




Breast Cancer Detection Algorithm Using Ensemble Learning

Sophia Sandoval Torres¹, Ana Paola Romero Espinoza¹,
Grisel Jhovana Castro Valles¹, and Carlos Eduardo Cañedo Figueroa² (✉) 

¹ Coordinación de ingenierías, Universidad La Salle Chihuahua, Prolongación, Lomas de Majalca 11201, Labor de Terrazas, 31020 Chihuahua, Chihuahua, Mexico

² Facultad de Medicina y Ciencias Biomédicas, Universidad Autónoma de Chihuahua, Circuito Universitario S/N, 31125 Chihuahua, Chihuahua, Mexico

ccanedo@uach.mx

Abstract. There are certain parameters in the human body that may be indicators of the presence of breast cancer, these can be assessed with different algorithms such as the Support-Vector Machine (SVM), the Naïve Bayes Algorithm (BA) and Artificial Neural Networks (ANN) to determine whether the laboratory tests are positive or not. Machine Learning (ML) has gained more uses across fields as it proposes a cost-effective classifier with versatility to be developed for any type of application, such as early breast cancer detection. This paper shows an ensemble of the algorithms previously mentioned that, based on a database consisting of 8 characteristics, can provide a high accuracy result. The highest F1 Score obtained was 78.261% from the BA, followed by the ANN's score of 77.273% and a 72.34% from the SVM, resulting in a compositive F1 Score of 80.851%. All the data used on this article was trained using supervised machine learning techniques and variables of interest for breast cancer proliferation.

Keywords: Breast cancer · Classifier · Ensemble learning

1 Introduction

Breast cancer (BC) is defined as the abnormal growth of cells in the breast that can be felt as a lump. It can start in different parts of the breast, such as the lobules, ducts, and connective tissue, and can also metastasize to other parts of the body. This disease is most common amongst women, but men are also capable of developing breast cancer [1].

Cancer itself doesn't have a particular set of causes, however, having high levels of certain substances in the body may be an indicator of its presence.

Women with a higher Body Mass Index (BMI) than recommended tend to have elevated levels of insulin, a hormone that stimulates the proliferation of certain human breast cancer cells when in elevated levels, since it decreases levels of high-density lipoproteins that increase BC risk. [2]; glucose may also play a role in the proliferation of cancerous cells [3] as glucose levels rise when either leptin or Homeostatic Model Assessment of Insulin Resistance (HOMA-IR) increase [4].

On the other hand, having a resistance to insulin that exceeds normal limits can also be a risk factor, since insulin levels would also rise as a mechanism to beat this resistance and incorporate glucose into the cell [5]. At the same time, HOMA and adiponectin levels are inversely related and could be indicators of increased mortality in BC patients [6].

Adipokines are a group of substances that regulate numerous physiological functions, two of its most notorious proteins are leptin and resistin. Both have been shown to be a key factor in endocrine, paracrine, and autocrine evolution of breast cancer and are the direct link between obesity and BC, due to the microenvironment they create for tumor cells [7].

Also, Monocyte Chemoattractant Protein-1 (MCP-1) has been shown to provide favorable conditions for proliferation of tumorous cells, as its expression in malignant cells has a direct relation with tumor associated macrophages accumulation in the tumor area [8].

BC is often easily diagnosed, however if the lump is not of a considerable size, diagnosis may be delayed, complicating any treatment option available. As with any other disease, early treatment is key, which makes it of great importance to optimize current detection methods. Machine Learning algorithms can develop a significant role in detection, confirmation, and follow-up tests for BC, improving patients' life chances and, therefore, recovery times.

The efficiency of the algorithms mentioned on this paper has been a subject of study in BC detection by many authors, Chaurasia, Vikas and Pal, Saurabh obtained and F1 Score of 98% using the BA [9]. However, the SVM is one of the preferred models for this type of application, as it provides high precision classifications, as shown in the paper done by Raweh, Abeer et al. where F1 Scores for both the BA and SVM, were 92.2% and 94.4%, respectively [10].

Other techniques have also been used, such as the Sequential Least Squares Programming Method based on the SVM, K-Nearest Neighbors algorithm (KNN), Decision Tree (DT) and Logistic Regression (LR) for BC detection; Gupta, Madhuri and Gupta, Bharat were able to develop an ensemble algorithm with an F1 Score of 95% [11].

On the other hand, mammogram images can also be processed by ML algorithms to classify it as benign or malignant. Research done by Chand, Satish and Yadavendra, where five different algorithms were used to obtain an ensemble learning model resulted in an F1 Score of 81% for BC tumor classification [12]. While Sun, Xianhe et al. used a Regional Convolutional Neural Network with an F1 Score of 73.6% [13].

The application of machine learning to obtain a better, faster, and more precise cancer diagnosis has already been put to test by various authors; Kourou, Exarchos, et al. provide a similar focus on their paper, where based on already existing studies using the SVM, ANN and BA, the authors were able to conclude that a multidisciplinary technique including feature selection computerized algorithms provides a promising tool not only for diagnosis, but for prognosis as well [14].

2 Methods

2.1 Database

The database used was obtained from the UCI Machine Learning Repository and contains 116 data rows. The characteristics listed on it were age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin and MCP.1, aside from a classification column, where healthy patients were identified with the number 1, and patients with BC with the number 2 [15].

The age column was not considered for the development of the algorithms, as no correlation between age and the presence of BC was found when analyzing the data. All the remaining characteristics were considered for the Naïve Bayes Algorithm and ANN, while for the SVM only resistin and leptin values were used, as they provided more relevant and consistent information.

Aside from this data selection, several rows were deleted to have symmetric classes of 52 data sets each, from which only 30 were used for training and development of algorithms.

2.2 Naïve Bayes Algorithm

This algorithm's purpose is to determine whether a sample vector belongs to a class or another considering the values on each characteristic [16].

The first step consists of calculating the probability of it being positive or negative ($Prob_{neg}$ and $Prob_{pos}$, respectively), using Eqs. (1) and (2). Where t_{neg} corresponds to the probability of the sample belonging to the negative class, and t_{pos} , to the positive class; both of which had values of 50%.

$$Prob_{neg} = \frac{t_{neg}}{(t_{neg} + t_{pos})} \quad (1)$$

$$Prob_{pos} = \frac{t_{pos}}{(t_{neg} + t_{pos})} \quad (2)$$

The mean (\bar{x}) and variance (σ^2) values were calculated for all characteristics on both classes. Then, with the Eq. (3) the probability of it belonging to one of the classes based on each characteristic was calculated.

$$Prob(Pos|C1, \dots, C8) = \frac{1}{\sqrt{2\pi\sigma^2}} \left(\frac{-(data-\mu)^2}{2\sigma^2} \right) \quad (3)$$

The relation between all probabilities of each class was calculated using Eq. (4) for both positive and negative classes

$$Pr_{neg,pos} = Prob_{neg,pos} * Prob_{(neg,pos)|(C1,\dots,C8)} \quad (4)$$

Then, the evidence was obtained adding both results given by the previous equation, which makes for Eq. (5). And the final equation's (6) result indicates the final probability

of classification of the sample, meaning that this equation should be done for positive and negative Pr values.

$$Evidence = Pr_{neg} + Pr_{pos} \quad (5)$$

$$P_s = \frac{Pr_{neg,pos}}{Evidence} \quad (6)$$

The greater number is the one corresponding to the prediction class following the rule:

$$P_{neg} > P_{pos} \text{ or } P_{pos} > P_{neg}$$

2.3 Artificial Neural Network

Artificial Neural Networks are a subset of machine learning which are inspired by the human brain and aim to mimic signaling between biological neurons. They consist of an input layer, an indefinite number of hidden layers and an output layer, and its action is triggered by a hard limit function [17].

The ANN designed contains 8 inputs, 1 hidden layer with 20 neurons and 2 outputs (Fig. 1).

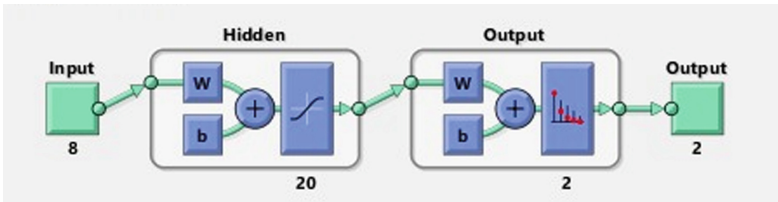


Fig. 1. ANN design with 8 inputs, one hidden layer of 20 neurons and two possible outputs.

The network's training was done following the Levenberg-Marquardt backpropagation algorithm with the hyperparameters obtained during testing:

- Learning rate: 0.01
- Number of epochs: 200,000
- Minimum error: $1e-59$
- Validation check: 1000

We selected these parameters so that the neural network would search for the minimum error, while avoiding over-fit. It should be noted that our values were experimental and that the starting values at the synaptic points were all random with each run.

We chose the network that showed the lowest error, however, the minimum error was never reached, being the closest one obtained in iteration 62, after which the error began to rise, which can be seen in Fig. 3.

Before training 30 rows from each class were selected, where 90% correspond to training values, 5% to validation and the remaining 5% to testing. Obtaining the following parameters for the already trained ANN, as shown on Fig. 2

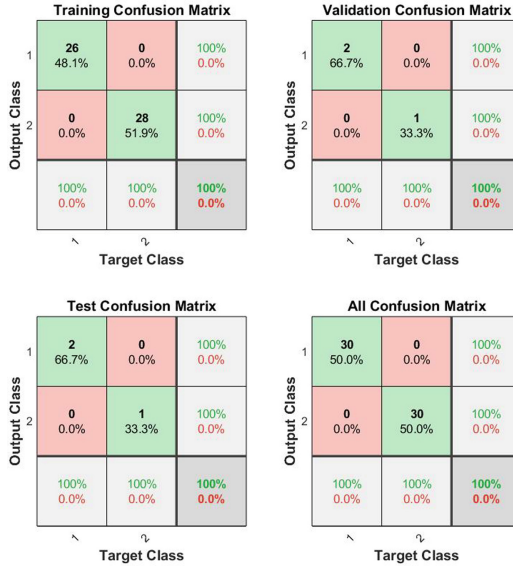


Fig. 2. Training, validation, test and global confusion matrixes obtained with a 100% accuracy.

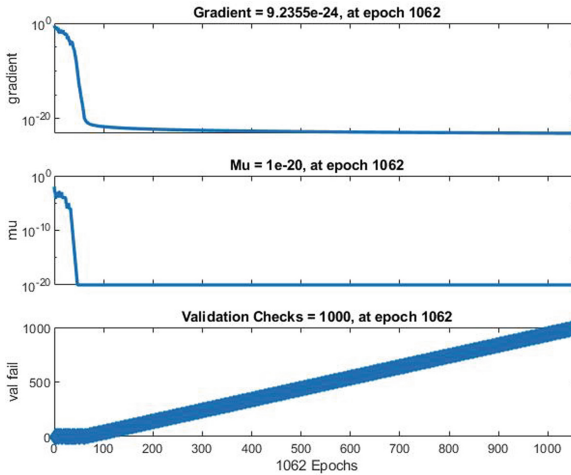


Fig. 3. Training state. The final value of gradient coefficient was 9.2355e-24, near the ideal error of 0.

2.4 Support-Vector Machine

A SVM is a supervised machine learning algorithm used for classification, where each data item is plotted and then separated into categories by a hyper-plane, usually defined as a line [18].

First, the two most significant categories information-wise were selected (resistin and leptin) and the distances between each vector from the negative class and each vector from the positive class were calculated to find the two smallest ones, and therefore find the shared vector, these are the support vectors:

- $S1 = [14.09, 7.64, 1]$
- $S2 = [15.1248, 9.1539, 1]$
- $S3 = [14.9037, 8.2049, 1]$

Where $S1$ and $S2$ correspond to the negative class and $S3$ to the positive one. Alphas were then calculated with equation system (7).

$$\begin{aligned}\alpha_1 S1 * S1 + \alpha_2 S2 * S1 + \alpha_3 S3 * S1 &= -1 \\ \alpha_1 S1 * S2 + \alpha_2 S2 * S2 + \alpha_3 S3 * S2 &= -1 \\ \alpha_1 S1 * S3 + \alpha_2 S2 * S3 + \alpha_3 S3 * S3 &= +1\end{aligned}\quad (7)$$

Then, W_x , W_y and W_b values were calculated with Eqs. (8), (9) and (10).

$$\alpha_1 * S1(1) + \alpha_2 * S2(1) + \alpha_3 * S3(1) = W_x \quad (8)$$

$$\alpha_1 * S1(2) + \alpha_2 * S2(2) + \alpha_3 * S3(2) = W_y \quad (9)$$

$$\alpha_1 * S1(3) + \alpha_2 * S2(3) + \alpha_3 * S3(3) = W_b \quad (10)$$

Finally, the resulting equation is as follows (11):

$$W_x(x) + W_y(y) + W_b = 0 \quad (11)$$

where:

$$W_x = 4.677570008$$

$$W_y = -3.197271579$$

$$W_b = -42.47980654$$

The hyper-plane was plotted to divide both classes (Fig. 4).

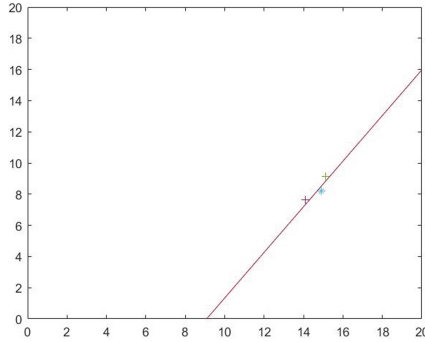


Fig. 4. (+) indicates data from the negative class, while (*) represents data from the positive class.

2.5 Composite Algorithm

A more complex algorithm was created using the outputs of all three previously mentioned algorithms, with the objective of determining if a data set was negative or positive based on the results given by the BA, the ANN and SVM.

If two of the three results point towards the positive class, the patient is likely a BC patient, on the contrary, if at least two results are negative, the person is probably healthy [19] (Table 1 and Fig. 5).

Table 1. Composite algorithm’s possible results.

Bayes	ANN	SVM	Result
P_pos	A_pos	SVM_pos	Positive
P_neg	A_pos	SVM_pos	Positive
P_pos	A_neg	SVM_pos	Positive
P_pos	A_pos	SVM_neg	Positive
P_neg	A_neg	SVM_pos	Negative
P_pos	A_neg	SVM_neg	Negative
P_neg	A_pos	SVM_neg	Negative
P_neg	A_neg	SVM_neg	Negative

ENSEMBLE LEARNING

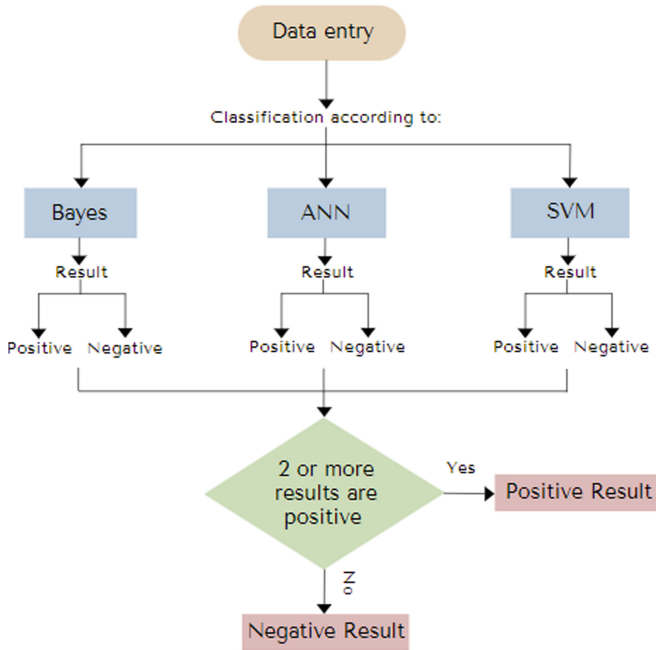


Fig. 5. Representative flowchart of the ensemble learning algorithm: data is processed by each algorithm separately, obtaining a positive or negative result. The three results are compared using an “if” condition. If two or more results fall in the positive category, the result will be positive. On the contrary, if two or more results are negative, the result will be negative.

3 Results

In Fig. 6 the results obtained for the BA are shown, with a precision of 81.81%, accuracy of 77.27% and recall of 75%, which gives an F1 Score of 78.261%.

BAYES			
Actual	Negative	18	4
	Positive	6	16
		Negative	Positive
		Prediction	

Fig. 6. Bayes algorithm confusion matrix, 18 true negatives and 16 true positives obtained.

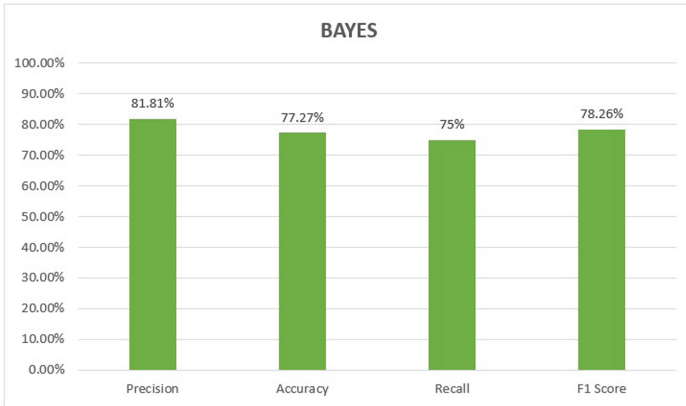


Fig. 7. Bayes algorithm values graph

Figure 8 displays the confusion matrix for the ANN, where precision, accuracy and recall were all 77.27%, giving an F1 Score of 77.27% as well.

ANN			
Actual	Negative	17	5
	Positive	5	17
	Negative	Positive	
	Prediction		

Fig. 8. Artificial Neural Network confusion matrix, with 17 true positives and negatives.

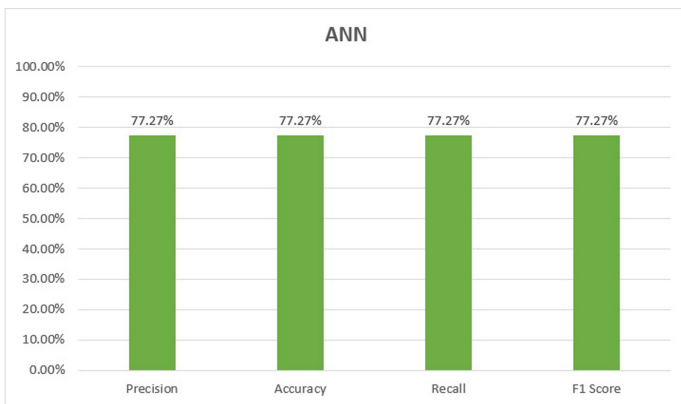


Fig. 9. Artificial Neural Network values graph

Data obtained from the SVM is shown on Fig. 10, with precision of 77.27%, accuracy of 70.45% and recall of 68%, resulting in an F1 Score of 72.34%.

SVM			
Actual	Negative	17	5
	Positive	8	14
		Negative	Positive
		Prediction	

Fig. 10. Support-Vector Machine confusion matrix, where true positives and negatives were 17 and 14, respectively.

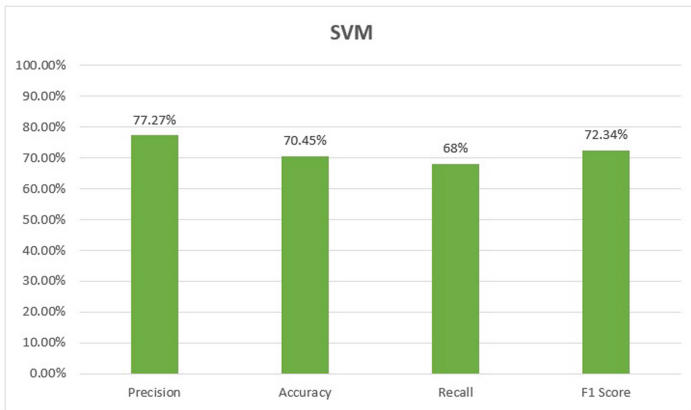


Fig. 11. Support-Vector Machine values graph

For the composite algorithm the values for precision, accuracy and recall were 86.36%, 79.54% and 76%, respectively, for a composite F1 Score of 80.851%.

COMPOSITE ALGORITHM			
Actual	Negative	19	3
	Positive	6	16
		Negative	Positive
		Prediction	

Fig. 12. Composite confusion matrix. A higher accuracy was achieved, where 19 data sets were classified as true negatives and 16 as true positives.

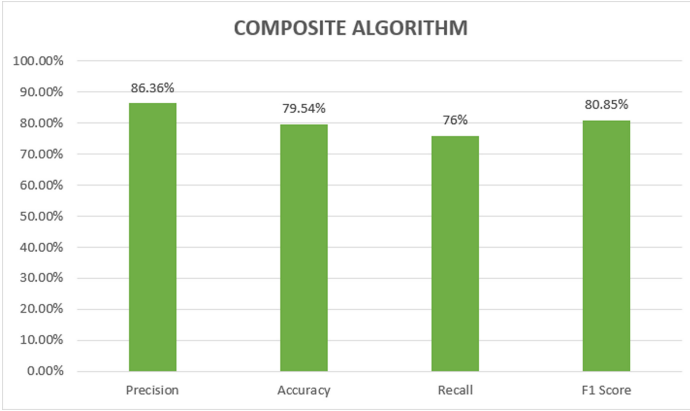


Fig. 13. Composite Algorithm values graph

As shown on Figs. 6, 7, 8, 9, 10, 11, 12 and 13, each algorithm has relatively low values of precision, accuracy, recall and F1 Score on their own, whereas the composite algorithm increases all of these metrics by the use of the three models working as a whole to process the same data sets.

4 Discussion

The analysis of the metrics has shown that the accuracy values are practically identical between SVM and ANN, with 77.27%, while the accuracy of BA was higher by 4.54%. This means that the composite algorithm is the most accurate, with 86.36%.

On the other hand, the accuracy of each algorithm showed similar variation, as the values were 77.27%, 70.45% and 77.27% for BA, SVM and ANN, respectively, for a composite accuracy of 79.54%.

Recall was low for SVM with 68%, while BA had 75% and ANN 77.27%, giving a composite recall of 76%.

All the metrics obtained conclude that the composite algorithm is the most reliable as all the results in it are higher than in the tests of singular algorithms, similar to the work [11], in which the F1 scores of the algorithms operating separately were lower than that obtained by the composite of all of them, with 95%. Table 2 shows the test metrics of various algorithms from the aforementioned paper.

Table 2. Metrics obtained for all the algorithms used by Gupta, Madhuri and Gupta Bharat on their paper on BC prediction using SLSQP.

Classification	Ensemble	SVM	K-NN	DT	LR
Accuracy	97.88	93.98	90.12	92.15	89.12
F-score	95	93	91	92	90
R ²	0.9	0.7	0.7	0.8	0.6
10-Fold	97.2	94.71	89.70	94.2	90.8

It is important to note that even though the purpose of this ensemble is to support medical professionals in diagnosis, it can't operate on its own and its results should be confirmed by a specialist.

5 Conclusions

Ensemble learning provides a different approach to the diagnosis of certain diseases and could represent more reliability than using a single algorithm alone. The F1 score obtained in this work could be greatly improved by using a larger database to train these algorithms.

The results obtained were obtained by analyzing only one database [15], so it is intended to perform a set of tests using the same methodology with other databases, in order to complement the results obtained. At the same time, we intend to conduct a study collaborating with local hospitals to test our development in order to make improvements in the data classification process.

Likewise, it should be noted that our algorithms work with characteristics, although, in comparison with other works it is in a medium ranking, if compared with works that work with image processing, our algorithm shows an advantage, since the metrics obtained are above itself, it does not require such a specialized computer equipment to run the algorithm.

References

1. Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm. Accessed 26 May 2022
2. Rose, D., Vona-Davis, L.: The cellular and molecular mechanisms by which insulin influences breast cancer risk and progression. *Endocr. Relat. Cancer* **19**, 228–233 (2012)
3. National Library of Medicine. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6567675/>. Accessed 26 May 2022
4. Patrício, M., Pereira J. et al.: Using Resistin, glucose, age and BMI to predict the presence of breast cancer, 2–8 (2018)
5. Zacharzewski C., Tibolla, M., et al.: Obesidad y resistencia a la insulina como factores de riesgo en el cáncer de mama, pp. 5–6 (2016)
6. Duggan, C., Irwin, M., et al.: Associations of insulin resistance and adiponectin with mortality in women with breast cancer, pp. 4–5 (2011)
7. National Library of Medicine. <https://pubmed.ncbi.nlm.nih.gov/31637624/#:~:text=Adipokines%20exert%20independent%20and%20joint,dysfunction%20characterized%20by%20chronic%20inflammation>. Accessed 26 May 2022
8. Saji, H., Koike, M., et al.: Significant correlation of monocyte chemoattractant protein-1 expression with neovascularization and progression of breast carcinoma, pp. 2–4 (2001)
9. Chaurasia V., and Pal, S.: Performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer disease, pp. 11–14 (2014)
10. Raweh, A., Nassef, M. and Badr, A.: A hybridized feature selection and extraction approach for enhancing cancer prediction based on DNA methylation, pp. 11–12 (2017)
11. Gupta, M. and Gupta, B.: An ensemble model for breast cancer prediction using Sequential Least Squares Programming Method (SLSQP), pp. 1–3 (2018)

12. Yadavedra and Chand, S.: A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method, pp. 7–10 (2020)
13. Sun, X., Cai, D., et al.: Efficient mitosis detection in breast cancer histology images by RCNN, pp. 3–4 (2019)
14. Kourou, K., Exarchos, T., Karamouzis, M., Fotiadis, D.: Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 1–10 (2014)
15. UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>. Accessed 01 May 2022
16. Java T Point. <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>. Accessed 27 May 2022
17. IBM Cloud Education. <https://www.ibm.com/cloud/learn/neural-networks>. Accessed 27 May 2022
18. Towards Data Science. <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>. Accessed 27 May 2022
19. García H, and Cañedo C.: Diseño de algoritmo compuesto por Machine Learning y un modelo probabilístico para la detección de diabetes