

## COMPARATIVE ANALYSIS OF RANDOM FOREST, KNN, AND SVM FOR TODDLER STUNTING CLASSIFICATION

Maritza Ayu Shula<sup>1\*</sup>, Sri Siswanti<sup>1</sup>

<sup>1</sup>Informatics, Universitas Tiga Serangkai

*email: \*22500047.maritza@tsu.ac.id*

**Abstract:** Stunting is a chronic nutritional condition in toddlers characterized by a Height-for-Age (HFA) measurement below the standard growth threshold, necessitating early detection to prevent long-term consequences. This study aims to classify toddler stunting status by comparing three machine learning methods: Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The dataset comprises 345 toddler records from Puskesmas Indramayu (2025), including weight, height, and nutritional status based on WFA, HFA, and WFH indicators. Preprocessing steps include data cleaning, StandardScaler normalization, One-Hot Encoding for categorical features, and splitting the training and testing data with a ratio of 80:20. The comparison results are that KNN achieved the best performance with an accuracy of 71.01%, a precision of 0.69, a recall of 0.69, and an F1 score of 0.67, while RF and SVM both had an accuracy of 69.57% with F1 scores of 0.67 and 0.68, respectively. Thus, KNN demonstrated superior effectiveness in classifying the stunting status of toddlers compared to RF and SVM on this dataset.

**Keywords:** KNN; Random Forest; SVM; Stunting; toddlers

**Abstract:** Stunting adalah kondisi gizi kronis pada balita yang ditandai dengan pengukuran Tinggi Badan menurut Usia (HFA) di bawah ambang batas pertumbuhan standar, sehingga memerlukan deteksi dini untuk mencegah konsekuensi jangka panjang. Penelitian ini bertujuan untuk mengklasifikasikan status stunting pada balita dengan membandingkan tiga metode pembelajaran mesin: Random Forest (RF), K-Nearest Neighbor (KNN), dan Support Vector Machine (SVM). Kumpulan data terdiri dari 345 catatan balita dari puskesmas indramayu (2025), termasuk berat badan, tinggi badan, dan status gizi berdasarkan indikator WFA, HFA, dan WFH. Langkah-langkah prapemrosesan meliputi pembersihan data, normalisasi StandardScaler, One-Hot Encoding untuk fitur kategorikal, serta pembagian data pelatihan dan pengujian dengan rasio 80:20. Hasil perbandingan adalah KNN mencapai kinerja terbaik dengan akurasi 71,01%, presisi 0,69, recall 0,69, dan skor F1 sebesar 0,67, RF dan SVM keduanya memiliki akurasi 69,57% dengan skor F1 masing-masing sebesar 0,67 dan 0,68. Dengan demikian, KNN menunjukkan keefektifan yang lebih unggul dalam mengklasifikasikan status stunting balita dibandingkan dengan RF dan SVM pada dataset ini.

**Kata kunci:** KNN; random forest; SVM; Stunting; Balita

## INTRODUCTION

Stunting is a chronic nutritional condition in toddlers resulting from prolonged nutritional deficiencies, especially

during the first 1,000 days of life [1]. In Indonesia, the national prevalence has declined from 21.6% (2022) to 19.8% (2024) yet remains high in certain regions, highlighting the need for systematic



ic early detection at community health centers (Puskesmas) [3]. Machine learning offers a promising approach to classify nutritional status more effectively and systematically.

Prior studies have explored this problem from various angles. Sholikhin & Atmojo achieved 83% accuracy using a single KNN model [4], while Wiratama & Aziz reported RF outperforming SVM at 99.97% accuracy on a large public dataset [5]. Pratama & Fajri identified class imbalance as a key limitation affecting minority-class prediction [6]. Non-computational approaches by Yunani [7] and Supriyadi et al. [8] further emphasize the role of accurate anthropometric measurements in early detection. Nevertheless, a research gap remains: most prior studies rely on pre-cleaned public datasets, test single models without comparison, and omit optimization techniques in real-world clinical settings where class imbalance and local data variability are prevalent.

This study addresses that gap by simultaneously comparing Random Forest, KNN, and SVM with hyperparameter tuning (GridSearchCV and RandomizedSearchCV) on a real Puskesmas Indramayu dataset. Evaluation is conducted via confusion matrix metrics (precision, recall, and F1-score) to minimize misclassification in extreme stunting cases, with the goal of recommending the most valid model for Puskesmas-level early detection.

**METHODS**

This study uses a quantity-based approach with machine learning methods to classify the stunting status of toddlers based on the height-for-age (HFA) indicator. The research data were obtained weighing dates, which were later re-

from measurements of toddlers at community health centers (Puskesmas), including initial and final weight and height, as well as initial nutritional status. The research stages consist of (1) data collection and label determination, (2) preprocessing and feature engineering, (3) splitting the data into training and test sets, (4) model training using Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM), (5) parameter optimization (hyperparameter tuning), (6) model evaluation using classification metrics and a confusion matrix, and (7) model implementation into a Flask-based web application. The workflow of the study is shown in Image 1.

**(Research Flowchart).**

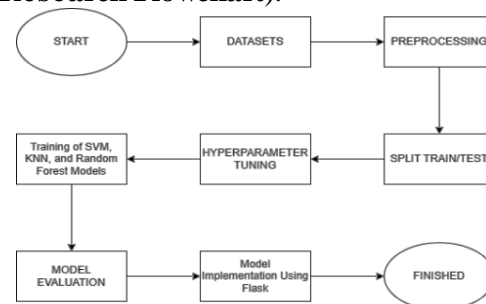


Image 1. Research Flowchart

**Dataset and Label Determination**

The dataset consists of 345 anthropometric records of toddlers from Puskesmas Indramayu (2025), collected through routine growth monitoring. Input attributes include initial and final body weight (BW) and height (H), along with initial nutritional status based on WFA, HFA, and WFH indicators. The target class is the final HFA status, with categories of normal, short, and very short. Irrelevant administrative columns, such as measurement dates, were removed during feature selection to optimize model performance. As shown in Table 1, the original dataset contained administrative attributes, including the initial and final

cess. As shown in Table 2, both the “Initial weighing date” and “Final weighing date” columns have been removed since they are not required for the training process of the mod.

**Pre-processing and Feature Engineering**

Pre-processing involved removing irrelevant data, checking for missing values, and cleaning numeric formats. Categorical features (WFA, HFA, WFH) were encoded using One-Hot Encoding, while numerical features were standardized via StandardScaler (z-score). Feature engineering was applied by computing weight ( $\Delta BW$ ) and height ( $\Delta H$ ) differences between initial and final measurements to capture toddler growth changes.

$$Z = \frac{(x - \mu)}{\sigma} \tag{1}$$

**Description:**

Z is the standardized value, x is the original feature value,  $\mu$  represents the mean of the feature, and  $\sigma$  is the standard deviation.

Where is the feature value, the feature mean, and the standard deviation?

Feature engineering is performed to capture changes in measurements, such as differences in weight and height.

$$\Delta BW = BW_{final} - BW_{initial} \tag{2}$$

$$\Delta H = H_{final} - H_{initial} \tag{3}$$

In this study,  $\Delta BW$  denotes the change in toddler body weight, calculated as the difference between the body weight at the final measurement ( $BW_{final}$ ) and the body weight at the initial measurement ( $BW_{initial}$ ). Likewise,  $\Delta H$  denotes the change in toddler height, calculated as the difference between the height at the final measurement ( $H_{final}$ ) and the height at the initial measurement ( $H_{initial}$ ).

When class distribution was imbalanced, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to reduce model bias toward the majority class.

**Data Split**

As shown in Table 3, the dataset was divided into training and testing sets using an 80:20 ratio

Table 3. Train Data and Test Data

Train Data	Test Data
80%	20%

Table 1. Dataset before removing the “Initial weighing date” and “Final weighing date”

Puskesmas	Child's Name	Parents' Names	Initial weighing date	...	HFA	WFH
Puskesmas 1	Child 1	Parent 1	30/06/2025	...	short	good nutrition
Puskesmas 2	Child 2	Parent 2	30/06/2025	...	short	good nutrition
Puskesmas 3	Child 3	Parent 3	Juni 2025	...	normal	good nutrition
Puskesmas 4	Child 4	Parents 4	Juni 2025	...	normal	good nutrition

Table 2. Dataset after removing the “Initial weighing date” and “Final weighing date”

Puskesmas	Child's Name	Parent's Name	...	TB/U	BB/TB
Puskesmas 1	Child 1	Parent 1	...	short	good nutrition
Puskesmas 2	Child 2	Parent 2	...	short	good nutrition
Puskesmas 3	Child 3	Parent 3	...	normal	good nutrition
Puskesmas 4	Child 4	Parent 4	...	normal	good nutrition

**Classification Model Development**

This study compare three

classification algorithms such as:

**Random Forest:** an ensemble model that combines multiple decision trees, where the final prediction is determined through majority voting, and can be generally expressed as:

$$\hat{Y} = mode\{h_1(x), h_2(x), \dots, h_t(x)\} \quad (4)$$

In the model,  $\hat{Y}$  represents the final prediction result obtained by aggregating the outputs of all decision trees. The prediction produced by the  $t$ -th decision tree is denoted as  $h_t(x)$ , where  $t$  indicates the index of the decision tree and  $x$  represents the input features used for prediction.

Where  $h_t(x)$  is the prediction of the  $t$ -th tree and  $t$  is the total number of trees.

**K-Nearest Neighbor (KNN):** classifies data points based on the proximity to their nearest neighbors, using Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2} \quad (5)$$

In this equation,  $(d(x_i, x_j))$  represents the distance between data point (i) and data point (j). The variable  $(x_{ik})$  denotes the value of the (k)-th feature in data point (i), while  $(x_{jk})$  denotes the value of the (k)-th feature in data point (j). The variable (n) represents the total number of features used in the distance calculation.

**Support Vector Machine (SVM):** constructs a separating hyperplane with maximum margin. The decision function is expressed as:

$$f(x) = w^t x + b \quad (6)$$

In the Support Vector Machine (SVM) model,  $(f(x))$  represents the decision function used to classify the input

data. The variable (w) denotes the weight vector that determines the orientation of the decision boundary, while (x) represents the input feature vector. The variable (b) denotes the bias term, which adjusts the position of the decision boundary relative to the origin.

### Hyperparameter Tuning

To improve performance, an optimization of the best parameters is performed using GridSearchCV with k-fold cross-validation (for example). Commonly tested parameters include: The machine learning algorithms evaluated in this study include Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). These algorithms were selected to compare their performance in predicting toddler nutritional status based on anthropometric data. The best parameters were selected based on the highest average cross-validation accuracy.

### Model Evaluation

Model evaluation was performed using accuracy, along with additional metrics such as precision, recall, and F1-score for each class. A confusion matrix was also employed to identify prediction error patterns. The general form of the confusion matrix is shown in Table 4.

Table 4. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

### Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

TP (True Positive) and TN (True Negative) refer to correctly predicted

positive and negative data, respectively. Conversely, FP (False Positive) refers to negative data that is incorrectly predicted as positive, and FN (False Negative) refers to positive data that is incorrectly predicted as negative.

**Precision**

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

TP (True Positive) is the number of positive data points that are correctly predicted, while FP (False Positive) is the number of negative data points that are predicted to be positive.

**Recall**

$$\text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

FN (False Negative) adalah jumlah data positif yang salah diprediksi sebagai negatif.

**F1-Score**

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

F1-Score is the harmonic mean of Precision and Recall, used to assess the balance between accuracy and completeness in classification results.

**RESULT AND DISCUSSION**

Classification of toddler stunting status (final HFA indicator) was performed using Random Forest, KNN, and SVM with an 80:20 train-test split, yielding 274 training samples and 69 test samples after preprocessing.

Figure 2 shows the distribution of data after the train-test splitting process, resulting in 274 training samples and 69 testing samples.

```

X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=42,
    stratify=y
)

print("\nData latih:", X_train.shape, " | Data uji:", X_test.shape)
...
Data latih: (274, 11) | Data uji: (69, 11)
    
```

Image 2. Train Data and Test Data

Evaluation was conducted using accuracy, classification report, and confusion matrix metrics across three classes: normal, short, and very short. As shown in Table 5, KNN achieved the highest accuracy of 0.710, followed by Random Forest at 0.696 and SVM at 0.667.

As shown in Table 5, KNN achieved the highest accuracy among the three evaluated models

Table 5. Model Accuracy Comparison on Test Data

Model	Accuracy
KNN	0,710
SVM	0,667
Random Forest	0,696

Confusion matrix analysis reveals that Random Forest and SVM frequently misclassify "normal" samples as "short" due to overlapping feature patterns, while KNN shows more stable predictions across both classes. Nevertheless, all three models struggle with the "very short" class owing to limited samples, leading to consistent minority-class misclassification. Figure 3 and 4 presents the confusion matrix of the Random Forest model, illustrating the classification performance across all classes.

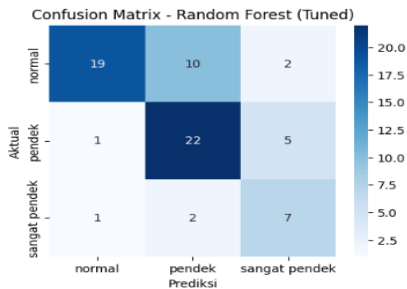


Image 3. Random Forest Confusion Matrix

	precision	recall	f1-score	support
normal	0.90	0.61	0.73	31
pendek	0.65	0.79	0.71	28
sangat pendek	0.50	0.70	0.58	10
accuracy			0.70	69
macro avg	0.68	0.70	0.67	69
weighted avg	0.74	0.70	0.70	69

Image 4. Random Forest Classification Report

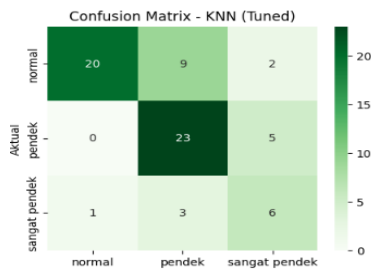


Image 5. KNN Cofusion Matrix

	precision	recall	f1-score	support
normal	0.95	0.65	0.77	31
pendek	0.66	0.82	0.73	28
sangat pendek	0.46	0.60	0.52	10
accuracy			0.71	69
macro avg	0.69	0.69	0.67	69
weighted avg	0.76	0.71	0.72	69

Image 6. KNN Classification Report

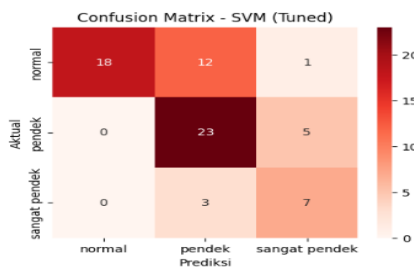


Image 7. SVM Cofusion Matrix

	precision	recall	f1-score	support
normal	1.00	0.58	0.73	31
pendek	0.61	0.82	0.70	28
sangat pendek	0.54	0.70	0.61	10
accuracy			0.70	69
macro avg	0.71	0.70	0.68	69
weighted avg	0.77	0.70	0.70	69

Image 8. SVM Classification Report

Across all three confusion matrices, the most frequent prediction error occurs when the "normal" class is misclassified as "short," while the "very short" class remains the hardest to predict due to its limited test samples. This underscores the need for more data and representative features to sharpen class boundaries.

As illustrated in Figure 9, KNN outperformed Random Forest and SVM in terms of overall classification accuracy.

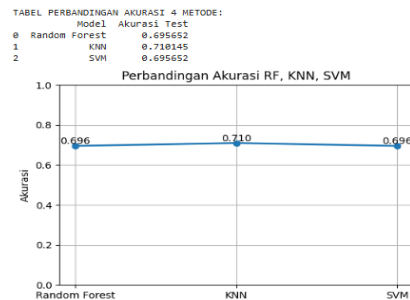


Image 9. Comparison of Methods

From the test results, the KNN method yielded the highest accuracy of 0.710, while Random Forest and SVM achieved 0.696. Although the difference in accuracy among the models was relatively small, KNN proved to be the most effective in capturing patterns of proximity in the normalized anthropometric data for classifying stunting status in toddlers.

## CONCLUSION

This study successfully identified

KNN as the most suitable algorithm for classifying toddler stunting status (normal, short, very short) based on the final HFA indikator, achieving the highest accuracy of 0.710 compared to Random Forest (0.696) and SVM (0.667). KNN's superiority stems from its local distance-based principle (Euclidean Distance), which is more sensitive to the variability of real-world clinical data than SVM's hyperplane approach or Random Forest's tree structure. However, confusion matrix analysis reveals persistent misclassification between the "short" and "very short" classes due to overlapping anthropometric values in practice. These findings imply that future stunting detection systems must go beyond simply increasing data volume — they must incorporate class imbalance handling techniques and additional non-anthropometric features to improve diagnostic accuracy.

## REFERENCES

- [1] Hadriani, Hadina, R. Arianty, A. Fatmawati Syamsu, F. Kolomboy, and N. Yulita Siregar, "Edukasi Stunting Sebagai Upaya Meningkatkan Pengetahuan Ibu Balita Dalam Pencegahan Stunting di Desa," *Ju Jurnal Kolaboratif Sains*, vol. 7, no. 11, pp. 4003–4012, 2024, doi: [10.56338/jks.v7i11.6617](https://doi.org/10.56338/jks.v7i11.6617).
- [2] M. E. Setiyawati, L. P. Ardhiyanti, E. N. Hamid, N. Ayu, T. Muliarta, and Y. J. Raihanah, "Studi Literatur: Keadaan Dan Penanganan Stunting Di Indonesia," *IKRAITH-HUMANIORA*, Jul. 2024, doi: [10.37817/ikraith-humaniora.v8i2](https://doi.org/10.37817/ikraith-humaniora.v8i2).
- [3] Kemenkes, "SSGI 2024: Prevalensi Stunting Nasional Turun Menjadi 19,8%," Kemenkes. Accessed: Dec. 21, 2025. [Online]. Available: <https://kemkes.go.id/id/ssgi-2024-prevalensi-stunting-nasional-turun-menjadi-198>
- [4] N. A. Sholikhin and S. Atmojo, "Aplikasi WEB Untuk KLASIFIKASI STUNTING Pada Balita Dengan Menggunakan Metode K-NEAREST NEIGHBOURS (Studi Kasus Posyandu Jawa Kidul)," *The Journal of System Engineering and Technological Innovation*, vol. 01, no. Vol 1 No 02 (2022): Oktober 2022, pp. 44–47, Oct. 2022, doi: <https://doi.org/10.38156/jisti.v1i02.23>.
- [5] Y. Wiratama and R. A. Aziz, "Perbandingan Prediksi Penyakit Stunting Balita Menggunakan Algoritma Support Vektor Machine dan Random Forest," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, pp. 1159–1168, Sep. 2024, doi: [10.47065/bits.v6i2.5543](https://doi.org/10.47065/bits.v6i2.5543).
- [6] S. R. Pratama and I. N. Fajri, "Comparison of K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) Algorithms in Predicting Customer Satisfaction," *Journal of Computer Science and Informatics Engineering*, vol. 4, no. 3, pp. 135–146, Jun. 2025, doi: [10.55537/cosie.v4i3.1160](https://doi.org/10.55537/cosie.v4i3.1160).
- [7] Y. Yunani, "Deteksi dini status kesehatan balita dan status kesehatan ibu balita resiko stunting pada anak balita," *MCHC THE JOURNAL OF Mother and Child Health Concerns*, no. Vol. 4 No. 5 (2025): June Edition 2025, Jun. 2025, doi: <https://doi.org/10.56922/mchc.v4i5.1066>.
- [8] Supriyadi, E. Oktavianto, D. Nur Adkhana Sari, and P. Studi Keperawatan Stikes Surya Global, "Pendidikan Kesehatan Tentang Pencegahan Stunting Pada Ibu Balita," *Jurnal Pengabdian Masyarakat Kesehatan Terkini*, no. Vol.

- 2 No. 2 (2023), Aug. 2023, doi: <https://doi.org/10.58516/4q9t8t83>.
- [9] F. Ramadhani *et al.*, “Sosialisasi 1000 HPK Sebagai Upaya Pencegahan Stunting Pada Balita Di Kabupaten Gorontalo,” *Insan Cita: Jurnal Pengabdian Kepada Masyarakat*, no. Vol. 5 No. 2 (2023), Aug. 2023, doi: <https://doi.org/10.32662/insancita.v5i2.2444>.
- [10] M. Mulyono, E. Budianita, A. Nazir, and F. Syafria, “Klasifikasi Status Stunting Balita Menggunakan Metode Naïve Bayes Gaussian Berbasis Web,” *Jurnal Informatika Universitas Pamulang*, vol. 8, no. 3, pp. 399–406, Sep. 2023, doi: [10.32493/informatika.v8i3.33399](https://doi.org/10.32493/informatika.v8i3.33399).
- [11] S. Marsya Finda and D. Wahyu Utomo, “Klasifikasi Stunting Balita menggunakan Metode Ensemble Learning dan Random Forest,” *Jl. Imam Bonjol No*, vol. 15, no. 02, 2024, doi: [10.35970/in\\_fotekmesin.v15i2.2326](https://doi.org/10.35970/in_fotekmesin.v15i2.2326).
- [12] T. Prasetya, I. Ali, C. L. Rohmat, and O. Nurdiawan, “Klasifikasi Status Stunting Balita Di Desa Slangit Menggunakan Metode K-Nearest Neighbor,” *INFORMATICS FOR EDUCATORS AND PROFESSIONALS*, vol. 4, no. 2, pp. 93–104, 2020.
- [13] Lia Lumbaa, Evangs Mailoa, and Magdalena A. Ineke Pakereng, “Implementasi Metode SVM Dan Gradient Boost Dalam Klasifikasi Bahasa Daerah (Halmahera, Kalimantan, Toraja),” *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 908–915, no. Vol 9 No 2 (2022): JATISI (Jurnal Teknik Informatika dan Sistem In-
- formasi), Jun. 2022, doi: <https://doi.org/10.35957/jatisi.v9i2.1663>.
- [14] RB Fajriya Hakim, “SVM (Support Vector Machine),” Medium. Accessed: May 30, 2026. [Online]. Available: <https://medium.com/@986110101/svm-support-vector-machine-6ee9fccd4222>
- [15] Y. Fauziah, F. Khairani, A. N. Nasution, S. Tinggi, I. K. Flora, and U. S. Utara, “Perubahan Pengetahuan Ibu Sebelum Dan Sesudah Membaca Media Leaflet Tentang Stunting Pada Ibu Anak Balita Stunting,” *Jurnal Kesehatan Ilmiah Indonesia (Indonesian Health Scientific Journal)*, vol. 9, no. Vol. 9 No. 1 (2024);, p. 220, 2024, doi: <https://doi.org/10.51933/health.v9i1.1287>.
- [16] S. Suhaemi, H. Hidayani, and A. S. Rini, “Hubungan Pola Asuh Status Gizi Pengetahuan Ibu Dengan Kejadian Stunting Pada Balita,” *SIMFISIS: Jurnal Kebidanan Indonesia*, vol. 3, no. 2, pp. 615–622, Nov. 2023, doi: [10.53801/sjki.v3i2.185](https://doi.org/10.53801/sjki.v3i2.185).
- [17] S. P. I. Hadi, R. I. Hakim, and S. Pabidang, “Pencegahan Stunting melalui Edukasi Kebutuhan Dasar Anak di Desa Sewukan, Kec. Dukun Kab. Magelang,” *Jurnal Kreativitas Pengabdian Kepada Masyarakat (PKM)*, vol. 8, no. 4, pp. 2040–2050, Apr. 2025, doi: [10.33024/jkpm.v8i4.18722](https://doi.org/10.33024/jkpm.v8i4.18722).
- [18] S. Sukati, S. Aisyah, W. Ernawati, and A. Zuitasari, “Faktor-Faktor Yang Mempengaruhi Kejadian Stunting Pada Anak Balita Di Wilayah Kerja UPTD Puskesmas Peninjauan Tahun 2022,” *Jurnal Ilmu Kebidanan dan Kesehatan*, vol. 15, no. 1, pp. 30–35, 2024, doi: [10.52299/jks.v15i1.217](https://doi.org/10.52299/jks.v15i1.217).