Vol. 4 No. (1) (June 2026) 19-26

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

K-Nearest Neighbors Algorithm in Classification of Stunting Detection Dataset

Lea Angelina[1], Saeful Amri[2], M. Al Haris[3], Rochdi Wasono[4], Erna Julia Nanga[5], Faninda Aidina Fitri[6]

[1,2,3,4,5,6]Universitas Muhammadiyah Semarang, Indonesia

DOI: https://doi.org/10.26714/jodi.v4i1.752

Article Information

Article History:
Submitted July 21, 2025
Revised May 21, 2026
Accepted June 01, 2026

Keywords:
Accuracy; Classification; KNN; Machine Learning; Stunting.

Abstract

Stunting is a nutritional problem that can affect children's physical growth and cognitive development and has a long-term impact on the quality of future generations. Early detection of stunting is crucial to enable timely and effective interventions. As technology advances, machine learning algorithms such as K-Nearest Neighbors (KNN) offer potential solutions to improve the accuracy of stunting risk classification. This study aims to design a classification model based on the KNN algorithm in the early detection of stunting risk in toddlers. This research uses the 2024 stunting dataset obtained from Kaggle. The data is analyzed through the stages of cleaning, transformation, and division into training and testing data. The KNN model was tested with various K values to determine the optimal value. The results showed that the KNN model with a value of K=8 resulted in an accuracy of 93.80%, F1-Score of 93.65%, precision of 93.63%, and recall of 93.79%. This shows that KNN is reliable in classifying the nutritional status of toddlers and can be applied in stunting prevention efforts using more accurate data. This research contributes to developing machine learning-based classification systems that can support decision-making in public health programs, especially in reducing stunting *rates.*

✉ Corresponding Author:
E-mail: saefulamri@unimus.ac.id

e-ISSN: 2988 - 2109

INTRODUCTION

Stunting is an indicator of nutritional status that reflects chronic growth disorders in children due to long-term nutritional deficiencies. Stunting is characterized by shorter height than age standards due to chronic malnutrition and inadequate nutrient intake (1). In 2022 the number of under-fives in Indonesia is estimated to reach around 31.8 million people (2). This data shows that attention to the growth and development of toddlers is very crucial, because it concerns the quality of the nation's future generation (3). Toddlerhood is an important period in the process of human growth and development because it takes place very quickly and will never be repeated, so it is often referred to as the golden age (4).

Stunting has serious and diverse impacts, including negative consequences in both the short and long term. In the short term, stunted children are at risk of having a weak immune system, being more susceptible to disease, and experiencing barriers to cognitive development. According to longitudinal research on intelligence development, about 50% of a child's cognitive abilities develop during the first four years of life, and this development reaches 100% when the child turns 18 years old (5). This means that stunting affects not only physical growth, but also cognitive development (6). Meanwhile, in the long term, stunting can lead to reduced economic productivity and contribute to the sustainability of the poverty cycle.

In Indonesia, stunting has become a serious public health problem because it can threaten the quality of future generations. According to data from the Indonesian Nutrition Status Survey (SSGI) in 2022, the prevalence of stunting in children under five reached 21.6% (7). This figure shows that almost one in five children under five experience growth delays, which have an impact on reducing learning abilities, future productivity, and increase the risk of metabolic diseases (8). Therefore, efforts to prevent stunting from an early age are very important so that the long-term impact on children's health and well-being can be minimized.

The development of information technology encourages the utilization of artificial intelligence in various fields, including health. One application is in the classification and early detection of stunting risk. The use of machine learning algorithms is proven to increase the effectiveness in detecting various medical conditions, including stunting, through the utilization of more comprehensive data (9). One of the popular algorithms in health data classification is K-Nearest Neighbors (KNN). K-Nearest Neighbors is a classification method that determines the class of an object based on its proximity to the training data, in this method, the calculation of the distance between data is carried out using the Euclidean Distance (10).

KNN has shown good performance in identifying stunting cases, making it an effective option for data-driven decision making in resource-constrained areas (11). In the evaluation of child nutritional status, factors such as age, weight, height, medical history, and environmental conditions can be used to identify whether a child is stunted. The KNN method serves to group symptom data based on similar characteristics, so as to identify patterns related to stunting (12).

Several relevant previous studies have tested the effectiveness of the KNN method in the classification of nutritional status of toddlers, where Ramadhani et al. (13) showed that the KNN algorithm was able to classify malnutrition into two conditions, namely marasmus and kwashiorkor, with optimal accuracy of 87%. In addition, research conducted by Ritonga and Muhandhis (14) revealed that the application of KNN in the detection of nutritional status of toddlers resulted in an accuracy of 74.73%, which shows the potential of KNN in child health applications.

In addition, Wahyudi et al. (15) also confirmed that in their research using the K-Nearest Neighbors (KNN) method was effective in classifying the nutritional status of toddlers, with accuracy results reaching 88% for the K = 3 value, which shows the reliability of this algorithm in analyzing the nutritional condition of children. However, most previous studies have focused on testing the accuracy of the algorithm only. Not many have examined in depth how this classification system can be implemented as a decision-making tool, especially in sustainable stunting prevention programs.

Therefore, this study aims to design a KNN based classification model to detect early stunting risk in toddlers and evaluate its performance using a representative dataset. The results of the classification process are expected to make a real contribution in efforts to accelerate the reduction of stunting rates through the use of technology as a supporting tool for more accurate and data-based analysis and decision making. In addition, the results of this study can be used as a basis for developing machine learning-based health information systems, as well as being a scientific reference in the development of classification algorithms in the field of public health.

METHODS

Data Source

This study uses secondary data obtained from kaggle "Stunting Toddler (Toddler) Detection (121K Rows)". Determination of stunting status is based on the Z-score formula according to the standards of the World Health Organization (WHO). Z-score is an indicator that shows how far a value deviates from the average in standard deviation units. This value represents the position of the data on the horizontal axis of the normal distribution (16). The dataset used includes approximately 121,000 data entries containing medical information and characteristics of toddlers, which are then used to predict the incidence of stunting using the KNN algorithm.

Research Variables

This study utilizes data covering four variables, all of which play an important role in identifying and intervening early in children who are at risk or experiencing stunting. These variables are used to analyze growth patterns and detect the risk of stunting in toddlers. The following are the variables used in the study.

Table 1. Research Variables

Variable	Description
Age	Age of toddlers (0-59 months)
Gender	Male, Female
Height	Height of toddlers (centimeters)
Nutritional Status	Classification based on standard deviation score for stunting risk identification: Severely Stunted, Stunted, Normal, High

Research Steps

The research stages carried out in this study are as follows:

1. Import Dataset
   Dataset import is the process of entering data from external sources into analysis software such as python. The dataset is the main source that contains the information to be analyzed. Usually, the data is stored in Comma Separated Values (CSV) format so that it is easy to read and process.
2. Data Preprocessing
   Preprocessing is the process of cleaning and preparing data before analysis. It involves deleting or dealing with blank or incomplete data, converting categorical data into numerical form if required, and standardizing or normalizing the data so that all features are on a uniform scale.
3. Split Data
   Split data is the process of dividing the dataset into two main parts, namely training and testing data, which are used for model training and testing. This stage aims to ensure that the model can be evaluated objectively using data that has never been seen before. In addition to data separation, it is important to maintain a balanced distribution of the data for more

accurate evaluation results. If necessary, re-standardization is also performed to ensure scale consistency between variables before the training process begins.

4. KNN Classification
   KNN Classification is the process of data classification using the KNN algorithm. In this process, the algorithm will identify a number of K closest data (nearest neighbors) of the data to be classified, based on a certain distance such as Euclidean Distance. The class that appears the most among these neighbors will be used as a prediction. The following are the stages in applying the KNN algorithm (17):
   a. Determining the value of K
      The K value can be determined using the following formula:

$$k = \sqrt{N} \tag{1}$$

      Where N is the number of samples in the training data.
   b. Measuring distance (Euclidean distance)
      Calculate the distance between each data in the training data and the data to be classified. The general formula for calculating the Euclidean distance is:

$$d_i = \sqrt{\left(x_{ki} - x_{kj}\right)^2 + \left(x_{ki} - x_{kj}\right)^2} + \cdots + \left(x_{ki} - x_{kj}\right)^2 \tag{2}$$

      Description:
      d_i : Euclidean distance between two points
      x_{ki} : Attribute in the 1st training data
      x_{kj} : Attribute in the 1st testing data
   c. Grouping data based on the calculated distance
      Test data is grouped into categories based on the training data with the closest distance according to the results of the Euclidean calculation.
   d. Identifying the nearest neighbors
      From all the calculated distances, select K training data with the smallest distance to the test data.
   e. Determining the classification result
      Classify the test data by looking at the majority class of the K nearest neighbors.
5. Model Accuracy
   Model accuracy is an evaluation stage that aims to measure the performance of the classification model that has been built. In this stage, accuracy is calculated based on the percentage of correct predictions against all test data. In addition to accuracy, other evaluation metrics such as precision, recall, and F1-score are also used to provide a more comprehensive picture of the model's performance. To find out how well the model classifies each class, the confusion matrix is used as a visualization tool for the classification results.

RESULTS AND DISCUSSION

Import Dataset

The data in this study was obtained from the Kaggle platform developed by Eduardo D'Anjour. The dataset is stored in Comma Separated Value (CSV) format, which facilitates the process of analyzing and processing data using various software such as python. The first step was to import the dataset using the pandas library. With the pd.read_csv() function, the data is read into a DataFrame structure which makes it easy to analyze more The data set includes 121,000 entries with 4 variables. The following is the sample obtained.

Table 2. Sample Data on Stunting in Toddlers

Age (month)	Gender	Height (cm)	Nutrition Status
0	Male	44.591973	Stunted
0	Male	56.705203	High
0	Male	46.863358	Normal
0	Male	47.508026	Normal
0	Male	42.743494	Severly Stunted

Data Preprocessing

Before the data is analyzed using the KNN model, it is necessary to preprocess the data to ensure that the data is in a condition that is suitable for use. This stage is very important because KNN is a distance-based algorithm, so the quality of the data greatly affects the accuracy of the results obtained. The following are the stages of data preprocessing carried out in this study.

1. The data collection stage involves taking raw data that will be used as analysis material in the KNN model.
2. Data cleaning, in this study 81,574 duplicate data were removed which were then not used in the study.
3. Data transformation was performed by converting categorical variables to numeric using the map() function. The Gender column was converted to 0 for males and 1 for females. Meanwhile, Nutritional Status was converted to 0 for severely stunted, 1 for stunted, 2 for normal, and 3 for high. This transformation is needed so that the data can be processed by the machine learning algorithm, KNN.

Split Data

The data division stage in the KNN algorithm aims to separate the dataset into 80% training data and 20% testing. Training data serves to build the model, while testing data is used to measure the accuracy of the model against new data. This proportional division is important to minimize overfitting and have good generalization power. After the split process, data standardization is performed using "StandardScaler" to standardize the features in the dataset so that they have a uniform scale. This is important because the KNN algorithm is very sensitive to scale differences between features, so standardization helps the model calculate the distance between data fairly and accurately.

KNN Classification

The classification stage using the K-Nearest Neighbors (KNN) algorithm begins with determining the optimal number of nearest neighbors (K value). The choice of K value is a crucial factor because it significantly affects the classification performance of the model. Based on the test results, the KNN model shows high accuracy, both on training and testing data. The following is the accuracy for all K values.

Table 3. Accuracy of K Value in Training Data

K Value	Akurasi Train
1	0.9515
2	0.9489
3	0.9550
4	0.9548
5	**0.9560**
6	0.9545
7	0.9550

8	0.9530
⋮	⋮
⋮	⋮
30	0.9450

Table 3 shows that using five nearest neighbors (K=5) produces the highest accuracy of 95.60%, so it was chosen as the best parameter for the KNN model. Although several other K values also produce similar accuracy, K=5 is considered the most optimal because it is able to capture local patterns without being too affected by noise, and avoid overfitting or underfitting. In addition, increasing the K value did not show a significant change in accuracy, indicating the stability of the model performance. Therefore, the nearest neighbor k value that will be used for the KNN model is K=5. Next, the KNN model will be evaluated to measure its performance by calculating the accuracy.

Model Accuracy

The evaluation of the classification model in this study is carried out using three main metrics, namely accuracy, precision, and recall. These three metrics were chosen to provide a comprehensive picture of the level of accuracy and reliability of the model in making predictions. To support a more in-depth analysis of model performance, a confusion matrix is also used as a visualization tool that shows the details of the classification results, including the number of correct and incorrect predictions in each class. The following table presents the confusion matrix of the best classification model in this study, namely:

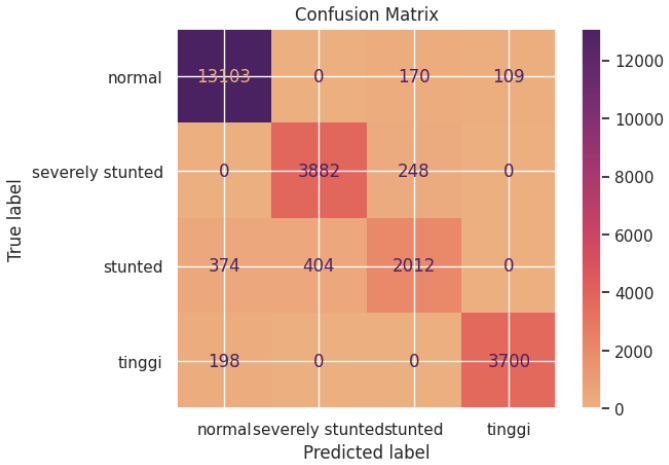

Figure 1. Confusion Matrix

Figure 1 showing the confusion matrix shows that the model has good classification capabilities in the normal and high classes, with correct predictions of 13,103 and 3,700 data, respectively. However, the model still has difficulty distinguishing between the stunted and severely stunted classes, as seen from the 2,012 stunted data that were misclassified as severely stunted, and 248 severely stunted data that were predicted as stunted. This is thought to be due to the similar characteristics between the two classes. In general, the accuracy of the model is high in the extreme class, but still needs improvement for the middle class. Furthermore, the performance evaluation was conducted by calculating the recall, accuracy, and precision values based on the confusion matrix.

Table 4. KNN Performance Results

Algorithm	Accuracy	Precision	Recall
KNN	0.9380	0.9363	0.9379

Table 4 shows that the K-Nearest Neighbors (KNN) algorithm is able to provide a fairly good performance, with an accuracy of 93.80%, precision of 93.63%, and recall of 93.79%. The high accuracy value indicates that the model is able to classify the overall data correctly. Meanwhile, the precision value reflects that the model rarely produces false positive predictions, and the high recall value indicates that the model successfully identifies most of the data that should belong to the positive category. These results show that the KNN algorithm has the potential to be an effective choice in solving classification problems that require a high level of accuracy and completeness of identification.

The accuracy difference between the training data and the testing data is only 1.80%. This small discrepancy indicates that the KNN model does not suffer from overfitting or noise, meaning the model performs well not only on the training data but also on the testing data. This demonstrates that the KNN algorithm with a value of K = 8 is capable of delivering stable, accurate, and consistent performance across both datasets. Given the relatively minor difference in accuracy, the model can be considered a suitable and optimal classification approach based on the evaluation of performance metrics.

CONCLUSION

This study shows that a classification model based on the K-Nearest Neighbors (KNN) algorithm can be used effectively to detect early stunting risk in toddlers. Using a dataset that includes variables of age, gender, height, and nutritional status, the K-Nearest Neighbors (KNN) model successfully classifies stunting conditions in toddlers with high accuracy. The highest accuracy value of 0.9380 was obtained at parameter K=8. In addition, the model evaluation showed excellent performance with high F1-score, precision, and recall values. These results indicate that KNN is an appropriate tool to support decision-making in stunting prevention in Indonesia. It is hoped that the application of this model can accelerate the reduction of stunting rates among children under five and contribute to improving the quality of health of future generations.

REFERENCES

1. Lestari WS, Saragih YM, Technology I, Mikroskil U. Multiclass Classification For Stunting Prediction Using Deep Neural Networks. 2024;10(2):386–93.
2. Kementerian Kesehatan Republik Indonesia. Profil Kesehatan Indonesia [Internet]. 2023. 1–550 p. Available from: https://kemkes.go.id/id/indonesia-health-profile-2022
3. Aryanti AD. Gambaran Pertumbuhan dan Perkembangan Balita di Wilayah Tempat Pembuangan Akhir (TPA) Antang Kota Makassar. 2024; Available from: https://repository.unhas.ac.id/id/eprint/36706/
4. Diyah HS, Sari DL, Nikmah AN. Hubungan Antara Pola Asuh dengan Status Gizi Pada Balita. J Mahasiwa Kesehat. 2020;1(2):151–8.
5. Kemenkes. Kemenkes. 2022. Faktor yang Mempengaruhi Pertumbuhan dan Perkembangan Anak. Available from: https://yankes.kemkes.go.id/view_artikel/1340/faktor-yang-mempengaruhi-pertumbuhan-dan-perkembangan-anak
6. Rohloff P, Flom P. Stunting: methodological considerations for improved study design and reporting. BMJ Paediatr open [Internet]. 2023 May 10;7(1):e001908. Available from: https://bmjpaedsopen.bmj.com/content/7/1/e001908
7. Kementrian Kesehatan RI. Buku Saku Hasil Survei Status Gizi Indonesia (SSGI) 2022 [Internet]. 2023. Available from: https://repository.badankebijakan.kemkes.go.id/id/eprint/4855
8. UNICEF. Levels and trends child malnutrition: UNICEF/WHO/World Bank Group Joint Child Malnutrition Estimates [Internet]. Vol. 24, Geneva: WHO. 2020. 1–16 p. Available from: https://www.who.int/publications/i/item/9789240003576
9. Lonang S, Yudhana A, Biddinika MK. Analisis Komparatif Kinerja Algoritma Machine Learning untuk Deteksi Stunting. J Media Inform Budidarma. 2023;7(4):2109.
10. Yudhana A, Muslim A, Wati DE, Puspitasari I, Azhari A, Mardhia MM. Human Emotion Recognition Based on EEG Signal Using Fast Fourier Transform and K-Nearest Neighbor. Adv Sci Technol Eng Syst J [Internet]. 2020;5(6):1082–8. Available from: https://astesj.com/v05/i06/p131/

11. Putri IP, Terttiaavini T, Arminarahmah N. Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak. MALCOM Indones J Mach Learn Comput Sci. 2024;4(1):257–65.
12. Azis MF, Kaesmetan YR. Penerapan K-NN (K-Nearest Neighbors) Pada Sistem Pakar Diagnosa Gejala Stunting Pada Balita Menggunakan Naïve Bayes Classifier. Sist J Ilm Sist Inf [Internet]. 2024 Oct 1;1(1):75–91. Available from: https://ejournal.rizaniamedia.com/index.php/sistematis/article/view/120
13. Ramadhani DH, Jumadi J, Sandi G. Implementasi Algoritma K-Nearest Neighbors (KNN) Untuk Prediksi Gizi Buruk. SMATIKA J [Internet]. 2024 Dec 16;14(02):326–36. Available from: https://jurnal.stiki.ac.id/SMATIKA/article/view/1360
14. Ritonga AS, Muhandhis I. Aplikasi Berbasis Website Untuk Mendeteksi Status Gizi Balita Menggunakan Metode K-Nearest Neighbors (KNN). J Syst Comput Eng [Internet]. 2024 Jan 22;5(1):44–55. Available from: https://journal.unpacti.ac.id/index.php/JSCE/article/view/1081
15. Wahyudi R, Orisa M, Vendyansyah N. Penerapan Algoritma K-Nearest Neighbors Pada Klasifikasi Penentuan Gizi Balita (Studi Kasus di Posyandu Desa Bluto). JATI (Jurnal Mhs Tek Inform [Internet]. 2021 Oct 24;5(2):750–7. Available from: https://ejournal.itn.ac.id/index.php/jati/article/view/3738
16. Yousan MM, Latuconsina R, Ansori ASR. Aplikasi Penentuan Gizi Anak Laki- Laki Sesuai Dengan Standar Who (World Health Organization) Menggunakan Metode Z-Score. eProceedings Eng [Internet]. 2020;7(1):1425–33. Available from: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/11630
17. Fasnuari HA, Dwi, Yuana H, Chulkamdi MT. Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Diabetes Melitus. Antivirus J Ilm Tek Inform [Internet]. 2022 Oct 18;16(2):133–42. Available from: https://ejournal.unisbablitar.ac.id/index.php/antivirus/article/view/2445