Journal of Fundamental and Applied Sciences

Research Article

ISSN 1112-9867

Available online at http://www.jfas.info

DETECTING BREAST CANCER THROUGH BLOOD ANALYSIS USING DECISION TREE (J48) CLASSIFICATION ALGORITHM

O. Oladimeji[1,2,*], A. Oladimeji[3], O. Oladimeji[2]

[1]Department of Computer Science and Information Technology, Bowen University, Iwo, Nigeria

[2]Department of Computer Science, University of Ibadan, Ibadan, Nigeria

[3]Department of Chemistry, University of Ibadan, Ibadan, Nigeria

Received: 03 February 2020 / Accepted: 29 July 2021 / Published online: 01 September 2021

ABSTRACT

Breast cancer is the second major cause of death in the world. Breast cancer accounts for 16% of all cancer deaths worldwide. Most of the methods of detecting breast cancer very expensive and difficult such as mammography. The objective of this research paper is detecting breast cancer through blood analysis using J48 algorithm which will serve as alternative to these expensive methods.

The J48 algorithm was used to classify 116 instances also,10-fold cross validation and holdout procedure were used coupled changing of random seed. Average accuracies of 84.65% and 89.99% were acquired for cross validation and holdout procedure. Although it was also discovered that Blood Glucose level is a major determinant in detecting breast cancer, it has to be combined with other attributes to make decision as a result of other health issues such as diabetes.

Keywords: J48 Algorithm, Breast Cancer, Decision Tree, Machine learning, Data Mining

Author Correspondence, e-mail: oladimejioladosu@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v13i3.8

1. INTRODUCTION

For the past decade cancer has been a major source of threat to human life [1], but out of the various types of cancer, it was discovered that women are the only group suffering breast cancer, hence has a high mortality rate in women [2]. Sadly, this rate is increasing daily, especially in developed and developing countries [3,4]. However, breast cancer has risen to be second biggest cause of death in the world [5]. As at 2013, it was estimated that 508,000 women died in 2011 as a result of breast cancer worldwide based on World Health Organization (WHO) data [6]. It was also noted that breast cancer is the most common cancer in women.

Generally, cancer is a form of sickness in the cell and then gradually spread into other parts of the body. This is why early detection is very important before it spreads. According to [7], early detection of breast cancer is the most important, expensive and difficult part of breast imaging. Although, many works have been done on early detection of breast cancer in which World Health Organization (WHO) also testified to it that "So far the only breast cancer screening method that has proved to be effective is mammography screening. Mammography screening is very costly and is cost-effective and feasible in countries with good health infrastructure that can afford a long-term organized population-based screening programmes" [6]. This led to this research work, with the aim of detecting breast cancer through blood analysis using J48 algorithm.

J48 (Iterative Dichotomiser 3) is a form of supervised learning algorithm [8], J48 algorithm falls under classification algorithms which is majorly used for prediction based on historical data [9]. that is used to generate decision tree which resembles a flow-chart structurally, whereby each node denotes the test on an attribute and branch denotes the outcome [10-12]

Thus, the main objective of the research paper is to apply machine learning algorithm to detect breast cancer. The second section of this paper discusses the methodology used in this research, while the third section showcase the result, followed by the discussion of the result in the fourth section and finally the conclusion is drawn at the fifth section.

2. MATERIALS AND METHODS

The dataset that was used to pinpoint this research was gotten from UCI Machine Learning Repository [13], Breast Cancer Coimbra dataset. The dataset is consisting of 116 rows with 10 attributes viz. "age (years), BMI (kg/m^2), Glucose (mg/dL), Insulin (µU/mL), HOMA, Leptin (ng/mL), Adiponectin (µg/mL), Resistin (ng/mL) and MCP1(pg/dL)". According to these input

features, target data can be classified as healthy or unhealthy. These features were measured from 64 patients with breast cancer and 52 healthy people [14,15]. This dataset differs from others in terms of the features it contains.

2.1 Data Preprocessing

Based on the dataset collected, all the 10 attributes are numeric, Table 1 shows some of the data before the data preprocessing. In order to make the dataset usable for a classification task, the class was transformed to two categories namely Healthy Control and Patient based on the data description 1= Healthy Controls and 2= Patient while the glucose attribute was transformed into four categories: optimal, excellent, good and danger. Table 2 below shows how the set of rules of glucose(mg/Dl) was classified.

Table 1: Some data used for breast cancer detection before preprocessing

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Class
48	23.5	70	2.707	0.467409	8.8071	9.7024	7.99585	417.114	1
83	20.69049	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
82	23.12467	91	4.498	1.009651	17.9393	22.43204	9.27715	554.697	1
45	20.83	74	4.56	0.832352	7.7529	8.237405	28.0323	382.955	2
49	20.95661	94	12.305	2.853119	11.2406	8.412175	23.1177	573.63	2
34	24.24242	92	21.699	4.924226	16.7353	21.82375	12.06534	481.949	2

Table 2: Categorization of Glucose Classes

Glucose (X)	Class
60<=X<84	OPTIMAL
84<=X<97	EXCELLENT
97<=X<108	GOOD
X>=108	DANGER

Finally, Age attribute was removed to obtain better result. Table 3 below shows some data used for breast cancer detection after preprocessing. Figure 1 shows the visualization of the attributes after

preprocessing. In this process it was discovered that the datasets were skewed (imbalanced), resample filter method was used to resolve the class imbalance problem.

Table 3: Some data used for breast cancer detection after preprocessing

BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Class
27.2	excellent	14.07	3.262364	35.891	9.34663	8.4156	377.227	Healthy controls
30.3	Good	8.34	2.098344	56.502	8.13	4.2989	200.976	Healthy controls
25.3	Optimal	3.508	0.519184	6.633	10.5673	4.6638	209.749	Healthy controls
21.30395	Good	13.852	3.485163	7.6476	21.05663	23.03408	552.444	Patient
20.83	Optimal	4.56	0.832352	7.7529	8.237405	28.0323	382.955	Patient
20.95661	excellent	12.305	2.853119	11.2406	8.412175	23.1177	573.63	Patient

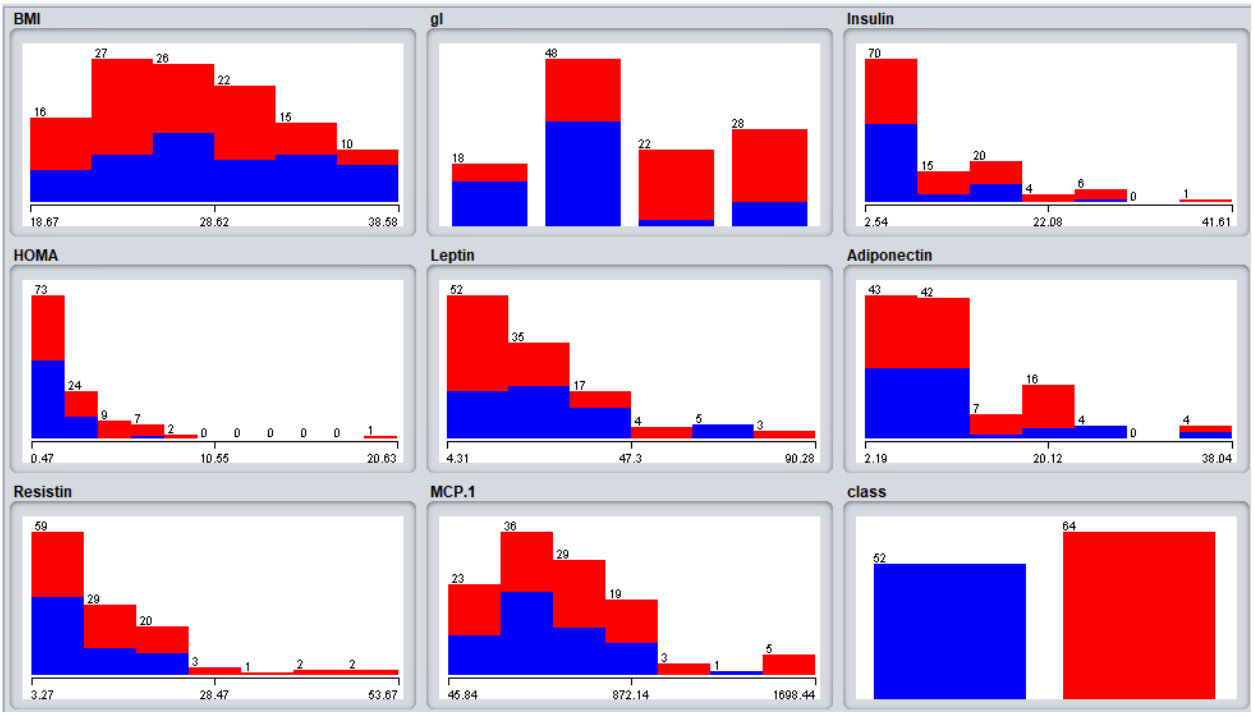

Fig.1. The visualization of the attributes after preprocessing

2.1 Classification

After the data preprocessing the J48 algorithm was implemented with Waikato Environment for Knowledge Analysis (WEKA) which is a tested and trusted open source software for machine learning which was developed at the University of Waikato, New Zealand [16]. Cross validation was selected as the test mode option with 10 as the number of folds and class was set as the target class. This process was done 10 times coupled with changing the random seed starting from 1 -10 for the process for internal validation purposes.

This process was also repeated for percentage split (hold out) test option which was set to 90% in essence, 90% of data was trained on and test was performed on the 10% remainder in order to serve as external validation.

3. RESULT

The algorithm was implemented as stated in the previous section. The performance measures which includes Recall, Precision and F-Measure which are gotten from the confusion matrix which is used to determine how well a classification has performed [17] by reporting the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), is shown in table 4 below while the mean and standard deviation shown in table 5 below.

Precision is given as the number of correctly classified positive examples divided by the number of examples labelled by the system as positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the number of correctly classified positive examples divided by the number of positive examples in the data.

$$Recall = \frac{TP}{TP + FN}$$

F-Measure score is just the harmonic mean of precision and recall.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Also, the decision tree which is the graphical representation of the classification tree for the classification is shown in the figure 2 below, the tree size is 21 and the number of leaves is 12.

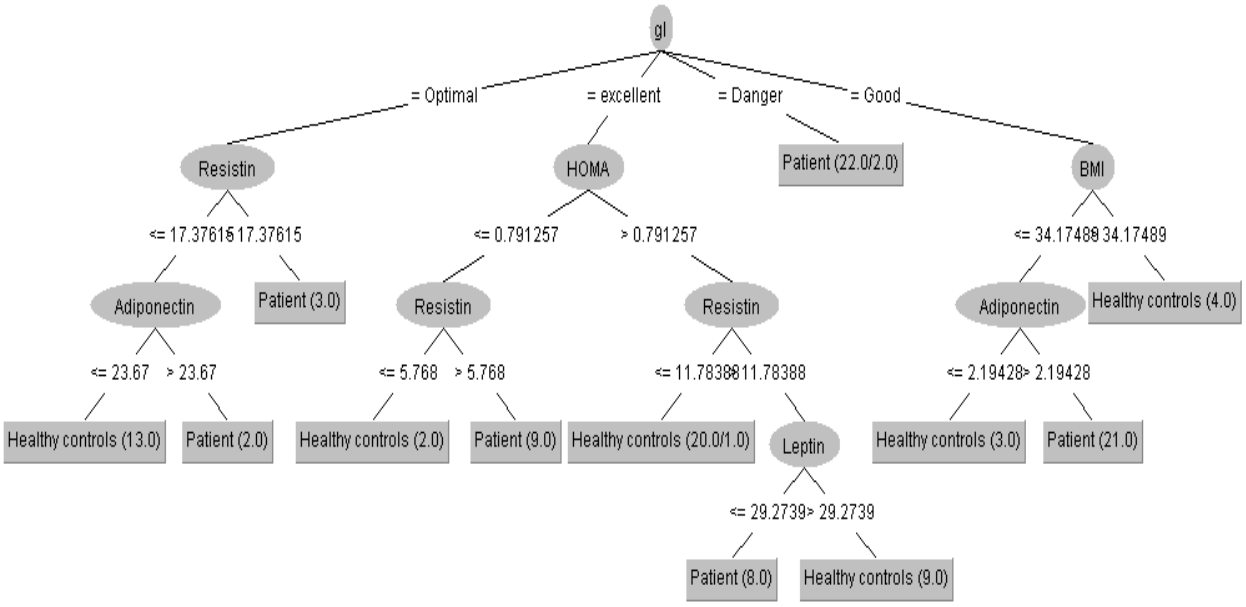

Fig.2. Decision Tree for the Classification

Table 4: Details of performance measure of the classification

	Cross Validation(10-fold)				Holdout(10%)			
Random of Seed	**% Accuracy**	**F-Measure**	**Precision**	**Recall**	**% Accuracy**	**F-Measure**	**Precision**	**Recall**
1	87.069	0.870	0.871	0.871	91.667	0.911	0.925	0.917
2	83.621	0.835	0.837	0.836	100	1.000	1.000	1.000
3	81.897	0.819	0.819	0.819	83.333	0.844	0.900	0.833
4	85.345	0.853	0.853	0.853	100	1.000	1.000	1.000
5	85.345	0.854	0.854	0.853	83.333	0.833	0.833	0.833
6	85.345	0.854	0.854	0.853	91.667	0.913	0.926	0.917
7	88.793	0.888	0.888	0.888	83.333	0.833	0.833	0.833
8	83.621	0.836	0.836	0.836	83.333	0.833	0.833	0.833
9	80.172	0.802	0.802	0.802	100	1.000	1.000	1.000
10	85.345	0.854	0.854	0.853	83.333	0.815	0.867	0.833

Table 5: The mean and standard deviation of the accuracy

	Cross Validation (10-fold)	**Holdout (10%)**
Mean	84.6553	89.999
Standard deviation	2.465247	7.657982

4. DISCUSSION

Based on the result obtained, it can be said Glucose level is major determinant in detecting breast cancer while Resitin, HOMA, BMI, Adiponectin and Leptin are other determinants in detecting breast cancer. Also, Insulin and MCP.1 do not have any effect in detecting breast cancer. Although The interpretation of the decision is given below.

IF Glucose Level = Excellent AND HOMA <= 0.791257 AND Resistin <= 5.768 THEN Class = Healthy Controls

IF Glucose Level = Excellent AND HOMA <= 0.791257 AND Resistin > 5.768 THEN Class = Patient

IF Glucose Level = Excellent AND HOMA > 0.791257 AND Resistin <= 11.78388 THEN Class = Healthy Controls

IF Glucose Level = Excellent AND HOMA > 0.791257 AND Resistin > 11.78388 AND Leptin <= 29.2739 THEN Class = Patient

IF Glucose Level = Excellent AND HOMA > 0.791257 AND Resistin > 11.78388 AND Leptin >29.2739 THEN Class = Healthy Controls

IF Glucose Level = Danger THEN Class = Patient

IF Glucose Level = Optimal AND Resistin <= 17.37615 AND Adiponectin <= 23.67 THEN Class = Healthy Controls

IF Glucose Level = Optimal AND Resistin <= 17.37615 AND Adiponectin >23.67 THEN Class = Patient

IF Glucose Level = Optimal AND Resistin > 17.37615 THEN Class = Patient

IF Glucose Level = Optimal AND Resistin > 18.35574 THEN Class = Healthy Controls

IF Glucose Level = Good AND BMI <= 34.17489 AND Adiponectin <= 2.19428 THEN Class = Healthy Controls

IF Glucose Level = Good AND BMI <= 34.17489 AND Adiponectin > 2.19428 THEN Class = Patient

IF Glucose Level = Good AND BMI > 34.17489 THEN Class = Healthy Controls

With this result, people advised to try maintain Blood Glucose level of excellent which is between 60 and 83 mg/dL inclusive.

5. CONCLUSION

In this paper, we applied Decision Tree (J48) Classification Algorithm to detect breast cancer through blood analysis, with the use of WEKA software. The dataset of 116 instances was acquired from UCI Machine Learning Repository, Breast Cancer Coimbra dataset. A 10-fold cross validation and holdout procedure were used coupled changing of random seed. Average accuracies of 84.65% and 89.99% were acquired for cross validation and holdout procedure. Although it was discovered that Blood Glucose Level is a major determinant in detecting Breast cancer, it has to be combined with other attributes to make final decision because many health conditions may affect glucose level for example diabetes, the same also runs for some of the other included attributes. In addition, this study may support the further work in this field.

6. REFERENCES

[1] Tang J., Rangayyan R. M., Xu J., Naqa I. E, and Yang Y. (2009)., "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 2, pp. 236-251.

[2] Muhammet Fatih Aslan, Yunus Celik, Kadir Sabanci and Akif Durdu. (2018). Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. International Journal of Intelligent Systems and Applications in Engineering. Vol 6(4). DOI: 10.18201/ijisae.2018648455

[3] Ahmad Z., Khurshid A., Qureshi A., Idress R., Asghar N., and Kayani N., (2009). "Breast carcinoma grading, estimation of tumor size, axillary lymph node status, staging, and nottingham prognostic index scoring on mastectomy specimens," Indian Journal of Pathology and Microbiology, vol. 52, no. 4, pp. 477.

[4] Acharya U. R., K. Ng E. Y.-, Tan J.-H., and Sree S. V. (2012). "Thermography based breast cancer detection using texture features and support vector machine," Journal of medical systems, vol. 36, no. 3, pp. 1503-1510.

[5] Ganesan K., Acharya U. R., Chua C. K., Min L. C., Abraham K. T., and Ng K.-H. (2013). "Computer-aided breast cancer detection using mammograms: a review," IEEE Reviews in biomedical engineering, vol. 6, pp. 77-98.

[6] WHO. "Breast cancer: prevention and control," http://www.who.int/cancer/detection/breastcancer/en/index1.html assessed 12 December,2019.

[7] Schreer I., and Lüttges J., (2005). "Breast cancer: early detection," Radiologic-Pathologic Correlations from Head to Toe, pp. 767-784: Springer.

[8] Kotsiantis, S.B. (2007) "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31. pp 249-268

[9] Ahishakiye E., Omulo E.O., Taremwa D. and Niyonzima I. (2017). Crime Prediction Using Decision Tree (J48) Classification Algorithm. International Journal of Computer and Information Technology. Vol 6(3)

[10] Rokach L, Maimon O. "Top – Down Induction of Decision Trees Classifiers – A Survey", IEEE Transactions on Systems

[11] Jiawei, H and Kamber, M (2011). Data Mining: Concepts and Techniques. Morgan Kaufman.

[12] Mahesh Pal, Paul M. Mather. (2003). An assessment of the effectiveness of decision tree methods for land cover classification, Remote Sensing of Environment 86(2003) .pp. 554-565, ScienceDirect.

[13] UCI. "Machine Learning Repository," https://archive.ics.uci.edu/ml/index.php.

[14] Patrício M., Pereira J., Crisóstomo J., Matafome P., Gomes M., R. Seiça, and Caramelo F. (2018). "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," BMC cancer, vol. 18(1), pp. 29.

[15] Crisóstomo J., Matafome P., Santos-Silva D., Gomes A. L., Gomes M., Patrício M., Letra L., Sarmento-Ribeiro A. B., Santos L. and Seiça R. (2016). "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer," International Journal of Basic and Clinical Endocrinology, vol. 53(2), pp. 433-442.

[16] www.cs.waikato.ac.nz/ml/weka assessed 12 December,2019.

[17] Pablo Diez (2018). Smart Wheelchairs and Brain-Computer Interfaces, ScienceDirect.

How to cite this article:
Oladimeji O, Oladimeji A, Oladimeji O. Detecting breast cancer through blood analysis using decision tree (J48) classification algorithm. J. Fundam. Appl. Sci., 2021, *13(3), 1275-1284.*