Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2023;15(2):91-101

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA DOI: 10.5336/biostatic.2023-97597

A Novel Variable Selection Procedure for Binary Logistic Regression Using Akaike Information Criteria Testing: An Example in Breast Cancer Prediction: Methodological Study (Validity Study)

Akaike Bilgi Kriterleri Testi Kullanılarak İkili Lojistik Regresyon için Yeni Bir Değişken Seçim Prosedürü: Meme Kanseri Tahmininde Bir Örnek: Metodolojik Çalışma (Geçerlilik Çalışması)

Oyebayo Ridwan OLANIRAN[a], Saidat Fehintola OLANIRAN[b]

[a]Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Ilorin, PMB 1515, Kwara State, Nigeria
[b]Department of Statistics and Mathematical Sciences, Faculty of Pure and Applied Science, Kwara State University, PMB 1530 Malete, Kwara State, Nigeria

ABSTRACT Objective: Breast cancer is a leading cause of cancer-related death among women worldwide, with approximately 2.3 million new cases and 685,000 deaths reported in 2020 alone. One critical step in developing effective classification and prediction models is variable selection, which involves identifying a subset of relevant variables from a larger set of potential predictors. Accurate variable selection is crucial for building interpretable and robust models that are not overfit to noise, leading to improved model performance and generalization ability. In this paper, we proposed an alternative objective approach for comparing two Akaike Information Criterions (AIC) that originated from two competing models, such that the magnitude of the difference is subjected to the statistical test of significance. **Material and Methods:** We developed a new backward elimination variable selection procedure similar in spirit to the existing "stepAIC" within the environment of R statistical software. We used both simulated and Wisconsin breast cancer diagnostic datasets to compare the proposed method's variable selection and predictive performances with "stepAIC" and LASSO. **Results:** The simulation showed that the proposed AIC procedure achieved higher variable selection sensitivity, specificity and accuracy when compared to stepAIC and LASSO. Also, the proposed AIC method's prediction results are relatively comparable with stepAIC and LASSO at various simulated data dimensions. Similar supremacy results were observed with the breast cancer dataset used. **Conclusion:** The AIC-based variable selection approach proposed is a promising method that integrates AIC with statistical testing for improved variable selection in breast cancer classification and prediction.

Keywords: Breast cancer; Akaike Information Criteria; variable selection; backward selection; LASSO

ÖZET Amaç: Göğüs kanseri, yalnızca 2020 yılında bildirilmiş yaklaşık 2,3 milyon yeni vaka ve 685.000 ölüm ile dünya çapında kadınlar arasında kanser ilişkili ölümlerin başında gelen sebeplerinden biridir. Etkili sınıflandırma ve tahmin modelleri geliştirmede kritik bir adım, daha geniş bir potansiyel öngörücü setinden, ilgili değişken alt seti tanımlamayı içeren değişken seçimdir. Doğru değişken seçimi, gürültüye fazla uyum sağlamayan, yorumlanabilir ve sağlam modeller oluşturmada çok önemlidir. Bu durum gelişmiş model performansı ve generalizasyon becerisi sağlar. Bu makalede, 2 rakip modelden oluşan 2 Akaike Bilgi Kriterleri'ni [Akaike Information Criterions (AIC)] karşılaştırdığımız alternatif objektif bir yaklaşım sunduk, öyle ki farkın büyüklüğü istatistiksel anlamlılık testine tabi tutulmuştur. **Gereç ve Yöntemler:** R istatistik yazılımı ortamında bulunan "stepAIC"ye benzer yeni bir geriye dönük eleme değişken seçme prosedürü geliştirdik. Sunulan metodun değişken seçimi ile "stepAIC" ve LASSO ile tahmini performanslarını karşılaştırmak için simüle edilmiş, Wisconsin meme kanseri tanı veri setlerini kullandık. **Bulgular:** Simülasyon, sunulan AIC prosedürünün stepAIC ve LASSO'ya kıyasla yüksek değişken seçim hassasiyeti, spesifitesi ve doğruluğu kazandığını göstermiştir. Ayrıca, sunulan AIC yönteminin tahmin sonuçları, simüle edilen çeşitli veri boyutlarında stepAIC ve LASSO ile görece karşılaştırılabilirdir. Kullanılan meme kanseri veri setinde de benzer üstünlük sonuçları gözlemlenmiştir. **Sonuç:** AIC temelli değişken seçim yaklaşımı, meme kanseri sınıflandırması ve tahmininde AIC'yi gelişmiş değişken seçimi için istatistiksel testlere entegre eden, umut verici bir metottur.

Anahtar Kelimeler: Meme kanseri; Akaike Bilgi Kriterleri; değişken seçimi; geriye dönük seçim; LASSO

Correspondence: Oyebayo Ridwan OLANIRAN
Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Ilorin, PMB 1515, Kwara State, Nigeria
E-mail: rid4stat@yahoo.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 22 Apr 2023 ***Received in revised form:*** 07 Jul 2023 ***Accepted:*** 07 Jul 2023 ***Available online:*** 13 Jul 2023

Breast cancer is a leading cause of cancer-related death among women worldwide, with approximately 2.3 million new cases and 685,000 deaths reported in 2020 alone.[1] Early and accurate detection of breast cancer is crucial for effective treatment and improved patient outcomes. Machine learning techniques, specifically classification and prediction models, have shown great promise in assisting healthcare professionals in the early diagnosis and prognosis of breast cancer. A common machine learning predictive model that is often applied in breast cancer prediction is logistic regression. Logistic regression is a widely used statistical method for breast cancer prediction. It is a binary classification algorithm that models the relationship between a set of independent variables (features) and the probability of a binary outcome, such as the presence or absence of breast cancer.

One critical step in developing effective logistic regression models is variable selection, which involves identifying a subset of relevant variables or variables from a larger set of potential predictors. Accurate variable selection is crucial for building interpretable and robust models that are not overfit to noise, leading to improved model performance and generalization ability.

One popular method for variable selection is the Akaike Information Criterion (AIC), which Akaike introduced in 1974.[2] AIC is a widely used model selection criterion that balances the trade-off between model complexity and goodness of fit, making it suitable for both model selection and prediction tasks. AIC is based on the principle of information theory and penalizes models with higher complexity, encouraging the selection of simpler models with fewer variables. The goal of AIC variable selection is to identify the subset of predictors that provides the best trade-off between goodness of fit and model complexity. Variable or model selection with AIC involves identifying variable(s) or model(s) with the least AIC value. In simple terms, the lower the AIC value for a model, the better the model.

Several studies have used AIC variable selection to improve breast cancer classification and prediction. One such study by Li et al. used the AIC for selecting the best model among competing models before performing breast cancer classification and prediction.[3] The authors performed variable selection with the use of AIC to identify relevant variables from high-dimensional breast cancer data.

In their study, Li et al. used a dataset containing gene expression profiles of breast cancer patients and applied their proposed AIC-based variable selection algorithm to identify a subset of genes that are strongly associated with breast cancer classification and prediction.[3] The authors compared the performance of their AIC-based approach with other commonly used variable selection methods, such as Lasso, Ridge, and Elastic Net, and demonstrated that their proposed method outperformed these methods in terms of prediction accuracy and model interpretability.

Another similar study by Yu et al. employed AIC variable selection for gene expression data in triple-negative breast cancer classification.[4] The authors used AIC to compare different models with different subsets of variables and selected the model with the lowest AIC value as the best model. They demonstrated that their AIC-based variable selection method outperformed other commonly used methods, such as stepwise selection and LASSO, in terms of classification accuracy and model interpretability.

Similarly, Chen et al. used AIC variable selection for early breast cancer prediction using radiomics variables extracted from medical images.[5] They applied AIC to evaluate the performance of different variable subsets and identified the optimal subset of variables that yielded the lowest AIC value. Their results showed that the AIC-based variable selection method improved the prediction accuracy compared to other variable selection methods, such as correlation-based and mutual information-based methods.

In addition to classification and prediction, AIC has also been used for variable selection in other types of cancer studies. For example, in a study by Liu et al., AIC was used to select the most informative variables from a large set of clinical, pathological, and imaging variables for predicting breast cancer recurrence.[6] The authors developed a predictive model using AIC-selected variables and demonstrated that their model had superior predictive performance compared to other models.

Another recent study by Li et al. utilized AIC variable selection technique for cervical cancer classification and prediction was proposed.[7] The authors developed a novel algorithm that combines a backward elimination procedure with AIC to select the most relevant variables from a large dataset of cervical cancer patients. The backward elimination procedure starts with the full variable set and iteratively removes the least significant variables based on their p-values from the logistic regression model. Then, AIC variable selection is performed to evaluate the performance of the model with different variable subsets and select the optimal variable set that minimizes the AIC value. The proposed method was compared with other variable selection techniques, including stepwise selection and LASSO regression, on a publicly available breast cancer dataset. The results showed that the proposed method outperformed other methods in terms of classification accuracy, sensitivity, specificity, and Area Under the Curve (AUC) of the Receiver Operating Characteristic curve, indicating its superior performance in breast cancer classification and prediction.

Furthermore, another recent study by Chen et al. and Zhang et al. also utilized AIC variable selection in breast cancer classification and prediction.[8,9] The authors proposed a novel procedure that integrates AIC-based variable selection with support vector machine (SVM) classification for breast cancer diagnosis. The AIC-based variable selection method was used to select the most informative variables from a high-dimensional dataset of gene expression profiles, and the selected variables were then used to train an SVM classifier for breast cancer diagnosis. The proposed method was evaluated on a real-world breast cancer dataset, and the results demonstrated that it achieved higher classification accuracy, sensitivity, specificity, and AUC compared to other variable selection methods, such as recursive variable elimination and random forest.

In summary, recent literature has shown that AIC-based variable selection methods can improve the accuracy and interpretability of breast cancer classification and prediction models. These methods offer a promising approach to identifying the most relevant variables from high-dimensional datasets and enhance the performance of breast cancer diagnosis and prediction models. However, all the studies mentioned above performed AIC variable selection using a subjective approach. By subjective, we mean the best models or variables were identified based on minimum AIC value. The big question is, what is the minimum or optimal AIC threshold difference for a specific dataset? Thus, in this paper, we propose an alternative objective approach for comparing two AICs that originated from two competing models, such that the magnitude of the difference is subjected to the statistical test of significance. Specifically, a new backward elimination variable selection procedure similar to the existing "stepAIC" procedure in R statistical software is developed. The method's variable selection and predictive performances are compared with "stepAIC" and LASSO using both simulated and real-life breast cancer datasets.

MATERIAL AND METHODS

Logistic regression model: Logistic regression is a statistical method used to analyze the relationship between a binary dependent variable (breast cancer outcome) and one or more independent variables (diagnostic or prognostic factors). The goal of logistic regression is to find the best-fitting model that can predict the probability of the binary outcome based on the values of the independent variables.[10-15]

The logistic regression model is formulated using a logistic function, which is a type of sigmoid function that maps any real-valued input to the range [0, 1]. The logistic function is defined as:

$$P(y = 1|\boldsymbol{X}) \; = \; 1 \, / \, (1 \; + \exp(-\boldsymbol{X}\beta)) \tag{1}$$

where $P(y = 1|\boldsymbol{X})$ also denoted by π is the probability of the dependent variable y (breast cancer outcomes) being equal to 1 given the values of the independent variables $\boldsymbol{X}$ and their associated coefficient β, and exp() is the exponential function. The likelihood of parameter β in equation (1) for n observed samples according to (7) is given by

$$L(\beta|\boldsymbol{X},y)=\prod_{i=1}^{n}\pi_i^{y_i}\,(1-\pi_i)^{1-y_i}$$

$$L(\beta|\boldsymbol{X},y)=\prod_{i=1}^{n}\left(\frac{\exp(\boldsymbol{X_i}\beta)}{1+\exp(\boldsymbol{X_i}\beta)}\right)^{y_i}\left(\frac{1}{1+\exp(\boldsymbol{X_i}\beta)}\right)^{1-y_i}$$

The corresponding log-likelihood follows as:

$$l(\beta|\boldsymbol{X},y)=\sum_{i=1}^{n}[y_i\log(\pi_i)+(1-y_i)\log(1-\pi_i)]$$

$$l(\beta|\boldsymbol{X},y)=\sum_{i=1}^{n}\left[y_i\log\left(\frac{\exp(\boldsymbol{X_i}\beta)}{1+\exp(\boldsymbol{X_i}\beta)}\right)+(1-y_i)\log\left(\frac{1}{1+\exp(\boldsymbol{X_i}\beta)}\right)\right]$$

$$l(\beta|\boldsymbol{X},y)=\sum_{i=1}^{n}[y_i\boldsymbol{X_i}\beta-\log(1+\exp(\boldsymbol{X_i}\beta))]$$

The logistic regression model is trained using a maximum likelihood estimation method, which estimates the parameters of the model that maximizes the likelihood of observing the data. The model parameters are usually estimated using numerical optimization algorithms, such as gradient descent or Newton-Raphson (3).

AIC Criteria for logistic regression model: In logistic regression, AIC is calculated using the following formula:

$$AIC\ =\ -2l(\beta|\boldsymbol{X},y)\ +\ 2k$$

$$AIC\ =\ -2\left\{\sum_{i=1}^{n}[y_i\boldsymbol{X_i}\beta-\log(1+\exp(\boldsymbol{X_i}\beta))]\right\}\ +\ 2k$$

where k is the number of parameters in the model, also known as the model's complexity. In logistic regression, the parameters include the regression coefficients for the predictor variables and an additional parameter for the intercept term. The AIC compares different logistic regression models, with lower values indicating better model fit. The AIC penalizes models with higher complexity (i.e., more parameters) to prevent overfitting. The term $2k$ acts as a penalty term in the AIC formula, discouraging overly complex models that may fit the data well but not generalize to new data. When comparing multiple logistic regression models, the model with the lowest AIC value is considered the best-fitting model, as it balances the goodness of fit and model complexity (3).

Test of difference of two AICs for logistic regression: Suppose $\widehat{AIC}_1$ and $\widehat{AIC}_2$ are two Akaike Information Criteria estimated from two competing nested models with the number of parameters $k_1$and k_2. If $k_1 > k_2$, we would expect $\widehat{AIC}_2 > \widehat{AIC}_1$, thus we can construct a z-like statistic as follows:

$$z=\frac{\widehat{AIC}_2-\widehat{AIC}_1-E[\widehat{AIC}_2-\widehat{AIC}_1]}{\sqrt{Var[\widehat{AIC}_2-\widehat{AIC}_1]}} \qquad (2)$$

where $E\left[\widehat{AIC}_2-\widehat{AIC}_1\right]$ and $Var\left[\widehat{AIC}_2-\widehat{AIC}_1\right]$ are the expected value and variance of the theoretical distribution of $\widehat{AIC}_2-\widehat{AIC}_1$. The statistic z can be used to test the following one-sided hypothesis:

$H_0: AIC_2 - AIC_1 \leq 0$

$H_a: AIC_2 - AIC_1 > 0.$

Under H_0, the statistic z can be approximated with the normal distribution $N(0,1)$, and under H_a, the statistic z follows a normal distribution with mean and variance $E\left[\widehat{AIC}_2-\widehat{AIC}_1\right]$ and $Var\left[\widehat{AIC}_2-\widehat{AIC}_1\right]$.

Proof:

Given that:

$$\widehat{AIC}_1 = -2l(\beta^1|\boldsymbol{X^1}, y) + 2k_1 \quad (3)$$

$$\widehat{AIC}_2 = -2l(\beta^2|\boldsymbol{X^2}, y) + 2k_2 \quad (4)$$

Equation (4) – (3) gives:

$$\widehat{AIC}_2 - \widehat{AIC}_1 = -2\left[l(\beta^2|\boldsymbol{X^2}, y) - l(\beta^1|\boldsymbol{X^1}, y)\right] + 2(k_2 - k_1) \quad (5)$$

Recall that $-2\left[l(\beta^2|\boldsymbol{X^2}, y) - l(\beta^1|\boldsymbol{X^1}, y)\right]$ is distributed $\chi^2_{k_1-k_2}$, and is the likelihood ratio test for comparing two nested models where $\boldsymbol{X^1}$ is the full model and $\boldsymbol{X^2}$ is the reduced model, thus,[16]

$$\widehat{AIC}_2 - \widehat{AIC}_1 = \chi^2_{k_1-k_2} + 2(k_2 - k_1) \quad (6)$$

Therefore, the expected value and variance of the difference between the two AICs, $\widehat{AIC}_2 - \widehat{AIC}_1$ are

$$E\left[\widehat{AIC}_2 - \widehat{AIC}_1\right] = E\left[\chi^2_{k_1-k_2}\right] + 2(k_2 - k_1)$$

$$E\left[\widehat{AIC}_2 - \widehat{AIC}_1\right] = k_1 - k_2 + 2(k_2 - k_1)$$

$$E\left[\widehat{AIC}_2 - \widehat{AIC}_1\right] = k_2 - k_1 \quad (7)$$

Consequently, the variance is;

$$Var\left[\widehat{AIC}_2 - \widehat{AIC}_1\right] = Var\left[\chi^2_{k_1-k_2}\right]$$

$$Var\left[\widehat{AIC}_2 - \widehat{AIC}_1\right] = 2(k_1 - k_2) \quad (8)$$

Now, under H_0 and backward elimination procedure, we can assume that $k_1 = k_2 + 1$, that is the model 1 is one additional parameter larger than model 2. Thus, $E\left[\widehat{AIC}_2 - \widehat{AIC}_1|H_0\right] = -1$, and the corresponding variance is $Var\left[\widehat{AIC}_2 - \widehat{AIC}_1|H_0\right] = 2$. Therefore, the mean and variance of the statistic z are:

$$E[z|H_0] = E\left[\frac{\widehat{AIC}_2 - \widehat{AIC}_1 - E\left[\widehat{AIC}_2 - \widehat{AIC}_1\right]}{\sqrt{Var\left[\widehat{AIC}_2 - \widehat{AIC}_1\right]}}|H_0\right]$$

$$E[z|H_0] = \left[\frac{E\left[\widehat{AIC}_2 - \widehat{AIC}_1\right] + 1}{\sqrt{2}}|H_0\right]$$

$$E[z|H_0] = \left[\frac{-1 + 1}{\sqrt{2}}|H_0\right]$$

$$E[z|H_0] = 0 \quad (9)$$

$$Var[z|H_0] = Var\left[\frac{\widehat{AIC}_2 - \widehat{AIC}_1 - E\left[\widehat{AIC}_2 - \widehat{AIC}_1\right]}{\sqrt{Var\left[\widehat{AIC}_1 - \widehat{AIC}_2\right]}}|H_0\right]$$

$$Var[z|H_0] = \left[\frac{Var\left[\widehat{AIC}_2 - \widehat{AIC}_1\right]}{Var\left[\widehat{AIC}_2 - \widehat{AIC}_1\right]}|H_0\right]$$

$$Var[z|H_0] = \left[\frac{2}{2}|H_0\right]$$

$$Var[z|H_0] = 1 \quad (10)$$

Equations (9) and (10) imply that the statistic $z \sim N(0,1|H_0)$.

Variable selection based on AIC testing: We develop a backward elimination procedure based on AIC testing in equation (2). We adapted the existing "stepAIC" procedure in R statistical software by updating the default AIC model comparison based on small thresholds Δ. Instead of predefining Δ, we compare the AICs of model $k+1$ and k at each step by calculating statistic z in (2) and the associated p-value. The p-value of z can be easily computed in R using the function "1-pnorm(z)". The new variable selection algorithm is presented below:

Algorithm 1: AIC testing variable selection algorithm

1. Let $M(\widehat{\beta^1|\boldsymbol{X^1}}, y)$ be the full model with k_1 variables.
2. Compute $\widehat{AIC}_1 = -2l(\beta^1|\boldsymbol{X^1}, y) + 2k_1$.
3. Create a null $\widehat{AIC}$ vector of size k_1.
4. For each k in $1, \ldots, k_1$ do:
   a. Compute $\widehat{aic}_k = -2l\left(\beta^k|\boldsymbol{X^k}, y\right) + 2$
   b. End For k in $1, \ldots, k_1$.
5. Sort the $\widehat{aic}_k$ in decreasing order and store them as $\widehat{AIC}_{sort}$.
6. Create a null $\widehat{AIC}_{step}$ and $\hat{p}_{step}$ vectors of size k_1
7. For l in $1, \ldots, k_1$ do:
   a. Fit an updated model $M(\beta^2|\boldsymbol{X^2}, y) = M(\beta^{k_1-l}|\boldsymbol{X}^{k_1-l}, y)$ using the maximum $\widehat{aic}_k$ in $\widehat{AIC}_{sort}$.
   b. Compute $\widehat{aic}_l = -2l(\beta^2|\boldsymbol{X^2}, y) + 2(k_1 - l)$
   c. Compute the test statistic $\hat{z}_l = \frac{\widehat{AIC}_1 - \widehat{aic}_l - 1}{\sqrt{2}}$
   d. Compute the p-value $\hat{p}_l = 1 - \Phi(z)$
   e. If $\hat{p}_l < \alpha$, break.
   f. Update $\widehat{AIC}_1 = \widehat{aic}_l$.
   g. Print the final best model $M(\beta^2|\boldsymbol{X^2}, y)$ with $\widehat{aic}_l < \widehat{aic}_{l-1}$.
   h. End For l in $1, \ldots, k_1$.

Simulation study: We simulated five predictors to be informative variables out of $k \in \{10, 20, 30\}$ variables, and set $n = 300$. We generate all predictors from a k-variate normal distribution with mean 0 and independent covariance variances $\boldsymbol{\Sigma}_{ij} = 0, \forall\, i \neq j$. The logistic regression model is simulated using $y|\boldsymbol{X}_i \sim Binomial(1, \hat{\pi})$ with $\hat{\pi} = 1 / (1 + \exp(-W))$, $W = \boldsymbol{X}\beta$; $\beta = (\beta_1, \ldots, \beta_k)$ and

$$\beta_i = \begin{cases} 2, & j \in [1,2,3,4,5] \\ 0, & j \in [6,7, \ldots, k] \end{cases}.$$

We compare the proposed AIC testing variable selection procedure with the traditional "stepAIC" procedure in R and LASSO.[17,18] The following variable selection performance metrics were used:

Sensitivity is one minus the false deletion rate. For example, suppose that the true coefficients are [0, 1, 1, 1, 0] so that the best model would include predictors 2, 3, and 4, but that our model includes only predictors 3 and 4. Then the false deletion rate is 33% (1/3 of the good predictors were lost), so sensitivity is 67% (2/3 were kept) (10). Therefore,

$$Sensitivity = \frac{n(select_{\mathbf{x}})}{n(true_{\mathbf{x}})}$$

where $n(select_{\mathbf{x}})$ is the number of the selected informative variables, and $n(true_{\mathbf{x}})$ is the number of true informative variables.

Specificity is one minus the false inclusion rate. For example, suppose that the true model would include predictors 2, 3, and 4 (and exclude predictors 1 and 5), but our estimate contains 1, 2, 3, and 4. Then the false inclusion rate and the specificity rate would both be 50%, since of the two inactive predictors, one was included, and one was not (10).

$$Specificity = \frac{n(select'_{\mathbf{x}})}{n(true'_{\mathbf{x}})}$$

where $n(select'_{\mathbf{x}})$ is the number of the selected uninformative variables, and $n(true'_{\mathbf{x}})$ is the number of true uninformative variables.

Correct model rate (Accuracy) is the proportion of simulations in which exactly the correct subset was identified (i.e., in which sensitivity and specificity were both 100%). If a model is not correctly identified, then it is overfit (has too many predictors), underfit (too few), or misfit (the wrong ones) (10).

$$Accuracy = \frac{n[(Sensitivity = 1) \cup (Specificity = 1)]}{t},$$

where t is the number of repetitions set as 100.

We calculated the sensitivity, specificity and accuracy attained by each simulation method and then averaged them. In addition, we also compare the predictive performance of the best model of each method using the misclassification error rate (mer) (10). The mer is computed as

$$mer = \frac{c_1 + c_0}{n}$$

where c_1: represents the actual number of correct class "1" predicted as class "1", c_0: represents the actual number of correct class "0" predicted as class "0" (10). In addition, we compared the computational time of the proposed method with the existing methods. The computational time was computed in R using the function "system.time()".

Database: Our study did not require ethical board approval because it was based on publicly available information from the UCI machine learning repository.[19] The Wisconsin Breast Cancer Diagnostic Data Set is a widely used dataset in machine learning and cancer research. It was originally obtained from the University of Wisconsin Hospitals, Madison, by William H. Wolberg. The data set contains a total of 569 instances, with each instance representing a breast cancer biopsy sample. The data set includes 30 variables describing various characteristics of the cell nuclei present in the biopsy samples, such as mean radius, texture, smoothness, and so on. The full variable description can be found in Table 1. The dataset is divided into two classes: 357 benign (representing non-cancerous samples) and 212 malignant (representing cancerous samples).[19-21]

TABLE 1: Variable descriptions in the Wisconsin Breast Cancer Diagnostic Data Set.

Variable code	Variable name	Variable code	Variable name	Variable code	Variable name
X1	Radius mean	X11	Radius severity	X21	Radius worst
X2	Texture mean	X12	Texture severity	X22	Texture worst
X3	Perimeter mean	X13	Perimeter severity	X23	Perimeter worst
X4	Area mean	X14	Area severity	X24	Area worst
X5	Smoothness mean	X15	Smoothness severity	X25	Smoothness worst
X6	Compactness mean	X16	Compactness severity	X26	Compactness worst
X7	Concavity mean	X17	Concavity severity	X27	Concavity worst
X8	Concave points mean	X18	Concave points severity	X28	Concave points worst
X9	Symmetry mean	X19	Symmetry severity	X29	Symmetry worst
X10	Fractal dimension mean	X20	Fractal dimension severity	X30	Fractal dimension worst

The Wisconsin Breast Cancer Diagnostic Data Set variables are computed from digitized images of fine needle aspirates (FNA) of a breast mass. These images were captured using a CellScan system, which generates images of stained FNA slides at magnifications of 40x and 100x. The variables were then extracted from the images using image processing techniques, and statistical measures were calculated to characterize the cell nuclei in the biopsy samples.[19]

RESULTS

Table 2 presents the performances of the variable selection methods using the performance metrics defined earlier and the average computational time in seconds. The results represent the average (mean) of 100 simulation runs. Algorithm 1 was implemented in R for the Proposed AIC_{test}, the "stepAIC" function in R package "stat" for the StepAIC method, and R package "glmnet" for LASSO.

TABLE 2: Performance measures and average computational time (s) for the Proposed AIC_{test}, R StepAIC, and LASSO based on the average of 100 simulation runs.

k	Performance metrics	Method		
		Proposed AICtest	R StepAIC	LASSO
10	Sensitivity	1.000	1.000	1.000
	Specificity	1.000	0.784	0.822
	Accuracy	1.000	0.330	0.470
	MER	0.100	0.098	0.098
	Time (s)	0.034	0.082	0.138
20	Sensitivity	1.000	1.000	1.000
	Specificity	0.999	0.789	0.862
	Accuracy	0.990	0.020	0.280
	MER	0.101	0.092	0.096
	Time (s)	0.089	0.469	0.189
30	Sensitivity	1.000	1.000	1.000
	Specificity	0.997	0.794	0.869
	Accuracy	0.970	0.000	0.260
	MER	0.103	0.088	0.097
	Time (s)	0.170	1.429	0.301

AIC: Akaike Information Criterion.

Application to Wisconsin Breast Cancer Diagnostic Data Set: We applied the three procedures to identify the most informative variables that will accurately predict the two classes. The dataset was partitioned into ten folds train:test cross-validation. The three methods were compared based on the average number of selected variables (ASV), test misclassification error rate (TMER) and the top selected variables (TSV). The TSV is computed by calculating the total number of times a variable is selected out of the ten folds. Variables that are selected at least 5/10 are reported. Table 3 presents the results of the breast cancer data analysis.

TABLE 3: ASV, TMER, and TSV for the 10-fold cross-validation of the Wisconsin breast cancer diagnostic dataset.

Method	ASV	TMER	TSV
Proposed AIC_{test}	11	0.036	X5(10), X7(10), **X8(10)**, X9(10), X10(10), X12(10), X14(10), **X15(10)**, **X19(10)**, **X25(7)**, **X27(6)**, **X30(5)**.
R StepAIC	20	0.053	X6(10), X18(10), X1(9), X20(9), **X8(8)**, X16(8) X17(8), **X19(8)**, **X29(8)**, **X30(8)**, X4(7), X11(7), X13(7), X14(7), **X27(7)**, X9(6), X21(6), **X22(6)**, **X15(5)**, X24(5).
LASSO	9	0.053	X11(10), **X22(10)**, **X25(10)**, **X29(10)**, **X27(9)**, X28(9), X21(8), **X8(7)**, **X15(7)**, X24(5).

Red: Variables selected by all three methods, **Black:** Variables selected by two methods. The ASV is the ASV out of the 30 variables over the ten folds cross-validation. The TMER is the average misclassification error rate based on prediction using the test dataset in 10 folds cross-validation. The TSV is computed by calculating the total number of times a variable is selected out of the ten folds. Variables that are selected at least 5/10 are reported. ASV: Average number of selected variables; TMER: Test misclassification error rate; TSV: Top selected variables; AIC: Akaike Information Criterion.

DISCUSSION

The simulation results in Table 2 showed that the three methods are equally sensitive across various data dimensions. This implies that they can all correctly identify the five informative variables. In terms of specificity, the proposed AIC_{test} is the most specific in terms of correctly identifying the $k-5$ uninformative variables at various levels of k. This result also translated to having the highest model accuracy in identifying the correct model with the five informative variables. As expected, the predictive performance depends on the number of model parameters; the higher the number of parameters in the model, the lower the misclassification error rate. Therefore, the prediction results observed for the three methods are relatively similar. Overall, the AIC_{test} achieved the highest performance score of 3 out of the 4 criteria, making it the best among the three methods. The variable selection consistency performance of AIC_{test} across various levels of k also indicates that it can be easily adapted to $k \gg n$ high-dimensional situation, as the increase in dimension does not substantially degrades its performance. Similarly, the computational time comparison presented in Table 2 showed that the proposed AIC_{test} is the best in terms of the lowest average execution time to perform model selection.

Table 3 presents the average 10-fold cross-validation. The results are similar to the ones observed in the simulation study. The average number of selected variable results (ASV) showed that the proposed AIC_{test} and LASSO are more precise with fewer variables selected, while StepAIC identified more variables. In addition, in terms of prediction accuracy (1-TMER), the proposed AIC_{test} is the best, with 96.4% accuracy in correctly predicting benign and malignant breast cancer classes with an average of 11 variables. On the other hand, both StepAIC and LASSO achieved 94.7% accuracy with an average of 20 and 9 variables, respectively. We also used the TSV to measure the selection consistency of the methods. The proposed AIC_{test}, LASSO and StepAIC consistently identified 9/30, 4/30, and 2/30 variables with a selection probability of 1. This implies that the proposed AIC_{test} is more valid in identifying the most important variables under differ-

ent conditions. The top nine variables that were consistently selected by the proposed AIC_{test} are smoothness mean, concavity mean, concave point mean, symmetry mean, fractal dimension mean, texture severity, area severity, smoothness severity, and symmetry severity. Also, these selected variables were similarly identified in, where the same breast cancer dataset was used.[20]

The observed discrepancies in the selected variables by various methods, stemming from differences in the selection procedure, are in line with findings from previous studies.[3-9,20] These studies have also reported similar observations of overlapping variables with different inclusion probabilities across different variable selection techniques. One such study is the research conducted by Haq et al., which specifically focused on comparing multiple variable selection techniques using the breast cancer dataset.[20] Their study examined the performance of different methods and highlighted the presence of overlapping variables with varying inclusion probabilities across these methods. This consistency in findings across studies suggests that variable selection is a complex task influenced by the specific algorithm and selection criteria employed. The choice of the method can impact which variables are deemed important or relevant for inclusion in the model. These variations underscore the importance of carefully considering the specific goals and characteristics of the dataset when selecting a variable selection method.

The results from the simulations and real-life breast cancer datasets showed that the proposed AIC_{test} is better than the default method (stepAIC) as well as the LASSO in terms of variable selection. This strength is an important aspect of disease prognosis and diagnosis as it ensures the identification of the most informative variables needed to perform diagnosis or prognosis. However, the method is limited in terms of predicting the model outcome as the result observed is not much different from that of the existing methods.

CONCLUSION

The AIC-based variable selection approach proposed in this paper has several advantages. First, it accounts for both model complexity and goodness of fit, making it suitable for variable selection in high-dimensional datasets where the number of predictors is much larger than the sample size. Second, it incorporates statistical testing to assess the significance of individual predictors, providing a more robust and reliable approach for variable selection. Finally, the proposed method has the potential to improve the interpretability of the resulting models, which is critical in healthcare applications where interpretability is essential. In conclusion, breast cancer classification and prediction are challenging tasks that require accurate variable selection methods to build effective models. The AIC-based variable selection approach proposed is a promising method that integrates AIC with statistical testing for improved variable selection in breast cancer classification and prediction. Further research and validation of this approach in diverse breast cancer datasets are needed to establish its utility and generalizability in clinical practice.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Oyebayo Ridwan Olaniran;* ***Design:*** *Oyebayo Ridwan Olaniran;* ***Control/Supervision:*** *Oyebayo Ridwan Olaniran;* ***Data Collection and/or Processing:*** *Saidat Fehintola Olaniran;* ***Analysis and/or Interpretation:*** *Oyebayo Ridwan Olaniran;* ***Literature Review:*** *Saidat Fehintola Olaniran;* ***Writing the Article:*** *Oyebayo Ridwan Olaniran, Saidat Fehintola Olaniran;* ***Critical Review:*** *Oyebayo Ridwan Olaniran;* ***References and Fundings:*** *Oyebayo Ridwan Olaniran.*

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49. [Crossref] [PubMed]
2. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19(6):716-23. [Crossref]
3. Li X, Li Y, Yu X, Jin F. Identification and validation of stemness-related lncRNA prognostic signature for breast cancer. J Transl Med. 2020;18(1):331. [Crossref] [PubMed] [PMC]
4. Yu S, Hu C, Liu L, Cai L, Du X, Yu Q, et al. Comprehensive analysis and establishment of a prediction model of alternative splicing events reveal the prognostic predictor and immune microenvironment signatures in triple negative breast cancer. J Transl Med. 2020;18(1):286. [Crossref] [PubMed] [PMC]
5. Cheng J, Ren C, Liu G, Shui R, Zhang Y, Li J, et al. Development of high-resolution dedicated PET-based radiomics machine learning model to predict axillary lymph node status in early-stage breast cancer. Cancers (Basel). 2022;14(4):950. [Crossref] [PubMed] [PMC]
6. Liu H, Li J, Koirala P, Ding X, Chen B, Wang Y, et al. Long non-coding RNAs as prognostic markers in human breast cancer. Oncotarget. 2016;7(15):20584-96. [Crossref] [PubMed] [PMC]
7. Li N, Yu K, Lin Z, Zeng D. Identifying a cervical cancer survival signature based on mRNA expression and genome-wide copy number variations. Exp Biol Med (Maywood). 2022;247(3):207-20. [Crossref] [PubMed] [PMC]
8. Chen R, Qi Y, Huang Y, Liu W, Yang R, Zhao X, et al. Diagnostic value of core needle biopsy for determining HER2 status in breast cancer, especially in the HER2-low population. Breast Cancer Res Treat. 2023;197(1):189-200. [Crossref] [PubMed] [PMC]
9. Zhang J, Zhang Z, Mao N, Zhang H, Gao J, Wang B, et al. Radiomics nomogram for predicting axillary lymph node metastasis in breast cancer based on DCE-MRI: a multicenter study. J Xray Sci Technol. 2023;31(2):247-63. [Crossref] [PubMed]
10. Olaniran OR, Olaniran SF, Popoola J, Omekam IV. Bayesian additive regression trees for predicting colon cancer: methodological study (validity study). Turkiye Klinikleri Journal of Biostatistics. 2022;14(2):103-9. [Crossref]
11. Banjoko AW, Yahya WB, Garba MK, Olaniran OR, Dauda KA, Olorede KO. Efficient support vector machine classification of diffuse large b-cell lymphoma and follicular lymphoma MRNA tissue sampless. annals. Computer Science Series. 2015;13(2):69-79. [Link]
12. Olaniran OR, Abdullah MAA. Gene selection for colon cancer classification using bayesian model averaging of linear and quadratic discriminants. Journal of Science and Technology. 2017;9(3):140-4. [Link]
13. Olaniran OR, Abdullah MAA. BayesRandomForest: an R implementation of bayesian random forest for regression analysis of high-dimensional data. Romanian Statistical Review. 2018;66(1):95-102. [Link]
14. Olaniran OR, Abdullah MAA. Bayesian variable selection for multiclass classification using Bootstrap Prior Technique. Austrian Journal of Statistics. 2019;48(2):63-72. [Crossref]
15. Olaniran OR, Abdullah MAA. Bayesian analysis of extended cox model with time-varying covariates using bootstrap prior. Journal of Modern Applied Statistical Methods. 2020;18(2):7. [Crossref]
16. Olaniran OR, Yahya WB. Bayesian hypothesis testing of two normal samples using bootstrap prior technique. Journal of Modern Applied Statistical Methods. 2017;16(2):618-38. [Crossref]
17. Garofoli R, Resche-Rigon M, Roux C, van der Heijde D, Dougados M, Moltó A. Machine-learning derived algorithms for prediction of radiographic progression in early axial spondyloarthritis. Clin Exp Rheumatol. 2023;41(3):727-34. [Crossref] [PubMed]
18. Engebretsen S, Bohlin J. Statistical predictions with glmnet. Clin Epigenetics. 2019;11(1):123. [Crossref] [PubMed] [PMC]
19. Wolberg WH, Street N, Mangasarian OL. Wisconsin diagnostic breast cancer (wdbc). U. o. California, Ed., ed. USA. 1995. [Link]
20. Haq AU, Li JP, Saboor A, Khan J, Wali S, Ahmad S, et al. Detection of breast cancer through clinical data using supervised and unsupervised variable selection techniques. IEEE Access. 2021;9:22090-105. [Crossref]
21. Dua D, Graff C. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science; 2019. [Link]