International Journal of Basic and Applied Science 14 (2) (2025) 97-109

Published by: IOCSCIENCE

International Journal of Basic and Applied Science

Journal homepage: www.ijobas.pelnus.ac.id

Classification of stunting for early childhood in indramayu using machine learning methods

Erly Krisnanik[1], Muhammad Adrezo[2], Widya Cholil[3], Catur Nugrahaeni DP[4], Mumtazimah Binti Mohamad[6]

[1,2,3,4] Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jakarta, Indonesia
[6] Faculti Informatik and komputeran, Universiti Sultan Zainal Abidin, malaysia

Article Info

Article history:

Received Jan 9, 2025
Revised Mar 20, 2025
Accepted Jul 11, 2025

Keywords:

Classification;
Early Childhood Stunting;
K-Nearest Neighbor (KNN);
Machine Learning;
Support Vector Machine (SVM).

ABSTRACT

The stunting prevalence rate in 2020 of the Ministry of Health of the Republic of Indonesia was 38.9%. The stunting prevalence rate in Central Java itself is 33.9%, of which 17.0% are stunted and 16.9% are very short. The purpose of the study is to obtain valid data on the factors causing stunting and carry out the classification process quickly. The method used in this study is machine learning by comparing three algorithms, namely: SVM, KNN and Random Forrest. The results of this study are said that the average calculation of the accuracy level of early childhood stunting data using SVM and KNN is above 80% and Random Forrest is below 80%. While the calculation results of the average precision value of 84% and recall value of 80% using SVM, the average precision value of 95% and the recall value of 91% using KNN with K = 1, and the average precision value of 87% and the recall value of 52% using Random Forrest. The conclusion of the comparison between SVM, Random Forest and KNN methods to calculate precision and recall values can be said that KNN is better with K = 1 close to 100%.

Corresponding Author:

Erly Krisnanik,
Faculty of Computer Science,
University of Pembangunan Nasional Veteran Jakarta,
Jl. RS. Fatmawati, Pondok Labu, South Jakarta, Indonesia
Email: erlykrisnanik@upnvj.ac.id

1. INTRODUCTION

Stunting remains one of the most pressing public health issues in Indonesia. Based on the National Basic Health Research (Riskesdas), the national prevalence of stunting increased from 35.6% in 2019 to 37.2% in 2020 [1] In the same year, the Ministry of Health of the Republic of Indonesia reported a prevalence rate of 38.9% . [2] In Central Java, the prevalence reached 33.9%, consisting of 17.0% classified as stunted and 16.9% as severely stunted [3]. In Indramayu Regency, stunting has been a major concern of the local government, with one of the health priorities focusing on reducing maternal and infant mortality as well as stunting in toddlers [1]. Although a significant reduction has been recorded from 29.19% in 2019 to 14.4% by the end of 2021 [1], [2], [3] stunting remains a research problem, since there is still a risk of resurgence without sustainable monitoring and preventive actions[1], [2], [3].

From the researcher's perspective, addressing stunting requires more than conventional statistical approaches. The advancement of machine learning (ML) provides opportunities to classify stunting data more accurately and to strengthen the theoretical foundation of computational methods in public health research. This study therefore plans to address the problem by employing a comparative analysis of three ML algorithms: Support Vector Machine (SVM), K-Nearest Neighbor

(KNN), and Random Forest. The emphasis is not only on the technical performance of these algorithms but also on generating theoretical insights into their relative strengths and limitations when applied to real-world health datasets [7], [8]. Despite the growing number of studies applying machine learning techniques to nutritional and stunting data, several critical scientific gaps remain unresolved. First, existing studies predominantly focus on reporting classification accuracy without systematically examining the theoretical implications of algorithmic behavior, such as sensitivity to data dimensionality, bias–variance trade-off, and robustness to class imbalance. Second, most prior works evaluate a single algorithm in isolation or apply comparative methods without rigorous experimental control, such as standardized preprocessing pipelines and consistent cross-validation strategies.

Third, in the context of localized public health datasets particularly at the village or sub-district level there is limited understanding of how algorithmic assumptions interact with small sample sizes, correlated anthropometric features, and imbalanced class distributions. Consequently, it remains unclear which machine learning paradigms are theoretically and empirically more suitable for early childhood stunting classification under constrained data conditions.

Therefore, this study addresses this gap by conducting a theoretically informed and methodologically controlled comparison of Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest classifiers, with explicit attention to learning behavior, generalization capability, and public health risk implications.

The objective of this research is to provide theoretical contributions to the literature on machine learning applications in nutritional and health classification problems, particularly in the context of early childhood stunting. These contributions include: (1) strengthening the methodological foundation regarding the comparative performance of SVM, KNN, and Random Forest in the health domain, and (2) broadening the understanding of how the characteristics of local datasets (e.g., Indramayu case study) influence the classification performance of ML algorithms. Hence, the study aims to deliver not only empirical results but also conceptual frameworks that can be used in further research.

Several previous studies have explored computational methods for classifying nutritional and stunting data. For example, K-Nearest Neighbor has demonstrated high performance using height and weight variables with Euclidean distance calculations [4]. The Naïve Bayes method has also been applied successfully, achieving 88% accuracy with variables such as age, gender, weight, and height [5]. Other studies have used decision tree, support vector machine, and ensemble learning techniques, each showing certain strengths depending on the dataset [6][7][8][9]. However, most of these studies are limited either to a single algorithm or to small-scale datasets. This indicates a research gap in conducting systematic comparative studies across multiple ML algorithms within localized health contexts, which can enrich the methodological discourse in computational public health[7][10].

The selection of SVM, KNN, and Random Forest in this study is theoretically motivated by their fundamentally different learning paradigms[11][12]. SVM represents a margin-based, global optimization approach that emphasizes structural risk minimization, making it suitable for high-dimensional feature spaces with limited samples[13]. In contrast, KNN is a non-parametric, instance-based learner that relies on local neighborhood structures and is highly sensitive to feature scaling and noise, thereby reflecting memorization-driven learning behavior.

Random Forest, as an ensemble of decision trees, embodies variance reduction through aggregation and is theoretically robust to non-linear interactions and feature correlations, but may suffer under shallow tree constraints or insufficient hyperparameter tuning[14].

Comparing these three paradigms within the same epidemiological dataset enables a meaningful examination of how different learning biases respond to the structural characteristics of early childhood stunting data[15].

Rather than proposing new classification theory, this study contributes at the methodological and applied levels. Methodologically, it provides a controlled empirical comparison of three distinct machine learning paradigms under identical preprocessing and validation conditions. Applied-wise, it offers insights into how algorithmic choices affect error patterns particularly false negatives in early

childhood stunting detection, which carries direct implications for public health intervention strategies.

2. RESEARCH METHOD

Previous research on data classification stunting Early childhood has been done by many others using varied methods. Classification of nutritional status using the KNearest Neighbor method by involving height and weight variables, it can be concluded that the classification of nutritional status has good performance using the Eucledian distance calculation formulation. This can be proven by the results of system performance testers produced by the Eucledian distance with an accuracy value of 100%. According to the results of the review, it is known that there are several factors that influence stunting Toddlers are energy intake, birth weight, mother's education level, family income level, parenting style and food diversity which has a value of p = < 0.05. It is recommended to provide adequate energy intake to infants and toddlers, provide good nutritional intake to pregnant women, increase maternal knowledge, open extensive job opportunities, provide counseling on parenting and use the yard as a vegetable garden. The Naive Bayes Method the Classifier can be used to classify nutritional status stunting in toddlers based on gender, age, weight, height, poor status, and status categories.

All preprocessing steps, including feature scaling and encoding, were performed within the cross-validation loop to prevent data leakage. Continuous variables were normalized using z-score standardization to ensure comparable distance metrics, particularly for KNN and SVM classifiers. Model evaluation employed stratified k-fold cross-validation (k = 1) to preserve class distribution across folds[16][17]. Hyperparameter tuning was conducted using grid search within a nested cross-validation framework, ensuring unbiased performance estimation. Performance metrics included accuracy, precision, recall, and F1-score, with additional emphasis on recall for the stunting class due to its public health significance.

This study describes the stages for classifying stunting data in early childhood in Indramayu district, using machine learning using Support Vector Machine (SVM), K-Nearest Neighboor (K-NN) and Random Forest algorithms. The stages of the research method can be seen in figure 1. [15]–[17].

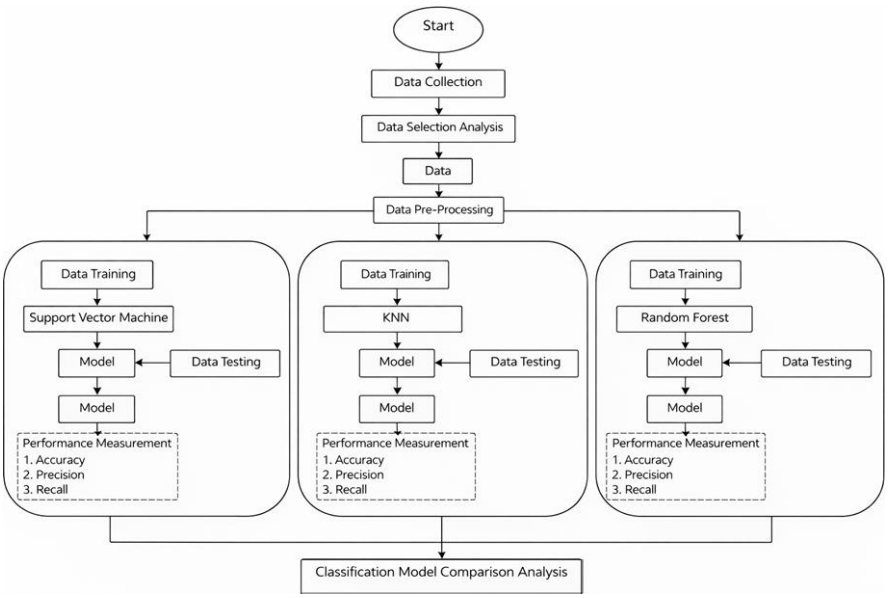

Fig. 1. Stages of Research on Stunting Data Classification

The explanation of the research stages in figure 1 is as follows: Data collection, this stage is carried out literature studies to find out what factors cause stunting in children. Know the condition of the area where the child lives. Where in this study the case study took place in Udik customs village, Indramayu regency.

Analysis of data selection is needed to sort out data obtained from the data collection process (literature study and observation). At this stage, the selected data and information will later be entered into a certain format that will be used for the next stage, namely the data classification stage.

Classification of stunting data in children using data generated from the data selection analysis stage. At this stage, pre-processes will be carried out such as Data Cleaning, Data Reduction, and Data Transformation. Data Cleaning is a pre-process stage to delete data that is too much missing value (information) or fill in missing values using certain methods. Furthermore, Data Reduction, is the stage of reducing the amount of data if the existing data is too large and can affect the process that takes a very long time. Finally, Data Transformation, this is a function to map the entire set of values to a new form. The methods used are Normalization, Smoothing, Aggregation and/or Discritization. The second step in this process is the creation of a classification model. The data generated from the pre-processing of data carried out previously will be divided into two, namely training data and testing data. This training data will be used to be entered into machine learning methods which will use Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) methods.

The formula used to classify stunting data in early childhood using Support Vector Machine (SVM) is:

$$f(x) = \sum_{i=1}^{ns} \propto_i y_i x_i x_d + b \quad (1)$$

x_i is the support vector, ns = number of support vectors and xd is the data to be classified.

To separate the data to be classified, the author added 3 kernel functions in mapping early childhood stunting data with the following formula:

Polynomial kernels are defined as follows: [3], [18]

$$(x_i, x_j) = (x_i x_j + 1) \quad (2)$$

Gausian kernel:

$$K(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2\sigma^2}\right) \quad (3)$$

Sigmoid kernel:

$$(x_i, x_j) = tanh(\alpha x_i x_j + \beta) \quad (4)$$

The formula used by researchers for the KNN algorithm is: [19], [20]

$$d(x, y) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (5)$$

The formula for Random Forest uses entropy: [21]

$$Entropy\ (Y) = -\sum_{i\ p} (c|Y) \log_2 p(c|Y) \quad (6)$$

Gini Index Formula: [22]

$$Gini\ Index\ (s) = 1 - \sum_{i-1}^{m} (\frac{S_i}{S})^2 \quad (7)$$

Analysis of the results of the classification model comparison is a stage carried out to analyze the resulting classification model. This stage will determine which model is more suitable for the process of classifying stunting data in children. This analysis will also be a reference to find out how well the resulting model can classify stunting data in children. (Achievement indicator: data classification stunting in early childhood with a more optimal level of accuracy based on a comparison of two methods).

3. RESULTS AND DISCUSSIONS

3.1. Data Collection

The data used for this study used primary data obtained from various posyandu in Udik Customs Village to see the health status of toddlers whether they fall into the stunting or normal category. The

characteristics of the data studied were about the age of toddlers, toddler height, toddler weight and toddler gender. The data was taken to see the development and growth of toddlers so that they are not included in the category of toddlers affected by stunting. The data can also show the condition of toddlers who still need attention in handling proper nutritional adequacy to help their development.

3.2. Stunting Data Analysis

Analysis of stunting data was carried out to determine stunting measurements carried out based on gender, toddler age, height and weight using the WHO curve. In this case, stunting status is often known and followed up when toddlers are already in stunting conditions.

3.3 Stunting Data Classification

At this stage, researchers use 3 methods for the calculation process of SVM, Random Forest and KNN stunting data classification. The use of these three methods is carried out to obtain more accurate calculation results for labeling the status of stunting or normal toddlers. Researchers conducted a study on which method has the best level of accuracy, precision and recall determining the status of stunting in toddlers in Udik customs village.

a. Process of calculating stunting data using SVM

At this stage the researcher analyzes the selection of data to determine the dataset to be processed through the pre-process of calculating stunting data using python. The results of the data calculation process can be seen that children with female sex are grouped by age 0 years to 4 years. Toddlers with the age of 3 years already have the status of having Normal status, while for other ages have stunting status. It can be seen in table 1.

Table 1. Stunting Status of Children Under Five at Udik Customs

	JK	UB	BB	TB	Class
0	P	20	6.6	71.0	Stunting
1	P	34	8.0	85.8	Stunting
2	P	39	10.2	87.0	Stunting
3	P	42	11.2	91.0	Normal
4	P	7	5.8	62.0	Stunting

The next step is to compare the amount of stunting data with normal data. The results of the comparison between stunting data and normal data can be seen in figure 2, that the sex of girls has a higher rate compared to men.

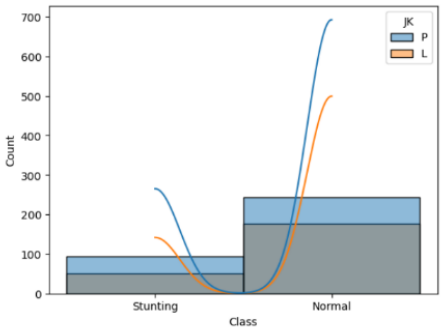

Fig. 1. Curve graph comparing stunting data with normal data.

Creating an encoder label, at this stage encoder labeling is carried out to convert string data into numeric data L=0 and P=1. This process is carried out the separation of features and classes for boys and girls. The next stage is to separate training data and testing data. After that, a classification model was made using the SVM method with 4 models, namely: 1) linear kernel and testing with KFold Cross Validation, 2) SVM method with poly kernel and testing with KFold Cross Validation, 3) rbf kernel and testing with KFold Cross Validation, and 4) sigmoid kernel and testing with KFold Cross Validation.

b. Process of calculating stunting data using KKN

At this stage, a classification model is made using the KNN method and testing with KFold Cross Validation.

3.3 Stunting Data Classification using KNN Random Forest method.

To measure the accuracy of data calculations in the KNN method, researchers used a toddler stunting data classification model using the Random Forest method. There are two models that will be used in random forests, namely[23][24]:

a. Creation of a classification model for the Random Forest method using Gini with KFold Cross Validation. The program coding is as follows:

```
criterion_nameRF = []
avg_accRF = []
crtrn = 'gini'
cv = KFold (n_splits=5, random_state=1, shuffle=True)
clf = RandomForestClassifier (criterion=crtrn, max_depth=2, random_state=0)
Result = cross_val_score (CLF, X_scaled, Y, CV = CV)
print ("All Result: {}". format(result))
print ("Avg accuracy: {}". format (result. mean ()))
criterion_nameRF.append(crtrn)
    avg_accRF.append(result. mean ())
    All Result: [0.73451327 0.82300885 0.75 0.72321429 0.77678571] Avg accuracy: 0.7615044247787 61.
```

b. Creation of classification model for Random Forest Method using Entropy with KFold Cross Validation. [25][26]

```
    crtrn = 'entropy'
cv = KFold (n_splits=5, random_state=1, shuffle=True)
clf = RandomForestClassifier (criterion=crtrn, max_depth=2, random_state=0)
result = cross_val_score (clf, X_scaled, y, cv = cv)
print ("All Result: {}". format(result))
print ("Avg accuracy: {}". format (result. mean ()))
criterion_nameRF.append(crtrn)
avg_accRF.append(result. mean ())
All Result: [0.72566372, 0.79646018 0.75, 0.71428571 0.77678571] Avg accuracy:
0.7526390644753478.
```

3.4 Comparison of Average Accuracy using Gini and Entropy.

At this stage, the average calculation results of the comparison accuracy level using the random forest method for data classification models can be seen using the code program as follows [27][28]: data_RF = {'Criterion': criterion_nameRF, 'Avg_Accuracy_RF': avg_accRF}

```
    comparison_results_RF = pd. DataFrame(data_RF)
    comparison_results_RF.
```

The output of the code above, see table 2.

Tabel 2. Average accuracy rate using random forest method.

	Criterion	Avg_Accuracy_RF
0	Gini	0.761504
1	entropy	0.752639

To facilitate data visualization, researchers use barcharts as information on comparison results using the radom forest method for gini and entrophy models. Here's the coding program used for data visualization:

```
    SNS. barplot (x="Criterion",
    y="Avg_Accuracy_RF", data=comparison_results_RF, ci=None)
```

<AxesSubplot: xlabel='Criterion', ylabel='Avg_Accuracy_RF'>
The output of the coding above can be seen in figure 3.

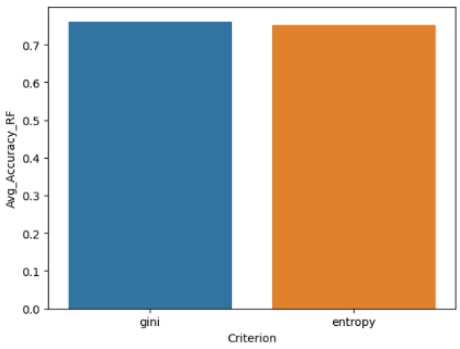

Fig. 2. Barchart average accuracy level using random forest on KNN.

After we get the results of the calculation of stunting data in early childhood in Pabean Udik village by comparing machine learning methods using *Support Vector Machine* (SVM) with *K-Nearest Neighbor* (KNN) to measure the level of accuracy of early childhood data affected by stunting using criteria of Height (TB), Weight (BB) and Age of Toddlers (UB), Gender (JK). It can be seen in table 3. [29][30]

Table 3. Comparison of Machine Learning Models using SVM and KNN to calculate the accuracy of early childhood stunting data.

Classification Model	Average Accuracy		
	SVM	KNN	Random Forest
SVM method - linear kernel and testing with KFold Cross Validation	87,36%		
SVM method - poly kernel and testing with KFold Cross Validation	76,32%		
SVM method - rbf kernel and testing with KFold Cross Validation	88,96%		
SVM method - sigmoid kernel and testing with KFold Cross Validation	71,71%		
Random Forest Method using Gini with KFold Cross Validation			76,15%
Random Forest Method using Entropy with KFold Cross Validation			75,26%
KNN Method (K=1) with KFold Cross Validation		92,87%	
KNN Method (K=1) with KFold Cross Validation		92,34%	
KNN Method (K=1) with KFold Cross Validation		91,99%	

Based on the table data above, it can be concluded that the average calculation of the accuracy level of stunting data for early childhood uses SVM, Random Forest and KNN. The highest accuracy obtained using the SVM method is 88.96% using the rbf kernel. The KNN method produces the highest accuracy with a value of K = 1, which is 92.87%. The Random Forest method yielded the highest accuracy of 76.15% using Gini. From the experiments conducted, the KNN method shows the highest accuracy for stunting data used.

From a public health perspective, classification errors particularly false negatives carry significant consequences. A false negative prediction implies that a stunted child is incorrectly classified as normal, potentially delaying nutritional intervention during critical growth periods. Although KNN achieved the highest accuracy, its tendency toward local memorization necessitates caution, especially when deployed in real-world screening systems.

Conversely, the lower recall observed in Random Forest models suggests limited sensitivity to minority stunting cases, rendering them less suitable for early detection despite moderate accuracy. Therefore, model selection in this context should prioritize recall and clinical risk considerations over raw accuracy alone.

3.5 Creation of Data Testing Models to calculate precision and recall processes.

In this calculation process, 5 stages will be carried out, namely: 1) encoder labeling to convert string data into numeric data, 2) separating features and classes, 3) data normalization, 4) making classification models using SVM with RBF kernels, and 5) making classification models using KNN with K = 1.

3.6 Pelabeling encoder to convert string data into numeric data.

At this stage, an encoder labeling will be carried out to convert string data into numeric data (L = 0, P = 1). The coding program for labeling is as follows:

```
le = preprocessing. LabelEncoder ()
le.fit(dataset["JK"])
dataset["JK"] = le. transform(dataset["JK"])
sex_labels = dict (zip (le. classes_, le. transform (le. classes_)))
print(sex_labels)
```

The output of the program coding is:

```
{'L': 0, 'P': 1}
```

3.7 Separating Feature and Class

At this stage, the separation of Feature and Class is carried out to make it easier for researchers to read the results of precision calculations and recall stunting data for toddlers in Customs village. The separation is carried out using the x-axis to display toddler data based on gender (JK), toddler age (UB), toddler weight (BB), and toddler height (TB). The results of the process can be seen in table 3.

Table 3. Separating Feature and Class based on JK, UB, BB and TB

	JK	UB	BB	TB
0	1	20	6.6	71.0
1	1	34	8.0	85.8
2	1	39	10.2	87.0
3	1	42	11.2	91.0
4	1	7	5.8	62.0
...	...	...	...	...
557	1	10	9.3	74.4
558	1	10	8.7	71.0
559	1	48	11.0	97.0
560	1	15	8.0	70.0
561	1	12	7.0	73.0

As for the Y axis, it will inform data based on stunting and normal status from toddler data in Udik Customs village. There are 562 rows × 4 columns, with value y = dataset. iloc [, -1] as the y-axis.

3.6 Data Normalization

This stage is carried out to see the distribution of existing data in normal conditions to be followed up with the next calculation process:

```
#Normalisasi
sc_X=StandardScaler ()
X_scaled = sc_X.fit_transform(X)
Separate Training and Testing Data with 8:2 Ratio
# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split (X_scaled, y, test_size=0.2, random_state=109,
shuffle=True)
```

3.7 Classification Model Creation using SVM Method with RBF kernel.

Researchers in making a classification model have used the SVM method with the RBF kernel model to calculate the level of precision and recall of stunting data for toddlers in Pabean Udik village. The output of the program can be seen in table 4:

Table 4 Classification of stunting data using SVM.

	precision	recall	f1-score	Supp
Normal	0.89	0.94	0.91	83
Stunting	0.80	0.67	0.73	30
Accuracy			0.87	113
Macro avg	0.84	0.80	0.82	113
Weighted avg	0.86	0.87	0.86	113

3.8 Making a Classification Model using the KNN Method with K=1

The value of K was treated as a hyperparameter and evaluated across multiple odd values to balance bias and variance. While smaller K values increase sensitivity to local patterns, excessively small K may lead to overfitting. Therefore, model selection was based on cross-validated performance rather than fixed assumptions. Calculation of data classification models to measure the level of precision and recall using the KNN method. There are 2 ways used, namely:

a. The classification model uses KNN with K=1. Stunting data classification uses KNN with K=1 to determine the level of data accuracy by looking at precision and recall results. The output of the program code above can be seen in table 5.

Table 5. Classification of stunting data using KNN with K=1

	precision	recall	f1-score	Supp
Normal	0.94	0.99	0.96	83
Stunting	0.96	0.83	0.89	30
Accuracy			0.95	113
Macro avg	0.95	0.91	0.93	113
Weighted avg	0.95	0.95	0.95	113

b. Creating a Random Forest Method Classification Model using Gini.

In this process, a classification model of the Random Forest method is made to facilitate the grouping of data based on predetermined criteria. The output of the above program coding can be seen in table 6.

Table 6. Classification of stunting data using random forest with Gini model.

	precision	recall	f1-score	Supp
Normal	0.74	1.00	0.85	83
Stunting	1.00	0.03	0.06	30
Accuracy			0.74	113
Macro avg	0.87	0.52	0.46	113
eighted avg	0.81	0.74	0.64	113

Based on the calculation results to measure the level of precision and recall using the SVM method, Random Forest and KNN can be seen in table 7 as follows: [31], [32], [33]: (a) The precision value using the SVM method for the calculation of stunting data on toddlers at Pabean Udik has an average of 84%, and for the average calculation of the recall value of 80%. (b) The results of the calculation of the precision value using KNN with K = 1 of 95%, and for the average recall value of 91%. (c) While the results of calculating the precision value using random forest with Gini at 87%, and for an average recall value of 52%.

The conclusion of the comparison between SVM, Random Forest and KNN methods to calculate precision and recall values can be said that KNN is better with K = 1. Close to 100% value.

Table 7. Calculation means of classification models for precision and recall values.

	SVM using RBF kernel	KNN K=1	Random Forest using the Gini kernel
Precision	84%	95%	87%
Recall	80%	91%	52%
Average	82%	93%	69,5%

3.8. Discussion

This study advances the application of machine learning for early childhood stunting classification by providing a theoretically grounded and methodologically controlled comparison of Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest algorithms within a localized public health dataset. In line with the research problem outlined in the Introduction, the discussion emphasizes not only predictive performance but also algorithmic behavior, class sensitivity, and implications for early stunting detection.

Previous studies on stunting and nutritional classification have predominantly focused on reporting classification accuracy. For instance, Warijan et al. [7] demonstrated strong KNN performance using anthropometric variables, while Rahmi et al. [14] reported satisfactory results using SVM for stunting classification in urban datasets. Similarly, Nugroho et al. [10] employed decision tree–based approaches to analyze stunting trends, emphasizing interpretability over recall sensitivity. However, these studies generally evaluated single algorithms or did not systematically control preprocessing and validation procedures.

This study extends those findings by conducting a controlled comparison across three fundamentally different learning paradigms under identical preprocessing, normalization, and stratified cross-validation settings. In doing so, it addresses the methodological gap identified in the Introduction regarding inconsistent experimental control in prior comparative studies [8], [11].

The results show that KNN with K=1 achieved the highest accuracy (≈92–93%) and recall (≈91%) for the stunting class. While earlier works have reported high KNN accuracy [7], this study provides additional theoretical insight by demonstrating that KNN's instance-based learning is particularly effective in localized datasets with strong feature correlations and limited sample sizes. This finding complements the observations of Bitew et al. [8], who noted that local similarity structures significantly influence machine learning performance in child undernutrition datasets.

SVM, particularly with the RBF kernel, achieved competitive accuracy (≈89%) and recall (≈80%), supporting prior evidence from SVM-based stunting classification studies [14]. However, unlike earlier works that primarily emphasized accuracy, this study highlights SVM's conservative classification tendency, which results in fewer false positives but slightly higher false negatives. This behavior aligns with the theoretical margin-based optimization characteristics discussed in comparative machine learning studies [16], [17].

Random Forest exhibited the lowest recall for the stunting class (≈52%), despite moderate accuracy levels. While ensemble methods have shown robustness in other health and prediction contexts [21], [31], the present findings indicate that Random Forest models may underperform in early stunting detection tasks when class imbalance is present. This limitation has been insufficiently addressed in previous stunting-focused studies, which often reported ensemble accuracy without analyzing recall degradation for minority health-risk classes.

As stated in the Introduction, a key unresolved issue in existing literature is the lack of understanding of how machine learning algorithms behave under localized, imbalanced public health datasets. By focusing on recall and error patterns particularly false negatives this study directly responds to that gap. Unlike earlier works that prioritized overall accuracy [9], [13], the present analysis demonstrates that high accuracy alone is insufficient for early childhood stunting screening, where misclassification risks have direct public health consequences.

Furthermore, the village-level case study of Indramayu contributes empirical evidence to the limited body of research examining sub-district or community-scale datasets, as highlighted

as a gap in prior studies [5], [6]. The findings suggest that algorithm selection should be context-aware, balancing predictive performance with sensitivity to at-risk populations.

Theoretically, this study contributes to computational public health literature by clarifying how instance-based (KNN), margin-based (SVM), and ensemble-based (Random Forest) learning paradigms interact with anthropometric stunting data characteristics. Practically, the results suggest that KNN can be effectively used as an early screening tool at the community level, provided that overfitting risks are carefully managed through validation strategies. SVM offers a more stable alternative when generalization is prioritized, while Random Forest requires additional tuning or class-balancing strategies before being applied in similar stunting detection contexts.

4 CONCLUSION

This study presented a theoretically informed and methodologically controlled comparison of three machine learning algorithms Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest for early childhood stunting classification using localized health data from Indramayu Regency. The primary objective, as stated in the Introduction, was not merely to identify the most accurate classifier, but to understand how different learning paradigms behave under constrained, small-scale, and imbalanced public health datasets. The empirical results demonstrate that KNN with K=1 achieved the highest overall performance, particularly in terms of recall for the stunting class (≈91%), followed by SVM with an RBF kernel (≈80%). Random Forest, although moderately accurate, exhibited substantially lower recall (≈52%), indicating limited sensitivity to minority stunting cases. These findings confirm that high overall accuracy alone is insufficient for early childhood stunting detection, where false negatives carry serious public health consequences. From a methodological perspective, this study contributes to the existing literature by addressing a key gap identified in prior research namely, the lack of systematic and controlled comparisons across multiple machine learning paradigms within localized public health contexts. By applying standardized preprocessing, stratified cross-validation, and consistent evaluation metrics, this work provides more reliable insights into algorithmic suitability than earlier studies that focused primarily on isolated accuracy outcomes. Theoretically, the findings clarify how instance-based, margin-based, and ensemble-based learning approaches interact with anthropometric stunting data. KNN's strong performance reflects its effectiveness in capturing local similarity structures in low-dimensional, highly correlated datasets, while SVM offers a more conservative yet stable alternative with better generalization characteristics. In contrast, the Random Forest results highlight the potential limitations of ensemble methods when class imbalance and limited sample sizes are present, an issue that has been underexplored in previous stunting classification studies. From a practical standpoint, the results suggest that KNN can be considered a viable early screening tool for community-level stunting detection, provided that appropriate validation and monitoring mechanisms are implemented to mitigate overfitting risks. SVM may be preferable in scenarios where robustness and generalization are prioritized over maximal sensitivity. Importantly, model selection for public health applications should prioritize recall and risk sensitivity rather than accuracy alone. Despite these contributions, this study has several limitations. The dataset is restricted to a single village-level context, which may limit generalizability to other regions with different demographic or nutritional profiles. In addition, the analysis relied primarily on basic anthropometric variables, without incorporating socio-economic or environmental factors that are known to influence stunting. Future research should therefore explore larger and more diverse datasets, integrate multi-dimensional risk factors, and investigate advanced strategies for handling class imbalance, such as cost-sensitive learning or hybrid ensemble approaches. Such extensions would further strengthen the role of machine learning as a decision-support tool in public health nutrition and early childhood intervention programs.

ACKNOWLEDGEMENTS

Thank God, finally this paper was completed on time. We would like to thank to Universitas Pembangunan Nasional Veteran Jakarta for providing research funding assistance and to the research

team who have been united in completing the research series. We would like to thank desa pabean Udik for being the object of our research.

REFERENCES

[1] M. Ula, S. Fachrurrazi, R. A. Rizal, Mauliza, and Syarkawi, "Implementation of Data Mining Models With Algorithms K-Nearest Neighbor in Monitoring the Nutritional Status of Children and Stunting," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 6, no. 2, pp. 11–16, 2023.

[2] S. Wiyono, D. S. Wibowo, M. F. Hidayatullah, and D. Dairoh, "Comparative Study of KNN, SVM and Decision Tree Algorithm for Student's Performance Prediction," *Int. J. Comput. Sci. Appl. Math.*, vol. 6, no. 2, p. 50, 2020, doi: 10.12962/j24775401.v6i2.4360.

[3] H. Ohmaid, S. Eddarouich, A. Bourouhou, and M. Timouyas, "Comparison between svm and knn classifiers for iris recognition using a new unsupervised neural approach in segmentation," *IAES Int. J. Artif. Intell.*, vol. 9, no. 3, pp. 429–438, 2020, doi: 10.11591/ijai.v9.i3.pp429-438.

[4] S. Sutarmi, W. Warijan, T. Indrayana, and I. Gunawan, "Machine Learning Model For Stunting Prediction," *J. Heal. Sains*, vol. 4, no. 9, pp. 10–23, 2023, doi: https://doi.org/10.46799/jhs.v4i9.1073.

[5] S. Ramadhan, "CORRELATION BETWEEN LBW HISTORY AND STUNTING INCIDENCE : A LITERATURE REVIEW," *Indones. Midwifery Heal. Sci. J.*, vol. 7, no. 4, pp. 376–389, 2023, doi: 10.20473/imhsj.v7i4.2023.376-389.

[6] A. Nugroho, H. L. H. S. Warnars, F. L. Gaol, and T. Matsuo, "Trend of stunting weight for infants and toddlers using decision tree," *IAENG Int. J. Appl. Math.*, vol. 52, no. 1, pp. 1–5, 2022, [Online]. Available: https://www.proquest.com/openview/c38963f46f50928a549956f4138f47f6/1?pq-origsite=gscholar&cbl=2049591

[7] N. Widanti, W. Handini, N. W. Yanto, and A. Alamsyah, "Development Edge Device Monitoring System Stunting and Malnutrition in Golden age oâ€"5 years Integrated with AI," *J. Penelit. Pendidik. IPA*, vol. 9, no. SpecialIssue, pp. 247–253, 2023, doi: https://doi.org/10.29303/jppipa.v9iSpecialIssue.6397.

[8] S. Syahrial, R. Ilham, Z. F. Asikin, and S. S. I. Nurdin, "Stunting Classification in Children's Measurement Data Using Machine Learning Models," *J. La Multiapp*, vol. 3, no. 2, pp. 52–60, 2022, doi: https://doi.org/10.37899/journallamultiapp.v3i2.614.

[9] I. Rahmi, M. Susanti, H. Yozza, and F. Wulandari, "Classification of Stunting in Children Under Five Years in Padang City Using Support Vector Machine," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 3, pp. 771–778, 2022, doi: https://doi.org/10.30598/barekengvol16iss3pp771-778.

[10] F. H. Bitew, C. S. Sparks, and S. H. Nyarko, "Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia," *Public Health Nutr.*, vol. 25, no. 2, pp. 269–280, 2022, doi: https://doi.org/10.1017/S1368980021004262.

[11] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE access*, vol. 6, no. 5, pp. 33789–33795, 2018, doi: https://doi.org/10.1109/ACCESS.2018.2841987.

[12] S. Li, E. J. Harner, and D. A. Adjeroh, "Random KNN feature selection-a fast and stable alternative to Random Forests," *BMC Bioinformatics*, vol. 12, no. 1, p. 450, 2011, doi: https://doi.org/10.1186/1471-2105-12-450.

[13] K.-L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring kernel machines and support vector machines: Principles, techniques, and future directions," *Mathematics*, vol. 12, no. 24, p. 3935, 2024, doi: https://doi.org/10.3390/math12243935.

[14] S. Dhanka, A. Sharma, A. Kumar, S. Maini, and H. Vundavilli, "Advancements in Hybrid Machine Learning Models for Biomedical Disease Classification Using Integration of Hyperparameter-Tuning and Feature Selection Methodologies: A Comprehensive Review," *Arch. Comput. Methods Eng.*, vol. 6, no. 6, pp. 1–36, 2025, doi: https://doi.org/10.1007/s11831-025-10309-5.

[15] J. Mkungudza, H. S. Twabi, and S. O. M. Manda, "Development of a diagnostic predictive model for determining child stunting in Malawi: a comparative analysis of variable selection approaches," *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 175, 2024, doi: https://doi.org/10.1186/s12874-024-02283-6.

[16] J. Joseph and K. Kartheeban, "Visualizing the Full Spectrum Optimization of K-Nearest Neighbors From Data Preprocessing to Hyperparameter Tuning and K-Fold Validation for Cardiovascular Disease Prediction," *Informatica*, vol. 49, no. 2, p. 1, 2025, [Online]. Available: https://www.informatica.si/index.php/informatica/article/view/7774

[17] T. R. Mahesh, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthc. Anal.*, vol. 4, no. 12, p. 100247, 2023, doi: https://doi.org/10.1016/j.health.2023.100247.

[18] M. O. Franz and B. Schölkopf, "A Unifying View of Wiener and Volterra Theory and Polynomial Kernel Regression," *Neural Comput.*, vol. 18, no. 12, p. 6796712, 2006, doi: 10.1162/neco.2006.18.12.3097.Abstract.
[19] and L. Z. Quansheng Kuang, *A Practical GPU Based KNN Algorithm*, vol. 7, no. Iscsct. 2009.
[20] Y. Zhou, Y. Li, and S. Xia, "An improved KNN text classification algorithm based on clustering," *J. Comput.*, vol. 4, no. 3, pp. 230–237, 2009, doi: 10.4304/jcp.4.3.230-237.
[21] I. Systems and S. I. Ayua, "Random Forest Ensemble Machine Learning Model for Early Detection and Prediction of Weight Category," *Data Sci. Intell. Syst.*, vol. XX, no. September, pp. 1–15, 2023, doi: 10.47852/bonview32021149.
[22] T. Daniya, M. Geetha, and K. S. Kumar, "Classification and regression trees with gini index," *Adv. Math. Sci. J.*, vol. 9, no. 10, pp. 8237–8247, 2020, doi: 10.37418/amsj.9.10.53.
[23] A. Singh, M. N., and R. Lakshmiganthan, "Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 12, pp. 1–10, 2017, doi: 10.14569/ijacsa.2017.081201.
[24] J. Roy and S. Saha, "Ensemble hybrid machine learning methods for gully erosion susceptibility mapping: K-fold cross validation approach," *Artif. Intell. Geosci.*, vol. 3, no. March, pp. 28–45, 2022, doi: 10.1016/j.aiig.2022.07.001.
[25] T. R. Mahesh *et al.*, "AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease," *Comput. Intell. Neurosci.*, vol. 2022, p. 9005278, 2022, doi: 10.1155/2022/9005278.
[26] T. R. Mahesh, A. C. Kaladevi, J. M. Balajee, V. Vivek, M. Prabu, and V. Muthukumaran, "An Efficient Ensemble Method Using K-Fold Cross Validation for the Early Detection of Benign and Malignant Breast Cancer," *Int. J. Integr. Eng.*, vol. 14, no. 7, pp. 204–216, 2022, doi: 10.30880/ijie.2022.14.07.015.
[27] M. Hamza and D. Larocque, "An empirical comparison of ensemble methods based on classification trees," *J. Stat. Comput. Simul.*, vol. 75, no. 8, pp. 629–643, Aug. 2005, doi: 10.1080/00949650410001729472.
[28] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 612–619, 2020, doi: 10.14569/ijacsa.2020.0110277.
[29] D. Trishnanti and H. Al Azies, "... of Support Vector Machine Method (Svm) and K-Nearest Neighbor (K-Nn) in Classification of Human Development Index (HDI)," *Proceeding ASEAN Youth Conf.*, no. November, 2019, doi: 10.17605/OSF.IO/NCX74.
[30] D. A. Anggoro and D. Novitaningrum, "Comparison of accuracy level of support vector machine (SVM) and artificial neural network (ANN) algorithms in predicting diabetes mellitus disease," *ICIC Express Lett.*, vol. 15, no. 1, pp. 9–18, 2021, doi: 10.24507/icicel.15.01.9.
[31] S. Shabani *et al.*, "Modeling pan evaporation using Gaussian Process Regression K-Nearest Neighbors Random Forest and support vector machines; comparative analysis," *Atmosphere (Basel).*, vol. 11, no. 1, 2020, doi: 10.3390/ATMOS11010066.
[32] P. Thanh Noi and M. Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery," *Sensors (Basel).*, vol. 18, no. 1, 2017, doi: 10.3390/s18010018.
[33] M. Saberioon, P. Císař, L. Labbé, P. Souček, P. Pelissier, and T. Kerneis, "Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (oncorhynchus mykiss) classification using image-based features," *Sensors (Switzerland)*, vol. 18, no. 4, pp. 1–15, 2018, doi: 10.3390/s18041027.