Gradient Boosting Machine Algorithm Optimization for Child Nutrition Classification at Al-Islam Hospital, Krian

Firman Rizky Ramadhan[1], Henni Endah Wahanani[2], Budi Mukhamad Mulyo[3]
[1,2,3] Universitas Pembangunan Nasional Veteran Jawa Timur
Email : 21081010238@student.upnjatim.ac.id , henniendah.if@upnjatim.ac.id , budi.m.mulyo.fasilkom@upnjatim.ac.id

The double burden of malnutrition in school-aged children (6–12 years old) remains a serious challenge for the healthcare sector in Indonesia. Conventional manual monitoring methods currently implemented in hospitals often face time constraints and are susceptible to human error. This study aims to optimize the architecture and parameters of the Gradient Boosting Machine (GBM) algorithm in classifying children's nutritional status digitally, quickly, and precisely at Al-Islam H.M. Mawardi Krian Hospital. The research dataset includes 257 medical records of pediatric patients from 2023–2025 containing anthropometric features such as age, gender, weight, height, and Body Mass Index (BMI). Data pre-processing stages are carried out through gender encoding, feature engineering, target binary transformation ("Good" and "Not Good"), feature selection, minority data balancing with SMOTE, and value standardization. The GBM model was configured using a parameter combination of 200 trees (n_estimators), a learning rate of 0.2, a maximum depth of 4 levels, and a subsample of 0.8. Test results demonstrated impressive and stable model performance, with an accuracy rate of 92% (0.92). In the 'Good' nutritional status category, the model achieved a precision of 0.94 and a recall of 0.98, while in the 'Poor' category, it achieved a precision of 0.80 and a recall of 0.57. This study concluded that the application of the boosting-based ensemble method is highly effective in improving the accuracy of clinical classification of children's nutritional status and has the potential to be integrated into hospital information systems as a real-time decision support system.

Keywords: Child Nutrition Classification, Gradient Boosting Machine, Medical Records, SMOTE, Decision Support System

This is an open access article under theCC BY-NClicense

Corresponding Author:
Henni Endah Wahanani
Universitas Pembangunan Nasional Veteran Jawa Timur
henniendah.if@upnjatim.ac.id

1. Introduction

Nutrition is a fundamental component that supports the growth, development, and maintenance of individual health (Nomira, 2024). Balanced nutritional intake is a key factor in supporting growth, development, and health, with individual needs varying based on age, gender, physical activity level, and physiological condition. For school-aged children (6–12 years), meeting nutritional needs is crucial because this is a period of rapid physical, cognitive, and psychosocial growth. Adequate nutritional intake not only plays a role in supporting optimal height and weight development but also contributes to improved learning ability, concentration, immune system, and productivity in the school environment. Conversely, an unbalanced consumption pattern can lead to the double burden of malnutrition, a condition where undernutrition and overnutrition occur simultaneously in a population. This condition has the potential to increase the risk of various chronic diseases, such as type 2 diabetes mellitus, hypertension, and cardiovascular disease in adulthood (Christian & Diana, 2023).

The double burden of malnutrition remains a serious challenge to health development in Indonesia. According to data from the National Agency for Research on Health (BRIN) (2023), the prevalence of overweight and obesity in children aged 6–12 years reached 19.7%, while the prevalence of wasting was

7.7% and underweight was 17.1%. These data indicate that Indonesia still faces complex nutritional challenges, where cases of undernutrition and overnutrition coexist. Suboptimal nutritional conditions during childhood have been shown to hinder motor development, cognitive function, and academic ability, as well as increase the risk of degenerative diseases in adulthood (Aulia, 2021). Therefore, regular monitoring and evaluation of children's nutritional status are necessary as a preventative measure to prevent future health complications.

Children's nutritional status is influenced by various interrelated factors, such as age, gender, weight, height, dietary patterns, physical activity, health conditions, and family socioeconomic factors. The complexity of the relationships between these variables makes it difficult to identify nutritional status optimally using conventional statistical approaches. In practice, healthcare professionals must consider various anthropometric indicators simultaneously to ensure that the resulting nutritional status classification is tailored to each individual's condition. Therefore, an analytical approach is needed that can more comprehensively identify patterns of relationships between variables, thereby improving the accuracy of the classification process.

Furthermore, advances in information technology have driven the digitization of medical records in various healthcare facilities. Electronically stored medical records not only serve as documentation of healthcare services but also hold significant potential as a data source for the development of Artificial Intelligence (AI)-based systems. Utilizing medical record data in the analysis process allows for the discovery of hidden patterns that are difficult to identify through manual observation. Therefore, the integration of artificial intelligence technology in healthcare can help improve service effectiveness, accelerate data analysis processes, and support more objective clinical decision-making.

In healthcare, early detection of children's nutritional status plays a crucial role in minimizing long-term health risks. However, current monitoring methods are still largely manual, requiring a relatively long time and being prone to human error. Furthermore, the increasing number of patients each year leads to a larger volume of medical record data that must be analyzed. This situation makes the manual process of classifying nutritional status inefficient when applied to large amounts of data. Therefore, the use of machine learning-based technology is an alternative solution capable of producing a fast, consistent, and highly accurate classification process.

In recent years, the application of machine learning in the healthcare sector has grown rapidly. Various algorithms have been utilized to assist in disease diagnosis, health risk prediction, patient condition classification, and the development of decision support systems. The advantage of machine learning lies in its ability to learn patterns from historical data without the need to manually develop classification rules. Through this learning process, the model is able to recognize complex relationships between variables, producing more accurate predictions than conventional approaches in various cases.

One machine learning algorithm widely used in classification problems is the Gradient Boosting Machine (GBM). GBM is an ensemble-based supervised learning algorithm that works by incrementally building multiple decision trees, with each new tree designed to correct the prediction errors of the previous model (Freund & Schapire, 1997). This approach allows the model to continuously improve its performance through iterative error optimization (loss function) (Suryana et al., 2021). These characteristics enable GBM to handle data with nonlinear relationships, reduce bias, and improve the model's generalizability to new data.

In addition to its high accuracy, GBM also offers flexibility in optimizing hyperparameters, such as learning rate, number of estimators, maximum depth, and subsample, allowing the model to be tailored to the characteristics of the dataset. These capabilities make GBM superior to several other classification

algorithms, especially when faced with structured data with a large number of variables. Therefore, this algorithm is widely used in various research fields in health, economics, and social data analysis (Natekin & Knoll, 2013).

Various previous studies have demonstrated the advantages of GBM in various fields. Sari and Prabowo (2025) demonstrated that GBM performed well in predicting regression data, with a coefficient of determination (R^2) of 0.7639. Other research also demonstrated that GBM yielded stable performance in predicting house prices and classifying social assistance recipients, outperforming several other classification algorithms (Dinata et al., 2022). Furthermore, research by Zhang et al. (2024) demonstrated that ensemble learning methods, including GBM, are excellent at handling structured data with complex characteristics, thus improving classification performance compared to conventional approaches.

However, the implementation of Gradient Boosting Machines (GBMs) in classifying children's nutritional status based on hospital medical records remains relatively limited. Most previous research has focused on disease prediction, social data classification, and economic analysis, while research on classifying children's nutritional status generally uses conventional statistical methods or simple classification algorithms. However, the characteristics of anthropometric data are quite complex, requiring algorithms capable of optimally learning the patterns of relationships between variables. Therefore, there is still research opportunity to evaluate the ability of GBMs to classify children's nutritional status using real-world medical record data.

This study aims to design and implement a Gradient Boosting Machine (GBM) algorithm to classify children's nutritional status at Al-Islam H.M. Mawardi Hospital, Krian, using a validated medical record dataset. The model development process includes data preprocessing, parameter optimization, model training, and performance evaluation using appropriate classification metrics. The research results are expected to produce a classification model with a high level of accuracy and can be implemented as a digital decision support system for medical personnel. In addition to contributing to the development of machine learning applications in the health sector, this research is also expected to help the early detection of malnutrition in children so that treatment can be carried out more quickly, precisely, and effectively.

2. Method

This study uses an experimental quantitative approach to classify children's nutritional status by applying the Gradient Boosting Machine (GBM) algorithm. The GBM algorithm was selected based on the structured nature of medical record data and its complex feature dimensions. This algorithm is considered superior due to its ability to minimize bias and variance through a sequential iteration process that gradually improves prediction errors from the previous decision tree. The primary focus of this methodology is to produce a stable and consistent model that accurately maps children's anthropometric variables into nutritional status categories. Figure 1 illustrates the research concept:

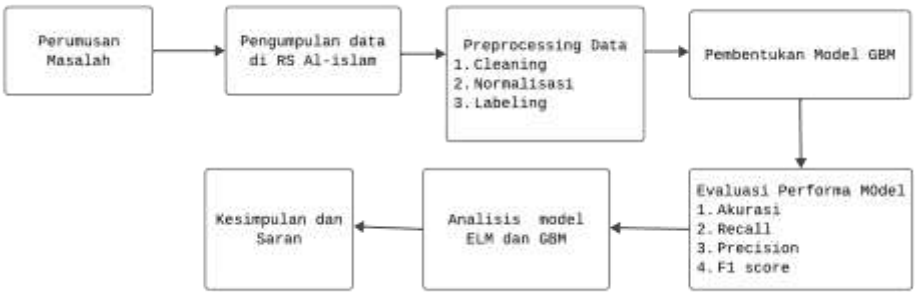


Figure 1. Research Concept

Data Collection

Research data was obtained from medical records of children aged 6-12 years at Al-Islam H.M. Mawardi Krian Hospital, Sidoarjo, for the period 2023-2025. The dataset consists of basic physical features including name, age, weight (BW), height (H), gender, and Body Mass Index (BMI).

Data Preprocessing

The next step is data preprocessing to ensure data quality before the modeling stage. Data preprocessing in this study includes several steps, such as gender encoding, feature engineering, data transformation, feature selection, balancing using SMOTE, and data standardization.

Gradient Boosting Machine (GBM) Modeling

GBM works iteratively by building a series of decision trees. The model configuration uses 200 trees (n_estimators) with a maximum depth of 4 levels. The learning rate parameter was set at 0.2 to control the strength of each new tree's correction to the previous model's errors, and a subsample of 0.8 was used to improve the model's generalizability.

Model Performance Evaluation

The model was evaluated using the Confusion Matrix, accuracy, precision, recall, and F1-score metrics to measure classification performance on the test data. The Confusion Matrix is highly effective for measuring classification model performance by presenting detailed visualizations of predicted and actual data. The Confusion Matrix provides a clear picture of the model's ability to distinguish between different nutritional categories (Good and Bad).

3. Result And Discussion

Data Retrieval

Statistik Data Anak RS HM Mawardi

Kelompok Usia	Total Anak	Perempuan (P)	Laki-laki (L)	Range BB (kg)	Range TB (cm)	Range IMT/BMI
6-7 Tahun	100	49	51	6 - 30	81 - 129	9.1 - 21.1
8-9 Tahun	81	43	38	19 - 42	114 - 144	9.6 - 23.7
10-12 Tahun	76	36	40	24 - 64	123 - 164	12.6 - 25.0
Total	**257**	**128**	**129**	-	-	-

Figure 2. Data Collection Results

Figure 2 shows the results of data collection conducted at H.M. Mawardi Krian Hospital, Sidoarjo, using two methods: access through the hospital's internal inpatient registry database system and direct observation of the hospital's inpatient registry, resulting in a total of 257 data points, consisting of 128 girls and 129 boys. This population was grouped into three age categories: 100 children aged 6–7 years, 81 children aged 8–9 years, and 76 children aged 10–12 years. In addition to showing the frequency distribution by sex in each group, this table also includes children's anthropometric values, including a range of overall weight (BW) from 6 to 64 kg, a range of height (H) from 81 to 164 cm, and a range of Body Mass Index (BMI) from 9.1 to 25.0..

ISSN 2828-9463

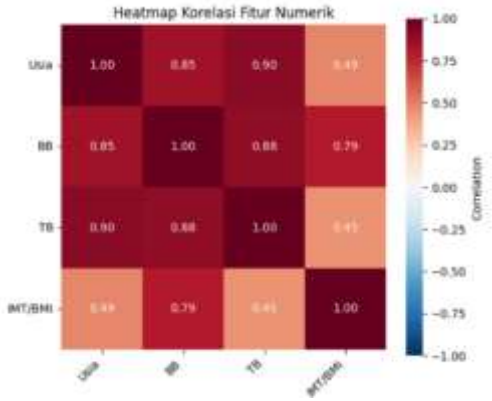


Figure 3. Heatmap of Feature Correlation

The initial data collected included attributes such as Name, Weight (BW), Height (H), Age, Gender, and Body Mass Index (BMI). Correlations between variables showed a very strong relationship between Height and Age (0.90), followed by the relationship between Weight and Height (0.88). Meanwhile, BMI showed a more moderate correlation with Age (0.49) and Height (0.45), indicating that children's body proportions do not always change linearly with physical growth.

Data Preprocessing

The next step was data preprocessing to ensure data quality before the modeling stage. Data preprocessing in this study included several steps, including:
Gender Encoding



Figure 4. Gender Encoding Results

Categorical gender data (M and F) were converted to numeric values (0 for Male and 1 for Female) because the Gradient Boosting Machine (GBM) model cannot directly process text data.
Feature Engineering



Figure 5. Feature Engineering Results

Adding lower and upper BMI thresholds based on health standards and calculating the difference, or "margin," between the child's actual BMI and these limits to provide specific information about deviations in the child's physical condition.

Binary Transformation



No	Nama	Usia	BB	TB	Target_Binary
1	Fahreza Subandono	12	24	138	Tidak Baik
2	Haura Aleska Ayudhari	8	20	122	Baik
3	Elvano Alfaroky Yudistira	7	25	113	Tidak Baik
4	Naranoyan Randka	8	17	105	Baik
5	Syarifah Rodhiyah	6	6	81	Tidak Baik
6	M. Farhan Hamdani	11	36	153	Baik
7	Gibran Triasdo	7	21	123	Baik
8	Athanasaka Rayan A	8	15	105	Baik
9	M Athallah Haaland	9	19	114	Baik
10	Rani Zahira Hibatillah	8	40	138	Baik

Figure 6. Binary Transformation

The target variable, which originally had three BMI status categories, was simplified into a binary classification: "Good" for the Ideal category, and "Not Good" for the Undernourished and Overnourished categories.



Feature Selection

Figure 7. Feature Selection Results

Determining the features to be used as input (such as Age, Weight, Height, BMI, and other feature engineering results) and removing identifying data such as names or serial numbers to prevent bias.

Data Splitting

The dataset is divided into training data (X train and y train) with a proportion of 80% and test data (X test and y test) with a proportion of 20%. The parameter test_size=0.2 is used to lock in data randomization. This ensures that the data division remains consistent and identical each time it is run.

Balancing using SMOTE

```
Sebelum SMOTE: Counter({'Ideal': 169, 'Gizi Kurang': 27, 'Gizi Berlebih': 9})
Sesudah SMOTE: Counter({'Ideal': 169, 'Gizi Kurang': 169, 'Gizi Berlebih': 169})
```

Figure 8. Balancing Results Using SMOTE

Because the "Bad" category was smaller, the Synthetic Minority Oversampling Technique (SMOTE) was used to balance the minority data to match the majority category (Chawla, 2022).

Standardization



Figure 9. Standardization Results

Because the features have different units and ranges of values, StandardScaler (Pedregosa, 2011) was used to transform each feature value into a standard distribution (mean zero and variance one) to prevent any single feature from dominating the model weight calculation.

Gradient Boost Machine (GBM) Classification Modeling

Constructs a series of decision trees incrementally to correct prediction errors from the previous model. The model configuration uses 200 trees with a learning rate of 0.2.

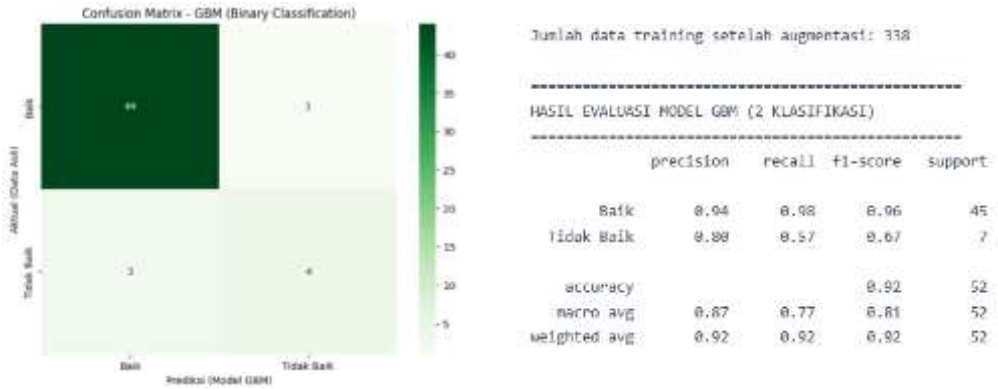


Figure 11. Gradient Boosting Machine Results

The Gradient Boosting Machine (GBM) model demonstrated impressive and stable performance in classifying children's nutritional status. It achieved an accuracy rate of 92% (0.92) across the test data, indicating its reliability in distinguishing nutritional status with minimal error. For the 'Good' category, the model performed very well, with a precision of 0.94 and a recall of 0.98, indicating that it almost never missed cases of Good nutritional status in its predictions. Meanwhile, for the 'Not Good' category, the model demonstrated a fairly accurate detection capability, with a precision of 0.80 (80%) and a recall of 0.57, indicating that the model was able to provide fairly accurate predictions in identifying cases of sub-optimal nutrition. Overall, with a weighted average of 0.92, the GBM model demonstrated consistent and superior generalization capability in classifying children's nutritional status.

4. Conclusion

Based on the research and analysis conducted on the application of the Gradient Boost Machine algorithm for child nutritional status classification at Al-Islam Hospital, Krian, the following conclusions can be drawn:

1. The implementation of the Gradient Boosting Machine (GBM) algorithm demonstrated excellent performance in classifying nutritional status. This was demonstrated by achieving the highest accuracy value of 0.92 (92%), a precision level of 0.80, a recall of 0.57, and an f1-score of 0.67. The

GBM's ability to perform incremental learning (boosting) has proven effective in minimizing classification errors.

2. Overall, the application of the Gradient Boosting Machine (GBM) has proven to produce superior and stable performance in the task of classifying child nutritional status. This confirms that the boosting model-based approach has significant potential in supporting a health analytics system based on medical record data for early detection of nutritional problems. Therefore, it can be concluded that the use of the GBM method is one of the most effective approaches in improving the accuracy and efficiency of the child nutritional status classification system at Al-Islam Hospital H.M. Mawardi Krian, and has the potential to become the basis for the development of a medical decision support system in handling cases of malnutrition and obesity in children more precisely.

5. References

Aulia, N. R. (2021). *Peran pengetahuan gizi terhadap asupan energi, status gizi dan sikap tentang gizi remaja*. *Jurnal Ilmiah Gizi dan Kesehatan (JIGK), 2*(2), 31–35. https://doi.org/10.46772/jigk.v2i02.454

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Christian, M. S., & Diana, A. T. (2023). Analisis faktor risiko stunting dan obesitas pada anak usia sekolah dasar. *Jurnal Gizi Klinik Indonesia, 19*(3), 112–121. https://doi.org/10.22146/ijcn.71204

Dinata, R. K., Hasdyna, N., & Retno, S. (2022). Optimasi algoritma klasifikasi ensemble untuk pemetaan penerima bantuan sosial. *Jurnal Edukasi dan Penelitian Informatika (JEPIN), 8*(2), 204–211. https://doi.org/10.26418/jp.v8i2.53912

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Kementerian Kesehatan Republik Indonesia. (2020). *Peraturan Menteri Kesehatan Republik Indonesia Nomor 2 Tahun 2020 tentang standar antropometri anak*. Kementerian Kesehatan Republik Indonesia.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7*, Article 21. https://doi.org/10.3389/fnbot.2013.00021

Nomira, S. (2024). Relevansi gizi dan kesehatan. *Public Health Journal, 1*(2), 1–16.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pusat Riset Kesehatan BRIN. (2023). *Tantangan dan strategi penanganan obesitas pada anak dan remaja*. Badan Riset dan Inovasi Nasional. https://brin.go.id/reviews/118905/tantangan-dan-strategi-penanganan-obesitas-pada-anak-dan-remaja

Sanjaya, H. L., & Faisal, M. R. (2023). Keterbatasan metode statistik konvensional dalam penentuan status nutrisi klinis. *Jurnal Teknologi Informasi dan Kesehatan, 8*(1), 22–30. https://doi.org/10.34012/jtik.v8i1.2941

Sari, B. W., & Prabowo, D. (2025). Analisis perbandingan prediksi harga rumah dengan *Random Forest*, *Gradient Boosting*, dan *XGBoost*. *4*(1), 42–51.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Suryana, S. E., Warsito, B., & Suparti. (2021). Penerapan *gradient boosting* dengan *Hyperopt* untuk

memprediksi keberhasilan telemarketing bank. *Jurnal Gaussian, 10*(4), 617–623. https://doi.org/10.14710/j.gauss.v10i4.31335

Zhang, J., Li, Y. M., & Park, J. H. (2024). Comparative analysis of ensemble learning methods for structured socio-economic data classification. *Journal of Big Data Analytics, 12*(3), 145–158. https://doi.org/10.1186/s40537-024-00891-2