CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 10(1) (2025), Pages 467-479
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Comparing Multivariate Adaptive Regression Splines and Machine Learning Methods for Classifying Pneumonia in Indonesian Toddlers

Ardi Kurniawan*, Nur Azizah, Sheila Sevira Asteriska Naura

Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

Email: ardi-k@fst.unair.ac.id

ABSTRACT

Pneumonia is a type of infectious and contagious respiratory disease that causes death in toddlers. According to the Indonesian Ministry of Health (2024), the coverage of pneumonia among toddlers in 2023 was 36.95% with a total of 416,435 cases. This study aims to model and classify the pneumonia status of toddlers in Indonesia using the Multivariate Adaptive Regression Splines (MARS) method and several machine learning methods, such as logistic regression, K-NN, random forest, and SVM. This study uses secondary data from the Survei Kesehatan Indonesia in 2023 and the Profil Kesehatan Indonesia in 2023, including variables such as the percentage of toddlers health service coverage, low birth weight babies, population density, percentage of malnutrition in toddlers, prevalence of smoking in the population aged larger than 10 years in the last 1 month, percentage of toddlers who are exclusively breastfed, and percentage of toddlers who have incomplete basic immunization. The best model obtained using the MARS method is with BF = 14, MI = 2, and MO = 3. This model produces a GCV value of 0.122 and R-Square of 82.9%, which shows good prediction performance. The classification results show that the MARS method is superior to the logistic regression, K-NN, random forest, and SVM methods with an accuracy rate of 97.06%.

Keywords: Classification Accuracy; Machine Learning; MARS; Pneumonia; Toddlers

INTRODUCTION

Pneumonia is an acute infection that attacks the alveoli. This disease can be caused by various types of microorganisms, such as bacteria, viruses, fungi, and parasites [1] with symptoms including chills, fever, headache, cough with phlegm, and shortness of breath. Pneumonia is a type of infectious and contagious respiratory tract infection (ARI) that causes the most deaths in toddlers. Data from UNICEF shows that pneumonia claims the lives of more than 700,000 toddlers, or about 2,000 lives in every day. In Indonesia, pneumonia is a disease with a very high number of death cases. In 2023, pneumonia accounted for 416,435 cases with 522 deaths among toddlers [2]. This figure tends to fluctuate over the past eleven years. This condition makes pneumonia one of the major challenges in achieving the Sustainable Development Goals (SDGs), particularly in the 3rd

Submitted: Pebruary 18, 2025 Reviewed: April 12, 2025 Accepted: May 12, 2025
DOI: http://dx.doi.org/10.18860/ca.v10i1.32418

goal of "Healthy and Prosperous Life" which targets reducing toddlers mortality to less than 3 per 1,000 live births, and reducing the incidence of severe pneumonia in toddlers by 75% compared to the incidence in 2019 [3].

Based on PMK No. 82/2014, the Indonesian Ministry of Health plays a role in increasing public awareness of pneumonia risk factors through socialization and advocacy on the importance of a healthy lifestyle as an effort to prevent and control infectious diseases. The Indonesian Ministry of Health together with WHO and the United States Center for Disease Control and Prevention (US CDC) collaborated to build a system for early detection and response to diseases that have the potential for outbreak. This system is also known as the Early Warning Alert and Response System (EWARS). This program was first implemented by the Ministry of Health since 2009 [4]. In addition, the Ministry of Health also established a Rapid Response Team or Tim Gerak Cepat (TGC) at the central, provincial, and district/city levels. The TGC is a team that conducts early detection, reports responses, and makes recommendations for outbreak management.

There are many risk factors that contribute to the high mortality rate of pneumonia among toddlers in developing countries. Based on several previous studies, these factors are divided into two groups, intrinsic factors and extrinsic factors. Intrinsic factors are factors that come from the individual themselves, such as age, gender, nutritional status, birth weight, exclusive breastfeeding history, and immunization status. While extrinsic factors are factors that come from outside the individual, such as health service coverage, population density, and smoking status of family members. Research by [5] and [6] revealed that the percentage of exclusively breastfed infants and the percentage of infants who received complete basic immunization significantly influenced the number of pneumonia cases among toddlers. In addition, low birth weight [7] and nutritional status [8] also affect the incidence of pneumonia in toddlers. In addition, [7] and [9] in their studies revealed that smoking status of family members and population density also had a significant effect on the incidence of pneumonia in toddlers.

MARS is a nonparametric regression method that is more flexible in handling the complexity of health data that is often nonlinear [10] because MARS can capture complex and nonlinear patterns of relationships between variables, even when the data shows high variability [11]. MARS has advantages of handling interactions between variables and simplifying the model without reducing the level of accuracy [12]. These advantages is obtained through the use of basis functions and an efficient variable selection process, namely a combination of forward and backward elimination based on the Generalized Cross Validation (GCV) value or classification accuracy, which distinguishes it from classical spline regression [13]. MARS is also effective in overcoming multicollinearity problems by selecting the most relevant basis functions, thereby improving its predictive performance [14]. Research with the MARS method has been conducted by [15] regarding modeling factors affecting ARI in Sidoarjo.

In addition, other statistical methods that can be used to classify and predict pneumonia in toddlers are machine learning, including logistic regression, K-NN, random forest, and SVM approaches. Logistic regression is a special type of regression where the response variable is binary [16]. Similar research with this method has been done by [17] about modeling the factors that influence ARI. K-NN is one of the supervised learning-based classification methods that is done by calculating the distance between training and testing data to determine the class of unknown objects. Its advantages are flexible, simple, and resistant to noise, making itu suitable for various types of datasets, even large ones [18]. Random forest is an effective algorithm in reducing the risk of overfitting [19]. Research related to this method has been conducted by [20] on predicting the possibility of early

stage diabetes. Whereas in SVM, the way it works is by finding the best separation function to separate classes. The concept of classification is to find the best hyperplane so that it can be used to separate two classes of data in the input space [21]. Research related to the classification of respiratory diseases with the SVM algorithm has been carried out by [22] who also found that SVM performance is affected by the number of classes.

Based on the description above, the authors are interested in conducting research related to the application and comparison of several statistical methods, namely MARS, logistic regression, K-NN, random forest, and SVM by looking at the performance of the five methods to classify and predict pneumonia cases in toddlers in Indonesia. The comparison of these methods is important as each offers a different approach to addressing data complexity issues. MARS method has the advantage of handling nonlinear relationships as well as complex interactions between predictor variables, while machine learning methods offer a more flexible and simpler approach in dealing with various types of datasets. The novelty lies in the application of MARS, which has not been widely utilized in previous studies for health classification problems in toddlers at the national scale. By identifying the most accurate predictive model, this study provides a practical tool for early detection and policy decision-making in child health management in Indonesia. By comparing the performance of each method, the results of this study are expected to provide an overview of the factors that influence the prevalence of pneumonia in toddlers, thus making an important contribution to pneumonia prevention and control efforts in Indonesia. In addition, this study also highlights the influential factors behind pneumonia prevalence, which provides valuable insights for future prevention and control efforts. This study begins by detailing the dataset and methodology used, followed by an analysis and comparison of the performance of each classification method then conclusions and suggestions for future research.

METHODS

The type of research used is quantitative research with the data used is secondary data obtained from the publication of the Indonesian Health Survey 2023 and the Indonesian Health Profile 2023. The variables used in this study are the prevalence of pneumonia among toddlers in Indonesia, the percentage of health service coverage among toddlers, low birth weight babies, population density, the percentage of malnutrition among toddlers, the prevalence of smoking in the population aged ≥10 years in the last 1 month, the percentage of toddlers who are exclusively breastfed, and the percentage of toddlers with incomplete basic immunization.

Data processing and analysis in this study were performed MARS version 2.0 and Phyton via Google Colab. For machine larning methods including logistic regression, K-NN, random forest, and SVM, the scikit-learn package version 1.6.1 was used. Additional library utilized include pandas, numpy, and matplotlib for data handling and visualization. Data analysis was carried out in several stages which are presented briefly as follows.

1. Change the data scale from ratio (percentage) to categorical binary data. Cities/districts that have a percentage of pneumonia among toddlers below the national percentage are categorized as 0 and above the national percentage are categorized as 1.
2. Describe the characteristics of pneumonia among toddlers in Indonesia in 2023 and factors that are thought to influence them using descriptive statistical analysis.
3. Perform MARS modeling with the first step of determining the basis function (BF), maximum interaction (MI), and minimum observation (MO). According to [23], the

maximum number of basis functions is 2 to 4 times the number of predictor variables. The maximum number of interactions (MI) is 1, 2, or 3 with the consideration that if more than 3 will get a very complex model. Minimum observation (MO) between knots is 0, 1, 2, and 3.

4. Obtain the best MARS model by determining knots in MARS using the forward stepwise and backward stepwise algorithms. The MARS model can be said to be the most optimal if the GCV value in the model has the minimum value.
5. Test the significance of the MARS model simultaneously and partially with the following hypothesis [24].
   a. Testing simultaneous regression coefficients with the following hypothesis.
      $H_0 = a_1 = a_2 = \cdots = a_M = 0$
      H_1 : there is at least one $a_m \neq 0$; $m = 1,2, \ldots, M$
      Critical region: reject H_0 if $F > F_{a(M-1,N-M)}$ or $\mathrm{p} - \mathrm{value} < a$
   b. Partial regression coefficient testing with the following hypothesis.
      H_0: $a_m = 0$
      H_1: $a_m \neq 0$; for each m, where $m = 1,2, \ldots, M$
      Critical region: reject H_0 if $t > t_{(\frac{a}{2},N-M)}$ or $\mathrm{p} - \mathrm{value} < a$
6. Determine and interpret the factors that have the most influence on the percentage of pneumonia in Indonesia based on MARS analysis.
7. Calculating classification accuracy with MARS accuracy value with confusion matrix. The main purpose of the confusion matrix is to assess the performance or accuracy of a classification system in classifying test data [25]. An example confusion matrix for binary classification is shown in Table 1 [26].

Table 1. Confusion Matrix

Actual Class	Prediction Class	
	1	**0**
1	True Positive	False Negative
0	False Positive	True Negative

8. Divide the data with k-fold cross validation of 5 folds. For every 1 k-fold of testing data will be cross validated with 4 training data [27].
9. Perform classification with logistic regression with the first step of determining the best parameter (C) using random search with a range of 10^{-4} to 10^{4}. After obtaining the optimal C value, the model is trained using these parameters, and then used for prediction and evaluation of classification performance.
10. Perform classification with KNN with the first step of determining the best number of neighbors (K) using grid search in the range of K = [3,5,7,9,11,13,15] [28]. After obtaining the optimal K, the model is trained using k-fold and used for prediction of the target class.
11. Performing classification with random forest begins with determining the best parameters such as n_estimator, max_depth, and criterion using random search. Then the model is trained using the optimal parameters obtained to be used in prediction.
12. Perform classification using SVM by determining the best C and the best kernel using random search. The following are the parameter values used in hyperparameter tuning based on the kernel [29].

Table 2. Parameter Value of Hyperparameter Tuning Based on Kernel

Kernel	Parameter	Test Value
Linear	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
Rbf	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
	gamma	[0.001, 0.01, 0.1, 1, 10, 100]
Poly	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
	gamma	[0.001, 0.01, 0.1, 1, 10, 100]
	degree	[2,3,4,5]
Sigmoid	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
	gamma	[0.001, 0.01, 0.1, 1, 10, 100]

After obtaining the best parameters, the model is trained and used to predict the target class based on the best hyperparameters selected.

13. Comparing the accuracy value of classification accuracy between MARS, logistic regression, KNN, random forest, and SVM methods.

RESULTS AND DISCUSSION

Descriptive Statistics

Descriptive statistics describe the general characteristics of the response and predictor variables which are presented in Table 3 and Table 4 below.

Table 3. Descriptive Statistics of Response Variable

Variable	Value	Count
Pneumonia in Toddlers	1	14
	0	20
	Total	**34**

Table 4. Descriptive Statistics of Predictor Variables

	Variables	Mean	Minimum	Maximum
X_1	Coverage of Toddler Health Services (%)	71.72	26.40	97.00
X_2	Low Birth Weight (%)	4.59	1.00	8.10
X_3	Population Density (Hectares)	752.65	10.00	16146.00
X_4	Malnutrition (%)	0.56	0.00	1.80
X_5	Smoking (%)	20.40	15.60	27.70
X_6	Exclusive Breastfeeding (%)	54.36	35.90	71.40
X_7	Incomplete Basic Immunization (%)	59.01	26.50	75.20

Based on Table 4, the coverage of toddler health services has an average of 71.72 with a minimum value of 26.40 in Papua Province and a maximum of 97.00 in Central Java Province. The low birth weight variable has an average of 4.59 with a minimum value of 1.00 in DKI Jakarta Province and a maximum of 8.10 in Gorontalo Province. The population density variable has an average of 752.65 with a minimum value of 10 in North Kalimantan Province and a maximum of 16146.00 in DKI Jakarta Province. The malnutrition variable has an average of 0.56 with a minimum value of 0.00 in Bangka Belitung Islands Province and a maximum of 1.80 in West Province. Smoking variable has an average of 20.40 with a minimum value of 15.60 in West Papua Province and a maximum of 27.70 in West Nusa Tenggara Province. The exclusive breastfeeding variable had an average of 54.36 with a minimum value of 35.90 in West Papua Province and a maximum of 71.40 in DI Yogyakarta Province. The incomplete immunization variable has an average value of 59.01 with a

minimum value of 26.50 in Bali Province and a maximum of 75.20 in West Sulawesi Province.

Modeling the Prevalence of Pneumonia in Toddlers with the MARS Method

The first step taken in MARS modeling is to determine the maximum number of Basis Function (BF), Maximum Interaction (MI), and Minimum Observation (MO). To get the best model, trial and error is carried out on the combination of BF, MI, and MO so that the MARS model with the minimum GCV value and maximum R-Square value is obtained. Based on the results of the trial and error MARS model and by considering the principle of parsimony, the best MARS model is obtained with a combination of BF = 14, MI = 2, and MO = 3. The GCV value for the best MARS model is 0.122 and R-Square is 82.9%. The model form is shown in the following equation.

$$\hat{y} = 0.930 - 0.001BF_2 + 0.165BF_4 - 0.002BF_5 + 0.001BF_7 + 0.006BF_{11} \quad (1)$$

with:
$BF_2 = \max(0, 788.0 - X_3)$
$BF_3 = \max(0, X_6 - 44.0)$
$BF_4 = \max(0, 44.0 - X_6)$
$BF_5 = \max(0, X_2 - 6.4)BF_2$
$BF_7 = \max(0, X_2 - 5.3)BF_2$
$BF_{11} = \max(0, X_5 - 23.5)BF_3$

The interpretation of the MARS model based on equation (1) above is as follows:

1. $BF_2 = \max(0, 788 - X_3)$ with a coefficient of -0.001
   This means that a one-unit increase in BF_2 will reduce the prevalence of pneumonia among toddlers in Indonesia by 0.001. BF_2 will be equal to $(788.0 - X_3)$ if the population density (X_3) is less than 788. However, if the population density is equal to or higher than 788, BF_2 becomes meaningless or in other words, it equals 0. Thus, for every one increase in the BF_2 at a population density of less than 788, the prevalence of pneumonia in Indonesia will decrease by 0.001.
2. $BF_4 = \max(0, 44 - X_6)$ with a coefficient of 0.165
   This means that a one-unit increase in BF_4 will increase the prevalence of pneumonia among toddlers in Indonesia by 0.165. BF_4 will be equal to $(44 - X_6)$ if exclusive breastfeeding (X_6) is less than 44, and it will be 0 if exclusive breastfeeding is equal to or higher than 44.
3. $BF_5 = \max(0, X_2 - 6.4)BF_2$ with a coefficient of -0.002
   This means that a one-unit increase in BF_5 will reduce the prevalence of pneumonia among toddlers in Indonesia by 0.002, assuming other basis functions remain constant. BF_5 will be equal to $(X_2 - 6{,}4)$ if low birth weight (X_2) is greater than 6.4 and population density (X_3) is less than 788. If low birth weight (X_2) is equal to or less than 6.4 and population density (X_3) is equal to or less than 788, it will be 0.
4. $BF_7 = \max(0, X_2 - 5.3)BF_2$ with a coefficient of 0.001
   This means that a one-unit increase in BF_7 will increase the prevalence of pneumonia among toddlers in Indonesia by 0.001, assuming other basis functions remain constant. BF_7 will be equal to $(X_2 - 5.3)$ if low birth weight (X_2) is greater than 5.3 and population density (X_3) is less than 788. If low birth weight (X_2) is

equal to or less than 5.3 and population density (X_3) is equal to or less than 788, then it will be 0 or in other words BF_7 has no meaning.

Furthermore, from the MARS model in equation (1), it can be seen that there are four predictor variables that enter the model, namely X_2, X_3, X_5, and X_6. To see the extent to which these variables affect the formation of the MARS model, it can be seen in Table 5.

Table 5. Level of Importance of Predictor Variables

Variables	Importance Level	GCV
Population Density (X_3)	100.000	0.304
Exclusive Breastfeeding (X_6)	78.860	0.235
Low Birth Weight (X_2)	66.268	0.202
Smoking (X_5)	9.462	0.123
Coverage of Toddler Health Services (X_1)	0.000	0.122
Malnutrition (X_4)	0.000	0.122
Incomplete Basic Immunization (X_7)	0.000	0.122

From the best MARS model, the following parameter significance testing can be performed.

1. Simultaneous regression testing
   Hypothesis:
   $H_0 : \alpha_2 = \alpha_4 = \alpha_5 = \alpha_7 = \alpha_{11} = 0$ (model is not significant)
   H_1 : there is at least one $\alpha_m \neq 0$; $m = 2,4,5,7,11$ (model is significant)
   Significant level: 5%
   The simultaneous test results are presented in Table 6.

Table 6. Simultaneous Testing Results

F	P-Value
27.231	0.609×10^{-9}

Based on Table 6, with a critical region that rejects H_0 if $F > F_{0.05(4,29)}$ or p-value < 0.05. Thus, the decision is to reject H_0 because the value of $F(27.231) > F_{0.05(4,29)}(2.70)$ which means that the model is significant and can be used to model pneumonia cases among toddlers.

2. Partial regression testing
   Hypothesis:
   $H_0 : \alpha_m = 0$ (coefficient γ_m has no effect on the model)
   $H_1 : \alpha_m \neq 0$ for each m, where $m = 2,4,5,7,11$
   Significant level: 5%
   Partial test results are presented in Table 7.

Table 7. Partial Test Results

Parameter	Estmation	S.E	t-value	p-value
Constant	0.930	0.086	10.795	0.173×10^{-10}
BF 2	-0.001	0.148×10^{-3}	-9.585	0.243×10^{-9}
BF 4	0.165	0.023	7.312	0.583×10^{-7}
BF 5	-0.002	0.342×10^{-3}	-6.321	0.775×10^{-6}
BF 7	0.001	0.185×10^{-3}	6.820	0.208×10^{-6}
BF 11	0.006	0.002	2.903	0.007

The partial test results in Table 7, show that the coefficient of α_2, α_4, α_5, α_7, and α_{11} have a significant effect on pneumonia in toddlers because the $|t_{value}| > t_{(0.025,29)}$ (2.045) or $p - value < \alpha(5\%)$, so that it can be used to model pneumonia cases in toddlers.

Classification Accuracy of Pneumonia in Toddlers with MARS

The following is the result of the classification accuracy of the MARS method.

Table 8. MARS Classification Accuracy

Observation Results	Prediction Results	
	0	**1**
0	19	1
1	0	14

$$\text{Accuracy} = \frac{19 + 14}{19 + 14 + 0 + 1} = \frac{33}{34} = 0.9706 \approx 97.06\%$$

Classification Accuracy of Pneumonia in Toddlers with Logistic Regression

The logistic regression method used parameter is C (regulation) to control the balance between model complexity and overfitting. To obtain the optimal parameters, hyperparameter tuning is performed using random search and validated with k-fold cross validation of 5 folds. The value of C used during tuning is between 10^{-4} and 10^{4} and obtained the best C parameter of 0.0001. Classification accuracy with the logistic regression method is presented in Table 9.

Table 9. Logistic Regression Classification Accuracy

Observation Results	Prediction Results	
	0	**1**
0	20	0
1	7	7

$$\mathbf{Accuracy} = \frac{\mathbf{7 + 20}}{\mathbf{7 + 20 + 0 + 7}} = \frac{\mathbf{27}}{\mathbf{34}} = \mathbf{0.7941 \approx 79.41\%}$$

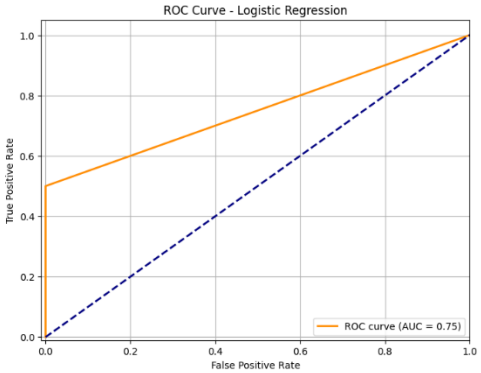

Figure 1. ROC Curve and AUC of Logistic Regression

Classification Accuracy of Pneumonia in Toddlers with K-NN

In the K-NN method, the main parameter used is the number of neighbors to determine the number of nearest neighbors during the classification process. To get the optimal parameters, hyperparameter tuning is done using Grid search so that the optimal number of neighbors is 7 with the K-Fold Cross Validation method. Classification accuracy with the K-NN method is presented in Table 10.

Table 10. K-NN Classification Accuracy

Observation Results	Prediction Results	
	0	**1**
0	20	0
1	7	7

$$\text{Accuracy} = \frac{7+20}{7+20+0+7} = \frac{27}{34} = 0.7941 \approx 79.41\%$$

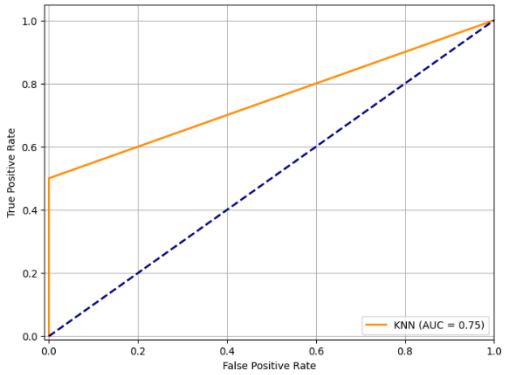

Figure 2. ROC Curve and AUC of K-NN

Classification Accuracy of Pneumonia in Toddlers with Random Forest

The optimal parameters are obtained through the hyperparameter tuning process with the random search method resulting in an entropy criterion with a maximum depth of 7. The classification accuracy with the random forest method is presented in Table 11.

Table 11. Random Forest Classification Accuracy

Observation Results	Prediction Results	
	0	**1**
0	19	1
1	5	9

$$\text{Accuracy} = \frac{9+19}{9+19+1+5} = \frac{28}{34} = 0.8235 \approx 82.35\%$$

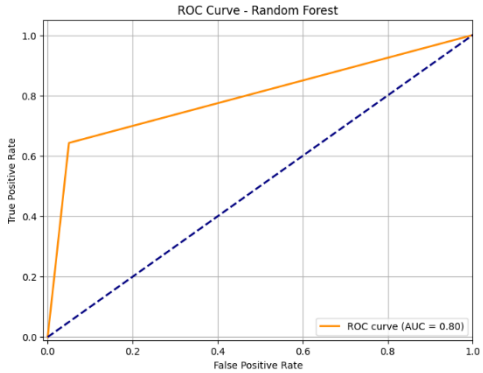

Figure 3. ROC Curve and AUC of Random Forest

Classification Accuracy of Pneumonia in Toddlers with SVM

Determining the optimal parameters is necessary to produce an accurate SVM model. Optimal parameters can be determined using random search algorithm and K-fold cross validation. After successfully identifying the optimal parameters through the random search algorithm, it was found that the best parameters for the SVM model were C = 1, gamma = 100, degree = 2 and polynomial kernel. The classification accuracy with the SVM method is presented in Table 12.

Table 12. SVM Classification Accuracy

Observation Results	Prediction Results	
	0	**1**
0	18	2
1	5	9

$$\text{Accuracy} = \frac{18+9}{18+9+2+5} = \frac{27}{34} = 0.7941 \approx 79.41\%$$

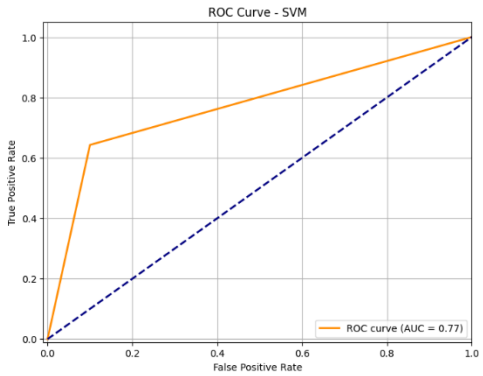

Figure 4. ROC Curve and AUC of Random Forest

Comparison of Classification Accuracy of Pneumonia in Toddlers with MARS, Logistic Regression, K-NN, Random Forest, and SVM Methods

Furthermore, the best method is selected by comparing the classification accuracy value (1-APER) which is presented in Table 13 below.

Table 13. Classification Accuracy Comparison

Classification Methods	Classification Accuracy $(1 - \text{APER})$
MARS	97.06%
Logistic Regression	79.41%
K-NN	79.41%
Random Forest	82.35%
SVM	79.41%

Based on the classification accuracy comparison results in Table 13, it is known that the (1-APER) value with the MARS method is higher than logistic regression, K-NN, random forest, and SVM, which is 97.06%. So, it can be concluded that MARS is more suitable for modeling and predicting pneumonia cases among toddlers in Indonesia.

Although this study successfully modeled the prevalence of pneumonia using data from 34 provinces in Indonesia, the limited number of observations may limit the application of more complex or data-intensive machine learning models. In addition, this study only uses cross-sectional data from a single year, which may not fully capture potential changes in risk factors over time. Future research could utilize longitudinal data to analyze trends and causality, explore additional health or environmental predictors, or compare more advanced algorithms to further improve classification accuracy.

CONCLUSIONS

The modeling of pneumonia cases in Indonesia using MARS method is $\hat{y} = 0{,}930 - 0{,}001BF_2 + 0{,}165BF_4 - 0{,}002BF_5 + 0{,}001BF_7 + 0{,}006BF_{11}$ has the largest influence with an importance level of 100%, followed by the percentage of toddlers who are exclusively breastfed (X_6) with an importance level of 78.86%, low birth weight (X_2) at 66.27%, and

the prevalence of smoking in the population aged ≥10 years in the last 1 month at 9. 46%. The interaction between predictor variables also provides an important insight, where the combination of low birth weight and population density (BF_5) tends to reduce pneumonia cases among toddlers, but in contrast, the combination of low birth weight and population density (BF_7) ends to increase the prevalence of pneumonia among toddlers. This indicates that low birth weight can have a different impact on the prevalence of pneumonia in toddlers based on environmental conditions, especially population density. Based on the model performance, the MARS method produced an accuracy of 97.06%. Classification accuracy using logistic regression, K-NN, and SVM methods resulted in an accuracy of 79.41%. While the classification accuracy with the random forest method resulted in an accuracy of 82.35%. Thus, the MARS method classification is better than logistic regression, K-NN, SVM, and random forest methods. This study has certain limitations that should be acknowledged. The use of data from only one year restricts the analysis to a cross-sectional perspective, which may not capture temporal dynamics. Moreover, the limited sample size based on 34 provinces may not support highly complex models. Future research could incorporate multi-year data, expand the set of predictor variables, or compare alternative classification algorithms to enhance predictive accuracy and generalizability.

ACKNOWLEDGMENTS

We would like to express our deepest gratitude to the Department of Mathematics, Universitas Airlangga, for providing the resources and support necessary for conducting this research. Additionally, we extend our appreciation to the authors and researchers whose works have significantly contributed to the theoretical and methodological framework of this study. Lastly, we thank all parties who provided data and other assistance, making this study possible.

REFERENCES

[1] Kemenkes RI, Profil Kesehatan Indonesia 2023, Jakarta: Kementerian Kesehatan Republik Indonesia, 2024.

[2] BKPK Kemenkes, "Survei Kesehatan Indonesia (SKI) 2023," 2023. [Online].

[3] Kemenkes Ditjen P2, "Pneumonia Menjadi Ancaman Kesakitan dan Kematian di Dunia," 2024. [Online]. Available: https://p2p.kemkes.go.id/pneumonia-menjadi-ancaman-kesakitan-dan-kematian-di-dunia/. [Accessed 26 Februari 2024].

[4] Kemenkes RI, PEDOMAN Sistem Kewaspadaan Dini dan Respon (SKDR) Penyakit Potensial KLB / Wabah, Jakarta: Kementerian Kesehatan RI, 2023.

[5] Y. L. Kang, Q. X. Zheng, X. Q. Chen and F. Zheng, "Effects of Exclusive Breastfeeding Duration on Pneumonia Occurrence and Course in Infants Up to 6 Months of Age: A Case-Control Study," *Journal of Community Health Nursing,* vol. 41, no. 4, pp. 256-264, 2024.

[6] V. N. Sutriana, M. N. Sitaresmi and A. Wahab, "Risk Factors of Childhood Pneumonia: a Case-Control Study in a High Prevalence Area in Indonesia," *Clin Exp Pediatr,* vol. 64, no. 11, pp. 588-595, 2021.

[7] H. Shi, T. Wang, Z. Zhao, D. Norback, X. Wang, Y. Li, Q. Deng, C. Lu, X. Zhang, X. Zheng, H. Qian, L. Zhang, W. Yu, Y. Shi, T. Chen, H. Yu, H. Qi, Y. Yang, L. Jiang, Y. Lin, J. Yao, J. Lu and Q. Yan, "Prevalence, Risk Factors, Impact and Management of Pneumonia Among

Preschool Children in Chinese Seven Cities: A Cross-Sectional Study with Interrupted Time Series Analysis," *BMC Medicine,* vol. 21, no. 227, pp. 1-14, 2023.

[8] U. Yuliniar, Y. Wijayanti and D. R. Indriyanti, "An Analysis Factors Affecting the Cases of Pneumonia in Toddlers at Public Health Center (Puskesmas) Pati I," *Public Health Perspective Journal,* vol. 6, no. 3, pp. 254-261, 2021.

[9] G. S. Prihant, K. C. Widati, P. T. Yovi, A. Z. Dewi, W. Kirtanti, A. M. I. Restu, S. E. Elvaretta, A. A. Susilo, P. T. J. Audiawiyanti, Friska and A. Putri, "The Effect of House Environmental Factors on the Incidence of Pneumonia in Toddlers," *KnE Medicine,* vol. 2, no. 3, pp. 296-306, 2022.

[10] C. C. Chang, J. H. Yeh, H. C. Chiu, T. C. Liu, Y. M. Chen, M. J. Jhou and C. J. Lu, "Assessing the Length of Hospital Stay for Patients with Myasthenia Gravis Based on the Data Mining MARS Approach," *Frontiers in Neurology,* vol. 14, p. 1283214, 2023.

[11] M. S. Abed, F. J. Kadhim, J. K. Almusawi, H. Imran, L. F. A. Bernardo and S. N. Henedy, "Utilizing Multivariate Adaptive Regression Splines (MARS) for Precise Estimation of Soil Compaction Parameters," *Applied Sciences,* vol. 13, no. 11634, pp. 1-21, 2023.

[12] A. Özmen, Y. Zinchenko and G. W. Weber, "Robust multivariate adaptive regression splines under cross-polytope uncertainty: an application in a natural gas market.," *Annals of Operations Research,* vol. 324, pp. 1337-1367, 2024.

[13] M. B. Adiguzel and M. A. Cengiz, "Model Selection in Multivariate Adaptive Regressions Splines (MARS) Using Alternative Information Criteria," *Heliyon,* vol. 9, no. 9, p. 19964, 2023.

[14] M. A. Sahraei, H. Duman, M. Y. Codur and E. Eyduran, "Prediction of Transportation Energy Demand: Multivariate Adaptive Regression Splines," *Energy,* vol. 224, no. 12, p. 120090, 2021.

[15] Mahfudhotin, "Pemodelan Penyakit Infeksi Saluran Pernafasan Akut di Daerah Sekitar Semburan Lumpur Lapindo Sidoarjo dengan Pendekatan Model Multivariate Adaptive Regression Spline," *JAMBURA: Journal of Probability and Statistics,* vol. 3, no. 2, pp. 86-96, 2022.

[16] H. Lee and H.-S. Kim, "Logistic Regression and Least Absolute Shrinkage and Selection Operator," *Cardiovascular Prevention and Pharmacotherapy,* vol. 2, no. 4, pp. 142-146, 2020.

[17] E. Matasina, "Penerapan Regresi Logistik untuk Kasus Infeksi Saluran Pernafasan Akut (ISPA) pada Balita," *Jurnal Diferensial,* vol. 2, no. 1, pp. 56-66, 2020.

[18] M. Bansal, A. Goyal and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decision Analytics Journal,* vol. 3, 2022.

[19] N. S. Thomas and S. Kaliraj, "An Improved and Optimized Random Forest Based Approach to Predict the Software Faults," *SN Computer Science,* vol. 5, no. 530, 2024.

[20] W. Apriliah, I. Kurniawan, M. Baydhowi and T. Haryati, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," *SISTEMASI: Jurnal Sistem Infromasi,* vol. 10, no. 1, pp. 163-171, 2021.

[21] F. Riandari, H. T. Sihotang, T. Tarigan and M. Rafli, "Classification of Book Types Using the Support Vector Machine (SVM) Method," *Jurnal Mantik,* vol. 6, no. 1, pp. 43-49, 2022.

[22] A. Achmad, Adnan and M. Rijal, "Klasifikasi Penyakit Pernapasan Berbasis Visualisasi Suara Menggunakan Metode Support Vector Machine," *Jurnal Ilmiah Ilmu Komputer,* vol. 8, no. 2, p. 119, 2022.

[23] J. H. Friedman, Multivariate Adaptive Regression Splines, vol. 19, The Annals of Statistics, 1991, pp. 1-67.

[24] D. Rahma, N. Amalita, Y. Kurniawati and Z. Martha, "Application of Multivariate Adaptive Regression Splines for Modeling Stunting Toddler on The Island of Java," *UNP Journal of Statistics and Data Science,* vol. 2, no. 3, pp. 338-343, 2024.

[25] R. Nurhidayat and K. E. Dewe, "Penerapan Algoritma K-Nearest Neighbor dan Fitur Ekstraksi N-Gram dalam Analisis Sentimen Berbasis Aspek," *KOMPUTA : Jurnal Ilmiah Komputer dan Informatika,* vol. 12, no. 1, pp. 91-100, 2023.

[26] A. Indriani, "Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, Yogyakarta, 2014.

[27] S. Ridho and D. Rusda, "Analisis Preferensi Konsumen dalam Memilih Produk Hortikultura Menggunakan Metode Algoritma C45 dan Naive Bayes," *Emitor,* vol. 24, no. 1, pp. 66-77, 2024.

[28] S. Sanjaya, M. L. Pura, S. K. Gusti, F. Yanto and F. Syafria, "K-Nearest Neighbor for Classification of Tomato Maturity Level Based on Hue, Saturation, and Value Colors," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM),* vol. 2, no. 2, pp. 101-106, 2019.

[29] U. Ulfitasari, A. Yonita, S. Siswanto and A. Kalondeng, "Pemodelan Indeks Kebahagiaan Negara dengan Metode Multivariate Adaptive Regression Spline," *MATHunesa: Jurnal Ilmiah Matematika,* vol. 13, no. 1, pp. 209-216, 2025.