AJMS
Asian Journal of Mathematical Sciences

RESEARCH ARTICLE

COMPARATIVE STUDY OF TUMOR STAGE PREDICTION IN BREAST CANCER THROUGH FACTOR ANALYSIS AND MULTINOMIAL LOGISTIC REGRESSION

AWOGBEMI CLEMENT ADEYEYE[1*]ILORIADETUNJI KOLAWOLE[1]MUHAMMEDTAHIRMUHAMMED [2] OYEYEMIGAFARMATANMI[3]

[1]StatisticsProgramme, National Mathematical Centre, Abuja, Nigeria
[2]Statistics Department, Federal Polytechnic, Idah, Kogi State, Nigeria
[3]Statistics Department, University of Ilorin, Ilorin, Nigeria

Corresponding Email: awogbemiadeyeye@yahoo.com

Received: 11-04-2025; Revised: 19-05-2025; Accepted: 05-06-2025

ABSTRACT

Breast cancer remains one of the leading causes of cancer-related morbidity and mortality among women globally. Early-stage detection of tumor of breast cancer assists in providing preventive measures against its spread to other parts of the body. In this study, factor analysis was employed in reducing the number of genes to fewer principal components. The Scree and Eigenvalue methods of selecting the number of principal components were employed for comparison purpose. Multinomial logistic regression model was employed to fit the stage of tumor of the breast cancer on the scores of the principal component variables along with the patient's age and tumor size. The findings of the study showed that the eigenvalue approach outperformed the Scree approach ranging from the percentage of variance explained to the accuracy level. Hence, the eigenvalue method of selecting number of components to include in factor analysis is recommended.

Keywords: Accuracy, Breast Cancer, Eigenvalue, Factor Analysis, Scree-plot.

INTRODUCTION

Breast cancer is a public health problem in the world, especially in developing countries where the majority of cases are diagnosed in the last stage (Adebiyi et al., 2022; Isara & Ojedokun, 2011). Additionally, Mares-Quiñones et al., 2024)supported that it is the most common malignancy in women and the second leading cause of cancer-related deaths in developed and industrialized countries. Breast cancer remains one of the leading causes of cancer-related morbidity and mortality among women globally(Azubuike et al., 2018; Oguntunde et al., 2017). The complexity of breast cancer biology is underscored by its heterogeneous nature, which is influenced by various genetic, environmental, and lifestyle factors. Understanding the genetic underpinnings of breast cancer is crucial for developing effective diagnostic tools and therapeutic strategies. Although, there is no preventive mechanism for breast cancer, but early discovery can greatly improve the diagnosis. In addition, early detection of symptoms of breast cancer is hard and challenging. However, mammograms and self-breast examinations are essential for detecting early anomalies before the malignancy progresses Adebiyi et al.(2022)**.**

Identifying and characterizing susceptibility genes for common complex human diseases typify one of the human greatest challenges. This challenge is partly due to the limitations of parametric-statistical methods for detecting gene that dependents solely or partially on interactions with other genes(Ritchie et al., 2001). In this context, statistical methods such as factor analysis and discriminates analysis have emerged as powerful techniques for analyzing gene expression data associated with breast cancer. This work explores the application of these methods in identifying and classifying genes related to breast cancer, and also highlighting their significance in enhancing our understanding of this disease based on the Scree plot criterion and the Eigenvalue criterion.

Factor analysis is a statistical method used to identify underlying relationships between variables by reducing the dimensionality of data. The goal of factor analysis is to explain the variance in the observed variables in terms of underlying latent factors or constructs. In the context of breast cancer research, factor analysis can help uncover latent structures within complex gene expression datasets. By grouping genes that exhibit similar expression patterns, researchers can identify key biological pathways involved in tumor genesis and progression. For instance, factor analysis has been employed to analyze gene expression profiles from various breast cancer subtypes, revealing distinct molecular signatures that correlate with specific clinical outcomes(Adebiyi et al., 2022; Ni et al., 2020). This approach not only aids in the identification of potential biomarkers for early detection but also enhances our understanding of the biological mechanisms driving breast cancer.

Moreover, factor analysis facilitates the exploration of interactions between multiple genes and their contributions to breast cancer phenotypes. By examining these interactions, researchers can identify critical gene networks that may serve as therapeutic targets.

Multinomial logistic regression is a powerful ideal tool for modeling data with more than two response categories. The independent variable under this could either be categorical or continuous.

In breast cancer research, multinomial logistic regression combined with factor analysis can be instrumental in differentiating between various stages of cancer based on gene expression profiles. By constructing a predictive model that maximizes the separation between classes (e.g., luminal A versus triple-negative breast cancer), researchers can achieve high classification accuracy(Adebiyi et al., 2022).

Despite the high mortality rate from breast cancer in Africa being compared to high-income countries, breast cancer has not been extensively studied in the region(Azubuike et al., 2018).The use of factor analysis and discriminant analysis in modeling breast cancer and its associated genes presents several methodological challenges that can hinder the effective identification and classification of genetic factors influencing the disease. Despite the potential of these statistical techniques to uncover complex relationships within high-dimensional gene expression data, their application in breast cancer research is often limited by issues such as data quality, dimensionality, and the interpretability of results.

This research utilized factor analysis and multinomial regression analysis to identify and model the genetic factors associated with stages of breast cancer, enhancing the understanding of tumor heterogeneity and improving prognostic and predictive capabilities in clinical settings. In addition to the contribution to existing literature, this work serves as a road map for policymakers in the detection of breast cancer based on genetic classifications.

Literature Review:

Yue et al.(2018) applied machine learning to breast cancer diagnosis and prognosis. In their work, they compared some machine learning techniques and their applications to breast cancer diagnosis and prognosis. These techniques are Artificial Neural Network (ANN), Support Vector Machines (SVN), Decision Trees (DTs) and K-Nearest Neighbors (KNNs).

Engel et al. (2014) carried out a study which was aimed at showcasing the efficiency of factor analysis and multinomial logistic regression in the analysis of a set of objective and subjective environmental noise data in the city of Curitiba, Brazil. From the total of 21 questions drafted for data collection, factor analysis was implemented to reduce them to 7 factors. The extracted data was added to the noise monitored at 23 points along three parallel streets (objective part). In addition, multinomial logistic regression was applied to regress the dependent variable (interviewees' symptoms and reactions to environmental noise) on the 7 factors extracted. In their finding, it was discovered that about 85% of the symptoms and reactions could be attributed to the combination of the seven factors with the noise measurement data.

Deng et al.(2019)investigated crucial genes and key pathways in breast cancer using bioinformatics analysis. The researchers' work was focused on identifying potential pathogenic and prognostic differentially expressed genes (DEGs) in breast adenocarcinoma through bioinformatics analysis of public dataset. After applying various descriptive analyses, including Kaplan–Meier (KM) plotter to analyze the expression levels and prognostic values of hub genes, their results revealed that mitotic cell cycle and epithelial-to-mesenchymaltransition pathway could be potential pathway accounting for the progression in breast cancer and 6 genes were identified as potential crucial genes. Finally, from the integrated bioinformatics analysis, the present study has identified 321 DEGs and six hub genes that are associated with breast cancer tumor genesis and progression.

Ni et al. (2020) worked on the prediction of the clinic pathological subtypes of breast cancer using a fisher discrimination analysis model based on radiomic features of diffusion-weighted magnetic resonance imaging (MRI). Data from institutional picture archiving and communication system (PACS) between March 2023 and September 2017 on patients who underwent breast magnetic resonance imaging were confirmed. Fisher's discriminant analysis was performed on the data for clinic pathological sub typing by using a backward selection method. The Receiver Operating Characteristic (ROC) curve analysis was performed to differentiate between immunohistochemical biomarker-positive and negative groups. In their conclusion, they applaud that the Fisher discriminant analysis model based on radiomic features of diffusion-weighted MRI reliable method for the prediction of clinic pathological breast cancer subtypes.

Pereira et al.(2016)investigated a comprehensive analysis of breast cancer subtypes using genomic and histopathological data from the METABRIC cohort. The study identifies ten distinct integrative clusters (Incrusts) characterized by unique genomic drivers, which are significantly associated with traditional clinic pathological features such as histological type, tumor grade, receptor status, and lymphocytic infiltration. One of the key strengths of this research is its robust methodology, which includes a central review of tumor pathology involving 1,643 cases from the METABRIC dataset. This allows for a more accurate correlation between the Incrust subtypes and conventional clinical parameters. The findings indicate that while there are notable associations between IntClusts and various histopathological characteristics, no single clinic pathological variable can adequately define an IntClust. This emphasizes the complexity of breast cancer biology and the necessity for genomic stratification to enhance clinical relevance. The results reveal that certain IntClusts are enriched for specific tumor types; for example, IntClust 3 is associated with tubular and lobular carcinomas, while IntClust5 is predominantly HER2 positive. Moreover, the study highlights the prognostic implications of these associations, suggesting that genomic stratification could lead to improved therapeutic strategies tailored to individual tumor profiles.

Pereira et al. (2016)contributed significantly to the study of breast cancer heterogeneity by integrating genomic data with traditional pathology. It advocates for a shift towards more personalized treatment approaches based on comprehensive molecular profiling rather than relying solely on histopathological features. The insights gained from this study could pave the way for future research aimed at optimizing breast cancer management and improving patient outcomes.

Adebiyi et al. (2022)presented a significant advancement in the use of machine learning for breast cancer diagnosis, focusing on methodologies that enhance diagnostic accuracy. The authors employ a systematic approach that integrates feature extraction and classification techniques, specifically utilizing Linear Discriminant Analysis (LDA), Random Forest (RF), and Support Vector Machine (SVM) algorithms.

The study utilizes the Wisconsin Breast Cancer Dataset, which is a well-established dataset for breast cancer research. The authors emphasize the importance of data preprocessing, which involves cleaning and preparing the dataset for analysis. This step is crucial as it ensures that the algorithms operate on high-quality data, thereby improving the reliability of the results. A central component of the methodology is the application of Linear Discriminant Analysis for feature extraction. LDA is employed to reduce dimensionality while preserving as much class discriminatory information as possible. This technique helps in identifying the most relevant features that contribute to distinguishing between benign and malignant tumors, thereby enhancing the classification performance of subsequent models.

Two machine learning classifiers: Random Forest and Support Vector Machine were implemented. Each classifier is trained on the dataset after LDA feature extraction. The choice of these algorithms is justified by their proven effectiveness in handling classification tasks in medical data. The performance of both classifiers is assessed using accuracy as a primary metric. The study reports impressive results, with SVM achieving an accuracy of 96.4% and RF achieving 95.6%. These results are indicative of the methodologies' effectiveness in accurately classifying breast cancer cases. The findings contributed valuable insights into how computational methods can aid in early detection and treatment decisions, ultimately aiming to reduce mortality rates associated with breast cancer.

Rezaeian et al.(2016)explored the use of machine learning models to predict breast cancer patient outcomes based on genomic and molecular data. The aim of their study was to predict patient survival outcomes after hormone therapy (HT) and chemotherapy (CT) in breast cancer using gene expression profiles. This was built on the premise that certain genetic markers can indicate resistance or sensitivity to specific cancer therapies, which can guide personalized treatment plans. The METABRIC dataset, a substantial breast cancer genomic repository, serves as the foundation for analysis, while the models developed were inspired by biological pathways related to drug action. The authors employed several machine learning models, primarily Support Vector Machine (SVM) and Random Forest (RF), to analyze the gene expression signatures related to CT agents like paclitaxel and HT agents like tamoxifen. They used feature selection techniques to refine relevant gene subsets and performed cross-validation to address model robustness. Additionally, they tackled challenges such as batch effects and overfitting by conducting cross-study validation and addressing heterogeneity across data sources. In their findings, the paclitaxel gene signature showed significant predictive power. The RF model achieved high accuracy (85.5%) in predicting survival for HT patients at a three-year threshold, while SVM models also performed well, particularly in the subgroup of patients receiving combined HT and CT. These findings underscore the potential of machine learning in identifying biomarkers for cancer treatment response, with paclitaxel models yielding the highest prediction accuracy among evaluated therapies.

Ritchie et al. (2001)introduces the multifactor-dimensionality reduction (MDR) method to analyze gene-gene interactions associated with breast cancer. By focusing on estrogen-metabolism genes, the authors identify complex interactions without needing large sample sizes or specific inheritance models, as MDR is nonparametric and model-free. Using simulated and real breast cancer datasets, they demonstrate MDR’s capability to capture gene interactions that traditional models might miss, finding a significant four-locus interaction linked to breast cancer risk. While MDR is promising for multidimensional genetic data, challenges like computational intensity and interpretive complexity highlight areas for refinement. The study represents a notable advancement in genetic epidemiology, particularly for multifactorial diseases like breast cancer, by showcasing MDR’s potential to uncover complex biological interactions in relatively small datasets.

Methodology:

The dataset for this work was extracted from Kaggle on the 7th of November, 2024 (https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/data). It is a product of the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database which is a Canada-UK Project which contains targeted sequencing data of 1,980 primary breast cancer samples. In addition, it is a combination of both Clinical and genomic data.

From the clinical attribute of the dataset, the age of the patient at diagnosis time, tumor size and stage of cancer base on the involvement of surrounding structures, lymph nodes and distant spread were extracted. The stages of breast cancer were categorized from 0 to 4.

From the genetics part of the dataset which contains m-RNA levels z-score for 331 genes, and mutation for 175 genes, due to missing data, only 489 genes were selected for analysis.

Due to the large number of variables (genes), they were reduced through a factor analysis, which identified the most important factors that reveal the information collected about the genes without significant loss of information.

The number of principal components of the reductions was ascertained using the Scree plot criterion and the Eigenvalue criterion. Thereafter, the scores of the selected features along with the age of the patient will serve as the independent variable and the stages of the breast cancer will be regression them using multinomial logistic regression analysis. Multinomial logistic regression analysis is suitable for the analysis of studies that deal with categorical response variables and its main objective is to describe the relationship between a dependent variable and a set of predictor variables.

Multinomial Regression Analysis (MRA)

Multinomial regression analysis is a statistical method used to model the relationship between a nominal dependent variable with more than two categories and one or more independent variables. In other words, this technique is used when the dependent variable is categorical with three or more unordered levels.The method also makes use of a linear combination of independent variables to explore correlations with outcome likelihoods and to predict outcomes using specific input conditions (Jim Frost, 2020). The modeling approach of multinomial regression involves estimation of the log odds of each category relative to a reference category. This requires creating multiple equations (X-1) for X categories, thereby allowing for a comprehensive analysis of the relationship between variables.
Let the probability of occurring and the probability of not occurring be denoted by the

$$\text{Odds} = \frac{p}{1-p} \tag{1}$$

$$\log\frac{p}{1-p} = a_0 + a_1x_1 + \cdots + a_px_p \tag{2}$$

Equation (2) is translated to (3) in terms of probabilitie (3)

where p represents the probability of an even occurring and a’s are the regression coefficients and x’s are the independent variables.
Suppose a dependent variable has N categories. The last or the value with the highest frequency of the dependent variable is used as the reference category so that for a dependent variable N categories, the computation of N-1 equations are required. This is carried out for each base relative to the reference base and the independent variables (Starkweather and Amanda, 2011).
For the reference category and $m = 2, \dots, M$ we have

$$\ln\left(\frac{p(Y_i=m}{p(Y_i=1}\right) = \beta_m + \sum_{b=1}^{B} \beta_{bm}X_{bi} \tag{4}$$

where$Y_i$and i^{th} respondents belong to M category, $\alpha \, and \, \beta$ are regression coefficients, x’s are independent variables, $i = 1,2, \dots, n \; and \; b = 1,2, \dots B$.
In case there more than two groups, we have more complex probabilitiesthan logistic regression.
Thus, for $m = 2,3 \dots, M$, we have

$$\mathrm{p}(\mathrm{Y_i} = \mathrm{m}) = \frac{\exp(\mathrm{K_{mi}})}{1+\sum_{\mathrm{j=2}}^{\mathrm{M}} \mathrm{K_{ji}}} \qquad (5)$$

where$K_{mi} = \gamma_m \sum_k^K \beta_{mk} X_{ik}$ and K_{ji} is the linear combination of independent variables of all outcomes other than m outcome.

For the reference category, we have

$$p(Y_i = 1) = \frac{1}{1+\sum_{j=2}^{M} K_{ji}} \qquad (6)$$

Assumptions of Multinomial Logistic Regression :

The data used in multinomial logistic regression analysis according to Jim Frost (2020) must satisfy the following conditions:

(i) Categorical Outcome: The dependent variable must have at least three unordered categories

(ii) Independence: An observation's outcome should not influence another observation.

(iii) Non-Perfect Multicollinearity: Independent variables should not be perfectly correlated as it can distort coefficient estimation.

(iv) Linearity in the Logit: There should be linear relationship between the predictors and the log odds of the outcomes.

(v) There is no presence of outliers.

Discussion of Results :

From the data matrix, the results generated showed 20 factor loading using the Scree criterion and 104 factor loading using the Eigenvalue criterion from the original 489 variables (genes). However, the Scree criterion of selecting the number of principal components only explained 44% of the total variance, while the eigenvalue criterion accounted for 70% of the total variance, as shown in Table 1.

Table 1: Total Variance Explained by the two Selection Criteria

Criteria	Number of Factors Selected	% Total Variance Explained
Scree	20	43.772
Eigenvalue	104	70.096

The number of principal components selected using the Scree criteria was obtained at the position where the Scree plot curve formed an elbow shape, as shown in Figure 1.

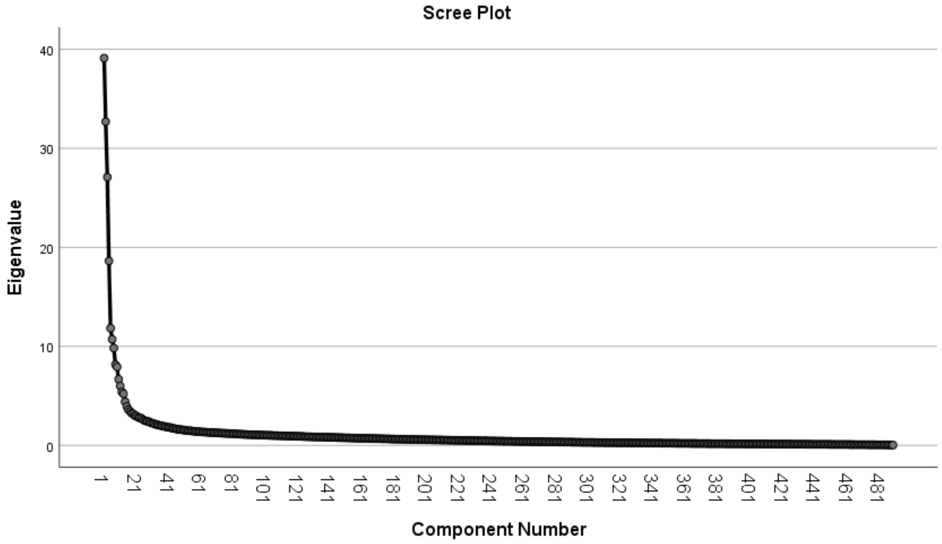

Figure 1: Scree Plot showing the possible number of components for the variables

Multinomial logistic regression

From the model fitted using multinomial logistics regression, based on the Chi-square test, it was found that, under the Scree method, G2(80)=235.811 with P=0.000. Likewise, for the Eigenvalue approach, G2(424)=107719 and P=0.000 as shown in Table 2. Hence, the null hypotheses that the models are not significant at the usual level of significance are rejected. In conclusion, under each criterion, there is at least one independent variable that significantly influences the dependent variable.

Table 2: Model Fitting Information

Model/Creiteria	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Scree	2404.995	235.811	80	0
Eigenvalue	1541.059	1077.719	424	0

Because multinomial logistic regression does not calculate R^2 in the same way as linear regression does, Pseudo-R^2s are calculated based on the maximum likelihood coefficients between variables(Engel et al., 2014). Pseudo R^2s are calculated through the Cox and Snell coefficient, which never reaches the maximum value of 1 (100 %). The Nagelkerke coefficient normalizes the Cox and Snell coefficient, so that the maximum value can reach 1 (100 %).

The purpose of the maximum likelihood fit is to obtain estimates of statistical parameters from a sample, ensuring consistency, efficiency and adjustment of model parameters(Engel et al., 2014).

In this study, the results of the pseudo R^2logistic regression are shown in table 3.

Table 3: Pseodo R-Square

	Scree	Eigenvalue
Cox and Snell	0.155	0.537
Nagelkerke	0.182	0.635
McFadden	0.089	0.412

Thus, under the Scree approach, the model indicates that 15.5% (the Cox and Snell R^2) of the interdependencies between the independent variables and the dependent variable can be explained. In other words, 15.5% of the answers of the factor genes levels in this study were correlated with the variable tumor stage. This is a poor result.

Conversely, under the Eigenvalue approach, the model indicates that 53.7% (the Cox and Snell R^2) of the interdependencies between the independent variables and the dependent variable can be explained. In other words, 53.7% of the answers of the factor genes levels in this study were correlated with the variable tumor stage. This is better compare to the initial Scree approach.

Table 4: Prediction Matrix under the Scree Approach

		Predicted Response Category				Total
		0	1	2	3	
tumor stage	0	1	0	3	0	4
	1	0	156	318	1	475
	2	0	99	701	0	800
	3	0	14	101	0	115
	4	0	0	9	0	9
	Total	1	269	1132	1	1403

The performance of the two approaches was evaluated via their accuracy levels. Accuracy is the ratio of the correct forecast provided by the model over the entire predictions of all the task.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}, \quad (1)$$

where TP = True Positive, FP = False Positive, TN = True Negative, and FN = False Negative respectively.

From table 4, the accuracy of the multinomial logistic regression under the Scree selection criterion is 0.6115 (61.15%). For the Eigenvalue selection criterion, the accuracy is 0.7786 (77.86%) as calculated using table 5. Based on this evaluation, the Eigenvalue approach outperformed the Scree approach of principal component selection.

Table 5: Prediction Matrix under the Eigenvalue Approach

		Predicted Response Category					Total
		0	1	2	3	4	
tumor_stage	0	3	0	0	0	0	3
	1	0	351	120	4	0	475
	2	0	108	679	13	0	800
	3	0	9	56	48	0	113
	4	0	0	0	0	9	9
Total		3	468	855	65	9	1400

Conclusion:

For the purpose of this study, the eigenvalue method of selecting principal components outperformed the Scree approach. Under the eigenvalue approach, 104 principal components which accounted for about 70% of the total variance of the variables were explained. While the Scree approach selected 20 principal components with 44% total variance explained.

Although, the multinomial logistic regression model obtained under each approach was significant. However, the eigenvalue approach outperformed the Scree method in terms of the interdependency between the independent variable and the dependent variable explained.

In addition, the accuracy level of the eigenvalue approach significantly outperformed that of the Scree approach. Hence the eigenvalue approach of selecting the number of principal components to include in an analysis is recommended.

Acknowledgements

We are most grateful to National Mathematical Centre (NMC), Abuja for providing an enabling environment and resources needed to carry out this research.

REFERENCES

1. Adebiyi, M. O., Arowolo, M. O., Mshelia, M. D.& Olugbara, O. O. (2022). A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. Applied Sciences (Switzerland), 12(22).

2. Azubuike, S. O., Muirhead, C., Hayes, and L. & McNally, R. (2018) .Rising Global Burden of Breast Cancer: The Case of Sub-Saharan Africa (With Emphasis on Nigeria) And Implications for Regional Development: A Review. World Journal of Surgical Oncology, 16(1), 1-13.

3. Deng, J. L., Xu, Y. H.& Wang, G. (2019). Identification of Potential Crucial Genes and Key Pathways in Breast Cancer Using Bioinformatic Analysis. Frontiers in Genetics, 10(JUL), 1-17.

4. Engel, M. S., De Vasconcelos Segundo, E. H.& Zannin, P. H. T. (2014). Statistical Analysis of a Combination of Objective and Subjective Environmental Noise Data using Factor Analysis and Multinomial Logistic Regression. Stochastic Environmental Research And Risk Assessment, 28(2), 393-399.

5. Isara, A. R. & Ojedokun, C. I. (2011). Knowledge of Breast Cancer and Practice of Breast Self Examination among Female Senior Secondary School Students in Abuja, Nigeria. Journal Of Preventive Medicine And Hygiene, 52(4), 186-190.

6. Jim Frost, M.S. (2020). Regression Analysis: An Intuitive Guide for using and Interpreting Linear Models. Jim Publishing.

7. Mares-Quiñones, M. D., Galan Vasquez, E., Pérez-Rueda, E., Pérez-Ishiwara, D. G., Medel-Flores, M. O.& Gómez-García, M. del C. (2024). Identification of Modules and Key Genes Associated with Breast Cancer Subtypes Through Network Analysis. Scientific Reports, 14(1), 1-18.

8. Ni, M., Zhou, X., Liu, J., Yu, H., Gao, Y., Zhang, X.& Li, Z. (2020). Prediction of the Clinic pathological Subtypes of Breast Cancer Using A Fisher Discriminant Analysis Model Based on Radiomic Features of Diffusion-Weighted MRI. BMC Cancer, 20(1), 1–11.

9. Oguntunde, P. E., Adejumo, A. O. & Okagbue, H. I. (2017). Breast Cancer Patients in Nigeria: Data Exploration Approach. Data in Brief, 15, 47-57.

10. Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S. J., Tsui, D. W. Y., Liu, B., Dawson, S. J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R. &Caldas, C. (2016). The Somatic Mutation Profiles of 2,433 Breast Cancers Refines their Genomic and Transcriptomic Landscapes. Nature Communications, 7(May).

11. Rezaeian, I., Mucaki, E. J., Baranova, K., Pham, H. Q., Angelov, D., Ngom, A., Rueda, L.& Rogan, P. K. (2016). Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Study by Biochemically-inspired Machine Learning. F1000Research.

12. Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F.& Moore, J. H. (2001). Multifactor-Dimensionality Reduction Reveals High-Order Interactions Among Estrogen-Metabolism Genes in Sporadic Breast Cancer. American Journal of Human Genetics, 69(1), 138-147.

13. Stark weather and Amanda (2011). Multinomial Logistic Regression. http://it.unt.edu/sites/default/files/mlr_jds-aug2011.pdf.

14. Yue, W., Wang, Z., Chen, H., Payne, A.& Liu, X. (2018). Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis, Designs, 2(2), 1-17.

15. Hailu, T., Berhe, H., Hailu, D. (2016) Awareness of Breast Cancer and Its Early Detection Measures among Female Students, Northern Ethiopia. Int. Journal of Public Health Sci. 5, 213.