Analysis of Stunting Prediction for Toddlers in Bekasi Regency Using the K-Nearest Neighbors and Random Forest Algorithms

1st Kamelia Khoirunnisa
School of Computing
Telkom University
Bandung, Indonesia
kameliakhoirunnisa@student.telkomuniversity.ac.id

2nd Putu Harry Gunawan
School of Computing
Telkom University
Bandung, Indonesia
phgunawan@telkomuniversity.ac.id

***Abstract*—Stunting, a condition where children are shorter than their age, is a serious nutritional issue in developing countries, including Indonesia. Research shows that Low Birth Weight (LBW) significantly affects children's growth. In Bekasi Regency, the stunting prevalence remains high at 17%, with a target reduction to 14%. Achieving this target requires prevention efforts focused on improving nutrition and regularly monitoring child growth. In such monitoring, innovative evaluation techniques using Machine Learning (ML) are needed to predict stunting potential. This study aims to develop a predictive model for early detection of stunting risks in Bekasi Regency, using machine learning techniques to analyze Low Birth Weight (LBW) and Low Birth Length (LBL) factors, which could potentially be integrated into the local health monitoring system for preventive intervention. RF excels in handling complex features and identifying important predictors, while KNN is effective at recognizing local patterns. The results show that RF achieved the best performance with 99.22% accuracy and an F1-score of 96.94%, compared to KNN with 96.19% accuracy and an F1-score of 87.16%, highlighting RF's greater stability and robustness over KNN in predicting stunting cases. This study is expected to provide an accurate predictive system that helps parents, health workers, and the government identify stunting potential early while also determining the appropriate ML algorithm for stunting case prediction in Indonesia. Future research is encouraged to test this model in other regions with different characteristics to ensure the generalizability and effectiveness of stunting prediction on a broader scale.**

***Index Terms*—Stunting, Machine Learning, K-Nearest Neighbors, Random Forest**

I. Introduction

Stunting, a condition where children's height is significantly lower than the average for their age, remains a critical nutritional challenge, particularly in developing and underdeveloped countries [1]. Research examining 26 stunting-related factors has identified Low Birth Weight (LBW) as a significant contributor that substantially impacts child growth and development [2]. Given LBW's role as a causative factor, addressing stunting becomes increasingly crucial for preventing adverse effects on child development.

Indonesia's commitment to reducing stunting rates to 14% by 2024, as stated by President Jokowi, has particular relevance for Bekasi Regency, which has demonstrated strong dedication to this cause. According to Acting Regent Dani Ramdan's report on *bekasikab.go.id*, the region has already achieved a notable reduction in stunting rates from 21% to 17% between 2021 and 2022. However, reaching the 14% target requires sustained effort and enhanced collaboration among stakeholders.

This commitment aligns with WHO's global target for stunting reduction by 2025. Indonesia, identified as one of 34 countries with high stunting prevalence, must contribute to WHO's goal of reducing stunting rates by 40% by 2025 [3]. Bekasi Regency's efforts directly support Indonesia's commitment to achieving this target.

Machine Learning (ML) offers an innovative approach to stunting prevention through improved nutrition monitoring and child development assessment. ML techniques can predict potential stunting cases, enabling more proactive and targeted preventive measures [4].

Previous research in 2021 explored ML applications in addressing childhood stunting, specifically using the K-Nearest Neighbors classification algorithm. The study achieved optimal performance with an accuracy of 97.31% using specific data partition ratios [5]. A more recent study in February 2024 implementing Random Forest algorithms and Cross Validation for stunting prediction demonstrated good accuracy at 77.55% [6].

Another recent study in 2024 compared several ML algorithms for stunting prediction, evaluating Naive Bayes (NB), K-Nearest Neighbors (KNN), and Random Forest (RF). In this study, NB achieved an accuracy of 83.2%, KNN reached 84.8%, and RF achieved the highest accuracy at 87.75% [4]. A similar study also compared stunting prediction accuracy using KNN, RF, and Boosted KNN (BK) algorithms. In this study, KNN achieved an accuracy of 93.55%, RF was higher with an accuracy of 97.70%, and BK achieved the highest accuracy of 98.00% [7].

Although algorithms like K-Nearest Neighbors (KNN) and Random Forest (RF) have been used in stunting prediction, no study has yet compared the effectiveness of these algorithms in the context of stunting data using LBW and LBL factors, specifically in Bekasi Regency, which has different demographic and socio-economic characteristics compared to other regions. This study aims to conduct a comparative analysis between K-Nearest Neighbors and Random Forest algorithms using infant data based on LBW and LBL for stunting prediction. The results could assist healthcare providers and government officials in determining the most reliable predictive tool, ultimately supporting the implementation of more effective stunting prevention policies and interventions.

II. METHODS

A. Research Design

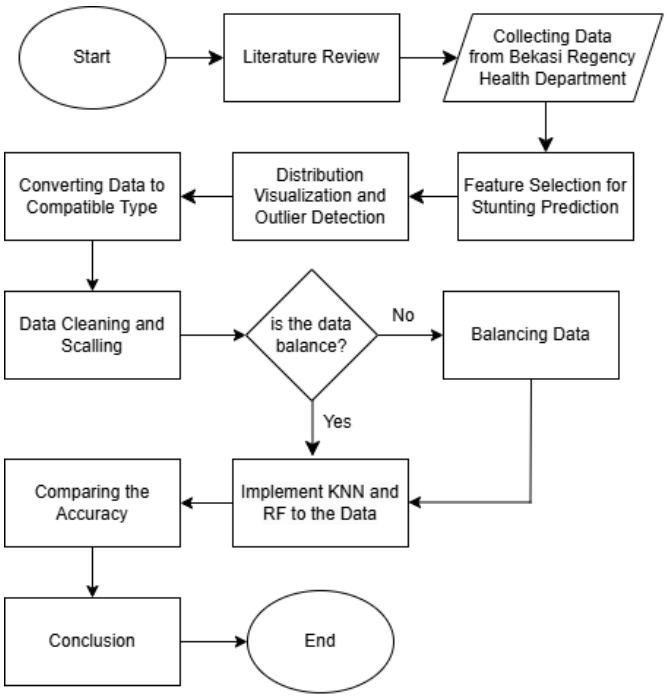

Fig. 1. Flowchart of research

The research process, as illustrated in Fig. 1, begins with a literature review to gather relevant theories and findings on stunting prediction. The next step is data collection from the Bekasi Regency Health Office. Following data collection, the process continues with feature selection, where key features related to stunting are identified for analysis. After feature selection, visualization and outlier detection are performed to explore data distribution and detect any anomalies. The dataset is then converted to a compatible type, ensuring it is suitable for analysis. Following this, data cleaning and scaling are applied, where normalization techniques are used to standardize the data values and mitigate the impact of outliers. If the dataset is imbalanced, techniques such as SMOTE are applied to ensure fair class representation. Finally, the K-Nearest Neighbors (KNN) and Random Forest (RF) models are implemented and compared to evaluate performance. The research concludes with a summary of findings and documentation of the analysis.

B. K-Nearest Neighbors

K-Nearest Neighbor (KNN) is a widely used non-parametric supervised learning algorithm for classification and regression that recognizes patterns in data by deriving input from a training set to generate an output model [8], [9]. In order to effectively classify or predict data, it is crucial to evaluate the proximity of data points. Specifically, the KNN method involves determining the shortest distance between the data being evaluated and its K nearest neighbors within the training dataset [10]. To calculate this distance, the Euclidean distance formula is commonly employed, as it provides the straight-line distance between points and is the most prevalent method used in KNN [11].

The Euclidean Distance can be formulated as follows:

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

where x_i represents the data sample, y_i represents the test data, and $d(x,y)$ is the distance between x_i and y_i. Following this foundational understanding, the next section will outline the step-by-step procedure of the K-Nearest Neighbor algorithm, highlighting the specific processes involved in classifying new data points based on their proximity to the existing training data. To illustrate this concept, Algorithm 1 provides a straightforward implementation of the KNN algorithm.

Algorithm 1 K-Nearest Neighbors

Input: X: Training data, Y: Class labels of X, K: number of nearest neighbors
Output: Predicted class of test data x

Start
Classification (X, Y, x)
1. *for* each x data point *do*
   Calculate the Euclidean distance between x and each point in X
   end for
2. Classify each test data point x into the majority class of its nearest neighbors

End

C. Random Forest

Random Forest (RF) is a popular supervised learning algorithm used for classification and regression, which leverages ensemble learning to handle complex problems by improving model accuracy through a combination of multiple decision trees applied to various dataset subsets [12], [13]. This technique averages the predictions from numerous trees, each independently trained on a random sample, to increase predictive power. Unlike single decision trees that are prone to high variance, Random Forest reduces this issue by maintaining a low-bias structure [14].

The RF algorithm functions as a hierarchical ensemble of tree-based classifiers, selecting the most relevant attributes for model construction based on a predetermined probability [15]. Particularly useful in high-dimensional datasets, Random Forest effectively filters out irrelevant features, focusing on informative ones that improve classifier performance [16]. Each tree in the Random Forest makes an independent prediction, and the final classification or regression output is determined by aggregating the outputs of all trees, typically through a majority vote, thereby ensuring a robust and accurate predictive model [12]. To see a straightforward implementation of the RF algorithm, please refer to [17].

D. Evaluation Metrics

The performance evaluation will be conducted using the Confusion Matrix method. The Confusion Matrix is used to assess the performance and accuracy of the classification process [6]. It consists of the values False Positive (FP), True Negative (TN), True Positive (TP), and False Negative (FN). These values in the Confusion Matrix can then provide performance measures. The following are the evaluation metrics that will be utilized:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\text{-}score = 2 \times \left(\frac{precision \times recall}{precision + recall} \right) \tag{5}$$

Accuracy is the ratio of correct predictions (TP and TN) to the total data, measuring overall model performance, Precision is the ratio of true positives to all predicted positives, assessing the accuracy of positive predictions, Recall measures the model's ability to identify true positives from all actual positives [18], and F1-Score is the harmonic mean of precision and recall, balancing both metrics when class distributions are uneven . In (2) to (5), TP (True Positive) refers to the cases of stunting that were accurately identified, TN (True Negative) signifies the cases of non-stunting that were correctly identified, FP (False Positive) denotes instances that were mistakenly classified as stunted, and FN (False Negative) indicates instances that were incorrectly classified as non-stunted.

III. RESULTS AND DISCUSSIONS

A. Exploratory Data Analysis

The dataset used in this study was obtained from the Bekasi Regency Health Office. The data collection, conducted in February 2024, includes real-time measurements from infants with Low Birth Weight (LBW) and Low Birth Length (LBL). The dataset consists of 2.255 entries. After feature selection, the following attributes were identified for inclusion in the analysis, including *Gender, Age_at_Measurement, Weight, Height, Weight/Age, Weight/Age_ZS, Height/Age, Height/Age_ZS, Weight/Height, Weight/Height_ZS*. Of these attributes, *Height/Age* is used as the target variable for the training model because it indicates whether a child's height is appropriate for their age, serving as a critical marker for identifying stunting conditions.

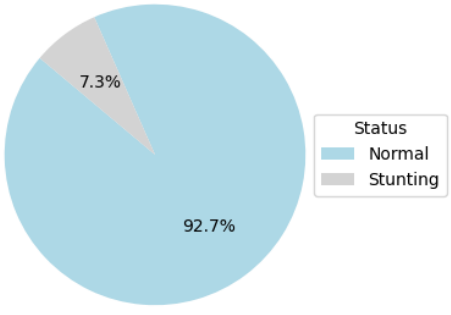

Fig. 2. Percentage of Toddlers with Normal and Stunting Status

As shown in Fig. 2, the distribution of the target feature is highly imbalanced, with the "Normal" class significantly outnumbering the "Stunting" class. This uneven distribution can lead to model bias toward the majority class, potentially reducing its accuracy for the minority class and impacting overall predictive performance. The next challenge in the analysis can be seen in Fig. 3, where the variables *Weight, Weight/Age_ZS*, *Height/Age_ZS*, and *Weight/Height_ZS* exhibit outliers, particularly at the upper and lower extremes, indicating the presence of extreme values in the distribution. In contrast, the variables *Age_at_Measurement* and *Height* do not contain outliers, with a more normally distributed pattern. These outliers have the potential to disrupt the analysis and reduce accuracy, making careful handling essential to improve model performance and ensure reliable predictions.

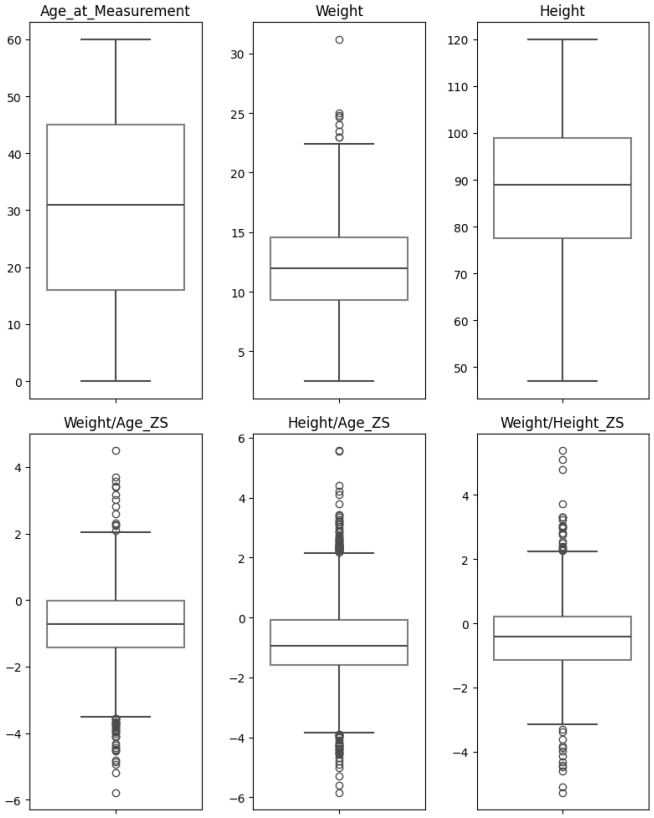

Fig. 3. Visual Data Distribution of Numerical Features Before Prepocessing

B. Data Prepocessing

The dataset contains 33 feature columns, but only 10 columns are used for analysis. Categorical data is transformed into numerical format through encoding. For example, the 'Gender' column is converted from 'M' and 'F' to 0 and 1. The 'Weight/Age' column is encoded with 'Severely Underweight' represented by 0, 'Underweight' as 1, 'Normal' as 2, and 'Risk of Overweight' as 3. The 'Height/Age' column is encoded with 'Normal' and 'Tall' both represented by 0, while 'Stunted' and 'Severely Stunted' are assigned a value of 1. The 'Weight/Height' column is encoded with 'Severely Wasted' as 0, 'Wasted' as 1, 'Normal' as 2, 'Risk of Overweight' as 3, 'Overweight' as 4, and 'Obese' as 5. The 'age_at_measurement' column is converted to represent the number of months.

Missing values in the 'Height/Age' and 'Weight/Height' columns are addressed by removing rows containing these missing values. Similarly, invalid age data in the 'Age_at_Measurement' column is also removed. Outliers are managed using the Interquartile Range (IQR) method, with values replaced by the median, and winsorization is applied to limit extreme values. Additionally, feature normalization is performed using MinMaxScaler to ensure the data is within

an appropriate range. As shown in Fig. 4, the outliers have been effectively handled.

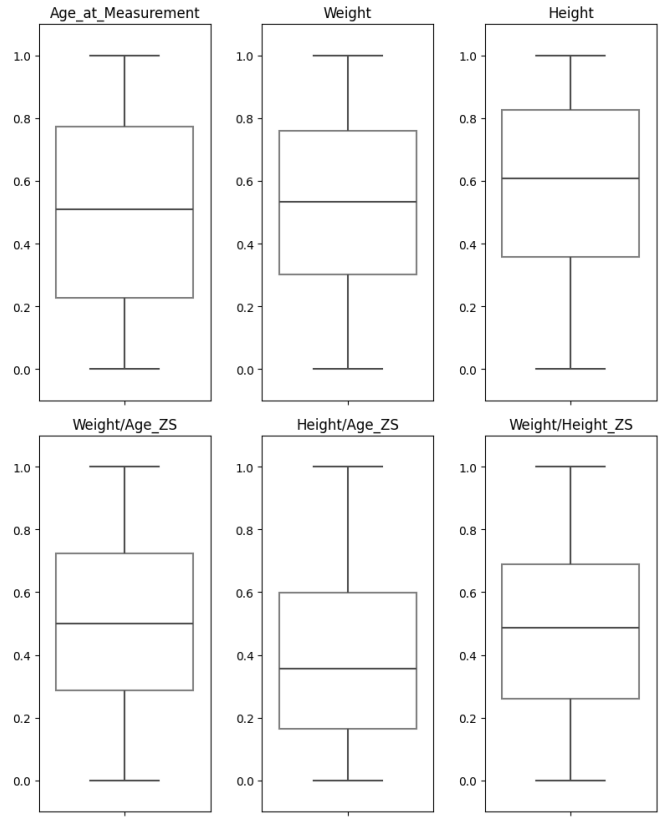

Fig. 4. Visual Data Distribution of Numerical Features After Prepocessing

C. Handling Imbalance

An issue related to data imbalance was identified, as shown in Fig. 2. Of the total 2.255 records, only 165 children are classified as stunted, while 2.090 are not experiencing growth impairment. To address this, handling is performed using the Synthetic Minority Over-sampling Technique (SMOTE).

Inspired by the approach in the study by [19], where SMOTE is applied to address data imbalance through an oversampling method that generates synthetic data while preserving the original data distribution, SMOTE was adopted in this analysis to manage the imbalance in the dataset, as it closely aligns with the research objectives and dataset characteristics. After applying SMOTE to the training dataset, both normal and stunted cases were balanced, resulting in 2.504 instances of each. Achieving this balanced distribution is essential for enhancing the performance of machine learning models, as they could otherwise become biased toward the majority class.

D. Implementation of K-Nearest Neighbors

For this analysis, the dataset was split into training and testing sets to develop and evaluate the models. Specifically, 60% of the data, totaling 1.353 records, was allocated to the training set, while the remaining 40%, totaling 902 records, was used for testing.

The K-Nearest Neighbors (KNN) model was developed using the Python library Scikit-learn, with a configuration of the distance metric set to 'euclidean', a neighborhood size of 5, and distance-based weighting. These parameters enhance the model's classification capability by assigning greater influence to closer neighbors within the feature space, which improves prediction accuracy.

As shown in Table I, the model consistently demonstrates strong performance across multiple runs, with the confusion matrix reflecting similar results in each run. The KNN model accurately classifies the majority of samples as either stunting or normal, indicating its high proficiency in identifying between stunting and normal cases for binary classification tasks.

TABLE I
CONFUSION MATRIX FOR THE KNN MODEL

		Predicted Values	
		Stunting	**Normal**
Actual Values	**Stunting**	55	8
	Normal	26	804

E. Implementation of Random Forest

In addition to KNN, the analysis was also conducted using Random Forest (RF) because research indicates that RF often produces superior performance, particularly in capturing complex patterns among variables. The RF model was developed using the Scikit-learn library, utilizing 100 estimators, the 'gini' criterion, without depth limitations, and without bootstrapping. The combination of high accuracy and resilience to overfitting makes RF a suitable choice for this task. The results of the analysis, as presented in Table II, confirm that RF is indeed an appropriate choice for this analysis. The table shows the confusion matrix values with the maximum value achieved over several tests, where the performance of RF remains higher and more consistent compared to KNN.

TABLE II
CONFUSION MATRIX FOR THE RANDOM FOREST MODEL

		Predicted Values	
		Stunting	**Normal**
Actual Values	**Stunting**	59	6
	Normal	4	828

F. Model Evaluation

TABLE III
ACCURACY AND F1-SCORE OF MODELS

Model	**Accuracy**	**F1-score**
K-Nearest Neighbors	96.19%	87.16%
Random Forest	99.14%	96.68%

*this values are average from 20 times running

Table III presents the accuracy and F1-score of the K-Nearest Neighbors (KNN) and Random Forest (RF) models evaluated in this study. The KNN model achieved an accuracy of 96.19% and an F1-score of 87.16%. Although the model performs well, its lower F1-score suggests that KNN may have limitations in handling the complexity of this dataset, particularly in balancing precision and recall. This result suggests that KNN, while effective in many predictive tasks, may face challenges in fully capturing the complexity of stunting prediction, even after addressing class imbalance.

This observation aligns with the findings in earlier research, such as the study by [4], where KNN demonstrated good performance but was better at identifying stunted cases due to its higher recall compared to other algorithms.

On the other hand, the RF model achieved a higher accuracy of 99.22% with an F1-score of 96.94%. This highlights RF's advantage in classifying data with high accuracy and a better balance between precision and recall. This superior performance can be attributed to RF's ability to capture complex patterns and interactions within the data through its ensemble-based decision tree approach. Random Forest models are particularly well-suited for tasks involving multiple features and complex relationships, which is evident in this study where the data involves a mix of continuous and categorical attributes. This result is consistent with findings in previous studies, such as those by [4] and [7], which reported RF's robust performance in similar stunting prediction tasks. These studies demonstrated that RF outperforms KNN, particularly in datasets that require effective handling of feature interactions, even after addressing class imbalances.

IV. Conclusion

A comparative study of the K-Nearest Neighbors (KNN) and Random Forest (RF) algorithms for predicting stunting in Bekasi Regency demonstrates that machine learning approaches can effectively support early detection of stunting, particularly in cases of Low Birth Weight (LBW) and Low Birth Length (LBL). The KNN model demonstrated stable results across 20 runs, with an accuracy of 96.19% and an F1-score of 87.16%, highlighting consistent performance but limitations in capturing the dataset's complexity. In contrast, the RF model achieved a higher range of accuracy, with values between 98.99% and 99.22%, alongside stronger F1-scores ranging from 96.13% to 96.99%. This superior performance, attributed to RF's ensemble-based decision tree approach, demonstrates its capability to capture intricate patterns and maintain a balanced classification performance, particularly after addressing class imbalance using SMOTE and implementing comprehensive data preprocessing. These findings contribute to the advancement of machine learning applications in healthcare and provide practical tools for practitioners and policymakers to achieve stunting reduction targets. However, future research should consider incorporating additional features, exploring other advanced algorithms, and developing more comprehensive real-time analysis systems to enhance stunting prevention across various demographic groups and geographic regions.

References

[1] S. B. Gaffar, N. N. M. B, and M. Asri, "PKM Pencegahan Stunting melalui Pendidikan Keluarga," in *Seminar National Results of Community Service*, 2021, pp. 22–25.

[2] I. M. Apriliani, N. P. Purba, L. P. Dewanti, H. Herawati, and I. Faizal, "Stunting Risk Factors in Children Under Five in Indonesia: A Scoping Review," *Indonesian Journal of Health Promotion*, vol. 5, no. 6, pp. 654–661, 2022.

[3] M. D. Onis *et al.*, "The World Health Organization's global target for reducing childhood stunting by 2025: Rationale and proposed actions," *Maternal & Child Nutrition*, vol. 9, no. S2, pp. 6–26, September 2013.

[4] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, "Analisis Perbandingan Machine Learning untuk Prediksi Stunting pada Anak," *MALCOM Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 257–265, 2024.

[5] H. H. Sutarno, R. Latuconsina, and A. Dinimaharawati, "Prediksi Stunting Pada Balita Dengan Menggunakan Algoritma Klasifikasi K-Nearest Neighbors," *e-Proceeding Engineering*, vol. 8, no. 5, p. 6657, 2021.

[6] F. Azzahra, N. Suarna, and Y. A. Wijaya, "Penerapan algoritma random forest dan cross validation untuk prediksi data stunting," *KOPERTIP: Scientific Journal of Informatics Management and Computer*, vol. 8, no. 1, pp. 1–6, 2024.

[7] M. G. Daffa and P. H. Gunawan, "Stunting classification analysis for toddlers in Bojongsoang: A data-driven approach," in *2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)*, Feb 2024, pp. 42–46.

[8] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, "Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, March 2019, pp. 679–684.

[9] A. T. A. Sibuea and P. H. Gunawan, "Classifying stunting status in toddlers using k-nearest neighbor and logistic regression analysis," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, July 2024, pp. 6–11.

[10] J. B. Chandra and D. Nasien, "Application of Machine Learning K-Nearest Neighbour Algorithm to Predict Diabetes," *International Journal of Electrical, Energy and Power System Engineering*, vol. 6, no. 2, pp. 134–139, 2023.

[11] A. Almomany, W. R. Ayyad, and A. Jarrah, "Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3815–3827, 2022.

[12] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House price prediction using random forest machine learning technique," *Procedia Computer Science*, vol. 199, pp. 806–813, 2022.

[13] C. Fannany, P. H. Gunawan, and N. Aquarini, "Machine Learning Classification Analysis for Proactive Prevention of Child Stunting in Bojongsoang: A Comparative Study," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, July 2024, pp. 1–5.

[14] M. Pal and S. Parija, "Prediction of heart diseases using random forest," *Journal of Physics: Conference Series*, vol. 1817, no. 1, p. 012009, March 2021.

[15] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *The Journal of Supercomputing*, vol. 77, no. 5, pp. 5198–5219, 2021.

[16] H. Janawisuta and P. H. Gunawan, "Early detection of stunting in indonesian toddlers: A machine learning approach," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, July 2024, pp. 12–16.

[17] H. Guo, H. Nguyen, D.-A. Vu, and X.-N. Bui, "Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach," *Resources Policy*, vol. 74, p. 101474, 2021.

[18] W. I. Rahayu, C. Prianto, and E. A. Novia, "Perbandingan Algoritma K-Means dan Naïve Bayes untuk Memprediksi Prioritas Pembayaran Tagihan Rumah Sakit Berdasarkan Tingkat Kepentingan Pada PT. Pertamina (Persero)," *Jurnal Teknik Informatika*, vol. 13, no. 2, pp. 1–8, 2021.

[19] G. A. F. Khansa and P. H. Gunawan, "Predicting Stunting in Toddlers Using KNN and Naïve Bayes Methods," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*. IEEE, July 2024, pp. 17–21.