International Journal of Big Data and Analytics in Healthcare
Volume 4 • Issue 1 • January-June 2019

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.

DOI: 10.4018/IJBDAH.2019010104

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 el 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 el 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 el 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). Shajahaan 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.

STATISTICAL DATA MINING TECHNIQUES

There are several techniques or tools that are used to model either by prediction or for classification purposes. These tools are used in building accurate and efficient classifiers for large dataset. This is the work of data mining and machine learning techniques. Data mining techniques are used for classification and to increase understanding of the domain or to improve predictions compared to unclassified data. Building effective classification systems is one of the central tasks of data mining (Chaurasia & Pal, 2017). Many different types of classification techniques have been proposed in literature that includes Decision Trees, Naive- Bayesian methods, Sequential Minimal Optimization (SMO), IBK, BF Tree, random forest tree, naïve bayes model, classification and regression tree (CART) and boosted tree. Some of the techniques are classified as traditional techniques. These traditional techniques include; Chi-Square test, Binary logistic, Analysis of Variance (ANOVA) among others. Due to the advancement of statistical modeling for accuracy purposes, data mining techniques have been development to enhance the work of the traditional models. Though they are known to be traditional models, yet they are still efficient in its usage. Data mining is regarded as an emerging technology that has made radical change in the information world. The term `data mining' (often called as knowledge discovery) refers to the method of analyzing data from different perspectives and summarizing it into valuable information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system. Technically, "data mining is the method of finding correlations or patterns among dozens of fields in large relational databases". Therefore, data mining consists of key functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyze data using application tools and techniques, and meaningfully present data to provide useful information (Goldschmidt, 2006; Bace, 2000).

The data mining consists of methods serving various purposes. Classification and clustering are the two most common techniques of data mining which are used in field of medical science. However, most data mining methods commonly used are of classification category as the applied prediction techniques assign patients to either a" benign" group that is non- cancerous or a" malignant" group that is cancerous and generate rules for the same. The aim of classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The data mining techniques were developed to provide accurate and precise results for decision making. This study used learning techniques such Random Forest Tree, Classification and Regression Technique (CART), Naive Bayes, and Boost techniques. Studies on Breast cancer analysis using data mining techniques have gain much attention in the developed countries. However, in developed countries (Ghana), there have been limited or no number of studies on the usage of data mining techniques in prediction and classifications of breast cancer. This motivated the researchers to embark on this study.

RANDOM FOREST TREE

The Random Tree is a kind of supervised Classifier developed by (Breiman et al., 2009). It is an ensemble learning algorithm which can generate many individual learners. It employs the principles of bagging to produce a random set of data for constructing of the decision tree. In standard tree each node is split using the best split among all variables. In a random forest, each node is split using the best among the subset of predicators randomly chosen at that node. Random Forest can take a different subset (sample) of data with replacement and it can even sample the features, that means it performs Data Sampling (Observations) as well as Feature sampling (Variables). Finally, the decision is taken by majority voting.

NAÏVE BAYES MODEL

This technique is a machine learning algorithm for classification purposes. It is a classification technique that was designed to classify high-dimensional training dataset. It is a probabilistic classifier. it is naïve because, it makes certain assumptions that has occurrence of a certain features independent of the occurrence of the other features. The Bayes model is based on the Bayes' Theorem. The Bayes' Theorem is stated as; $P(A \setminus B) = \frac{P(B \setminus A)P(A)}{P(B)}$. The aim of the naïve Bayes learning model is to calculate the conditional probability of an object which as the vector $y_1 y_2 \ldots y_n$ belonging to a specific class say $c_1 c_2 c_3 \ldots c_n$.

CLASSIFICATION AND REGRESSION TREE (CART)

The CART is a data mining technique suggested by (Breiman,2017). The decisions tress that is generated by the classification algorithms are mainly binary. The technique works by partitioning the dataset in to training and testing. The CART algorithm grows the tree by conducting for each decision node, an exhaustive search of all available variables and all possible splitting values, selecting the optimal split according to the following criteria as indicated by (Kennedy,1997). Below is the mathematics behind the learning technique. Let $Y_{1.} \ldots Y_m$, P be random variables where Y_1 has domain Dom (Y_i). The random variable P has domain Dom (P) = {1, . . ., J}. $Y_{1.} \ldots Y_m$ is known as the predictor attributes, m becomes the number of predictor attributes and C becomes the class label. A classifier C is a function C: Dom (Y_1 x…x Dom (Y_m) $\rightarrow$ Dom (C). Let Ω = Dom (Y_1)x…x Dom (Y_m) $\times$ Dom (C) representing set of events. For a given classifier C and a given probability measure P over Ω we can introduce a functional $R_p(C) = P\left[C\left(X_{1,\ldots\ldots,} X_n\right) \neq C\right]$ called the misclassification rate of the classifier C.

BOOSTED TREE

Boosting is one of the learning models that is used for classification it is a technique which is used for improving the performance of any learning algorithm. It is used significantly reduce the error of any "weak" learning algorithm that consistently generates better classifiers as compared with random guessing. Freund and Schapire (1996) defined Boosting as a learning model that is used to fit many large or small trees to reweighted training dataset which is classified by weighted majority vote. The boosting works by maintaining the weights of bootstrap samples rather than drawing independent sample. The higher the weight it draws the better the classifier. this is done to enhance the performance of the classifier in order to increase the misclassification weights. This is how the technique works; Random dataset is selected from the original dataset without replacement to obtain A1 and train the weak learner. N samples from A1 with half of the samples misclassified by A1 to obtain B1 and train weak learner B2. Select samples from A1, BI and B2 and those who disagree train the weaker learner to form B3. Final classifier is vote of weak learners. If the bootstrap replicated approximations were correct, then bagging would reduce variance without changing bias.

TRAINING AND TESTING OF THE DATASET

The data has one dependent variable-DV (diagnose of cancer or not) and the other nine independent variables-IDVs (associated variables of breast cancer) for the learning models. The IDVs are as follows;

Current age of patient, Age of first child, Age of menarche, BMI, Usage of contraceptive, Family history, Alcohol intake, Parity. The dataset was divided into training and testing dataset. The training dataset has 70% of the dataset and the remining 30% used for the testing and evaluating of the model.

MODEL EVALUATION

The confusion matrix was used to evaluate the learning models' performance. It compares the decisions made by the model with the actual decisions. It provides accuracy level of the model and determine how the model will perform on new, previously and unseen dataset. The accuracy measures of the model are shown in the Table 1 below;

RESULTS AND DISCUSSION

The results in the Figure 1 shows the correlation analysis between the input variables and the output variable. It shows the association that exist among the factors. The results show that, there is a strong correlation approximately $\left(r_x = 0.740\right)$ between the current age and age of menopause of the women. However, from the Figure 1, it could be observed that, there is statistically significant association $\left(r_x = 0.280, p < 0.05\right)$ between Cancer status and Age of menopause among women. Also, there was significant association between Cancer status and current age of the patients $\left(r_x = 0.20, p < 0.05\right)$. In the determination of the risk factors associated with breast cancer among women, the Chi-Square $\left(\chi^2\right)$ for independence was used to ascertain this and the test results in the Table 2 shows that, eleven (11) variables were statistically significant. However, family history had no influence on breast cancer status among women. Ford et al. (1995); Sasco (2003) indicated that the amount of breast cancer associated factors attributed hereditary is quite low as compared to social and environmental factors.

RESULTS OF THE LEARNING MODELS TO THE BREAST CANCER DATA

Studies have shown that, learning models (Data mining techniques) have been used in the diagnosis of breast cancer among women. However, this study is adopting these learning techniques to classify and predict the cancer status among the patients(women). The classification and the prediction are based on the associated breast cancer factors. The data was analysed using four (4) main learning techniques such as; Random Forest, Naive Bayes, Classification and Regression Model (CART) and Boosted Tree model. Each learning model results would be compared with one another and the best model with the highest accuracy index and high sensitivity value would be the appropriate technique that, could be used to model the breast cancer dataset gathered for the study. The results in the table 3 shows the confusion matrix and the diagnostics statistics of each of the learning model used for

Table 1. Model adequacy measure

Confusion Matrix		**Positive**	**Negative**		
	Positive	**I**	**II**	**Positive Predictive Value**	**I/I+II**
	Negative	**III**	**IV**	**Negative Predictive Value**	**III/III+IV**
		Sensitivity	Specificity	Accuracy = I+IV/ (I+II+III+IV)	
		I/I+III	II/II+IV		

Table 2. Test of association of breast cancer associated factors

Variable	χ^2_{df}	$p-value$
Body Mass Index (BMI)	554.329(31)	0.000
Age	187.179(74)	0.000
Age at Menarche	28.930(14)	0.001
Age at First Child	86.276(32)	0.000
Age at Menopause	284.074(30)	0.000
Parity	307.954(12)	0.000
Occupation	7.363(1)	0.007
Marital Status	22.046(4),	0.000
Contraceptive	20.320(1)	0.000
Alcohol	11.238(1)	0.001
Breast Feeding	136.456(1)	0.000
Family History	0.296(1)	0.586

Note: p-value < 0.05 indicates significance at 5%, values in bracket are degree of freedom

Figure 1. Correlation analysis, deep colour in the figure indicates high correlation

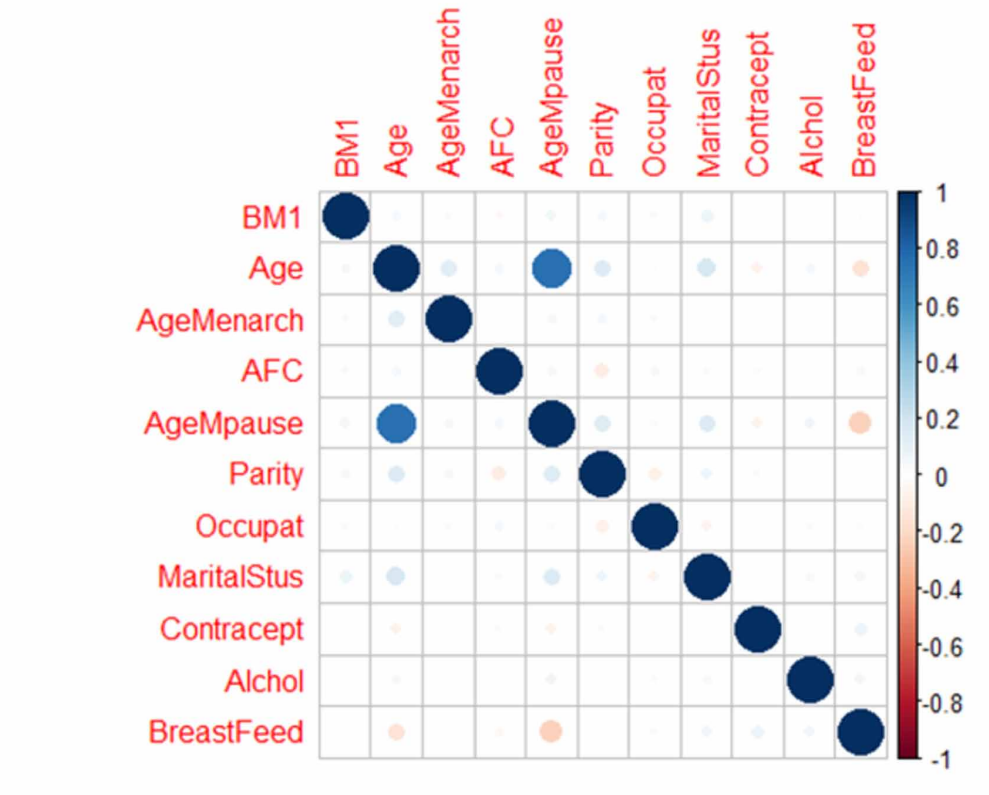


Table 3. Confusion matrix and diagnostic statistics

	Random Forest	Naïve Bayes	CART	Boosted Tree
Accuracy	0.9892	0.8702	0.8697	0.8697
95% CI	(0.9832, 0.9935)	(0.8536, 0.8856)	(0.853, 0.8851)	(0.853, 0.8851)
Sensitivity	0.9170	0.0306	0.8210	0.0010
Specificity	1.0000	0.9961	1.0000	1.0000
Pos Predictive Value	1.0000	0.5385	NaN	NaN
Neg Predictive Value	0.9877	0.8727	0.8697	0.8697

the study. It has the overall accuracy value, the sensitivity value, specificity value, Positive (Pos) and Negative (Neg) predictive values. The parameters estimated in the Random Forest model includes; 98% accuracy (95%, CI= (0.9832,0.99935)), 100% Positive Predictive Value of 100%, Negative Predictive Value, Sensitivity = 92% and specificity = 100%. The Naïve model has an accuracy value of approximately 87% (95% CI = 0.8536,0.8856), Sensitivity value of 3%, Specificity value of 99%, Positive predictive value (PPV,54%), Negative predictive value (NPV,87%). The Classification and Regression Tree (CART) has an overall accuracy value of 87% with (95% CI=0.853,0.8851), Sensitivity value of (82%), Specificity value of 100%, Positive predictive value (PPV=NA), Negative Predictive Value (87%). Finally, the Boosted Tree learning technique has an overall accuracy value of approximately 87% with (95% CI = 0.8530,0.8851), Sensitivity value of 0.0010 (%), Specificity value of 100%, PPV = NaN and NPV = 0.8697. The result for the CART is quite different from the Boosted Tree model as indicated in the table with accuracy value of 0.8697, (95% CI = 0.853, 0.8851), Sensitivity value of (0.0010), Specificity value of 100%, PPV = NaN and NPV = 0.8697.

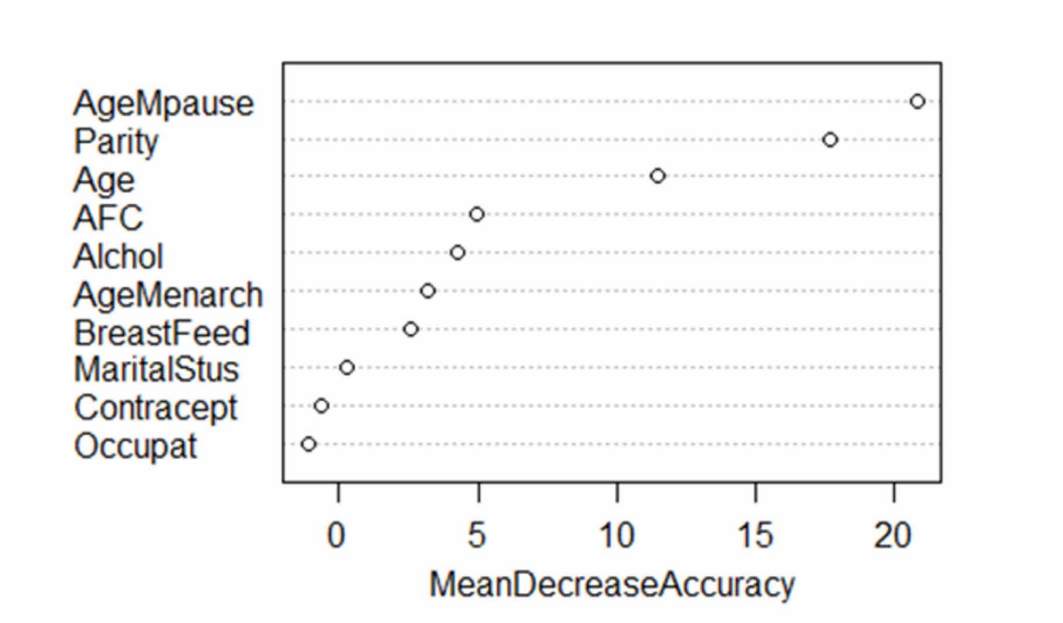
VARIABLE IMPORTANCE

The variable importance test using the Random Forest technique was used determine the order of significant variable contributing to the model adequacy. The results in the Figure 2 shows the most important variables that contributes to the accuracy measures as indicated in the in Table 3. The order of importance of the variables have been expressed in percentages (%). From the figure, the age at menopause of a woman has significant impact on the accuracy rate of the model. Its impact is about 30.37%, Parity is the second most importance with an impact of 29.52%, current age of the women is the third most important variable with an impact of 13.68%, age at first child (8.18%). The least importance variable that contribute to the model adequacy is occupational status of the women which has an impact of about negative (2.84%).

CONCLUSION

The data mining technique analysis (Random Forest, Naïve Bayes, Boosting and CART) shows that, Random Forest technique had the highest or better accuracy value of 0.9892(95% CI,0.9832 -0.9935) and a sensitivity value of about 92%. Which means that Random Forest learning model is the best model that could be used to classify and predict breast cancer based on associated factors. The results obtained and its Interpretation are limited in this analysis because it is difficult to determine whether they reflect general differences that hold true for all other health facilities that record breast cancer data at any point in time.

Figure 2. Variable accuracy measures



LIMITATIONS OF STUDY

Many of the cases in our study had incomplete data related to some risk factors such as age at menopause, age at menarche, present age, age at first child and family history. There is clearly a need to improve documentation of demographic / clinical data in patients' medical records.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

COMPLIANCE WITH ETHICAL STANDARDS

This research has been assessed and approved by the School of Allied Health Sciences, University of Ghana Ethics and Protocol Review Committee with identification number: SAHS-ET. / AA/1A/2013-2014. Patients consent was not sought, but the study was granted a waiver for informed consents by the committee, due to the nature of the study and likelihood that many of the patients whose information we studied were now deceased or lost to follow-up.

ACKNOWLEDGEMENTS

We are thankful to the radiotherapy unit staff of the Korle-Bu Teaching Hospital, who kindly provided me with all the materials that I required for this study and for giving me the permission to browse through the files for the collection of the data. We would like to express our sincere thanks to Messers E. Enchil, Philip Oduro, Isaac Aidoo, Isaac Appiah, Samuel Egyir and Mrs Eva Asiamah, who provided me with assistance in the data collection.

REFERENCES

Abdelaal, M. M. A., Sena, H. A., Farouq, M. W., & Salem, A. B. M. (2010). Using data mining for assessing diagnosis of breast cancer. In *Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT)* (pp. 11-17). IEEE. doi:10.1109/IMCSIT.2010.5679647

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064–1069. doi:10.1016/j.procs.2016.04.224

Bace, R. G. (2000). *Intrusion detection*. Sam's Publishing.

Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age*, *58*(13), 10–110.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a Cancer Journal for Clinicians*, *68*(6), 394–424.

Breiman, L. (2017). *Classification and regression trees*. Routledge. doi:10.1201/9781315139470

Breiman, L., Friedman, J. H., & Olshen, R. A. (2009). *Stone, cj (1984) classification and regression trees*. Belmont, CA: Wadsworth.

Chang, W. P., & Liou, D. M. (2008). Comparison of three data mining techniques with genetic algorithm in the analysis of breast cancer data. *Journal of Telemedicine and Telecare*, *9*(1), 26.

Chaurasia, V., & Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering, 2*(1).

Christobel, A., & Sivaprakasam, Y. (2011). An empirical comparison of data mining classification methods. *International Journal of Computer Information Systems*, *3*(2), 24–28.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., & Bray, F. (2012). *Cancer incidence and mortality worldwide: IARC CancerBase* (Vol. 10, p. 11). GLOBOCAN.

Ford, D., Easton, D. F., & Peto, J. (1995). Estimates of the gene frequency of BRCA1 and its contribution to breast and ovarian cancer incidence. *American Journal of Human Genetics*, *57*(6), 1457. PMID:8533776

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In ICML (Vol. 96, pp. 148-156).

Goldschmidt, P. (2007). Managing the false alarms: A framework for assurance and verification of surveillance monitoring. *Information Systems Frontiers*, *9*(5), 541–556. doi:10.1007/s10796-007-9048-1

Kennedy, R. L. (1997). *Solving data mining problems through pattern recognition*. Prentice Hall PTR.

Lavanya, D., & Rani, K. U. (2012). Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services*, *2*(1), 17–24. doi:10.5121/ijitcs.2012.2103

Liou, D. M., & Chang, W. P. (2015). Applying data mining for the analysis of breast cancer data. In *Data Mining in Clinical Medicine* (pp. 175–189). New York, NY: Humana Press. doi:10.1007/978-1-4939-1985-7_12

Saleema, J. S., Shenoy, P. D., Venugopal, K. R., & Patnaik, L. M. (2014). Cancer prognosis prediction model using data mining techniques. *Data Mining and Knowledge Engineering*, *6*(1), 21–29.

Sasco, A. J., Kaaks, R., & Little, R. E. (2003). Breast cancer: Occurrence, risk factors and hormone metabolism. *Expert Review of Anticancer Therapy*, *3*(4), 546–562. doi:10.1586/14737140.3.4.546 PMID:12934666

Shajahaan, S. S., & Shanthi, S., & ManoChitra, V. (2013). Application of data mining techniques to model breast cancer data. *International Journal of Emerging Technology and Advanced Engineering*, *3*(11), 362–369.

Zand, H. K. K. (2015). A comparative survey on data mining techniques for breast cancer diagnosis and prediction. *Ind. J. Fundam. Appl. Life Sci*, *5*, 4330–4339.

Alice Constance Mensah is a Senior Lecturer at the Department for Mathematics and Statistics. She is currently the Dean for the faculty of Applied Sciences, Accra Technical University, and lectures in Economics and Statistics. She holds a PhD in Applied Statistics, MSc in Statistics, a PGDE and a B. A (Economics and Statistics). She is also a reviewer for Dove Medical Press and Redfame Publishing. Dr. Mensah has a number of articles to her credit and has delivered at national and International Conferences. Her interest is in the application of statistical techniques and models to scientific researches.

Isaac Ofori Asare is a Research and Data Analysis Consultant at Vita Verde Consult, Accra, Ghana. He obtained an MSc in Applied Statistics, a BSc Statistics, and a Higher National Diploma (HND) in Statistics. His research interests are data science (Big Data), multivariate data analysis, longitudinal, and scholarly data analysis.