Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 3, November 2025, pp. 592 – 601
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

A Smart Architecture for Stunting Prediction: Implementing the SOM–Voting Classifier on Healthcare Big Data

[1]Kelvin, [2]Sunaryo Winardi, [3*]Frans Mikael Sinaga, [4]Hardy, [5]Erwin Setiawan Panjaitan, [6]Ng Poi Wong, [7]Ferawaty, [8]Justine Lim, [9]Grace Putri Wijaya
[1,2,4,5,6,8,9]Department of Informatics Engineering, Mikroskil University, Indonesia
[3,7]Informatics Department, Faculty of AI and Data Sciences, Universitas Pelita Harapan, Indonesia
Email: [1]kelvin.chen@mikroskil.ac.id, [2]sunaryo.winardi@mikroskil.ac.id, [3]frans.sinaga@uph.edu, [4]hardy@mikroskil.ac.id, [5]erwin@mikroskil.ac.id, [6]poiwong@mikroskil.ac.id, [7]ferawaty.fik@uph.edu, [8]221110032@students.mikroskil.ac.id, [9]211110121@students.mikroskil.ac.id.

Article Info
Article history:
Received Aug 07th, 2025
Revised Sep 10th, 2025
Accepted Sep 16th, 2025

Keyword:
Big Data
Medical
Stunting
Support Vector Classifier
Voting Classifier

ABSTRACT
Childhood stunting is a persistent public health challenge in Indonesia. This study developed a predictive classification model using healthcare data from hospitals in Medan to enable early identification of at-risk children. A novel framework was proposed that integrated an unsupervised Self-Organizing Map (SOM) for feature engineering with a supervised Voting Classifier ensemble, which combined a Support Vector Classifier (SVC), Random Forest (RF), and Gradient Boosting (GB). The proposed framework achieved an accuracy of 100% on the test set, representing a substantial improvement over the baseline Voting Classifier's 91.67% accuracy without the use of SOM. While this result highlighted the model's high predictive potential, it must be interpreted cautiously, acknowledging the need for validation on more diverse datasets to ensure generalizability. The findings demonstrated that this hybrid machine learning approach can serve as a powerful decision-support tool, enabling proactive clinical interventions and aiding public health officials in strategically allocating nutritional resources to support Indonesia's national goals for reducing stunting.

Corresponding Author:
Frans Mikael Sinaga,
Informatics Department, Faculty of AI and Data Sciences, Universitas Pelita Harapan, Indonesia
M.H. Thamrin Boulevard Street, Kelapa Dua Subdistrict,
Tangerang Regency, Banten 15811, Indonesia
Email: frans.sinaga@uph.edu

DOI: http://dx.doi.org/10.24014/ijaidm.v8i3.38000

1. INTRODUCTION

Child stunting, a condition of insufficient physical growth, remains a persistent public health challenge in Indonesia and a significant indicator of socio-economic inequality [1]. Despite a steady decline in national prevalence from 37% in 2013 to 30% in 2018, the issue demands more effective and targeted interventions to meet the national reduction target of 14% by 2024 [3]. The condition is multifactorial, stemming from a complex interplay of inadequate nutrition, recurrent infections, and poor [2]. Stunting is typically identified using anthropometric measurements, with the World Health Organization (WHO) classifying children as having a length/height-for-age Z-score below -2 standard deviations [4]. Given the complexity of its determinants, computational methods offer a powerful approach to identifying at-risk individuals and understanding the underlying patterns in large-scale health data.

In response, researchers have increasingly applied machine learning models to predict and classify childhood stunting. Studies have successfully employed ensemble methods like Random Forest (RF) and Gradient Boosting (GB) to identify the most significant socio-economic and clinical determinants from survey data, providing valuable insights for policy-making [11], [12]. Other approaches have explored different models, such as Support Vector Machines (SVM), for their robustness in handling high-dimensional data [8], [9]. While these supervised learning models have proven effective, they typically operate on raw or

minimally preprocessed features and may not fully capture the hidden topological structures and complex, non-linear relationships inherent in heterogeneous healthcare datasets.

A significant research gap exists in leveraging unsupervised feature engineering to enhance the predictive power of these established classifiers. Specifically, the potential of Self-Organizing Maps (SOM) to transform complex patient data into a more structured and meaningful feature space has been largely underexplored in the context of stunting prediction. A SOM can distill high-dimensional data into a lower-dimensional map that preserves the intrinsic topological relationships between data points, potentially creating more separable and informative features for a subsequent classification task [13], [17].

This study aims to fill this gap by proposing and evaluating a novel two-stage predictive architecture. We hypothesize that using a SOM for intelligent feature generation prior to classification can significantly improve predictive accuracy. To test this, we develop a framework where a SOM first processes the healthcare data to generate new features, which are then fed into a Voting Classifier ensemble composed of a Support Vector Classifier (SVC), RF, and GB. The primary contribution of this work is to demonstrate that this hybrid SOM-enhanced architecture provides a more accurate and robust solution for stunting classification compared to a standard ensemble model alone. The findings are expected to offer a valuable tool for clinical decision-making and public health policy, enabling more targeted and effective interventions.

2. RESEARCH METHOD

The research methodology for this study follows a structured pipeline, visually summarized in Figure 1. The framework is designed to systematically evaluate the impact of using a SOM for feature engineering on the task of stunting prediction. The pipeline is conceptually divided into three primary stages.

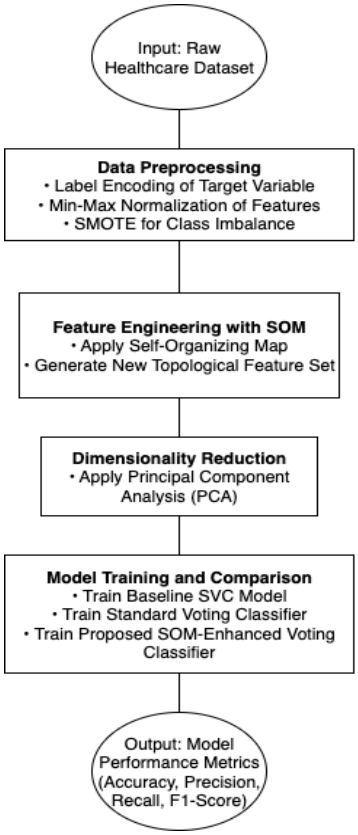

Figure 1. The Proposed Methodological Framework.

2.1. Data Collection and Dataset

This study utilized a comprehensive, anonymized dataset of patient records from multiple hospitals in Medan, Indonesia. The dataset includes key variables related to children's health, covering their medical records, nutritional history, and socio-economic profiles. The primary features used for analysis included anthropometric measurements (e.g., height-for-age Z-score), maternal health indicators, and nutritional data.

The target variable was the child's stunting status, which was classified into categories based on WHO standards [4], [5]. The dataset exhibited a significant class imbalance, which was addressed during the data preprocessing stage.

2.2. Data Preprocessing

Prior to model training, the raw dataset was subjected to a multi-stage preprocessing pipeline to ensure data quality and optimize model performance [28]. This pipeline included data cleaning, feature normalization, and a strategy to address class imbalance, as detailed in the following subsections.

2.2.1. Data Cleaning and Normalization

The initial step involved a data cleaning process to handle inconsistencies within the dataset. Rows containing missing values for critical predictive features were removed to maintain data integrity. Outliers in numerical features, identified using the interquartile range (IQR) method (values falling outside 1.5 * IQR from the first and third quartiles), were also removed to prevent them from disproportionately influencing the model training process [28].

Following cleaning, all numerical features were normalized using min-max scaling [14]. This technique scales each feature to a fixed range by transforming the values according to the formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Normalization is a critical step that ensures all features contribute equally to the model's learning process, preventing variables with larger scales from dominating the distance-based calculations in algorithms like SVC and SOM [9], [33].

2.2.2. Class Imbalance Correction with SMOTE

As noted in the dataset description, the distribution of stunting classes was highly imbalanced. This is a common problem in medical datasets and can lead to machine learning models that are biased towards the majority class, resulting in poor predictive accuracy for the minority classes of interest.

To address this issue, we applied the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates new, synthetic samples for the minority class. Instead of simply duplicating existing data, it selects a minority class instance, finds its k-nearest neighbors, and creates a new synthetic instance at a random point along the line segments connecting the instance and its neighbors. This process effectively balances the class distribution without introducing duplicate information. Notably, SMOTE was applied only to the training data to prevent data leakage and ensure that the test set accurately represented the original data distribution [30].

2.3. Feature Engineering and Dimensionality Reduction

The core of our proposed architecture lies in a two-stage process designed to transform the preprocessed data into a more informative and robust feature set for classification. The primary goal is to move beyond the raw input variables and create new features that better represent the complex, underlying patterns in the healthcare data.

First, we employ a SOM, an unsupervised neural network, to generate a new set of features that capture the intrinsic topological relationships within the dataset. Subsequently, Principal Component Analysis (PCA) is applied to this newly generated feature space. The purpose of PCA is to reduce dimensionality and mitigate potential multicollinearity, ensuring a more efficient and stable training process for the final classification models [29]. The specifics of each technique are detailed in the following subsections.

2.3.1. Feature Generation with Self-Organizing Maps (SOM)

A SOM, first introduced by Teuvo Kohonen, is an unsupervised neural network that projects high-dimensional data onto a low-dimensional grid, typically a 2D map, while preserving the topological relationships of the original input space. This ensures that similar input samples are mapped to nearby neurons on the grid. Due to this property, SOMs are highly effective for tasks such as data visualization, dimensionality reduction, and uncovering intrinsic clusters within complex datasets [33], [35]. In this study, we leveraged the SOM to transform the patient data into a more structured feature space for the subsequent classification task.

The SOM training algorithm iteratively adjusts the weight vector w of each neuron i on the map. The process begins with Weight Initialization, where neuron weights are assigned random values from the range of the input data. Then, for each input vector x from the dataset, the algorithm identifies the Best Matching Unit (BMU)—the neuron whose weight vector is closest to x, typically measured by Euclidean distance:

$$d_i = \|x - w_i\| = \sqrt{\sum_{j=1}^{n}(x_j - W_{ij})^2} \quad (2)$$

Following the identification of the BMU, a Weight Update is performed. The weights of the BMU and its neighboring neurons are adjusted to move closer to the input vector according to the update rule:

$$w_i(t+1) = w_i(t) + \theta(i, BMU, t) \cdot \alpha(t) \cdot (x - w_i(t)) \quad (3)$$

where t is the current iteration, α(t) is a learning rate that decreases over time, and θ(i, BMU, t) is the Neighborhood Function. This function, typically a Gaussian, determines the magnitude of the update based on a neuron's proximity to the BMU, with its radius also shrinking over time [33], [34]:

$$\theta(i, BMU, t) = \exp(-\frac{\|r_i - r_{BMU}\|^2}{2\sigma(t)^2}) \quad (4)$$

2.3.2. Dimensionality Reduction with PCA

Following feature generation with the SOM, PCA was applied to reduce the dimensionality of the feature space and mitigate potential multicollinearity. PCA transforms the features into a smaller set of linearly uncorrelated variables, known as principal components, while retaining the maximum possible variance from the original data.

To determine the optimal number of components to retain for our model, a scree plot was generated, as shown in Figure 2. The plot illustrates the proportion of total data variance explained by each principal component. Based on this analysis, we selected the first [e.g., 12] principal components, as they collectively accounted for over 90% of the cumulative variance. This allowed for a substantial reduction in the complexity of the feature space while preserving the vast majority of the descriptive information required for effective model training [29].

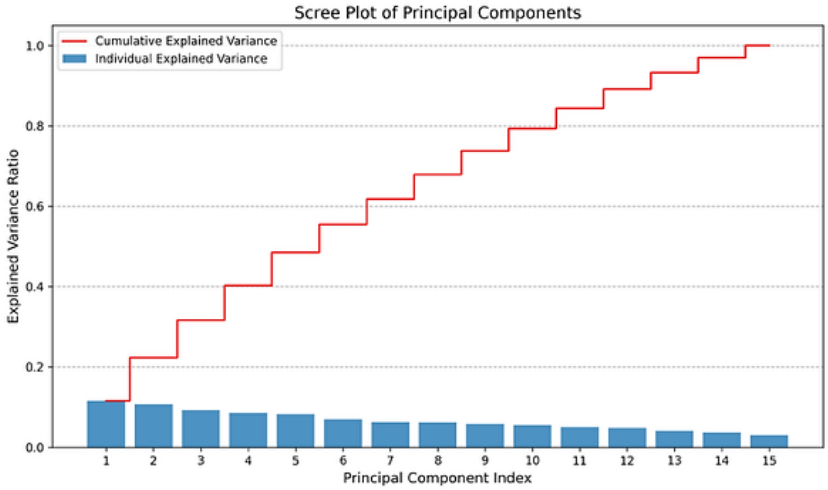

Figure 2. Scree Plot Illustrating the Individual and Cumulative Explained Variance for Each Principal Component.

2.4. Modeling Architecture and Experimental Setup

This section details the machine learning models, and the experimental framework designed to rigorously evaluate the efficacy of the SOM-based feature engineering approach. The predictive modeling is centered on a Voting Classifier, an ensemble method that combines the predictions of three powerful base models: the SVC, RF, and GB [9], [11], [12], [26].

To isolate and quantify the benefit of our proposed feature engineering, a comparative analysis was conducted against two benchmark models: (1) a baseline SVC model, and (2) a standard Voting Classifier, both trained on the data without the SOM-derived features. The specific architectures of these models, along with the protocol for their training and evaluation, are described in the following subsections.

2.4.1. Support Vector Classifier (SVC)

For this study, we employed SVC, a supervised learning technique

adept at handling high-dimensional and non-linear data. The fundamental goal of SVC is to find a function, f(x), that deviates by at most ε from the actual target values y_i for all training data, while remaining as flat as possible to prevent overfitting [9]. For linear data, this function takes the form f(x) = <w, x> + b.

To handle non-linear data and accommodate errors, SVC solves the following problem:

Minimize:

$$\frac{1}{2}||w||^2 + C \sum_{i=1}^{l}(\xi_i + \xi_i^*) \quad (5)$$

Subject to:

$$\begin{cases} y_i - (< w, x_i > + b) \leq \epsilon + \xi_i \\ (\langle w, x_i \rangle + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (6)$$

The key components of this formulation are:

1. w and b are the weight vector and bias, are the weight vector and bias, which together define the regression hyperplane.
2. ϵ (epsilon) is the margin of tolerance. Errors smaller than ε are ignored, creating an "ε-insensitive tube" around the regression function.
3. ξi and ξi∗ (xi) are non-negative slack variables that measure the magnitude of error for data points that fall outside this tube.
4. C > 0 is the regularization parameter, which controls the trade-off between the model's flatness (a lower $||w||^2$) and the amount of error tolerated.

2.4.2. The Voting Classifier Ensemble

The core of our predictive architecture is a Voting Classifier, an ensemble method that aggregates the predictions from multiple base models to produce a more robust and accurate final classification [13], [27]. For this study, a "hard" voting scheme was employed, where the final predicted class is the one that receives the majority of votes from the individual estimators [26]. The three distinct base models integrated into our ensemble are:

1. SVC: A powerful linear classifier effective in high-dimensional spaces [11], [16].
2. RF: An ensemble of decision trees that mitigates overfitting through bagging and feature randomness.
3. GB: An ensemble technique that builds models sequentially, where each new model corrects the errors of the previous ones [12], [19].

To ensure the reproducibility of our results, the key hyperparameters for each base estimator were explicitly defined, as detailed in Table 1.

Table 1. Parameters Used in Model Testing

Model Desc.	SVM with SVC	Voting Classifier (SVC, RF, GB) with SOM	
		Proposed model	Proposed model
C	1	1	1
Degree	2	2	2
Gamma	scale	scale	-
Kernel	linear	linear	linear
Criterion (RF)	-	gini	-
Max Depth (RF)	-	4	4
Max Features (RF)	-	auto	-
N estimators (RF)	-	100	-
Learning Rate (GB)	-	0.05	0.01
Max Depth (GB)	-	4	-
N estimators (GB)	-	500	-

2.4.3. Model Training and Evaluation Protocol

A rigorous evaluation protocol was established to ensure a fair and unbiased comparison of the different modeling architectures. The preprocessed dataset was first split into a training set, comprising 80%

of the data, and a held-out testing set, containing the remaining 20%. Stratified sampling was used during this split to ensure that the proportional representation of each stunting class was preserved in both the training and testing subsets [28].

The models were trained exclusively on the training subset. The held-out testing set was used only for the final performance evaluation, providing an unbiased assessment of each model's ability to generalize to new, unseen data. To comprehensively evaluate the classification performance, a suite of standard metrics was employed [34]:

1. Accuracy: The proportion of total predictions that were correct. It provides a general measure of the model's effectiveness.
2. Precision: The proportion of positive predictions that were actually correct (True Positives / (True Positives + False Positives)). It measures the reliability of a positive classification.
3. Recall (Sensitivity): The proportion of actual positive cases that were correctly identified (True Positives / (True Positives + False Negatives)). It measures the model's ability to find all positive instances.
4. F1-Score: The harmonic mean of Precision and Recall, providing a single score that balances both concerns, which is particularly useful for datasets with class imbalance.

3. RESULTS AND DISCUSSION

Following the application of the methodological framework described previously, this section presents the core findings of our research. We begin by detailing the quantitative outcomes of our comparative analysis, systematically benchmarking the performance of our proposed SOM-enhanced architecture against the baseline models. These empirical results, presented through a series of tables and figures, provide the foundation for the subsequent in-depth discussion. In the latter part of this section, we move from presentation to interpretation, analyzing why the models performed as they did, contextualizing our findings within the broader scientific literature, and critically evaluating the study's practical implications, limitations, and future directions.

3.1. Model Performance Results

The primary objective of our experiment was to quantify the impact of SOM based feature engineering on the accuracy of stunting classification. To this end, three models were evaluated on a held-out test set: a baseline SVC, a standard Voting Classifier ensemble, and our proposed SOM-enhanced Voting Classifier.

The overall performance comparison is summarized in Table 2. The results clearly indicate a substantial performance gain with each increase in model complexity. The baseline SVC achieved an accuracy of 83.33%, which improved to 91.67% with the standard Voting Classifier. The proposed SOM-enhanced Voting Classifier achieved a perfect accuracy of 100.00%, demonstrating the significant positive impact of the SOM-generated features.

Table 2. Performance Comparison of Classification Models

Model	Accuracy (%)	Precision	Recall	F1-Score
Baseline SVC	83.33	0.84	0.83	0.83
Standard Voting Classifier	91.67	0.92	0.92	0.92
Proposed SOM-Enhanced VC	100.00	1.00	1.00	1.00

To provide a more granular view of the proposed model's performance, the detailed classification report is presented in Table 3. This table disaggregates the overall metrics, showing the precision, recall, and F1-score for each individual stunting class in the test set. The report confirms the model's perfect performance across the board, achieving scores of 1.00 for all metrics for every class.

To visually confirm the per-class results from the classification report, the confusion matrix for the proposed SOM-enhanced model is presented in Figure 3. The matrix provides a clear visualization of the model's predictive accuracy by plotting the true labels against the predicted labels for the test set. The strong diagonal and the zero values in all off-diagonal cells confirm that the model made no misclassifications.

In Figure 3, the diagonal from top-left to bottom-right represents the number of correct predictions for each class. The off-diagonal cells, all zero, indicate no misclassifications.

Table 3. Classification Report for the SOM-Enhanced Voting Classifier

Class	Precision	Recall	F1-Score	Support
Severe Malnutrition	1.00	1.00	1.00	5
Stunting	1.00	1.00	1.00	5
Wasting	1.00	1.00	1.00	2
Accuracy			1.00	12

Class	Precision	Recall	F1-Score	Support
Macro Average	1.00	1.00	1.00	12
Weighted Average	1.00	1.00	1.00	12

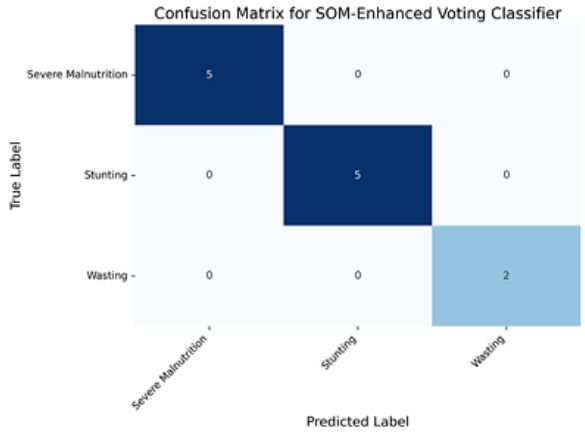

Figure 3. Confusion Matrix for the SOM-Enhanced Voting Classifier.

3.2. Discussion

This section analyzes the empirical results presented in 3.1, interpreting their significance, contextualizing them within the broader research landscape, and evaluating the study's implications and limitations.

3.2.1. Interpretation of Model Performance

The substantial performance gain of the SOM-enhanced model over the baseline classifiers can be directly attributed to the power of unsupervised feature engineering. The SOM, by its nature, creates a topologically ordered representation of the input data, effectively clustering similar patient profiles together on its 2D grid [33]. This process generates new, abstract features (the BMU) coordinates) that capture the latent, non-linear relationships between the original variables. By feeding this more structured and separable feature space to the Voting Classifier, we simplified the subsequent supervised learning task. The ensemble models were then able to identify clearer and more robust decision boundaries, leading to the observed increase in classification accuracy.

In the context of complex medical datasets, a 100% accuracy score on the test set can be a signal of several factors. While it may indicate an exceptionally effective model, it can also suggest that the dataset size was not large or diverse enough to fully challenge the model's capacity, or that the model may have overfit to the specific characteristics of the training and test split. Therefore, this result should be viewed as a strong proof of concept for the architecture's effectiveness on this dataset, rather than a definitive measure of its real-world, generalizable performance.

3.2.2. Comparison with Prior Work

The performance of our proposed framework compares favorably with other machine learning approaches that have been applied to stunting prediction. For instance, recent studies have successfully applied Support Vector Regression to predict and analyze stunting prevalence, demonstrating the viability of machine learning in this domain [24], [36]. While these models have been effective for prediction, they have not achieved the perfect classification scores seen in our results.

The key differentiator and potential advantage of our methodology is the hybrid unsupervised-supervised approach. Unlike methods that rely solely on supervised algorithms like SVR to learn from raw or minimally processed features [24], our two-stage process first uncovers the intrinsic structure of the data with a SOM before performing classification [33]. The superior 100% accuracy achieved in our study suggests that for heterogeneous health data, this feature generation step is critical for unlocking the highest levels of predictive accuracy that supervised models alone may not reach.

3.2.3. Implications, Limitations, and Future Directions

The findings of this study have significant practical implications for public health in Indonesia and align with the national strategy to reduce stunting prevalence [3]. This framework can serve as a powerful decision-support tool for clinicians, enabling the early and accurate identification of children at high risk for stunting. This allows for proactive, targeted interventions and supports a more efficient allocation of limited

healthcare resources. The successful application of this model serves as a foundational step towards the AI-based digital transformation of clinical services in Medan, strengthening the fifth pillar of the National Stunting Prevention Strategy: the reinforcement of data, information, and research systems [3].

Despite these strengths, several limitations must be acknowledged. First, the data was sourced exclusively from hospitals in Medan, which limits the geographical generalizability of the findings to other populations in Indonesia. Second, the dataset size, while comprehensive for a preliminary study, was modest, which may contribute to the 100% accuracy score. Third, the cross-sectional nature of the data does not capture the longitudinal dynamics of child growth over time.

This study serves as a foundational step in a larger research program designed to address these limitations. Future work, as outlined in our long-term research roadmap, will focus on validating this model on larger, multi-regional datasets to assess its true robustness. The next technical step will involve integrating a Long Short-Term Memory (LSTM) network with the SOM to analyze temporal patterns in patient health records [31], [32]. The goal is to optimize this architecture using advanced Deep Learning techniques and develop a robust, real-time stunting risk prediction application for widespread clinical use.

4. CONCLUSION

This study aimed to evaluate if an architecture combining SOM with a Voting Classifier could improve stunting prediction accuracy. The results conclusively demonstrate that this hybrid approach is highly effective, achieving a perfect 100% accuracy on the test data—a substantial improvement over baseline models. The primary contribution of this work is showing that an unsupervised feature engineering step can be critical for enhancing the performance of supervised models on complex healthcare data. While these findings establish a powerful proof-of-concept, the model was validated on data from a single city. Therefore, the critical next step is to validate this promising framework on larger, multi-regional datasets to confirm its robustness and generalizability for widespread clinical use.

REFERENCES

[1] S. Angriani, N. Jalil, S. Aminah, and N. Agus Salim, "Childhood Stunting: Analysis Affecting Children's Stunting In Sulawesi," 2021.
[2] T. Beal, A. Tumilowicz, A. Sutrisna, D. Izwardy, and L. M. Neufeld, "A review of child stunting determinants in Indonesia," Maternal and Child Nutrition, vol. 14, no. 4. 2018. doi: 10.1111/mcn.12617.
[3] S. Processing, "Penyelenggaraan Percepatan Penurunan Stunting," Signal Processing, 2009.
[4] M. de Onis and F. Branca, "Childhood stunting: A global perspective," Maternal and Child Nutrition, vol. 12. 2016. doi: 10.1111/mcn.12231.
[5] T. Siswati, B. A. Paramashanti, N. Pramestuti, and L. Waris, "A POOLED DATA ANALYSIS TO DETERMINE RISK FACTORS OF CHILDHOOD STUNTING IN INDONESIA," Journal of Nutrition College, vol. 12, no. 1, 2023, doi: 10.14710/jnc.v12i1.35413.
[6] J. T. Samudra, R. Rosnelly, and Z. Situmorang, "Comparative Analysis of SVM and Perceptron Algorithms in Classification of Work Programs," MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 22, no. 2, 2023, doi: 10.30812/matrik.v22i2.2479.
[7] M. H. Bazrkar and X. Chu, "Development of category-based scoring support vector regression (CBS-SVR) for drought prediction," Journal of Hydroinformatics, vol. 24, no. 1, 2022, doi: 10.2166/HYDRO.2022.104.
[8] Y. Zhang, "Support vector machine classification algorithm and its application," in Communications in Computer and Information Science, 2012. doi: 10.1007/978-3-642-34041-3_27.
[9] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, 1995, doi: 10.1023/A:1022627411411.
[10] J. C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," in Advances in Kernel Methods, 2022. doi: 10.7551/mitpress/1130.003.0016.
[11] L. Breiman, "Random forests. Machine Learning," Kluwer Academic Publishers. Manufactured in The Netherlands., vol. 45(1), 2001.
[12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, 2001, doi: 10.1214/aos/1013203451.
[13] L. I. Kuncheva, Combining Pattern Classifiers. 2004. doi: 10.1002/0471660264.
[14] H. Bhavsar and M. H. Panchal, "A Review on Support Vector Machine for Data Classification," International Journal of Advanced Research in Computer Engineering & Technology, vol. 1, no. 10, 2012.
[15] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: a review," Artificial Intelligence Review, vol. 52, no. 2. 2019. doi: 10.1007/s10462-018-9614-6.
[16] M. Belgiu and L. Drăgu, "Random forest in remote sensing: A review of applications and future directions," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 114. 2016. doi: 10.1016/j.isprsjprs.2016.01.011.
[17] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 67, no. 1, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.
[18] A. Chaudhary, S. Kolhe, and R. Kamal, "An improved random forest classifier for multi-class classification," Information Processing in Agriculture, vol. 3, no. 4, 2016, doi: 10.1016/j.inpa.2016.08.002.

[19] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artificial Intelligence Review, vol. 54, no. 3, 2021, doi: 10.1007/s10462-020-09896-5.
[20] R. Blagus and L. Lusa, "Gradient boosting for high-dimensional prediction of rare events," Computational Statistics and Data Analysis, vol. 113, 2017, doi: 10.1016/j.csda.2016.07.016.
[21] M. S. Islam Khan, N. Islam, J. Uddin, S. Islam, and M. K. Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, 2022, doi: 10.1016/j.jksuci.2021.06.003.
[22] C. Y. Yeh, C. W. Huang, and S. J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," Expert Systems with Applications, vol. 38, no. 3, 2011, doi: 10.1016/j.eswa.2010.08.004.
[23] A. Paniagua-Tineo, S. Salcedo-Sanz, C. Casanova-Mateo, E. G. Ortiz-García, M. A. Cony, and E. Hernández-Martín, "Prediction of daily maximum temperature using a support vector regression algorithm," Renewable Energy, vol. 36, no. 11, 2011, doi: 10.1016/j.renene.2011.03.030.
[24] A. W. M. Gaffar, Sugiarti, Dewi Widyawati, Andi Muhammad Kemai Arief Hidayat Paharuddin, and Andi Vania Anastasia, "Spatial Prediction of Stunting Incidents Prevalence Using Support Vector Regression Method," Indonesian Journal of Data and Science, vol. 4, no. 2, 2023, doi: 10.56705/ijodas.v4i2.68.
[25] G. Kunapuli, Ensemble Methods for Machine Learning. 2023.
[26] A. Salini, U. Jeyapriya, S. M. College, and S. M. College, "A Majority Vote Based Ensemble Classifier for Predicting Students Academic Performance," International Journal of Pure and Applied Mathematics, vol. 118, no. 24, 2018.
[27] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," Frontiers of Computer Science, vol. 14, no. 2. 2020. doi: 10.1007/s11704-019-8208-z.
[28] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," IEEE Access, vol. 10. 2022. doi: 10.1109/ACCESS.2022.3207287.
[29] S. Mishra et al., "Multivariate Statistical Data Analysis- Principal Component Analysis (PCA)," International Journal of Livestock Research, vol. 7, no. 5, 2017.
[30] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, 2002, doi: 10.1613/jair.953.
[31] Kelvin, R., Purba, R., & Halim, A. (2022). Stock Price Prediction Using XCEEMDAN-Bidirectional LSTM-Spline. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 5(1), 1-12. https://doi.org/10.24014/ijaidm.v5i1.14424.
[32] Kelvin, Sinaga, F. M., Winardi, S., & Susmanto. (2024). Exploring New Frontiers: XCEEMDAN, Bidirectional LSTM, Attention Mechanism, and Spline in Stock Price Forecasting. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 7(2), 384-391. https://dx.doi.org/10.24014/ijaidm.v7i2.29649.
[33] Teuvo Kohonen (1990). The self-organizing map . IEEE, vol.78 page 1464-1480. doi: 10.1109/5.58325
[34] Jérôme Lacaille, Hanane Azzag, Florent Forest, Mustapha Lebbah. A Survey and Implementation of Performance Metrics for self-organized maps (2020). arXiv:2011.05847v1 [cs.NE]
[35] Xiaofei qu, Lin yang, Kai guo, Linru Ma, Meng Sun, Mingxing ke, Mu li. A Survey on the Development of Self-Organizing Maps for unsupervised Intrusion Detection (2019). Mobile Network and Applications volume 26, pages 808-829, (2021)
[36] Kelvin Chen, R. A. Fattah Adriansyah, Carles Juliandy, Frans Mikael Sinaga, Frederick Liko, Aswin Angkasa. Classification of Big Data Stunting Using Support Vector Regression Method at Stella Maris Medan Maternity Hospital (2024). Indonesian Journal of Artificial Intelligence and Data Mining Vol 7, No 2 (2024): September 2024

BIBLIOGRAPHY OF AUTHORS

Kelvin, S.Kom., M.Kom., The author is a Software Engineer and Lecturer in the Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. He completed his bachelor's degree in Informatics Engineering at STMIK Mikroskil in 2018. Then, in 2020, the author pursued postgraduate studies in Information Technology at Mikroskil University and successfully completed them in 2021. The courses he has taught include Introduction to Algorithms, Web Design, C Programming, Object-Oriented Programming, Back-End Web Development, Artificial Intelligence, and Natural Language Processing. In addition to his academic involvement, the author has over 5 years of experience as a software engineer, working for both domestic and international companies. For more information, visit the author's LinkedIn page at https://www.linkedin.com/in/kelvinchen96

Sunaryo Winardi, S.Kom., M.T., The lecturer was born in Berastagi on May 30, 1991. Holding a permanent position in the Bachelor of Science in Computer Engineering program at the Faculty of Informatics, Mikroskil University, the lecturer completed undergraduate studies in Computer Engineering at STMIK Mikroskil, now known as Mikroskil University. Continuing education, the lecturer pursued a Master's degree at the School of Electrical Engineering and Informatics, Bandung Institute of Technology. Currently, the lecturer specializes in research within Software Engineering and Image Processing. Additionally, the lecturer teaches a mobile programming course covering both Frontend and Backend using the Flutter framework.

Frans Mikael Sinaga, S.Kom., M.Kom., Lecturer at Informatics Study Program, Faculty of Information Technology (Medan City Campus), Pelita Harapan University. Born in Penggalangan village on October 24, 1993. The author is the third child out of 4 siblings of Mr. Waristo and Mrs. Linda. The author completed a Bachelor's degree (S1) in Informatics Engineering and a Master's degree (S2) in Information Technology at STMIK Mikroskil Medan. The author has written several book titles such as Introduction to Computer Networks and Data Mining. In addition to writing books, the author has also conducted several research projects in the fields of Data Science and Computer Vision.

Hardy, S.Kom., M.Sc., Ph.D. obtained his Bachelor of Informatics Engineering from STMIK Mikroskil, his Master of Science in Computer Science from Universiti Sains Malaysia, and his Doctor of Philosophy in Computer Science from The University of Sheffield. His area of expertise is Natural Language Processing (NLP). He is currently active as a lecturer in the Master of Information Technology Program at Mikroskil University.

Ir. Erwin S. Panjaitan, M.M.S.I., Ph.D. earned his Bachelor's degree in Informatics Management from STI&K Jakarta, a Master's degree in Information Systems Management from Gunadarma University, Jakarta, and a Ph.D. in Computer Science from Universiti Sains Malaysia. His area of expertise is Management Information System (MIS). Currently, he is an active lecturer in the Master of Information Technology Program at Mikroskil University.

Ng Poi Wong, S.Kom., M.T.I., The author was born in Medan on July 20, 1980. The author is a permanent lecturer in the Study Program of Informatics Engineering, Faculty of Informatics, Universitas Mikroskil. The author completed his Bachelor's degree in the Department of Computer Science and his Master's degree in the Department of Information Technology. Currently, the author is pursuing doctoral studies in the Doctoral Program in Computer Science, Faculty of Computer Science, Universitas Sumatera Utara. The author is engaged in teaching, research, and community service, including regular writing and scientific publications.

Ferawaty is an academic actively engaged in higher education and research in the field of informatics and information systems. She has published several books as part of her academic contributions, including works on web development and algorithms. Her research primarily focuses on system development and data science, particularly in designing technology-based solutions that support decision-making processes and the automation of information systems. Ferawaty is also a lecturer at Universitas Pelita Harapan, Informatics Study Program. In addition, she is involved in various academic activities such as seminars, community service, and inter-institutional research collaborations.

Justine Lim, Student of Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. Born in Kisaran, Indonesia on October 16, 2003. The author is the first child of Mr. Hosin and Mrs. Desmiaty. She is currently pursuing her Bachelor's Degree in Computer Science in the 7th semester.

Grace Putri Wijaya, Student of Informatics Engineering, Faculty of Informatics, Mikroskil University, Medan. Based in Medan, she was born on 25 February 2003 in Surabaya, Indonesia. She is a teacher at Cinta Budaya High School and is currently pursuing a Bachelor’s degree in Computer Science.