ASIAN JOURNAL OF INFORMATION TECHNOLOGY
MEDWELL PUBLICATIONS

Predictive Model for Likelihood of Survival among Breast Cancer Patients using Machine Learning Techniques

Ajinaja Micheal Olalekan and Mobolaji Olarinde
Department of Computer Science, Federal Polytechnic, Ile-Oluji, Ondo State, Nigeria

Key words: Breast cancer, survival, Nigeria, predictive model, Naive Baye, machine learning

Corresponding Author:
Ajinaja Micheal Olalekan
Department of Computer Science, Federal Polytechnic, Ile-Oluji, Ondo State, Nigeria

Page No.: 263-269
Volume: 19, Issue 11, 2020
ISSN: 1682-3915
Asian Journal of Information Technology

Abstract: Providing a prediction model that can give survival rate of breast cancer patients among women based on past records collected over the years in an underdeveloped country like Nigeria poses a challenge. This is because of their poor data collection habit and underdeveloped health care system. Machine Learning (ML) offers a different approach and cheaper alternative of identifying survival rate among breast cancer patients among women. The purpose of this study is to provide survival rate or mortality rate of breast cancer patients after treatments has been administered. Naive Baye's Machine learning techniques was used in developing a predictive model to predict survival rate of breast cancer patients among women. Data was gathered from 30 different health centre location ranging from hospitals and institute. The data included all women who have been diagnosed with breast cancer from 2000-2005 and all death cases encountered so far. The simulation of the model was done using R Studio software. The result of the model was good as survival rate was above 85% showing incredible in the model used. Comparisons were made between some of the factors affecting breast cancer and survival rate using box plot. The results showed there is high survival rate in breast cancer patients among women in Nigeria. Other ML techniques can also be considered using same data to further improve the model.

INTRODUCTION

Cervical and breast cancers are the two most common cancers among women in Nigeria and other developing countries contributing significantly to a high morbidity and mortality rate in the country. Breast cancer is the most common cancer in women worldwide in Nigeria with population of about 187 million people and it represents about 12% of all new cancer cases and 25% of all cancers in women. According to Centre for disease control and prevention, breast cancer is a disease in which cellsin the breast grow out of control. It is a type of cancer that starts in the breast and gradually grows to and wreak havoc on its host. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumor that can often be seen on an x-ray or felt as a lump. Breast

cancer occurs almost entirely in women but men can get breast cancer too. Breast cancer can begin in different parts of the breast.

A breast is made up of three main parts: lobules, ducts and connective tissue. The lobules are the glands that produce milk. The ducts are tubes that carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together.

Most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized. Breast cancers has better prognosis if diagnosed and treated early. Ajekigbe[1] did some studies in breast cancer that suggested large number of women who got diagnosed early could beat breast cancer. Some of the reasons that were listed in the study include low literacy levels, high rates of poverty, cultural and religious traditions, poor geographical access to cancer care, low level of awareness of breast and cervical cancers, lack of screening and poor diagnostic procedure and treatment among health-care provider[2]. The effects of late presentation include complicated diagnosis and treatment, poor prognosis, increased risks of side effects from the use of second- or third-line therapies, huge costs of treatment, loss of productivity and increased mortality rates.

Most breast lumps are benign and not cancer (malignant). Non-cancerous breast tumors are abnormal growths but they do not spread outside of the breast. They are not life threatening but some types of benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a health care professional to determine if it is benign or malignant (cancer) and if it might affect the patient in future.

As stated by the Division of Cancer Prevention and Control, Centres for Disease Control and Prevention, there are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer.The most common kinds of breast cancer are Invasive ductal carcinoma in which the cancer cells grow outside the ducts into other parts of the breast tissue. Invasive cancer cells can also spread or metastasize to other parts of the body:

- Invasive lobular carcinoma. Cancer cells spread from the lobules to the breast tissues that are close by
- These invasive cancer cells can also spread to other parts of the body

The American Cancer Society stated that breast cancers can start from different parts of the breast. Most breast cancers begin in the ducts that carry milk to the nipple (ductal cancers). Some start in the glands that make breast milk (lobular cancers). There are also other types of breast cancer that are less common like phyllodes tumor and angiosarcoma. A small number of cancers start in other tissues in the breast. These cancers are called sarcomas and lymphomas and are not really thought of as breast cancers. Although, many types of breast cancer can cause a lump in the breast, not all do. Many breast cancers are also found on screening mammo grams which can detect cancers at an earlier stage, often before they can be felt and before symptoms develop.

Literature review: Elwood *et al*.[3] worked on the development and validation of a new predictive model for breast cancer survival in New Zealand and comparison to the Nottingham prognostic index. The team developed a model to predict 10-year breast cancer-specific survival using data collected prospectively in the largest population-based regional breast cancer registry in NZ (Auckland, 9182 patients) and assessed its performance in this data set (internal validation) and in an independent NZ population-based series of 2625 patients in Waikato (external validation). The data included all women with primary invasive breast cancer diagnosed from 1 June 2000-30 June, 2014 with follow up to death or Dec 31, 2014. Multivariate Coxproportional hazards regression to assess predictors and to calculate predicted 10-year breast cancer mortality was used and therefore survival, probability for each patient. The team also assessed observed survival by the Kaplan Meier method. Elwood and his team also assessed discrimination by the C statistic and calibration by comparing predicted and observed survival rates for patients in 10 groups ordered by predicted 10 years survival. The team compared the NZ model with the Nottingham Prognostic Index (NPI) in this validation data set. The data collected prospectively through the two largest and longest-established population-based regional breast cancer registries in NZ, in the Auckland and Waikato regions. These two regional registries are linked to include over 40% of all patients with breast cancer in NZ and are representative of NZ women in terms of socioeconomic, demographic and ethnic background. The result of the team work stated that for the 9182 eligible women in the Auckland database, there were 864 breast cancer specific deaths over the 14-year time period; median follow up time was 67.6 months and mean age of patients 56.9 years. Patients were predominantly Stage 1(43%) and 2 (39%), ER and PR positive (79 and 68%), HER-2 negative (69%) without lymphovas cularinvasion (73%),and with ductal tumours (81%). Of the patients, 71% were of NZ European ethnic group with 8% Maori, 7% Pacific and 14% other(such as

Asian countries). The risk of breast cancer mortality within 10 years of diagnosis increased significantly with age being over 70 years; higher tumour grade, larger tumour size, greater number of positive lymph nodes, presence ofmetastases at diagnosis and with ER or PR negative tumours. Also discrimination was good, the C statistics were 0.84 for internal validity and 0.83 for an independent external validity. For calibration, for both internal and external validity the predicted 10-year survival probabilities in all groups of patients, ordered by predicted survival were within the 95% Confidence Intervals (CI) of the observed Kaplan-Meier survival probabilities. The NZ Model showed good discrimination even within the prognostic groupsdefined by the NPI.

Stark[4] worked on predicting breast cancer risk using personalhealth data and machine learning models. The team developed machine learning models that used highly accessible personal health data to predict 5 year breast cancer risk. They created machine learning models using only the Gail model inputs and models using bothGail model inputs and additional personal health data relevant to breast cancer risk. Forboth sets of inputs, six machine learning models were trained and evaluated on the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial data set. The machine techniques used in the research work includes the logistic regressions, Naive Bayes, Decision trees, linear discriminant and Support vector machines.Models were trained and evaluated on the PLCO data set. This data set was generated aspart of a randomized, controlled, prospective study that sought to determine the effectivenessof different prostate, lung, colorectal and ovarian cancer screenings. From November, 1993-July, 2001, participants enrolled inthe study and filled out a baseline questionnaire detailing their previous and current healthconditions. All processing of this data set was completed in Python (Version 3.6.7). This data set consists of 78.215 women ages 50-78. The team chose to exclude women who met any of the following conditions: were missing data regarding whether they had been diagnosed with breastcancer and/or the timing of the diagnosis were diagnosed with breast cancer before thebaseline questionnaire did not self-identify as White, Black or Hispanic identified as Hispanic but did not have information available about where they were born or were missing data for any of the thirteen selected predictors.The team excluded women who were diagnosedwith breast cancer before the baseline questionnaire. The team implemented the logistic regression, naive Bayes, decision tree, support vector machine and linear discriminant analysis models using the Python scikit-learn package (Version 0.20.1). For both sets of inputs, the selected neural network hyper parameters were those that produced the highest mean minus standard deviation AUC across 10 iterations of neural networks trained using 10-fold cross-validation on the training data set. For neural networks with only BCRAT[5] inputs, these hyperparameters were 6 neurons per hidden layer, 2 hidden layers, an 0.01 learning rate and 5000 steps of backpropagation. For the neural networks with abroader set of inputs, the hyperparameters were 12 neurons per hidden layer, 2 hidden layers, a 0.01 learning rate and 2500 steps of backpropagation. The logistic regression and linear discriminant analysis models tied for the highest testing data set AUC (0.613). The logistic regression was significantly stronger than the decision tree, support vector machine and naive Bayes models but not stronger than the lineardiscriminant analysis or the neural network models. Similarly, the linear discriminant analysiswas significantly stronger than the decision tree, support vector machine and naive Bayes models but not stronger than the logistic regression or neural network models. The logistic regression, linear discriminant analysis and neural network models had very different values for sensitivity. The logistic regression had a low sensitivity of 0.476 whereas the linear discriminant analysis had a sensitivity of 0.688. The neural network sensitivity lay between these two values at 0.599. For specificity, the logistic regression had the highest value at 0.691, followed by the neural network (0.562) and the linear discriminant analysis (0.467). The precisions for all three of these models were low. The logistic regression had a precisionof 0.0323, slightly higher than the neural network (0.0287) and linear discriminant analysis (0.0272) precisions. Again, we saw that the machine learning model sensitivities, specificities and precisions were generally comparable to those of the BCRAT[5].

Huang *et al*. worked on the development of a prediction model forbreast cancer based on the national cancerregistry in Taiwan. The study aimed to develop a prognostic model to predict the breast cancer-specific survival andoverall survival for breast cancer patients in Asia and to demonstrate a significant difference in clinical outcomesbetween Asian and non-Asian patients. The team developed our prognostic models by applying a multivariate Cox proportional hazards model toTaiwan Cancer Registry (TCR) data. A data-splitting strategy was used for internal validation and a multivariablefractional polynomial approach was adopted for prognostic continuous variables. Subjects who were Asian, blackor white in the US-based Surveillance, Epidemiology and End Results (SEER) database were analyzed for externalvalidation. Model discrimination and calibration were evaluated in both internal and external datasets.In the internal validation, both training data and testing data calibrated well and generated good area under the ROC

curves (AUC; 0.865 in training data and 0.846 in testing data). In the external validation, although the AUCvalues were >0.85 in all populations, a lack of model calibration in non-Asian groups revealed that racialdifferences had a significant impact on the prediction of breast cancer mortality. For the calibration of breast cancer specific mortality, $p<0.001$ at 1 year and 0.018 at 4 years in whites and $p\leq0.001$ at 1 and 2 years and 0.032 at 3 years in blacks, indicated that there were significant differences ($p<0.05$) between the predicted mortality and the observed mortality. Our model generally underestimated the mortality of the black population. In the white population, the model underestimated mortality at 1 year and overestimated it at 4 years. And in the Asian population, all $p>0.05$ indicating predicted mortality and actual mortality at 1-4 years were consistent. The team developed and validated a pioneering prognostic model that especially benefits breast cancer patients in Asia. The study could serve as an important reference for breast cancer prediction in the future.

MATERIALS AND METHODS

Data source and sample selection: The original data for our primary analysis were retrieved from different medical centres in Nigeria who specializes in breast cancer diagnosis and treatment. The data were collected from Lagos state University teaching hospital Ikeja, Crystal Hospital Akowonjo, OrileAgege Government Hospital Agege, Onikan Health Centre, General HospitalLagos, General HospitalIsolo, General Hospital Ajerom, Awadiora Health Centre, Oluwaseun Medical Centre, Harvey Road Health Cent, Mike Medics Hospital, Longing Medical Centre, Lag Path Consulting, General Hospital Badagry, General Hospital Surulere, Lagos State Government Alausa, General Hospital Ikorodu, Merit HospitalGeneral Hospital Alimos, Olatunwa Clinic, Alagba General Hospital Epe, Osuntuyi Medical Centre, EKO Hospital, Island MaternityLagos, General Hosp. Randle, General HospitalGbagada, General Hospital Ibeju, NAF Base Ikeja and Cancer Institute of Nigeria.. The data included all women who have been diagnosed with breast cancer from 2000-2005 and all death cases encountered so far. It consisted of 927 records of different women who had been diagnosed with breast cancer. From the dataset, 11 attributes were used for the analysis which are represented as: Status of patient (Stat). Position of the tumour (Top). Marital Status (maritstatus). Types of Malignant tumour (Mor). Religion investigative method carried out on breast (Bas) tumor difference after examination (tumourdiff) Symptoms diagnosis of the breast (tdiag) Method of treatment of the patient (treatmt) (Table 1-9):

Table 1: Number 1 on the data spread represent patient who are alive while 2 represent dead patient after diagnosis and treatment

Parameters	Values
Alive	1
Dead	0

Table 2: Represent the position of the tumour in the breast where it was diagnosed

Parameters	Values
C50.5 Lower-outer quadrant of breast	505
C50.6 Axillary tail of breast	506
C50.9 Breast, NOS	509
C50.2 Upper-inner quadrant of breast	502
C50.0 Nipple	500
C50.1 Central portion of breast	501
C50.4 Upper-outer quadrant of breast	504
C50.8 Overly lesion of breast	508
C50.3 Lower-inner quadrant of breast	503

Table 3: Represent the marital status of the patient

Martial status	Values
Married	1
Single	2
Divorced	3

Table 4: Represent the types of malignant tumour in the breast

Parameters	Values
Cystadenocarcinoma, NOS	8440
Pleomorphic carcinoma	8022
Infiltrating duct carcinoma, NOS	8500
Intracystic carcinoma, NOS	8504
Infiltrating ductular carcinoma	8521
Carcinoma in pleomorphic adenoma	8941
Carcinoma, NOS	8010
Adenocarcinoma, NOS	8140
Duct carcinoma, desmoplastic type	8514
Malignant tumor, giant cell type	8003
Secretory carcinoma of breast	8502
Atypical medullary carcinoma	8513
Lobular carcinoma, NOS	8520
Comedocarcinoma, NOS	8501
Adenoid cystic carcinoma	8200
Infiltrating duct mixed with other types of car	8523
Infiltrating duct and lobular carcinoma	8522
Lipid-rich carcinoma	8314
Inflammatory carcinoma	8530
Paget disease and intraductal carcinoma of breast	8543
Cystic hypersecretory carcinoma	8508
Malignant tumor, small cell type	8002
Lymphoepithelial carcinoma	8082
Fibrosarcoma, NOS	8810
Infiltrating lobular mixed with other types of	8524
Fibromyxosarcoma	8811
Papillary carcinoma, NOS	8050
Mucinous adenocarcinoma	8480
Neoplasm, malignant	8000
Phyllodestumor, malignant	9020
Intraductal papillary adenocarcinoma with invas	8503
Medullary carcinoma, NOS	8510
Squamous cell carcinoma, large cell, nonkeratin	8072
Granular cell carcinoma	8320
Myxosarcoma	8840
Spindle cell sarcoma	8801
Squamous cell carcinoma, small cell, nonkeratin	8073
Metaplastic carcinoma, NOS	8575
Clear cell adenocarcinoma, NOS	8310
Squamous cell carcinoma, keratinizing, NOS	8071
Tubular adenocarcinoma	8211

Table 4: Continue

Parameters	Values
Desmoplastic small round cell tumor	8806
Myosarcoma	8895
Trabecular adenocarcinoma	8190
Liposarcoma, NOS	8850
Cribriform carcinoma, NOS	8201
Fibroblastic liposarcoma	8857
Paget disease, mammary	8540
Giant cell carcinoma	8031
Papillary adenocarcinoma, NOS	8260
Squamous cell carcinoma, NOS	8070
Leiomyosarcoma, NOS	8890
Mast cell sarcoma	9740
Granular cell tumor, malignant	9580
Epithelial-myoepithelial carcinoma	8562
Glycogen-rich carcinoma	8315
Malignant tumor, clear cell type	8005
Pseudosarcomatous carcinoma	8033
Acinar cell carcinoma	8550
Adenocarcinoma with apocrine metaplasia	8573
Stromal sarcoma, NOS	8935
Composite Hodgkin and non-Hodgkin lymphoma	9596
Tumor cells, malignant	8001
Mesenchymoma, malignant	8990
Malignant tumor, spindle cell type	8004
Hodgkin lymphoma, mixed cellularity, NOS	9652
Paget disease and infiltrating duct carcinoma	8541
Squamous cell carcinoma, micro invasive	8076

Table 5: Religion of the patient

Religion	Values
Muslim	1
Christianity	2

Table 6: Investigative method used in diagnosing breast cancer in patient

Parameters	Values
Histology of primary	7
Unknown	9
Cytology	5
Histology of metastases	6
Clinical only	1
2Clinical Invest/Ultra Sound	

Table 7: Tumor difference after examination (tumourdiff)

Parameters	Values
Well	1
Moderate	2
Poor	3
Undifferentiated	4

Table 8: Method of symptom diagnosis of the breast

Parameters	Values
Symptomatic	1
Screening	3
Unknown	4
Asymptomatic	2

Table 9: Describes the method of treatment applied to treat breast cancer of patient

Parameters	Values
Surgery	1
Chemotherapy	3
None	0
Radiotherapy	2
Palliative care	4
Others	5

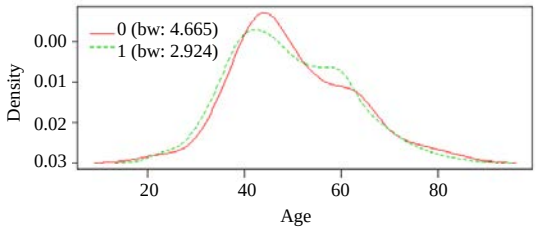

Fig. 1: Density plot of age attribute in the dataset

- Age (age): general age of each of the patient
- Address (Addr): home address of each patient

Each of the attributes were represented in numeral forms to enable pre-processing on R-studio (Fig. 1).

RESULTS AND DISCUSSION

The R-studio machine learning toolkit Version 1.1.456 was used for analysis on breast cancer patients. The confusion matrix is at able that presents the number of correctly and incorrectly classified instances of the actual and predicted class instances on the data set distribution. Also, the dataset was partitioned into two training and testing. About 193 observables were used for testing while 734 observables were used for training the model. The density plot is a representation of the distribution of a numeric variable which shows the probability density function of the variable.

Figure 1 present the output of the numeric variable distribution of age attribute in the dataset showing peak at 40 year olds. This indicate most patient were in their late 30 sec and the rest been distributed in the plot.

Figure 2 present the output plot showing the chances of a patient having the tumour of the cancer at the left outer quadrant of the breast (505) has a 100% chance of survival or been alive after treatment. Also, those with the tumour positioned at the central of the breast (501) and at the nipple (500) had a lower chance of survival after treatment.

Figure 3 shows the output plot of the investigative method carried out on the breast (Bas). Women who were investigated using the histology of primary mode had a better chance of survival while Histology of metastases mode had the highest chance of death in patients.

Figure 4 shows the output distribution of Tumour difference after examination (tumourdiff). There is high chance of a patient surviving if the tumour difference is poor and low chance if it cannot be categorized.

Figure 5 shows the distribution output of Symptoms diagnosis of the breast (tdiag) whether it was symptomatic, screening and asymptomatic or an unknown

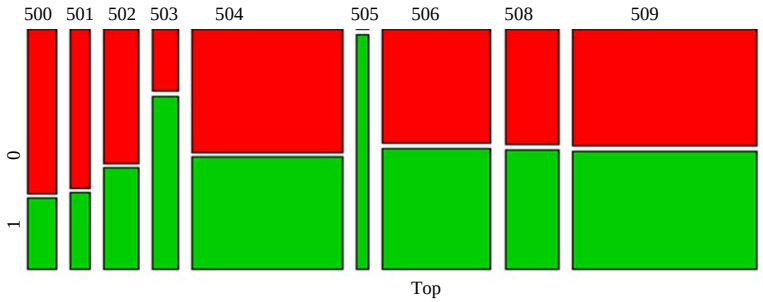

Fig. 2: Output plot distribution showing chances of survival based on the position of tumour in the breast

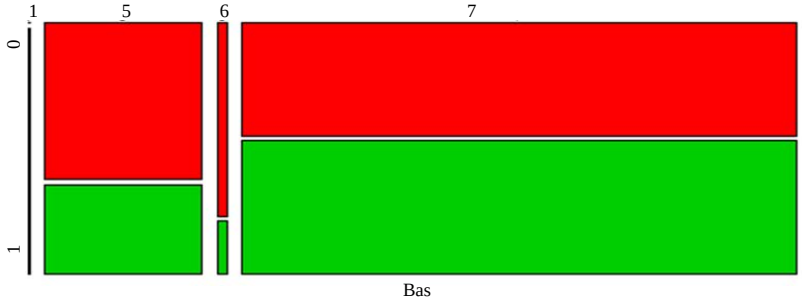

Fig. 3: Output plot distribution showing chances of survival based on the investigative method carried out on the breast

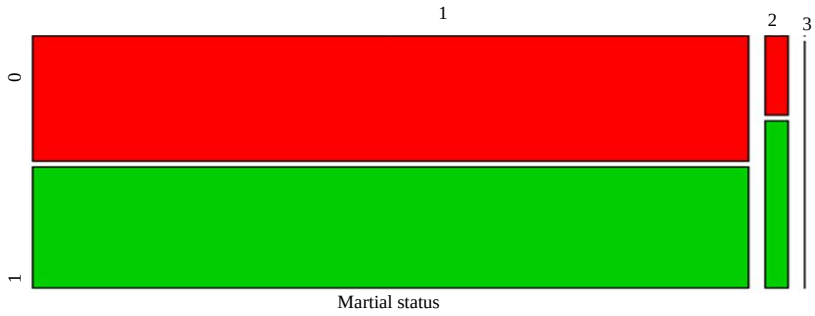

Fig. 4: Output plot of marriage status of the dataset

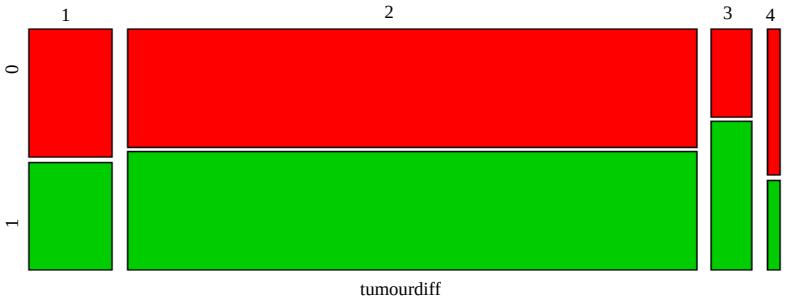

Fig. 5: Output distribution of tumour difference among patients

symptoms diagnosis. Patient who showed symptomatic symptoms had a worst chance of survival most. Figure 6 shows the method of treatment of the patient (treatmt). Patient who were treated using radiotherapy had a better chance of surviving. Table 2 shows the confusion matrix for both the trained and test data and the accuracy of the algorithm when implemented on both dataset (Table 10).

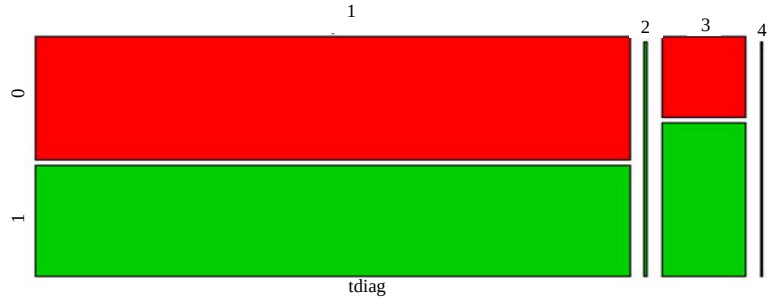

Fig. 6: Output distribution of symptoms diagnosis of the breast (tdiag)

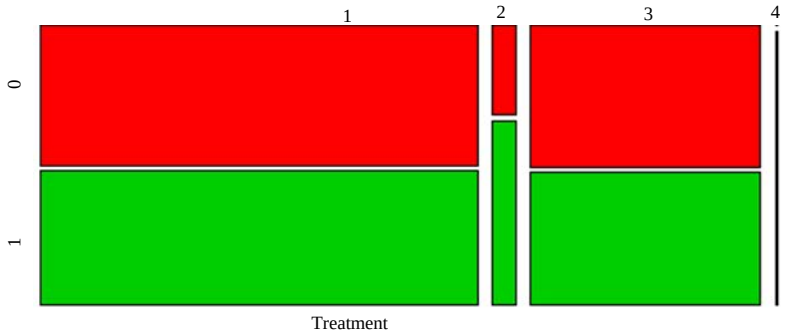

Fig. 7: Method of treatment

Table 10: Confusion matrix of the trained and test data

Matrix	----------Data----------		
Accuracy	Trained data		
99.9386921	0	1	
	0	2	0
	1	45	687
Accuracy	Test data		
99.9171	0	1	
	0	1	5
	1	11	176

It was observed from this study that the Naive Bayes have good performance accuracy on the data sets. The result will help to identify patients that are susceptible or chances of survival of breast cancer patients, so that, appropriate measures can be taken early enough to improve survival rate.

CONCLUSION

This research was able to study trends in breast cancer patients among women and further build models that can be used for prediction of survival rate of breast cancer patients using Naïve Bayes machine learning techniques. The result gave a high accuracy after modelling and suitable for mining data.

REFERENCES

01. Ajekigbe, A.T., 1991. Fear of mastectomy: The most common factor responsible for late presentation of carcinoma of the breast in Nigeria. Clin. Oncol. (R. Coll. Radiol. (Great Britain)), 3: 78-80.
02. Wang, F., S. McLafferty, V. Escamilla and L. Luo, 2008. Late-stage breast cancer diagnosis and health care access in Illinois. Prof. Geogr., 60: 54-69.
03. Elwood, J.M., E. Tawfiq, S. TinTin, R.J. Marshall and T.M. Phung *et al*., 2018. Development band validation of a new predictive model for breast cancer survival in New Zealand and comparison to the Nottingham prognostic index. BMC Can., Vol. 18, 10.1186/s12885-018-4791-x.
04. Stark, G.F., G.R. Hart, B.J. Nartowt and J. Deng, 2019. Predicting breast cancer risk using personal health data and machine learning models. PLoS One, Vol. 14, No. 12.
05. National Cancer Institute, 2019. Breast cancer treatment-patient version. National Cancer Institute, Bethesda, Maryland.