Proximate Breast Cancer Factors Using Data Mining Classification Techniques

Alice Constance Mensah, Accra Technical University, Accra, Ghana
Isaac Ofori Asare, Vita Verde Consult, Accra, Ghana

ABSTRACT
Breast cancer is the most common of all cancers and is the leading cause of cancer deaths in women worldwide. The classification of breast cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors. Data mining technology helps in classifying cancer patients and this technique helps to identify potential cancer patients by simply analyzing the data. This study examines the determinant factors of breast cancer and measures the breast cancer patient data to build a useful classification model using a data mining approach. In this study of 2397 women, 1022 (42.64%) were diagnosed with breast cancer. Among the four main learning techniques such as: Random Forest, Naive Bayes, Classification and Regression Model (CART), and Boosted Tree model were used for the study. The Random Forest technique had the better accuracy value of 0.9892(95%CI,0.9832 -0.9935) and a sensitivity value of about 92%. This means that the Random Forest learning model is the best model to classify and predict breast cancer based on associated factors.

KEYWORDS
Boosted Tree Model, Breast Cancer, Classification and Regression Model, Naive Bayes, Random Forest

INTRODUCTION
Breast cancer is a malignant tumor which develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. Breast cancer is the most common cause of death among women with cancer (522,000 deaths in 2012), the incidence rate stood at approximately 17 percent and type of cancer most attacked women in 140 of 184 countries in the world (Ferley et al. 2013). Breast cancer is on the increase in Ghana. Bray et al (2018) report also indicates that a total of 13,807 cancer cases were recorded in females of all ages during the research period in Ghana. Breast Cancer led with 4,645 (33.6 percent), Cervix Uteri Cancer followed with 3,151 (22.8 percent), then Ovary Cancer with 861 (6.2 percent), Liver Cancer was fourth with 737 (5.3 percent) and Colorectum Cancer had 570 cases (4.1 percent). The rest of the other cancer cases recorded 3,843 which constituted 27.8 percent. The incidence rate of breast cancer in women currently stands at 43 percent whiles the mortality rate is 17.7 percent. The report estimates that over 4600 new cases of breast cancer will be diagnosed in Ghana this year and that more than 1,800 women will lose their lives to this cancer. Scientists do not know the exact causes of most breast cancer, however there are some known risk factors that increase the likelihood of a woman developing breast cancer. These factors contain such attributes as age, family history and genetic risk.
LITERATURE REVIEW

Breast cancer occurrence is increasing globally and one of the major causes of death in women compared to all other cancers. Chaurasia and Pal (2017) breast cancer is a major health problem and represents a significant worry for many women. To reduce life losses, detecting breast cancer early is very essential and it calls for accurate and reliable diagnosis procedure. One of the major problems in medical applications is medical diagnosis Liou and Chang (2015). The application of machine learning methods is widely used nowadays in medical diagnosis for prediction. One of the most interesting and challenging tasks is to develop data mining applications in the prediction of an outcome of a disease. Saleema et al. (2014) posited that, the production and availability of large volumes of the medical data by the medical research groups has resulted in making data mining techniques a popular research tool. This tool is used to identify and exploit patterns and relationships among large number of variables and also to predict an outcome of a disease using the historical datasets.

Various studies have been done on the application of data mining techniques in diagnosing breast cancer. One of such studies was by (Bellaachia & Guven, 2006), who reported that C4.5 algorithm, gave the best performance of 86.7% accuracy having used the SEER data to compare three prediction models for detecting breast cancer. The use of the genetic algorithm model on the data of breast cancer patients explored by (Chang & Liou, 2008) yielded a better result than other data mining models for the analysis of the overall accuracy of the patient classification, expression and complexity of the classification rule. Investigation by (Abdelaal et al., 2010) revealed that SVM techniques show a promising result for increasing diagnostic accuracy of classifying the cases witnessed by the largest area under the ROC curve comparable to values for tree boost and tree forest. The approach by (Christobel & Sivaprakasam, 2011) decision tree classifier (CART) for breast cancer diagnosis, attained an accuracy of 69.23%. Comparing the classification accuracy of Support Vector Machine (SVM), IBK, BF Tree algorithms, the SVM had the best accuracy (Lavanya and Rani, 2012). Asri et al. (2016) applied the performance of Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (kNN) on the Wisconsin Breast Cancer datasets. The results indicated that SVM had the best performance in term of accuracy (97.13%). A study on breast cancer comparing data mining techniques for breast cancer shows that C4.5 is the best classification technique for breast cancer as it had an accuracy rate of 86.70% (Zand, 2015). Shajahan et al. (2013) in their study shown that, Random Tree was the best data mining technique to classify and predict breast cancer with an accuracy rate of 100%.

METHODOLOGY

The study was carried out at the Korle Bu Breast Clinic and the National Centre for Radiotherapy and Nuclear Medicine, both located in the Korle Bu Teaching Hospital (KBTH). KBTH is the leading national referral Centre in Ghana receiving patients from across the country, but mostly from the southern part. The Radiotherapy Centre serves as the cancer Centre for the hospital. The Breast clinic, run by a multidisciplinary team, receives referrals but is a walk-in clinic that admits women who desire to be screened for breast cancer without a formal referral. This study used Two thousand three hundred and ninety-seven (2397) women.

INCLUSION AND EXCLUSION CRITERIA

All Ghanaian women who visited the National Centre for Radiotherapy and Nuclear Medicine and the walk-in clinic for breast screening were eligible. Cases were required to have histologically proven breast cancer. Patients with incomplete information, other malignancies (e.g. sarcomas) and aged less than twenty (20) years were excluded.
The Role of Learning Analytics in Global Higher Education
www.igi-global.com/chapter/the-role-of-learning-analytics-in-global-higher-education/147048?camid=4v1a

Thematic Analysis in Qualitative Research
Anindita Majumdar (2019). Qualitative Techniques for Workplace Data Analysis (pp. 197-220).
www.igi-global.com/chapter/thematic-analysis-in-qualitative-research/207798?camid=4v1a