OPEN ACCESS Freely available online

Journal of Data Mining in Genomics & Proteomics

Research Article

Application of Machine Learning Techniques in Predicting of Breast Cancer Metastases Using Decision Tree Algorithm, in Sokoto Northwestern Nigeria

Abdulrahaman A. Musa[1]*, Usman Malami Aliyu[2]

[1]*Department of Histopathology, Usmanu Danfodiyo University Teaching Hospital, Sokoto State, Nigeria;* [2]*Department of Radiotherapy and Oncology, Usmanu Danfodiyo University Teaching Hospital, Sokoto State, Nigeria*

ABSTRACT

According to international agency for research on cancer, female breast cancer was the leading type of cancer worldwide in terms of the number of new cases (approximately 2.1 million) diagnosed in 2018.

Predicting outcome of a disease is a challenging task. Data mining techniques tends to simplify the prediction segment. Automated tools have made it possible to collect large volumes of medical data, which are made available to the medical research groups. This study aimed to apply machine learning algorithms using decision three classifier and descriptive statistics to evaluate the performance of the model in predicting the probability of cancer metastasis in patients that present late.

Materials and method: The breast cancer disease dataset has been taken from the department of Radiotherapy and Oncology of Usmanu Danfodiyo University Teaching Hospital, Sokoto state, Nigerian. Dataset has 259 instances and 10 attributes. The experimental results of this study used, decision three classifier in IMB SPSS (version 23) software environment. In the experiment, two classes were used and therefore a 2 × 2 confusion matrix was applied. Class 0=Not Metastasized, Class 1=Metastasized. We applied supervised machine learning approach in which dataset were divided into two classes that is training and testing using 10 fold cross validation.

Results: Shows that 259 instance of breast cancer, 218(84.2%) cases were not metastasized while 41(15.8%) cases were metastasized to the other region of the body. The overall accuracy of the model was found to be 87%, with the sensitivity of 88%, specificity 75% and the precision of 98%

Conclusion: Based on these findings, the machine learning algorism using decision three classifiers predicted that 87% of the tumor presented at stage IV, indicating that the tumour can spread to the other region of the body.

Keywords: Breast cancer; Machine learning; Prediction; Decision trees

INTRODUCTION

The global cancer burden is estimated to have risen to 18.1 million new cases and 9.6 million deaths in 2018. One in 5 men and one in 6 women worldwide develop cancer during their lifetime, and one in 8 men and one in 11 women die from the disease. Worldwide, the total number of people who are alive within 5 years of a cancer diagnosis is estimated to be 43.8 million [1].

According to international agency for research on cancer stated that female breast cancer was the leading type worldwide in terms of the number of new cases approximately 2.1 million diagnoses are estimated in 2018[1]. Breast cancer is a malignant disease that initiates in the breast cells. The patients with a family history of breast or ovarian cancer have possibility of developing breast cancer [2].

Some of the risk factors for breast cancer are gender (more in females), hereditary, genetic mutation, Smoking, alcohol consumption, obesity (As in sedentary life style), canned foods, chemicals carcinogens used as preservatives and in cosmetics [3-5].

Correspondence to: Abdulrahaman A. Musa, Department of Histopathology, Usmanu Danfodiyo University Teaching Hospital, Sokoto state, Nigeria, E-mail: abdulmusa910@gmail.com

Received: February 13, 2020; **Accepted:** February 28, 2020; **Published:** March 05, 2020

Citation: Musa AA, Aliyu UM (2020) Application of Machine Learning Techniques in Predicting of Breast Cancer Metastases Using Decision Tree Algorithm, in Sokoto Northwestern Nigeria. J Data Mining Genomics Proteomics. 11:220. DOI: 10.35248/2153-0602.20.11.220.

The high burden of breast cancer was attributed to the low or lack of cancer awareness among population as well as delay in cancer screening and detection. Increasing cancer cases in developing countries is also linked to the ageing population, and change in lifestyle such as unhealthy dietary practice and lack of physical activities [3-6].

Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data which is stored in databases or data warehouse using various data mining techniques such as, artificial intelligence (AI) and statistics [7]. Data mining techniques tends to simplify the prediction segment of the process. Automated tools have made it possible to collect large volumes of medical data, which are made available to the medical research groups [8]. Machine learning technique is a statistical tool that is used to predict the outcome of a disease base on the time of presentation [8].

Decision tree provides a powerful technique for classification and prediction in Breast Cancer diagnosis. Various decision tree algorithms are available to classify data, these include ID3, C4.5, C5, J48, CART and CHAID [8].

Several other similar studies were conducted to predict breast cancer outcome using various types of machine learning techniques, these include:

Artificial neuron network (ANN): It is a computational non-linear statistical data model based on the structure and functions of biological neural networks. It trains the neurons based on the experience and learning of input and output [9].

Artificial Neural networks (ANN) consists of an input layer, a hidden layer and an output layer. The input layer represents the elements of the dataset and the output layer consists of one node. Weights between the layers are adjusted using the training data using feed forward neural network and back propagation learning algorithm [10].

The Naive Bayes Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high [11,12].

This study aimed to apply machine learning algorithms using decision three classifier and descriptive statistics to evaluate the performance of the model in predicting the probability of cancer metastasis in patients that present late.

METHODOLOGY

The breast cancer disease dataset has been taken from the department of Radiotherapy and Oncology of Usmanu Danfodiyo University Teaching Hospital, Sokoto state, Nigerian. Dataset has 259 instances and 10 attributes. The experimental results of this study used, decision three classifier in IMB SPSS (version 23) software environment. In the experiment, two classes were used and therefore a 2 × 2 confusion matrix was applied. Class 0=Not Metastasized, Class 1=Metastasized. We applied supervised machine learning approach in which dataset were divided into two classes that is training and testing using 10 fold cross validation.

Evaluation of the model accuracy of decision three algorism

The performance of model was evaluated using four performance measures: accuracy, sensitivity, specificity, and precision. These measures are defined by four decisions: true positive (TP), true negative (TN), false positive (FN), and false negative (FN). TP decision occurs when tumour instance was not metastasized are predicted rightly. TN decisions when tumour instances were metastasized are predicted rightly. FP decision occurs when tumour instances are predicted as metastasized incorrect classification. FN decision occurs when tumour instances are predicted as not metastasized incorrect classification.

RESULTS

There were a total of 259 instances with 10 attributes recorded. The mean age of patients was 48.3 with SD ± 11 and the age range of 26 – 80 years respectively. The peak age of incidence occurred in those with age less than 50 year 144(55.6%), while age ≥ 50 year accounted for 115(44.4%) (Figure 1).

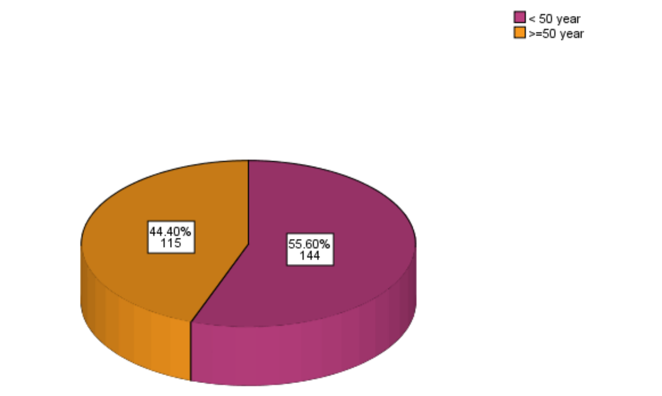

Figure 1: Age Distribution of breast cancer patients.

One hundred and eighty two (70.3%) of the study population were unemployed, 51(19.7%) were civil servants and 26(10%) were business women (Table 1).

Table 1: Demographical characteristics.

Occupation	Frequency	Percent
Civil Servant	51	19.7
Trader	26	10
Un-employed	182	70.3
Tribe		
Hausa	168	64.9
Yoruba	40	15.4
Igbo	40	15.4

Others	11	4.2
Religion		
Islam	177	68.3
Christian	82	31.7
Marital status		
Married	249	96.1
Divorce	10	3.9
Total	259	100

Majority 96.1% (249) of the breast cancer patients were married, only 10(3.9%) were divorce. One hundred and fifty one (58.3%) cases of breast cancer were on left breast, while 108(41.7%) occurred on right breast (Figure 2).

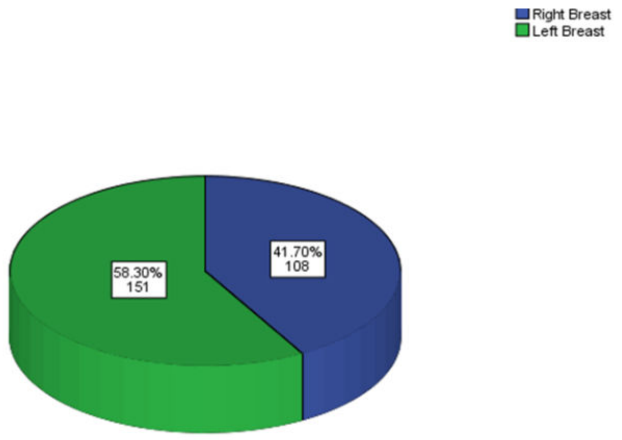

Figure 2: Classification of tumors by anatomical site.

Two forty one breast cancer patients (93.1%) had mastectomy as a form of surgery, others had lumpectomy 11(4.3%), quadractomy 7 (2.7%) (Figure 3).

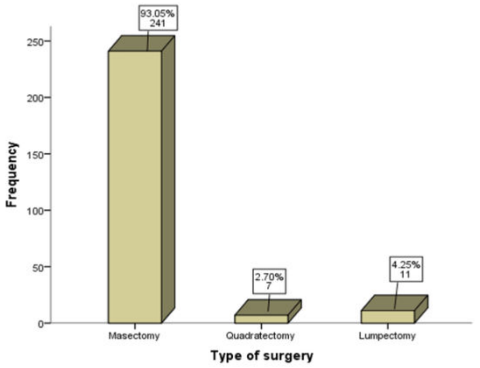

Figure 3: Classification of surgery done on breast cancer patients.

Almost all 258(99.6%) the patients presented with Invasive ductal carcinoma, while 1(0.4%) had Invasive lobular carcinoma (Figure 4).

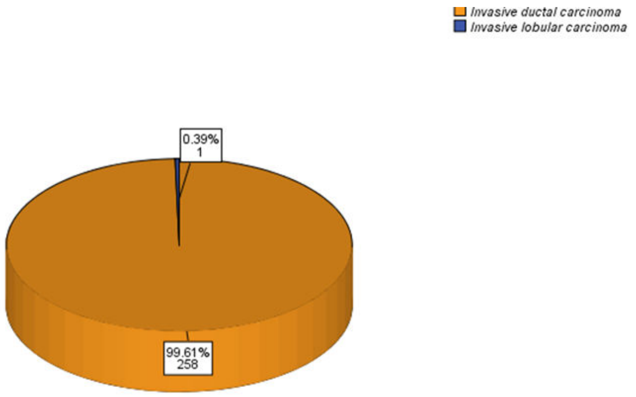

Figure 4: Classification of tumors by histological type.

Using supervised machine learning technique, decision three diagrams shows that 259 instance of breast cancer form the parent node of which 218(84.2%) cases were not metastasized while 41(15.8%) cases metastasized to the other parts of the body (Table 2).

Table 2: Confusion matric.

	Predicted		
Observed	**Not metastases**	**Metastases**	**Percent correct**
Not Metastases	214	4	98.20%
Metastases	29	12	29.30%
Overall Percentage	93.80%	6.20%	87.30%

The child nodes (leafs of the three) were classified in to three (3) nodes depicting the stages of the tumour presented to the oncology unit; these stages are [II, III, IV]. Late presentation (stage IV) was found to be significant ($p<0.05$)

The first child note shows that; out of 80 instances 79(98.8%) cases where did ' not metastasize while, 1(1.2%) case metastasized. The second child node shows that 163 cases presented at stage III, 135(82.8%) were did not metastasize, while 28(17.2%) cases metastasized, meanwhile the third child node shows that 16 cases presented at stages IV, 12(75.0%) metastasized, 4(25%) did not (Figure 5).

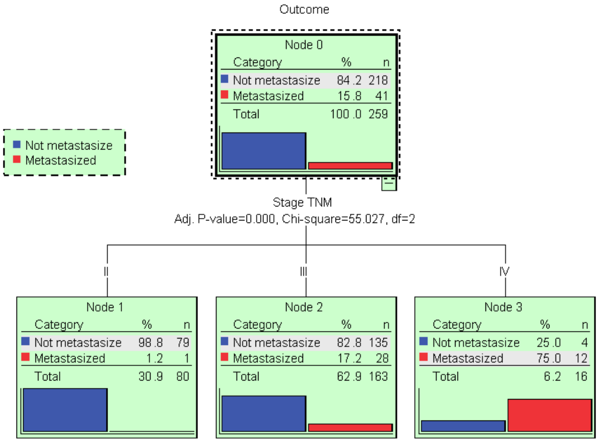

Figure 5: Decision three diagrams of breast cancer.

A value of overall accuracy, sensitivity, specificity, and precision was computed in below equation.

Accuracy can be calculated as:

$$\frac{TP+TN}{TP+TN+FP+FN}\frac{214+12}{214+12+4+29}=87.3\%$$

Sensitivity= $\frac{TP}{TP+FN}\frac{214}{214+29}=88\%$

Specificity= $\frac{TN}{TN+FP}\frac{12}{12+4}=75\%$

Precision= $\frac{TP}{TP+FP}\frac{214}{214+4}=98\%$

DISCUSSION

The mean age of population was 48 ± 11 year. This is in consonant with similar studies earlier conducted [13,14]. In this study the proportion of breast cancer was higher in those less than fifty years of age (55.6%). this corresponded to the finding from other scholars that reported similar result [13,15]. Metastatic breast cancer is associated with a severe burden both to the patient and to the healthcare delivery system [16]. In this study, 41(15.8%) cases metastasized to the other region of the body, and reason for this spread was, majority of these cases presented at late stages. The overall accuracy of the model in this study was found to be 87%. This agreed with a similar study in a Nigerian tertiary hospital [16]. There are many factors responsible for late presentation. In Africa, socioeconomic factors are noted to influence the choice and outcome of treatment among women with late stage breast cancer [17,18]. In our society, medical decision-making among these relatively young women is further compounded by the cultural influence of the husband and family members on their choice, acceptance and adherence to treatment options [19,20].

CONCLUSION

Automatic prediction the outcome of the disease is an important in real-world medical treatment. This paper shows that decision trees were used to model actual outcome of the tumour, the effectiveness of the model predicted that, 87% of tumour spread to the other part of the body presented at late stages. Thus detection of breast cancer in early stages is the key factors to be consider as well as public enlightenment on the consequences of late presentation.

REFERENCES

1. International Agency for Research on Cancer. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018.
2. Hartwell LH, Kastan M. Cell cycle control and cancer. Science. 1994;266(5192):1821-1823.
3. American Cancer Society. Cancer Facts & Figures 2018.
4. Samat N, Ghazali S, Atang C. Awareness and knowledge of cancer: A community survey in Kedah and Perlis. Asian Social Science. 2014;10(21):10-18.
5. WHO. Newsroom 2020.
6. Akram M, Iqbal M, Daniyal M, Khan AU. Awareness and current knowledge of breast cancer. BioMed Central. 2017;50(33):1-23.
7. Nathan N, Kabari L, Francis A. Prediction of breast cancer disease using decision tree algorithm. Int J Innovative Info Systems & Tech Res. 2019;7(1):34-38.
8. Sumbaly R, Vishnusri N, Jeyalatha S. Diagnosis of breast cancer using decision tree data mining technique. Int J of Computer Applications. 2014;98(10):0975-8887.
9. Samundeeswari ES, Saranya PK. An artificial neural network model for prediction of survival time of breast cancer dataset. Int J of Research in Engineering and Applied Sciences. 2016;6(1): 161-168.
10. Chi CL, Street WH, Wolberg WH. Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets. AMIA Annu Symp Proc. 2007: 130-134.
11. Abdelghani B, Guven E. Predicting Breast Cancer Survivability using Data Mining Techniques. Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining 2006.
12. Rathi M, Singh AK. Breast Cancer Prediction using Naïve Bayes Classifier. International Journal of Information Technology & Systems. 2012;1(2):2277-9825.
13. Quayson SE, Wiredu EK, Adjei DN, Anim JT. Breast cancer in Accra, Ghana. Journal of Medical and Biomedical Sciences. 2014;3(3):21-26.
14. Stark A, Kleer CG, Martin I, Awuah B, Nsiah-Asare A, Takyi V, Braman M, Quayson SE, Zarbo R, Wicha M, Newman L. African Ancestry and Higher Prevalence of Triple-Negative Breast Cancer: Findings from an International Study. Cancer. 2010; 116(21): 4926–4932.
15. Anim JT. Breast diseases: Review of surgical material in Korle-Bu Teaching Hospital, 1977-1978. Ghana Med J. 1997;18:30-33.
16. Adisa AO, Arowolo OA, Akinkuolie AA, Titiloye NA, Alatise OI, Lawal OO, et al. Metastatic breast cancer in a Nigerian tertiary hospital. African Health Sciences. 2011;11(2):279-284.
17. Bradley CJ, Given CW, Roberts C. Race, socioeconomic status, and breast cancer treatment and survival. J Natl Cancer Inst. 2002;94:490-496.
18. Adisa AO, Lawal OO, Adesunkanmi ARK. Paradox of wellness and nonadherence among Nigerian women on breast cancer chemotherapy. Journal of Cancer Research and Therapeutics. 2008; 4(3):107-110.

19. Nour A. Breast-conserving therapy in lowliteracy patients in a developing country. Breast Journal. 2003;9(2):71-73.

20. Ajekigbe AT. Fear of Mastectomy: the most common factor responsible for late presentation of carcinoma of the breast in Nigeria. Clinical Oncology. 1991;3:78-80.