STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, June 2026, pp 4736–4756.
Published online in International Academic Press (www.IAPress.org)

Penalized Spline Semiparametric Logistic Regression for Modelling Coronary Heart Disease Risk Based on Demographic and Lifestyle Factors

Naufal Ramadhan Al Akhwal Siregar[1], Nur Chamidah[2,3,*], Marisa Rifada[2,3], Budi Lestari[4], Dursun Aydin[5,6]

[1]*Mathematics Master Study Program, Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, 60115, Indonesia*
[2]*Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya, 60115, Indonesia*
[3]*Research Group of Statistical Modeling in Life Science, Faculty of Science and Technology, Airlangga University, Surabaya, 60115, Indonesia*
[4]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, The University of Jember, Jember 68121, Indonesia*
[5]*Department of Statistics, Faculty of Science, Muğla Sıtkı Koçman University, Muğla 48000, Turkey*
[6]*Department of Mathematics, University of Wisconsin, Oshkosh Algoma Blvd, Oshkosh, WI 54901, USA*

Abstract This study presents a novel application of Penalized Splines Semiparametric Binary Logistic Regression (PS-SBLR) to evaluate Coronary Heart Disease (CHD) risk. By combining parametric and nonparametric components, the established PS-SBLR method extends classical logistic regression to effectively model both linear and non-linear relationships simultaneously. To estimate the nonparametric component, a penalized spline estimator is used to produce smooth adaptive curves. At the same time, Generalized Approximate Cross Validation (GACV) is employed for smoothing parameter selection to bypass the nonconvergence issues often found in standard CV or GCV methods. While the theoretical foundation of PS-SBLR has shown strong potential in medical research, it has not yet been applied within an integrated framework for both CHD prediction and prevention. To address this gap, we developed a specific PS-SBLR predictive framework using real-world data to enhance the accuracy and efficiency of CHD risk prediction. This applied approach provides valuable, practical insights for the management and mitigation of CHD risk. The resulting predictive model achieved 84.38% accuracy on the training data with an AUC of 0.90, and 87.5% accuracy on the test data with an AUC of 0.98, demonstrating its excellent performance in distinguishing CHD risk profiles. The analysis revealed that, while age and sugar consumption show a linear positive correlation with CHD, continuous variables such as body weight and stress levels exhibit significant non-linear relationships.

Keywords Coronary Heart Disease, Penalized Spline Estimator, Semiparametric Binary Logistic Regression.

AMS 2010 subject classifications 62G08, 62J12, 62P10

DOI: 10.19139/soic-2310-5070-3334

1. Introduction

Logistic regression represents one of the most fundamental tools in statistical modeling due to its strong performance in both prediction and classification tasks. A common variant, Binary Logistic Regression (BLR), assumes a functional relationship between the covariates and the logit of the response [1, 2]. Despite its popularity, the parametric formulation is constrained by the assumption of logit linearity, which may lead to biased or inefficient estimates when the true relationship deviates from this structure. To address these limitations,

*Correspondence to: Nur Chamidah (Email: nur-c@fst.unair.ac.id). Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya, Indonesia (60115).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

nonparametric approaches provide a flexible alternative because they do not impose a predefined functional form on the association between predictors and the response [3]. Instead, these methods construct smooth curves entirely driven by the observed data, thereby minimizing potential bias in modeling [4]. A wide range of smoothing techniques has been developed for non-parametric regression, including kernel estimators [5, 6], local linear estimators [7], truncated spline estimators [8, 9, 10], penalized spline estimators [11, 12, 13], and Fourier series estimators [14]. Among these techniques, penalized spline estimators are particularly advantageous because they effectively capture complex nonlinear patterns while ensuring smooth estimation curves through the Penalized Least Squares (PLS) criterion, regulated by a smoothing parameter (λ) [15].

However, in practice, the relationship between the response variable and certain predictors may exhibit specific patterns, while others may not. Although the nonparametric approach offers significant flexibility, it can suffer from dimensionality issues and reduced interpretability as the degree of the polynomial in each predictor variable increases [16, 17]. To balance model flexibility and simplicity, Semiparametric Binary Logistic Regression (SBLR) emerges as an alternative solution [18, 19]. This approach combines parametric and nonparametric components within a single model, allowing some predictors to enter linearly while others are modeled via data-driven smooth functions. This maintains the interpretability of parametric effects while accommodating nonlinear relationships where necessary. Therefore, semiparametric regression models are better suited to capture such data structures.

In the SBLR model, the relationship between the binary response and its predictors is formulated through a decomposition of linear and smooth functional components. These smooth components are commonly represented using penalized spline bases, which ensure both flexibility and smoothness through appropriate regularization [20]. Semiparametric models based on penalized spline estimators have gained considerable attention due to their computational efficiency and favorable theoretical properties. These models incorporate a spline-based representation of the nonparametric component, controlled by a smoothing parameter that balances fit and smoothness [21]. By doing so, semiparametric logistic regression is capable of capturing complex nonlinear patterns without sacrificing the stability associated with parametric estimation [22]. This hybrid structure is particularly advantageous in biomedical and epidemiological applications, where certain risk factors exhibit well-established linear effects, whereas others demonstrate inherently nonlinear behavior. The development of a flexible and accurate prediction model by combining the SBLR method with one of the data smoothing techniques, namely the penalized spline estimator, can be applied to the prediction of disease risks such as cardiovascular diseases (CVDs), namely coronary heart disease (CHD). CVDs are one of the leading non-communicable diseases (NCDs) globally. In many developed countries, the proportion of deaths caused by cardiovascular disease decreased from around 48 in 1990 to about 43 in 2010. In contrast, developing countries experienced an increase over the same period, with the share rising from 18 percent to 25 percent [23]. Indonesia, as one of the developing countries, has shown a similar upward trend, with cardiovascular disease becoming a major contributor to mortality over the past two decades [24]. Among the various types of CVDs, coronary heart disease (CHD) stands out as the leading cause of death, accounting for approximately 32 percent of global mortality. According to the World Health Organization (WHO), there were 17.9 million deaths caused by cardiovascular diseases in 2018, of which 85 percent were due to CHD [25]. Data from Indonesia's Basic Health Research (Riskesdas) in 2013 and 2018 also indicate a rising trend in heart disease prevalence, increasing from 0.5 percent in 2013 to 1.5 percent in 2018 [26]. One potential strategy for early diagnosis and intervention is the development of risk prediction models for CHD. Numerous studies have demonstrated that CHD risk is influenced by multiple predictors, shaped by a complex interplay of demographic, metabolic, dietary, and psychosocial determinants. As atherosclerosis is the primary pathological mechanism underlying CHD, the disease is often associated with degenerative processes and is more common among individuals with established risk factors and older adults. Age, in particular, is consistently identified as one of the strongest predictors of CHD due to cumulative vascular degeneration over time [27, 28], with the highest prevalence observed in the 65–74-year age group [29]. Hereinafter, anthropometric measures, especially body weight, have also been shown to correlate strongly with CHD risk, reflecting an individual's metabolic load and potential predisposition to cardiometabolic disorders [30]. Furthermore, lifestyle-related behaviors further contribute to this burden. The Indonesian Cardiology Society (PERKI) emphasizes that the rising prevalence of CHD in Indonesia is mainly attributable to unhealthy lifestyle patterns, noting that approximately 50 percent of CHD patients are at risk of sudden cardiac arrest or sudden cardiac death [31]. Dietary habits also play a

critical role, high sugar intake is associated with insulin resistance and metabolic syndrome, whereas excessive consumption of saturated and trans fats accelerates the formation of atherosclerotic plaques [32]. In addition, psychological stress has been shown to exacerbate CHD risk through prolonged activation of neuroendocrine and inflammatory pathways [33], indicating that both physiological and behavioral responses contribute to overall disease progression. Given the multifactorial and potentially nonlinear nature of these risk factors, previous studies have increasingly emphasized the need for statistical models capable of capturing both linear and nonlinear relationships among CHD predictors, thereby improving estimation accuracy and predictive performance. The novelty of this research lies on the development of a semiparametric binary logistic regression method, which is an extension of classical logistic regression that integrates both parametric and nonparametric components, allowing it to model linear and nonlinear relationships simultaneously. The penalized spline estimator is one of the smoothing techniques used in the nonparametric approach, as it produces smooth, adaptive curves for fluctuating data. Research by [34] highlights that penalized splines can produce smooth curves that are also adaptive to data pat-terns, making them highly suitable for the classification of heterogeneous medical data. Semiparametric logistic regression models with penalized splines (PS-SBLR) have shown strong potential in analyzing complex data, particularly in medical research. However, existing research has not yet applied this approach in an integrated frame-work for both CHD prediction and prevention. This study develops and applies a penalized spline estimator based on a semiparametric logistic regression (PS-SBLR) model using real-world data to enhance the accuracy and efficiency of CHD risk prediction. The proposed approach not only advances statistical modeling methodology but also provides practical insights for managing and mitigating CHD risk. Moreover, this re-search aligns with Sustainable Development Goal (SDG) 3, which targets a reduction in premature mortality from non-communicable diseases by 2030.

2. Material and Methods

In this study, we first explain the estimation of the Semiparametric Binary Logistic Regression (SBLR) model parameters theoretically using the penalized likelihood method. If this process produces implicit equations, so the solution cannot be carried out directly, then we need a numerical method to obtain parameter estimates, namely, the Iteratively Reweighted Penalized Least Squares (IRPLS) iteration method. Next, we apply the SBLR model's estimation results to real data to predict CHD risk, accounting for the effects of patient age, body weight, body height, sugar and fat intake, and stress level, using the following steps: (i). Determine data characteristics; (ii). Carry out correlation testing using the Eta correlation test; (iii). Select the optimal smoothing parameter value of the PS-SBLR model based on Generalized Approximate Cross Validation (GACV) criterion; (iv). Determine the estimation values of the PS-SBLR model on the real data by creating R-code; (v). Determine the Deviance value and investigate the stability of the classification model; (vi). Determine the optimal classification threshold based on Youden's J index; (vii). Determine the values of accuracy, sensitivity, specificity, and AUC. In the following, we briefly review the concept of the Binary Logistic Regression (BLR) model that, in this study, would be extended to a semiparametric regression term such that it yields a Semiparametric Binary Logistic Regression (SBLR) model, concepts of penalized spline (P-Spline) estimator and GACV that would be used in the estimation process of the SBLR model.

2.1. Semiparametric Binary Logistic Regression (SBLR) Model

Semiparametric binary logistic regression is a statistical method for analyzing response variables that have binary scale and combines both parametric and nonparametric approaches. This model can be estimated by integrating the framework of the Generalized Linear Model (GLM) and the Generalized Additive Model (GAM), namely the Generalized Additive Partially Linear Model (GAPLM) [35].

Let y_i be a binary response variable for the i-th observation $(i = 1, 2, ..., n)$, assumed to follow a Bernoulli distribution, $y_i \sim b(\pi_i)$, where $\pi_i = P(y_i = 1|\mathbf{x}_i, \mathbf{t}_i)$. Let $\mathbf{x}_i = (1, x_{i1}, ..., x_{ip})^T$ denote the vector of predictors modeled parametrically and $\mathbf{t}_i = (t_{i1}, ..., t_{iq})^T$ denote the vector of predictors modeled nonparametrically using smooth functions. The relationship between the conditional mean of the response variable and the predictors is

modeled through the logit link function. The systematic component of the SBLR model is expressed as an additive model [36]:

$$\eta_i(\mathbf{x}_i, \mathbf{t}_i) = \ln\left(\frac{\pi_i(\mathbf{x}_i, \mathbf{t}_i)}{1 - \pi_i(\mathbf{x}_i, \mathbf{t}_i)}\right) = \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^{q} g_j(t_{ji}) \tag{1}$$

Where η_i is the logit function, $g(.)$ represents the unknown smooth functions associated with the continuous predictors (t_{ji}) in the nonparametric component.

2.2. Truncated Spline Basis

A common method for representing a spline function is through the truncated spline basis [20]. For a polynomial spline of degree d with K knots located at $\xi_{j1}, \xi_{j2}, \xi_{j3}, \ldots, \xi_{jr_j}$, the function $g(t_{ji})$ can be expressed as:

$$g_j(t_{ji}) = \theta_{j0} + \theta_j t_{ji} + \cdots + \theta_{jd_j} t_{ji}^{d_j} + \sum_{k_j=1}^{r_j} \theta_{jk_j} \left(t_{ji} - \xi_{jk_j}\right)_+^{d_j} \tag{2}$$

In this study, the unknown functions $g_j(t_{ji})$ are approximated using Penalized Splines (P-Splines) based on the Linear Spline Truncated Basis. For a specific predictor t_{ji} with (degree = 1), and a set of knots $\xi_{j1}, \xi_{j2}, \xi_{j3}, \ldots, \xi_{jr_j}$, the function can be represented as:

$$g_j(t_{ji}) = \theta_{j0} + \theta_j t_{ji} + \sum_{k_j=1}^{r_j} \theta_{jk_j} \left(t_{ji} - \xi_{jk_j}\right)_+ \tag{3}$$

Where,

$$\left(t_{ji} - \xi_{jk_j}\right)_+ = \begin{cases} \left(t_{ji} - \xi_{jk_j}\right), & t_{ji} \geq \xi_{jk_j} \\ 0, & t_{ji} < \xi_{jk_j} \end{cases}$$ and θ_{jk_j} represents the coefficients

for the basis functions associated with the knots.

2.3. Penalized Likelihood Method

Penalized splines (P-splines) have emerged as a flexible and robust tool for modeling nonlinear relationships in regression analysis. They are constructed from piecewise polynomial segments smoothly joined at specific knot locations to achieve a desired degree of continuity. In most applications, the knot locations are determined using the sample quantiles of the unique values of the explanatory variable, which allows the method to adapt to the data distribution. To address the inherent trade-off between flexibility and overfitting, P-splines generally employ a relatively large number of knots to capture complex data structures. However, to avoid excessive fluctuations, a penalty term is introduced into the likelihood function to control smoothness and stabilize estimation [21]. In spline-based regression models, the estimation process can be formulated through the penalized log-likelihood approach. This method modifies the standard log-likelihood function by incorporating a penalty term that constrains the roughness of the fitted curve. The general form is given by:

$$\ell_p(\boldsymbol{\omega}) = \ell(\boldsymbol{\omega}) - \frac{\lambda}{2} \boldsymbol{\omega}^T \boldsymbol{S} \boldsymbol{\omega} \tag{4}$$

2.4. Generalized Approximate Cross Validation (GACV)

The estimation of the semiparametric model parameters relies heavily on the choice of the smoothing parameter λ. This parameter plays a crucial role in controlling the trade-off between the goodness-of-fit of the model and the smoothness of the spline curve. A small value of λ tends to produce a rough curve that overfits the data, whereas a large value of λ results in an overly smooth curve that may lead to underfitting (high bias). Therefore, an optimal selection method is required to determine the value of λ that balances bias and variance.

In the context of non-Gaussian data, such as the binary response in this study, the ordinary Cross-Validation (CV) or Generalized Cross-Validation (GCV) methods used in linear regression are not directly applicable due to the non-quadratic nature of the log-likelihood function. To address this, we employ the Generalized Approximate Cross Validation (GACV) criterion, which is an approximation of the Leave-One-Out Cross-Validation (LOOCV) adapted for Generalized Linear Models (GLMs). The optimal smoothing parameter λ is obtained by minimizing the value of GACV function. Based on the formulation by Xiang and Wahba [37], the GACV function for a penalized likelihood model is defined as follows:

$$GACV(\lambda) = -\frac{1}{n} l(\hat{\boldsymbol{\omega}}) + \frac{tr(\boldsymbol{H})}{n} \frac{\boldsymbol{y}^T(\boldsymbol{y} - \hat{\boldsymbol{\pi}})}{n - tr(\boldsymbol{W}^{1/2}\boldsymbol{H}\boldsymbol{W}^{1/2})} \tag{5}$$

can be written as follows:

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i\eta_\lambda(\boldsymbol{x}_i, \boldsymbol{t}_i) + \log(1 + e^{\eta_\lambda(\boldsymbol{x}_i, \boldsymbol{t}_i)})\right) + \frac{tr(\boldsymbol{H})}{n} \frac{\sum_{i=1}^{n} y_i(y_i - \pi_\lambda(\boldsymbol{x}_i, \boldsymbol{t}_i))}{n - tr(\boldsymbol{W}^{1/2}\boldsymbol{H}\boldsymbol{W}^{1/2})} \tag{6}$$

To determine the best PS-SBLR model, we should select the optimal smoothing parameters, namely, the optimal knot points and smoothing parameters (λ) based on GACV criterion that is minimum value of GACV function. The GACV function is presented in Equation (6). Next, below we provide an algorithm to determine the optimal smoothing parameter values (see **Algorithm 1**).

Algorithm 1. Determining the Optimal Smoothing Parameters by GACV Value

1. **Create** the "ginverse" function for the general inverse.
2. **Create** the "quant" function to determine the optimal knot point.
3. **Create** the "trun" function to determine the value of the knot point location for each variable.
4. **Create** the "matrikx" function to obtain optimal knot point iterations and knot locations.
5. **Create** the "matrikd" function to construct the penalty matrix
6. **Create** the "hitung.gacv" function to obtain the minimum GACV value.
7. **Create** the "plot_gacv" function to obtain the plot of GACV value.
8. **Define** q for i-th predictor for nonparametric components.
9. **Define** a vector of nonparametric predictor variable (t).
10. **Define** a vector of response variable (y).
11. **Sorting** the data on each predictor will obtain the optimum knot point and knot location.
12. **Make iterations** to obtain the minimum GACV value with the syntax "carioptimal_logistic_gacv", so that we, obtain the optimal knot point, knot location, and lambda (λ) according to the minimum GACV value.

2.5. Data Analysis Steps

After we obtain optimal smoothing for all predictor variables, the next step is to iterate the initial values by using the IRPLS method and by creating R-code to obtain the estimation result of the multipredictor PS-SBLR model using training data. Below, we provide an algorithm to estimate the PS-SBLR model on the real data. (see **Algorithm 2**).

Algorithm 2. Steps of Estimating the PS-SBLR Model and Analyzing Data

1. **Define** n of the number of observations in the data
2. **Define** a vector of parametric predictor variable x
3. **Define** a vector of nonparametric predictor variable t

4. **Define** a vector of response variable y.
5. **Construct** and combined of the semiparametric design matrix $\mathbf{C}$ and penalty matrix $\mathbf{S}$
6. **Select** the optimal smoothing parameter (k_{opt}, λ_{opt}) using the GACV method
7. **Iterate** the Iteratively Reweighted Penalized Least Square (IRPLS) algorithm k_{opt}, λ_{opt} to obtain final parameters $\hat{\boldsymbol{\omega}}$
8. **Extract** the final estimator $\hat{\boldsymbol{\omega}}$, separating the parametric coefficients $(\widehat{\boldsymbol{\beta}})$ and the nonparametric spline coefficients $(\widehat{\boldsymbol{\gamma}})$.
9. **Calculate** the vector of estimated probabilities $\widehat{\boldsymbol{\pi}}$ based on **step 8**.
10. **Determine** the optimal classification threshold (c_{opt}) based on Youden's J index
11. **Perform** the final classification using c_{opt}
12. **Calculate** the Deviance, Press-Q, and MC-Nemar test
13. **Calculate** the final Confusion Matrix, and model evaluation metrics, namely Accuracy, Sensitivity, Specificity, and Area Under the Curve (AUC)

2.6. Performance Evaluation Metrics

To assess the effectiveness of the predictive models, a "2 × 2 confusion matrix" is employed. The disease status is categorized into two groups: "CHD" and "Non-CHD." In this context, "CHD" is considered a positive event, while "Non-CHD" is regarded as a negative event. By applying these two classes to the confusion matrix, we can obtain counts for true negatives, true positives, false negatives, and false positives. To further assess the performance of the models using clinical variables recommended by the WHO and the American Heart Association, we will examine several metrics for all models, such as "accuracy, sensitivity, specificity, and the area under the curve (AUC) associated with the receiver operating characteristic (ROC) curve" [39]. The following are the detailed formulas for the performance evaluation metrics utilized in this research:

$$\text{Recall or Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{7}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \tag{8}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{False Negative} + \text{True Negative}} \tag{9}$$

$$\text{AUC} = \sum_{i=1}^{n} (FPR_{i+1} - FPR) \times \frac{TPR_{i+1} + TPR_i}{2} \tag{10}$$

A general reference according to [38] for interpreting the AUC value of a diagnostic test is presented in Table 1.

Table 1. Categorical Value of AUC

Categorical	**Range AUC Values**
Excellent Classification	$0.90 < AUC \leq 1.00$
Very Good Classification	$0.80 < AUC \leq 0.90$
Good Classification	$0.70 < AUC \leq 0.80$
Sufficient Classification	$0.60 < AUC \leq 0.70$
Bad Classification	$0.50 < AUC \leq 0.60$
Test Not Useful	$AUC \leq 0.5$

2.7. Data Source and Research Variables

This research utilizes primary data collected at Universitas Airlangga Hospital (RSUA) in 2024. The data were obtained through patient questionnaires and interviews conducted during medical care visits. Specifically, medical record data from 80 post-cardiac catheterization patients were used, with 64 observations designated for model training and 16 observations for model testing. The sample comprised 43 individuals without coronary heart disease (non-CHD) and 37 individuals diagnosed with coronary heart disease (CHD). The primary outcome variable was the incidence of CHD, whereas the predictor variables encompassed a range of factors potentially associated with CHD. A comprehensive description of all variables is provided in Table 2.

Table 2. Research Variables

Notation	Variable Name	Scale	Detail
y	Incidence of CHD	Nominal	0: Non-CHD 1: CHD
x_1	Gender	Nominal	0: Male 1: Female
x_2	Age	Ratio	In (Years)
t_1	Body Weight	Ratio	In (Kg)
t_2	Sugar Consumption	Ratio	Amount of sugar intake per day (grams)
t_3	Stress Level Scoring	Ratio	Total score on the questionnaire

Data collection for this study was conducted using an accidental sampling technique, where subjects were selected based on their availability and willingness to participate. The study period encompassed the interval from [Month, Year] to [Month, Year]. The variables observed in this study include the dependent variable, Incidence of CHD (y), and several independent variables Gender (x_1), Age (x_2), Body Weight (t_1), Sugar Consumption (t_2), and Stress Level Scoring (t_3). To measure dietary habits, specifically Sugar Consumption (t_2), a Semi-quantitative Food Frequency Questionnaire (SQ-FFQ) was used to estimate daily intake in grams. Additionally, the Stress Level Scoring (t_3) was derived from the questionnaire total score. This research adheres to ethical standards for human subject research. Ethical approval was obtained from the "Rumah Sakit Universitas Airlangga (RSUA) Committee" with Ethical Clearance Number: 089/KEP/2024 before data collection.

3. Results

This study applies the SBLR-PS model to Coronary Heart Disease (CHD) risk cases at Universitas Airlangga Hospital (RSUA). This study encompasses the description of the characteristics of factors suspected to influence CHD incidence among outpatients at the RSUA cardiology clinic, CHD risk modeling using the PS-SBLR approach, and model analysis and risk prediction of CHD incidence using the obtained PS-SBLR model.

3.1. Descriptive Analysis of Patient Characteristics

The descriptive analysis aims to identify the distributional patterns and fundamental characteristics of the research subjects concerning CHD risks. To facilitate interpretation, the predictor variables in this study are classified into three primary categories. The first category concerns objective factors, including age and body weight. The second category encompasses biochemical parameters, specifically sugar levels, which serve as crucial metabolic indicators. The third category involves psychological factors, quantified by scoring the stress level variable. Stress levels were categorized based on the knot points identified in the spline regression analysis, which align with the clinical progression from normal to severe stress states.

Descriptive analysis was conducted to examine the characteristics of the predictor variables, both numerically and visually, according to each variable's scale. For numerical predictor variables, descriptive statistics included the mean, variance, minimum, and maximum values. These statistics are presented separately by response variable category, namely CHD status. Meanwhile, for binary categorical variables, the analysis was performed visually using bar charts to compare categories. The summary of numerical predictor variable characteristics is presented in Table 3 as follows.

Table 3. Description of Numerical Predictor Variable Characteristics

Variable	Status	Minimum	Maximum	Mean	Variance
Age	Non-CHD	18	78	39.7	380.25
	CHD	31	74	59.3	104.04
Body Weight	Non-CHD	36	97	63.8	176.89
	CHD	38	100	69.2	141.61
Sugar Consumption	Non-CHD	0.63	32.27	10.42	68.66
	CHD	1.57	40.29	14.69	125.02
Stress Level	Non-CHD	0	14	5.27	11.56
	CHD	0	12	4.167	12.92

Based on **Table 3** presents the descriptive statistics of predictor variables classified by the response variable status, namely the Non-CHD group (healthy) and the CHD group (Coronary Heart Disease patients). Based on the table, the CHD group has a significantly higher mean age (59.3 years) than the Non-CHD group (39.7 years). Furthermore, the variance in the Non-CHD group (380.25) is much larger than that in the CHD group (104.04). This indicates that the Non-CHD group represents a broader age range, whereas CHD cases tend to be concentrated in the elderly demographic. In terms of anthropometric measurements, the CHD group shows a higher mean body weight (69.2 kg) than the Non-CHD group (63.8 kg). Specific patterns were also identified in the sugar consumption variable and the psychological factor, namely stress level. The CHD group recorded a higher mean sugar consumption (14.69) compared to the Non-CHD group (10.42). In the aspect of psychological factors, the Non-CHD group showed a slightly higher mean Stress Level score (5.27) compared to the CHD group (4.167). Nevertheless, the variances of stress levels in both groups are relatively similar (11.56 and 12.92, respectively), suggesting a consistent spread of stress scores across the sample, regardless of disease status.

Based on this descriptive analysis, the data indicate the presence of complex, potentially non-linear relationships between dietary habits, particularly sugar consumption, and disease status. This complexity underscores the need to apply the Semiparametric Binary Logistic Regression (SBLR) model. Unlike parametric and nonparametric binary logistic regression methods, SBLR offers the necessary flexibility to accommodate such irregular data patterns.

Furthermore, the descriptive analysis of characteristics was also conducted visually using a bar chart for the binary categorical predictor variable, namely gender, as shown in Figure 1.

In the gender variable, respondents with non-CHD status are predominantly female. This is evidenced by Figure 1, which illustrates that the 'Female' bar is higher than the 'Male' bar, with 23 and 20 respondents respectively, resulting in a frequency difference of 3 respondents. Conversely, respondents with CHD status are predominantly male. This is evidenced by Figure 1, which shows that the 'Male' bar is higher than the 'Female' bar, with 31 and 6 respondents, respectively, a difference of 25.

After conducting descriptive analyses of all predictor variables, the next step is to split the data. The data is divided into training data and testing data with proportions of 80 and 20 percent of the total dataset, respectively. The number of training data in this study is 64 respondents, and the number of testing data is 16 respondents. The overall division can be tabulated in Table 4 as follows.

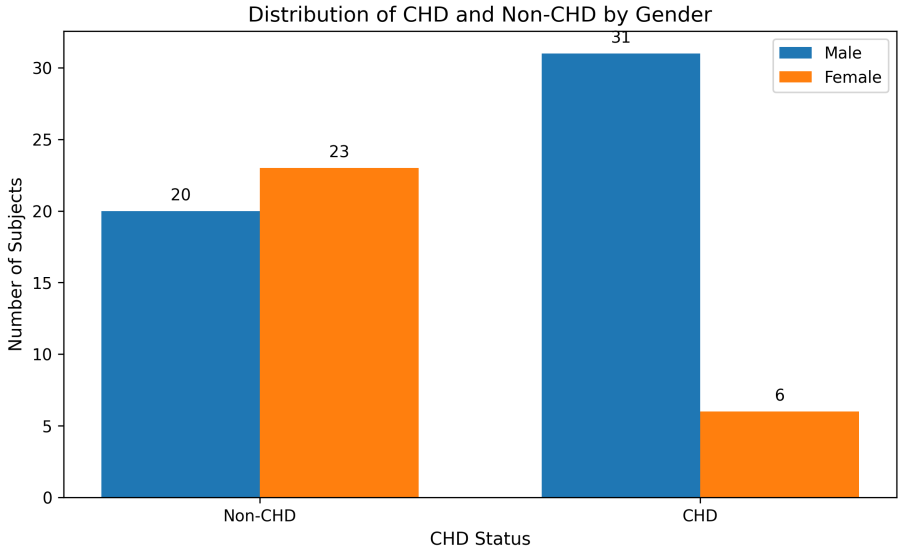

Figure 1. Distribution of CHD and Non-CHD based on Gender.

Table 4. Tabulation of Research Data Distribution

Data Partition	**Categorization**		**Total**
	Non-CHD	**CHD**	
Training Data	36	28	64
Testing Data	7	9	16
Total	**43**	**37**	**80**

3.2. Modeling Coronary Heart Disease Risk Using Semiparametric Binary Logistic Regression Based on the Penalized Spline Estimator

The initial step in modeling the risk of Coronary Heart Disease (CHD) using semiparametric binary logistic regression based on the penalized spline estimator is to examine the relationship between each predictor variable and the response variable within the training data, given each variable's scale. To test the linearity assumption inherent in logistic regression, the relationship between continuous predictors and the observed log-odds (logit) of Coronary Heart Disease (CHD) is examined. The resulting scatterplots are presented in Figure 2 as follows:

While the graphical exploration via observed logit plots in the previous section provided visual indications of fluctuating and nonlinear patterns, interpretations based solely on visual inspection are inherently subjective. A pattern that appears nonlinear visually must be statistically verified to ensure that the deviation from linearity is significant and not merely due to random noise. Therefore, to overcome this subjectivity and objectively validate the assumption for using nonparametric functions, this study employs the Eta Correlation (η) test statistic [40]. The calculation results of the Eta correlation test for the predictors are presented in Table 5

Based on Table 5, it can be seen that the predictor variable Age has an η statistic value and a p-value < 0.05. Therefore, the decision is to reject H_0, concluding that there is a dependency between the predictor variable Age and the CHD incidence variable. Thus, this predictor variable is used in the semiparametric binary logistic regression model as a parametric component.

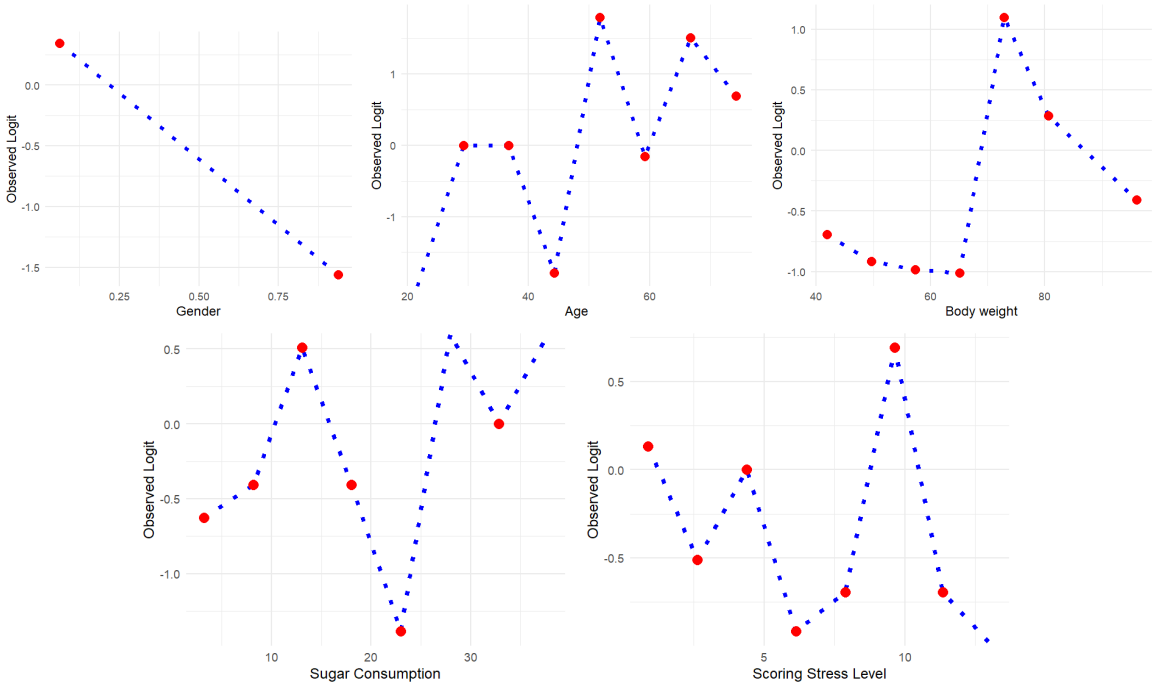

Figure 2. Observed Logit Plot For Each Predictor Variable.

Table 5. Eta Correlation Test

Variable	Value of Eta (η)	Value of Eta (η^2)	Pvalue	Decision
Age	0.541	0.293	0.000	Reject H_0 (Linear)
Body Weight	0.191	0.036	0.131	Not Reject H_0 (Non-Linear)
Sugar Consumption	0.179	0.032	0.155	Not Reject H_0 (Non-Linear)
Scoring Stress level	0.103	0.011	0.416	Not Reject H_0 (Non-Linear)

Meanwhile, some other predictor variables have η statistic values and p-values > 0.05, specifically Body Weight, Sugar Consumption, and Stress Level. Therefore, the decision is to accept H_0, concluding that there is no dependency between the predictor variables Body Weight, Sugar Consumption, and Stress Level and the CHD incidence variable. Thus, these three predictor variables can be used as nonparametric components in the semiparametric binary logistic regression model.

Based on the similarity of scales and data types, the relationship between the response variable and categorical predictor variable was tested using the Chi-square independence test statistics. The calculation of test statistics and critical regions was performed computationally in RStudio, with the results presented in Table 6 as follows.

Table 6. Chi-square (χ^2) Independence Test Results on Categorical Predictor Variables

Predictor Variable	Statistic (χ^2)	Critical Region	*P-Value*
(X_1) Gender	8.533	3.841	0.003

Based on the Chi-square independence test in Table 6, it can be seen that the predictor variable Gender has a χ^2 statistic value (8.533) $> \chi^2_{(0.05;1)}$ (3.841) and a p-value < 0.05. Therefore, a decision can be made to reject H_0, concluding that there is a dependency between the predictor variable Gender and the incidence of Coronary Heart Disease (CHD). Thus, the predictor variable Gender can be used in semiparametric binary logistic regression modeling as a parametric component. Thus, the predictor variables used in this study can be treated as both parametric and nonparametric components in the semiparametric binary logistic regression model.

Before modeling the risk of CHD incidence using Semiparametric Binary Logistic Regression (SBLR) based on the Penalized Spline estimator, three smoothing parameters consisting of order, knots, and λ (Lambda) were determined for each predictor variable that acts as a nonparametric component. The determination of these smoothing parameters was based on the Generalized Approximate Cross Validation (GACV) criterion, as given in Equation (6). The optimal smoothing parameter selection was chosen based on the minimum GACV value. In this study, the polynomial order used for each nonparametric component predictor variable was order one (linear). Thus, the parameters for the number of knots, optimal knot locations, and λ values for each nonparametric component predictor variable were obtained in Table 7

Table 7. Selection of Optimal Smoothing Parameters

Variable	**Number of Knots**	**Knot Locations**	**Lambda (λ)**	**Minimum GACV Value**
Body Weight	2	63; 71	0.011	0.6224
Sugar Consumption	3	4.34; 8.40; 18.83	0.078	0.6730
Stress Level	2	3; 6	1.1	0.6834

In Table 7, a summary of the optimal smoothing parameter selection results for each continuous predictor variable is presented. This selection is based on the minimum *Generalized Approximate Cross Validation* (GACV) value criterion, where the lowest GACV value indicates the best balance between *goodness of fit* and curve *smoothness*.

For the Body Weight variable, a minimum GACV value of 0.6224 was obtained using 2 knot points located at 63 kg and 71 kg, as well as a smoothing parameter (λ) of 0.011. Furthermore, for the sugar consumption variable, the optimal configuration was achieved at a GACV value of 0.6730 with 3 knot points (locations: 4.34 g; 8.40 g; and 18.83 g) and a smoothing parameter (λ) of 0.078. Meanwhile, the scoring stress level variable yielded a minimum GACV of 0.6834 with 2 knot points at 3 and 6, and a smoothing parameter (λ) of 1.1. These knot points and smoothing parameters will be used in the RLBS modeling with the penalized spline estimator.

Based on the testing of the relationship between predictor variables and response variables on the training data (see in the Table 5 and Table 6), as well as the determination of the optimum smoothing parameters for the nonparametric components (Table 7), it is established that this study uses two parametric systematic component predictor variables, namely age and gender, and three nonparametric systematic component predictor variables, namely body weight, sugar consumption, and stress level. The estimation results of the semiparametric binary logistic regression model can be written for each predictor variable as follows.

$$\left.\begin{aligned}
\hat{\theta}_0 &= 4.0340 \\
\hat{\beta}_1 &= -0.7374 \\
\hat{\beta}_2 &= 0.0957 \\
\hat{\boldsymbol{\gamma}}_{\mathbf{1}} &= [-0.1945 \quad 0.6994 \quad -0.5149]^T \\
\hat{\boldsymbol{\gamma}}_{\mathbf{2}} &= [0.0242 \quad 0.0467 \quad -0.0828 \quad 0.3369]^T \\
\hat{\boldsymbol{\gamma}}_{\mathbf{3}} &= (-0.3748 \quad 0.4898 \quad 0.1096)^T
\end{aligned}\right\} \tag{11}$$

For the first parametric predictor variable, namely **Gender**, the link function estimation of the CHD risk model using semiparametric binary logistic regression is obtained by assuming that other variables are constant, so it can

be written as $\hat{f}(x_{1i(1)})$ for the gender variable in Equation (12) as follows:

$$\begin{aligned}\hat{f}(x_{1i(1)}) &= \hat{\beta}_{1(1)}(x_{1i(1)}) = -0.7374(x_{1i(1)})\\ \hat{f}(x_{1i(1)}) &= \begin{cases}-0.7374, & \text{if } x_{1i}=1 \text{ (Female)}\\ 0, & \text{if } x_{1i}=0 \text{ (Male)}\end{cases}\end{aligned} \tag{12}$$

For the second parametric predictor variable, namely **Age**, the link function estimation of the CHD risk model using semiparametric binary logistic regression is obtained by assuming that other variables are constant, so it can be written as $\hat{f}(x_{2i})$ for the age variable in Equation (13) as follows:

$$\hat{f}(x_{2i}) = \hat{\beta}_2(x_{2i}) = 0.0957(x_{2i}) \tag{13}$$

For the first nonparametric predictor variable, namely **Body Weight**, the link function estimation of the CHD risk model using semiparametric binary logistic regression is obtained by assuming that other variables are constant, so it can be written as $\hat{g}(t_{1i})$ for the body weight variable in Equation (14) as follows:

$$\begin{aligned}\hat{g}(t_{1i}) &= \hat{\boldsymbol{\gamma}}_{\mathbf{1}}(t_{1i})\\ &= -0.1945(t_{1i}) + 0.6994(t_{1i}-63)_+ - 0.5149(t_{1i}-71)_+\\ \hat{g}(t_{1i}) &= \begin{cases}-0.1945t_{1i}, & \text{if } t_{1i}<63\\ -44.0622+0.5049t_{1i}, & \text{if } 63\le t_{1i}<71\\ -7.5043-0.01t_{1i}, & \text{if } t_{1i}\ge 71\end{cases}\end{aligned} \tag{14}$$

For the second nonparametric predictor variable, namely **Sugar Consumption**, the link function estimation of the CHD risk model using semiparametric binary logistic regression is obtained by assuming that other variables are constant, so it can be written as $\hat{g}(t_{2i})$ for the sugar consumption variable in Equation (5) as follows:

$$\begin{aligned}\hat{g}(t_{2i}) &= \hat{\boldsymbol{\gamma}}_{\mathbf{2}}(t_{2i})\\ &= 0.0242(t_{2i}) + 0.0468(t_{2i}-4.34)_+ - 0.0828(t_{2i}-8.40)_+\\ &\quad + 0.3370(t_{2i}-18.83)_+\\ \hat{g}(t_{2i}) &= \begin{cases}0.0242t_{2i}, & \text{if } t_{2i}<4.34\\ -0.2031+0.071t_{2i}, & \text{if } 4.34\le t_{2i}<8.40\\ 0.4924-0.0118t_{2i}, & \text{if } 8.40\le t_{2i}<18.83\\ -5.8533+0.3252t_{2i}, & \text{if } t_{2i}\ge 18.83\end{cases}\end{aligned} \tag{15}$$

For the third nonparametric predictor variable, namely **Stress Level**, the link function estimation of the CHD risk model using semiparametric binary logistic regression is obtained by assuming that other variables are constant, so it can be written $\hat{g}(t_{3i})$ for the stress level variable in Equation (16) as follows:

$$\begin{aligned}\hat{g}(t_{3i}) &= \hat{\boldsymbol{\gamma}}_{\mathbf{3}}(t_{3i})\\ &= -0.3748(t_{3i}) + 0.4898(t_{3i}-3)_+ + 0.1096(t_{3i}-6)_+\\ \hat{g}(t_{3i}) &= \begin{cases}-0.3748t_{3i}, & \text{if } t_{3i}<3\\ -1.4694+0.115t_{3i}, & \text{if } 3\le t_{3i}<6\\ -2.127+0.2246t_{3i}, & \text{if } t_{3i}\ge 6\end{cases}\end{aligned} \tag{16}$$

After obtaining the link function estimates through Equations (11-16), the construction of the probability estimation model for CHD risk prediction using PS-SBLR can be performed, which is presented in Equation (17) as follows.

$$\hat{\pi}_i(x_i,t_i) = \frac{\exp(4.034 - 0.7374(x_{i(1)}) + \cdots + 0.1096(t_{3i}-6)_+)}{(1+\exp(4.034 - 0.7374(x_{i(1)}) + \cdots + 0.1096(t_{3i}-6)_+))} \tag{17}$$

Based on the model parameter estimation results obtained in Equations (1-6), the model interpretation can be performed by calculating the *Odds Ratio* value. In the semiparametric binary logistic regression model, the parametric component predictor variables have a single *odds ratio*, whereas the nonparametric component predictor variables can have *odds ratio* values greater than or equal to two. The number of *odds ratio* values produced depends on the number of optimal knots for each variable. Thus, the calculation of the *odds ratio* for each predictor variable as a whole can be presented in Table 8 as follows.

Table 8. Odds Ratio Values for Each Predictor Variable

Variable	**Knot Point**	**Parameter Estimate**	**Odds Ratio [OR]**	**CHD Risk**
Gender (x_{1i})	-	-0.7374	0.4784	Decreases
Age (x_{2i})	-	0.0957	1.1004	Increases
Body Weight (t_{1i})	$t_{1i} < 63$	-0.1945	0.8232	Decreases
	$63 \leq t_{1i} < 71$	0.5049	1.6568	Increases
	$t_{1i} \geq 71$	-0.01	0.9900	Decreases
Sugar Consumption (t_{2i})	$t_{2i} < 4.34$	0.0242	1.0245	Increases
	$4.34 \leq t_{2i} < 8.40$	0.071	1.0736	Increases
	$8.40 \leq t_{2i} < 18.83$	-0.0118	0.9883	Decreases
	$t_{2i} \geq 18.83$	0.3252	1.3843	Increases
Stress Level (t_{3i})	$t_{3i} < 3$	-0.3748	0.6874	Decreases
	$3 \leq t_{3i} < 6$	0.115	1.1219	Increases
	$t_{3i} \geq 6$	0.2246	1.2518	Increases

Based on Table 8, show that the predictor variables reveal that CHD risk is driven by significant linear associations by age and gender, and also non-linear associations across body weight, sugar consumption, and stress factors.

1. **Gender**
   Based on Table 8, it is known that the OR value for the gender variable is $e^{(-0.7374)} = 0.4784$. The interpretation of this OR value is that the probability of CHD occurrence for a "Female" patient is 0.4784 times lower compared to a "Male" patient, assuming other predictor variables are constant.

2. **Age**
   Based on Table 8, it is known that the OR value for the age variable is $e^{(0.0957)} = 1.1004$. The interpretation of this OR value is that the increase in the probability of CHD occurrence if a patient's age increases by one year is 1.1004 times higher, assuming other predictor variables are constant.

3. **Body Weight**
   Based on Table 8, it is known that the OR value for the body weight variable less than 63 kg is $e^{(-0.1945)} = 0.8232$. The interpretation of this OR value is the magnitude of the decrease in the probability of CHD occurrence if a patient's body weight increases by 1 kg is 0.8232 times lower assuming other predictor variables are constant. However, if the patient's body weight is in the range of 63 kg to less than 71 kg, the OR is $e^{(0.5049)} = 1.6568$, which means the increase in the probability of CHD occurrence if a patient's body weight increases by one kg is 1.6568 times higher, assuming other predictor variables are constant.

4. **Sugar Consumption**
   Based on Table 8, in general, all sugar consumption intervals show a tendency for increased risk (OR $>$ 1), but the magnitude varies. It is known that the OR value for the sugar consumption variable greater than or equal to 18.83 grams is $e^{(0.3252)} = 1.3843$. The interpretation of this OR value is the magnitude of the

increase in the probability of CHD occurrence if a patient's sugar consumption increases by 1 gram is 1.3843 times higher, assuming other predictor variables are constant.

5. **Stress Level Scoring**
   Based on Table 8, there is a clear pattern of risk change (*threshold effect*) on the stress level variable. At low stress levels (score < 3), the parameter estimation value is negative and OR < 1, indicating the risk of CHD decreases. It is known that the OR value for the stress level variable if it is greater than 3 and less than 6 is $e^{(0.115)} = 1.1219$. The interpretation of this OR value is the magnitude of the increase in the probability of CHD occurrence if a patient's stress level increases by one unit is 1.1219 times higher assuming other predictor variables are constant. Furthermore, if the stress level is greater than or equal to 6, then the OR value is $e^{(0.2246)} = 1.2518$, which means that the increase in the probability of CHD occurrence if a patient's stress level score increases by one unit is 1.2518 times higher assuming other predictor variables are constant.

3.3. Model Performance Analysis

Once the probability estimation model is established in Equation (17), the next stage is predictive performance analysis. This process includes testing the model's goodness of fit using the deviance test statistic, as well as finding the optimal classification threshold classification threshold tuning. Model evaluation is performed by constructing a confusion matrix to calculate three main evaluation metrics: accuracy, sensitivity, and specificity. This model evaluation is complemented by the creation of an ROC curve and the calculation of the AUC.

Based on the estimation results of the Coronary Heart Disease risk probability using semiparametric binary logistic regression in Equation (17), goodness of fit testing is performed. This test is based on the deviance statistic [41].

- H_0: The logistic regression model is appropriate.
- H_1: The logistic regression model is not appropriate.

The critical region of the *Deviance* test is H_0 is rejected if $D > \chi^2_{(\alpha,(n-d-1))}$ or when p-value $< \alpha$, where n is the number of samples, d is the number of predictors used (excluding the *intercept* variable), and α is the significance level. The calculation of the deviance test statistic is performed computationally using RStudio software so that the model goodness of fit can be obtained through Table 9 as follows.

Table 9. Model Goodness of Fit Test Results using Deviance Test Statistic

Goodness Criteria	**Binary Logistic Regression**	
	Parametric	**Semiparametric**
Deviance	57.5199	48.0781
P-value	0.4931	0.6158

Based on Table 9, it is known that the parametric binary logistic regression model has a deviance value of **57.5199** which is outside the critical region with p-value > 0.05. Therefore, the decision to **accept** H_0 is obtained, with the conclusion that the parametric logistic regression model is appropriate. Furthermore, it is also known that the semiparametric binary logistic regression model has a deviance value of **48.0781** which is outside the critical region with p-value > 0.05. Therefore, the decision to **accept** H_0 is obtained, with the conclusion that the **semiparametric binary logistic regression model is appropriate**.

Based on the comparison of the deviance values for both models, the semiparametric binary logistic regression model has a lower deviance than the parametric binary logistic regression model. Therefore, the semiparametric approach is more suitable for predicting CHD risk in this study.

In this study, optimal threshold selection is used to classify the estimated CHD risk probability for each patient, $\hat{\pi}_i(x_i, t_i)$, into the binary categories 0 and 1. Optimal threshold selection is performed using Youden's J statistic in Equation (18) [42].

$$J_{\max(c)} = \arg \max_{c\in[0,1]} \{\text{Sensitivity}(c) + \text{Specificity}(c) - 1\} \tag{18}$$

The top five pairs of classification threshold and Youden's J values are shown in Table 10 below.

Table 10. Classification Threshold and Youden's J Values

Threshold	Accuracy	Sensitivity	Specificity	Youden's J
0.62	84.375	0.7143	0.9444	0.6587
0.6	82.8125	0.7143	0.9167	0.631
0.61	82.8125	0.7143	0.9167	0.631
0.63	82.8125	0.6786	0.9444	0.623
0.64	82.8125	0.6786	0.9444	0.623

Based on Table 10, it is known that the optimal *classification threshold* is when $c = 0.62$. Class estimation through prediction probability obtained in Equation (17) using threshold ($c = 0.62$), is presented in Equation (19) as follows:

$$\hat{y}_i = \text{class}(\hat{\pi}_i(\boldsymbol{x}_i, \boldsymbol{t}_i)) = \begin{cases} 1 \text{ (CHD)}, & \text{if } \hat{\pi}_i(\boldsymbol{x}_i, \boldsymbol{t}_i) \geq 0.62 \\ 0 \text{ (No CHD)}, & \text{if } \hat{\pi}_i(\boldsymbol{x}_i, \boldsymbol{t}_i) < 0.62 \end{cases} \tag{19}$$

The estimation results of the CHD occurrence class for each patient in the *training* data are performed computationally using RStudio software and presented in Table 11 as follows:

Table 11. Confusion Matrix on Training Data

Actual Class	Predicted Class		Total
	No CHD	CHD	
No CHD	34	2	36
CHD	8	20	28
Total	42	22	64

Based on Table 11, we calculate the PRESS's Q statistic to evaluate predictive accuracy and McNemar's test to evaluate the symmetry of misclassifications. The PRESS's Q statistic tests the null hypothesis that the classification accuracy is not significantly better than chance [45].

$$Q = \frac{[N - (n \times K)]^2}{N(K-1)} \tag{20}$$

Where:

- N = Total sample size
- n = Number of correct classifications ($TN + TP$)
- K = Number of groups (classes)

$$Q = \frac{[64 - (54 \times 2)]^2}{64(2-1)} = \frac{[64 - 108]^2}{64(1)} = \frac{[-44]^2}{64} = \frac{1936}{64} = 30.25 \tag{21}$$

The calculated PRESS's Q value is 30.25. We compare this to the critical value of the Chi-square distribution with 1 degree of freedom ($\chi^2_{0.05,1} = 3.841$). Since $30.25 > 3.841$, we reject the null hypothesis. This indicates that the model's predictive accuracy is statistically significant and performs substantially better than random guessing.

Additionally, McNemar's test is used to determine if there is a significant difference between the discordant classifications (False Positives vs. False Negatives) [44].

$$\chi^2_{McN} = \frac{(b-c)^2}{b+c} \tag{22}$$

Where:

- b = False Positives (Actual Non-CHD, Predicted CHD)
- c = False Negatives (Actual CHD, Predicted Non-CHD)

$$\chi^2_{McNemar} = \frac{(2-8)^2}{2+8} = \frac{(-6)^2}{10} = \frac{36}{10} = 3.6 \tag{23}$$

The calculated McNemar's value is 3.6. Comparing this to the critical Chi-square value ($\chi^2_{0.05,1} = 3.841$). Since $3.6 < 3.841$, we fail to reject the null hypothesis ($p > 0.05$). This suggests that there is no significant difference between the rate of False Positives (2) and False Negatives (8). The model does not show a significant bias towards one specific type of error over the other.

After that, Based on Table 11, several model performance evaluation metric calculations can be performed on the *training* data. The calculations are performed computationally using RStudio software, and the results are obtained in Table 12 as follows.

Table 12. Model Performance Evaluation Metric Values on Training Data

Model Performance Evaluation Metric	**Result**	**95% Confidence Interval**
Accuracy(%)	84.375%	73.14% - 92.24%
Sensitivity (*True Positive Rate*)	0.7143	0.5133 - 0.8678
Specificity (*True Negative Rate*)	0.9444	0.8134 - 0.9932

The following is the interpretation of each evaluation metric based on Table 12.

a) The **Accuracy** metric of 84.375% states that the proportion of the number of actual CHD occurrences successfully classified correctly by the model, including the "No CHD" class and "CHD" class, compared to the total number of observations is 84.375%. The 95% confidence interval for accuracy is in the range of 73.14% to 92.24%, which indicates that the true accuracy value in the population is estimated to be within that interval.

b) The **Sensitivity** metric of 0.7143 represents that approximately 71.43% of patients who factually suffer from CHD (category 1) were successfully classified correctly by the model, with a 95% confidence interval ranging from 0.5133 to 0.8678.

c) The **Specificity** metric of 0.9444 means that 94.44% of individuals who do not suffer from CHD (category 0) were successfully identified correctly as non-CHD, with a 95% confidence level that the population specificity is in the high range, namely 0.8134 to 0.9932.

The optimal threshold ($c = 0.62$) can be used to calculate the AUC and to visualize it through the ROC curve presented in Fig.3 as follows.

Based on Fig. 3, the area of the ROC curve through the AUC value obtained can be categorized as outstanding classification (**very good**). This means that the semiparametric binary logistic regression model based on the

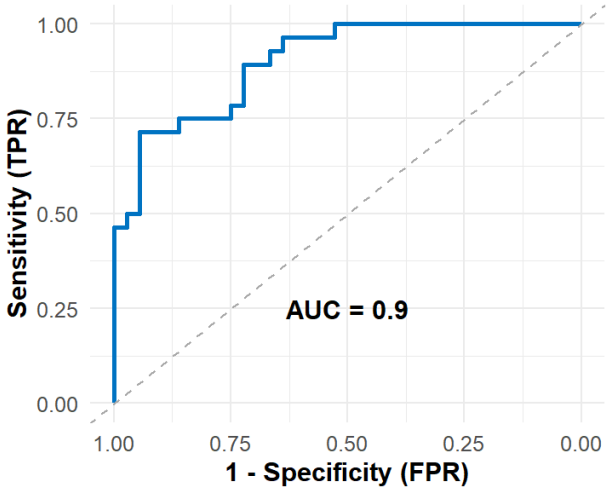

Figure 3. Threshold Visualization and ROC Curve

penalized spline estimator is capable of preventing misclassification and classifying the majority of observations in the training data correctly.

To validate that the built model does not experience overfitting, it needs to be applied to new data that has never been recognized by the model before, namely, testing data. The results of the CHD occurrence class estimation for each patient in the *testing* data are performed computationally using RStudio software, with the results presented in Table 13 as follows:

Table 13. Confusion Matrix on Testing Data

Actual Class	**Predicted Class**		**Total**
	No CHD	**CHD**	
No CHD	7	0	7
CHD	2	7	9
Total	9	7	16

Based on Table 13, the calculation of several model performance evaluation metrics can be performed on the testing data. The calculation is performed computationally using RStudio software, and the results are obtained in Table 14 as follows.

Table 14. Model Performance Evaluation Metric Values on Testing Data

Model Performance Evaluation Metric	**Calculation Result**	**95% Confidence Interval**
Accuracy	87.500%	61.65% - 98.45%
Sensitivity (*True Positive Rate*)	0.778	0.3999 - 0.9719
Specificity (*True Negative Rate*)	1.000	0.5904 - 1.000

The following is the interpretation of each evaluation metric based on Table 14.

a) The **Accuracy** metric of 87.5% states that the proportion of the number of actual CHD occurrences successfully classified correctly by the model, including the "No CHD" class and "CHD" class, compared

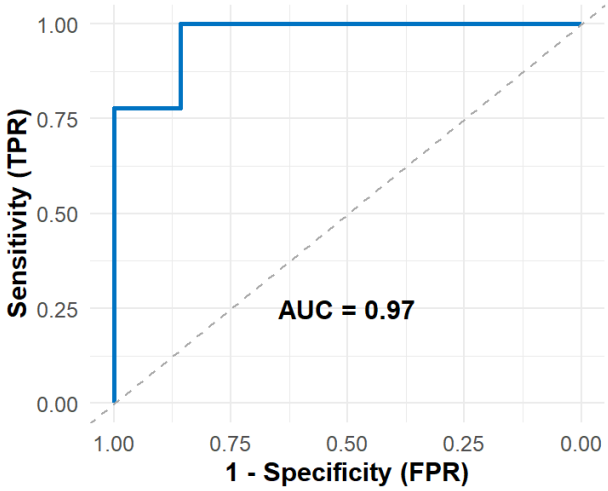

Figure 4. Threshold Visualization and ROC Curve

to the total number of observations is 87.5%. The 95% confidence interval for accuracy is in the range of 61.65% to 98.45%, which indicates that the true accuracy value in the population is estimated to be within that interval.

b) The **Sensitivity** metric of 0.7778 indicates that approximately 77.78% of patients who actually have CHD (category 1) were correctly classified by the model, with a 95% confidence interval of 0.3999 to 0.9719.

c) The **Specificity** metric of 1.000 means that 100% of individuals who do not suffer from CHD (category 0) were successfully identified correctly as non-CHD, with a 95% confidence level that the population specificity is in the high range, namely 0.5904 to 1.000.

The optimal threshold ($c = 0.62$) can be used to calculate the AUC and to visualize it through the ROC curve presented in Fig. 4. Based on the AUC value obtained from the ROC curve in Fig. 4, the classification performance can be categorized as Excellent. This implies that the semiparametric binary logistic regression model based on the penalized spline estimator can prevent classification errors, including overfitting, and correctly classify observations in the testing data.

4. Discussion

4.1. Interpretation

The primary objective of this study was to develop and evaluate a Semiparametric Binary Logistic Regression (SBLR) model using a Penalized Spline estimator for predicting Coronary Heart Disease (CHD) risk. The results demonstrate that the proposed SBLR algorithm achieves superior predictive performance, with an Area Under the Curve (AUC) of 0.97 and a testing accuracy of 87.5%. These findings suggest that the semiparametric approach offers a significant improvement over traditional parametric logistic regression, particularly in medical datasets where the relationship between physiological predictors and disease outcomes is rarely strictly linear [46, 47]. By utilizing Generalized Approximate Cross Validation (GACV) for smoothing parameter selection, the model effectively minimized the risk of overfitting, a common challenge in flexible modeling approaches.

Regarding the parametric components, the study confirms established epidemiological evidence. The analysis indicates that CHD risk increases linearly with age (OR 1.1004) and is significantly higher in males compared to females (OR 0.4784 for females). This aligns with global cardiovascular statistics, which attribute the lower risk

in females to the cardioprotective effects of estrogen, particularly in pre-menopausal populations [48, 49]. The linear positive association with age is also consistent with the progressive nature of arterial stiffness and plaque accumulation over time [50].

A key contribution of this study is the identification of non-linear threshold effects in modifiable risk factors. Contrary to the assumption that risk increases uniformly with body weight, our model revealed a specific high-risk window between 63 kg and 71 kg (OR 1.6568). This finding supports recent literature suggesting that metabolic health is not solely determined by BMI magnitude but also by specific body composition thresholds [51]. Similarly, sugar consumption and stress levels exhibited distinct "tipping points." The sharp increase in risk for sugar consumption exceeding 18.83 grams and stress scores ≥ 6 highlights the importance of identifying critical cut-off points for clinical intervention, rather than assuming a constant dose-response relationship [52, 53].

4.2. Limitations of the Study

Despite the robust insights provided, several limitations must be acknowledged. First, the sample size was relatively restricted ($n = 80$). While the penalized spline method with GACV selection is mathematically designed to handle smaller datasets by penalizing roughness to enhance generalization capability, the limited number of observations may still constrain the statistical power and the generalizability of the findings to the broader population. Second, the data was collected from a single center, Universitas Airlangga Hospital (RSUA), which may introduce geographic or demographic biases. Finally, potential confounding variables such as smoking history, genetic predisposition, and lipid profiles were not included due to data availability constraints. Future studies should validate these non-linear thresholds using larger, multi-center longitudinal datasets further to refine the predictive precision of the SBLR model.

5. Conclusion

This study successfully developed a robust estimation algorithm for Semiparametric Binary Logistic Regression (SBLR) using a Penalized Spline estimator. The estimation procedure integrates the Local Scoring iterative method within the Generalized Linear Models (GLM) framework and the Backfitting algorithm for Generalized Additive Models (GAM) components. The computational implementation, developed using R statistical software, effectively automates the selection of optimal smoothing parameters (λ) by minimizing the Generalized Approximate Cross Validation (GACV) score. This approach has proven to be computationally efficient and capable of producing precise model parameter estimates and risk probabilities.

The empirical application of the proposed model to Coronary Heart Disease (CHD) patient data at Universitas Airlangga Hospital (RSUA) reveals critical insights into the disease's etiology. The study identified a mixed functional relationship between predictors and the response variable. Demographic factors, specifically **Age** and **Gender**, were identified as parametric components, whereas **Body Weight**, **Sugar Consumption**, and **Stress Levels** exhibited complex, non-linear patterns best modeled through a semiparametric approach. Based on the updated Odds Ratio (OR) analysis, the dynamics of CHD risk vary significantly across predictors. Regarding **Linear Effects**, **Age** contributes positively to increased CHD risk with an OR of **1.1004**, while conversely, **Gender** (female patients) demonstrates a significantly lower probability of risk (**0.4784** times) compared to males.

Furthermore, regarding **Non-Linear Threshold Effects**, the continuous variables show distinct patterns. **Body Weight** shows a distinct risk spike while weights are less than 63 kg and more than or equal to 71 kg. tend to lower the risk, a sharp increase of **1.6568 times** is observed specifically in the 63–71 kg range. **Sugar Consumption** exhibits fluctuating effects, with the most notable risk increase (OR **1.3843**) occurring when consumption exceeds **18.83**. Finally, **Stress Levels** display a clear progression; scores < 3 are associated with lower risk, whereas the risk increases by **1.1219 times** for scores between 3–6 and climbs to **1.2518 times** for scores ≥ 6.

Finally, the predictive performance of the SBLR model is exceptional. On the training data, the model achieved a classification accuracy of 84.375% and an Area Under Curve (AUC) of 0.9. Crucially, the model demonstrated excellent generalization capabilities on the testing data, achieving an improved accuracy of **87.5%** and an AUC of

0.97. These results classify the model's performance as outstanding, confirming its reliability and effectiveness as a predictive tool for Coronary Heart Disease risk in outpatient settings.

Acknowledgement

The authors thank the Directorate of Research and Community Service (DPPM), Directorate General of Higher Education, Research and Technology, Ministry of Higher Education, Research and Technology, the Republic of Indonesia, for funding this research trough a Magister Theses Research grant (hibah Penelitian Tesis Magister - PTM), with a contract number: 2435/B/UN3.LPPM/PT.01.03/2025.

REFERENCES

1. D.W. Hosmer Jr, S. Lemeshow, and R.X. Sturdivant, *Applied Logistic Regression*, John Wiley & Sons, 2013.
2. D.G. Kleinbaum and M. Klein, *Introduction to Logistic Regression*, Logist. Regres. Self-Learn. Text, pp. 1–39, 2010.
3. A. Yara and Y. Terada, *Nonparametric Logistic Regression with Deep Learning*, ArXiv Prepr. ArXiv240112482, 2024.
4. M. Rifada, N. Chamidah, and V. Ratnasari, *Estimation of Nonparametric Ordinal Logistic Regression Model Using Local Maximum Likelihood Estimation*, Commun. Math. Biol. Neurosci., vol. 2021, Article-ID 28, 2021.
5. T.H. Ali, *Modification of the Adaptive Nadaraya-Watson Kernel Method for Nonparametric Regression (Simulation Study)*, Commun. Stat. - Simul. Comput., vol. 51, pp. 391–403, 2022.
6. N. Chamidah and T. Saifudin, *Estimation of Children Growth Curve Based on Kernel Smoothing in Multi-Response Nonparametric Regression*, Appl. Math. Sci., vol. 7, pp. 1839–1847, 2013.
7. Y. Linke, I. Borisov, P. Ruzankin, V. Kutsenko, E. Yarovaya, and S. Shalnova, *Universal Local Linear Kernel Estimators in Nonparametric Regression*, Mathematics, vol. 10, p. 2693, 2022.
8. A. Islamiyati, A. Kalondeng, N. Sunusi, M. Zakir, and A.K. Amir, *Biresponse Nonparametric Regression Model in Principal Component Analysis with Truncated Spline Estimator*, J. King Saud Univ.-Sci., vol. 34, p. 101892, 2022.
9. M. Setyawati, N. Chamidah, A. Kurniawan, and D. Aydin, *Confidence Interval for Semiparametric Regression Model Parameters Based on Truncated Spline with Application to COVID-19 Dataset in Indonesia*, Salud Cienc. Tecnol., vol. 3, 2024.
10. M. Setyawati, N. Chamidah, and A. Kurniawan, *Confidence Interval of Parameters in Multiresponse Multipredictor Semiparametric Regression Model for Longitudinal Data Based on Truncated Spline Estimator*, Commun. Math. Biol. Neurosci., vol. 2022, Article-ID 107, 2022.
11. N. Chamidah, B. Lestari, H. Susilo, T.K. Dewi, T. Saifudin, N.R.A.A. Siregar, and D. Aydin, *Modeling Coronary Heart Disease Risk Based on Age, Fatty Food Consumption and Anxiety Factors Using Penalized Spline Nonparametric Logistic Regression*, MethodsX, p. 103320, 2025.
12. Z.N. Amalia, D.R. Hastuti, F. Istiqomah, and N. Chamidah, *Hypertension Risk Modeling Using Penalized Spline Estimator Approach Based on Consumption of Salt, Sugar, and Fat Factors*, In Proceedings of the AIP Conference Proceedings, vol. 2264, AIP Publishing, 2020.
13. T. Adiwati and N. Chamidah, *Modelling of Hypertension Risk Factors Using Penalized Spline to Prevent Hypertension in Indonesia*, In Proceedings of the IOP Conference Series: Materials Science and Engineering, vol. 546, p. 052003, IOP Publishing, 2019.
14. M. Zulfadhli, I.N. Budiantara, and V. Ratnasari, *Nonparametric Regression Estimator of Multivariable Fourier Series for Categorical Data*, MethodsX, vol. 13, p. 102983, 2024.
15. A. Islamiyati and N. Chamidah, *Penalized Spline Estimator with Multi Smoothing Parameters in Bi-Response Multi-Predictor Nonparametric Regression Model for Longitudinal Data*, Songklanakarin J. Sci. Technol., vol. 42, 2020.
16. M. Alswaitti, K. Siddique, S. Jiang, W. Alomoush, and A. Alrosan, *Dimensionality Reduction, Modelling, and Optimization of Multivariate Problems Based on Machine Learning*, Symmetry, vol. 14, p. 1282, 2022.
17. T.A. Reddy and G.P. Henze, *Parametric and Non-Parametric Regression Methods*, In Applied data analysis and modeling for energy engineers and scientists, Springer, pp. 355–407, 2023.
18. R.J. Carroll and M.P. Wand, *Semiparametric Estimation in Logistic Measurement Error Models*, J. R. Stat. Soc. Ser. B Methodol., vol. 53, pp. 573–585, 1991.
19. F. Fang, J. Li, and X. Xia, *Semiparametric Model Averaging Prediction for Dichotomous Response*, J. Econom., vol. 229, pp. 219–245, 2022.
20. M.A.S. Mullah, J.A. Hanley, and A. Benedetti, *LASSO Type Penalized Spline Regression for Binary Data*, BMC Med. Res. Methodol., vol. 21, p. 83, 2021.
21. J. Yu, J. Shi, A. Liu, and Y. Wang, *Smoothing Spline Semiparametric Density Models*, J. Am. Stat. Assoc., vol. 117, pp. 237–250, 2022.
22. X. Zheng, Y. Rong, L. Liu, and W. Cheng, *A More Accurate Estimation of Semiparametric Logistic Regression*, Mathematics, vol. 9, p. 2376, 2021.
23. A. Maharani and G. Tampubolon, *Unmet Needs for Cardiovascular Care in Indonesia*, PloS One, vol. 9, p. e105831, 2014.
24. F. Savira, B.H. Wang, A.R. Kompa, Z. Ademi, A.J. Owen, D. Liew, and E. Zomer, *The Impact of Coronary Heart Disease Prevention on Work Productivity: A 10-Year Analysis*, Eur. J. Prev. Cardiol., vol. 28, pp. 418–425, 2021.
25. World Health Organization, *Noncommunicable Diseases Country Profiles 2018*, 2018.
26. Kemenkes RI, *Laporan Nasional RISKESDAS 2018*, Kementrian Kesehatan RI, 2018.

27. K. Hosseini, S.H. Mortazavi, S. Sadeghian, A. Ayati, M. Nalini, A. Aminorroaya, H. Tavolinejad, M. Salarifar, H. Pourhosseini, and A. Aein, *Prevalence and Trends of Coronary Artery Disease Risk Factors and Their Effect on Age of Diagnosis in Patients with Established Coronary Artery Disease: Tehran Heart Center (2005–2015)*, BMC Cardiovasc. Disord., vol. 21, p. 477, 2021.
28. Y.-T.H. Lee, J. Fang, L. Schieb, S. Park, M. Casper, and C. Gillespie, *Prevalence and Trends of Coronary Heart Disease in the United States, 2011 to 2018*, JAMA Cardiol., vol. 7, pp. 459–462, 2022.
29. I. Rethemiotaki, *Global Prevalence of Cardiovascular Diseases by Gender and Age during 2010–2019*, Arch. Med. Sci. Atheroscler. Dis., vol. 8, p. e196, 2023.
30. J.F. Meyer, S.B. Larsen, K. Blond, C.T. Damsgaard, L.G. Bjerregaard, and J.L. Baker, *Associations between Body Mass Index and Height during Childhood and Adolescence and the Risk of Coronary Heart Disease in Adulthood: A Systematic Review and Meta-analysis*, Obes. Rev., vol. 22, p. e13276, 2021.
31. P.D.S.K. Indonesia, *Pedoman Tatalaksana Sindrom Koroner Akut*, Jkt. Cent. Commun., 2015.
32. N. Temple, *Fat, Sugar, Whole Grains and Heart Disease: 50 Years of Confusion*, Nutrients, vol. 10, p. 39, 2018.
33. P.H. Wirtz and R. Von Känel, *Psychological Stress, Inflammation, and Coronary Heart Disease*, Curr. Cardiol. Rep., vol. 19, p. 111, 2017.
34. G. Marra and R. Radice, *Penalised Regression Splines: Theory and Application to Medical Research*, Stat. Methods Med. Res., vol. 19, pp. 107–125, 2010.
35. R. Liu and W.K. Härdle, *Statistical Inference for Generalized Additive Partially Linear Model*, ArXiv:2009.04793, 2020.
36. R.F. Manghi, F.J.A. Cysneiros, and G.A. Paula, *Generalized Additive Partial Linear Models for Analyzing Correlated Data*, Comput. Stat. Data Anal., vol. 129, pp. 47–60, 2019.
37. D. Xiang and G. Wahba, *A General Approximate Cross Validation for Smoothing Splines with Non-Gaussian Data*, Stat. Sin., vol. 6, pp. 675–692, 1996.
38. A.-M. Šimundić, *Measures of diagnostic accuracy: basic definitions*, EJIFCC, vol. 19, no. 4, p. 203, 2009.
39. R. Trevethan, *Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice*, Front. Public Health, vol. 5, p. 307, 2017.
40. M.K. Uçar, *Eta Correlation Coefficient Based Feature Selection Algorithm for Machine Learning: E-Score Feature Selection Algorithm*, J. Intell. Syst. Theory Appl., vol. 2, pp. 7–12, 2019.
41. J.K. Harris, *Primer on Binary Logistic Regression*, Fam. Med. Community Health, vol. 9, p. e001290, 2021.
42. M.D. Ruopp, N.J. Perkins, B.W. Whitcomb, and E.F. Schisterman, *Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection*, Biom. J., vol. 50, pp. 419–430, 2008.
43. J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. De Mendonça, *Data Mining Methods in the Prediction of Dementia: A Real-Data Comparison of the Accuracy, Sensitivity and Specificity of Linear Discriminant Analysis, Logistic Regression, Neural Networks, Support Vector Machines, Classification Trees and Random Forests*, BMC Res. Notes, vol. 4, p. 299, 2011.
44. M.Q.R. Pembury Smith and G.D. Ruxton, *Effective Use of the McNemar Test*, Behav. Ecol. Sociobiol., vol. 74, p. 133, 2020.
45. J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. De Mendonça, *Data Mining Methods in the Prediction of Dementia: A Real-Data Comparison of the Accuracy, Sensitivity and Specificity of Linear Discriminant Analysis, Logistic Regression, Neural Networks, Support Vector Machines, Classification Trees and Random Forests*, BMC Res. Notes, vol. 4, p. 299, 2011.
46. T.J. Hastie and R.J. Tibshirani, *Generalized Additive Models*, London: Chapman and Hall/CRC, 1990.
47. S.N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed., Boca Raton: CRC Press, 2017.
48. L. Mosca et al., "Effectiveness-based guidelines for the prevention of cardiovascular disease in women—2011 update," *Circulation*, vol. 123, no. 11, pp. 1243–1262, 2011.
49. E.J. Benjamin et al., "Heart disease and stroke statistics—2019 update," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.
50. B.J. North and D.A. Sinclair, "The intersection between aging and cardiovascular disease," *Circulation Research*, vol. 110, no. 8, pp. 1097–1108, 2012.
51. C.K. Kramer, B. Zinman, and R. Retnakaran, "Are metabolically healthy overweight and obese individuals at increased risk of cardiovascular disease and mortality?," *Annals of Internal Medicine*, vol. 159, no. 11, pp. 758–769, 2013.
52. S. Cohen, D. Janicki-Deverts, and G.E. Miller, "Psychological stress and disease," *JAMA*, vol. 298, no. 14, pp. 1685–1687, 2007.
53. V.S. Malik et al., "Sugar-sweetened beverages and risk of metabolic syndrome and type 2 diabetes," *Diabetes Care*, vol. 33, no. 11, pp. 2477–2483, 2010.