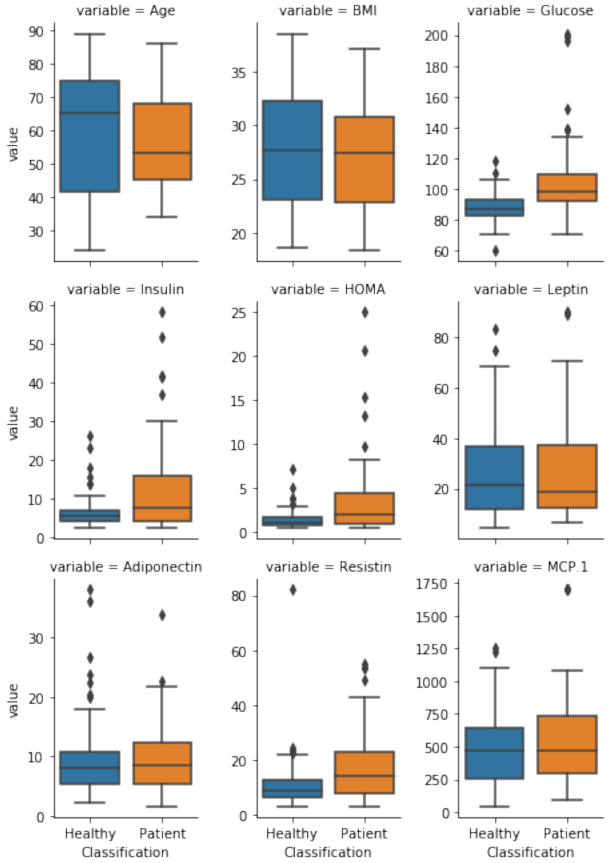

Fig. 2. Class-wise boxplots of the variables

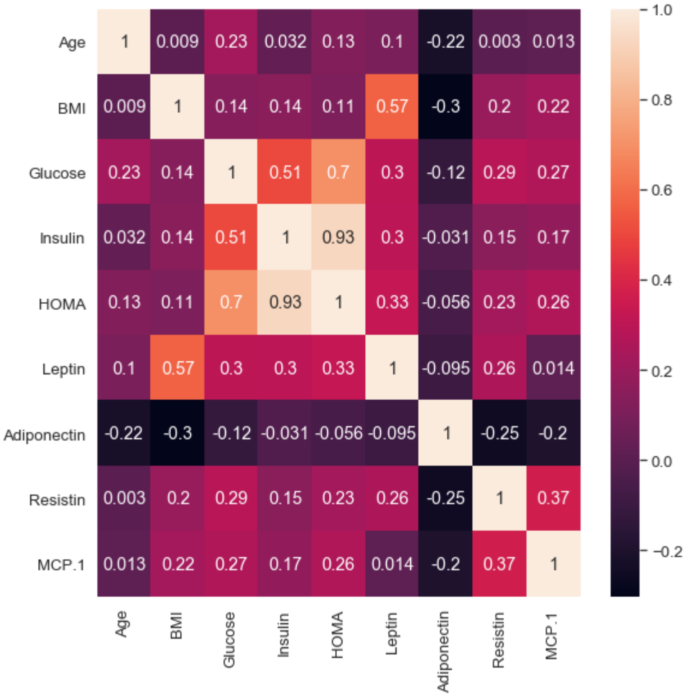

Fig. 3. Correlation matrix of the predictors

E. Principal Component Analysis

The Principal Components (PC) of a set of data can be defined as a set of orthogonal components computed from linear combinations of the original variables of the dataset such that they maximize variance [9]. For a matrix $n \times m$, where n is the number of samples (rows) and m is the number of predictors (columns), there are m PCs, as stated in (2).

$$PC_i = \phi_{i,1}x_1 + \phi_{i,2}x_2 + ... + \phi_{i,m}x_m \tag{2}$$

Each new computed PC captures a portion of the overral variance of the data in decreasing magnitude ($\sigma^2_{PC_1} > \sigma^2_{PC_2} > ... > \sigma^2_{PC_m}$) and its orthogonal to the previous PC ($PC_1 \perp PC_2 \perp ... \perp PC_m$). By selecting a small subset of $k < m$ PCs we can form an orthonormal basis through which a change of basis may be performed on the data. This results in a transformed matrix with k columns, and its basis are now along the directions with the most variance in the space.

The analysis of the PCs may be able to bring out the main dimensions through which the data points vary, or help to filter out noise in systems and provide insights on the nature of the data [10]. It is also a possibly valuable pre-processing step for statistical methods, as its dimensionality reduction may allow faster computations on costly functions. Two important algebraic solutions for calculating the PC can be pointed out [10].

1) Eigendecomposition: The covariance matrix allows us to view the covariance between the predictors of the dataset. As its possible to conclude from its formula (3), $Cov(\mathbf{X}, \mathbf{X})$ is a square matrix, and along is diagonal there are the variance values of the columns, and the off-diagonal values represent the covariance of the columns matrix X

$$Cov(\mathbf{X}, \mathbf{X}) = \mathbf{C_X} = \frac{\mathbf{X}\mathbf{X}^T}{n} \tag{3}$$

For a column-wise normalized (centered and scaled) dataset $\mathbf{X}$ with m predictor rows by n sample columns assume there is a matrix of orthonormal basis vectors $\mathbf{P}$ in $\mathbf{Y} = \mathbf{PX}$ such that $\mathbf{Y}$ is a diagonal matrix, which means off-diagonal values are zero (components are orthogonal). This means that the covariance matrix $Cov(\mathbf{Y}, \mathbf{Y})$ is diagonal. The algebraic manipulation (4) describes $Cov(\mathbf{Y}, \mathbf{Y})$ in terms of $Cov(\mathbf{X}, \mathbf{X})$ [10].

$$\begin{aligned} Cov(\mathbf{Y}, \mathbf{Y}) &= \frac{\mathbf{Y}\mathbf{Y}^T}{n} \\ &= \frac{(\mathbf{PX})(\mathbf{PX})^T}{n} \\ &= \frac{\mathbf{P}\mathbf{X}\mathbf{X}^T\mathbf{P}^T}{n} \\ &= \mathbf{P}\left(\frac{\mathbf{X}\mathbf{X}^T}{n}\right)\mathbf{P}^T \\ Cov(\mathbf{Y}, \mathbf{Y}) &= \mathbf{P}\mathbf{C_X}\mathbf{P}^T \end{aligned} \tag{4}$$

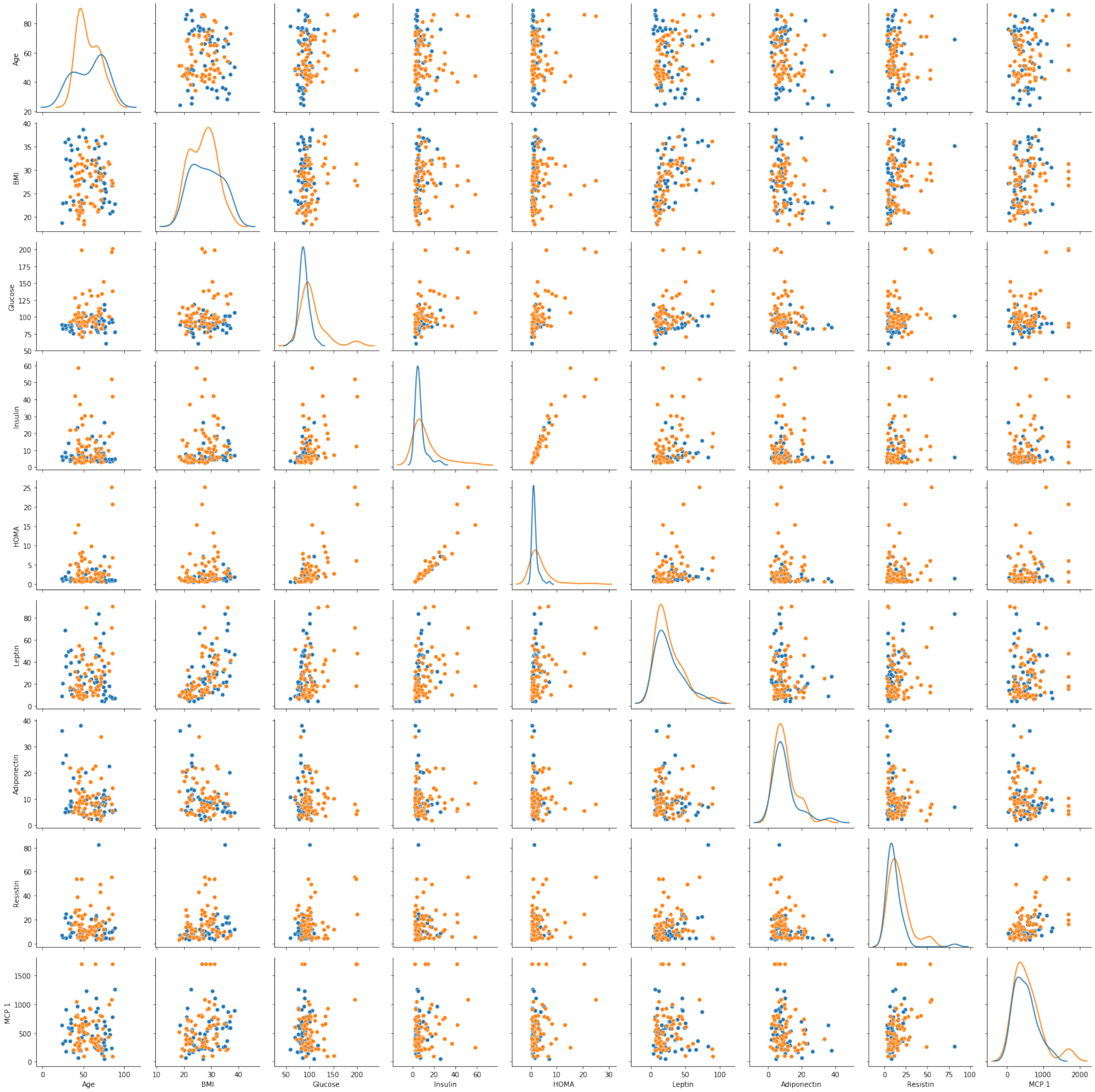

Fig. 4. Class-wise scatter plots

Taking the eigendecomposition $\mathbf{C_X} = \mathbf{Q}\Lambda\mathbf{Q}^T$ and defining $\mathbf{P} = \mathbf{Q}^T$

$$\begin{aligned} Cov(\mathbf{Y},\mathbf{Y}) &= \mathbf{P}(\mathbf{Q}\Lambda\mathbf{Q}^T)\mathbf{P}^T \\ &= \mathbf{P}(\mathbf{P}^T\Lambda\mathbf{P})\mathbf{P}^T \\ &= (\mathbf{P}\mathbf{P}^{-1})\Lambda(\mathbf{P}\mathbf{P}^{-1}) \\ Cov(\mathbf{Y},\mathbf{Y}) &= \Lambda \end{aligned} \tag{5}$$

The solution (5) demonstrates that the eigenvectors of $\mathbf{C_X}$ do provide a basis that make the covariance matrix of $\mathbf{Y}$ diagonal, and the eigenvectors of $\mathbf{C_X}$ can be said to be the Principal Components of $\mathbf{X}$ [10].

2) Singular Value Decomposition (SVD): For a column-wise normalized (centered and scaled) dataset $\mathbf{X}$ with m predictor rows by n sample columns and a $\mathbf{Y}$ matrix with $n \times m$ dimensions defined as 6, where each column has zero mean, the singular value decomposition $\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^T$ yields a $\mathbf{V}$ matrix which contains the eigenvectors of $\mathbf{Y}^T\mathbf{Y}$.

$$\mathbf{Y} = \frac{\mathbf{X}^T}{\sqrt{n}} \tag{6}$$

It is evident that $\mathbf{Y}^T\mathbf{Y} = Cov(\mathbf{X},\mathbf{X})$, and $\mathbf{V}$ contains the

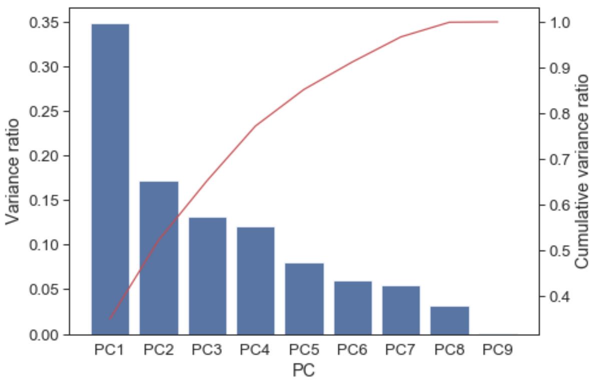

Fig. 5. Variance ratio and cumulative variance ratio of the Principal Components

eigenvectors of $\mathbf{C_X}$, which are the principal components of $\mathbf{X}$. SVD guarantees us that $\mathbf{V}$ spans the row space of $\mathbf{Y}$ and therefore, $\mathbf{V}$ spans the column space of $\frac{\mathbf{X}}{\sqrt{n}}$. So $\mathbf{V}$ is an orthonormal basis that spans the column space of $\mathbf{X}$ [10].

The resulting PC variances can be summed and have a variance ratio calculated. This ratio indicates the amount of variance explained by each PC and it can be used to decide on how many PCs will be used for dimensionality reduction.

III. Results

The univariate analysis conducted on the BC dataset has shown that most of the predictors have a right-skewed distribution, meaning that a minority of samples have a high magnitude. Such behavior indicates the presence of an abnormal subset of samples which might belong to an specific class. This point is further driven by the class-wise plots. The Figure 2 denotes the recurrence of these high magnitude samples in the Patient class, and Figure 3 shows a number of Patient class individuals having abnormal values in pairwise variable comparisons. This shows a picture of an anomalous biomarker profile for BC patients.

The Figure 4 matrix shows a fair correlation between the variables, which indicates the necessity of evaluating the potential of dimensionality reduction. As it is a required preprocessing step, the skewed variables from the BC dataset (all except Age, BMI and Glucose) had the Box-Cox transformation applied and were centered and scaled prior to Principal Component Analysis.

The PCs were calculated on the pre-processed data and the results can be seen in Figure 5. The first two PCs amount to an explained variance of approximately 0.521, where PC_1 by itself has a variance ratio of approximately 0.35. PC_8 has a ratio of less that 0.04 and PC_9 is less that 0.001, and in scenario of dimensionality reduction, these could be discarded with little loss.

This analysis shows that there is no single or few number of components responding to all of the difference between the classes. Figure 6 shows that when projected on the 2 main PCs, there are a few Patient samples out of the general trend in the center, but the classes are not easily separable, having many samples from both classes habitating very closely in the middle.

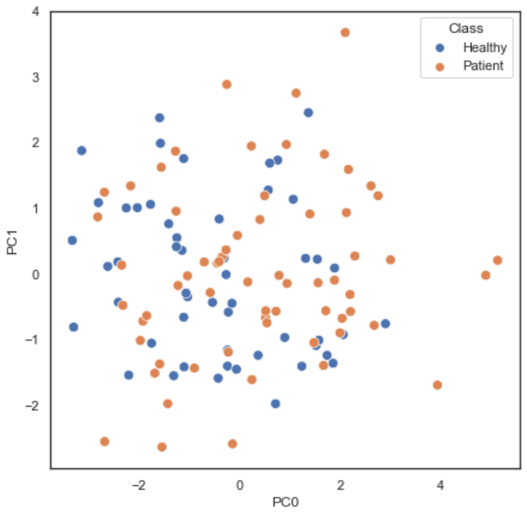

Fig. 6. BC dataset projected into PC_1 and PC_2

IV. Conclusion

It is visible that there is a profile of abnormal indicators in patients of breast cancer and the monitoring of these biomarkers can certainly provide very reasonable evidences in the scenario of evaluating a diagnose. But while a myriad of predictor attributes can amount to a general higher probability, no single one provides a definitive conclusion. Certainly the resarch and investigation of new indicative clinical predictors should provide an even more clear picture, and the inclusion of other outside informations will surely make the prediction more realiable.

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. Mulas, "Data pre-processing," December 2020.

[10] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.