Vol. 7, No. 4, December 2025

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Comparative Analysis of Random Forest, Logistic Regression and SVM for Stunting Prediction Using Anthropometric Data

Shalsa Bela Dwi Widyawati[1], Purwadi[2], Ika Romadoni Yunita[3]

[1,2]Postgraduate Program, Amikom Purwokerto University, Purwokerto, Indonesia

[3]Faculty of Computer Sciences, Amikom Purwokerto University, Purwokerto, Indonesia

Email: salsabeladwi12@gmail.com[1], purwadi@amikompurwokerto.ac.id[2], ikarom@amikompurwokerto.ac.id[3]

Received: Nov 3, 2025

Revised: Nov 23, 2025

Accepted: Dec 6, 2025

Published: Dec 26, 2025

Corresponding Author:

Author Name*:

Purwadi

Email*:

purwadi@amikompurwokerto.ac.id

DOI:

10.63158/journalisi.v7i4.1387

Abstract. Stunting remains a critical nutritional issue in Indonesia, significantly impacting the physical and cognitive development of children under five. Prompt and accurate detection of nutritional status is essential for early intervention. This study aims to predict toddlers' nutritional health using the Random Forest algorithm, based on age and height data. From an initial dataset of 120,998 anthropometric records, preprocessing steps—such as duplicate removal and nutritional status recategorization—resulted in a final dataset of 39,425 entries. The research methodology includes data collection, preprocessing, exploratory analysis, model training, handling class imbalance, and performance evaluation using accuracy, precision, recall, and F1-score. The study also compares the Random Forest model with Logistic Regression and Support Vector Machine (SVM). Results show that Random Forest outperforms the other models, achieving perfect classification metrics: Accuracy (1.00), Recall (1.00), F1-Score (1.00), and Cross-validation Accuracy (99.74%). These outcomes highlight Random Forest's robustness in classifying under-five nutrition data, making it an effective tool for rapid and reliable stunting risk detection. This research supports efforts to reduce Indonesia's stunting rate to below 20% by 2024, contributing to national health improvement strategies through technology-driven early diagnosis.

Keywords: Stunting, Nutritional Status, Machine Learning, Random Forest, Logistic Regression, Support Vector Machine (SVM)

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Stunting, or Impaired growth in toddlers, is a serious problem experienced by childern In Indonesia due to prolonged malnutrition. This Issue Ia a serious concern In Indonesia, given Its Impact on childern's physical and cognitive development, this could potentially have an determine the quality of human resources in the years to come [1]. The high stunting rate Indicates that many childern suffer from chronic malnutrition, which affects their growth and development, requiring Immediate action [2].

According to the results of the Indonesian Nutrition Status Study report that comes from the Ministry of Health, the stunting rate in Indonesia went down from 27.7% in 2019 to 24.4% In 2021, and futher decreased to 21.6% In 2022 [3]. The majority of instances are seen in kids between the ages of 3-4, with a prevalence of up to 6%. However, this figure still does not meet WHO standard, which sets the Ideal stunting prevalence at below 20%. To overcome this, the government Is targeting a reduction In stunting to 17% In 2023 and 14% In 2024 [4].

There are many factors associated with stunting. These include mothers who are malnourished during pregnancy, mother who are short in stature, and poor parenting, especially in terms of behavior and feeding practices [5]. Other factors that cause stunting in childern include maternal Infection, teenage pregnancy, closely spaced briths, economic condotion, and poor accers to health services and clean water, all of which greatly affect child growth [6]. Studies have shown that early Intervention Is crucial to prevent stunting and its long-term adverse development, which can affect a child's quality of life into adulthood. Unfortunately, the Identification of childern at risk Is still oftendone manually and is ineffective. Therefore, a fast and accurate system Is needed to classify toodlers to detect whether they are stunted or not. Based on this, scientific methods and technologies such as machine learning algorithms can play a significant role [7].

In recent years, machine leraning methods have been widely used to predict childern's nutritional status. One algorithm that's is quite effective is Random Forest, which is known to predictions on various types of datasets. Random Forest work by randomly

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

constructing a large number of decision tree, then combining the result to produce better predictions [8].

Previous studies, the classification of stunting in toddlers using the K-Nearest Neighbor algorithm and Logistic Regression produced the performance of both models, each contributing to the classification of stunting, with the K-Nearest Neighbor algorithm used to handle complex data, which provided high accuracy results of 0.980 and an F1-Score of 0.987. Meanwhile, the Logistic Regression algorithm contributed to understanding the impact of various health indicators with an accuracy of 0.877 and an F1-Score of 0.894. This study combined two machine learning algorithms to assist in the prevention of stunting [9].

Research on the classification of stunting in toddlers using the Support Vector Machine (SVM) algorithm based on toddler anthropometric data shows that the SVM algorithm has an accuracy of 82%, meaning that positive stunting predictions have a model recall value of 86%. Researchers applied the switching multiple variables algorithm with a linear kernel to create a stunting classification model, resulting in a model with good classification values [10].

Other research has also conducted test using Random Forest algorithm with stunted toddler data, showing that the best scenario Is to use 80% of the data for training and 20% for testing, with the highest accuracy of 90,1%, precision of 71,4% and recall of 62.5%. Therefore, there is still a need to improve the classification of stunting in toddlers by using other machine learning algorithms to provide more accurate results [11].

This study aims to predict the nutritional status of toddlers based on age and height data using three algorithms, namely Random Forest, SVM, and Logistic Regression. The dataset used is toddler anthropometric data that has undergone preprocessing o guarantee the accuracy of the analysis. The research methodology included a data collection process, data preprocessing, model training, and performance evaluation of the algorithms using metrics such as accuracy, precision, and recall. The expected results will provide a comparison of the accuracy of the algorithms to determine which is best to use as a basis for preventing stunting in the future.

2. METHOD

The research flow as shown in Figure 1, which outlines the systematic steps taken during the study—from initial data acquisition to model evaluation. Each stage ensures that the data is properly prepared, explored, modeled, and assessed for accuracy and effectiveness.

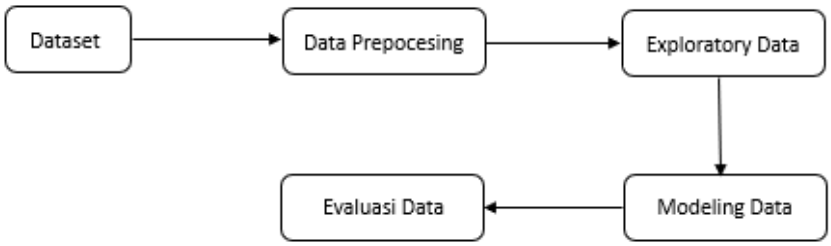

Figure 1. Research Method

2.1. Dataset

A dataset Is a collection of data organized in a table or specific format that Is used as a basis for analysis or research, becausee the quality and completeness of the data contained therein will greatly affect the results of the analysis or model produced. Datasets contains Information from various sources, such as numbers and Images. Within the dataset, recognizing and handling duplicate information is crucial because it can significantly impact how data is processed. Furthermore, comprehending how balanced the dataset is will assist in the modeling phase, alongside incorporating descriptive statistical tables displaying values such as the average, lowest value, highest value, and standard deviation [12].

2.2. Data Preprocessing

Data preprocessing begins with sorting the data after it has been succsessfully collected. The main objectives at this stage are to select relevant features and to handle missing or duplicate data. To minimize potential errors, dependent variables that are not Involved In the testing stage can be eliminated. An essential step in this study is label consolidation, which was conducted to simplify the target variable for a binary classification task. Consolidating related categories ensures conceptual alignment with international standards and enhances the interpretability and consistency of the classification model.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Next process for numerical feature preprocessing, continuous variables were processed using Min–Max Normalization, transforming the feature range to a standardized interval between 0 and 1 [13]. This step is crucial as several algorithms used in the experiment, such as Logistic Regression and Support Vector Machine (SVM), are sensitive to varying feature scales and may produce biased decision boundaries if normalization is not applied [14]. The final step is the SMOTE (Synthetic Minority Oversampling Technique) process of the dataset to support dataset balance, so that the modeling results will not be biased towards data with large values [15].

2.3. Exploratory Data Analysis (EDA)

The process esploring a dataset to understand it consent, components, and characteristicsso that we can Identify data patterns. Exploratory Data Analysis (EDA) was Introduced by Jonh Tukey and aims to encourage statisticans to explore data and formulate hypothesis [16]. At this process, data analyst Is also carried out to understand the characteristic, patterns, or relationships between variables. Activities carried out include data visualization and descriptive statistical analysis such as mean, median, standard deviation, etc. The results of EDA will help provide Initial Insights before the data is used for the next stage, which Is the modeling stage [17].

2.4. Data Modeling

Data modeling is process to Involves selecting and appliying machine learning algorithms thar are appropriate for the results to be tested [18]. In sthis study, the algorithms used are Random Forest, Logistic Regression, and SVM. The results of this phase are used to compare each algorithm that can provide accurate results so that they can provide Information related to the research. To ensure a systematic modeling process, the following workflow was implemented, as shown in Figure 2.

In Figure 2, model training workflow diagram showing the complete pipeline from raw data to final model selection. The initial phase involves refining the unprocessed dataset through actions like managing absent entries, getting rid of extreme data points, and adjusting feature magnitudes using StandardScaler. Subsequently, the refined data is divided into a training group and a testing group, with the training group comprising 80% and the testing group comprising 20% of the data [19]. To guarantee that the outcomes can be consistently replicated, a fixed random seed value of 42 was used throughout all

stages involving stochasticity, such as the partitioning of training and testing data, the setup of the model, and the processes of cross-validation. This measure ensures that any user executing the identical code on the identical dataset will achieve the same results, a pivotal aspect for both verification and assessment by other experts. To enhance the effectiveness of the model, a meticulous search for the most suitable hyperparameters for every algorithm was carried out through the utilization of Grid Search Cross-Validation [20].

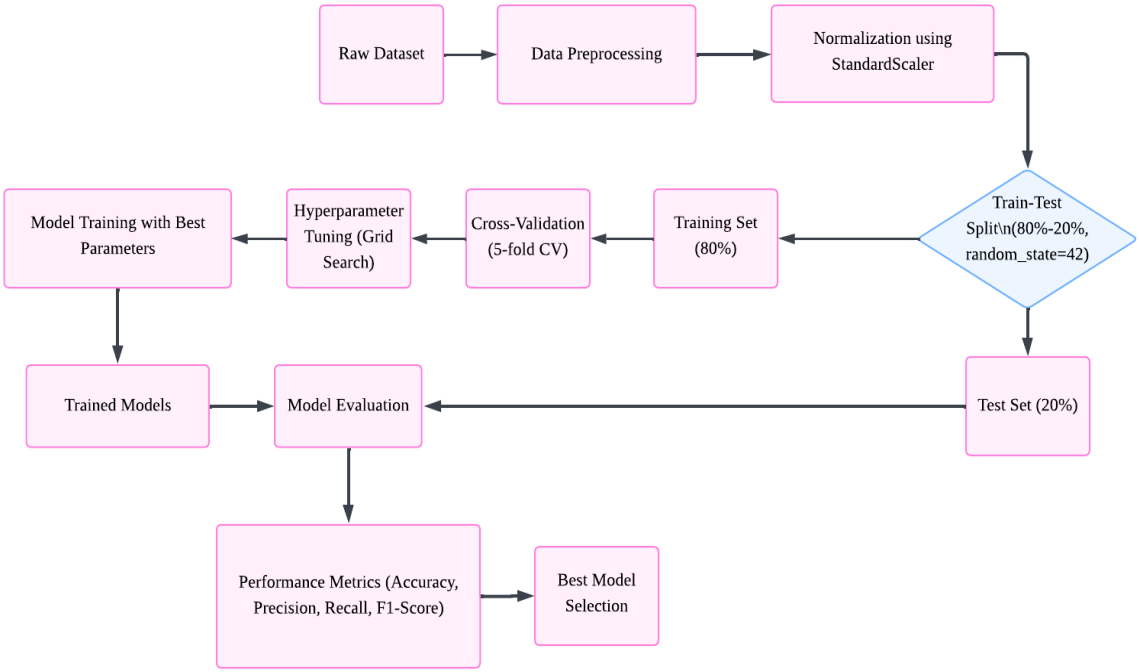

Figure 2. Model Training Workflow Diagram

To guarantee a reliable assessment and avoid overfitting, each model underwent a training phase that included 5-fold Stratified Cross-Validation on the training dataset [21]. After finding the ideal hyperparameters by employing Grid Search, every model was retrained utilizing these best settings on the complete training dataset. To offer a thorough assessment of each model's ability to categorize nutritional status, performance was evaluated using a variety of measures, including accuracy, precision, recall, and F1-score. Comparing the three algorithms makes it possible to choose the best model for classifying nutritional status, which offers trustworthy data to help with decision-making in programs for monitoring and intervening in children's health.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2.5. Data Evaluation

One Important step in assessing model performance is to evaluate classification using a matrix. The goal of this procedure is conducted to evaluate how effectively the model performs in accurately categorizing the data. The five main components in classification matrix evaluation Includee accuracy, precision, recall, F1-score, support, and consistency. A confusion matrix table also known as a confusion matrix, serves the purpose of evaluatibg how wll supervised learning algorithms are performing [22]. This table consist of two rows representing actual class examples and one column showing the class predictions generated by the model [23]. To calculate the confusion matrix measurement, as shown in Equation 1 to 4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precission = \frac{TP+TN}{TP+FN} \quad (3)$$

$$F1\ Score = 2\ \times \frac{Recall \times Precission}{Recall+Precission} \quad (4)$$

In the context of understanding matrix calculations, TN signifies True Negative, representing the count of negative instances accurately identified. Subsequently, TP represents True Positive, denoting the count of positive instances accurately categorized. Conversely, FP indicates False Positive, signifying the count of negative instances erroneously classified as positive, whereas FN denotes False Negative, indicating the count of positive instances erroneously classified as negative. [22].

3. RESULTS AND DISCUSSION

3.1. Dataset

Data Collection Dataset of Stunting Balita Detection was obtained from Kaggle (https://www.kaggle.com/code/rojindarafarin/stunting-balita-detection). This study uses a dataset in csv format yo perform the classification process by comparing the accuracy results of the four methods used, namely Random Forest, Logistic Regression, and SVM.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The dataset can be seen In Figure 3, which contains data on toddlers with variables of Age, Gender, Height in cm and nutritional status, watotaling 120,998 original data points.

	Age (months)	Gender	Height (cm)	Nutritional Status
0	0	male	44.591973	stunted
1	0	male	56.705203	tall
2	0	male	46.863358	normal
3	0	male	47.508026	normal
4	0	male	42.743494	severely stunted
...	...	...	...	...
120994	60	female	100.600000	normal
120995	60	female	98.300000	stunted
120996	60	female	121.300000	normal
120997	60	female	112.200000	normal
120998	60	female	109.800000	normal

Figure 3. Toddlers Stunting Dataset

Then, at Figure 4 below, there is a descriptive statistics table of the dataset which explains the dataset is composed of 39,425 data points, showcasing a near-even split between genders (average=0.49, suggesting nearly the same number of males and females). The age breakdown exhibits a skew towards the higher end, averaging 28.30 months (SD=19.26), spanning from birth 0 to 60 months, with a greater number of younger participants, evidenced by the 25th percentile mark at 10 months. Measurements of height follow a roughly standard distribution, averaging 86.04 cm (SD=19.77), varying from 40.01 cm to 128.00 cm, illustrating the anticipated differences in physical growth across the spectrum of ages. The variable of interest, which is the state of nutrition, displays a significant disparity in categories, with about 72% of the data falling into a single category and the remaining 28% in another (average=0.28, SD=0.45), this requires careful attention when creating and assessing models.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

	Age (months)	Gender	Height (cm)	Nutritional Status
count	39425.000000	39425.000000	39425.000000	39425.000000
mean	28.301332	0.492099	86.040251	0.277413
std	19.260394	0.499944	19.766565	0.447728
min	0.000000	0.000000	40.010437	0.000000
25%	10.000000	0.000000	72.100000	0.000000
50%	29.000000	0.000000	88.400000	0.000000
75%	45.000000	1.000000	100.800000	1.000000
max	60.000000	1.000000	128.000000	1.000000

Figure 4. Descriptive Statistic Table

3.2. Data Preprocessing

Data preprocessing represents a crucial step within the process of data analysis when employing machine learning algorithms. After the data were collected, data is prepared through a preprocessing stage to guarantee it is accurate and set for utilization. This phase Includes handling missing value by Imputing or deleting missing or duplicate data. For the data preprocessing stage, Include:

1) Data Cleansing

In the dataset used, namely data_balita.csv, duplicate data, outlier data, empty or null data and minor data changes were handled to Improve data accuracy. In this dataset, there were 81,574 duplicate data points, which were then deleted, leaving 39,425 data points.

2) Encoding Data

At this stage, the nutritional status attribute is changed with the descriptions stunted and severely stunted changed to the number 0, while the descriptions normal and tall are changed to the number 1. Next, change the Gender attribute by changing the description male to the number 0 and female to the number 1 to facilitate the analysis of stunting predictions in toddlers. The preprocessing results can be seen in Figure 5.

3) Splitting Data

Following the initial data preparation steps, the dataset underwent a partitioning process, resulting in an 80/20 split, designating 80% for training purposes and the remaining 20%

for testing the model. The primary reason for dividing the dataset was to handle possible inconsistencies present within the Nutritional Status feature. The aim was to maintain a consistent ratio of each category across both the training set and the testing set, a crucial step considering the substantial disproportions observed in the dataset. [24].

	Umur (bulan)	Jenis Kelamin	Tinggi Badan (cm)	Status Gizi
0	0	1	44.591973	stunted
1	0	1	56.705203	normal
2	0	1	46.863358	normal
3	0	1	47.508026	normal
4	0	1	42.743494	stunted
...	...	...	...	...
120959	60	0	100.700000	normal
120967	60	0	113.700000	normal
120968	60	0	107.500000	normal
120972	60	0	127.600000	normal
120993	60	0	116.100000	normal

39425 rows × 4 columns

Figure 5. Dataset after preprocessing

4) Standardization data

To guarantee that every attribute possesses an equivalent scale before the modeling phase, the data normalization step is executed. Because the optimization and distance computation methods rely on numeric values, data standardization is implemented on the Logistic Regression and SVM models, as a result, StandardScaler is used for normalization to convert each attribute to have a mean of 0 and a standard deviation of 1. Conversely, the Random Forest model doesn't need data standardization since it makes use of decision trees to segment data depending on threshold values.

Figure 6 presents the outcomes of data standardization for the numerical attributes, specifically Age and Height, demonstrating that the distribution of the Age and Height attributes stays unchanged both before and after normalization. The Age attribute, originally spanning from 0 to 60 months prior to standardization, was effectively converted to a standard scale ranging roughly from -1.5 to 1.5. Simultaneously, the Height attribute, which previously ranged from 40 to 130 cm, shifted to a range of about -2 to 2. These results confirm that standardization only changes the scale size, not the distribution shape.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

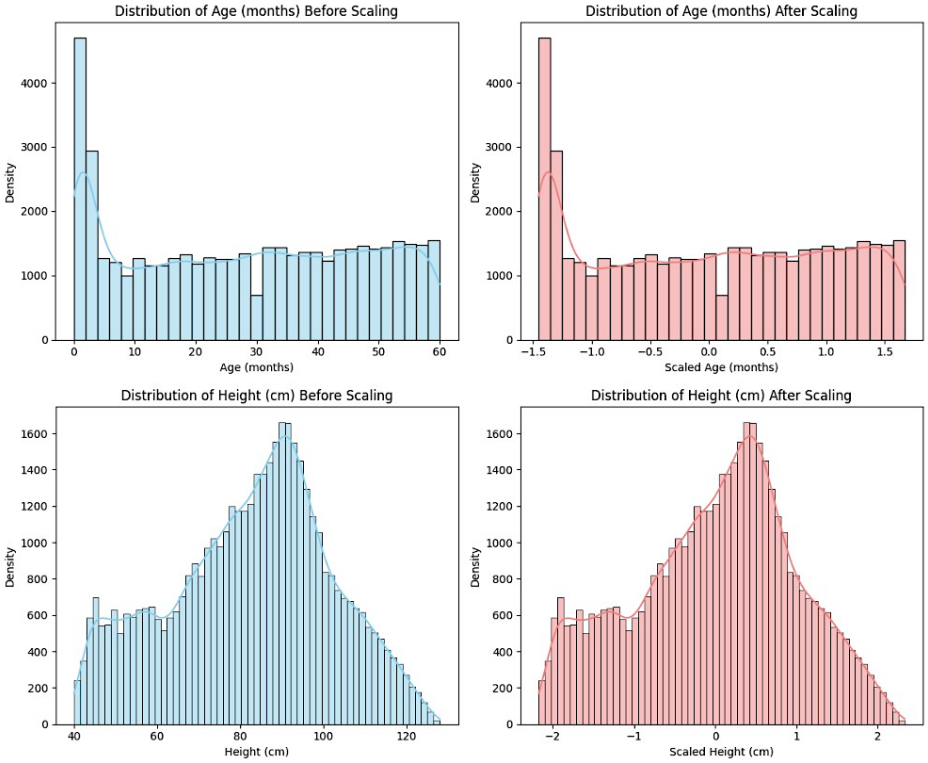

Figure 6. Data Standardization of variable Age and Height

3.3. Exploratory Data

The outcomes of the exploratory data analysis are illustrated in the correlation matrix In Figure 7.

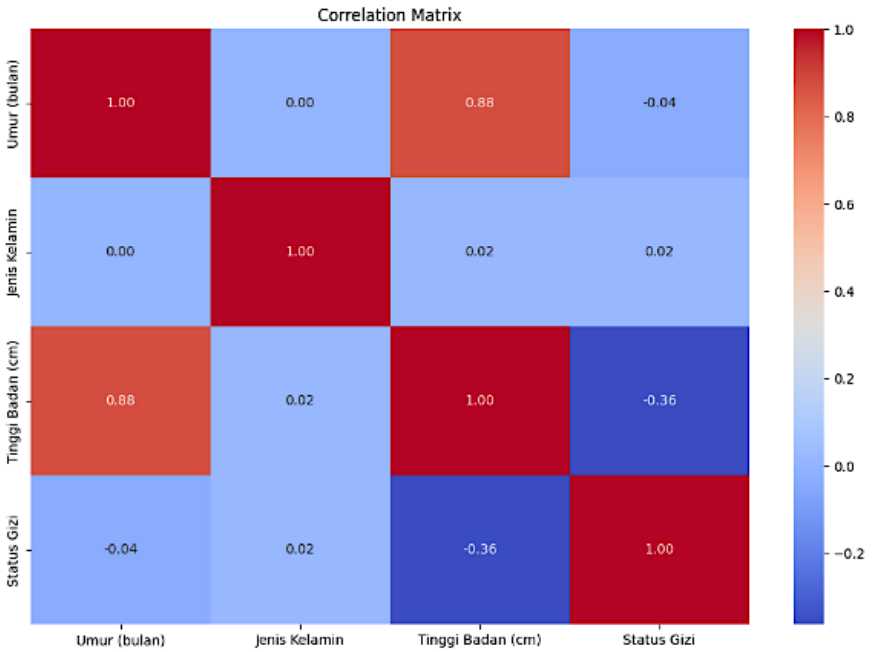

Figure 7. Corelation Matrix of Stunting Dataset

Based on the results of the Corelation Matrix above, the analysis reveals a significantly strong positive relationship between age (in months) and height with a correlation value of 0.88. This suggests that children's height increases substantially as they grow older, reflecting normal physical growth.

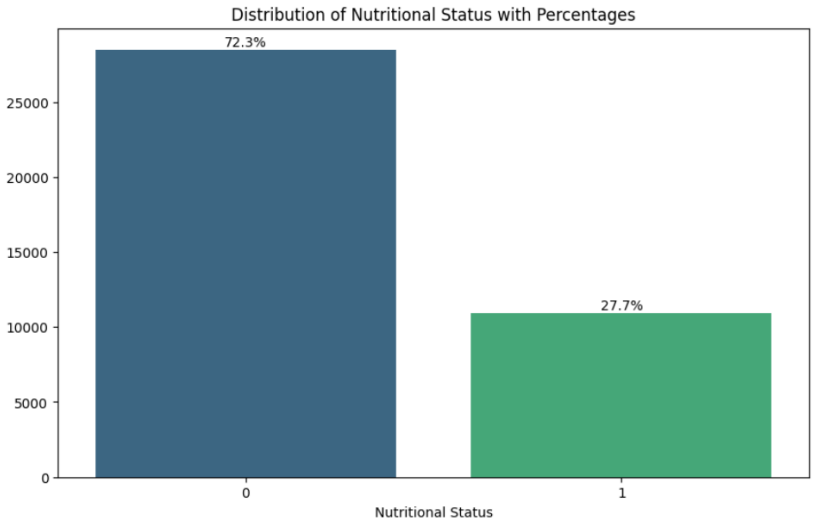

Figure 8. Distribution of Nutritional Status

Figure 8. shows the distribution of nutritional status, it's evident that group 0, which includes those with stunted and severely stunted growth, possesses a significantly greater number of data points in comparison to group 1, encompassing individuals with normal or above-average height. The disproportionate amount of data highlights a higher prevalence of stunting instances within the dataset compared to cases of normal or optimal nutritional status. The consequences of this imbalance are considerably significant when developing machine learning models. Consequently, this distribution validates the dataset's imbalanced nature, rendering the Synthetic Minority Oversampling Technique (SMOTE) an important approach to enhance the accuracy of models predicting the occurrence of stunting [25].

3.4. Data Modeling

This study used three different algorithms to classify the nutritional status of toddlers, with the following results for each algorithm.

1) Logistic Regression

Testing using the Logistic Regression has produced the results shown In Table 1.

Table 1. Logistic Regression Testing

	Precision	Recall	F1-score	Support
0	0.96	0.86	0.90	5698
1	0.71	0.90	0.79	2187
Accuracy	-	-	0.87	7885
Macro avg	0.83	0.88	0.85	7885
weighted avg	0.89	0.87	0.87	7885

The results based on research using the Logistic Regression algorithm show results in clasess 0 and 1. in class 0, the precision Is 0.96, recall is 0.86, F1-score Is 0.90, and it has a support of 5698. Meanwhile, for class 1, has precision 0.71, recall 0.90, F1-score 0.79, and support 2187. The accuracy produced by this algorithm is 0.87 with a Cross-validation Accuracy of 0.8801 or 88%.

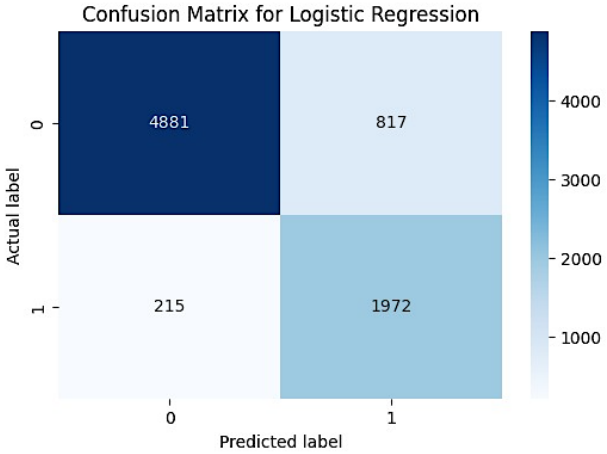

Figure 9. Confusion Matrix of Logistic Regression

Figure 9 presented subsequently, displays a confusion matrix derived from the Logistic Regression method. A True Positive (TP) value of 4,881 denotes the count of toddlers genuinely classified as normal (class 0) and suitably identified as normal by the model, thereby illustrating that a substantial TP value signifies the model's strong capability in detecting toddlers who do not have stunted growth. The False Positive (FP) value of 215 represents the count of infants genuinely identified as normal, yet mistakenly predicted

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

as having stunted growth. A True Negative (TN) value of 1,972 indicates the count of infants genuinely diagnosed with stunted growth (class 1), and suitably predicted as having stunted growth. The concluding False Negative (FN) value of 817 signifies the count of infants genuinely experiencing stunted growth, yet erroneously predicted as normal.

Conversely, several elements could obstruct the model from achieving peak accuracy, including the existence of variables lacking linearity with the model, along with exceptionally intricate data that impede ideal functioning. Analyzing the outcomes of the confusion matrix, the Logistic Regression technique demonstrates adequate effectiveness in differentiating between toddlers with stunted growth and those without, as evidenced by the elevated TP and TN numbers. The model's effectiveness is seen as substantially deficient due to its restriction to representing only direct connections, unlike the tendency of growth retardation data structures to exhibit non-direct associations and feature mixed groups. Nonetheless, Logistic Regression has the capacity to create coefficients that are helpful in pinpointing the traits that exert the greatest effect on the chances of young children being classified as healthy height or above-average height.

2) Random Forest

Testing using the Random Forest algorithm has produced the results shown In Table 2.

Table 2. Random Forest Testing

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	5698
1	0.99	1.00	0.99	2187
Accuracy	-	-	1.00	7885
Macro avg	0.99	1.00	1.00	7885
weighted avg	1.00	1.00	1.00	7885

The result based on research using the Random Forest algorithm have shown results in classes 0 and 1. In class 0, precision is 1.00, recall 1.00, F1-score 1.00, and has a support of 2187. The accuracy produced by this algorithm Is 1.00, with a Cross-validation Accuracy of 0.9974 or 99%.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

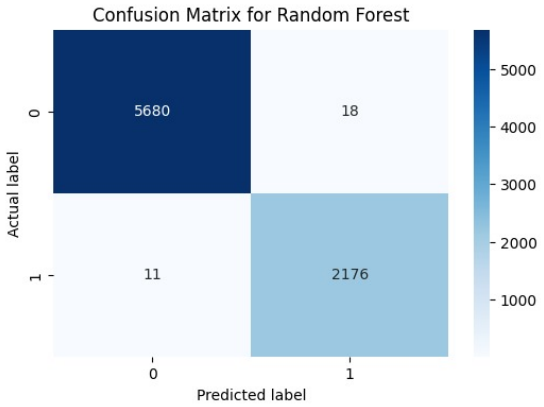

Figure 10. Confusion Matrix of Random Forest

Figure 10 is confusion matrix obtained from the Random Forest algorithm is True Positive (TP) result of 5687 is the number of toddlers who are actually normal (class 0) and correctly predicted as normal by the model, thus indicating that a high TP value results in a model that is very good at recognizing toddlers who are not stunted. The False Positive (FP) result of 11 shows the number of toddlers who are actually normal but were incorrectly predicted as stunted. The True Negative (TN) result of 2179 shows the number of infants who are actually stunted (class 1) and correctly predicted as stunted. The final False Negative (FN) result of 8 shows the number of infants who are actually stunted but incorrectly predicted as normal. Looking at the confusion matrix results in the Random Forest model shows that the model has very high performance and is almost perfect in classifying stunting status in toddlers, with a cross-validation value reaching 09974 or 99%.

The confusion matrix for the Random Forest model demonstrates exceptionally strong results, with very few misclassifications observed across all categories. Despite this impressive accuracy, it's possible that the model is overfitting the data, particularly due to the uneven distribution of data points in category 0 and category 1. Random Forest algorithms have a natural tendency to excel at identifying common trends, which heightens the chance that the model is overly reliant on the specific attributes of the data it was trained on. To confirm that this level of performance is genuinely reliable and not simply a feature of the training set, it is advisable to incorporate a learning curve or a validation curve into the analysis. By visualizing these curves, it can be verified if the

model maintains its ability to generalize effectively and to determine if the outstanding results are due to genuine learning or are a sign of overfitting.

3) Support Vector Machine (SVM)

Testing using the SVM algorithm has produced the results presented in Table 3.

Table 3. SVM Testing

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	5698
1	0.99	0.99	0.99	2187
Accuracy	-	-	1.00	7885
Macro avg	0.99	0.99	0.99	7885
weighted avg	1.00	1.00	1.00	7885

The results based on research using the SVM algorithm have shown results in classes 0 and 1. in class 0, achieved a precision of 1.00, a recall of 0.97, an F1-score of 0.99, and a support of 5698. Meanwhile, for class 1, had precision 0.93, recall 1.00, F1-score 0.96, and has a support 2187. The accuracy produced by this algorithm Is 98%, with a cross-validation accuracy Is 0.9836.

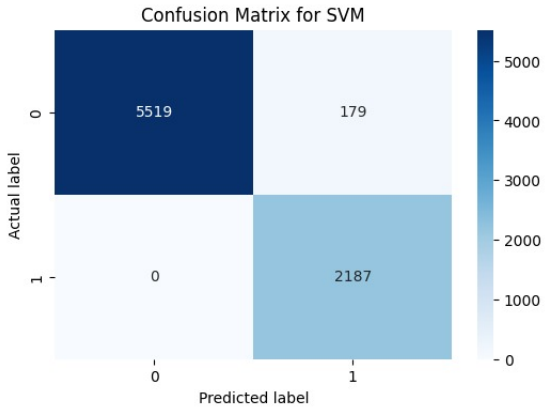

Figure 11. Confusion Matrix of SVM

The following Figure 11. Is confusion matrix obtained from for the SVM algorithm Is True Positive (TP) result of 5683 is the number of toddlers who are actually normal (class 0) and correctly predicted as normal by the model, thus indicating that a high TP value

results in a model that is very good at recognizing toddlers who are not stunted. The False Positive (FP) result of 15 shows the number of toddlers who are actually normal but were incorrectly predicted as stunted. The True Negative (TN) result of 2168 shows the number of infants who are actually stunted (class 1) and correctly predicted as stunted. The final False Negative (FN) result of 0 shows the number of infants who are actually stunted but incorrectly predicted as normal. The results of the confusion matrix in the SVM model show that the most significant mistake was observed in category 0, where 179 instances were incorrectly classified as category 1. The similarity in characteristics between the underdeveloped and typical groups might be the reason for this, placing some category 0 instances near the SVM's separation line. Furthermore, utilizing a curved RBF kernel might create a limit that is more inclined towards category 1, leading the model to prioritize minimizing mistakes in category 1 (FN = 0), which in turn raises the number of incorrect positive results in category 0.

4. Discussion

Based on the results of the classification of three classification algorithms, specifically Logistic Regression, Random Forest, and SVM, it can be concluded that the Random Forest algorithm achieved the highest accuracy in this study. The comparison of the accuracy of each algorithm as Indicated in Table 5.

Table 1. Model Evaluation Comparison

	Model	Accuracy	Precision	Recall	F1-score
0	Logistic Regression	0.87	0.83	0.87	0.84
1	Random Forest	1.00	0.99	0.99	0.99
2	SVM	0.98	0.96	0.98	0.97

For the comparison results in the table above considering the three implemented machine learning algorithms, the Random Forest algorithm demonstrated the highest performance, achieving superior accuracy of 1.00, and for precision, recall, and F1-score results, it was also superior to other algorithms.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

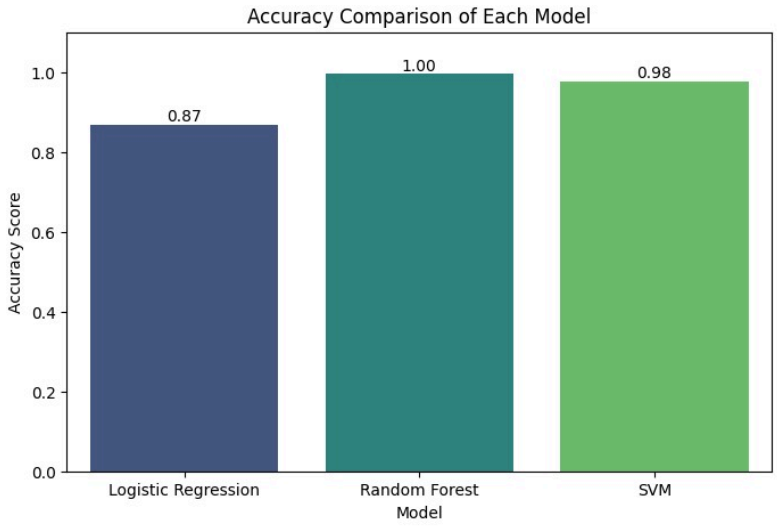

Figure 12. Comparison of Accuracy for Each Algorithm

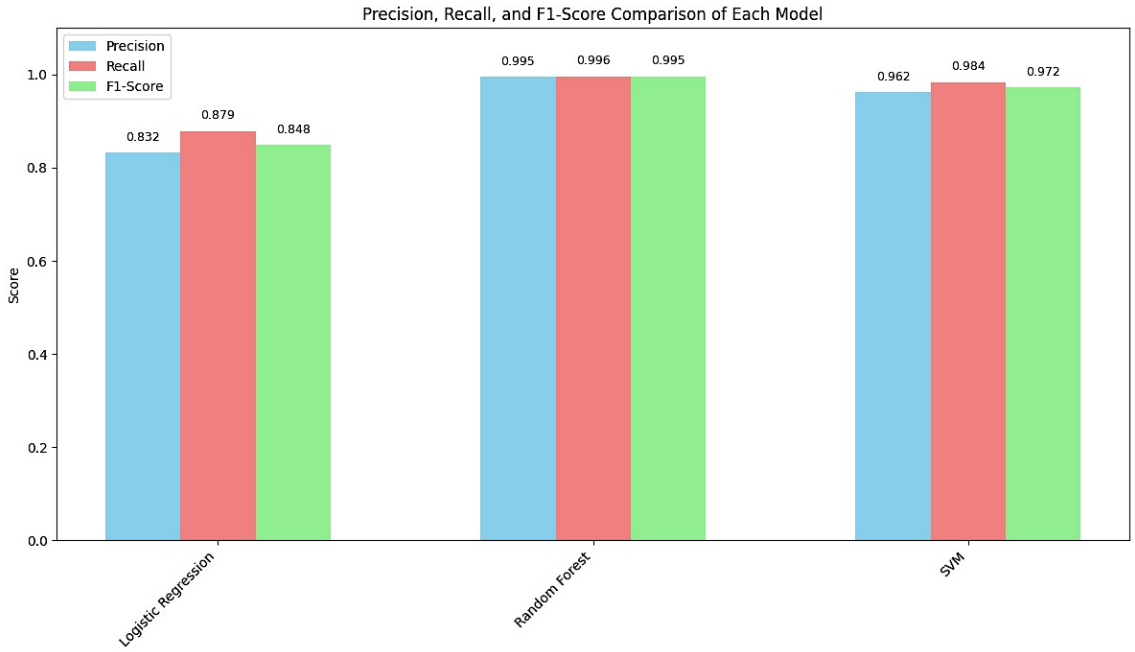

Figure 13. Visualization of Comparison Each Model

Figure 12 and Figure 13 shows that the accuracy of the three machine learning algorithms, Logistic Regression, Random Forest, and Support Vector Machine, indicates that each model has a different level of accuracy in predicting test data, with Random Forest demonstrating the highest accuracy compared to the other models, araoun 1.0 or 100%, Includeing that this algorithm Is capable of predecting data more accurately. Meanwhile, Logistic Regression has the lowest accuracy, In the range or 0.87 or 87%, compared to

SVM, which has an accuracy of 0.98 or 98%. Therefore, Random Forest remains the most appropiate and superior algorithm to use.

The analysis reveals that the effectiveness of the three machine learning techniques varies when assessing childhood stunting. The Random Forest technique emerges as the most effective overall, with the Support Vector Machine following closely, while the Logistic Regression approach demonstrates the weakest performance. This reflects the unique attributes of each algorithm alongside the intricacies of the dataset. The Random Forest algorithm is particularly effective because it can identify non-linear associations and multifaceted interactions among different data characteristics. Still, there are hints of the model being too closely fit to the training data when the training and testing results are put side by side, specifically within datasets where categories are not equally represented. In spite of this, the Random Forest algorithm's performance on fresh testing data holds steady and dependable. The Support Vector Machine, particularly when employing the RBF kernel, also demonstrates noteworthy effectiveness. The RBF kernel's strength lies in its ability to accurately represent non-linear connections, making it capable of handling intricate fluctuations in the data. Due to its linear nature, the Logistic Regression model displayed the weakest performance. As the dataset contains more intricate and non-linear trends, the Logistic Regression approach struggles to fully understand the interconnections between different features. Understanding the coefficients highlights the significant features, but they are not sufficient to significantly enhance predictions.

Statistical evaluations further validate that the Random Forest approach provides a noticeably improved level of accuracy relative to Logistic Regression, nevertheless, the enhancement is not consistently significant when juxtaposed with the SVM. This suggests that despite the superiority of the RF model, the SVM is still a competitive alternative. This research demonstrates that sophisticated models, such as the Random Forest and SVM, are better suited to handle complex stunting prediction datasets. In contrast, constraints, such as datasets with disproportionate class representation and the possibility of overfitting, should be carefully assessed when evaluating and implementing models.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

5. CONCLUSION

The results of this research used Random Forest algorithm to forecast the nutritional condition of toddlers determined by anthropometric data, such as age and height, as part of the intervention efforts aimed at reducing stunting in Indonesia. With an accuracy score of 100%, the Random Forest algorithm exhibited superior performance compared to the others in categorizing the nutritional well-being of toddlers, achieving better performance than Logistic Regression and Support Vector Machines, in classifying the nutritional status of toddlers. The dataset used came from Kaggle, consisting of 120,998 data points, and underwent preprocessing to ensure data quality, including the removal of duplicates, handling of missing values, and simplification of nutritional status categories.

The data analysis showed a strong positive correlation between age and height, reflecting normal growth in children with good nutrition. The Random Forest algorithm demonstrated not just a strong degree of correctness, but consistently reliable functioning as assessed by cross-validation, achieving 99.74% accuracy. This study supports the rapid and accurate early detection of stunting, contributing significantly to government programs aimed at reducing stunting rates. This section provides a brief conclusion about the research discussed in this article, along with suggestions for further development or follow-up research.

This research has several shortcomings that must be taken into account. Due to a notable disparity in numbers between the underdeveloped and healthy groups in the data collection, the possibility of skewed predictions remains, particularly for the smaller group, despite the use of SMOTE. Furthermore, the model's capacity to comprehend the more intricate reasons of growth retardation is still restricted since it only makes use of traits like age, gender, and height. The ideal accuracy attained by the Random Forest model also points to the chance of overfitting, mostly as a result of the comparatively straightforward data patterns. Since the relationship patterns in the dataset were non-linear, Logistic Regression performed the worst because linear models struggled to differentiate between classes. The fact that the dataset only came from one source (Kaggle) restricts how well the findings can be applied to Indonesian toddlers in general.

By incorporating more pertinent characteristics like socioeconomic factors, consumption habits, sanitation, and the history of maternal and child health, further study may be conducted to enable the model to account for more all-encompassing aspects. To improve accuracy in smaller classes, more advanced imbalance management strategies like SMOTE-Tomek and cost-sensitive learning can be used. To assess the model's stability and generalizability, model evaluation using external data from various locations is also required. Furthermore, using more sophisticated models like Gradient Boosting, XGBoost, or hybrid ensemble models has the potential to boost performance. Future studies may also incorporate interpretability analysis to clarify the contribution of each trait and create the best model-based early detection application to help healthcare professionals in the area.

REFERENCES

[1] R. Ratnasari, A. J. Wahidin, and T. H. Andika, "Early Detection of Stunting in Children Based on Anthropometric Indicators Using Machine Learning Algorithmsg," *J. Algoritm.*, vol. 21, no. 2, pp. 378–387, 2024, doi: 10.33364/algoritma/v.21-2.2122.

[2] N. Rusliani, W. R. Hidayani, and H. Sulistyoningsih, "Literature Review: Factors Associated with Stunting in Toddlers," *Bul. Ilmu Kebidanan dan Keperawatan*, vol. 1, no. 01, pp. 32–40, 2022, doi: 10.56741/bikk.v1i01.39.

[3] S. N. Azizah and Z. Fatah, "Implementation of the K-Nearest Neighbor (K-NN) Method in the Classification of Stunting in Toddlers ," *Gudang J. Multidisiplin Ilmu*, vol. 2, no. 10, pp. 282–288, 2024.

[4] T. R. . Lestari, "Stunting in Indonesia: The Root of the Problem and Its Solution ," *Info Singk. Kaji. Singk. Terhadap Isu Aktual dan Strateg.*, vol. XV, no. 14, pp. 21–25, 2023.

[5] F. M. Mulyaningrum, M. M. Susanti, and U. A. Nuur, "Factors Affecting Childhood Stunting," pp. 74–84, 2021.

[6] N. L. Rambe, "Indonesian Health Magazine," *J. Ilm. Kebidanan Imelda*, vol. 1, no. 2, pp. 45–49, 2020.

[7] J. Aurima, S. Susaldi, N. Agustina, A. Masturoh, R. Rahmawati, and M. Tresiana Monika Madhe, "Factors Associated with Stunting in Indonesian Toddlers," *Open Access Jakarta J. Heal. Sci.*, vol. 1, no. 2, pp. 43–48, 2021, doi: 10.53801/oajjhs.v1i3.23.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[8] S. Marsya Finda and D. Wahyu Utomo, "Classification of Stunting in Toddlers using Ensemble Learning and Random Forest Methods," *Jl. Imam Bonjol No*, vol. 15, no. 02, pp. 287–295, 2024, doi: 10.35970/infotekmesin.v15i2.2326.

[9] A. T. Armando Sibuea, P. Harry Gunawan, and Indwiarti, "Classifying Stunting Status in Toddlers Using K-Nearest Neighbor and Logistic Regression Analysis," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, 2024, pp. 6–11. doi: 10.1109/ICoDSA62899.2024.10652063.

[10] A. Jalil, A. Homaidi, and Z. Fatah, "Implementation of Support Vector Machine Algorithm for Classification of Stunting Status in Toddlers," *G-Tech J. Teknol. Terap.*, vol. 8, no. 3, pp. 2070–2079, 2024, doi: 10.33379/gtech.v8i3.4811.

[11] M. R. Akbar Ariyadi, S. Lestanti, and S. Kirom, "Classification of Stunted Toddlers Using Random Forest Classifier in Blita Districtr," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3846–3851, 2024, doi: 10.36040/jati.v7i6.7822.

[12] J. Juwariyem, S. Sriyanto, S. Lestari, and C. Chairani, "Prediction of Stunting in Toddlers Using Bagging and Random Forest Algorithms," *Sinkron*, vol. 8, no. 2, pp. 947–955, 2024, doi: 10.33395/sinkron.v8i2.13448.

[13] R. Fauzan and A. Rosita, "Comparison of KNN and Naïve Bayes Classification Algorithms for Predicting Stunting in Toddlers in Banjaran District," *Jurnal Teknologi dan Sistem Komputer,* vol. 9, no. 5, pp. 2711–2717, 2025.

[14] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, 4th ed., Elsevier, 2022.

[15] A. Smote and D. A. N. Neighbor, "Classification of Unbalanced Data Using SMOTE," *Jurnal Data Science Indonesia* (*DSI*), vol. 3, no. 1, pp. 44–49.

[16] F. Azimah and K. Rizky Nova Wardani, "Early Covid-19 Symptom Detection System Using the Al Project Cycle Method," *J. Locus Penelit. dan Pengabdi.*, vol. 1, no. 6, pp. 405–418, 2022, doi: 10.36418/locus.v1i6.135.

[17] P. K. Neighbors and D. A. N. LightGBM, "Combining K-Nearest Neighbors and LightGBM for Diabetes Prediction on the Pima Indians Dataset," *Jurnal Ilmu Komputer dan Aplikasi*, vol. 9, no. 3, pp. 1133–1144, 2024.

[18] F. Duran, F. Wijaya, Y. R. Hulu, M. Harahap, and A. Prabowo, "Performance Comparison of Random Forest Classifier and LightGBM Classifier Algorithms for Heart Disease Prediction," *Data Sci. Indones.*, vol. 3, no. 2, pp. 98–103, 2024, doi: 10.47709/dsi.v3i2.3831.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[19] K. Science, E. Journal, and M. A. Engin, "Parkinson's Disease Detection via Machine Learning Using Data Splitting and Validation Methods," *Karaelmas Fen ve Mühendislik Dergisi* (*KaraelmasFen*), vol. 14, no. 2, pp. 134–147, 2024, doi: 10.7212/karaelmasfen.1484222.

[20] F. O. Awalullaili et al., "Classification of Hypertension Using the SVM Grid Search Method and SVM Genetic Algorithm," *Jurnal Gaussian* (*J. Gauss*), vol. 11, no. 4, pp. 488–498, 2023, doi: 10.14710/j.gauss.11.4.488-498.

[21] T. Burzykowski, M. Geubbelmans, A. Rousseau, and D. Valkenborg, "Validation of machine learning algorithms," *Am. J. Orthod. Dentofac. Orthop.*, vol. 164, no. 2, pp. 295–297, doi: 10.1016/j.ajodo.2023.05.007.

[22] N. R. Muntiari, K. H. Hanif, and L. Herawati, "Detection of Stunting in Toddlers Using Comparison Method," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 13, no. 1, pp. 17–23, 2025.

[23] N. Nurdiansyah et al., "Mental Health Analysis to Prevent Mental Disorders in Students Using K-Nearest Neighbor and Random Forest Algorithms," *Jurnal Ilmiah Teknologi dan Komputer*, vol. 5, no. 1, pp. 1–9, 2025.

[24] M. I. Elim and E. Utami, "Performance Comparison of Child Stunting Prediction: Support Vector Machine vs Random Forest with Grid Search Optimization," *Jurnal Teknologi dan Sistem Informasi*, vol. 6, no. 5, pp. 5305–5319, 2025.

[25] A. Syukron, "Application of the SMOTE Method to Overcome Class Imbalance in Heart Failure Prediction," *Jurnal Gaussian* (*J. Gauss*), vol. 10, no. 1, pp. 47–50, 2023.