BAREKENG: Journal of Mathematics and Its Applications
December 2025 Volume 19 Issue 4 Page 2877–2890
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss4pp2877-2890

MODELING FACTORS CAUSING ALZHEIMER'S DISEASE USING LOGIT, PROBIT, AND GOMPIT LINK FUNCTIONS IN GENERALIZED LINEAR MODEL

Ardi Kurniawan[1*], Gabriella Agnes Budijono[2], Kimberly Maserati Siagian[3], Adrian Wahyu Abdillah[4]

[1,2,3,4]*Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga*
Jln. Dr. Ir. H. Soekarno, Mulyorejo, Surabaya, 60115, Indonesia

*Corresponding author's e-mail: *ardi-k@fst.unair.ac.id*

Article History:

Received: 9th March 2025
Revised: 2nd May 2025
Accepted: 17th June 2025
Available online: 1st September 2025

Keywords:

Alzheimer;
Generalized Linear Model;
Gompit;
Gompit Link Functions
Logit.

ABSTRACT

This study addresses the ongoing challenge of clarifying the risk factors contributing to Alzheimer's disease, a neurodegenerative condition marked by progressive cognitive decline and memory dysfunction, with cases rising globally. To provide a more accurate and comprehensive understanding of the predictors associated with the disease, this research models the contributing factors using logit, probit, and gompit link functions within the Generalized Linear Model (GLM). Utilizing secondary data from 2024, which includes predictor variables such as age, family history, head injury, hypertension, memory complaints, and behavioral disturbances, this research models the relationship between these variables and Alzheimer's diagnosis. The analysis finds that the logit, probit, and gompit link functions yield significant results in identifying risk factors associated with Alzheimer's diagnosis, particularly memory complaints and behavioral disturbances. The gompit link is selected as the best model due to its highest deviance R-squared value of 30.01%, indicating better reliability in predicting Alzheimer's diagnosis than other models. This GLM approach provides insights to support early prevention and intervention efforts for Alzheimer's disease and contribute to achieving Sustainable Development Goals (SDGs) number 3 on good health and well-being.

How to cite this article:

A. Kurniawan, G. A. Budijono, K. M. Siagian and A. W. Abdillah., "MODELING FACTORS CAUSING ALZHEIMER'S DISEASE USING LOGIT, PROBIT, AND GOMPIT LINK FUNCTIONS IN GENERALIZED LINEAR MODEL," *BAREKENG: J. Math. & App.*, vol. 19, iss. 4, pp. 2877-2890, December, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Alzheimer's disease is a form of neurodegenerative illness that develops gradually and progressively, characterized by the death of neurons in the brain. This condition often leads to declining cognitive functions, including memory, thinking ability, and speech, affecting patients' daily activities [1]. This disease can not only happen to older people, but it can also happen to young people. Although the exact cause of Alzheimer's disease remains unknown, research indicates that the accumulation of beta-amyloid protein in the brain plays a critical role in the formation of plaques that affect the nervous system. These plaques lead to significant apoptosis of brain cells, causing the brain to shrink and atrophy [2].

The global prevalence of Alzheimer's disease continues to rise. A 2020 report by Alzheimer's Disease International (ADI) revealed that in that year, the number of people living with Alzheimer's worldwide reached 55 million, with projections increasing to 78 million by 2030 and 139 million by 2050. The Asia region recorded the highest figures, with 29.23 million cases in 2020 and a projected increase to 42.71 million cases by 2030. Specifically, Southeast Asia contributed 4.31 million cases. In Indonesia, the number of Alzheimer's cases has also risen significantly. According to data from the Ministry of Health via the Online Hospital Information System, there were 1.2 million Alzheimer's patients in 2019. Moreover, between 2019 and 2022, the number of patients accessing healthcare services for dementia and Alzheimer's through BPJS Kesehatan increased from 5,583 to 10,414 patients, representing an 87% growth.

Previous studies on predictive analysis of Alzheimer's disease have employed various statistical and machine learning approaches. Research by [3] utilized the Naïve Bayes Classifier and binary logistic regression with eight genetic predictor variables, achieving an accuracy of 79.8% and an AUC of 0.809, although the sensitivity was only 69.4%. Another study by [4] tested several machine learning algorithms using data from Kaggle and found that the Logistic Regression model achieved the highest accuracy at 85.71%. However, it lacked consideration of clinical and biological factors. Meanwhile, [5] demonstrated the effectiveness of the Generalized Linear Model (GLM) in predicting malaria cases using climatic and historical variables, highlighting GLM's potential in health-related predictive modeling. Given the rising number of Alzheimer's cases, the GLM approach with binary logistic regression is crucial. It enables simultaneous analysis of categorical and numerical variables while accounting for interaction effects and non-linear relationships. This study compares explicitly three common GLM link functions: logit, probit, and gompit, each with its advantages in different contexts [6]. The logit function is helpful for its interpretability via odds ratios, the probit function is suitable under the assumption of normal distribution, and the gompit function is effective for modeling rare events. Choosing the appropriate link function can enhance model accuracy and calibration in predicting Alzheimer's risk.

This research also supports Sustainable Development Goal (SDG) 3, which aims to reduce premature mortality from non-communicable diseases and improve mental health and well-being. Therefore, this research aims to analyze the factors contributing to Alzheimer's disease using the Generalized Linear Model (GLM) with binary logistic regression, focusing on identifying variables that significantly increase the risk of the disease. By incorporating categorical and numerical variables, this study goes beyond previous approaches by considering interaction effects, non-linear relationships, and clinical factors to enhance prediction accuracy [7], [8]. Additionally, it addresses challenges such as class imbalance and prioritizes sensitivity improvement, offering a more precise and comprehensive framework for Alzheimer's risk prediction. The results of this research are expected to provide deeper insights for prevention strategies, support tailored intervention efforts, and enhance healthcare services related to Alzheimer's in the future.

2. RESEARCH METHODS

2.1 Alzheimer

Dementia is a syndrome caused by chronic or progressive brain disease characterized by decreased memory, communication, and daily activities [9]. Alzheimer's disease is the most common type of dementia. It is a neurodegenerative condition that affects the brain, leading to neuron loss and impairing abilities such as language, executive functions, attention, and visuospatial functions, with progressive cognitive decline, disrupting daily activities. Early symptoms include episodic short-term memory loss with minimal long-term memory impairment, followed by difficulties in problem-solving, judgment, executive functions, lack of motivation, and disorganization, causing multitasking and abstract thinking problems. Age is a significant

risk factor for Alzheimer's. Other triggers include traumatic head injuries, depression, cardiovascular and cerebrovascular diseases, smoking, family history of dementia, and elevated homocysteine levels, among others [10].

2.2 Generalized Linear Model

Generalized Linear Models (GLM) quantify the relationship between the response and predictor variables. GLM extends traditional linear regression models, enabling data analysis where the response variable need not follow a normal distribution. Link functions allow GLM to address non-linear relationships between predictors and the response variable [11], [12]. Sampling-based methods such as OSUMC can significantly improve estimation efficiency in large-scale GLMs under measurement constraints. Generally, the link function is expressed as:

$$g(\mu_i) = \eta_i \tag{1}$$

where, $g(\mu_i)$ is the link function, which maps the mean μ_i to the linear predictor η_i ; $\mu_i = E(Y_i)$ is the expected value of the response variable Y_i conditioned on the independent variable X_i. In GLM, the linear predictor is expressed as a combination of independent variables X with parameter β, namely [13]:

$$n_i = \beta_0 + \beta_0 x_{1i} + \beta_0 x_{2i} + \cdots + \beta_p x_{pi} \tag{2}$$

where, η_i is the linear predictor (a linear combination of independent variables); β_0 is the intercept; $\beta_1, \beta_2, \ldots$, are regression coefficients for the independent variables $x_{1i}, x_{2i}, \ldots, x_{pi}$; $x_{1i}, x_{2i}, \ldots, x_{pi}$ are the independent variables for the i-th observations. The canonical link function for the binomial distribution is $\eta = \ln\left(\frac{p}{1-p}\right)$, which also called the logistic or logit function. The relationship between p and η is non-linear, resembling an S-curve. If η increases, p will also increase within the interval $0 \leq p \leq 1$.

2.3 Binary Logit Model

The binary logit model is a regression model where the response variable consists of two categories, while the predictor variables can be either categorical or continuous. The probability of success in an experiment is based on the logistic distribution [14]. The binary logit model is expressed as follows:

$$g(\pi_i) = X_i\beta \; ; i = 1,2, \ldots, N \tag{3}$$

Where $g(\pi_i)$ is the logit function: $X_i = (1, X_{i1}, X_{i2}, \ldots, X_{ip})$ is the predictor variables for the i-th observation, and $\beta = \beta_0, \beta_1, \ldots, \beta_p)'$ is the parameter vector corresponding to the predictor variables. From equation (3), the probability of success for the i-th observation based on the predictor variables X_i is obtained as follows:

$$\pi_i = \frac{exp(X_i\beta)}{1 + exp\,(X_i\beta)} \tag{4}$$

with $\pi_i = \Pr\,(Y + 1|X_i)$.

2.4 Binary Probit Model

The binary probit model is a nonlinear model with a normal distribution. The binary probit model is a modeling approach used to explain the relationship between two or more variables, where the response variable is categorical and has an error factor [15], [16]. The binary probit model is expressed as follows [17]:

$$g(\pi_i) = \Phi^{-1}(\pi_i) = X_i\beta; i = 1,2, \ldots, n \tag{5}$$

Where $\boldsymbol{g(\pi_i) = \Phi^{-1}(\pi_i)}$ is the probit link function, $\boldsymbol{\Phi}$ is the cumulative standard normal distribution function, $\boldsymbol{X_i = (1, X_{1i}, X_{2i}, \ldots, X_{pi})}$ is the vector of predictor variables for the i-th observation, and $\boldsymbol{\beta = (\beta_0, \beta_1, \ldots, \beta_p)'}$ is the vector of parameters. From the above equation, the probability of success for the i-th observation based on the predictor variables $\boldsymbol{X_i}$ is obtained as follows:

$$\pi_i = \Phi(\mathrm{X}_i\beta) \tag{6}$$

where $\boldsymbol{\pi_i} = \mathbf{Pr}\,(\boldsymbol{Y} = \mathbf{1}|\boldsymbol{X_i})$.

2.5 Binary Gompit Model

The binary Gompit model is a regression model with a response variable consisting of two categories, namely $\boldsymbol{Y} = (\mathbf{0}, \mathbf{1})$, and the probability of success in the trial is based on the Gompit distribution. This binary logit model is a regression model that is based on the concept of probabilities. The Gompit model is:

$$g(\pi_i) = ln(-\,ln(1-\pi_i)) = X_i\beta, i = 1,2,\dots,n \tag{7}$$

where $\boldsymbol{g(\pi_i)}$ is the gompit function; $\boldsymbol{X_i}$ is the vector of predictor variables for the *i*-th observation; β is the parameter vector corresponding to the predictor variables. From the model above, the probability success in a trial for the *i*-th observation given the predictor variable $\boldsymbol{X_i}$ is:

$$\pi_i = 1 - exp\,[-\,exp(X_i\beta)] \tag{8}$$

where $\boldsymbol{\pi_i} = \mathbf{Pr}\,(\boldsymbol{Y} = \mathbf{1}|\boldsymbol{X_i})$.

2.6 Parameter Estimation of Binary Regression Model

The estimation method used for the parameters is the Maximum Likelihood Estimation (MLE) method [18] [19]. MLE is commonly employed for model parameter estimation when the distribution is known [20]. The response variable (Y) is assumed to follow a binary criterion with two possible values, 0 and 1, and is modeled using a Bernoulli distribution. Below are the steps involved in the MLE process.

1. A random sample of binary observations $\boldsymbol{Y_1, Y_2, \dots Y_n}$ from surveys or classifications is collected.
2. Determining the likelihood function of the random variable $\boldsymbol{Y}$, $\mathrm{L}(\boldsymbol{L(\beta|Y_1, Y_2, \dots, Y_n)} = \boldsymbol{f(Y_1, Y_2, \dots Y_n)}$, is log-transformed to facilitate the estimation of the parameter β.
3. The log-likelihood is differentiated with respect to the parameter $\boldsymbol{\beta}$, and the resulting equation is set to zero to find the value of $\boldsymbol{\beta}$ that maximizes the likelihood.
4. If a closed-form solution is not obtained, the maximum likelihood estimator is found using numerical optimization methods such as the Newton-Raphson iteration, which involves calculating the gradient vector $\boldsymbol{g(\beta)}$ (1st derivative) and the Hessian matrix $\boldsymbol{H(\beta)}$ (2nd derivative) of the log-likelihood function.

2.7 Inference on Binary Regression Models

To test whether a group of predictor variables simultaneously influences the response variable, the following hypothesis is used [21]: $\boldsymbol{H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0}$; $\boldsymbol{H_1}$: at least one $\boldsymbol{\beta_j \neq 0, for\ j = 1, 2, \dots, p}$. The statistic used to test this hypothesis is based on the likelihood ratio test, as follows:

$$G = -2\,ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}(\hat{\pi}_i)^{y_i}(1-\hat{\pi}_i)^{1-y_i}}\right] \tag{9}$$

Where $\boldsymbol{n_0} = \sum_{i=1}^{n}(\mathbf{1} - \boldsymbol{y_i})$ is the number of observations with $\boldsymbol{Y = 0}$, and $\boldsymbol{n_1} = \sum_{i=1}^{n}\boldsymbol{y_i}$ is the number of observations with $\boldsymbol{Y = 1}$. The test statistics G follows a $\boldsymbol{\chi^2}$ distribution with $\boldsymbol{p}$ degreed of freedom under $\boldsymbol{H_0}$. If the calculated $\boldsymbol{G > X_\alpha^2(p)}$, where α is the significance level, then $\boldsymbol{H_0}$ is rejected, indicating that the model is significant. For individual variable significance, the hypothesis test is $\boldsymbol{H_0: \beta_j = 0; H_1}$: at least one $\boldsymbol{\beta_j \neq 0, j = 1, 2, \dots, p}$. To test this hypothesis, the Wald test statistic is used:

$$Z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} j = 1,2,\dots,p \tag{10}$$

The critical region for significance is to reject $\boldsymbol{H_0}$ if $\boldsymbol{|Z_j| > Z_{1-\alpha/2}}$. For large samples, the value of $\boldsymbol{Z_j}$ confidence interval $\boldsymbol{(1-\alpha)100\%}$ for $\boldsymbol{\beta_j}$ is obtained as follows

$$\hat{\beta}_j \pm Z_{\frac{a}{2}} s(\hat{\beta}_j) \tag{11}$$

approximately standard normal $\boldsymbol{(N(0,1))}$.

2.8 Multicollinearity Test

The multicollinearity test refers to the linear relationship between independent variables in multiple regression. The multicollinearity test is used to examine the relationship/correlation between each variable. A good regression model should not have any correlation among the independent variables. The Tolerance value and Variance Inflation Factor (VIF) are used to detect the presence of multicollinearity. Below are the decision guidelines [22].

1. If the Tolerance value is greater than 0.10, then multicollinearity does not occur in the regression model."
2. If the VIF value is < 10.00, then multicollinearity does not occur in the regression model.

2.9 Association Measures

Association measures are parameters used to estimate the magnitude of risk associated with determinants of disease occurrence. This risk measure represents the proportion of cases among those exposed to the determinant of disease occurrence. An increase in the estimated risk value is directly proportional to the dominance of the factor influencing the disease occurrence [23]. A pair of observations is considered concordant if the observation with an observed response value of 1 has a higher predicted probability and is categorized as 1 according to the model. A pair of observations is said to be discordant if the observation with an observed response value of 1 has a lower predicted probability and is categorized as 0 according to the model. A pair is considered a tie if the observations have the same predicted probability. The following formula can calculate the concordant, discordant, and ties values.

$$Somers'D = \frac{nc - nd}{nc + nd + nt} \tag{12}$$

$$Goodman - Kruskal\ Gamma = \frac{nc - nd}{nc + nd} \tag{13}$$

$$Kendall's\ Tau - a = \frac{nc - nd}{(0{,}5 \times N \times N - 1)} \tag{14}$$

where, $\boldsymbol{nc}$ = the number of concordant pairs; $\boldsymbol{nd}$ = the number of discordant pairs; $\boldsymbol{nt}$ = the number of tied pairs; $\boldsymbol{N}$= the total number of observations.

2.10 Apparent Error Rate

Apparent Error Rate (APPER) is a value used to assess the likelihood of errors in classifying objects. The APPER value can be calculated using the following formula [24]:

$$APPER = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}\ x\ 100\% \tag{15}$$

where, $\boldsymbol{n_{11}}$= the number of failure events observed and classified as failures in the predictions; $\boldsymbol{n_{12}}$ = the number of failure events observed and classified as successes in the predictions; $\boldsymbol{n_{21}}$ = the number of success events observed and classified as failures in the predictions; $\boldsymbol{n_{22}}$ = the number of success events observed and classified as successes in the predictions. Therefore, the classification accuracy of the binary logit model is given by 1-$\boldsymbol{APPER}$.

3. RESULTS AND DISCUSSION

3.1 The Data Used

This study uses secondary data obtained from the Kaggle website [25]. The data utilized is from 2024 and pertains to the diagnosis of Alzheimer's disease in elderly individuals (aged 60 years and above) worldwide, along with several factors that may contribute to Alzheimer's disease, such as age, family history, head injury, hypertension, memory complaints, and behavioral disturbances. The dataset used in this study consists of 75 observations, as presented in Table 1.

Table 1. Research Data

Y	X1	X2	X3	X4	X5	X6
0	82	0	0	0	0	0
0	73	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	75	1	0	0	1	0
0	82	0	0	0	0	0

***Data source**: website kaggle*

Although the dataset comprises only 75 observations, which is relatively small for a Generalized Linear Model (GLM) with six predictor variables, the sample size still meets the minimum threshold for statistical modeling, considering the rule-of-thumb of having at least 10 observations per predictor variable.

3.2 Research Variables

The variables used in this study consist of a response variable and six predictor variables. The response variable is a binary categorical variable, representing the diagnosis of Alzheimer's disease in elderly individuals worldwide in 2024. Meanwhile, the predictor variables are those suspected to contribute to the development of Alzheimer's disease. The variables used are presented in Table 2, as follows.

Table 2. Research Data

Code	Variable	Scale/Category	Count
		Response Variable	
Y	Diagnosis of Alzheimer's Disease	0 = Does not have an Alzheimer's diagnosis	49
		1 = Has an Alzheimer's diagnosis	26
		Variables Predictor	
X_1	Age	Ratio	75
X_2	Family History	0 = No family history of Alzheimer's	57
		1 = Family history of Alzheimer's	18
X_3	Head Injury	0 = No head injury	66
		1 = Has had a head injury	9
X_4	Hypertension	0 = Not a hypertension sufferer	66
		1 = Hypertension sufferer	99
X_5	Memory Complaints	0 = No memory complaints	56
		1 = Has memory complaints	19
X_6	Behavioral Disturbances	0 = No behavioral disturbances	65
		1 = Has behavioral disturbances	10

3.3 Descriptive Statistics

Table 3. Descriptive Statistics

	X1	X2
Y	1	26
	0	49
	Total	75

Based on Table 3, it can be seen that the total number of data used in this study is 75. Additionally, the coding for the response variable, where a diagnosis of Alzheimer's is represented by 1 and no diagnosis is represented by 0, resulted in 26 elderly individuals being diagnosed with Alzheimer's.

3.4 Model Fit Test for Logit, Probit, and Gompit

Before performing the analysis with the logit, probit, and gompit link approaches, it is necessary to conduct a model fit test to determine whether using the logit, probit, and gompit models is statistically appropriate. The hypothesis formulation used for testing model: H_0: The logit/probit/gompit model is a good fit (appropriate model); H_1 : The logit/probit/gompit model is not a good fit (inappropriate model).

Table 4. Model Fit Test for Logit, Probit, and Gompit

Test	Logit	Probit	Gompit
Deviance	0.532	0.532	0.542
Pearson	0.283	0.320	0.270
Hosmer-Lemeshow	0.471	0.776	0.709

Table 4 shows that the p-value for all three model fit tests (logit, probit, and gompit) is greater than the significance level of α = 5%. Therefore, the decision is to fail to reject H_0. As a result, it can be concluded that the logit, probit, and gompit models are appropriate for modeling the data.

Table 5. R-Square and AIC for Logit, Probit, and Gompit Models

Value	Logit	Probit	Gompit
Deviance R-Square	31.40%	31.41%	31.71%
AIC	80.41	80.40	80.10

Table 5 shows the best model is the gompit model, with the most significant deviance R-squared value of 31.71%. This indicates that the predictor variables included in the model can explain 31.71% of the variability in the response variable. The remaining 68.29% of the variability is explained by factors that were not included in the model. This suggests that, while the Gompit model provides a good fit, other factors still influence the response variable that the model does not capture.

3.5 Beta Coefficients and Variance Inflation Factor (VIF)

The tolerance value and the Variance Inflation Factor (VIF) value can be used to determine whether multicollinearity is present in the regression model. A tolerance value of 0.10 or a VIF value greater than 10 is the cutoff value that is employed. The logit, probit, and gompit models' regression coefficient analysis results, standard errors, and VIF values are shown below.

Table 6. Parameter Coefficient and VIF Value

Term	Logit			Probit			Gompit		
	Coefficient	SE	VIF	Coefficient	SE	VIF	Coefficient	SE	VIF
Constant	0.42	2.66		0.23	1.50		-0.82	1.96	
X_1	-0.0274	0.0351	1.05	-0.0161	0.0197	1.04	-0.0127	0.0257	1.04
X_2	-0.679	0.822	1.07	-0.324	0.441	1.06	-0.608	0.620	1.05
X_3	-0.61	1.03	1.07	-0.358	0.613	1.05	-0.213	0.699	1.04

Term	Logit			Probit			Gompit		
	Coefficient	SE	VIF	Coefficient	SE	VIF	Coefficient	SE	VIF
X_4	0.30	1.01	1.08	0.114	0.596	1.07	0.483	0.741	1.07
X_5	3.031	0.727	1.13	1.812	0.409	1.09	2.201	0.488	1.24
X_6	2.159	0.857	1.06	1.293	0.512	1.04	1.715	0.617	1.22

Table 6 shows that the VIF values for all predictor variables in the logit, probit, and gompit models are less than 10. Therefore, it can be concluded that there is no multicollinearity between the predictor variables, and the data analyzed meet the model assumptions. The model parameters are as follows:

Table 7. Logit, Probit, and Gompit Model Parameters

i	Logit		Probit		Gompit	
	$\widehat{\beta}_i$	SE	$\widehat{\beta}_i$	SE	$\widehat{\beta}_i$	SE
0	0.42	2.66	0.23	1.50	-0.82	1.96
1	-0.0274	0.0351	-0.0161	0.0197	-0.0127	0.0257
2	-0.679	0.822	-0.324	0.441	-0.608	0.620
3	-0.61	1.03	-0.358	0.613	-0.213	0.699
4	0.30	1.01	0.114	0.596	0.483	0.741
5	3.031	0.727	1.812	0.409	2.201	0.488
6	2.159	0.857	1.293	0.512	1.715	0.617

The logit, probit, and gompit model equations generated to model the effect of age, hereditary history, head injury, hypertension, memory complaints, and behavioral disturbances on the diagnosis of alzheimer's disease.

Table 8. Logit, Probit, and Gompit Model Equations

Test	Logit
Logit	$ln\left(\frac{\pi_i}{1-\pi_i}\right) = 0.42 - 0.0274X_1 - 0.679X_2 - 0.61X_3 + 0.30X_4 + 3.031X_5 + 2.159X_6$
Probit	$\Phi^{-1}(\pi_i) = 0.23 - 0.0161X_1 - 0.324X_2 - 0.358X_3 + 0.114X_4 + 1.812X_5 + 1.293X_6$
Gompit	$ln(-ln(1-\pi_i)) = -0.42 - 0.0127X_1 - 0.608X_2 - 0.213X_3 + 0.483X_4 + 2.201X_5 + 1.715X_6$

3.6 Simultaneous Parameter Significance Test

To test whether the predictor variables simultaneously affect the response variable, the simultaneous parameter significance test is used, with the following hypothesis: H_0: $\beta_1 = \cdots = \beta_6 = 0$; H_1: at least one $\beta_i \neq 0$, for $i = 1,2,3,4,5,6$.

Table 9. Simultaneous Parameter Significance Test

Model	G	*P-value*
Logit	30.40	0.000
Probit	30.40	0.000
Gompit	30.70	0.000

Table 9 shows that all three models (logit, probit, and gompit) have p-values below the 5% significance level, leading to the rejection of H_0, indicating at least one significant predictor. The Gompit model has the highest G value, suggesting it has the most significant β coefficient.

3.7 Partial Parameter Significance Test

To identify which predictor variables significantly affect the response, a partial parameter significance test is conducted following the confirmation from the simultaneous test. H_0: $\beta_1 = 0$; H_1: $\beta_i \neq 0$, for $i = 1,2,3,4,5,6$.

Table 10. Partial Parameter Significance Test

β	Logit	Probit	Gompit
β_1	p-value: 0.432; wald test: -0.781	p-value: 0.412; wald test: -0.817	p-value: 0.622; wald test: -0.494
β_2	p-value: 0.400; wald test: -0.083	p-value: 0.458; wald test: -0.735	p-value: 0.304; wald test: -0.980
β_3	p-value: 0.548; wald test: -0.592	p-value: 0.552; wald test: -0.584	p-value: 0.757; wald test: -0.305
β_4	p-value: 0.762; wald test: 0.297	p-value: 0.848; wald test: 0.191	p-value: 0.523; wald test: 0.652
β_5*	p-value: 0.000; wald test: 4.169	p-value: 0.000; wald test: 4.430	p-value: 0.000; wald test: 4.510
β_6	p-value: 0.008; wald test: 2.519	p-value: 0.009; wald test: 2.525	p-value: 0.006; wald test: 2.780

Table 10 shows that in the logit, probit, and gompit models the variables that have a significant effect on the diagnosis of alzheimer's are memory complaints and habit problems variables that have a significant effect on the diagnosis of alzheimer's are memory complaints and behavioral disturbances variables because they have a p-value that is less than α = 5% and a test statistic value that is greater than $Z_{\frac{\alpha}{2}} = 1.96$.

3.8 Re-Modeling

Once it is established that the only variables that significantly impact the diagnosis of Alzheimer's are memory complaints and behavioral disturbances, re-modeling is done by eliminating the variables that do not significantly impact the diagnosis, the results of which are in Table 11.

Table 11. Partial Parameter Significance Test

Value	Logit	Probit	Gompit
Deviance R-Square	29.78%	29.68%	30.01%
Deviance R-Square (Adajusted)	27.71%	27.79%	27.95%
AIC	73.98	73.90	73.75

The decrease in Deviance R-Square after removing insignificant predictor variables occurs because, although these variables are not statistically significant, they still contribute slightly to explaining the variability of the response variable. Removing them simplifies the model and reduces the explained variation, leading to a lower Deviance R-Square. Parameter results after re-modeling are in Table 12.

Table 12. Parameter Re-Modeling

i	Logit		Probit		Gompit	
	$\widehat{\beta}_i$	SE	$\widehat{\beta}_i$	SE	$\widehat{\beta}_i$	SE
0	-1.810	0.408	-1.076	0.220	-1.871	0.377
1 (Memory Complaints)	2.944	0.696	-1.770	0.397	2.199	0.487
2 (Memory Complaints)	2.160	0.839	1.295	0.504	1.709	0.630

The resulting logit, probit, and gompit model equations for modeling memory complaints and behavioral disturbances to the diagnosis of alzheimer's disease are in Table 13.

Table 13. Model Equation After Re-Modeling

Test	Logit
Logit	$ln\left(\frac{\pi_i}{1-\pi_i}\right) = -1.810 + 2.944X_1 + 2.160X_2$
Probit	$\Phi^{-1}(\pi_i) = -1.076 + 1.770X_1 + 1.295X_2$
Gompit	$ln(-ln(1-\pi_i)) = -1.871 + 2.199X_1 + 1.709X_2$

The logit model can be analyzed to compare the probability of success and the probability of failure, namely by looking at the odds ratio. The odds ratio value for each predictor variable after re-modeling is shown in Table 14.

Table 14. Odds Ratio

Predictor Variables	Categorical	
	Odds Ratio	95%CI
X_1	18.9998	(4.8571 ; 74.3232)
X_2	8.6736	(1.6737 ; 44.9489)

Based on the table above, the odds ratio value for each predictor variable is obtained as follows.

1. Odds ratio value for predictor variable X_1

$$OR_1 = \frac{\pi(1)}{1-\pi(1)} : \frac{\pi(0)}{1-\pi(0)} = exp(\beta_1) = exp(2.944) = 18.99$$

The odds of patients diagnosed with alzheimer's disease who have memory complaints are 18.99 times greater than patients who do not have memory complaints greater than patients who do not have memory complaints.

2. Odds ratio value for predictor variable X_2

$$OR_2 = \frac{\pi(1)}{1-\pi(1)} : \frac{\pi(0)}{1-\pi(0)} = exp(\beta_2) = exp(2.160) = 8.67$$

The odds of a patient being diagnosed with Alzheimer's disease were 8.67 times higher among those with behavioral disturbances such as [e.g., smoking or irregular sleep patterns] compared to those without such habits.

3.9 Marginal Effect

The marginal effect value is needed to determine the effect of significant predictor variables on the response variable on the response variable [26]. The value of the marginal effect on the first observation with the value of $X_1 = 0$ and $X_2 = 0$, is in Table 15.

Table 15. Marginal Effect Value

Marginal Effect	Logit	Probit	Gompit
$\partial P(Y=0)/\partial X_1 = -\phi(\gamma - \beta^T x)\beta_1$	-0.22816	-0.3958	-0.1524
$\partial P(Y=0)/\partial X_1 = \phi(\gamma - \beta^T x)\beta_1$	0.22816	0.3958	0.1524
$\partial P(Y=0)/\partial X_2 = -\phi(\gamma - \beta^T x)\beta_2$	-0.1674	-0.2896	-0.1184
$\partial P(Y=0)/\partial X_1 = \phi(\gamma - \beta^T x)\beta_2$	0.1674	0.2896	0.1184

The initial patient's chance of not receiving a diagnosis of Alzheimer's disease will drop if their level of memory or behavioral disturbances increases by one unit, according to the negative marginal impact value. On the other hand, the first patient's positive marginal impact suggests that the likelihood of the patient receiving an Alzheimer's disease diagnosis will rise with every unit increase in memory complaints or behavioral disturbances.

3.10 Classification Accuracy

Classification Accuracy using the APPER (Apparent Error Rate) value is as follows.

Table 16. Marginal Effect Value

Model	Logit
Logit	81.3%
Probit	81.3%
Gompit	81.3%

Table 16 explains that the logit, probit, and gompit models have the same classification accuracy of 81.3% which is 81.3%. Overall, the three models were able to correctly classify observations with 81.3% of the total cases, both those in the category "No (0)" diagnosed with alzheimer's disease and in the category "Yes (1)" diagnosed with alzheimer's disease alzheimer's disease or in the category "Yes (1)" diagnosed with

alzheimer's disease. Furthermore, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) values were also examined to assess the discriminative ability of the models. The results obtained are attached below.

Table 17. Area Under the Curve Values for Logit, Probit, and Gompit Models

Model	AUC
Logit	0.8567
Probit	0.8579
Gompit	0.8532

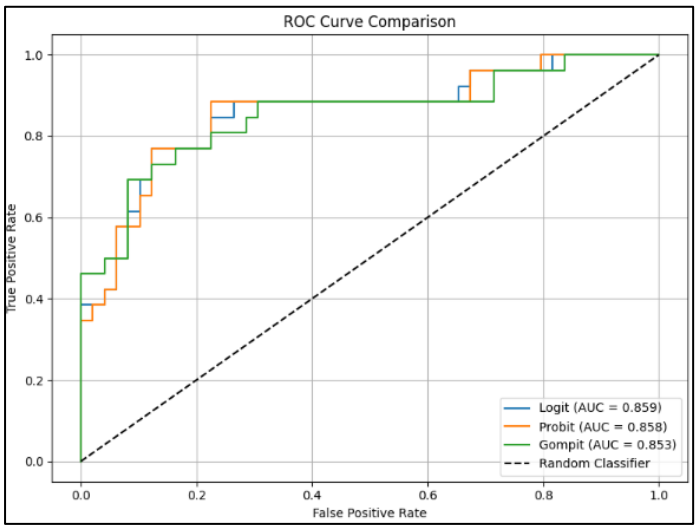

Figure 1. ROC Curve Comparison of Logit, Probit, and Gompit Regression Models

As presented in **Table 17** and **Figure 1**, the logit model obtained the highest AUC value (0.8587), followed closely by the probit model (0.8579), and then the gompit model (0.8532). These results indicate that all models demonstrate excellent classification performance (AUC > 0.85), with the logit model performing slightly better in distinguishing between classes.

3.11 Selection of the Best Model

After the analysis, including re-modeling on the three models, the best model can be selected to model data on the diagnosis of alzheimer's disease and factors suspected of causing alzheimer's, using the r-square value factors that are thought to cause alzheimer's, using R-squared value. Based on **Table 11**, the best model is the Gompit model with an R-squared value of 30.01%, greater than the logit model of 29.78% and the probit model of 29.86%.

In addition to having the highest R^2 value, logistic regression with the Gompit link also showed the best performance in explaining data variability (highest Deviance $R^2 = 30.01\%$) and model efficiency (lowest AIC = 73.75). Although its AUC value (0.8532) was slightly lower than the logit model (0.8587), the difference was very small and not practically significant. Therefore, the Gompit model is considered the most appropriate for analyzing factors that influence the incidence of Alzheimer's disease in this study.

3.12 Association Size

How well the model predicts the data is shown by the concordant and discordant pairs. The ability to forecast the model improves with the number of concordant pairs.

Table 18. Marginal Effect Value

Pair	Number	Percentage	Summary Measures	Value
Concordant	855	67.1%	Somer's D	0.62
Discordant	65	5.1%	Goodman-Kruskal Gamma	0.86
Ties	354	27.8%	Kendall's Tau-a	0.28
Total	1274	100.0%		

Based on **Table 18**, the Somers' D value of 0.62 indicates moderate discriminatory power, consistent with its use in evaluating binary logistic regression models. The Goodman-Kruskal Gamma value of 0.86 shows a strong association between observed and predicted outcomes, in line with findings that Gamma values above 0.80 indicate strong predictive ability. The Kendall's Tau-a value of 0.28, although lower due to 27.8% tied pairs, still reflects a meaningful relationship. These findings validate the model's effectiveness in identifying Alzheimer's risk factors and support its theoretical foundation.

4. CONCLUSION

Based on the results of the analysis, the following conclusions were drawn:

1. Of the 75 people diagnosed with Alzheimer's disease, 26 are elderly.
2. In the logit, probit, and gompit models, the variables of memory complaints and behavioral disorders showed a significant effect on Alzheimer's diagnosis.
3. The deviance R-square values of the three models used were 29.78% (logit), 29.86% (probit), and 30.01% (gompit), respectively, with the same level of classification accuracy of 81.3%.
4. Among the three models tested, the gompit model performed best because it had the highest deviance R-square value. This indicates that the gompit model can better explain data variation than the logit and probit models.
5. The findings of this study have important real-world applications in public health and clinical practice. The identification of memory complaints and behavioral disorders as significant predictors of Alzheimer's disease allows medical personnel to conduct early and targeted screening. These indicators can also be utilized in developing clinical decision support systems to assess patient risk more accurately.

AUTHOR CONTRIBUTIONS

Ardi Kurniawan: Conceptualization, Data Curation, Supervision, Funding Acquisition. Gabriella Agnes Budijono: Formal Analysis, Investigation, Visualization, Writing – Original Draft, Writing – Review and Editing, Software, Validation. Kimberly Maserati Siagian: Project Administration, Visualization, Writing – Review and Editing, Methodology. Adrian Wahyu Abdillah: Resources, Writing - Original Draft, Methodology. All authors discussed the results and contributed to the final manuscript.

FUNDING STATEMENT

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

ACKNOWLEDGMENT

The authors sincerely thank the Department of Statistics, Faculty of Science and Technology, Universitas Airlangga, for its academic support, and the Kaggle platform for providing access to valuable data. Appreciation is also given to all individuals and institutions who contributed, either directly or indirectly, to the successful completion and publication of this research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] L. Arwin and J. N. Pratiwi, "PERAN NEUROPROTEKTOR ASTAXANTHIN DALAM PENCEGAHAN PENYAKIT ALZHEIMER," *Jurnal Ilmu Keperawatan Jiwa,* pp. 47-52, 2020, doi: https://doi.org/10.32584/jikj.v3i1.469.

[2] S. Sahana, R. Kumar, S. Nag, R. Paul, I. Chatterjee and N. Guha, "A REVIEW ON ALZHEIMER DISEASE AND FUTURE PROSPECTS," *World Journal of Pharmacy and Pharmaceutical Sciences,* vol. 9, no. 9, pp. 1276-1285, 2020, doi: https://doi.org/10.55544/jrasb.1.2.9.

[3] R. W. Werdhana, "KLASIFIKASI GEN YANG TERKAIT SINDROM ALZHEIMER MENGGUNAKAN MTODE NAIVE BAYES CLASSIFIER, BINARY LOGISTIC REGRESSION DAN LOGISTIC REGRESSION ENSEMBLE," Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Speuluh Nopember, Surabaya, 2017.

[4] F. Akbar and Rahmaddeni, "KOMPARASI ALGORITMA MACHINE LEARNING UNTUK MEMPREDIKSI PENYAKIT ALZHEIMER," *Jurnal Politeknik Caltex Riau,* 2022, doi: https://doi.org/10.35143/jkt.v8i2.5713

[5] O. Diao, P.-A. Absil and M. Diallo, "GENERALIZED LINEAR MODELS TO FORECAST MALARIA INCIDENCE IN THREE ENDEMIC REGIONS OF SENEGAL," *International Journal of Environmental Research and Public Health,* vol. 20, no. 13, pp. 1-27, 2023, doi: https://doi.org/10.3390/ijerph20136303.

[6] R. B. Prasetyo, H. Kuswanto, N. Iriawan and B. S. S. Ulama, "A COMPARISON OF SOME LINK FUNCTIONS FOR BINOMIAL REGRESSION MODELS WITH APPLICATION TO SCHOOL DROP-OUT RATES IN EAST JAVA," in *AIP Conference Proceedings*, Melville, NY, 2019, doi: https://doi.org/10.1063/1.5139815.

[7] T. Zhang, Y. Ning and D. Ruppert, "OPTIMAL SAMPLING FOR GENERALIZED LINEAR MODELS UNDER MEASUREMENT CONSTRAINTS," *Journal of the American Statistical Association,* vol. 117, no. 540, pp. 1896-1910, 2020, https://doi.org/10.48550/arXiv.1907.07309.

[8] K. F. Arnold, V. Davies, M. d. Kamps, P. W. Tennant, J. Mbotwa and M. S. Gilthorpe, "REFLECTION ON MODERN METHODS: GENERALIZED LINEAR MODELS FOR PROGNOSIS AND INTERVENTION," *International Journal of Epidemiology,* vol. 49, no. 6, pp. 2074-2082, 2020, doi: https://doi.org/10.1093/ije/dyaa049.

[9] A. Al-Finatunni'mah and T. Nurhidayati, "PELAKSANAAN SENAM OTAK UNTUK PENINGKATAN FUNGSI KOGNITIF PADA LANSIA DENGAN DEMENSIA," *Ners Muda,* vol. I, 2020, doi: https://doi.org/10.26714/nm.v1i2.5666.

[10] A. Rosyida and T. B. Sasongko, "DETEKSI DINI PENYAKIT ALZHEIMER DENGAN ALGORITMA C4.5 BERBASIS BPSO (BINARY PARTICLE SWARM OPTIMIZATION)," *SISFOKOM (Sistem Informasi dan Komputer),* 2023, doi: https://doi.org/10.32736/sisfokom.v12i3.1716.

[11] M. Delporte, S. Fieuws, G. Molenberghs, G. Verbeke, S. S. Wanyama, E. Hatziagorou and C. D. Boeck, "A JOINT NORMAL-BINARY (PROBIT) MODEL," *International Statistical Review,* vol. 90, no. 3, pp. 410-434, 2022, doi: https://doi.org/10.1111/insr.12532.

[12] S. Saidi, N. Herawati and K. Nisa, "MODELING WITH GENERALIZED LINEAR MODEL ON COVID-19: CASES IN INDONESIA," *International Journal of Electronics and Communications System,* vol. 1, no. 1, pp. 25-32, 2021, doi: https://doi.org/10.24042/ijecs.v1i1.9299.

[13] H. Turner, "INTRODUCTION TO GENERALIZED LINEAR MODELS," ESRC National Centre for Research Methods, UK and Department of Statistcs University of Warwick, UK, 2008.

[14] E. C. Norton and E. D. Bryan, "LOGODDSANDTHEINTERPRETATION OF LOGIT MODELS," *Health Services Research,* vol. 53, no. 2, pp. 859-878, 2018, doi: https://doi.org/10.1111/1475-6773.12712.

[15] C. Johnston, J. McDonald and K. Quist, "A GENERALIZED ORDERED PROBIT MODEL," *Communications in Statistics - Theory and Methods,* vol. 48, no. 20, p. 5128–5144, 2019, doi: https://doi.org/10.1080/03610926.2019.1565780.

[16] C. J. McCabe, M. A. Halvorson, K. M. King, X. Cao and D. S. Kim, "INTERPRETING INTERACTION EFFECTS IN GENERALIZED LINEAR MODELS OF NONLINEAR PROBABILITIES AND COUNTS," *Multivariate Behavioral Research,* vol. 57, no. 2, pp. 243-263, 2022, doi: https://doi.org/10.1080/00273171.2020.1868966.

[17] D. Ariyanto, A. Sofro, A. N. Hanifah, J. Prihanto, D. A. Maulana and R. W. Romadhonia, "LOGISTIC AND PROBIT REGRESSION MODELING TO PREDICT THE OPPORTUNITIES OF DIEBETES IN PROSPECTIVE ATHLETES," *BAREKENG : Journal of Mathematics and Its Application,* vol. 18, no. 3, pp. 1391-1402, 2024, doi: https://doi.org/10.30598/barekengvol18iss3pp1391-1402.

[18] X. Peng, C. Lei and X. Sun, "COMPARISON OF LETHAL DOSES CALCULATED USING LOGIT/PROBIT–LOG(DOSE) REGRESSIONS WITH ARBITRARY SLOPES USING R," *Journal of Economic Entomology,* vol. 114, no. 3, pp. 1345-1352, 2021, doi: https://doi.org/10.1093/jee/toab044.

[19] C. Rainey and K. McCaskey, "ESTIMATING LOGIT MODELS WITH SMALL SAMPLES," *Political Science Research and Methods,* vol. 9, no. 4, pp. 754-769, 2021, doi: https://doi.org/10.1017/psrm.2021.9.

[20] I. Kosmidis, E. C. K. Pagui and N. Sartori, "MEAN AND MEDIAN BIAS REDUCTION IN GENERALIZED LINEAR MODELS," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 84, no. 2, pp. 380-405, 2022, doi: https://doi.org/10.1007/s11222-019-09860-6.

[21] D. W. Hosmer and S. Lemeshow, APPLIED LOGISTIC REGRESSION, New York: Wiley, 2000, doi: https://doi.org/10.1002/0471722146.

[22] S. D. Puspa, J. Riyono and F. Puspitasari, "ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI PEMAHAMAN KONSEP MATEMATIS MAHASISWA DALAM PEMBELAJARAN JARAK JAUH PADA MASA PANDEMI COVID-19," *Jurnal Cendekia: Jurnal Pendidikan Matematika,* 2021, doi: https://doi.org/10.31004/cendekia.v5i1.533.

[23] Amelia, E. Agustina, M. Aziyannoor and Rifaldi, "ASOSIASI LINGKUNGAN FISIK RUMAH SEBAGAI FAKTOR RISIKO KEJADIAN TB PARU DI INDONESIA," *Jurnal Kesehatan Tambusai,* vol. 4, no. 3, 2023, doi: https://doi.org/10.31004/jkt.v4i3.15481.

[24] M. A. Suhendra, I. Dwi and S. , "KETEPATAN KLASIFIKASI PEMBERIAN KARTU KELUARGA SEJAHTERA DI KOTA SEMARANG MENGGUNAKAN METODE REGRESI LOGISTIK BINER DAN METODE CHAID," *Jurnal Gaussian,* 2020, doi: https://doi.org/10.14710/j.gauss.v9i1.27524.

[25] R. E. Kharoua, "ALZHEIMER'S DISEASE DATASET," 2024. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data. [Accessed 5 September 2024].

[26] Ö. İ. Güneri and B. Durmuş, "DEPENDENT DUMMY VARIABLE MODELS: AN APPLICATION OF LOGIT, PROBIT AND TOBIT MODELS ON SURVEY DATA," *International Journal of Computational and Experimental Science and Engineering (IJCESEN),* vol. 6, no. 1, pp. 63-74, 2020, https://doi.org/10.22399/ijcesen.666512.18