Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram
https://e-journal.undikma.ac.id/index.php/prismasains/index
e-mail: prismasains.pkpsm@gmail.com

July 2026. Vol. 14, No. 3
p-ISSN: 2338-4530
e-ISSN: 2540-7899
pp. 2016-2032

Optimization of the SVM Algorithm with Chi-Square and SMOTE for High-Dimensional Stunting Data in Samarinda

[1]*Taghfirul Azhima Yoga Siswa, [2]Naufal Azmi Verdikha
Department of Informatics, Muhammadiyah University of East Kalimantan, Samarinda, Indonesia
*Corresponding Author e-mail: tay758@umkt.ac.id

Received: June 2026; Revised: July 2026; Published: July 2026

Abstract

Stunting is still considered a serious problem and has become a focus of government attention in the national research priorities for 2020-2024, with a stunting prevalence of 21.6% in 2022. The city of Samarinda ranked second highest after the Kutai Kartanegara district in East Kalimantan Province in 2022, with a percentage of 25.3%. Based on previous data mining research trends, the use of classification methods such as KNN, Naïve Bayes, Random Forest, Neural Network, CART, and SVM has generally produced quite high accuracy but still focuses on low-dimensional data, which can lead to significant information loss, potential overfitting, and difficulty in interpretation. Whereas in research topics related to high-dimensional stunting data, the majority still yield low accuracy. This is further supported by the still prevalent class imbalance found in other studies, which can affect the accuracy and recall values of the built model's performance. The objective of this research is to apply the Support Vector Machine (SVM) algorithm with Chi-Square feature selection and the Synthetic Minority Over-sampling Technique (SMOTE) to address high-dimensional stunting data and handle class imbalance. The dataset for this research is sourced from the Samarinda City Health Office, consisting of 26 community health centers with 20 attributes and 102,534 records. The division of training and testing data uses the k-fold cross-validation technique with k=10. The research results show the model's performance is very good, with an accuracy of 96.6%, supported by precision, recall, and f1-score values of 97% each.

Keywords: Support vector machine; Chi-square; Synthetic minority over-sampling technique; High-dimensional; Stunting

How to Cite: Siswa, T. A. Y., & Verdikha, N. A. (2026). Optimization of the SVM Algorithm with Chi-Square and SMOTE for High-Dimensional Stunting Data in Samarinda. *Prisma Sains: Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, *14*(3), 2016–2032. https://doi.org/10.33394/j-ps.v14i3.21848

https://doi.org/10.33394/j-ps.v14i3.21848

INTRODUCTION

Stunting remains a fundamental public-health problem because it reflects chronic or recurrent constraints on linear growth during the most sensitive period of child development. The World Health Organization defines stunting as a height-for-age value below minus two standard deviations from the median of the WHO Child Growth Standards; therefore, stunting is not merely a description of short stature but a population-level marker of cumulative nutritional, infectious, environmental, and socioeconomic deprivation (World Health Organization, 2025). In Indonesia, the 2022 Indonesian Nutritional Status Survey reported a national stunting prevalence of 21.6%, meaning that approximately one in five children under five was affected (Health Development Policy Agency, 2022). The burden was even more pronounced in Samarinda, where the prevalence reached 25.3% and placed the city among the areas with a substantial nutritional burden in East Kalimantan (Health Development Policy Agency, 2022). This local situation is strategically important because Samarinda is closely connected to the development corridor surrounding Indonesia's new capital and is expected to experience rapid demographic, economic, and health-service transitions. Persistent stunting

can compromise physical growth, cognitive development, educational achievement, and later-life human capital, while its determinants extend beyond food intake to maternal health, infection, sanitation, poverty, education, and access to services (Cameron et al., 2021; Pizzol et al., 2021; Victora et al., 2021). Consequently, timely and methodologically valid identification of children at risk is essential for directing preventive services and allocating limited public-health resources.

A central scientific issue, however, is that the terms classification, screening, diagnosis, and prediction are not interchangeable. Classifying a child's current stunting status from measurements recorded at the same visit is a contemporaneous classification task; screening seeks to identify probable cases when the definitive anthropometric indicator is unavailable; and prospective prediction estimates the probability of stunting at a later age from information available earlier in life. Each task requires a different outcome definition, predictor set, temporal structure, and validation strategy. Recent Indonesian research, for example, has developed prospective models that use newborn and household characteristics to estimate stunting at two years of age, which is conceptually different from reproducing a status already determined at the same measurement occasion (Azriani et al., 2025). The present study is therefore framed as a retrospective classification of recorded stunting status in routine health-center data, rather than as a causal explanation or a validated forecast of future stunting. This distinction is necessary so that a high numerical score is not interpreted automatically as evidence of early prediction or clinical readiness.

Data-mining and machine-learning approaches have increasingly been used to support the classification of child nutritional status. Indonesian studies have applied K-Nearest Neighbor with feature elimination, Naive Bayes with cross-validation, Random Forest, CART, and combinations of feature-selection methods with Support Vector Machine (SVM), often reporting accuracy values above 90% in their respective datasets (Arisandi et al., 2022; Hakimah et al., 2022; Lonang & Normawati, 2022; Perdana et al., 2021; Sasmita et al., 2022). International evidence has similarly compared logistic regression, Random Forest, gradient boosting, SVM, and related algorithms using demographic and health-survey data from Bangladesh and Rwanda (Khan et al., 2021; Ndagijimana et al., 2023; Rahman et al., 2021). These studies demonstrate the potential of machine learning to capture nonlinear relationships and interactions among child, maternal, household, and environmental factors. Nevertheless, reported accuracy values cannot be compared directly across studies because performance is influenced by population characteristics, outcome prevalence, feature availability, class definition, missing-data handling, and validation design. A lower score in a study with an untouched imbalanced test set may represent more credible generalization than a higher score obtained after information from the evaluation data has influenced preprocessing. Thus, the unresolved problem is not simply how to maximize accuracy, but how to obtain performance estimates that remain valid under realistic data conditions.

The earlier version of this manuscript positioned the problem primarily as a contrast between low-dimensional and high-dimensional data. That distinction requires greater precision. In the conventional statistical sense, high-dimensional data generally refer to settings in which the number of predictors is very large, sometimes approaching or exceeding the number of observations; the present administrative dataset, which contains approximately 20-21 original variables and more than 27,000 retained records, does not satisfy that strict definition. The term is used here operationally to describe a heterogeneous multivariable health dataset that combines demographic information, measurement metadata, anthropometric values, nutritional indices, and service-location variables, rather than a classical p-greater-than-n problem. Using too few variables can omit meaningful determinants, whereas including redundant, weak, or target-derived variables can increase computational burden, instability, and the risk of overfitting (James et al., 2021; Velliangiri et al., 2019). Feature selection is therefore relevant not because a small predictor set is intrinsically deficient, but because a

parsimonious model should retain information that is independently informative, reproducible, and available at the intended point of use.

Previous stunting studies have identified potentially informative variables such as child age and sex, birth weight and length, maternal and paternal age or height, parental education, breastfeeding, vitamin A supplementation, recent illness, sanitation, household wealth, and geographic context (Azriani et al., 2025; Khan et al., 2021; Ndagijimana et al., 2023; Rahman et al., 2021). These variables can be useful when they precede the outcome or provide information that is not mathematically embedded in the stunting definition. By contrast, height-for-age categories and height-for-age z-scores are normally used to define stunting itself. If the target label in a dataset is derived from the same height-for-age z-score recorded at the same visit, then using that z-score or a directly derived category as a predictor creates target leakage: the model learns the rule used to construct the outcome rather than an independent pattern that could support screening or prediction. Data leakage is a well-established source of overoptimistic machine-learning results and can occur when outcome information, validation-fold information, or preprocessing statistics enter model training inappropriately (Kapoor & Narayanan, 2023). For this reason, the scientific value of a stunting model depends not only on the algorithm selected but also on a transparent target definition, separation of outcome-defining variables from predictors, and a validation design consistent with the intended use.

SVM was selected because it can construct a maximum-margin decision boundary and, through a nonlinear kernel such as the radial basis function, model complex class separation in multivariable data (James et al., 2021). Chi-Square feature selection was included to rank candidate attributes according to their association with the target and to reduce the predictor set, while the Synthetic Minority Over-sampling Technique (SMOTE) was introduced to address the under-representation of stunted children by generating synthetic minority-class observations within the training data (Chawla et al., 2002). Prior stunting research has shown that SVM can perform competitively and that feature selection may alter classification performance, although the best feature-selection strategy is dataset-dependent (Hakimah et al., 2022; Khan et al., 2021). The combination of SVM, Chi-Square, and SMOTE is therefore a methodological strategy rather than novelty by itself. Its validity depends on applying encoding, scaling, supervised feature selection, and SMOTE only within each training fold, followed by evaluation on untouched observations that retain the original class prevalence. In an imbalanced health problem, accuracy alone is insufficient; class-specific recall, precision, F1-score, specificity, balanced accuracy, and discrimination should be considered to determine whether improved overall performance is accompanied by reliable detection of stunted children.

On the basis of this state of the art, three interrelated gaps remain. First, evidence is limited regarding the performance of machine-learning classification on a large routine dataset collected across 26 community health centers in Samarinda, where data quality, repeated measurements, and heterogeneous service contexts may differ from curated survey datasets. Second, it remains uncertain whether routinely available, non-diagnostic variables can identify stunting while maintaining sensitivity under the original class distribution and avoiding target and preprocessing leakage. Third, few local studies have isolated the contribution of individual pipeline components through baseline and ablation comparisons, making it difficult to determine whether an apparent improvement originates from SVM, feature selection, resampling, or the evaluation procedure itself. These gaps shift the emphasis from reporting a single high accuracy value to demonstrating a transparent, leakage-aware, and practically interpretable classification process.

Accordingly, the substantive contribution sought in this study is not the simple juxtaposition of three established techniques, but the use of a large multicenter administrative dataset from Samarinda to examine how feature reduction and class-imbalance handling interact with an SVM classifier in a locally relevant setting. The study aims to apply an SVM

model with Chi-Square feature selection and SMOTE to classify recorded stunting status, identify the attributes retained by the selection procedure, and evaluate internal classification performance using 10-fold cross-validation and confusion-matrix-based metrics. The unit of analysis is the available child measurement record after data preparation, and the scope is an internal retrospective algorithmic assessment rather than prospective risk prediction, causal inference, external validation, or immediate deployment as a clinical decision-support system. The findings are expected to provide an empirical basis for improving subsequent leakage-free modeling, comparator analyses, and external or temporal validation before the model can be considered for operational screening.

METHOD

This study was conducted through several stages, including problem identification, literature review, dataset collection, data preparation, feature selection, model development, and performance evaluation using a confusion matrix. The study employed a classification approach in which the target class was available in the historical dataset, allowing the classifier to learn the relationship between predictor attributes and the observed class labels and subsequently assign observations to the appropriate class based on the learned pattern (James et al., 2021).

Support Vector Machine (SVM) is a machine learning method that operates based on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that separates two classes in the input space. SVM is a learning system that uses a hypothesis space in the form of linear functions in a high-dimensional feature space, trained with a learning algorithm based on optimization theory by implementing a learning bias derived from statistical learning theory. The goal of the SVM algorithm is to find a hyperplane in an N-dimensional space (where N is the number of features) that separates observations while maximizing the margin between classes (James et al., 2021).

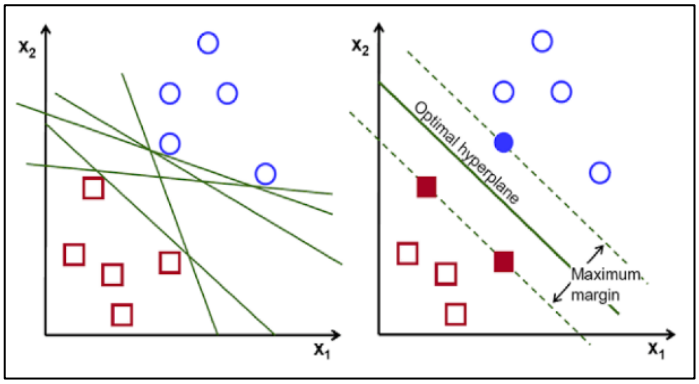


Figure. 1 (a) Hyperlane Non Optimal (b) Hyperlane Optimal

Figure 1(a) shows several possible hyperplane choices for the dataset, while Figure 1(b) is the hyperplane with the maximum margin. Although in figure (a) any hyperplane could also be used, the hyperplane with the maximum margin will provide better generalization for the classification method. The concept of SVM classification can be explained as the search for an optimal hyperplane that separates two classes in the input space with the largest feasible margin (James et al., 2021). To obtain the hyperplane in SVM, the following equation can be used.

$$(w \,.\, x_i) + b = 0 \qquad (1)$$

In the data xi, which belongs to class -1, it can be formulated as follows:

$$(w \,.\, x_i) + b \leq 1, y_i = -1 \qquad (2)$$

Meanwhile, the data xi, which belong to class +1, can be formulated as follows:

$$(w \,.\, x_i) + b \geq 1, y_i = 1 \qquad (3)$$

Explanation:
Xi = data to-i
W = the weight value of the support vector that is perpendicular to the hyperplane
b = the bias value
Yi = the class of data to - i

In the classification process with SVM, it is often encountered that the linear kernel does not work optimally, resulting in poor classification results for the data. This can be addressed by using a non-linear kernel with the help of the kernel trick. By utilizing the kernel trick, the input data are implicitly mapped into a higher-dimensional feature space in which nonlinear class boundaries can be represented by a linear separating hyperplane (James et al., 2021).

SVM Kernel Functions

Kernel	Mercer kernel	Purpose
RBF	$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$	Nonlinear similarity (σ sets width)
Linear	$K(x_1, x_2) = x_1^\top x_2$	Linear separation
Polynomial	$K(x_1, x_2) = (x_1^\top x_2 + 1)^\rho$	Degree-ρ interactions
Sigmoid	$K(x_1, x_2) = \tanh(\beta_0 x_1^\top x_2 + \beta_1)$	Neural-like mapping

Figure 2. SVM Kernel Equation

Previous machine-learning studies of child malnutrition have shown that nonlinear classifiers, including SVM configurations, can capture complex relationships among predictors, although the most effective model remains dataset-dependent (Khan et al., 2021; Rahman et al., 2021). Therefore, the RBF kernel will be applied in this research. The application of the SVM algorithm with Chi-Square feature selection and the SMOTE technique is carried out to address high-dimensional stunting data and class (label) imbalance.

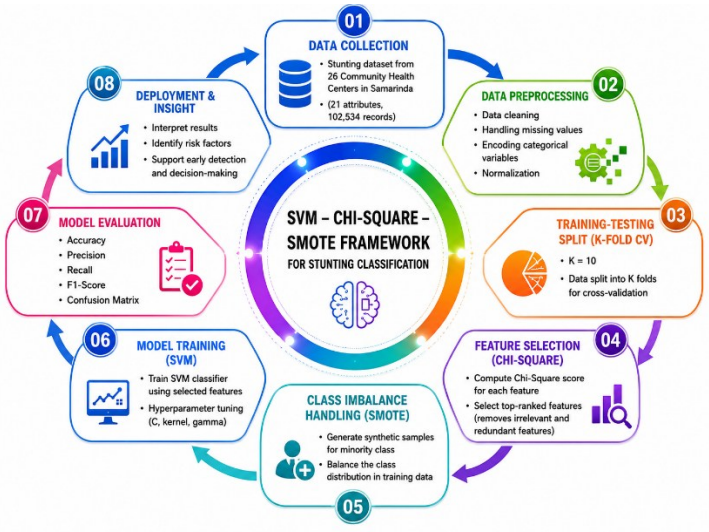


Figure 3. Flowchart of SVM-Chi Square-SMOTE

The workflow of the proposed research methodology systematically begins with the stages of problem identification and literature review, followed by data collection and preparation. To mitigate bias that may arise due to class distribution imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) approach is applied at the preprocessing stage, which is then optimized thru the Chi-Square feature selection method to retain the most statistically relevant predictive attributes. The modeling and testing stage is

executed using a 10-fold Cross-Validation scheme to ensure robust training and testing data partitions while minimizing variance and bias (overfitting). At the core stage, the Support Vector Machine (SVM) algorithm is implemented as the main classification model, where the effectiveness and reliability of its predictions are comprehensively evaluated at the final stage using performance metrics derived from the Confusion Matrix.

RESULTS AND DISCUSSION

Dataset Collection

The initial stage of this research process is data preparation or the preparation of stunting case data in Samarinda City. The dataset was obtained from 26 community health centers within the Samarinda City Health Office. The obtained data has 21 columns and a total of 102,534 records. The data obtained is as follows.

Table 1. Dataset Description.

No	Attribute	Data Type	Remarks
1	National ID Number	*string*	National Identification Number
2	Name	*string*	Name
3	Sex	*string*	Sex
4	Date of Birth	*date*	Date of Birth
5	Parent's Name	*string*	Parent's Name
6	Province	*string*	Province
7	Regency/City	*string*	Regency or City
8	Subdistrict	*string*	Subdistrict
9	Community Health Center	*string*	Community Health Center Location
10	Integrated Health Post	*string*	Integrated Health Post Location
11	Total Measurements	*integer*	Total Measurements
12	Measurement Date	*integer*	Measurement Date
13	Weight	*integer*	Weight
14	Height	*integer*	Height
15	WFA	*numeric*	Weight-for-Age
16	WAZ	*numeric*	Weight-for-Age Z-Score
17	HFA	*numeric*	Height-for-Age
18	HAZ	*numeric*	Height-for-Age Z-Score
19	WFH	*numeric*	Weight-for-Height
20	WHZ	*numeric*	Weight-for-Height Z-Score

Table 1 presents the details of the initial dataset structure that was acquired, consisting of 20 attributes with various data types (string, date, integer, and numeric). Conceptually, these raw attributes can be classified into three main groups of information: (1) demographic and administrative data that includes patient identity and healthcare facility location (such as the national identification number, sex, and community-health-center and integrated-health-post locations); (2) observation metadata that records the time and frequency of measurements; and (3) anthropometric metrics and nutritional indices that contain core clinical variables, including weight, height, and specific calculations such as Z-scores for weight-for-age (WFA), height-for-age (HFA), and weight-for-height (WFH). This set of attributes represents high-dimensional data before entering the preprocessing and feature selection stages.

Data Preparation

Data Selection

At this stage, data is collected by selecting relevant attributes or columns for the research, while attributes deemed irrelevant will be removed. In Table 2, the initial data obtained from the Samarinda City Health Office has 21 columns, then there are 8 columns that are considered irrelevant for predicting stunting in children, so the columns originally numbered 21 are reduced to 12 columns used as attributes and 1 column used as the target or class after the data selection process.

Table 2. Data Selection

No	Initial Attribute	Selected Attribute	Remarks
1	National ID Number	Name	-
2	Name	Sex	Attribute
3	Sex	Subdistrict	-
4	Date of Birth	Community Health Center Location	-
5	Parent's Name	Village or Urban Ward	-
6	Province	Integrated Health Post Location	-
7	Regency/City	Age at Measurement	-
8	Subdistrict	Child Measurement Date	-
9	Community Health Center	Weight	Attribute
10	Integrated Health Post	Height	Attribute
11	Total Measurements	Mid-Upper Arm Circumference	Attribute
12	Measurement Date	Weight-for-Age	Attribute
13	Weight	Weight-for-Age Z-Score	Attribute
14	Height	Height-for-Age Z-Score	Attribute
15	WFA	Weight-for-Height	Attribute
16	WAZ	Weight-for-Height Z-Score	Attribute
17	HFA	Child Weight Gain	Attribute
18	HAZ	Supplementary Food Received by the Child	Attribute
19	WFH	Number of Vitamin A Supplements Received by the Child	Attribute
20	WHZ	Developmental Prescreening Questionnaire	Attribute

Table II illustrates the stages of data selection, where relevant attributes or columns for the research are chosen, while attributes that do not directly contribute will be removed. In this process, there are 8 initial columns such as identity and demographic location (e.g., Name, Subdistrict, and Health Center Location), that are eliminated because they are considered irrelevant for predicting stunting in children. Conversely, this selection results in 12 columns containing anthropometric data and children's health history (such as Weight, Height, and Z-Score) that are retained and specifically used as the main modeling attributes.

Data Cleaning

The data cleaning performed on the stunting dataset of Samarinda City involved removing data with #N/A values (no value is available) or missing values, as well as duplicated data. In this process, a total of 69,912 records were cleaned for duplicated data, and a total of 5,602 records were cleaned for #N/A values, resulting in 27,020 records remaining after the cleaning process.

Data Transformation

The data transformation step applied is converting categorical attribute values into numerical form. This action is necessary because the sklearn library can only accept attributes with numerical values. Some of the transformed data includes gender, weight gain, weight according to age, and weight according to height. The data transformation process carried out can be seen in Table 3 for the data before transformation and in Table 4 for the data after transformation. The data transformation process was carried out with the help of a Python library.

Table 3. Dataset Before Transformation

No	Sex	WFA	WFH	Weight Gain
1	M	Overweight Risk	Risk of Overnutrition	O
2	M	Normal Weight	Normal Nutritional Status	T
3	M	Underweight	Undernutrition	N
…	…	…	…	…
27018	M	Normal Weight	Normal Nutritional Status	T
27019	F	Normal Weight	Normal Nutritional Status	N

No	Sex	WFA	WFH	Weight Gain
27020	M	Normal Weight	Normal Nutritional Status	O

Table 3 and Table 4 illustrate the data-transformation stage, in which categorical attribute values are encoded as numerical representations. This step is required because machine-learning libraries such as sklearn process numerical inputs. In Table 3 (Dataset Before Transformation), variables such as sex, weight-for-age (WFA), weight-for-height (WFH), and weight-gain history are represented using categorical labels, such as "M/F" or "Normal Weight." Table 4 (Dataset After Transformation) presents the corresponding integer encodings; for example, "M" in the Sex variable is encoded as 0 and "F" as 1. Through this process, the dataset becomes computationally valid and ready for model training.

Table 4. Dataset After Transformation

No	Sex	WFA	WFH	Weight Gain
1	0	2	5	2
2	0	0	0	3
3	0	1	2	1
…	…	…	…	…
27018	0	0	0	3
27019	1	0	0	1
27020	0	0	0	2

Data Balancing

The data transformation step applied is converting categorical attribute values into numerical form. This action is necessary because when using the sklearn library, it can only accept attributes with numerical values. Some of the transformed data include gender, weight gain, weight by age, and weight by the dataset used in this study, which has a class imbalance, with 22,458 data points for children who do not experience stunting, while only 4,562 data points for children who do experience stunting.

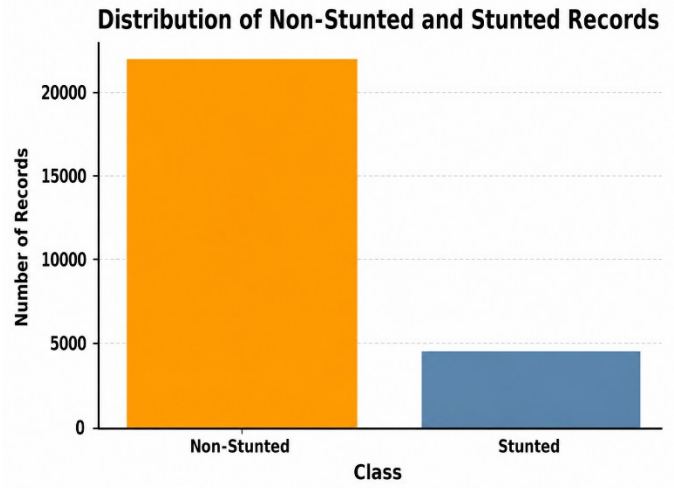


Figure 4. Visualization of classes in the stunting dataset

In Figure 5, the comparison of the number of stunted and non-stunted children classes is shown. This will result in the model's performance on the majority class (non-stunting) usually being better than the minority class (stunting) which has a smaller number. The SMOTE (Synthetic Minority Over-sampling Technique) oversampling technique is used to balance the amount of data between the majority class (non-stunting) and the minority class (stunting) by generating synthetic data in the minority class until its quantity matches that of the majority class. The application of SMOTE in this research uses the SMOTE library in the Python programming language.

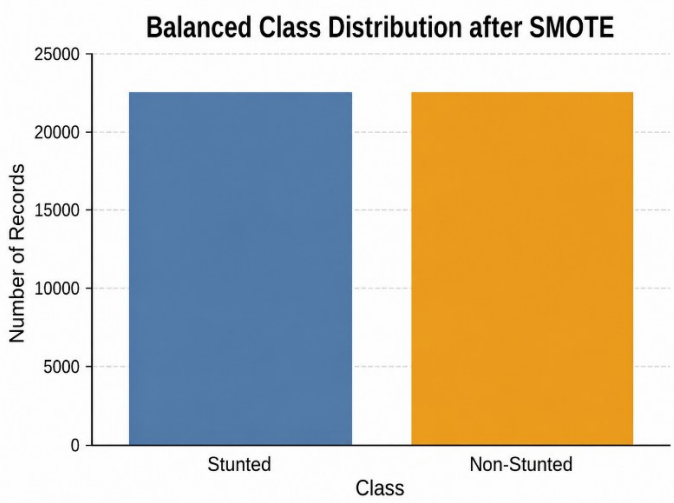
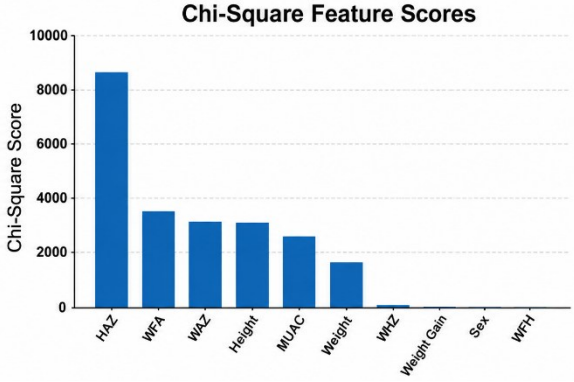


Figure 5. Visualization of Balanced Class Resulting from SMOTE Technique

Feature Selection

In the feature selection stage, a ranking is performed on the existing attributes, starting from the attribute with the greatest influence (receiving the highest chi2 value) on the prediction result, which is ranked 1, to the attribute with the least influence or no influence on the prediction (receiving the lowest chi2 value), which is ranked 10. The process of ranking the attributes uses the sklearn library. Here are the feature ranking results using Chi Square.




```
chi2_scores.sort_values(ascending=False)
```

Feature	Chi-Square
HAZ	8692.633678
WFA	3538.736155
WAZ	3155.719473
Height	3117.852136
MUAC	2658.031559
Weight	1678.581347
WHZ	69.916099
Weight Gain	7.776799
Sex	6.679535
WFH	2.648965

Figure 6. Feature Selection with Chi-Square

Figure 6 illustrates the results of Chi-Square feature selection. The bar chart compares the Chi-Square scores of the candidate attributes, while the accompanying panel reports their numerical values. HAZ has the highest score (8692.63), followed by WFA, WAZ, Height, MUAC, and Weight. In contrast, WHZ, Weight Gain, Sex, and WFH have substantially lower scores, indicating comparatively limited contributions to the classification outcome and supporting their exclusion from the subsequent SVM modeling stage.

Table 5. Chi-Square Ranking Results for the Samarinda Stunting Dataset

Attribute	Chi-Square Value	Ranking
HAZ	8692	1
WFA	3538	2
WAZ	3155	3
Height	3117	4
MUAC	2658	5
Weight	1678	6
WHZ	69	7
Weight Gain	7	8
Sex	6	9
WFH	2	10

Based on the attribute ranking results shown in Table 5, it was determined that the attributes ranked 1-6 are used as attributes in the SVM modeling because they have high Chi Square values. Therefore, only 6 attributes are used in the modeling, namely HAZ, WFA, WAZ, Height, MUAC, and Weight.

Model Evaluation

This research uses the k-fold cross-validation technique to divide the training and testing data. The value of k used is k=10, while the attributes used after feature selection using Chi Square are 6 attributes, and previously the SMOTE technique was used to address the class imbalance present in the stunting dataset of Samarinda City. The results of the model performance testing using the confusion matrix technique can be seen as follows.

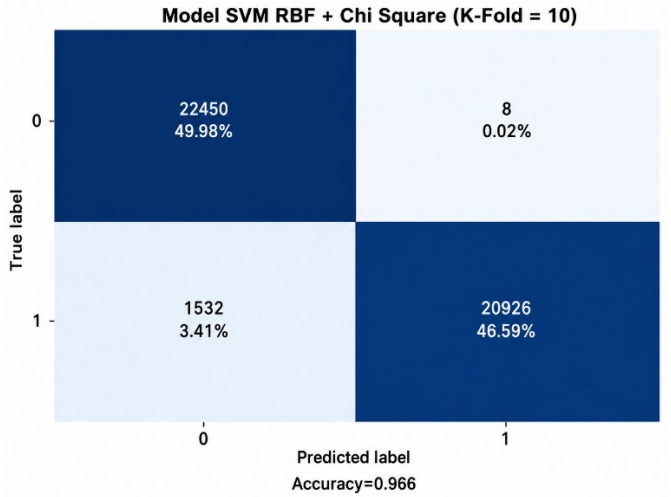


Figure 7. Visualization of the Confusion Matrix

Figure 7 presents a visualization of the confusion matrix that evaluates the final performance of the Support Vector Machine (SVM) model with a Radial Basis Function (RBF) kernel thru a 10-fold cross-validation testing scheme. Based on the computational results, the model optimized with the SMOTE technique and reduced to 6 dominant attributes via Chi-Square demonstrates excellent classification capabilities, achieving an overall accuracy of 96.6%. Specifically, the model accurately predicted 22,450 observations as class 0 (True Negative, representing 49.98%) and 20,926 observations as class 1 (True Positive, representing 46.59%). The misclassification rate is very minimal, as indicated by the low number of False Positives at 8 observations (0.02%) and False Negatives at 1,532 observations (3.41%). The distribution of this matrix empirically confirms that the proposed model has a reliable level of precision and sensitivity in proportionally distinguishing classes.

Discussion

The proposed SVM-RBF pipeline, using Chi-Square feature selection and SMOTE, produced an apparent overall accuracy of 96.6% after the predictor set was reduced to HAZ, WFA, WAZ, height, MUAC, and weight. At face value, this result indicates that the selected variables contain strong information for separating the recorded stunting and non-stunting labels in the analyzed dataset. The large number of records obtained from 26 community health centers is an important strength because routine administrative data capture variation in service locations, measurement practices, and child characteristics that is usually absent from small, highly curated datasets. Nevertheless, the result should be interpreted as internal retrospective classification performance under the reported preprocessing conditions, rather than as evidence of prospective prediction, causal explanation, or readiness for clinical deployment. Contemporary guidance for health prediction models emphasizes that the scientific value of a high score depends on a clearly defined intended use, transparent predictor timing,

representative evaluation data, and separation between model development and performance assessment (Collins et al., 2024; Moons et al., 2025).

The confusion matrix provides a more informative picture than the rounded statement that accuracy, precision, recall, and F1-score were all 97%. Based on the reported counts, the sensitivity for the stunting class is 93.18% [20,926/(20,926 + 1,532)], the positive predictive value is 99.96% [20,926/(20,926 + 8)], the specificity is 99.96% [22,450/(22,450 + 8)], and the stunting-class F1-score is approximately 96.45%. Thus, 1,532 of the 22,458 records labeled as stunting were classified as non-stunting, corresponding to a false-negative proportion of 6.82% among actual stunting records. This difference is substantively important because, in a screening context, missed cases may delay nutritional assessment and intervention even when the overall accuracy appears excellent. Studies of imbalanced clinical classification consistently show that aggregate accuracy can conceal clinically meaningful minority-class errors; sensitivity, specificity, precision-recall performance, calibration, and uncertainty intervals should therefore be reported separately and interpreted according to the consequence of each error type (Ke et al., 2024; Mosquera et al., 2024).

A critical issue emerges from the size of the reported confusion matrix. The four cells sum to 44,916 observations, exactly twice the original majority-class count of 22,458 and substantially more than the 27,020 real records retained after cleaning. This numerical pattern strongly suggests that evaluation was performed on the fully balanced dataset generated by SMOTE, rather than on untouched observations preserving the original prevalence of 22,458 non-stunting and 4,562 stunting records. If this interpretation is correct, the reported precision and accuracy do not estimate performance under the real Samarinda class distribution, because synthetic observations are part of the evaluation population and the event prevalence has been changed from approximately 16.9% to 50%. Resampling is intended to modify only the training data; validation data should remain unaltered so that discrimination, predictive values, and calibration reflect the intended application population. Recent clinical modeling research found that the benefit of SMOTE is strongly dependent on the learner and evaluation metric, and explicitly retained the original event rate in the validation set because resampling the evaluation data produces misleading estimates (Ke et al., 2024). Consequently, the 96.6% accuracy should be described as performance under a balanced internal evaluation until the pipeline is repeated with SMOTE restricted to each training fold and predictions pooled only from untouched real records.

The very large Chi-Square score for HAZ (8,692), compared with WFA (3,538), WAZ (3,155), and the remaining variables, is not surprising from a clinical or mathematical perspective. Stunting is conventionally defined using height-for-age, specifically a height-for-age z-score below -2 standard deviations from the WHO growth standard. Therefore, if the target label was derived from HAZ measured on the same date, including HAZ as a predictor gives the model direct access to the variable used to construct the outcome. In that situation, the highest ranking does not reveal a new risk factor; it demonstrates circular classification or target leakage, because the algorithm is reproducing the diagnostic rule embedded in the dataset. Height and a height-for-age category may also encode closely related outcome information, especially when age at measurement is available elsewhere in the record. Leakage through outcome-defining predictors can generate performance that is numerically impressive but does not generalize to a setting in which the outcome is genuinely unknown (Kapoor & Narayanan, 2023; Bernett et al., 2024). Empirical work has further shown that leakage through feature selection and repeated observations can substantially inflate cross-validated performance and alter the apparent importance of predictors (Rosenblatt et al., 2024).

The other retained predictors also require careful interpretation. WFA, WAZ, weight, height, and MUAC measure overlapping dimensions of current body size and nutritional status; their high association with a contemporaneous stunting label may therefore reflect the shared biological and computational structure of anthropometric indicators rather than independent

pre-diagnostic information. These variables may be useful for reproducing current nutritional status when all measurements are already available, but their practical value for early screening is limited because measuring height and age is already sufficient to calculate HAZ and determine stunting status directly. A scientifically stronger early-prediction design would use information available before the diagnostic measurement, such as birth size, parental height, maternal health, feeding practices, infection history, sanitation, socioeconomic conditions, and previous growth trajectory. The recent Indonesian newborn study by Azriani et al. (2025), for example, predicted stunting at two years from earlier parental and household characteristics and compared multiple algorithms using F1-score and predictive values. That design answers a prospective risk-stratification question, whereas the present study primarily classifies status recorded at the same measurement occasion. Making this distinction explicit prevents the current high accuracy from being interpreted as evidence that future stunting can be predicted before growth faltering occurs.

Chi-Square ranking is a univariate association procedure: it quantifies the relationship between each candidate feature and the class label but does not establish causal importance, independent contribution, stability across samples, or incremental clinical utility. A high score may arise from direct derivation of the target, strong correlation with another anthropometric variable, category frequency, or the large sample size. Likewise, selecting the top six variables solely because they occupy ranks 1-6 does not demonstrate that six is the optimal model size. The reduction from ten to six predictors should be supported by fold-wise comparisons of performance, computational cost, and stability across alternative numbers of features. An ablation analysis should compare at least SVM alone, SVM with Chi-Square, SVM with SMOTE, and SVM with both components, while a no-feature-selection model would show whether the reduction actually improves generalization or simply preserves outcome-proximal variables. Because supervised feature selection uses the class label, it must be fitted separately inside each training fold; selecting features once on the full dataset allows validation-fold information to influence the model and can materially exaggerate performance (James et al., 2021; Rosenblatt et al., 2024).

The observed performance cannot currently be attributed specifically to SMOTE, Chi-Square, or the RBF-SVM because no baseline or component-wise comparison is reported. SMOTE can increase representation of the minority class by interpolating between minority observations, but it does not automatically improve generalization, and synthetic cases may be unrealistic when variables combine continuous anthropometry with encoded categorical states. In a large retrospective clinical cohort, Ke et al. (2024) found that resampling effects varied substantially across algorithms and metrics; some combinations improved precision-recall performance, whereas others reduced discrimination or worsened calibration. This evidence supports treating SMOTE as a hypothesis to be tested rather than an optimization that can be assumed effective. The current study would be substantially strengthened by reporting the change in stunting recall, precision-recall area, balanced accuracy, Matthews correlation coefficient, and calibration before and after SMOTE, together with fold-level means, standard deviations, and confidence intervals. A comparison with class-weighted SVM, logistic regression, and a tree-based model would also show whether the complexity of the proposed pipeline produces a meaningful gain over simpler and more interpretable alternatives.

Direct comparison of the 96.6% accuracy with values from previous stunting studies is inappropriate unless the studies share the same target definition, prevalence, predictor timing, resampling strategy, and validation design. Azriani et al. (2025) reported an F1-score of 84.5% for a k-nearest-neighbor model predicting stunting at two years from 23 earlier-life candidate variables, with six final predictors centered on parental height, education, sanitation, and waste management. The lower numerical score does not imply an inferior model because the task was temporally harder and did not rely on contemporaneous HAZ as a predictor. Conversely, Ayele et al. (2025) reported very high performance after applying SMOTE to multiclass stunting data

from Ethiopia, illustrating that post-resampling class balance, model family, and evaluation design can strongly influence reported metrics. Earlier studies from Bangladesh and Rwanda similarly demonstrate that selected algorithms and important variables vary across populations and data sources (Khan et al., 2021; Ndagijimana et al., 2023; Rahman et al., 2021). The distinctive contribution of the present study is therefore not that its accuracy is numerically higher, but that it analyzes a large, locally relevant dataset spanning 26 Samarinda health centers. That contribution will become more defensible when the evaluation is leakage-free and the model is compared with transparent baselines under identical folds.

The reduction from 102,534 initial records to 27,020 analyzed records is another finding that deserves substantive discussion because approximately three quarters of the data were excluded as duplicates or incomplete observations. Routine child-growth databases often contain legitimate repeated measurements taken at different ages; if records were defined as duplicates without considering child identity and measurement date, clinically meaningful longitudinal information may have been removed. Conversely, if multiple visits from the same child remained and random record-level cross-validation was used, records from one child could appear in both training and validation folds, allowing the model to exploit within-child similarity. Subject-level dependence is a recognized source of optimistic performance, and leakage from repeated subjects has been shown to inflate machine-learning estimates substantially (Rosenblatt et al., 2024). Complete-case deletion may also introduce selection bias if excluded children differ systematically in age, health-center attendance, nutritional status, or socioeconomic vulnerability. The manuscript should therefore describe the duplicate rule, number of unique children, measurements per child, missingness by variable, and characteristics of excluded versus retained records. Grouped cross-validation by child identifier, followed by health-center-based or temporal validation, would provide a more credible estimate of generalization to new children, locations, and periods.

Although the dataset covers 26 community health centers, all records originate from one city and may share common data-entry systems, anthropometric equipment, staff training, referral patterns, and local population characteristics. The model may consequently learn center-specific measurement or coding patterns that do not transfer to other districts. Before operational use, performance should be examined by sex, age group, health center, and measurement period to identify subgroup disparities and distribution shift. External validation should then be conducted in an independent district without refitting the model, followed by recalibration if predicted probabilities are intended to guide action. For a screening model, discrimination alone is insufficient: decision-curve analysis, calibration plots, calibration slope and intercept, and threshold-specific net benefit are needed to determine whether using the model improves decisions compared with measuring HAZ directly or applying existing growth-monitoring procedures. TRIPOD+AI and PROBAST+AI emphasize transparent reporting of participant flow, predictor availability, missing data, clustering, model specification, evaluation data, calibration, and risk of bias before implementation claims are made (Collins et al., 2024; Moons et al., 2025).

Taken together, the results show that the SVM pipeline can closely reproduce the recorded stunting labels in the processed dataset, but the current evidence does not yet demonstrate an independent early-warning model. The principal strengths are the large multicenter administrative source, the explicit attempt to reduce features, and attention to class imbalance. The principal threats to validity are the likely inclusion of HAZ as an outcome-defining predictor, the apparent evaluation on a SMOTE-balanced dataset, possible preprocessing outside cross-validation, uncertain handling of repeated child measurements, extensive record exclusion, and absence of external validation. The most informative next analysis is therefore not the immediate deployment of the model, but a complete leakage-aware reanalysis that removes target-derived variables, nests encoding, scaling, feature selection, and SMOTE within grouped training folds, evaluates predictions on untouched records with the

original prevalence, and compares the full pipeline against simpler baselines. Under those conditions, any retained improvement in sensitivity, precision-recall performance, and calibration would provide substantially stronger evidence that the model offers practical value for stunting surveillance in Samarinda.

CONCLUSION

The results of applying the Chi Square feature selection method and handling class imbalance with the SMOTE technique on the SVM algorithm with high-dimensional stunting data from Samarinda City produced very good model performance, as evidenced by an accuracy of 96.6%, with precision, recall, and f1-score values of 97% each. This reflects that both methods have effectively addressed the class imbalance issue and improved the model's ability to classify data with a high level of accuracy.

RECOMMENDATION

Based on the robust findings of this study, several specific recommendations are proposed to advance the application of machine learning in stunting detection and to further optimize the current methodology:

a. Integration of Explainable AI (XAI) for Clinical Trust: Although the Support Vector Machine (SVM) model with an RBF kernel achieved an excellent accuracy of 96.6%, nonlinear SVMs inherently operate as "black-box" models. To bridge the gap between computational metrics and clinical applicability, future research should integrate Explainable Artificial Intelligence (XAI) frameworks, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations). This integration will provide local interpretability, allowing healthcare professionals to understand exactly how heavily weighted features like HAZ, WFA, and MUAC influence individual toddler predictions.
b. External Validation Across High-Risk Demographics: The current framework was trained and evaluated exclusively using data from 26 community health centers in Samarinda City. To ensure the model avoids spatial bias and proves its generalizability, future studies must perform external validation using independent datasets from other municipalities. Testing the model on datasets from Kutai Kartanegara, which currently holds the highest stunting prevalence in East Kalimantan would rigorously verify the model's robustness across different demographic and nutritional landscapes.
c. Transition to a Clinical Decision Support System (CDSS): Because the Chi-Square feature selection successfully narrowed down the high-dimensional data to only 6 critical attributes (HAZ, WFA, WAZ, Height, MUAC, and Weight), the data collection burden for healthcare workers is significantly reduced. It is highly recommended to deploy this optimized SVM-SMOTE model as a web-based or mobile-based Clinical Decision Support System (CDSS). This deployment would empower integrated health-post cadres to perform real-time, early-warning screenings using minimal input parameters.
d. Longitudinal Data and Temporal Modeling: Stunting is a chronic condition characterized by progressive growth faltering rather than an acute event. While this study successfully classified static historical records, future data collection should focus on longitudinal, time-series data that tracks a child's growth trajectory across multiple periodic measurements. Consequently, future algorithmic implementations could explore temporal machine learning architectures (such as Long Short-Term Memory networks or temporal ensemble methods) to predict the future trajectory of a child's nutritional status based on sequential growth patterns.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for funding this research through the 2023 New Lecturer Research Grant scheme.

FUNDING INFORMATION
This work was financially supported by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia under the 2023 New Lecturer Research Grant scheme (*Penelitian Dosen Pemula*) pursuant to Decree No. 0217/E5/PG.02.00/2023 and Contract No. 187/E5/PG.02.00.PL/2023. The Article Processing Charge (APC) was fully covered by the same research grant allocation.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Taghfirul Azhima Yoga Siswa	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Nauval Azmi Verdikha		✓		✓		✓	✓	✓	✓	✓	✓			

CONFLICT OF INTEREST STATEMENT
The authors declare that they have no competing financial, personal, or professional interests that could have inappropriately influenced or biased the objectivity, findings, or interpretation of the research presented in this manuscript.

INFORMED CONSENT
Informed consent was not required because this study involved a retrospective analysis of secondary administrative health data provided by the Samarinda City Health Office. All individual identifiers were fully anonymized prior to feature selection and model development, ensuring complete data privacy and participant confidentiality.

ETHICAL APPROVAL
This study was conducted in full compliance with national regulatory frameworks, institutional governance policies, and the tenets of the Declaration of Helsinki. Approval for the research protocol and secondary data utilization was granted by the Institute for Research and Community Service (*Lembaga Penelitian dan Pengabdian Masyarakat* – LPPM) at Universitas Muhammadiyah Kalimantan Timur (Ref. No. 190/LPPM/C.5/C/2023). Additionally, formal administrative authorization to access and analyze the routine pediatric health records was officially issued by the Samarinda City Health Office.analyze the routine child health records was granted by the Samarinda City Health Office.

DATA AVAILABILITY
The data that support the findings of this study are available from the corresponding author, [T.A.Y.S.], upon reasonable request. The data are not publicly available due to privacy restrictions protecting participant health records and institutional governance policies of the Samarinda City Health Office.

REFERENCES

Arisandi, R. R. R., Warsito, B., & Hakim, A. R. (2022). Application of the Naive Bayes Classifier (NBC) to classify stunting nutritional status among children under five using k-fold cross-validation. *Jurnal Gaussian, 11*(1), 130-139.

Azriani, D., Agustian, D., Zuhairini, Y., Yulita, I. N., & Dhamayanti, M. (2025). Prediction models for stunting at 2-years-old from Indonesian newborn population. *BMC Pediatrics, 25*, Article 718. https://doi.org/10.1186/s12887-025-06096-4

Ayele, M. K., Baye, G. A., Yesuf, S. H., Engda, A. A., & Mitiku, E. T. (2025). Predicting stunting status among under five children in Ethiopia using ensemble machine learning algorithms. *Scientific Reports, 15*, Article 27907. https://doi.org/10.1038/s41598-025-03206-1

Bernett, J., Blumenthal, D. B., Grimm, D. G., et al. (2024). Guiding questions to avoid data leakage in biological machine learning applications. *Nature Methods, 21*, 1444-1453. https://doi.org/10.1038/s41592-024-02362-y

Cameron, L., Chase, C., Haque, S., Joseph, G., Pinto, R., & Wang, Q. (2021). Childhood stunting and cognitive effects of water and sanitation in Indonesia. *Economics & Human Biology, 40*, Article 100944. https://doi.org/10.1016/j.ehb.2020.100944

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

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction

models that use regression or machine learning methods. *BMJ, 385*, Article e078378. https://doi.org/10.1136/bmj-2023-078378

Health Development Policy Agency. (2022). *Pocketbook of the 2022 Indonesian Nutritional Status Survey (SSGI) results*. Ministry of Health of the Republic of Indonesia.

Hakimah, M., Prabiantissa, C. N., Rozi, N. F., Yamani, L. N., & Puspitasari, I. (2022). Determination of relevant feature combinations for detection stunting status of toddlers. In *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 324-329). IEEE. https://doi.org/10.1109/ISRITI56927.2022.10053069

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804

Ke, J. X. C., DhakshinaMurthy, A., George, R. B., & Branco, P. (2024). The effect of resampling techniques on the performances of machine learning clinical risk prediction models in the setting of severe class imbalance: Development and internal validation in a retrospective cohort. *Discover Artificial Intelligence, 4*, Article 91. https://doi.org/10.1007/s44163-024-00199-0

Khan, J. R., Tomal, J. H., & Raheem, E. (2021). Model and variable selection using machine learning methods with applications to childhood stunting in Bangladesh. *Informatics for Health and Social Care, 46*(4), 425-442. https://doi.org/10.1080/17538157.2021.1904938

Lonang, S., & Normawati, D. (2022). Classification of stunting status among children under five using K-Nearest Neighbor with backward-elimination feature selection. *Jurnal Media Informatika Budidarma, 6*(1), 49-56. https://doi.org/10.30865/mib.v6i1.3312

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur, N. C., Dhiman, P., et al. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ, 388*, Article e082505. https://doi.org/10.1136/bmj-2024-082505

Mosquera, C., Ferrer, L., Milone, D. H., Luna, D., & Ferrante, E. (2024). Class imbalance on medical image classification: Towards better evaluation practices for discrimination and calibration performance. *European Radiology, 34*(12), 7895-7903. https://doi.org/10.1007/s00330-024-10834-0

Ndagijimana, S., Kabano, I. H., Masabo, E., & Ntaganda, J. M. (2023). Prediction of stunting among under-5 children in Rwanda using machine learning techniques. *Journal of Preventive Medicine and Public Health, 56*(1), 41-49. https://doi.org/10.3961/jpmph.22.388

Perdana, A. Y., Latuconsina, R., & Dinimaharawati, A. (2021). Prediction of stunting among children under five using the Random Forest algorithm. *e-Proceeding of Engineering, 8*(5), 6650-6656.

Pizzol, D., Tudor, F., Racalbuto, V., Bertoldo, A., Veronese, N., & Smith, L. (2021). Systematic review and meta-analysis found that malnutrition was associated with poor cognitive development. *Acta Paediatrica, 110*(10), 2704-2710. https://doi.org/10.1111/apa.15964

Purbasari, A., Rinawan, F. R., Zulianto, A., Susanti, A. I., & Komara, H. (2021). CRISP-DM for data quality improvement to support machine learning of stunting prediction in infants and toddlers. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. IEEE. https://doi.org/10.1109/ICAICTA53211.2021.9640294

Rahman, S. M. J., Ahmed, N. A. M. F., Abedin, M. M., Ahammed, B., Ali, M., Rahman, M. J., & Maniruzzaman, M. (2021). Investigate the risk factors of stunting, wasting, and underweight among under-five Bangladeshi children and its prediction based on machine learning approach. *PLOS ONE, 16*(6), Article e0253172. https://doi.org/10.1371/journal.pone.0253172

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., et al. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications, 15*, Article 1829. https://doi.org/10.1038/s41467-024-46150-w

Sasmita, W. G., Nugroho, W. H., & Astuti, A. B. (2022). CART method approach and high dimension simulation data selection and random under-sampling method in stunting case. *Journal of Theoretical and Applied Information Technology, 100*(24), 4810-4817.

Velliangiri, S., Alagumuthukrishnan, S., & Joseph, S. I. T. (2019). A review of dimensionality reduction techniques for efficient computation. *Procedia Computer Science, 165*, 104-111. https://doi.org/10.1016/j.procs.2020.01.079

Victora, C. G., Christian, P., Vidaletti, L. P., Gatica-Dominguez, G., Menon, P., & Black, R. E. (2021). Revisiting maternal and child undernutrition in low-income and middle-income countries: Variable progress towards an unfinished agenda. *The Lancet, 397*(10282), 1388-1399. https://doi.org/10.1016/S0140-6736(21)00394-9

World Health Organization. (2025). *Global nutrition targets 2030: Stunting brief*. https://www.who.int/publications/i/item/B09383