Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:120

https://doi.org/10.28919/cmbn/9534

ISSN: 2052-2541

MULTIVARIATE ADAPTIVE BIVARIATE REGRESSION SPLINES (MABRS) BINARY RESPONSE FOR MODELING STROKE AND HYPERTENSION IN RSKD DADI CITY MAKASSAR

SRI SULASTRI, BAMBANG WIDJANARKO OTOK*, ACHMAD CHOIRUDDIN

Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Kampus ITS-Sukolilo, Surabaya 60111, Indonesia

Abstract: Classification of health conditions with more than one correlated response variable is an important challenge in medical data analysis. This study proposes a Multivariate Adaptive Bivariate Regression Splines (MABRS) approach to classify stroke type (ischemic vs. hemorrhagic) and hypertension status simultaneously. Utilizing clinical data from stroke patients, the MABRS model was built based on the optimal parameter combination for each response. The results showed that the stroke type classification achieved a fairly good performance (accuracy 76.82%), with the most influential variables being obesity, hypercholesterolemia, and diabetes mellitus. In contrast, the hypertension classification model performed poorly (51.21% accuracy), although diabetes mellitus, gender, and age were identified as the main predictors. The MABRS approach has the advantage of capturing nonlinear relationships and interactions between predictor variables, while considering the correlation between two binary response variables in a unified model framework. These findings confirm the potential of MABRS in uncovering complex relationships between clinical variables and supporting data-driven medical decision-making, particularly in the management of comorbidities in stroke patients.

Keywords: MABRS; stroke; hypertension; binary response.

2020 AMS Subject Classification: 62G08, 62J12.

*Corresponding author

E-mail address: bambang_wo@its.ac.id

Received August 05, 2025

1. INTRODUCTION

Stroke is a condition characterized by focal or global cerebral dysfunction lasting 24 hours or more, which can lead to death caused by spontaneous bleeding or inadequate blood supply to brain tissue [1]. There are two types of stroke disease status, namely ischemic stroke and hemorrhagic stroke. Ischemic stroke is characterized by too little blood to supply enough oxygen and nutrients to the brain, while hemorrhagic stroke is characterized by too much blood in the closed cranial cavity. There are 80% of all strokes are ischemic strokes, and 20% are hemorrhagic strokes [2].

According to the World Health Organization (WHO), stroke is the second leading cause of death and the third leading cause of disability in the world (WHO, 2020). According to the results of the 2013 Riskesdas report, the prevalence of stroke in Indonesia in 2007 was 8.3% and then in 2013 it increased to 12.1%. According to information from Riskesdas 2018, there are 713,783 stroke victims aged 15 or 10.9% of the population each year. Most stroke victims are over 75 years old or 50.2%. This shows that the prevalence of stroke in Indonesia is still high based on Riskesdas statistics from 2007, 2013, and 2018.

Hypertension is a disease that occurs when blood pressure increases beyond normal limits. Hypertension is commonly called The Silent Killer because the symptoms are not visible and if not given treatment it can cause complications such as stroke, blood vessels and heart disease which can lead to disability and even death [3]. Riskesdas data in 2013, people with hypertension in Indonesia were around 25.8%, then in 2018 people with hypertension increased to 34.11%. The number of people with hypertension continues to increase every year. It is estimated that by 2025 there will be 1.5 billion people who are famous for hypertension, and it is estimated that every year 9.4 million people die from hypertension and its complications.

Based on the high rates of both diseases, it shows that stroke and hypertension are very serious diseases and should be taken seriously. Predictive models play an important role in identifying individuals at high risk of stroke and hypertension, thus supporting informed clinical decision-making. Many predictive models have been proposed assessing and quantifying stroke

and hypertension risk factors, such as Yang et.al [4] developed a prediction model of stroke risk in hypertensive patients using the XGBoost machine learning algorithm. Meanwhile Zheng et.al [5] built a logistic regression-based prediction model which showed that classical logistic regression could be comparable or better than machine learning.

Binary response Multivariate Adaptive Regression Splines (MARS) has been widely applied in data analysis in recent years. This method can overcome high-dimensional problems such as having predictor variables that tend to be many, as well as large sample sizes [6]. The MARS method is built with a stepwise algorithm consisting of forward stepwise and backward stepwise. Forward stepwise builds a model by adding truncated spline basis functions (knots and interactions) until a model with the maximum number of basis functions is obtained and backward stepwise is used to obtain a parsimony model by selecting the forward stepwise basis function whose contribution is most significant to the estimated response based on the minimum Generalized Cross Validation (GCV) value.

MARS modeling has been widely applied by various researchers, such as [7] who used MARS for transportation energy demand prediction, and [8] who developed a combination of MARS and Random Forest in the assessment of pile driving ability. Other developments were carried out through the integration of MARS with PLS-SEM [9] and the Inverse Gaussian approach to MARS [10]. However, in many real cases, there is a correlation between two response variables, which prompts the need to develop a MARS model with two responses. Although some studies, such as [11] and [12], have applied MARS with two response variables, the approach is still limited to quantitative responses. In fact, in practice, binary response variables are often encountered, so the existing continuous biresponse MARS model has not been able to optimally handle cases with binary responses. Therefore, this study aims to model stroke and hypertension in Dadi General Hospital, Makassar City, South Sulawesi Province using Multivariate Adaptive Bivariate Regression Splines (MABRS) on binary responses.

2. PRELIMINARIES

This section provides information about the dataset we used for the analysis and some literature

reviews for modeling the data.

2.1 Data

The data used were medical records of patients with stroke and hypertension who underwent hospitalization at Dadi Makassar Hospital based on predetermined criteria consisting of 271 patients. Table 1 provides an operational definition of the research variables.

Table 1. Operational Definition of Variables

Variable Name	Scale	Description
Type of stroke (Y_1)	Nominal	0 = Ischemic Stroke
		1 = Hemorrhagic Stroke
Hypertension (Y_2)	Nominal	0 = No Hypertension
		1 = Hypertension
Age (X_1)	Ratio	Patient's Age
Education Level (X_2)	Nominal	0 = Low Education
		1 = Higher Education
Gender (X_3)	Nominal	0 = Male
		1 = Female
Urid Acid (X_4)	Nominal	0 = No Urid Acid
		1 = Urid Acid
Cholesterol (X_5)	Nominal	0 = No Cholesterol
		1 = Cholesterol
Diabates Mellitus (X_6)	Nominal	0 = No Diabetes Mellitus
		1 = Diabetes Mellitus
Obesity (X_7)	Nominal	0 = No Obesity
		1 = Obesity
History of Heart Disease (X_8)	Nominal	0 = No History
		1 = History

2.2 Bivariate Binary Logistic Regression

Bivariate binary logistic regression is a logistic regression model that has two response variables where each response variable takes two categories (binary). Each binary response variable is associated with a certain number of independent variables. If there are interrelated

bivariate random variables (Y_1, Y_2) where variables Y_1 and Y_2 express an event 'success' or 'failure' then the event can be modeled with bivariate binary logistic regression. The marginal probabilities for each response variable are denoted by:

$$\pi_1 = P(Y_1 = 1) \text{ and } P(Y_2 = 1)$$

If there are k predictor variables, $x_1, x_2, \dots, x_k$ then the values of $\pi_1(x), \pi_2(x), \dots, \pi_k(x)$ are as follows:

$$\pi_1(\boldsymbol{x}) = \frac{exp\ (\beta_{01} + \beta_{11}x_1 + \cdots + \beta_{k1}x_k)}{1 + exp\ (\beta_{01} + \beta_{11}x_1 + \cdots + \beta_{k1}x_k)}$$
$$\pi_2(\boldsymbol{x}) = \frac{exp\ (\beta_{02} + \beta_{12}x_1 + \cdots + \beta_{k2}x_k)}{1 + exp\ (\beta_{02} + \beta_{12}x_1 + \cdots + \beta_{k2}x_k)} \quad (1)$$

The bivariate binary logistic regression model is expressed by the logit $\pi_1(\boldsymbol{x})$ and logit $\pi_2(\boldsymbol{x})$ equations as linear functions of $\boldsymbol{\beta}_1^T\boldsymbol{x}, \boldsymbol{\beta}_2^T\boldsymbol{x}$ and log $\Psi = \Theta$, where $\boldsymbol{\beta}_1 = [\beta_{01}, \beta_{11}, \dots, \beta_{k1}]^T$, $\boldsymbol{\beta}_2 = [\beta_{02}, \beta_{12}, \dots, \beta_{k2}]^T$ and $\boldsymbol{x} = [x_0, x_1, \dots, x_k]^T$.

By taking n mutually independent random samples, the bivariate binary random variable (Y_{1i}, Y_{2i}) where $i = 1,2, \dots, n$ will be identical to $(Y_{11}, Y_{10}, Y_{01}, Y_{00})$, binomially distributed with probability values $\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}$.

The likelihood function of a bivariate random variable is as follows:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(Y_{11} = y_{11}, Y_{10} = y_{10}, Y_{01} = y_{01}, Y_{00} = y_{00})$$
$$= \prod_{i=1}^{n} {\pi_{11}}^{y_{11}}\ {\pi_{10}}^{y_{10}} {\pi_{01}}^{y_{01}} {\pi_{00}}^{y_{00}} \quad (2)$$

Log-Likelihood function

$$\ln L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(Y_{11} = y_{11}, Y_{10} = y_{10}, Y_{01} = y_{01}, Y_{00} = y_{00})$$
$$= \sum_{i=1}^{n} (y_{11} \ln \pi_{11} + y_{10} \ln \pi_{10} + y_{01} \ln \pi_{01} + y_{00} \ln \pi_{00}) \quad (3)$$

Calculating the first derivative

$$\frac{\partial lnL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ \frac{y_{11}}{\pi_{11}} \frac{\partial \pi_{11}}{\partial \boldsymbol{\beta}} + \frac{y_{10}}{\pi_{10}} \frac{\partial \pi_{10}}{\partial \boldsymbol{\beta}} + \frac{y_{01}}{\pi_{01}} \frac{\partial \pi_{01}}{\partial \boldsymbol{\beta}} + \frac{y_{00}}{\pi_{00}} \frac{\partial \pi_{00}}{\partial \boldsymbol{\beta}} \right\} \quad (4)$$

Furthermore, solving the equation using the iteration procedure. The iteration method used is Newton-Raphson Iteration [13]. The principle of the Newton Raphson method is to find the value of $\widehat{\boldsymbol{\beta}}$ by updating the iteration process against $\boldsymbol{\beta}$ with the following equation [14].

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\boldsymbol{H}(\boldsymbol{\beta})^{(t)}\right)^{-1} \boldsymbol{g}(\boldsymbol{\beta})^{(t)} \quad (5)$$

2.3 Multivariate Adaptive Regression Splines (MARS)

MARS is a flexible and adaptive approach in non-linear regression modeling. The model can capture the complex relationship between predictor and response variables by constructing adaptively generated basis functions. The following equation describes the MARS basis function used to predict the response value:

$$y_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)i} - t_{km} \right) \right]_+ + \varepsilon_i \ , i = 1,2, \dots, n \quad (6)$$

Where α_0 is a constant basis function parameters, α_m is non-constant $m-th$ basis function paramater, m is a number of non-constant basis functions, K_m is a maximum interaction of the $m-th$ basis function, s_{km} is a sign of the basis function in basis function m and in the $k-th$ iteration, $x_{v(k,m)}$ is the $v-th$, predictor variable, where is v the variable indekx of the predictor corresponding to the $k-th$ interaction and the $m-th$ base In the MARS function, and t_{km} is a knot value at $k-th$ interaction and $m-th$ basis function.

Where:

$\boldsymbol{x}_i = (x_{1i}, x_{2i}, \dots, x_{ki})^T \quad ; i = 1,2, \dots, n$

$$\boldsymbol{B}_{mi}(\boldsymbol{x}, \boldsymbol{t}) = \prod_{k=1}^{K_m} \left[s_{km} \left(x_{v(k,m)i} - t_{km} \right) \right]_+, \quad (7)$$

With:

$$x_{v(k,m)i} \in \left\{ x_{ji} \right\}_{j=1}^{k} = \{ x_{1i}, x_{2i}, \dots, x_{ki} \} \text{ and} \quad (8)$$

$t_{km} \in \left\{ x_{v(k,m)i} \right\}_{i=1}^{n} = \left\{ x_{v(k,m)1}, x_{v(k,m)2}, \dots, x_{v(k,m)n} \right\}$

When $s_{km} = +1$, then

$$\left[+\left(x_{v(k,m)i} - t_{km}\right)\right]_+ = \begin{cases} x_{v(k,m)i} - t_{km}, \text{if } x_{v(k,m)i} > t_{km} \\ 0, \quad \text{more} \end{cases} \tag{9}$$

When $s_{km} = -1$, then

$$\left[-\left(x_{v(k,m)i} - t_{km}\right)\right]_+ = \begin{cases} t_{km} - x_{v(k,m)i}, \text{if } t_{km} > x_{v(k,m)i} \\ 0, \quad \text{more} \end{cases} \tag{10}$$

When combined, they form:

$$B_m(\boldsymbol{x}, \boldsymbol{t}) = \left[max\left(s_{km}\left(x_{v(k,m)i} - t_{km}\right), 0\right)\right]_+$$

The MARS equation that expresses the relationship between k predictors and a single continuous response involving n observations obtained based on equation (6) can be expressed in matrix form as follows:

$$\boldsymbol{y} = \mathbf{B}(\boldsymbol{x}, \boldsymbol{t})\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{11}$$

$$\mathbf{B}(\boldsymbol{x}, \boldsymbol{t}) = \begin{bmatrix} 1 & \prod_{k=1}^{K_1}\left[s_{k1}\left(x_{v(k,1)1} - t_{k1}\right)\right]_+ & \dots & \prod_{k=1}^{K_M}\left[s_{kM}\left(x_{v(k,M)1} - t_{kM}\right)\right]_+ \\ 1 & \prod_{k=1}^{K_1}\left[s_{k1}\left(x_{v(k,1)2} - t_{k1}\right)\right]_+ & \dots & \prod_{k=1}^{K_M}\left[s_{kM}\left(x_{v(k,M)2} - t_{kM}\right)\right]_+ \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \prod_{k=1}^{K_1}\left[s_{k1}\left(x_{v(k,1)n} - t_{k1}\right)\right]_+ & \dots & \prod_{k=1}^{K_M}\left[s_{kM}\left(x_{v(k,M)n} - t_{kM}\right)\right]_+ \end{bmatrix}$$

The response vector y in equation (11) is a response vector of size $(n \times 1)$ and $\mathbf{B}(\boldsymbol{x}, \boldsymbol{t})$ is a matrix of order $\left(n \times (M+1)\right)$. $\boldsymbol{\alpha}$ is a vector containing regression of size $((M+1) \times 1)$. $\boldsymbol{\varepsilon}$ is a vector of random errors of size $(n \times 1)$ which is assumed to be mutually independent normally distributed with $E(\boldsymbol{\varepsilon}) = 0$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

$$\underset{\boldsymbol{\alpha} \in R^{M+1}}{Min}\ (\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}) = \underset{\boldsymbol{\alpha} \in R^{M+1}}{Min}\left((\boldsymbol{y} - \mathbf{B}(\boldsymbol{x}, \boldsymbol{t})\boldsymbol{\alpha})^T\left((\boldsymbol{y} - \mathbf{B}(\boldsymbol{x}, \boldsymbol{t})\boldsymbol{\alpha}\right)\right) \tag{12}$$

The optimization solution of equation (12) is obtained by partially deriving the translation of $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$ with respect to $\boldsymbol{\alpha}$ and making the equation equal to zero, thus obtaining an estimator for $\boldsymbol{\alpha}$ as:

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{B}^T(\boldsymbol{x}, \boldsymbol{t})\mathbf{B}(\boldsymbol{x}, \boldsymbol{t})\right)^{-1}\mathbf{B}^T(\boldsymbol{x}, \boldsymbol{t})\boldsymbol{y} \tag{13}$$

The estimator form of the continuous response MARS regression function can then be obtained based on equation (13) with Eq:

$$\begin{aligned}\hat{f}(\boldsymbol{x}) &= \mathbf{B}(\boldsymbol{x},\boldsymbol{t})\hat{\boldsymbol{\alpha}} \\ &= \mathbf{B}(\boldsymbol{x},\boldsymbol{t})\left(\mathbf{B}^T(\boldsymbol{x},\boldsymbol{t})\mathbf{B}(\boldsymbol{x},\boldsymbol{t})\right)^{-1}\mathbf{B}^T(\boldsymbol{x},\boldsymbol{t})\boldsymbol{y} \\ &= \mathbf{S}(.)\boldsymbol{y}\end{aligned} \tag{14}$$

The matrix $\mathbf{S}(.)$ in equation (14) is a hat matrix that depends on the matrix $\mathbf{B}(\boldsymbol{x},\boldsymbol{t})$ which contains the basis function $B_1(\boldsymbol{x},\boldsymbol{t}), B_2(\boldsymbol{x},\boldsymbol{t}), \ldots, B_M(\boldsymbol{x},\boldsymbol{t})$. The selection of basis functions must be done optimally to get a MARS estimator that fits the data. A stepwise procedure (forward and backward) is performed to determine the optimal basis function based on the minimum GCV value.

MARS is also a modern statistical classification method that utilizes the flexibility of the model and estimates the distribution within each class, ultimately providing a clustering rule. Classification in MARS is based on the logistic regression approach. Thus, the logit relationship function in the MARS model is as follows.

$$logit\ \pi(\boldsymbol{x}) = \alpha_0 + \sum_{m=1}^{M}\alpha_m \prod_{k=1}^{K_m}\left[s_{km}\left(x_{v(k,m)i} - t_{km}\right)\right]_+ + \varepsilon_i \tag{15}$$

The best model selection in the MARS model uses the smallest GCV value. Friedman used the GCV method in a backward stepwise procedure to select the optimal basis function in the MARS algorithm.

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}(\boldsymbol{x}_i)\right]^2}{\left[1 - \frac{\tilde{C}(M)}{n}\right]^2} = \frac{n\sum_{i=1}^{n}\left[y_i - \hat{f}(\boldsymbol{x}_i)\right]^2}{\left(n - \tilde{C}(M)\right)^2} \tag{16}$$

2.4 Multivariate Adaptive Bivariate Regression Splines (MABRS)

Multivariate Adaptive Bivariate Regression Splines (MABRS) Binary response is a development of Multivariate Adaptive Regression Splines (MARS) designed to handle two binary response variables simultaneously in one model. This model combines MARS with bivariate binary logistic regression, so as to capture non-linear relationships and interactions between predictor variables flexibly.

$$logit\left(\frac{\pi_\ell(x)}{1-\pi_\ell(x)}\right) = \alpha_0^{(\ell)} + \sum_{m=1}^{M} \alpha_m^{(\ell)} \mathbf{B}_m(\boldsymbol{x}, t); \qquad \ell = 1,2\ ;\ i = 1,2,\dots,n \qquad (17)$$

Where $logit\left(\frac{\pi_\ell(x)}{1-\pi_\ell(x)}\right)$ is the logit function that links the probability of occurrence of the ℓ -th response.

2.5 Stages of Analysis

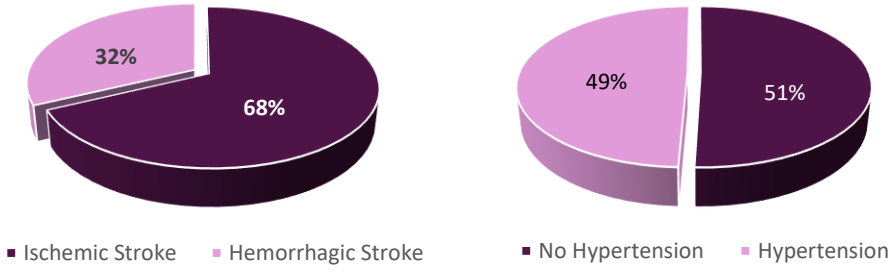
This research followed the following stages of data analysis

a. Exploring data from response and predictor variables to determine the descriptive statistics of each variable.
b. Testing the relationship between response variables using the chi-square test.
c. Divide the data into two parts, namely in-sample data (70% of the total data) and out-sample data (30% of the total data).
d. Perform a combination of BF, MI, and MO to determine the best model for each response.
e. Form a basis function for each response variable.
f. Estimating the response variable parameters
g. Testing the level of importance of predictor variables on each response.
h. Evaluate the accuracy of classification using accuracy, sensitivity, and specificity.

3. MAIN RESULTS

3.1 Descriptive Analysis

Descriptive analysis was conducted to obtain an overview of the research data.



(a) (b)

Figure 1. Response Variable Percentage (a) Type of Stroke and (b) Hypertension

Of the 271 patients recorded at Dadi General Hospital in Makassar City in 2022, the percentage of ischemic stroke was 68% or 184 patients, while the percentage of hemorrhagic stroke was 32% or 87 patients. While the percentage of patients who were not affected by hypertension was 51% or 138 patients, while the percentage of patients affected by hypertension was 49% or 132 patients.

The age of patients ranged from 13 years to 88 years, with an average of 59 years. Of the total data, patients had a high level of education 67% and a low level of education 33%. Male and female patients were balanced at 50%. A history of gout was present in 13%, while 39% had hypercholesterolemia, 29% had diabetes mellitus, 17% were obese, and 11% had a history of heart disease.

3.2 Correlation Test

The results of the correlation test between response variables are as follows:

Table 2. Correlation test between response variables

χ^2	df	$p-value$
2.8776	1	0.08982

Based on Table 2, a p-value of 0.08982 was obtained, indicating a statistically strong association between stroke type and hypertension at a significant 10% level. In addition to correlation tests, the relationship between response variables is also visualized using scatter plots against predictor variables.

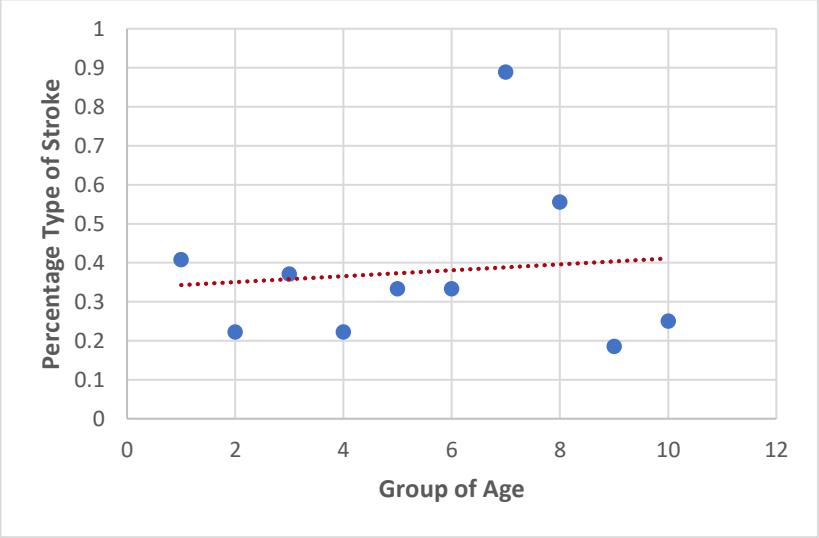


Figure 2. Scatter Plot for The Group of Age Against The Type of Stroke

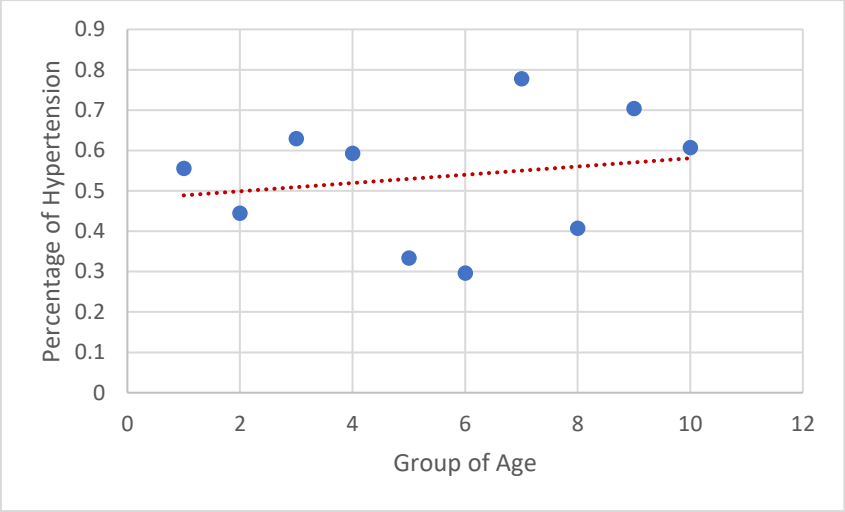


Figure 3. Scatter Plot for The Group of Age Against the Hypertension

Figures 2 and 3 show that stroke type (Y_1) and hypertension (Y_2) have a nonlinear relationship with the age variable X_1 (age). Meanwhile, other categorical variables cannot be visualized using scatter plots.

3.3 Binary Response MABRS Analysis

Three things need to be considered in the process of forming a MARS model, namely the basis function (BF), maximum interaction (MI), and minimum observation (MO). Basis functions (BF) generally use two to four times the number of predictor variables, MI is one, two or three with consideration if more than three will produce a very complex model and MO is zero, one, two, and three [15]. The following are the results of the combination of all BF, MI, and MO.

Table 3. Combination BF, MI, and MO

Model	Combination			Y1		Y2	
	BF	MI	MO	GCV	R^2	GCV	R^2
1	16	1	0	0.1235231	0.439628	0.2124247	0.149133
2		1	1	0.1237292	0.438693	0.2124247	0.149133
3		3	0	0.1081381	0.509423	*0.1963922	0.213352
4		3	1	*0.1081136	0.509534	0.1963922	0.213352
5	24	1	2	0.1237292	0.438693	02124247	0.149133
6		1	3	0.1239359	0.437755	0.2124247	0.149133
7		2	0	0.1129386	0.487645	0.1999570	0.199073
8		2	1	0.1128938	0.487848	0.1997363	0.199957

9	32	2	2	0.1128938	0.487848	0.1996530	0.200290
10		2	3	0.1128938	0.487848	0.1999570	0.199073
11		3	0	0.1081381	0.509423	0.1963922	0.213352
12		3	1	0.1081136	0.509534	0.1963922	0.213352

* Smallest GCV of 36 models formed

Based on Table 3, the best model is selected for variable Y_1 namely the combination of BF = 16, MI = 3, and MO = 1 with the GCV value which tends to be the minimum of 0.3045058 and the R^2 value which tends to be the maximum. In variable Y_2 the combination of BF = 16, MI = 3, and MO = 0 with a GCV value that tends to be the minimum of 0.1963922 and R^2 value that tends to be the maximum.

Estimation in MABRS modeling is done simultaneously on both response variables. Thus, this approach considers the covariance structure between the two responses, and produces two regression parameter estimation systems simultaneously, for variables Y_1 and Y_2, respectively, in one bivariate model framework.

Table 4. Parameter Estimation of Binary Response MABRS (Type of Stroke)

Response Variable	Basis Function (BF)	Parameter Estimation
Y_1	Intercept	0.22416627
	X_7	- 17.4026124
	X_5	- 3.43947382
	X_6	- 2.94352692
	$X_4 * X_7$	36.7578802
	$X_5 * X_6$	2.69644944
	$X_2 * X_7$	22.1247575
	$h(X1 - 74)$	0.14914755
	$h(74 - X_1)$	0.02096498
	$X_2 * X_4 * X_7$	9.99404687
	X_8	1.75296823
	$X_6 * X_8$	3.77462549

Based on the results of the parameters estimator in Table 4, the Binary Response MABRS model for the Stroke Type response is obtained, as follows:

$$\hat{\pi}_1(x) = 0.22416627 - 17.4026124BF_1 - 3.43947382BF_2 - 2.94352692BF_3 + 36.7578802BF_4 + 2.69644944BF_5 + 22.1247575BF_6 + 0.14914755BF_7 - 0.02096498BF_8 - 9.99404687BF_9 - 1.75296823BF_{10} + 3.77462549BF_{11}$$

$BF_4 = X_4 * X_7$

$BF_5 = X_5 * X_6$

$BF_6 = X_2 * X_7$

$BF_7 = h(X1 - 74)$

$BF_8 = h(74 - X_1)$

$BF_9 = X_2 * X_4 * X_7$

$BF_{11} = X_6 * X_8$

The combination results of BF = 16, MI = 3, and MO = 1 show that there are 7 variables that affect the classification of stroke types, namely age (X_1), education level (X_2), uric acid (X_4), cholesterol (X_5), diabetes mellitus (X_6), obesity (X_7), and heart disease (X_8). These seven variables have a good influence on the model, both individually and when interacting with other variables. Based on the modeling results obtained, where gender (X_3) has no influence on the model, it can be concluded that gender does not affect the classification of stroke patients. The basis function formed also shows the interaction between several predictor variables and the presence of a cut-off point at the age of 74 years. Identify the variables that make the most significant contribution in the formation of the stroke type classification model.

Table 5. Level of Importance (Type of Stroke)

Type of Stroke	
Variable	**Variable**
X_7	100.0%
X_5	65.8%
X_6	54.7%
X_2	46.0%
X_4	46.0%
X_1	25.8%
X_8	21.4%

Furthermore, for the second response variable, hypertension (Y_2)

Table 6. Parameter Estimation of Binary Response MABRS (Hypertension)

Response Variable	Basis Function (BF)	Parameter Estimation
Y_2	Intercept	- 0.13602919
	X_6	- 1.39730838
	$X_3 * X_6$	- 0.96893194
	$X_5 * X_6$	- 13.0991641
	$h(X_1 - 61)$	0.25015284
	$h(61 - X_1)$	0.07706338
	$X_2 * X_5 * X_6$	15.6696398
	X_5	- 0.31927988
	$h(X_1 - 61) * X_3$	- 0.27291480
	$h(X_1 - 61) * X_5$	0.19196235
	$h(X_1 - 68)$	- 0.15932668
	$h(61 - X_1) * X_5$	- 0.06358357

Based on the results of estimating the parameters in Table 6, the Binary Response MABRS model for the Hypertension response is obtained, as follows:

$$\hat{\pi}_2(x) = -0.13602919 - 1.39730838BF_1 - 0.96893194BF_2 - 13.0991641BF_3 + 0.25015284BF_4 + 0.07706338BF_5 + 15.6696398BF_6 - 0.31927988BF_7 - 0.27291480BF_8 + 0.19196235BF_9 - 0.15932668BF_{10} - 0.06358357BF_{11}$$

$BF_2 = X_3 * X_6$

$BF_3 = X_5 * X_6$

$BF_4 = h(X_1 - 61)$

$BF_5 = h(61 - X_1)$

$BF_6 = X_2 * X_5 * X_6$

$BF_8 = h(X_1 - 61) * X_3$

$BF_9 = h(X_1 - 61) * X_5$

$BF_{10} = h(X_1 - 68)$

$BF_{11} = h(61 - X_1) * X_5$

The combination result of BF = 16, MI = 3, and MO = 0 shows that there are 5 variables that affect the classification of hypertension, namely age (X_1), education level (X_2), gender (X_3), cholesterol (X_5), diabetes mellitus (X_6). These five variables have a good influence on the model, both individually and when interacting with other variables. Based on the modeling results obtained, where uric acid (X_5), obesity (X_7) dan heart disease (X_8) have no influence on the model, it can be concluded that these three variables do not affect the classification of hypertension. The basis function formed also shows the interaction between several predictor variables and the presence of a cut-off point at the age of 61 and 68 years.

Identify the variables that make the most significant contribution in the formation of hypertension classification models.

Table 7. Level of Importance (Hypertension)

Diagnosis of Hypertension	
Variable	**Level of Importance**
X_6	100.0%
X_3	84.9%
X_1	78.7%
X_2	66.7%
X_5	66.7%

After the parameter estimates for both response variables are obtained, the next step is to evaluate the performance of the classification model through measuring accuracy, sensitivity, and specificity. The results of the evaluation are presented in Table 8.

Table 8. Classification Accuracy Results

$\boldsymbol{Y_1}$		$\boldsymbol{Y_2}$	
Accuracy	76.82%	Accuracy	51.21%
Sensitivity	87.71%	Sensitivity	51.28%
Specificity	52.00%	Specificity	51.16%

The classification model for variable Y_1 type of stroke showed a fairly good performance with an accuracy of 76.82%, which means that the model was able to correctly classify the type of stroke in about 77% of cases. The sensitivity of 87.71% indicates that the model is very good at recognizing hemorrhagic stroke cases (positive), i.e. it is able to detect most patients who actually have hemorrhagic stroke. Meanwhile, the specificity of 52.00% reflects the model's ability to identify ischemic stroke cases (negative), although it is still relatively low, so there are quite a number of ischemic cases that are misclassified as hemorrhagic. In contrast, the classification model for variable Y_2 hypertension showed less than optimal performance, with an accuracy of 51.21%, meaning the model was only slightly better than random guessing in distinguishing hypertensive and non-hypertensive patients. The sensitivity of 51.28% indicates that the model is only able to recognize about half of the patients who actually have hypertension, while the specificity of 51.16% indicates that the model's ability to identify non-hypertensive patients is also low. This indicates that the model for Y_2 has not been able to capture an effective classification pattern between hypertensive and non-hypertensive patients.

4. CONCLUSION

This study successfully built a Multivariate Adaptive Bivariate Regression Splines (MABRS) model to classify two binary response variables simultaneously, namely stroke type and hypertension. The best model for stroke type was obtained from the combination of parameters BF = 16, MI = 3, and MO = 1, while for hypertension from the combination of BF = 16, MI = 3, and MO = 0. Estimation was done bivariate so as to consider the correlation between the two responses. Modeling results showed that stroke type classification was influenced by seven variables, with the most important variables being obesity (X_7), cholesterol (X_5), and diabetes mellitus (X_6). This model has good classification performance with 76.82% accuracy, 87.71% sensitivity, and 52.00% specificity. Meanwhile, hypertension classification is influenced by five variables, with the most important variables being diabetes mellitus (X_6), gender (X_3), and age (X_1). However, the performance of the hypertension model was still low (accuracy 51.21%, sensitivity 51.28%, specificity 51.16%). Overall, the MABRS approach is effective in identifying the main

determinants of stroke type, but further improvements are needed in the hypertension classification model, especially in overcoming data imbalance and improving the discriminative ability of the model. Future research can focus on improving the accuracy of hypertension classification, one of which is by integrating patient spatial information and applying a spatial point process approach estimated through regularization techniques, as developed in a recent study [16].

ACKNOWLEDGMENTS

The authors would like to thank the research funding support provided by the “Master's Education Program Towards Doctorate for Excellent Graduates (PMDSU)” which is a program of the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia. Through Grant Number 2158/E4/DT.04.02/2024.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

REFERENCES

[1] Y. Haiga, I. Prima Putri Salman, S. Wahyuni, Perbedaan Diagnosis Stroke Iskemik dan Stroke Hemoragik Dengan Hasil Transcranial Doppler di RSUP Dr. M. Djamil Padang, Sci. J. 1 (2022), 391-400. https://doi.org/10.56260/sciena.v1i5.72.

[2] A. Nurkhafifah, Determinants of Quality of Life of Stroke Patients in RSKD Dadi, Makassar City, Thesis, Hasanuddin University, 2022. https://repository.unhas.ac.id/id/eprint/18805.

[3] Y. Aprillia, Gaya Hidup dan Pola Makan Terhadap Kejadian Hipertensi, J. Ilm. Kesehat. Sandi Husada 12 (2020), 1044-1050. https://doi.org/10.35816/jiskh.v12i2.459.

[4] Y. Yang, J. Zheng, Z. Du, Y. Li, Y. Cai, Accurate Prediction of Stroke for Hypertensive Patients Based on Medical Big Data and Machine Learning Algorithms: Retrospective Study, JMIR Med. Inform. 9 (2021), e30277. https://doi.org/10.2196/30277.

[5] X. Zheng, F. Fang, W. Nong, D. Feng, Y. Yang, Development and Validation of a Model to Estimate the Risk of Acute Ischemic Stroke in Geriatric Patients with Primary Hypertension, BMC Geriatr. 21 (2021), 458. https://doi.org/10.1186/s12877-021-02392-7.

[6] J.H. Friedman, Multivariate Adaptive Regression Splines, Ann. Stat. 19 (1991), 1-67. https://doi.org/10.1214/aos/1176347963.

[7] M.A. Sahraei, H. Duman, M.Y. Çodur, E. Eyduran, Prediction of Transportation Energy Demand: Multivariate Adaptive Regression Splines, Energy 224 (2021), 120090. https://doi.org/10.1016/j.energy.2021.120090.

[8] W. Zhang, C. Wu, Y. Li, L. Wang, P. Samui, Assessment of Pile Drivability Using Random Forest Regression and Multivariate Adaptive Regression Splines, Georisk: Assess. Manag. Risk Eng. Syst. Geohazards 15 (2019), 27-40. https://doi.org/10.1080/17499518.2019.1674340.

[9] H.H. Dukalang, B.W. Otok, Purhadi, Modified Partial Least Square Structural Equation Model with Multivariate Adaptive Regression Spline: Parameter Estimation Technique and Applications, MethodsX 14 (2025), 103381. https://doi.org/10.1016/j.mex.2025.103381.

[10] I.D. Nisa, B.W. Otok, Sutikno, Multivariate Adaptive Inverse Gaussian Regression Spline: Parameter Estimation and Statistical Hypothesis Testing, AIP Conf. Proc. 3231 (2024), 060011. https://doi.org/10.1063/5.0231185.

[11] A.P. Ampulembang, B.W. Otok, A.T. Rumiati, Budiasih, Bi-responses Nonparametric Regression Model Using Mars and Its Properties, Appl. Math. Sci. 9 (2015), 1417-1427. https://doi.org/10.12988/ams.2015.5127.

[12] M. Meilisa, B.W. Otok, J.D.T. Purnomo, Estimation Curve of Multivariate Adaptive Biresponse Fuzzy Clustering Means Regression Splines Approach to Stunting and Wasting Cases in Southeast Sulawesi, MethodsX 12 (2024), 102775. https://doi.org/10.1016/j.mex.2024.102775.

[13] S.L. Cessie, J.C.V. Houwelingen, Logistic Regression for Correlated Binary Data, Appl. Stat. 43 (1994), 95-108. https://doi.org/10.2307/2986114.

[14] A. Agresti, M. Kateri, Categorical Data Analysis, in: International Encyclopedia of Statistical Science, Springer, Berlin, Heidelberg, 2025: pp. 408--411. https://doi.org/10.1007/978-3-662-69359-9_94.

[15] S. Hidayati, B.W. Otok, Purhadi, Parameter Estimation and Statistical Test in Multivariate Adaptive Generalized Poisson Regression Splines, IOP Conf. Ser.: Mater. Sci. Eng. 546 (2019), 052051. https://doi.org/10.1088/1757-899x/546/5/052051.

[16] A. Choiruddin, J.A. González, J. Mateu, A. Fadlurohman, R. Waagepetersen, Variable Selection for Spatio-Temporal Conditionally Poisson Point Processes, Comput. Stat. Data Anal. 212 (2025), 108238. https://doi.org/10.1016/j.csda.2025.108238.