Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models

Naveen
School of VLSI and Embedded Systems Design, NIT Kurukshetra
naveen151094@gmail.com

Dr. R. K. Sharma
Department of Electronics and Communication, NIT Kurukshetra
rksharma@nitkkr.ac.in

Dr. Anil Ramachandran Nair
R&D Divisions, Toshiba Software (India) Pvt. Ltd., Bangalore
anil.nair@toshiba-tsip.com

***Abstract*— Breast cancer is the second most exposed cancer in the world. When the growth of breast tissues are out of control is called breast cancer. Breast cancer prediction and prognosis are major challenge to medical community. Breast cancer are prominent cause of death for women. Recurrence of cancer is the biggest fears for cancer patient and this can affect their quality of life. The aim of this research is to predict breast cancer from cancer features with high accuracy. The breast cancer Coimbra dataset taken from UCI (University of California Irvine) [1, 5] to build a most efficient ensemble machine learning models. The major steps we follow, here are feature scaling, cross validation and various ensemble machine learning models with bagging technique. Decision tree and KNN gives highest 100% accuracy. Decision tree model gives 100% accuracy if we split train-test dataset in ratio of 90:10 and also used 300 bags of trees. KNN gives maximum accuracy 100%, for k= 1 to 7 in seven loops with 90% is train data and 10% is test data. Here k is the nearest neighbors. And we also evaluate its prediction by accuracy, confusion matrix and classification report. Our aim is to build a most accurate and efficient machine learning model. So as prediction result, patient can take treatment on the early stage.**

***Keywords*— Breast Cancer, Significant Feature, Ensemble Machine Learning Models, Accuracy, Confusion Matrix, Classification Report.**

I. INTRODUCTION

Breast cancer is the most frequent types of skin cancer disease in women around the world and this cancer may occur in men also. Breast cancer produces from breast tissues. In United States, one out of eight women have breast cancer in their lifetime. The study shows that average ages of women in between 45 to 59 has more chance to cancer. Breast cancer is the second most commonly diagnosed cancer. In US, every year diagnoses cases of breast cancer is more than 266,000. According to Global Health Estimates, WHO 2013 report that more than 508,000 women died in 2011 due to breast cancer. Approximately 50% of breast cancer cases and 58% of death occurs in developing nations (GLOBOCAN 2008 Report) [9]. Breast cancer survival rates differ in worldwide, most survival rates reaching 80% or over in North America, Sweden and Japan and to around 60% in developing countries and less than 40% in undeveloped countries due to lack of early prediction [4,6].

As previous works shows that there is a various research gone on breast cancer prediction with different machine learning approach. However, for Coimbra breast cancer dataset we could found only few studies. Coimbra breast cancer dataset gives nearly 65-85% accuracy using different models of machine learning. Hare we use ensemble bagging technique in machine learning models. So it give maximum accuracy 100% by decision tree and KNN when data splitting ratios is 90:10.

Breast cancer can be cured with current medical treatments and new innovative techniques [1, 16]. It is very useful to early predict breast cancer with high accuracy since patients can get treatment on time and can survive. In this approach, we first perform feature importance techniques on breast cancer dataset to select the most significance features. Then we perform standard scaling with zero mean and standard deviation one to scale all features in a range. After that we send this scaled features to various ensemble machine learning models. Here, we use 10 folds cross validation to train the ML models with each folds, it reduce over fitting of model, as model become more efficient. Here we use bagging techniques in ML models. Bagging technique makes the bags of samples, each bags may contain identical samples also. Each bag is test by different trained ML model, we performs voting in each bag prediction result, highest voting of class is our final prediction results, and we conclude prediction results of ML model. We also evaluate its prediction results with confusion matrix, classification report and accuracies. Models developed through these techniques are very helpful for medical community to take right decisions.

II. DATASET

The breast cancer Coimbra dataset taken from UCI (University of California Irvine) [1, 17] to developed a most efficient ensemble machine learning models. These dataset contains 116 samples. There are two set of samples (52 samples: healthy peoples, 64 samples: breast cancer patients).It consist of nine features and two classes as healthy and patients. These nine features are: Age (years), BMI (kg/m2), Glucose (mg/dL), Leptin (ng/mL), Adiponectin (ug/mL), Resistin (ng/mL), Insulin (Uu/mL), HOMA and MCP-1 (pg/dL). In this approach, we performed feature importance on breast cancer datasets for selection the significant features from dataset.

As fig.1, line plot show that the behavior of Coimbra breast cancer dataset. Here we see that MCP-1 attribute is highly varying. Glucose is less varying feature. Except this all are linear data. To reduce this variance of feature, we performs feature scaling with standard scale.

978-1-7281-0630-4/19/$31.00 ©2019 IEEE

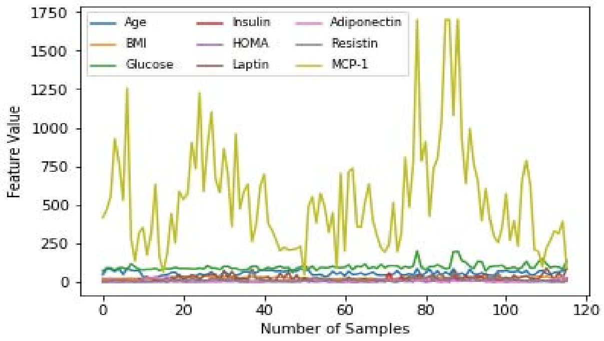

Fig.1: Line Plot of Breast Cancer Dataset.

III. Procedure and Methods

To classify the breast cancer dataset, a hybrid technique with three steps, has been proposed. As in initial step we normalized features with standard scale to keep in a range, in second step we build the different ensemble machine learning models with crossover validation and bagging technique and in third step we evaluate its prediction results with accuracy, confusion matrix and classification reports and thus concluded that which model is the most efficient for breast cancer prediction. Classification report values are in range of (0.0, 1.0). Confusion matrix is the heat maps of the actual class and predicted class.

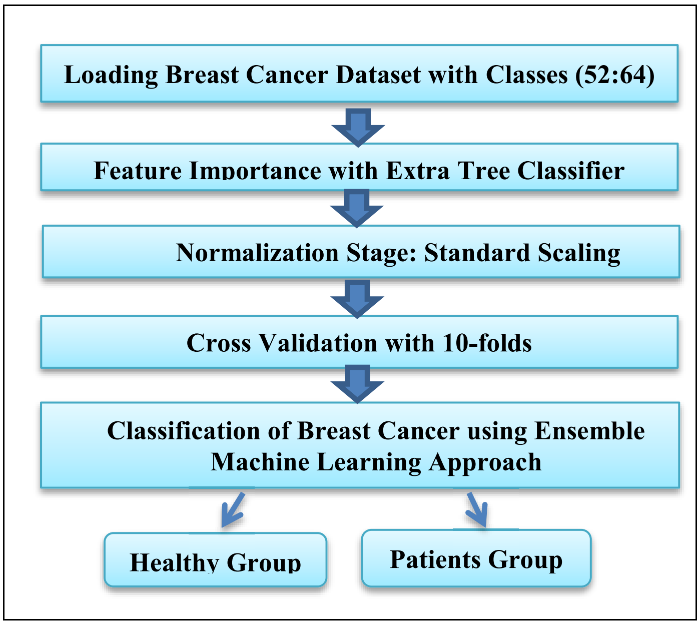
Loading Breast Cancer Dataset with Classes (52:64)
Feature Importance with Extra Tree Classifier
Normalization Stage: Standard Scaling
Cross Validation with 10-folds
Classification of Breast Cancer using Ensemble Machine Learning Approach
Healthy Group
Patients Group

Fig.2: Procedure of Ensemble ML Model Approach.

A. *Significant Features*

Feature importance improve the model's performance. It is quiet important to know the effect of a certain feature on the model's performance. Some are very less important that can reduce model performance. Significant features are evaluate by extra tree classifier algorithm with 250 estimation trees of forest. This extra tree classifier worked based on standard deviation of data. And indices of this algorithm sort features in decreasing order ranking.

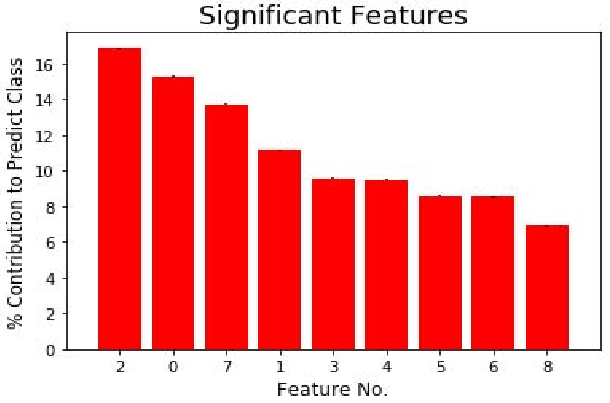

Fig.3: Significant Features in Ranking Wise

With significant features plot, we see glucose is the most significant feature for prediction of bread cancer and MCP-1 is the least significance feature. Below table shows importance of features in the decreasing order ranking wise.

Feature No.	Feature's Name	Contribution to Predict Class (%)
2	Glucose	16.85%
0	Age	15.23%
7	Resistin	13.74%
1	BMI	11.14%
3	Insulin	09.55%
4	HOMA	09.47%
5	Laptin	08.56%
6	Adiponectin	08.55%
8	MCP-1.0	06.91%

Table.1: Feature Significance ranking in decreasing order.

B. *Standard Scale or Z- Normalization*

Features may different in scale or units, it is difficult for classifier or regression to give optimal results. The way to overcome this difficultly we need to scale it, in one specific range. Here, we see MCP-1 is highly varying features, since we normalize it with standard scaling or z-score normalized. For standard scaling, first subtract its value with its mean, to bring its mean around zero then divided by its standard deviation value, to bring its standard deviation near to one.

$$Z - Normalized = \frac{\mathrm{Xi} - \mathrm{mean(X)}}{\mathrm{Stdev(X)}} \ldots\ldots\ldots\ldots (1)$$

C. *Cross Validation (CV)*

Cross validation is a re-sampling method used to evaluate machine learning models on a constrained data samples. Generally k is chosen as 10-fold. The general strategies follow in crossover validation is:

- Mix the dataset arbitrarily.

- Split the dataset into k groups (k =10)
- For every group have different kinds of samples.
- Summarize results of each group to utilizing the model accuracy.

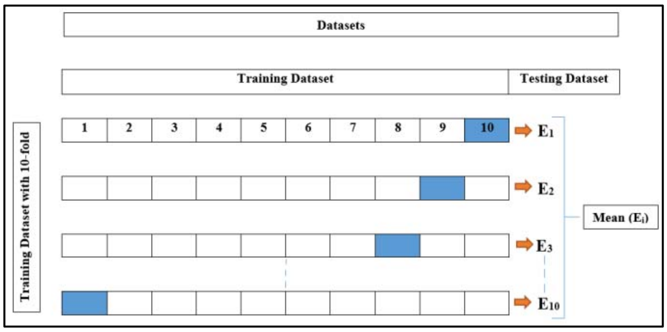

Fig.4: Working of the 10-fold Cross Validation

D. *Machine Learning (ML) Approach*

Machine learning (ML) techniques utilize statistics, probabilities, Boolean logics, and unconventional optimization techniques to build an ensemble machine learning classifier and predict the class based on the highest voting.

1) *Decision Tree:*

Decision tree algorithms uses the hierarchical tree approach, there every node represents a feature, and branch represents a decision and leaves represents an outcome (class). In this approach we predict the breast cancer with 300 trees and dataset split in to 90:10 as train and test dataset. Here this model gives 100% accuracy with 300 trees and dataset are split in to 90:10 as train and test.

2) *Support Vector Machine (SVM):*

Support vector machine (SVM) is a discriminative classifier characterized by an isolating hyper plane. In two dimensional space this hyper plane is a line separating a plane in two sections where in each class lay in either side. SVM gives 83.33% accuracy with linear kernel.

3) *Multilayer Perceptron:*

Multilayer perceptron (MLP) has a numbers of layers, two terminating layers are called input-output layer and intermediate layers are called hidden layers. It uses a stimulation function in all neurons. Here we use 60 hidden layers and each layer has 10 neurons with different weights. MLP gives 83.33% accuracy.

$$f(x_i) = b + \sum_{i=1}^{n} x_i w_i \qquad \dots\dots\dots..\dots\dots \qquad (2)$$

where x_i = inputs of incoming layers.
w_i = weights of hidden layers neurons.
b = initial weight.

4) *K-Nearest Neighbor (KNN):*

KNN is very useful for a large dataset do not use mathematical analysis. In the worst scenario, KNN needs more memory to check all data sets. Here we used k=1 to 7 neighbor in a 7 random loop. It is total 7x7=49 evaluation. Its maximum accuracy is 100% at k=5 in loop =3. And its average accuracy in all loop is 89.29%.

$$D(a.b) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2} \qquad \dots\dots\dots\dots\dots\dots \qquad (3)$$

- Where a, b are coordinates of two point.

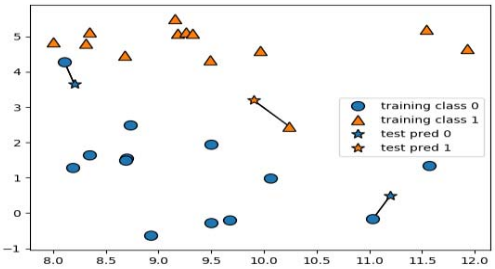

Fig.5: K-Nearest Neighbor (KNN) model

5) *Logistics Regression:*

Logistic regression is a statistical machine learning model that classifies the data by considering outcome variables on extreme end class and tries to makes a logarithmic line that separates between them. Logistic regression gives 91.67 % accuracy in our approach.

$$P(y|x) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \qquad \dots\dots\dots\dots\dots\dots\dots \qquad (4)$$

- Where α and β are the model parameters

6) *Random Forest:*

Random forests is an ensemble machine learning algorithm of decision tree. Random forest build a set of decision trees based on random chosen samples gets expectation from each tree and choose the best prediction results from the voting of each tree. In our aproach, it gives 90.91% accuracy of predictions.

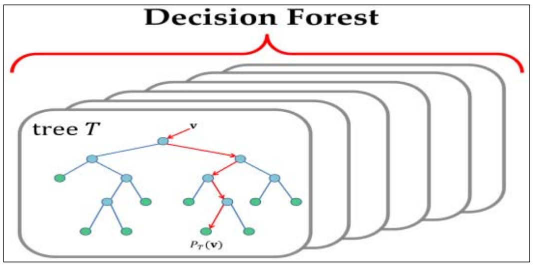

Fig.6: Random Forest model

IV. Algorithm

For this approach we use anaconda navigator 3 (64-bit) for python programming. Following algorithmic steps, we follow:

1) Import all the modules for feature selection, normalization, data splitting, ML models, accuracy score, confusion matrix, classification report and some other required modules.
2) Load the breast cancer dataset.
3) Divide the datasets as feature and class.
4) Check the significant features for prediction of class.
5) Normalize the features to scale in one range with Standard scaling.
6) Split the dataset in two training and testing set in 90:10 respectively.
7) Build a various machine learning models using bagging techniques with 10-fold cross-validation with different estimation trees.
8) Print accuracy and classification reports of different models as comparing the true and predicted class
9) Plot confusion matrices of different models comparing the true and predicted class.

V. Experimental Results

A. Accuracy

Accuracy is inversely proportional to the difference between true class and predicted class. If any model has same class predict as true class, since the model has highly accurate.

There are accuracies of various machine learning models.

Machine Learning Models	Accuracy (without ensemble)	Accuracy (with ensemble)
Decision Tree with 300 Trees	66.67 %	100.0 %
Support Vector Machine (SVM)	83.33 %	83.33 %
Multilayer Perceptron (MLP)	66.67 %	83.33 %
K-Nearest Neighbor-Average	70.58 %	83.33%
K-Nearest Neighbor-Maximum	89.90 %	100.0 %
Logistics Regression	75.00 %	91.67 %
Random Forest	75.00 %	90.91 %

Table.2: Accuracy of ML models

In this approach decision tree and KNN gives highest accuracy. KNN gives the maximum accuracy 100% when we evaluate for first seven (k= 1 to 7) neighbors in seven loops, it is total 7 X7=49 loops, in k=5 and loop=3 its gives 100% accuracy and average accuracy is 83.33% with 90:10 ratios of train and test split dataset respectively. It is 89.17% accuracy in k=10. If training and test dataset are 80:20 then maximum accuracy of KNN models is 87.5%.

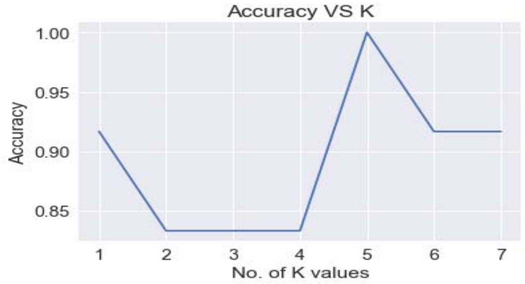

Fig.7: Accuracy plot of KNN with k = 1 to 7

B. Confusion Matrix

A confusion matrix is an outline of prediction. The number of accurate and inaccurate predictions are précised with count values and broken down by each class. The confusion matrix is the methods in which ML model is confused when it is predicted. It gives us intuition not only about the inaccuracies of classifier but also tells about the types of errors in which class.

Confusion Matrix of different ML model as below:

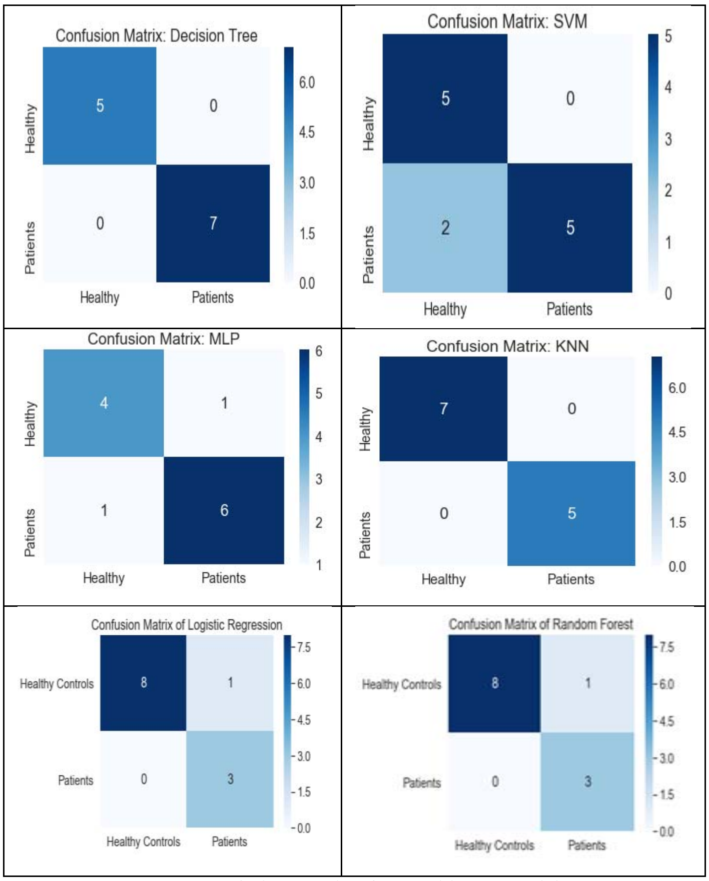

Fig.8: Confusion Matrix of Ensemble ML models

C. Classification Report

The classification report gives the information about precision, recall, F1-score and sample support of models. So it helps in problems identification. All heat maps are in the range (0.0, 1.0). Classification Report describe about the models behaviour and its reports. Among all models, KNN and decision tree algorithm gives the high precision, recall, f1-score and support samples as shown below table.

ML Models		**Classification Reports**			
		Precision	Recall	F1-Score	Support Sample
Decision Tree	C-1	1.00	1.00	1.00	7
	C-2	1.00	1.00	1.00	5
	avg/total	1.00	1.00	1.00	12
SVM	Healthy	0.71	1.00	0.83	5
	Patients	1.00	0.71	0.83	7
	avg/total	0.88	0.83	0.83	12
MLP	Healthy	0.80	0.80	0.80	7
	Patients	0.86	0.86	0.86	5
	avg/total	0.83	0.83	0.83	12
KNN	Healthy	1.00	1.00	1.00	7
	Patients	1.00	1.00	1.00	5
	avg/total	1.00	1.00	1.00	12
Logistics Regression	Healthy	1.00	0.89	0.94	9
	Patients	0.75	1.00	0.86	3
	avg/total	0.88	0.94	0.90	12
Random Forest	Healthy	1.00	0.89	0.94	9
	Patients	0.75	1.00	0.86	3
	avg/total	0.88	0.94	0.90	12

Table.3: Classification Reports of Ensemble ML Models

VI. CONCLUSIONS AND FUTURE WORKS

Ensemble ML models predict the breast cancer with high accuracy compare with without ensemble models. Ensemble model improves the system performance with un-biasing. Here we used different six machine learning algorithms such as decision tree, support vector machine, multilayer perceptron, K- nearest neighbors, logistics regression and random forest and compare its prediction evaluation with ensemble and without ensemble techniques. Decision tree and KNN gives 100% accuracy with ensemble technique. K-nearest neighbor gives maximum accuracy as 100% when we evaluate it for first seven (k= 1 to 7) neighbors in seven random loops, It is total 7X7=49 loops, in k=5 and loop=3 its gives 100% accuracy with 90:10 of train and test split dataset. It accuracy is 89.17% when we k=7 neighbors. If training and test dataset are 80:20 then accuracy of KNN models is 87.5%.

As extension of this work, we can provide this model to medical community, there doctor or diagnosis people put patients feature diagnosis report and predict breast cancer of the patients. This model give highly accurate results, since they can take right decision.

REFERENCES

[1] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seica, R., & Caramelo, F., "Using Resistin, glucose, age and BMI to predict the presence of breast cancer, BMC Cancer", 18(1), 2018.

[2] Mohamed NEMISSI, Halima SALAH, Hamid SERIDI, "Breast cancer diagnosis using an enhanced Extreme Learning Machine based-Neural Network ", 2018 International Conference on Signal, Image, Vision and their Applications (SIVA).

[3] Predict the presence of breast cancer", research article, Patricio et al. BMC Cancer (2018).

[4] Kemal Polat, Ümit Şentürk, "A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoost Classifier", IEEE 2018, ISMSIT Turkey.

[5] Polat, K., "Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets, Neural Computing and Applications", 30 (3), 987–1013, 2018.

[6] Jaber Alwidian, Bassam H. Hammo, Nadim Obeid, "WCBA: Weighted classification based on association rules algorithm for breast cancer disease", Applied Soft Computing, Volume 62, 536-549, 2018.

[7] "Breast Cancer Coimbra dataset from UCI", (https://archive.ics.uci.edu/ml/datasets /Breast+Cancer+Coimbra).

[8] Naresh Khuriwal, Nidhi Mishra, "Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm", IEEE 2018, IEEMA Engineer Infinite Conference (eTechNxT).

[9] "WHO data about breast cance", (https://www.who.int/cancer/detection/ breastcancer /en/index1.html).

[10] Noushin Jafarpisheh, Nahid Nafisi, Mohammad Teshnehlab, "Breast Cancer Relapse Prognosis by Classic and Modern Structures of Machine Learning Algorithms", IEEE 2018, 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS).

[11] Meriem Amrane, Saliha oukid, Ikram Gagaoua, Tolga ENSAR, "Breast Cancer Classification Using Machine Learning ", IEEE 2018, 2018 Electric Electronics, Computer Science, Biomedical Engineering's' Meeting (EBBT) Turkey.

[12] Muhammad Imran Faisal, Saba Bashir, Zain Sikandar Khan, Farhan Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer", IEEE 2018, 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST).

[13] Y.K. Anupama, S. Amutha, and D. R. Ramesh Babu, "Survey on data mining techniques for diagnosis and prognosis of breast cancer" Int. J. Recent and Innovation Trends in Computing and Communication, vol.5, pp. 33–37, 2017.

[14] C. E. DeSantis, J.Ma, A. Sauer Goding, L. A. Newman and A. Jemal, "Breast cancer statistics, 2017, racial disparity in mortality by state", CA: a cancer journal for clinicians, vol. 67, no 6, pp. 439-448. 2017.

[15] Joana Crisostomo, Paulo Matafome, Daniela Santos-Silva, Ana L. Gomes, Manuel Gomes, Miguel Patrı´cio, Liliana Letra, Ana B. Sarmento Ribeiro, Lelita Santos, Raquel Seica, "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer ", Springer Science Business Media New York 2016.

[16] Abreu, Pedro Henriques and Santos, Miriam Seoane and Abreu, Miguel Henriques and Andrade, Bruno and Silva, Daniel Castro, "Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review, ACM Comput. Survey", Volume 49 Issue 3, December 2016.

[17] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis", Cancer Informatics, vol. 2, pp. 59–77, 2006.