



Available online at <http://scik.org>

Commun. Math. Biol. Neurosci. 2026, 2026:102

<https://doi.org/10.28919/cmbn/10189>

ISSN: 2052-2541

ADVANCING DENGUE MORBIDITY PREDICTION IN INDONESIA: A HYBRID MARS–RANDOM FOREST FRAMEWORK FOR JAVA AND BALI

BAMBANG WIDJANARKO OTOK*, EVYANA WIJAYANTI, SRI SULASTRI

Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Kampus

ITS-Sukolilo, Surabaya 60111, Indonesia

Copyright © 2026 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: Dengue hemorrhagic fever (DHF) is an endemic disease in Indonesia with a steadily rising number of cases, particularly on the islands of Java and Bali, which account for the highest number of cases nationwide. The complexity of the relationships between environmental, social, and health factors means that modeling DHF morbidity rates requires methods capable of capturing nonlinear relationships. This study aims to compare the performance of the Multivariate Adaptive Regression Splines (MARS), Random Forest, and MARS-Random Forest hybrid methods in modeling DBD morbidity rates on the islands of Java and Bali. The data used are aggregated regency/city-level data from 2024 sourced from the Central Statistics Agency, with the response variable being the DBD morbidity rate and seven predictor variables. The preprocessing stage included handling missing data, removing outliers, applying a natural logarithmic transformation to the response variable, Z-score standardization, and splitting the data using a stratified split with an 80:20 ratio. The MARS model was built using a combination of basis function parameters, interaction order, and minimum observations, while the Random Forest and hybrid models were optimized via grid search based on the Out-of-Bag Mean Squared Error (OOB MSE) value. The results of the study show that the MARS

*Corresponding author

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

Received August 2, 2026

model produces four basis functions that describe the nonlinear relationships among the variables. On the test data, Random Forest delivered the best performance with an MSE of 0.9675, an RMSE of 0.9836, an MAE of 0.8014, and an R^2 of -0.0258. Meanwhile, the MARS-Random Forest hybrid model produced an MSE of 1.2454, and the MARS model produced an MSE of 1.7225. Thus, Random Forest is the best model for modeling dengue fever incidence on the islands of Java and Bali, while MARS retains an advantage in terms of interpretability through the basis functions it generates.

Keywords: morbidity rates; dengue hemorrhagic fever; MARS-random forest hybrid; multivariate adaptive regression splines; random forest.

2020 AMS Subject Classification: 62G08, 62J12.

1. INTRODUCTION

Dengue hemorrhagic fever (DHF) remains a significant global public health issue, particularly in tropical and subtropical regions. Globally, about half of the world's population is at risk of dengue infection, with an estimated 100–400 million cases each year, and more than 100 countries report endemic transmission [1]. The high incidence rate and the potential for severe complications make DBD a disease that requires serious attention in control and prevention efforts.

In Indonesia, dengue fever is an endemic disease that occurs in nearly all provinces and exhibits fluctuating trends each year. As of July 2024, there were 149,866 reported cases with 884 deaths spread across 465 regencies and cities [2]. Compared to the same period the previous year, the number of cases has nearly tripled. This situation indicates that the risk of dengue fever remains high and is influenced by various factors such as population density, community mobility, urbanization, and environmental conditions. The Java and Bali regions have the highest population and density rates nationwide [2] and have recorded the highest number of dengue fever cases compared to other regions. Their demographic characteristics and high mobility make these regions relevant for analysis in modeling dengue fever morbidity rates.

The spread of dengue fever is influenced by social, economic, and environmental factors such as population density, poverty levels, sanitation, access to clean water, and the availability of health facilities. Population density increases the potential for human contact with vectors [3],

while poor economic and sanitation conditions contribute to an increased risk of dengue fever outbreaks [4], [5]. However, most studies still focus on absolute case numbers, which may introduce a bias toward densely populated areas and do not fully represent the epidemiological risk level—which is more accurately measured by morbidity rates.

Several studies in Indonesia have applied linear, nonparametric, and spatial regression approaches to model dengue fever incidence. Linear regression has been used to predict the number of dengue cases; however, its ability to capture complex nonlinear relationships is limited [6]. Nonparametric spline regression can accommodate nonlinear associations among variables, but previous studies have not primarily emphasized the optimization of predictive performance [7]. Meanwhile, parametric approaches, including generalized Poisson regression and spatial negative binomial regression, are well suited to count data and overdispersion. Nevertheless, these models have generally focused on identifying significant determinants rather than maximizing prediction accuracy [8], [9].

Multivariate Adaptive Regression Splines (MARS) has been applied to dengue case modeling and has demonstrated its ability to capture nonlinear relationships while selecting relevant predictors [10]. Nevertheless, when used as a standalone approach, MARS may have limited predictive accuracy in the presence of highly complex data patterns [10]. However, the use of MARS as a standalone model still has limitations in improving accuracy for highly complex data. Advances in machine learning methods have driven the use of ensemble algorithms such as Random Forest, which are highly capable of handling nonlinear relationships and complex interactions among variables. Several studies have shown that Random Forest delivers good predictive performance in various health-related cases [11-13]. However, this method has limitations regarding model interpretability and has not yet been widely integrated with nonparametric variable selection approaches, such as MARS, in the context of modeling dengue fever morbidity rates in Indonesia.

Based on this literature review, this study proposes a hybrid MARS-Random Forest approach to produce a model that is both more accurate and informative in identifying the factors that influence dengue fever morbidity rates. Given that the distribution of dengue fever morbidity rates

is heavily skewed to the right due to extreme values in several regions, a natural logarithmic transformation of the response variable is required prior to the modeling process to stabilize the variance and improve the model's ability to capture data patterns more evenly.

This study aims to model dengue fever morbidity rates on the islands of Java and Bali using a hybrid MARS-Random Forest method. This approach is expected to identify the dominant factors influencing the risk of dengue fever outbreaks and to produce a predictive model that is more adaptable to the complexity of epidemiological data.

2. PRELIMINARIES

This study uses secondary data at the regency/city level in Java and Bali for the year 2024, sourced from the Central Statistics Agency (BPS). The response variable used is the incidence rate of Dengue Hemorrhagic Fever (DHF), while the predictor variables consist of population density, the percentage of the population living in poverty, the percentage of households with adequate sanitation, the percentage of households with access to safe drinking water, the number of health facilities, the percentage of households with their own latrines, and total annual rainfall. The initial dataset consisted of 128 observations. It then underwent a preprocessing stage that included handling missing data, removing outliers using the Interquartile Range (IQR) method, applying a natural logarithmic transformation to the response variable, standardizing the data using Z-scores, and splitting the data into training and test sets using a stratified split in an 80:20 ratio.

Modeling was performed using three methods: Multivariate Adaptive Regression Splines (MARS), Random Forest, and a MARS-Random Forest hybrid. The optimal parameters for the MARS model were determined via a grid search based on the Generalized Cross-Validation (GCV) score, while the parameters for the Random Forest and hybrid models were determined via a grid search based on the Out-of-Bag Mean Squared Error (OOB MSE). Next, the best model from each method was built using training data and evaluated on test data using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The evaluation results were then compared to determine which method provided the best performance in modeling dengue fever incidence on the islands of Java and Bali.

2.1 Multivariate Adaptive Regression Splines (MARS)

MARS is a flexible and adaptive approach in non-linear regression modeling. The model can capture the complex relationship between predictor and response variables by constructing adaptively generated basis functions. The following equation describes the MARS basis function used to predict the response value [14], [15]:

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

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

2.2 Random Forest

The random forest model is the preeminent method for regression and classification within decision tree learning. It is highly effective and concurrently exhibits superior regression accuracy compared to alternative regression techniques. The random forest (RF) model was introduced by L. Breiman in 1984 [16]. The stochastic model generates a multitude of uncorrelated decision trees during the training process. Upon constructing several decision trees, the model's output is derived by calculating the mean of the output values from each individual tree. The random forest model employs the learner bagging algorithm for the training of each individual tree [17]. In this process, bagging consistently chooses B bootstrap samples from the training dataset and constructs t_b trees utilizing the Gini impurity within these samples. Upon completion of the training procedure, the anticipated values for novel instances are determined by averaging the prediction outcomes from all regression trees through:

$$y = \frac{1}{B} \sum_{b=1}^B t_b(x) \quad (2)$$

3. MAIN RESULTS

3.1 Characteristics and Distribution Patterns of Dengue Fever Morbidity Rates

Understanding the characteristics of the data is an important first step before modeling. This stage aims to identify the nature of the data distribution, detect the presence of outliers, and understand the patterns of relationships among variables, which will subsequently form the basis for selecting the appropriate analytical approach. In this study, data characteristics were examined through three aspects: descriptive statistics for all study variables, the distribution pattern of dengue fever incidence rates across regencies and cities on the islands of Java and Bali, and an examination of the data distribution for each variable. The variables analyzed consist of one response variable—the dengue fever incidence rate (Y)—and seven predictor variables (X1–X7) representing four dimensions: demographic, socioeconomic, environmental sanitation, and health service availability.

Descriptive statistical analysis was conducted to provide an initial overview of the numerical characteristics of the dengue fever incidence data and all predictor variables across the 97 regencies/cities that served as the units of analysis following the data preprocessing stage. The statistical measures used included the mean as a measure of central tendency, the standard deviation as a measure of dispersion, and the minimum and maximum values to determine the range of the data. The complete results of the descriptive statistical analysis are presented in Table 1.

Table 1. Descriptive Statistics of Research Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
Y	118,326	128,777	5,300	840,940
X1	3.318,131	4.612,894	383,000	21.955,000
X2	8,804	3,915	2,230	20,830
X3	85,465	13,714	45,880	99,940
X4	96,093	4,603	77,920	100,000
X5	40,485	23,185	8,000	115,000
X6	89,923	6,374	59,790	100,000
X7	2.372,634	661,407	1.292,500	4.812,200

Based on Table 1, the incidence rates of dengue fever on the islands of Java and Bali in 2024 showed considerable variation, with an average of 118.326 per 100,000 population and a standard deviation of 128.777. The minimum value of 5,300 and the maximum value of 840,940 indicate significant differences in dengue fever incidence rates across regions. This variation suggests that regional characteristics and the factors influencing dengue fever incidence are not homogeneous. Among the predictor variables, population density (X1) exhibits the greatest variation, with values ranging from 383 to 21,955 people/km², while the number of health facilities (X5) ranges from 8 to 115 units. Variables expressed as percentages—such as the percentage of the poor (X2), access to adequate sanitation (X3), access to safe drinking water (X4), and toilet ownership (X6)—also show significant variation across regions. In addition, total annual rainfall (X7) ranges from 1,292,500 to 4,812,200 mm/year, reflecting differences in climatic conditions across the study area. This diversity in characteristics indicates that the relationship between dengue fever incidence and the factors influencing it is potentially complex and nonlinear, thus requiring a modeling method capable of accommodating these characteristics.

A correlation analysis was conducted to determine the strength and direction of the linear relationship between the response variable Y and each predictor variable, as well as the relationships among the predictor variables themselves in the training data. The results of the correlation analysis are presented in the form of a correlation matrix in Table 2 and visualized using a scatter plot along with the correlation values in Figure 1.

Table 2. Correlation Matrix Among Variables

	Y	X1	X2	X3	X4	X5	X6	X7
Y	1,000	0,049	- 0,197	- 0,008	0,006	- 0,152	0,225	0,059
X1	0,049	1,000	- 0,458	0,196	0,392	0,440	- 0,267	- 0,042
X2	- 0,197	- 0,458	1,000	- 0,300	- 0,287	- 0,187	- 0,171	- 0,068
X3	- 0,008	0,196	- 0,300	1,000	0,473	- 0,016	0,261	- 0,166
X4	0,006	0,392	- 0,287	0,473	1,000	0,047	- 0,048	- 0,137
X5	- 0,152	0,440	- 0,187	- 0,016	0,047	1,000	- 0,180	- 0,111
X6	0,225	- 0,267	- 0,171	0,261	0,048	- 0,180	1,000	0,136
X7	- 0,059	- 0,042	- 0,068	- 0,166	- 0,137	- 0,111	0,136	1,000

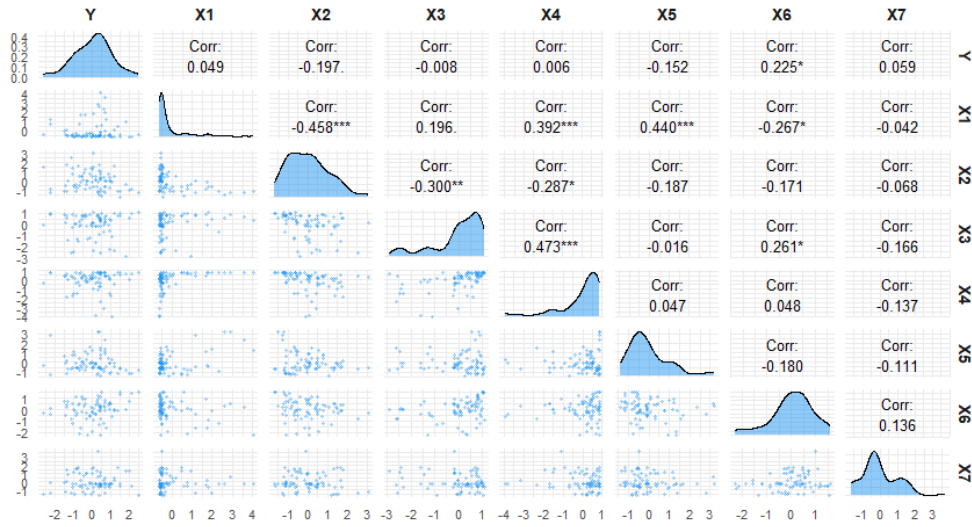


Figure 1. Scatter Plot and Correlation Among Variables

Correlation analysis shows that the linear relationship between dengue fever incidence and all predictor variables is relatively weak, with correlation coefficients ranging from -0.197 to 0.225. The strongest positive correlation with the incidence of dengue fever was found in the percentage of households with latrines (X6) at 0.225, while the strongest negative correlation was found in the percentage of the poor population (X2) at -0.197. These results indicate that there is no strong linear relationship between the response variable and each of the predictor variables.

Among the predictor variables, the strongest correlation was found between adequate sanitation (X3) and access to adequate drinking water (X4) at 0.473, followed by the relationship between population density (X1) and the number of health facilities (X5) at 0.440, as well as the negative relationship between population density (X1) and the percentage of the poor (X2) at -0.458. In general, there were no pairs of variables with very high correlation coefficients, indicating that there is no strong linear relationship among the predictor variables. These findings suggest that the factors influencing dengue fever morbidity rates are complexly interrelated; therefore, approaches capable of capturing nonlinear patterns—such as MARS, Random Forest, and the MARS-Random Forest hybrid—are relevant for use.

3.2 Preprocessing

Before the modeling process began, the data underwent a preprocessing stage that included data

quality checks, handling of missing data, outlier handling, data transformation, standardization, and data partitioning. For the total annual precipitation variable (X7), 16 missing data points were found out of a total of 128 observations. Missing data were handled using the listwise deletion method, reducing the number of observations to 112 regencies/cities. Next, outliers in variable X7 were detected using the Interquartile Range (IQR) method, as this method is robust against extreme values and does not require the assumption of a normal distribution.

Table 3. Results of Outlier Detection for Variable X7 Using the IQR Method

Component	Value
Q1	2.025,075
Q3	3.140.190
IQR	1.115,115
Lower Bound	352,403
Upper Bound	4.812,863
Number of Detected Outliers	15

Based on the IQR method, the lower bound was determined to be 352.403 mm/year and the upper bound 4,812.863 mm/year, resulting in the identification of 15 observations as outliers in the total annual rainfall variable (X7). After removing missing data and outliers, the number of data points used in the study was 97 regencies/cities.

An examination of the data distribution using histograms shows that the variables for dengue fever incidence (Y), population density (X1), and the number of health facilities (X5) exhibit distributions that are skewed to the right. Meanwhile, the percentage of access to safe drinking water (X4) and the percentage of households with latrines (X6) are skewed to the left. The variables percentage of the poor population (X2) and total annual rainfall (X7) show distributions that are relatively close to symmetrical, while the variable adequate sanitation (X3) exhibits a distribution pattern that forms two clusters. These results indicate that the distribution characteristics across variables do not fully satisfy the linearity assumption, thus requiring a transformation and standardization step prior to the modeling process.

Next, the response variable (Y) was transformed using $\ln(Y+1)$ to reduce the skewness of the distribution and stabilize the variance, while all variables were then standardized using the Z-

score method to ensure they were on the same scale before the modeling process began. Examples of the transformation and standardization results for the first five observations are presented in Table 4.

Table 4. The First Five Observations of the Data After Z-Score Standardization

Observation	Y	X1	X2	X3	X5	X6	X7
1	0,189	- 0,616	0,096	- 1,590	- 0,021	- 0,433	- 0,257
2	1,044	- 0,624	- 0,093	- 1,111	0,281	- 0,620	- 0,257
3	0,691	- 0,002	- 0,576	0,369	1,187	0,577	0,440
4	- 0,883	- 0,468	- 1,097	- 0,292	- 0,064	- 0,506	- 0,243
5	- 0,387	1,668	- 0,862	0,423	1,187	0,971	- 0,364

Based on Table 4, all variables are now on a comparable scale, with mean values close to zero and standard deviations close to one. A positive Z-score indicates that an observation is above the mean, while a negative Z-score indicates that the observation is below the mean for the variable in question. This transformation and standardization process aims to reduce the influence of scale differences between variables so that the model-building process proceeds more stably. The preprocessed data is then used in the data partitioning and modeling stages.

The data that had undergone all preprocessing stages was then divided into two subsets—the training set and the testing set—using the holdout method with a ratio of approximately 80:20. The data was split using a stratified split based on tertiles of the transformed response variable, ensuring that the distribution of response values in the training and testing sets remained representative of the overall data distribution. This approach aims to reduce distribution imbalances between the subsets so that the model training and evaluation processes can be conducted more representatively. The results of the data split are presented in Table 5.

Table 5. Results of the Division of Training Data and Test Data

Partition	Number of Observations	Proportion
Training Data	78	80,4%
Test Data	19	19,6%
Total	97	100%

Based on Table 5, the data were divided into 78 observations (80.4%) as training data and

19 observations (19.6%) as test data. The training data is used to build the model, while the test data is used to evaluate the model's ability to generalize to data not used during the training process. Although the number of observations decreased from 128 to 97 due to the handling of missing data and outliers, the sample size is still sufficient for modeling and evaluating the three methods used.

3.3 MARS Modeling

MARS modeling was performed to model the nonlinear relationship between dengue fever incidence and predictor variables. Optimal parameters were determined through a grid search, varying the maximum number of basis functions (BF), the degree of interaction, and the minimum span between knots (minspan). The best model was selected based on the lowest Generalized Cross Validation (GCV) value. The results of the parameter search are presented in Table 6.

Table 6. Results of the MARS Parameter Grid Search Based on GCV Values

No	BF	MI	MO	GCV	Number of Basis Functions
1	14	1	0	0.906645	3
2	14	1	1	0.888301	4
3	14	1	2	0.903081	4
4	14	1	3	0.889077	4
5	14	2	0	0.920450	2
6	14	2	1	0.909161	4
7	14	2	2	0.920450	2
8	14	2	3	0.909459	4
9	14	3	0	0.915492	5
10	14	3	1	0.862862	5
11	14	3	2	0.877946	5
12	14	3	3	0.863265	5
13	21	1	0	0.902409	4
14	21	1	1	0.888301	4
15	21	1	2	0.903081	4
16	21	1	3	0.889077	4
17	21	2	0	0.911266	6
18	21	2	1	0.885479	4
19	21	2	2	0.896446	4
20	21	2	3	0.885698	4
21	21	3	0	0.893338	7

No	BF	MI	MO	GCV	Number of Basis Functions
22	21	3	1	0.862516	5
23	21	3	2	0.877946	5
24	21	3	3	0.863265	5
25	28	1	0	0.893810	4
26	28	1	1	0.893692	3
27	28	1	2	0.903081	4
28	28	1	3	0.893702	3
29	28	2	0	0.911266	6
30	28	2	1	0.885479	4
31	28	2	2	0.896446	4
32	28	2	3	0.885698	4
33	28	3	0	0.914837	7
34	28	3	1	0.895079	4
35	28	3	2	0.905741	2
36	28	3	3	0.867625	8

Based on Table 6, the best combination of parameters was obtained at BF = 21, MI = 3, and MO = 1, with a GCV value of 0.8625. These results indicate that the relationship between the incidence of dengue fever and the predictor variables is nonlinear and involves interactions among the variable.

The final model was built using these optimal parameters on the training data. After the forward selection and backward pruning processes, four active basis functions were obtained involving five predictor variables: adequate sanitation (X3), access to safe drinking water (X4), number of health facilities (X5), toilet ownership (X6), and annual rainfall (X7), while population density (X1) and the percentage of the population living in poverty (X2) were not retained in the model.

The MARS model equation obtained

$$\widehat{Y}_{std} = 0,2777 - 0,5069 \cdot BF1 + 0,4285 \cdot BF2 - 5,2048 \cdot BF3 - 2,1314 \cdot BF4$$

with

$$BF1 = h(0,964316 - X6) \times h(-0,288527 - X7)$$

$$BF2 = X4 \times h(0,964316 - X6) \times h(X7 + 0,288527)$$

$$BF3 = X5 \times h(X6 - 0,964316)$$

$$BF4 = h(X3 - 0,512235)$$

The variable importance analysis shows that X5 and X6 are the most influential variables, with an importance value of 100, followed by X3 and X7 (51.2), and X4 (38.7). These results are consistent with the final model, which retains all five variables in the construction of the basis function. The complete results are presented in Table 7.

Table 7. MARS Model Variable Importance

Rank	Variable	Importance Value (GCV)
1	X5	100
2	X6	100
3	X7	51,2
4	X3	51,2
5	X4	38,7

Next, the model's performance was evaluated on the training and test data. The evaluation results are presented in Table 8.

Table 8. MARS Model Evaluation Results

Metric	Training Data	Test Data
-GCV	0,86245	-
MSE	0,6364	1,7225
RMSE	0,7977	1,3124
MAE	0,6394	1,0069
R ²	0,3564	-0,8264
Adjusted R ²	0,2920	-1,9886

Based on Table 8, the MARS model produced a GCV value of 0.8625 on the training data, with MSE, RMSE, and MAE values of 0.6364, 0.7977, and 0.6394, respectively. The R² value of 0.3564 indicates that the model is able to explain approximately 35.64% of the variation in dengue fever incidence rates in the training data. On the test data, the MSE, RMSE, and MAE values increased to 1.7225, 1.3124, and 1.0069, respectively, while R² was -0.8264. These results show that the MARS model is able to capture some of the patterns in the training data; however, its generalization ability to new data is still limited, so its predictive performance on the test data is not yet optimal.

3.4 Random Forest Modeling

Random Forest modeling was performed as a comparison to the MARS model. Optimal parameters were determined using a grid search with variations in the mtry and ntree parameters. The best model was selected based on the smallest Out-of-Bag Mean Squared Error (OOB MSE) value. The results of the parameter search are presented in Table 9.

Table 9. Results of the Random Forest Parameter Grid Search Based on OOB MSE Values

No	mtry	ntree	OOB MSE	OOB RMSE
1	2	100	1,050143	1,024765
2	2	300	1,016148	1,008042
3	2	500	1,024782	1,012315
4	3	100	1,075273	1,036954
5	3	300	1,033691	1,016706
6	3	500	1,028899	1,014347
7	7	100	1,052829	1,026074
8	7	300	1,016575	1,008253
9	7	500	1,048233	1,023833

Based on Table 9, the best combination of parameters was obtained at mtry = 2 and ntree = 300, with an OOB MSE value of 1.016148. This combination was then used to build the Random Forest model. The resulting model is a regression model with 300 decision trees, in which two predictor variables are randomly selected at each node splitting. Unlike MARS, Random Forest does not produce an explicit mathematical equation because the prediction is derived from the average of the predictions made by all the decision trees. A summary of the model is presented in Table 10.

Table 10. Summary of the Random Forest Model

Component	Value
Model type	Random Forest Regression
Number of trees (ntree)	300
Number of variables per split (mtry)	2
Mean of Squared Residuals	1,047525
% Variance Explained	-5,94%

An analysis of variable importance shows that X6 is the most important variable based on

%IncMSE, while X3 has the largest IncNodePurity value. The complete results are presented in Table 11 and Table 12.

Table 11. Variable Importance Based on %IncMSE

Rank	Variable	%IncMSE
1	X6	5,7237
2	X1	2,9091
3	X7	1,9345
4	X2	0,8097
5	X3	-0,0486
6	X4	-0,4662
7	X5	-2,0994

Table 12. Variable Importance Based on IncNodePurity

Rank	Variable	%IncMSE
1	X3	12,0817
2	X6	10,7983
3	X7	10,6657
4	X2	10,4904
5	X1	9,3481
6	X5	8,7091
7	X4	6,0714

Next, the model's performance was evaluated on the training and test data. The evaluation results are presented in Table 13.

Table 13. Random Forest Model Evaluation Results

Metric	Training Data	Test Data
OOB MSE	1,0161	-
MSE	0,2129	0,9675
RMSE	0,4614	0,9836
MAE	0,3620	0,8014
R ²	0,7847	-0,0258
Adjusted R ²	0,7632	-0,6787

Based on Table 13, the Random Forest model produced an OOB MSE of 1.0161 on the training data, with MSE, RMSE, and MAE values of 0.2129, 0.4614, and 0.3620, respectively, and an R² of 0.7847. On the test data, the MSE, RMSE, and MAE values increased to 0.9675, 0.9836, and 0.8014, respectively, while R² was -0.0258. Compared to the MARS model, Random Forest

produced smaller prediction errors and a higher R^2 value on the test data, indicating better predictive performance even though its generalization ability is still not optimal.

3.5 MARS-Random Forest Hybrid Modeling

The MARS-Random Forest hybrid model was constructed by using the basis functions from the MARS model as predictor variables in the Random Forest model. Based on the best MARS model, four basis functions (BF1–BF4) were obtained and used as new features in the modeling process. Random Forest parameter tuning was performed via grid search using a combination of the *mtry* and *ntree* parameters, with selection criteria based on the smallest Out-of-Bag Mean Squared Error (OOB MSE) value. The parameter search results are presented in Table 14.

Table 14. Results of the Grid Search for the MARS-RF Hybrid Model Parameters

mtry	ntree	OOB MSE	OOB RMSE
2	100	0,842665	0,917968
2	300	0,872126	0,933877
2	500	0,856671	0,925565
4	100	0,875881	0,935885
4	300	0,896348	0,946756
4	500	0,913286	0,955660

Based on Table 14, the best combination of parameters was obtained at *mtry* = 2 and *ntree* = 100, with an OOB MSE of 0.8427. The final model was built using the four basis functions obtained from MARS (BF1–BF4) as predictor variables, thereby enabling the model to incorporate the information on nonlinear relationships obtained during the MARS stage into the Random Forest model.

Table 15. Important Variables in the MARS-RF Hybrid Model

Rank	Basis Function	%IncMSE
1	BF1	7,1075
2	BF3	6,4549
3	BF4	2,1668
4	BF2	0,7631

Based on Table 15, BF1 is the basis function that has the greatest impact on the hybrid model, with an %IncMSE value of 7.1075, followed by BF3, BF4, and BF2. All basis functions have

positive %IncMSE values, meaning they all contribute to improving the model's accuracy.

Table 16. Evaluation Results for the MARS-RF Hybrid Model

Metric	Training Data	Test Data
OOB MSE	0,8427	-
MSE	0,3029	1,2454
RMSE	0,5503	1,1160
MAE	0,4250	0,9569
R ²	0,6937	-0,3205
Adjusted R ²	0,6769	-0,6977

Based on Table 16, the hybrid model produced an OOB MSE of 0.8427 on the training data, with MSE, RMSE, and MAE values of 0.3029, 0.5503, and 0.4250, respectively, and an R² of 0.6937. On the test data, the MSE, RMSE, and MAE values increased to 1.2454, 1.1160, and 0.9569, respectively, while R² was -0.3205. These results indicate that the hybrid model has better predictive capability than the MARS model but has not yet been able to surpass the performance of the Random Forest model on the test data, suggesting that its generalization ability remains limited.

After obtaining the MARS, Random Forest, and MARS-Random Forest hybrid models, a performance comparison was conducted based on the GCV/OOB MSE, MSE, RMSE, MAE, and coefficient of determination (R²) metrics on the training and test datasets. The results of the comparison on the training dataset are presented in Table 17, while the evaluation results on the test dataset are presented in Table 18.

Table 17. Comparison of Model Performance on the Training Data

Model	GCV/OOB MSE	MSE	RMSE	MAE	R²
MARS	0,8625	0,6364	0,7977	0,6394	0,3564
Random Forest	1,0161	0,2129	0,4614	0,3620	0,7847
MARS-Random Forest Hybrid	0,8427	0,3029	0,5503	0,4250	0,6937

Based on Table 17, Random Forest delivered the best performance on the training data, with the lowest MSE, RMSE, and MAE values and the highest R² value. Although the hybrid model produced the smallest OOB MSE value (0.8427), its performance based on prediction error metrics

still fell short of Random Forest. Meanwhile, the MARS model showed the lowest performance compared to the other two models.

Table 18. Comparison of Model Performance on the Test Data

Model	MSE	RMSE	MAE	R ²	Adjusted R ²
MARS	1,7225	1,3124	1,0069	-0,8264	-1,9886
Random Forest	0,9675	0,9836	0,8014	-0,0258	-0,6787
MARS-Random Forest Hybrid	1,2454	1,1160	0,9569	-0,3205	-0,6977

Based on Table 18, Random Forest once again delivered the best performance on the test data, with the lowest MSE, RMSE, and MAE values and the highest R² value among the three models. The MARS-Random Forest hybrid model performed better than the MARS model but was not yet able to outperform Random Forest. These results indicate that the Random Forest approach is more effective in predicting dengue fever morbidity rates in the study data.

Overall, Random Forest was selected as the best model because it consistently produced the lowest prediction error rate and the highest R² value on both the training and test datasets. These findings indicate that the ensemble mechanism in Random Forest is better able to capture the patterns of relationships among variables than either the MARS model or the MARS–Random Forest hybrid model.

4. CONCLUSION

The MARS model achieved its best performance with BF = 21, MI = 3, and MO = 1 (GCV = 0.8625), while the optimal Random Forest model used mtry = 2 and ntree = 300 (OOB MSE = 1.0161). The MARS–Random Forest hybrid model achieved an OOB MSE of 0.8427 with mtry = 2 and ntree = 100. Variable importance analysis identified X5 and X6 as the key predictors in MARS, X6 in Random Forest, and BF1 in the hybrid model. Despite the hybrid model's lower OOB MSE, Random Forest showed the best overall predictive performance on both training and test data, with the lowest error values and the highest R². Therefore, Random Forest was selected as the most suitable model for predicting DHF incidence in Java and Bali in 2024.

Future studies should incorporate broader spatial and temporal coverage, along with climatic, environmental, and socioeconomic predictors. Comparing Random Forest with other machine-learning methods and improving dengue surveillance data may further support accurate prediction and the development of an early warning system for targeted dengue control.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude for the financial support provided by the Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS), under the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, through the Academic Excellent Research Funding Scheme (Department Fund, Batch 1) of Institut Teknologi Sepuluh Nopember in 2026. This research was funded under Researcher Funding Agreement No. 2946/PKS/ITS/2026.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

REFERENCES

- [1] U. Maharani, D. Nababan, H.A. Bangun, J. Manurung, The Influence of the Health Belief Model on the Risk Factors of Dengue Fever Occurrence in the Working Area of Sei Suka Public Health Center, Batubara Regency 2025, *J. Kesmas Gizi* 8 (2025), 61-66. <https://doi.org/10.35451/hgfmrf80>.
- [2] B. Haryanto, Indonesia Dengue Fever: Status, Vulnerability, and Challenges, in: *Current Topics in Tropical Emerging Diseases and Travel Medicine*, IntechOpen, (2018). <https://doi.org/10.5772/intechopen.82290>.
- [3] M. Nakhaie, A. Shahpar, M. Rezaei Zadeh Rukerd, N. Farsiu, J. Charostad, et al., Dengue Fever: Viral, Environmental, and Human Factors Driving Expansion and Pandemic Risk, *Rev. Med. Virol.* 36 (2026), e70088. <https://doi.org/10.1002/rmv.70088>.
- [4] O. Telle, B. Nikolay, V. Kumar, S. Benkimoun, R. Pal, et al., Social and Environmental Risk Factors for Dengue in Delhi City: A Retrospective Study, *PLOS Negl. Trop. Dis.* 15 (2021), e0009024. <https://doi.org/10.1371/journal.pntd.0009024>.

- [5] X. Qi, Y. Wang, Y. Li, Y. Meng, Q. Chen, et al., The Effects of Socioeconomic and Environmental Factors on the Incidence of Dengue Fever in the Pearl River Delta, China, 2013, *PLOS Negl. Trop. Dis.* 9 (2015), e0004159. <https://doi.org/10.1371/journal.pntd.0004159>.
- [6] A. Ikbali, A.I. Purnamasari, I. Ali, Analisis Klasterisasi untuk Prediksi Jumlah Kasus DBD Berdasarkan Jenis Kelamin dan Kabupaten/Kota di Jawa Barat, *JATI (Jurnal Mhs. Tek. Inform.)* 7 (2023), 3789-3796. <https://doi.org/10.36040/jati.v7i6.8296>.
- [7] N. Kamila, Pemodelan Jumlah Penderita Demam Berdarah Dengue (DBD) di Jawa Timur Menggunakan Analisis Spline Truncated Tahun 2023, *MATHunesa* 13 (2025), 109-117.
- [8] A. Ramayanti, Baharuddin, B. Abapihi, Makkulau, Ruslan, et al., Pemodelan Jumlah Penderita Demam Berdarah Dengue (DBD) Menggunakan Metode Generalized Poisson Regression untuk Mengatasi Overdispersi, in: *Seminar Nasional Sains dan Terapan VI*, 2022, pp. 17-22.
- [9] D.S. Ayundasari, S.H. Hastuti, K. Kertanah, Pemetaan Kasus DBD di Pulau Lombok Menggunakan Regresi Binomial Negatif Berbasis Geografis, *Edumatic: J. Pendidik. Inform.* 8 (2024), 497-506. <https://doi.org/10.29408/edumatic.v8i2.27460>.
- [10] R. Sriningsih, B.W. Otok, Sutikno, Factors Affecting the Number of Dengue Fever Cases in West Sumatra Province Using the Multivariate Adaptive Regression Splines (MARS) Approach, *J. Phys.: Conf. Ser.* 1722 (2021), 012094. <https://doi.org/10.1088/1742-6596/1722/1/012094>.
- [11] N.R. Febriyanti, K. Kusriani, A.D. Hartanto, Analisis Perbandingan Algoritma SVM, Random Forest dan Logistic Regression untuk Prediksi Stunting Balita, *Edumatic: J. Pendidik. Inform.* 9 (2025), 149-158. <https://doi.org/10.29408/edumatic.v9i1.29407>.
- [12] A. Hidayanti, A.M. Siregar, S.A.P. Lestari, Y.C. Cahyana, Model Analisis Kasus COVID-19 di Indonesia Menggunakan Algoritma Regresi Linier dan Random Forest, *PETIR* 15 (2021), 91-101. <https://doi.org/10.33322/petir.v15i1.1487>.
- [13] A. Rozy, Penerapan Random Forest Untuk Prediksi Virus Hepatitis C, *FIMERKOM J. Inf. Syst. Technol.* 1 (2024), 19-23.
- [14] J.H. Friedman, Multivariate Adaptive Regression Splines, *Ann. Stat.* 19 (1991), 1-141. <https://doi.org/10.1214/aos/1176347963>.

- [15] J.H. Friedman, C.B. Roosen, An Introduction to Multivariate Adaptive Regression Splines, *Stat. Methods Med. Res.* 4 (1995), 197-217. <https://doi.org/10.1177/096228029500400303>.
- [16] L. Breiman, Random Forests, *Mach. Learn.* 45 (2001), 5-32. <https://doi.org/10.1023/A:1010933404324>.
- [17] D.R. Cutler, T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, et al., Random Forests for Classification in Ecology, *Ecology* 88 (2007), 2783-2792. <https://doi.org/10.1890/07-0539.1>.