Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Detecting Breast Cancer through Blood Analysis Data using Classification Algorithms

Oladosu Oladimeji[1*] and Olayanju Oladimeji[2]

1. Department of Computer Science, University of Ibadan, Ibadan, Nigeria.
2. Department of Computer Science and Information Technology, Bowen University, Iwo, Nigeria.

Article Info

Article History:
Received 27 June 2020
Revised 12 March 2021
Accepted 27 May 2021

DOI:10.22044/JADM.2021.9839.2116

Keywords:
Classification Algorithm, Breast Cancer, Machine Learning, Data Mining.

**Corresponding author: oladimejioladosu@gmail.com (O. Oladimehi).*

Abstract

Breast cancer is the second major cause of death, and it accounts for 16% of all cancer deaths worldwide. Most of the methods for detecting breast cancer such as mammography are very expensive and difficult to interpret. There are also limitations like cumulative radiation exposure, over-diagnosis, and false positives and negatives in women with a dense breast that pose certain uncertainties in the high-risk populations. The objective of this work is to create a model that detects breast cancer through blood analysis data using the classification algorithms. This serves as a complement to the expensive methods. High-ranking features are extracted from the dataset. The KNN, SVM, and J48 algorithms are used as the training platform in order to classify 116 instances. Furthermore, the 10-fold cross-validation and holdout procedures are used coupled with changing of random seed. The results obtained show that the KNN algorithm has the highest and best accuracies of 89.99% and 85.21% for the cross-validation and holdout procedures, respectively. This is followed by the J48 algorithm with accuracies of 84.65% and 75.65% for the two procedures, respectively. The SVM algorithm has the accuracies of 77.58 and 68.69%, respectively. Although, it has also been discovered that the blood glucose level is a major determinant in detecting the breast cancer, it has to be combined with other attributes to make decisions as a result of other health issues like diabetes. With the results obtained, women are advised to do regular check-ups including blood analysis to know which blood components are required to be worked on in order to prevent breast cancer based on the model generated in this work.

1. Introduction

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

Generally, cancer is the uncontrolled growth of abnormal cells anywhere in the body. Breast cancer is the cancer that can develop in the breast cells [7]. If not treated, it extends to other parts of the body. This is why early detection is very important before it spreads. Also [8] has explained that the risk for breast cancer increases with age, and most breast cancers are diagnosed after reaching the menopause age. According to [9], an early detection of breast cancer is very important to have a better chance of survival.

However, many works have been done on the early detection of breast cancer. WHO likewise testified to it saying: "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 the countries with a good health infrastructure that can afford a long-term organized population-based screening program" [6]. Apart from being costly, there are also limitations such as cumulative radiation exposure, over-diagnosis, and false positives and negatives in women with a dense breast. As a result, there are certain uncertainties in high-risk populations [10, 11].

This led to this research work. An early detection of breast cancer helps to increase the survival rate. This research work aims to get biomarkers from blood analysis data for the detection of breast cancer. It aims at detecting breast cancer through the blood analysis data. This is by collecting values of the level of glucose, insulin, HOMA, leptin, adiponectin, resistin, MCP-1, age, and body mass index (BMI). These parameters are believed to be a good set of components. This is because [12] has recently verified a deregulation in their profile in the obesity-associated breast cancer. Then creating a model by using the classification algorithms such as the J48, K-nearest neighbor, support vector machine algorithms that can be used to create a biomarker for the breast cancer prediction. The model will help in supporting the medical decisions.

Classification is a machine learning technique in which the data is categorized into a given number of classes. For example, the study aims at classifying a given data to either the breast cancer or healthy category. J48 (iterative dichotomiser 3) is a form of supervised learning algorithm [13]. The J48 algorithm falls under the classification algorithms, and is majorly used for prediction based on the historical data [14]. It is used to generate a decision tree that resembles a flow chart structurally, whereby each node denotes the test on an attribute, and branch denotes the outcome [15-17]. The J48 algorithm works by generating rules for predicting the target variable based on the dataset supplied. These rules are generated based on the values of the attributes of the dataset. For this study, the J48 algorithm will generate the rules based on the values of the blood attributes (resistin, leptin, glucose, and others) in order to classify the data into cancer positive or negative.

The Support Vector Machine (SVM) is a supervised machine learning algorithm used for the pattern classification and non-linear regression of the features. It is an estimated implementation of the method of structural risk minimization that provides a good generalization on a pattern classification problem. Given a set of training examples (blood analysis data), each marked as belonging to either the positive or negative category. A SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier [14]. The K- Nearest Neighbours (KNNs), also known as the case-based reasoning, has been used in many applications like pattern recognition and statistical estimation. It is a simple, lazy, and non-parametric classifier.

Since the past decades, the machine learning and data mining have become very popular in the prominent research works in virtually every aspect of the human activities [18]. The importance of machine learning cannot be overemphasized. Few instances in which machine learning has been used include crime rate prediction using the decision tree (J48) algorithm [14], in which 94% accuracy has been achieved, which is fair enough for the system to be relied on for prediction of the future crimes. Also the machine learning technique has been applied by [19] for an accurate diagnosis of the coronary artery disease, in which 93% accuracy has been achieved. Also machine learning has been applied for the credit card fraud detection purposes [20] with an over 99.6% accuracy, among others.

Likewise, machine learning has been applied to detect breast cancer such as [2], in which the authors have compared four machine learning algorithms in order to predict the breast cancer. They used datasets on the Support Vector Machine (SVM), artificial neural networks, Extreme Learning Machine (ELM), and K-nearest neighbor, and it was observed that ELM performed best with an accuracy of 80%. Also in [21], the authors have used Neural Networks (NNs), Decision Tree (DT), naïve bayes, and K-nearest neighbors in order to build the machine learning models in which artificial neural networks has the highest accuracy of 80%. Likewise, in [22], the authors have built their models based on DT, SVM, RF, LR, and NN, in which RF has performed best with an accuracy of 74.3%.

It was observed that one of the major challenges of machine learning is high dimensionality of the dataset [23]. This is as a result of a large memory required due to the analysis of many features that leads to overfitting. Therefore, the weighting features reduce the redundant data and processing time, thereby, improving the performance of the algorithm [24].

Thus the main objective of the research paper is to apply the machine learning algorithm to detect the breast cancer using feature selection, which eliminates the unnecessary and unimportant features in the dataset [24] in order to obtain better results compared to [2, 21-22]. The second section discusses the methodology used in this research work, while the third section showcases the results, followed by the discussion of the results in the fourth section, and finally, the conclusions are made in the fifth section.

2. Methods

The description of the proposed methodology is given below:

1. Pre-processing (data manipulation and normalization): Numerical attributes (class) changed to nominal values, glucose changed to ordinal values.
2. Feature selection: performed using the ReliefF algorithm coupled with the ranker search method.
3. Classification-3 classifiers were tested: kNN, SVM, J48.
4. Evaluation of results-based on confusion matrix (accuracy, recall, precision, and F-measure metrics).

The proposed methodology for this work was formulated using the WEKA software, an open source software for machine learning that was developed at the University of Waikato. The dataset that was used to pinpoint this research work was obtained from UCI Machine Learning Repository [20], Breast Cancer Coimbra dataset, which was loaded into WEKA. The dataset contained 64 breast cancer positive and 52 negative [25, 26]. In order to obtain a better result, feature selection was used for selecting the attributes to be used for the classification. In this research work, the cross-validation and holdout methods were used. For the hold-out method, the dataset was divided into 80% training dataset and 20% test examples. However, the entire training dataset was used for the cross-validation method. The J48, Support Vector Machine (SVM) (LibSVM), and K-Nearest Neighbour (KNN) (IBK) algorithms were used in this work.

2.1. Data Description

The dataset consists of 116 rows with 10 attributes viz. “age (years), BMI (kg/m^2), glucose (mg/dL), insulin ($\mu U/mL$), HOMA, leptin (ng/mL), adiponectin ($\mu G/mL$), resistin (ng/mL), and MCP1 (pg/dL)”. Glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP1 can be collected in the routine blood analyses. BMI (kg/m^2) was obtained by taking the ratio of weight and square height, HOMA = $\log((If)*(Gf))/22.5$, where (If) is the fasting insulin level ($\mu U/mL$) and (Gf) is the fasting glucose level ($mmol/L$).

2.2. Data Pre-processing

Based on the dataset collected, all the 10 attributes are numeric. Table 1 shows some of the data before data pre-processing. In order to make the dataset usable for a classification task, the class was transformed into two categories namely Healthy Control and Patient. Based on the data description 1 = Healthy Controls and 2 = Patient. The glucose attribute was transformed into four categories: optimal, excellent, good, and dangerous.

Table 2 below shows the range of glucose (mg/dL) classification.

Table 1. Some data used for breast cancer detection before pre-processing.

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)	Glucose class
$60 \leq X < 84$	Optimal
$84 \leq X < 97$	Excellent
$97 \leq X < 108$	Good
$X \geq 108$	Dangerous

Table 3 below shows some data used for breast cancer detection after pre-processing. Figure 1 shows the visualization of the attributes after pre-processing, in which the red colour denotes the positive class and the blue colour denotes the negative class. In this process, it was discovered that the datasets were skewed (imbalanced), and the resample filter method was used to resolve the class imbalance problem.

Table 3. Some data used for breast cancer detection after pre-processing.

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Class
83	20.69049454	Excellent	3.115	0.706897333	8.8438	5.429285	4.06405	468.786	Healthy controls
71	30.3	Good	8.34	2.098344	56.502	8.13	4.2989	200.976	Healthy controls
78	25.3	Optimal	3.508	0.519184	6.633	10.5673	4.6638	209.749	Healthy controls
45	21.30395	Good	13.852	3.485163	7.6476	21.05663	23.03408	552.444	Patient
46	20.83	Optimal	4.56	0.832352	7.7529	8.237405	28.0323	382.955	Patient
49	20.95661	Excellent	12.305	2.853119	11.2406	8.412175	23.1177	573.63	Patient

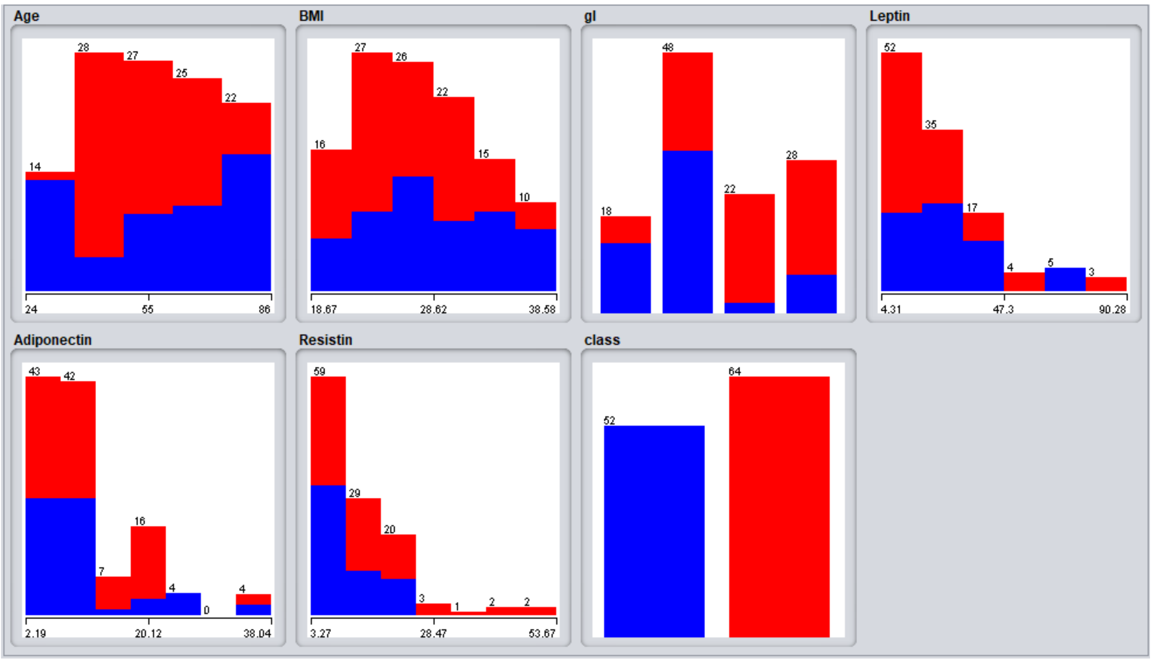

Figure 1. Visualization of the attributes after pre-processing.

2.2.1. Data Selection

The data selection phase involves understanding the datasets and selecting the attributes that will produce the necessary data required to infer the knowledge sought. This is also known as feature selection, which is a process for identifying the subset of data from a large dimension of data [24, 27]. The attributes contributing more to the development of the model were derived using ReliefFAttributeEvaluator (RF) coupled with the ranker algorithm. ReliefF was selected since it could deal with both the nominal and numerical attributes, and it was a robust algorithm [28].

Table 4 presents a summary of the attributes and how ReliefFAttributeEvaluator (RF) ranked them.

Table 4. Summary of the evaluator's ranking of each attribute of the dataset.

Attributes	Ranking of ReliefFAttributeEvaluator (RF)
Glucose	0.1689
Age	0.0846
BMI	0.0265
Leptin	0.0244
Resistin	0.0191
Adiponectin	0.0171
MCP.1	0.0158
HOMA	0.0033
Insulin	0.0008

The first six highest ranked attributes by the evaluator as the best influencing breast cancer detection are glucose, age, BMI, leptin, resistin, and adiponectin. Hence, they are selected for the classification problem.

2.3. Classification

After data pre-processing, the J48, KNN (IBK), and (LibSVM) SVM algorithm were implemented using Waikato Environment for Knowledge Analysis (WEKA). It is a tested and trusted open source software for machine learning developed at the University of Waikato, New Zealand [29]. Cross-validation was selected as the test mode option with 10 as the number of folds, and the class attribute was set as the target to be predicted for the classification. This process was done 5 times coupled with changing the random seed starting from 1-5 for the process for internal validation purposes.

This process was also repeated for percentage split (hold out) test option, which was set to 80% in essence. 80% of the data was trained on and the test was performed on the 20% remainder in order to serve as the external validation.

3. Results

The algorithms were implemented as stated in the previous section. The performance measures including the recall, precision, and F-measure, which were obtained from the confusion matrix, were used in order to determine how well a classification performed [30] by reporting the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Table 5). The mean and standard deviations are in Table 6, as shown below.

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

$$\Pr ecision = \frac{TP}{TP + FP} \quad (1)$$

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

$$\operatorname{Re} call = \frac{TP}{TP + FN} \quad (2)$$

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

$$F - measure = \frac{2 * \Pr ecision * Recall}{P\,\mathrm{r}\,ecision + \operatorname{Re} call} \quad (3)$$

Similarly, the decision tree, which is the graphical representation of the classification tree for the classification, is shown in the Figure 2; the tree size was 25 and the number of leaves was 14.

4. Discussion

Based on the result obtained from the tree generated from the feature selection and J48 algorithm, it can be said that the glucose level is major determinant in detecting the breast cancer. Age, resitin, HOMA, BMI, adiponectin, and leptin are the other determinants in detecting the breast cancer. However, insulin and MCP.1 have no effect in detecting the breast cancer. Hence, the biomarker for breast cancer detection is the combination of glucose, age, BMI, adiponectin, and leptin. It was likewise discovered that better accuracies were obtained compared to [2, 21, 22] due to the feature selection of the variables that will help for a better decision-making. One of the major advantages of the proposed methodology is that the limitations such as the cumulative radiation exposure, over-diagnosis, and false positives and negatives in women with a dense breast that pose certain uncertainties in the high-risk populations in mammography is not a limitation here. Another major advantage is that this method is not difficult to interpret. The disadvantage of this methodology is that the

features of the blood analysis data have to be combined together in order to make a decision.
The intention of this model is not to create an alternative to mammography but to complement it. This is done through blood analysis, which is not expensive compared to mammography.

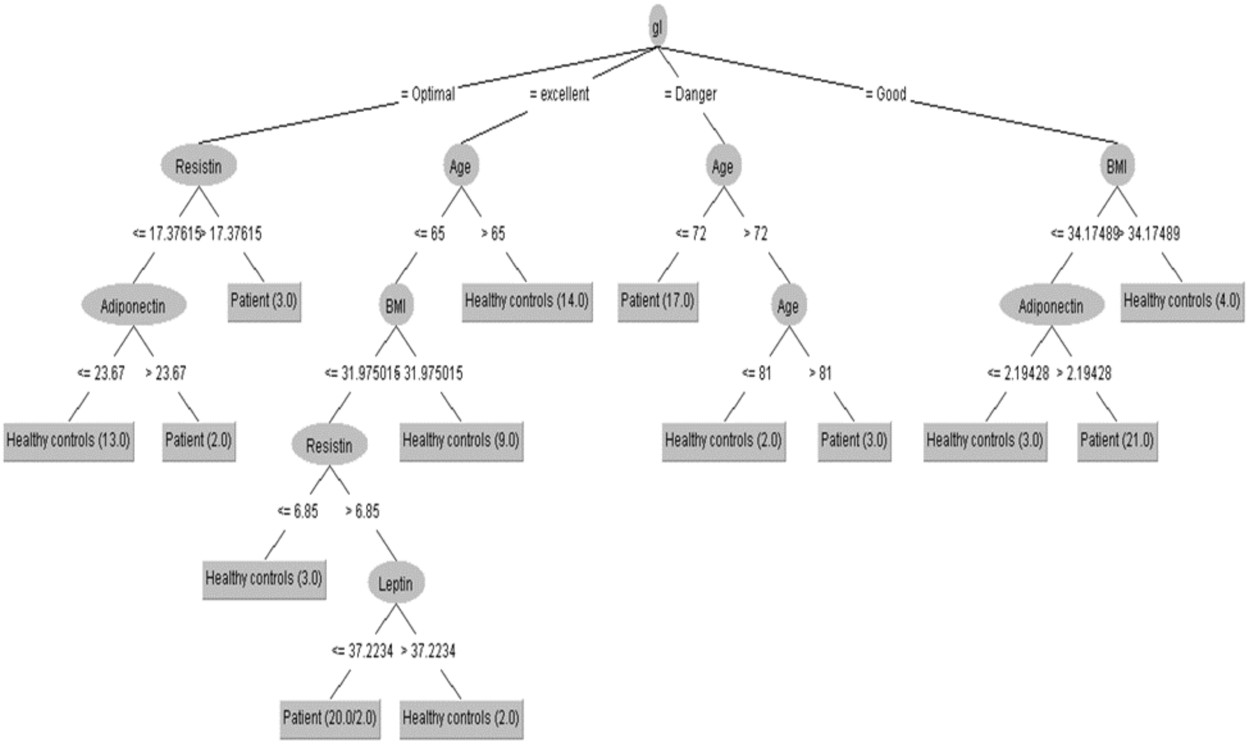

Figure 2. Decision tree for the classification.

Table 5. Details of the performance measure of the classification.

	Test option	Random seed/metrics	1	2	3	4	5
J48 Algorithm	Cross-validation (10-fold)	% Accuracy	87.931	82.759	83.621	85.345	83.621
		F-measure	0.879	0.825	0.836	0.853	0.836
		Precision	0.879	0.832	0.837	0.855	0.836
		Recall	0.879	0.828	0.836	0.853	0.836
	Holdout (20%)	% Accuracy	86.957	78.261	86.957	60.869	65.217
		F-measure	0.857	0.781	0.872	0.607	0.653
		Precision	0.890	0.782	0.878	0.608	0.665
		Recall	0.870	0.783	0.870	0.609	0.652
K- nearest neighbor (IBk)	Cross-validation (10-fold)	% Accuracy	91.379	88.793	87.931	90.517	91.379
		F-measure	0.914	0.888	0.880	0.905	0.914
		Precision	0.914	0.888	0.885	0.907	0.916
		Recall	0.914	0.888	0.879	0.905	0.914
	Holdout (20%)	% Accuracy	91.304	82.609	86.957	82.609	82.609
		F-measure	0.913	0.827	0.874	0.825	0.823
		Precision	0.913	0.840	0.909	0.837	0.833
		Recall	0.913	0.826	0.870	0.826	0.826
Support vector machine (LibSvm)	Cross-validation (10-fold)	% Accuracy	77.586	74.138	79.310	79.310	77.586
		F-measure	0.757	0.714	0.778	0.778	0.757
		Precision	0.841	0.824	0.850	0.850	0.841
		Recall	0.776	0.741	0.793	0.793	0.776
	Holdout (20%)	% Accuracy	82.609	78.261	52.174	69.565	60.870
		F-measure	0.801	0.764	0.502	0.670	0.566
		Precision	0.861	0.843	0.814	0.814	0.794
		Recall	0.826	0.783	0.522	0.696	0.609

Table 6 .The mean and standard deviations of accuracy.

		Cross-validation (10-fold)	Holdout (20%)
J48 algorithm	Mean	84.6554	75.6522
	Standard deviation	2.0579	12.1433
K-nearest neighbor	Mean	89.9998	85.2176
	Standard deviation	1.5659	3.8886
Support vector machine	Mean	77.586	68.6958
	Standard deviation	2.1114	12.4503

4. Discussion

Based on the result obtained from the tree generated from the feature selection and J48 algorithm, it can be said that the glucose level is major determinant in detecting the breast cancer. Age, resitin, HOMA, BMI, adiponectin, and leptin are the other determinants in detecting the breast cancer. However, insulin and MCP.1 have no effect in detecting the breast cancer. Hence, the biomarker for breast cancer detection is the combination of glucose, age, BMI, adiponectin, and leptin. It was likewise discovered that better accuracies were obtained compared to [2, 21, 22] due to the feature selection of the variables that will help for a better decision-making.

One of the major advantages of the proposed methodology is that the limitations such as the cumulative radiation exposure, over-diagnosis, and false positives and negatives in women with a dense breast that pose certain uncertainties in the high-risk populations in mammography is not a limitation here. Another major advantage is that this method is not difficult to interpret. The disadvantage of this methodology is that the features of the blood analysis data have to be combined together in order to make a decision.

The intention of this model is not to create an alternative to mammography but to complement it. This is done through blood analysis, which is not expensive compared to mammography.

5. Conclusion

In this work, we applied the classification algorithms in order to detect breast cancer through blood analysis using the WEKA software. The datasets of 116 instances were acquired from the UCI Machine Learning Repository, Breast Cancer Coimbra dataset. A 10-fold cross-validation and the holdout procedure were used coupled with changing of random seed. The results obtained showed that the KNN algorithm had the highest and the best accuracies of 89.99% and 85.21% for cross-validation and the holdout procedure, respectively. This was followed by the J48 algorithm with the accuracies of 84.65% and 75.65% for the two procedures, respectively. The SVM algorithm had the accuracies of 77.58% and 68.69%, respectively. Although it was discovered that the blood glucose level was a major determinant in detecting breast cancer, it had to be combined with other attributes before arriving at the final decision. This is because many health conditions such as diabetes may affect the glucose level. The same thing also goes for some of the other included attributes. In addition, the present work did not have any data on the irisin or visfatin level from the blood analysis data. Therefore, it would be interesting to include it in a future work. Similarly, further work could be of interest in extending to the other forms of cancer.

References

[1] J. Tang, R. M. Rangayyan, J. Xu, I. E. Naqa, and Y. Yang, “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, 2009.

[2] M. F. Aslan, Y. Celik, K. Sabanci, and A. Durdu, “Breast Cancer Diagnosis by Different Machine Learning Methods using Blood Analysis Data,” *International Journal of Intelligent Systems and Applications in Engineering*, vol. 6, no. 4, 2018.

[3] Z. Ahmad, A. Khurshid, A. Qureshi, R. Idress, N. Asghar, and N. Kayani, “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, 2009.

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

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

[6]http://www.who.int/cancer/detection/breastcancer/en/index1.html.

[7] U. Raghavendra, A. Gudigar, N. T. Rao, E. J. Ciaccio, E. Y. Ng, and U. R. Acharya, “Computer-aided diagnosis for the identification of breast cancer using thermogram images: A comprehensive review,” Infrared Physics and Technology, vol. 102, 2019

[8]https://www.cdc.gov/cancer/breast/basic_info/risk_factor.htm.

[9] I. Schreer and J. Lüttges, “Breast cancer: early detection,” In Radiologic-Pathologic Correlations from Head to Toe, Germany, pp. 767-784, 2005.

[10] J. Melnikow, J. J. Fenton, E. P. Whitlock, D. L. Miglioretti, M. S Weyrich, and J. H. Thompson, “Supplemental screening for breast cancer in women with dense breasts: a systematic review for the U.S. Preventive Services Task Force,” Ann Intern Med. Vol. 164, 2016.

[11] A. B Miller, C. Wall, C. J. Baines, P. Sun, T. To, and S. A. Narod, Twenty-five-year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomized screening trial,” BMJ, 2014.

[12] J. Crisóstomo, P. Matafome, D. Santos-Silva, A. Gomes, M. Gomes, M. Patricio, L. Letra, A. Sarmento-Ribeiro, L. Santos, R. Seica, “Hyperresistinemia and metabolic dysregulation: the close crosstalk in obese breast cancer,” Endocrine, vol. 53, no. 2, 2016.

[13] S.B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” Informatica, vol. 31, pp 249-268, 2007.

[14] E. Ahishakiye, E. O. Omulo, D. Taremwa, and I. Niyonzima, “Crime Prediction Using Decision Tree (J48) Classification Algorithm,” International Journal of Computer and Information Technology, vol. 6, no. 3, 2017.

[15] L. Rokach and O. Maimon, “Top – Down Induction of Decision Trees Classifiers–A Survey,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 35, no. 4, pp. 476-487, 2005.

[16] H. Jiawei and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufman, 2011.

[17] M. Pal and P. M. Mather, “An assessment of the effectiveness of decision tree methods for land cover classification,” *Remote Sensing of Environment*, vol. 86, pp. 554-565, 2003.

[18] O. O. Oladimeji and O. O. Oladimeji, “Exploring Data Mining Research in West Africa: A Bibliometric Analysis,” SLIS Connecting, vol. 9, no. 2, 2020.

[19] M. Abdar, W. Ksiazek, U. R. Acharya, R. Tan, V. Makarenkov, and P. A. Plawiak, “A New Machine Learning Technique for an Accurate Diagnosis of Coronary Artery Disease,” *00*, vol. 179, 2019.

[20] S. Maniraj, A. Saini, S. D. Sarka, and S. Ahmed, “Credit Card Fraud Detection using Machine Learning and Data Science,” *International Journal of Engineering Research and Technology*, vol. 8, no. 9, 2019.

[21] M.U. Ghani, T.M. Alam, and F.H. Jaskani, “Comparison of Classification Models for Early Prediction of Breast Cancer,” *In 2019 International Conference on Innovative Computing (ICIC)*, pp. 1-6, 2019.

[22] Y. Li and Z. Chen, “Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction,” Applied and Computational Mathematics, vol. 7, no. 4, pp. 212 -216, 2018.

[23] P. Domingos, “A few useful things to know about machine learning,” Commun ACM, vol. 55, no. 10., 2012.

[24] R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, “Supervised feature selection with a stratified feature weighting method.” *IEEE Access*, vol. 6, 2018.

[25] https://archive.ics.uci.edu/ml/index.php.

[26] 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, 2018.

[27] J. G. Santillan-Benitez, H. Mendieta-Zeron, L. M. Gomez-Olivan, J. J. Torres-Juarez, J. M. Gonzalez-Banales, L. V. Hernandez-Pena, and A. Ordonez-Quiroz, “*The tetrad BMI, Leptin, Leptin/Adiponectin (L/a) ratio and CA 15-3 are reliable biomarkers of breast cancer*,” J Clin Lab Anal., vol. 27, no. 1, pp. 12–20, 2013.

[28] R. Durgabai and Y. RaviBhushan, “Feature selection using ReliefF Algorithm,” International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, 2014.

[29] www.cs.waikato.ac.nz/ml/weka.

[30] P. Diez, “Smart Wheelchairs and Brain-Computer Interfaces,” Elsevier; 2018

5. Appendix

The interpretation of the decision tree gotten from the J48 algorithm is given below:

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 = excellent and age <=65 and BMI <= 31.975015 and resitin <= 6.85, then class = healthy controls.

If glucose level = excellent and age <=65 and BMI <= 31.975015 and resitin > 6.85 and Leptin <= 37.2234, then class = patient.

If glucose level = excellent and age <= 65 and BMI <= 31.975015 and resitin > 6.85 and leptin > 37.2234, then class = healthy controls.

If glucose level = excellent and age <= 65 and BMI > 31.975015, then class = healthy controls.

If glucose level = excellent and Age > 65, then class = healthy controls.

If glucose level = dangerous and age <= 72, then class = patient.

If glucose level = dangerous and Age > 72 and Age <= 81, then class = healthy controls.

If glucose level = dangerous and age >72 and age > 81, then class = patient.

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.

تشخیص سرطان پستان از طریق داده‌های تجزیه و تحلیل خون با استفاده از الگوریتم‌های طبقه‌بندی

Olayanju Oladimeji [۱,*] و Oladosu Oladimeji [۲]

[۱] گروه علوم کامپیوتر، دانشگاه ابادان، ایبادان، نیجریه.

[۲] گروه علوم کامپیوتر و فناوری اطلاعات، دانشگاه بوون، ایوو، نیجریه.

ارسال ۲۰۲۰/۰۶/۲۷؛ بازنگری ۲۰۲۱/۰۳/۱۲؛ پذیرش ۲۰۲۱/۰۵/۲۷

چکیده:

سرطان پستان دومین علت اصلی مرگ و میر است و ۱۶٪ از کل مرگ‌های سرطانی را در سراسر جهان تشکیل می‌دهد. بیشتر روش‌های تشخیص سرطان پستان مانند ماموگرافی بسیار گران است و تفسیر آن دشوار است. همچنین محدودیت هایی مانند قرار گرفتن در معرض اشعه تجمع، تشخیص بیش از حد و مثبت و منفی کاذب در زنان دارای پستان متراکم وجود دارد که عدم قطعیت خاصی را در جمعیت‌های پرخطر ایجاد می‌کند. هدف از این کار ایجاد مدلی است که با استفاده از الگوریتم‌های طبقه‌بندی، سرطان پستان را از طریق داده‌های آنالیز خون تشخیص می‌دهد. این به عنوان مکمل روشهای گران قیمت عمل می‌کند. ویژگی‌های رده بالا از مجموعه داده استخراج می‌شود. به منظور طبقه بندی ۱۱۶ نمونه، از الگوریتم‌های KNN، SVM و J48 به عنوان بستر آموزشی استفاده می‌شود. علاوه بر این ، ۱۰ برابر اعتبار سنجی و روش‌های نگهداری همراه با تغییر بذر تصادفی استفاده می‌شود. نتایج به دست آمده نشان می‌دهد که الگوریتم KNN دارای بالاترین و بهترین دقت به ترتیب ۹۹،۸۹٪ و ۲۱۱،۸۵٪ برای روش‌های اعتبار سنجی متقابل و نگهداری است. به دنبال آن الگوریتم J48 با دقت ۸۴،۶۵٪ و ۷۵،۶۵٪ برای دو روش به ترتیب دنبال می‌شود. دقت الگوریتم SVM به ترتیب ۷۷/۵۸ و ۶۸/۶۹ درصد است. اگرچه همچنین مشخص شده است که سطح گلوکز خون یک عامل اصلی تعیین کننده در تشخیص سرطان پستان است، اما برای تصمیم‌گیری در نتیجه سایر موارد بهداشتی مانند دیابت، باید با سایر خصوصیات ترکیب شود. با نتایج به دست آمده، به زنان توصیه می‌شود معاینات منظمی از جمله تجزیه و تحلیل خون را انجام دهند تا بدانند که برای جلوگیری از سرطان پستان بر اساس الگوی تولید شده در این کار، بر روی کدام اجزای خونی باید کار شود.

کلمات کلیدی: الگوریتم طبق بندی، سرطان پستان، یادگیری ماشین، داده‌کاوی.