Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15380
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Analysis of Factors Causing Toddler's Malnutrition in Medan City Using the Random Forest Method

Windi Saputri Simamora[1)*], Siti Sarah Harahap[2)], Andre Pratama[3)]
[1,2,3)] Program Studi Informatika, Universitas Satya Terra Bhinneka, Indonesia
[1)] windisimamora@satyaterrabhinneka.ac.id , [2)]sarahharahap@satyaterrabhinneka.ac.id,
[3)]andrepratama@satyaterrabhinneka.ac.id

Submitted :Oct 30, 2025 | **Accepted** : Nov 10, 2025 | **Published** : Jan 02, 2026

Abstract: Malnutrition and severe malnutrition in toddlers remain critical public health concerns that impair physical growth, cognitive development, and long-term productivity. Deficiencies in essential nutrients increase the risks of stunting, weakened immunity, and developmental delays. Although interventions such as supplementation and routine anthropometric monitoring are implemented, comprehensive identification of multidimensional causal factors is still limited, reducing the effectiveness of targeted policies. This study aims to predict toddler nutritional status using a quantitative data mining approach. A dataset consisting of 328 samples and 17 features was collected from health facilities in Medan City, including *Puskesmas*, the Health Office, and *Posyandu*. A Random Forest Classifier was developed with missing-value handling, feature engineering, and feature importance analysis to identify dominant predictors of nutritional outcomes. The model achieved an overall accuracy of 92.42 percent and showed strong performance in identifying the "Normal" class, although predictive sensitivity for minority classes such as "Gizi Kurang" and "Gizi Buruk" remained comparatively lower. Feature importance analysis indicated that complete immunization and health insurance ownership were the most influential determinants of nutritional status. This research provides a machine learning–based tool for early nutritional risk prediction and offers data-driven insights to support more precise malnutrition interventions. Future enhancement may include expanding feature diversity and applying advanced interpretability techniques to strengthen model reliability. The findings reinforce the importance of evidence-based nutrition policy strategies that prioritize early prevention and improved child health outcomes.
Keywords: Classification; Feature Importance; Machine Learning**;** Nutritional Status; Random Forest; Toddlers;

INTRODUCTION

Nutrition is a basic food substance needed for health and body development. Malnutrition in toddlers remains a critical public health issue. Toddlers who lack essential nutrients like vitamin A, iron, and protein have a higher risk of stunting, impaired immunity, and hindered cognitive development (WHO, 2015).Child Malnutrition is a multidimensional public health challenge that threatens healthy growth, cognitive development, and future productivity of children. Both acute and chronic forms increase disease vulnerability, heighten mortality risk, and impede physical and intellectual development (Badan Kebijakan Pembangunan Kesehatan, 2024). The negative impact extends beyond individuals to the national burden through higher healthcare costs and reduced economic potential. In North Sumatra, 1.96 percent of toddlers were recorded with poor nutritional status in 2023 (Utara, 2023), while in Medan City, 521 of 119,225 toddlers, or 0.6 percent, were classified as having poor growth performance based on weight-for-age indicators (DINKES Kota Medan, 2022). Although health authorities continuously provide vitamin supplementation and routine anthropometric monitoring to improve child growth, these efforts have not fully mapped the combined influence of socioeconomic, environmental, and educational determinants. A more comprehensive analytical approach is crucial to ensure earlier detection and more precise intervention strategies.

Previous studies have demonstrated the potential of machine learning, particularly Random Forest, in assessing and predicting nutrition-related outcomes in early childhood(Aprilia et al., n.d.; Pratama et al., 2023;

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15380
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Setiawan & Triayudi, 2022). Nevertheless, the majority of these studies have concentrated primarily on accuracy improvement or algorithmic performance rather than creating robust models that incorporate real, regional, and multidimensional datasets (Khusna et al., 2024). A study by Jajang Jaya Purnama yielded results that focused on the method. The random forest algorithm, combined with the undersampling-resample technique, is highly suitable for classifying malnutrition in imbalanced data, but the dataset was not derived from a specific real-time local context (Purnama, 2020). Similarly, Candra et al. improved performance using a genetic algorithm for hyperparameter optimization, yet the focus remained methodological rather than interpretative(Candra et al., 2022). Laia et al. (2023) identified family income as a dominant determinant but did not apply machine learning for broader factor analysis (Laia et al., 2023). Meanwhile, studies by Cappelli et al. and Gebeye et al. demonstrated the usefulness of feature importance techniques such as Permutation Feature Importance (PFI) and SHAP in identifying influential environmental and socioeconomic variables, although applied in different health domains(Cappelli et al., 2024; Gebeye et al., 2023)..

Addressing these research gaps, this study adopts a quantitative, data mining-based approach using the Random Forest algorithm to analyze the nutritional status of toddlers in Medan City. The dataset integrates multiple dimensions including socioeconomic, health, education, and environmental information sourced directly from Puskesmas, the Health Office, and Posyandu. The approach is nonparametric, enabling the extraction of complex relationships among variables without strong statistical distribution assumptions. The inclusion of feature importance analysis allows systematic identification of the most influential predictors, generating interpretable evidence to support decision-making by public health practitioners.

The contribution of this study lies in the comprehensive integration of real-time multidimensional data, advancing the use of machine learning beyond pure classification performance through actionable insights derived from feature importance modeling. The expected outcomes include improving early identification of toddlers at risk, strengthening surveillance systems, and supporting more targeted nutritional interventions tailored to the specific conditions of Medan City. Through its methodological innovation and practical relevance, the study reinforces the urgency of data-driven nutrition governance and provides empirical evidence to enhance policy formulation aimed at reducing long-term health inequities and promoting sustainable human capital development in the region.

LITERATURE REVIEW

Nutritional Status of Toddlers

Toddler nutritional status is a physical condition that can be caused by several factors and can impact their cognitive level (Handayani & Charis Fauzan, 2024). A toddler's nutritional status can be influenced by various direct factors, such as infectious diseases, birth history, exclusive breastfeeding, and food quality and quantity. It can also be caused by indirect factors, such as socioeconomic status, education, knowledge, and attitudes toward health services. These factors determine a toddler's mental and physical development. Malnutrition can cause serious impacts, such as impaired physical growth, an increased risk of disease, and delayed mental development (Pangumbara'an, 2021). Measuring toddler nutrition requires a threshold for interpretation. Therefore, each toddler's weight and age are converted into a Z-score using the WHO's standard anthropometric formula. The Z-score calculation is as shown in the following equation(Setiawan & Triayudi, 2022):

$$Z - score = \frac{Individual\ Subject\ Value - Reference\ Standard\ Median\ Value}{Reference\ Standard\ Deviation\ Value} \quad (1)$$

Table 1 Toddler Nutritional Status Categories

Index	Nutritional Status	Threshold
Body Weight for Age (BB/U)	Severe Malnutrition	<-3,0 SD
	Malnutrition	<-2,0 SD to >= -3,0 SD
	Normal Nutrition	>= -2 SD to 2,0 SD
	Overnutrition	>2,0 SD
Height for Age (TB/U)	Very Short	<-3,0 SD
	Short	<-2,0 SD to >= -3,0 SD
	Normal	>= -2 SD to 2,0 SD
	Tall	>2,0 SD
Height to Weight (TB/BB)	Severe Malnutrition	<-3,0 SD
	Malnutrition	<-2,0 SD to >= -3,0 SD
	Normal Nutrition	>= -2 SD to 2,0 SD
	Overnutrition	>2,0 SD

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15380
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Random Forest Method

Random Forest is a machine learning technique that combines data from multiple decision trees to achieve more accurate results. The random forest method works by creating a decision tree from randomly selected data and features (Juwariyem et al., 2024). The resulting decision trees are numerous, hence the name forest. The data testing method for a random forest involves inputting data into all the resulting decision trees, and the results are determined by the highest number of answers (Aprilia et al., n.d.).

The flowchart for creating a decision tree using a random forest can be seen in the following figure (Verikas et al., 2011):

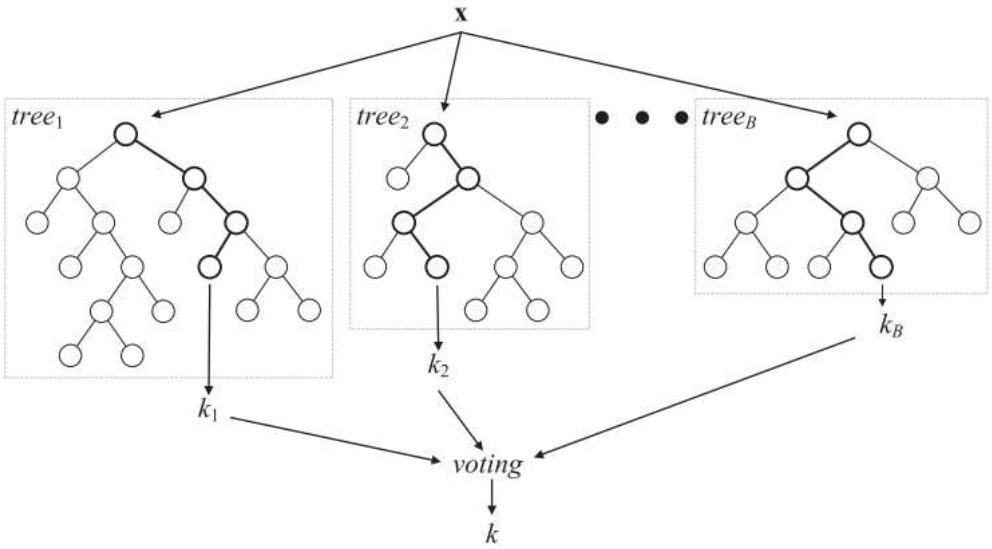

Fig. 1. A general architecture of a random forest.

This random forest algorithm can calculate entropy, which can determine information gain using the following equation:

$$Entropy\ (S) = \sum_{i=1}^{n} -pi \log 2\ pi \quad (2)$$

Where S (Case Set) is the entire dataset being analyzed, while A (Features) are the individual attributes or characteristics of each data point. n (Number of Partitions) indicates the number of divisions of the dataset, and pi (Proportion of samples for class i) is the percentage of samples belonging to a specific class.

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Sv|}{|S|} * Entropy(Si) Entropy(Si) \quad (3)$$

Where A (Attribute) refers to a specific feature or characteristic of the data. V (State a possible value for attribute A) is one of the potential values that attribute A can take. |Si| (Number of cases in the i-th partition) is the count of data points in a specific subset, while |S| (Total number of data samples) is the total number of data points in the entire dataset. Lastly, Entropy(Si) is a measure of the randomness or impurity within the samples that have value 'i'.

Previous Study

To strengthen the positioning of the proposed Random Forest model within the current body of research, a comparative review of recent machine learning studies on malnutrition prediction was conducted. Table 2 summarizes key publications from 2023–2025, highlighting variations in methods, dataset sizes, evaluation metrics, and methodological limitations.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15380
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2 Comparative Performance of Machine Learning Models for Malnutrition Prediction

Authors	Method	Dataset Size	Main Results (Accuracy / F1 / AUC)	Limitations
(Qasrawi et al., 2024)	Random Forest, Gradient Boosting, SVM, Logistic Regression	National-level Palestinian child nutrition survey dataset	Random Forest achieved the best performance with **Accuracy 97.7%**, **AUC 0.996**, and **F1 0.976**	The dataset contains highly detailed nutritional information (e.g., micronutrient intake, food consumption), which is not directly comparable to socio-economic datasets. The potential risk of overfitting due to rich feature granularity.
(Janssen et al., 2025)	XGBoost, GLMNet, RF, SVM	412 NICU patients (Turkey)	XGBoost: **AUC 0.79**, Accuracy 81%, F1 0.74	Focus on NICU patients. Limited external validation. The dataset did not apply balancing techniques.
(Nirmani & Kudagamage, 2024)	Ensemble Stacking (SVM + AdaBoost + XGBoost)	574 children under 5 (Sri Lanka)	Accuracy **93%**, F1 0.91	Only 8 predictor variables and limited socio-economic factors. The sample from a single region.
(Gol et al., 2024)	Random Forest, Logistic Regression, SVM, KNN	1,038 geriatric hospital patients (Saudi Arabia)	RF: **Accuracy 94.1%**, **AUC 0.96**, F1 0.92	Focus on **malnutrition-related anemia** in elderly adults not directly applicable to toddlers and relies on clinical/biomarker features.
(Kaur & Neeru, 2024)	Random Forest, CNN, Deep Learning models (review)	Multi-domain datasets (review study)	DL models typically >95% accuracy; ML models (RF/SVM) 85–93%	General review and not specific to malnutrition. The performance varies depending on dataset size.

METHOD

This research method uses the Random Forest algorithm approach. The following is a diagram of the algorithm process in the random forest.

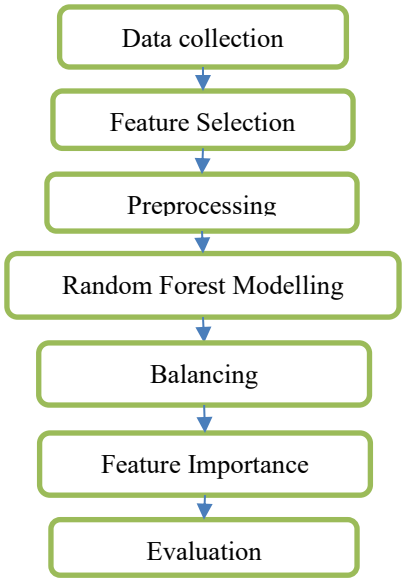

Fig. 2. Research Flowchart

The following is the analysis process flow:

a. Data collection: Data was obtained from several community health centers (Puskesmas), the Medan City Health Office, and integrated health posts (Posyandu) in Medan City. The data collected included toddler nutritional status, such as weight and height, along with various contributing factors such as father's occupation, father's medical history, mother's occupation, mother's medical history, birth weight (g), toddler's

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15380
e-ISSN : 2541-2019
p-ISSN : 2541-044X

medical history, immunization status, health insurance, height (cm), income, floor area, ceiling, ventilation/windows, toilet, waste disposal area, and nutritional status.

b. Data analysis/feature selection: selecting/adding features to be used in the analysis process by removing unrelated categories.
c. Data preprocessing: the process of cleaning data from duplicates or missing data that could interfere with the analysis results. Missing data can be addressed by deleting or filling in missing values with other data. Duplicate data can be addressed by removing duplicates (Putri Ayu Firnanda et al., 2025).
d. Random Forest Modeling: After the data was cleaned and balanced, the dataset was divided into training and testing subsets, ensuring that the class distribution proportions in each subset matched the original data. A predictive model was then developed using the Random Forest algorithm, a popular and effective machine learning method for managing and analyzing large and complex data sets (Mubarok et al., 2025).
e. Balancing: At this stage, class_weight='balanced' is added to the RandomForestClassifier to automatically adjust the weight of each class based on their frequency in the training data, handling class imbalance by giving more attention to minority classes.
f. Feature Importance: A multi-factor feature importance modeling approach was used to identify factors influencing the nutritional status of toddlers.
g. Evaluation: Finally, the results were evaluated and interpreted to formulate intervention recommendations using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

RESULT

Dataset

The secondary data that has been collected and has been subjected to feature selection has 328 rows and 17 columns including information on the nutritional status of toddlers and several influencing factors. The features used in this study are Pekerjaan Ayah (Father's Occupation), Riwayat Penyakit Ayah (Father's Medical History), Pekerjaan Ibu (Mother's Occupation), Riwayat Penyakit Ibu (Mother's Medical History), BB Lahir (Birth Weight (g)), Riwayat Penyakit Balita (Toddler's Medical History), Status Imunisasi (Immunization Status), Jaminan Kesehatan (Health Insurance) , BB (Weight (kg)), Panjang (Length (cm)), Pendapatan (Income), Lantai (Floor), Plafon (Ceiling), Ventilasi/Jendela (Ventilation/Windows), Jamban (Toilet), Tempat Pembuangan Sampah (Waste Disposal Site), and Status Gizi (Nutritional Status). In this dataset, several NaN values are visible in several columns. Therefore, further data cleaning is necessary.

Table 3 Data Set of Toddler Nutritional Status Categories

No	Pekerjaan Ayah	Riwayat Penyakit Ayah	BB Lahir (g)	Status Imuinsasi	BB	Jaminan Kesehatan	...	Status Gizi
0	Wiraswasta	Tidak Ada	2000	Tidak Lengkap	4.5	Ada	...	Gizi Kurang
1	Nelayan	Tidak Ada	2300	Tidak Lengkap	4.12	Tidak Ada	...	Gizi Buruk
2	Buruh	Tidak Ada	2400	Tidak Lengkap	10.2	Tidak Ada	...	Gizi Buruk
3	Penjual Keliling	Tidak Ada	2500	Tidak Lengkap	9.4	Tidak Ada	...	Gizi Buruk
...	...	...	...	...	...	...	...	...
327	Wiraswasta	Tidak Ada	NaN	Lengkap	7.7	Ada	...	Normal

Preprocessing data

In this stage, values are filled in for columns with missing values (NaN). Columns with numeric values are filled with the average value (mean), while categorical columns are filled with the most frequently occurring value (mode). This stage produces the dataset shown in Table 3.

Table 4 Data Set of Toddler Nutritional Status Categories After Preprocessing

No	Pekerjaan Ayah	Riwayat Penyakit Ayah	BB Lahir (g)	Status Imuinsasi	BB	Jaminan Kesehatan	...	Status Gizi
0	Wiraswasta	Tidak Ada	2000	Tidak Lengkap	4.5	Ada	...	Gizi Kurang
1	Nelayan	Tidak Ada	2300	Tidak Lengkap	4.12	Tidak Ada	...	Gizi Buruk
2	Buruh	Tidak Ada	2400	Tidak Lengkap	10.2	Tidak Ada	...	Gizi Buruk
3	Penjual Keliling	Tidak Ada	2500	Tidak Lengkap	9.4	Tidak Ada	...	Gizi Buruk
...	...	...	...	...	...	...	...	...
327	Wiraswasta	Tidak Ada	2877.77	Lengkap	7.7	Ada	...	Normal

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15380
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Random Forest Tree Modeling

The preprocessed dataset was divided into training data (80%) and testing data (20%) to maintain the proportion of the target class. Dividing data into training and testing sets is a standard practice in machine learning to evaluate model performance on previously unseen data. The use of entropy criteria refers to how the decision tree divides the data based on information gain. The Random Forest Classifier model was developed using 100 trees (n_estimators=100), the entropy criterion (criterion="entropy") and class_weight='balanced' to evaluate the quality of each split. To ensure consistent and reproducible results, the random_state parameter was set to 0. The model was trained with the Out-of-Bag (OOB) score enabled (oob_score=True) to estimate generalization performance without the need for additional validation data. The Out-of-Bag Score was 0.9237, providing an estimate of performance on unseen data during training. A visualization of the Random Forest Decision Tree can be seen in the following image:

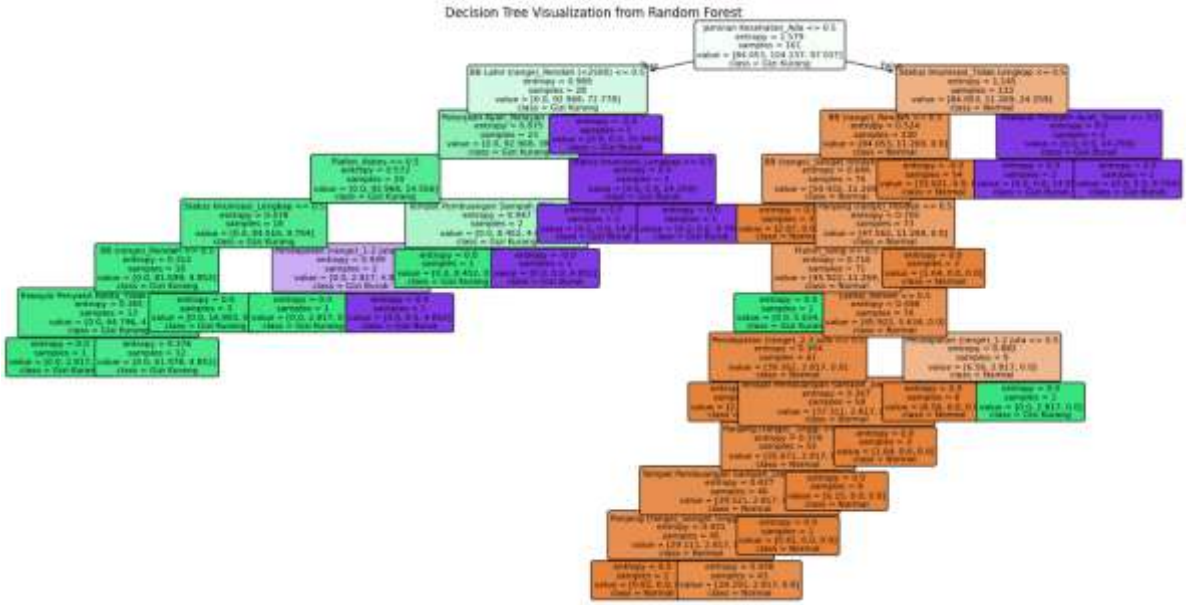

Fig.3. Decision Tree Visualization from Random Forest

The visualization illustrates one of the decision trees generated within the Random Forest model used to classify toddler nutritional status. Each internal node represents a decision rule based on a specific predictor, such as parental employment, household income, health insurance availability, immunization status, sanitation facilities, or environmental conditions. The branches indicate the path taken according to the feature threshold, leading to leaf nodes that display the predicted nutritional category along with entropy, sample count, and class distribution. The color variations of the nodes distinguish different predicted classes, demonstrating how combinations of socioeconomic and health-related factors contribute to determining whether a toddler falls into the “Normal,” “Gizi Kurang” (Malnutrition) or “Gizi Buruk” (Severe Malnutrition) category. This tree visualization enhances interpretability by showing how key variables influence model decisions and contribute to predicting nutritional outcomes.

Feature Importance

The importance of each feature in the Random Forest model was calculated and ranked in descending order. The features Jaminan Kesehatan_Ada, Status Imunisasi_Lengkap, Status Imunisasi_Tidak Lengkap, dan Jaminan Kesehatan_Tidak as the most important features based on their percentage importance, as visualized in the following figure:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15380
e-ISSN : 2541-2019
p-ISSN : 2541-044X

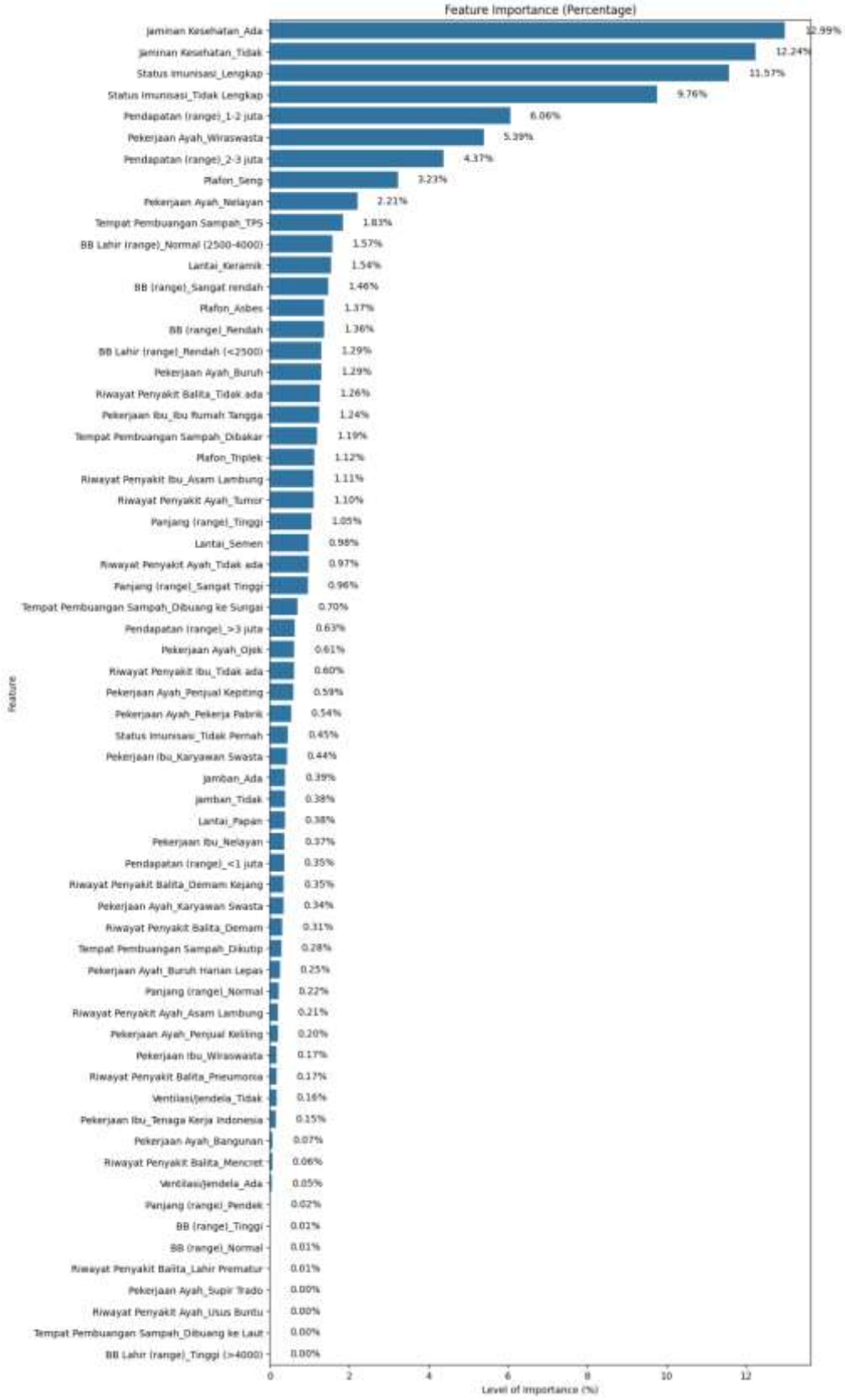

Fig.4. Feature Importance (Percentation)

Evaluation

The next step was testing using a Confusion Matrix to determine the Recall, Precision, Accuracy, and F1-score values obtained from the Random Forest method. The evaluation results showed an accuracy of 92.42%. Precision

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15380
e-ISSN : 2541-2019
p-ISSN : 2541-044X

is used to measure the model's accuracy when predicting a class. Precision for "Gizi Kurang" indicates the percentage of toddlers predicted as "Gizi Kurang " are actually " Gizi Kurang." Recall (Sensitivity) measures the model's ability to find all samples from a class. Recall for " Gizi Kurang " indicates the percentage of toddlers who are actually " Gizi Kurang" that the model correctly predicted.

The F1-Score is the harmonic mean of Precision and Recall, providing a balance between the two metrics. A high F1-Score indicates good model performance in balancing Precision and Recall. Support indicates the actual number of samples in each class in the testing data. This is important because model performance on classes with low support may be less reliable. The Confusion Matrix provides a visual summary of the model's predicted results compared to the actual values. Rows represent actual classes, while columns represent predicted classes. The Classification Report and Confusion Matrix have been generated, as shown in Table 4 and Figure 5.

Table 4 Classification Report

	Precision	recall	f1-score	Support
0	1.00	1.00	1.00	54
1	0.80	0.50	0.62	8
2	0.43	0.75	0.55	4
Accuracy			0.92	66
Macro avg	0.74	0.75	0.72	66
Weighted avg	0.94	0.92	0.93	66

The Balanced Accuracy, which assigns equal importance to each class, was recorded at 0.7499. Furthermore, the Macro F1-Score, representing the average F1-score per class regardless of sample distribution, was 0.72. From the Classification Report, it can be seen that the model has excellent performance in predicting the "Normal" class (Precision, Recall, and F1-Score 1.00). However, performance decreased slightly in the "Gizi Kurang" (Precision 0.80, Recall 0.50, F1-Score 0.62) and "Gizi Buruk" (Precision 0.43, Recall 0.75, F1-Score 0.55) classes. This indicates that the model has more difficulty distinguishing between " Gizi Kurang " and "Gizi Buruk" compared to "Normal".

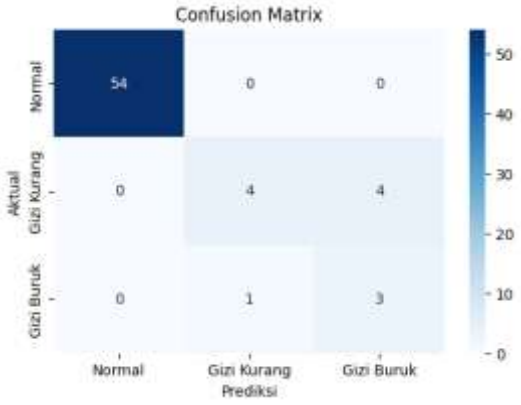

Fig. 5. Confusion Matrix

The main diagonal (from top left to bottom right) shows the number of correct predictions (True Positives) for each class. The model correctly predicted 54 toddlers as “Normal”, 4 toddlers as “Gizi Kurang” and 3 toddlers as “Gizi Buruk”. Values outside the main diagonal indicate prediction errors. The number in the “Gizi Kurang” row, “Normal” column (4) indicates that 4 toddlers who were actually “Gizi Kurang” were incorrectly predicted as “Normal”. The number in the “Gizi Kurang” row, “Normal” column (1) indicates that 1 toddler who was actually “Gizi Kurang” was incorrectly predicted as “Normal”. The number in the “Gizi Kurang” row, “Gizi Kurang” column (0) indicates that no “Gizi Kurang” toddlers were incorrectly predicted as “Gizi Buruk”. The number in the “Gizi Kurang” row, “Gizi Kurang” column (0) indicates that no “Gizi Kurang” toddlers were incorrectly predicted as “Gizi Kurang”.

DISCUSSIONS

In this study, a Random Forest Classifier model was developed to predict the nutritional status of toddlers based on a dataset collected from Puskesmas, Dinas Kesehatan Kota Medan, and Posyandu in Medan City. The data included various factors such as parental occupation and health history, birth weight, child's health history, immunization status, health insurance, body weight and length, income, and housing characteristics. Missing data

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

were handled using imputation, applying the mean for numerical attributes and the mode for categorical variables. The dataset was then stratified and split into training (80 percent) and testing (20 percent) subsets to maintain class distribution, particularly for the target variable. The Random Forest model was configured with 100 estimators and the entropy criterion (RandomForestClassifier(n_estimators=100, criterion='entropy', random_state=0, class_weight= 'balanced')) to optimize information gain during splitting and address class imbalance.

The trained model achieved an overall accuracy of 92.42 percent, which indicates strong global performance. However, a more granular evaluation demonstrated uneven model behavior across classes. The model performed exceptionally well for toddlers categorized as having Normal Nutritional Status, achieving a Precision, Recall, and F1-score of 1.00. Conversely, performance declined on the minority classes: for Malnutrition, the Precision was 0.80, Recall 0.50, and F1-score 0.62, whereas Severely Malnutrition achieved a Precision of 0.43, Recall 0.75, and F1-score 0.55. These findings were reinforced by the Confusion Matrix, which showed that half of actual Malnutrition cases (4 of 8) and one Severely Malnutrition case (1 of 4) were misclassified as Normal. This discrepancy is primarily driven by class imbalance, where the Normal class has substantially higher support (n=54) compared with Malnutrition (n=8) and Severely Malnutrition (n=4). Limited sample representation makes minority class boundaries less distinct, increases sensitivity to noise, and constrains the ability of the classifier to generalize rare case patterns. In addition, the available variables may not fully capture context-sensitive clinical cues relevant to malnutrition severity, which aligns with previous findings that socioeconomic and environmental indicators strongly influence child nutrition outcomes (Gebeye et al., 2023; Laia et al., 2023).

The Feature Importance analysis revealed that Health Insurance Availability and Complete Immunization Status were the most influential predictors of nutritional outcomes. These findings reflect key principles in public health and epidemiology. Access to health insurance facilitates regular healthcare visits, nutritional counseling, and timely medical intervention, thereby reducing infection-related growth disturbances. Similarly, complete immunization protects against recurrent illnesses such as diarrhea and respiratory infections, which impair nutrient absorption and metabolism. This pattern aligns with the World Health Organization's (WHO) life-course framework, which identifies the first 1,000 days of life as a critical window for preventing chronic growth failure through continuous healthcare access, adequate nutrition, and sanitation. Supporting local epidemiological evidence (Syamdarniati, 2024), this study affirms that socioeconomic constraints especially poverty, limited sanitation, and poor access to maternal-child health services remain the strongest determinants of malnutrition among Indonesian toddlers.

Compared with a related study by Akbar Ariyadi et al., which achieved 90.1 percent accuracy in stunting classification using Random Forest, the present work demonstrates slightly higher overall accuracy. However, differences in classification scope and dataset characteristics suggest caution when making direct benchmarks. The error patterns highlight the need for strategies to strengthen minority class representation and discriminative learning(Akbar Ariyadi et al., 2024). The implications of these findings are highly relevant for public health interventions in Medan City. The model's interpretability enables local authorities and Posyandu officers to use risk-prediction tools for early screening, prioritizing targeted outreach to toddlers lacking immunization and health insurance support. Integrating machine learning based alerts into routine nutrition monitoring could reduce late detection of nutritional decline and strengthen resource allocation toward vulnerable groups.

Future research should strengthen the model's robustness by expanding the feature scope with additional socioeconomic, dietary, and environmental indicators, as well as conducting advanced hyperparameter optimization to refine predictive performance. Explicit methodological extensions include the implementation of SHAP interpretability to deepen causal insight into feature contributions, systematic testing of alternative ensemble models such as Gradient Boosting or XGBoost for comparative performance evaluation, and the incorporation of longitudinal child-growth data to enhance temporal predictive validity and support stronger novelty in subsequent model development. Overall, this study demonstrates that the Random Forest model provides high predictive accuracy and actionable insights into nutritional risk factors in toddlers. Despite challenges in identifying minority cases, the results contribute meaningful evidence to support data-driven policies for addressing malnutrition and safeguarding child health in Medan City.

CONCLUSION

This study successfully developed a Random Forest Classifier model for predicting the nutritional status of toddlers using multidimensional data collected from various health facilities in Medan City. The preprocessing steps, including missing-value treatment and feature engineering, enabled the construction of a model that achieved an overall accuracy of 92.42 percent. The classifier performed exceptionally well in identifying toddlers with Normal nutritional status, although its performance remained comparatively lower for the minority classes, namely Underweight and Severely Underweight. The feature importance analysis highlighted the dominant role of health insurance availability and complete immunization, underscoring the strong influence of healthcare access on child nutrition outcomes. Although imbalance-handling strategies were incorporated during model development, the predictive resolution for minority classes indicates opportunities to improve feature diversity and model

*name of corresponding author

sophistication. Future research may enhance model capability by integrating additional socioeconomic, environmental, and dietary indicators, testing alternative ensemble architectures, and implementing interpretability tools such as SHAP to deepen understanding of feature contributions. Overall, the study provides meaningful evidence on the determinants of toddler nutritional status and establishes a data-driven foundation to support more targeted and effective malnutrition interventions.

ACKNOWLEDGMENT

We would like to express our gratitude for the funding of this activity from the DIKTI Grant of the Ministry of Higher Education, Science, and Technology through the Directorate General of Research and Development for the 2025 Fiscal Year. We also express our appreciation to Satya Terra Bhinneka University and the LPPM Team of Satya Terra Bhinneka University for their continued support in the implementation of this grant.

REFERENCES

Akbar Ariyadi, M. R., Lestanti, S., & Kirom, S. (2024). Klasifikasi Balita Stunting Menggunakan Random Forest Classifier Di Kabupaten Blitar. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3846–3851. https://doi.org/10.36040/jati.v7i6.7822

Aprilia, Y. N., Sani, D. A., Anggadimas, N. M., Studi, P., Informatika, T., Informasi, F. T., & Pasuruan, U. M. (n.d.). *Klasifikasi Status Penderita Gizi Stunting Pada Balita Menggunakan Metode Random Forest (Studi Kasus di Kelurahan Petamanan Kota Pasuruan)*. 143–154.

Badan Kebijakan Pembangunan Kesehatan, K. K. R. (2024). *Survei Status Gizi Indonesia (SSGI) 2024.* https://www.badankebijakan.kemkes.go.id/survei-status-gizi-indonesia-ssgi-2024/

Candra, E. N., Cholissodin, I., & Wihandika, R. C. (2022). Klasifikasi Status Gizi Balita Menggunakan Metode Optimasi Random Forest Dengan Algoritme Genetika (Studi Kasus: Puskesmas Cakru). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *6*(5), 2188–2197. http://j-ptiik.ub.ac.id

Cappelli, F., Castronuovo, G., Grimaldi, S., & Telesca, V. (2024). Random Forest and Feature Importance Measures for Discriminating the Most Influential Environmental Factors in Predicting Cardiovascular and Respiratory Diseases. *International Journal of Environmental Research and Public Health*, *21*(7). https://doi.org/10.3390/ijerph21070867

DINKES Kota Medan. (2022). *Profil Kesehatan Tahun 2022 Kota Medan*. 100.

Gebeye, L. G., Dessie, E. Y., & Yimam, J. A. (2023). Predictors of micronutrient deficiency among children aged 6–23 months in Ethiopia: a machine learning approach. *Frontiers in Nutrition*, *10*(January), 1–13. https://doi.org/10.3389/fnut.2023.1277048

Gol, M., Akturk, C., Talan, T., Vural, M. S., & Turkbeyler, I. H. (2024). Predicting malnutrition-based anemia in geriatic patients using machine learning methods. *Journal of Evaluation in Clinical Practice*, 10. https://doi.org/https://doi.org/10.1111/jep.14142

Handayani, P., & Charis Fauzan, A. (2024). KLIK: Kajian Ilmiah Informatika dan Komputer Machine Learning Klasifikasi Status Gizi Balita Menggunakan Algoritma Random Forest. *Media Online)*, *4*(6), 3064–3072. https://doi.org/10.30865/klik.v4i6.1909

Janssen, S. M. W., Bouzembrak, Y., Yalcin, N., & Tekinerdogan, B. (2025). Machine learning models for predicting malnutrition in NICU patients : A comprehensive benchmarking study. *Computers in Biology and Medicine*, *192*(PB), 110326. https://doi.org/10.1016/j.compbiomed.2025.110326

Juwariyem, J., Sriyanto, S., Lestari, S., & Chairani, C. (2024). Prediction of Stunting in Toddlers Using Bagging and Random Forest Algorithms. *Sinkron*, *8*(2), 947–955. https://doi.org/10.33395/sinkron.v8i2.13448

Kaur, S., & Neeru, N. (2024). *Comparative Insights into Machine Learning and Deep Learning Models : Applications and Performance*. *44*(3), 10806–10816.

Khusna, N. F., Rahmah, A., & Nur, R. K. (2024). *Implementasi Random Forest dalam Klasifikasi Kasus Stunting pada Balita dengan Hyperparameter Tuning Grid Search*. *2024*(Senada), 791–801.

Laia, Y., Nasution, Z., & Asriwati. (2023). Analisis faktor risiko kejadian kurang gizi pada balita di Puskesmas Pembantu Tanjung Sari. *Jurnal Kesehatan Dan Fisioterapi*, *3*(1), 27–36.

Mubarok, A. H., Pujiono, P., Setiawan, D., Wicaksono, D. F., & Rimawati, E. (2025). Parameter Testing on Random Forest Algorithm for Stunting Prediction. *Sinkron*, *9*(1), 107–116. https://doi.org/10.33395/sinkron.v9i1.14264

Nirmani, H., & Kudagamage, U. P. (2024). *Ensemble Approach for Early Prediction of Malnutrition Level of Children : A Case study on Children Under Five Years Old*.

Pangumbara'an, M. S. G. et al. (2021). KLASIFIKASI STATUS GIZI BALITA MENGGUNAKAN ALGORITMA K- NEAREST NEIGHBOR DAN NAÏVE BAYES. *Infotech: Journal of Technology Information*, *7*(1), 55–62.

Pratama, J., Fauziah, F., & Sholihati, I. D. (2023). Metode K-Nearest Neighbor Dan Naive Bayes Dalam Menentukan Status Gizi Balita. *Brahmana: Jurnal ...*, *4*(2), 214–221.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

http://tunasbangsa.ac.id/pkm/index.php/brahmana/article/view/197%0Ahttp://tunasbangsa.ac.id/pkm/index.php/brahmana/article/viewFile/197/196

Purnama, J. J. (2020). *Prediksi Child Malnutrition Dengan Algoritma Random Forest*.

Putri Ayu Firnanda, Litasya Shofwatillah, Fauziah Rahma, & Fatkhurokhman Fauzi. (2025). Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket. *Emerging Statistics and Data Science Journal*, *3*(1), 445–461. https://doi.org/10.20885/esds.vol3.iss.1.art2

Qasrawi, R., Sgahir, S., Nemer, M., Halaikah, M., Badrasawi, M., & Amro, M. (2024). *Machine Learning Approach for Predicting the Impact of Food Insecurity on Nutrient Consumption and Malnutrition in*. 1–16.

Setiawan, R., & Triayudi, A. (2022). Klasifikasi Status Gizi Balita Menggunakan Naïve Bayes dan K-Nearest Neighbor Berbasis Web. *Jurnal Media Informatika Budidarma*, *6*(2), 777. https://doi.org/10.30865/mib.v6i2.3566

Syamdarniati. (2024). *BAB I Konsep Dasar Epidemiologi Pada Stunting*.

Utara, B. P. S. (2023). *Profil Kesehatan Provinsi Sumatera Utara*. 1–462. https://dinkes.sumutprov.go.id/unduhan/downloadfile?id=2799

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, *44*(2), 330–349. https://doi.org/10.1016/j.patcog.2010.08.011

WHO. (2015). *Stunting in a nutshell*. https://www.who.int/news/item/19-11-2015-stunting-in-a-nutshell

*name of corresponding author