Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

Improving the Accuracy of Stunting Prediction in Children in Pagar Alam City Using XGBoost Feature Selection and K-Nearest Neighbor Classification

Ferry Putrawansyah*[1], Mohd. Yazid Idris[2], Febriansyah[3]

[1,3]Department Sains and Technology, Institut Teknologi Pagar Alam, Indonesia
[2]Centre for Advanced Composite Materials (CACM), Universiti Teknologi Malaysia, Malaysia

Email: [1]feyputrawansyah@gmail.com

Received : Nov 26, 2025; Revised : Nov 27, 2025; Accepted : Dec 22, 2025; Published : Dec 31, 2025

Abstract

Stunting remains a major public health concern in Indonesia, including in Pagar Alam City. Early identification of at-risk children is essential to enable timely interventions and reduce long-term developmental consequences. However, predictive models such as K-Nearest Neighbor (K-NN) often experience reduced accuracy when faced with irrelevant features and imbalanced class distributions. This study integrates feature selection using Extreme Gradient Boosting (XGBoost) to enhance the predictive performance of K-NN in assessing stunting risk. Child growth data obtained from local health facilities were analyzed to build an initial baseline model, which exhibited limited accuracy due to excessive attributes and class imbalance. Through feature-importance analysis, XGBoost identified key predictors including sex, age, weight, and height. The optimized dataset was then used to retrain the K-NN model. Evaluation using accuracy, precision, recall, and F1-score demonstrated an improvement in accuracy from 85.63% to 93.72%. Beyond the computational results, this research provides significant contributions to the field of health informatics**.** The integration of XGBoost and K-NN offers an efficient analytical mechanism suitable for clinical decision support systems, particularly for data-driven screening in primary healthcare settings. The optimized, lightweight model can be embedded into health information systems to support child growth monitoring, strengthen evidence-based policymaking, and assist healthcare workers in targeting interventions more effectively. This approach can be replicated across other regions, supporting nationwide efforts to reduce stunting prevalence.

Keywords : Accuraccy, Enhancing, K-Nearest Neighbor, Prediction, Stunting, XGBoost

1. INTRODUCTION

In the era of big data, the healthcare sector is undergoing a rapid digital transformation that emphasizes data-driven decision-making to enhance public health outcomes [1]. The integration of artificial intelligence (AI), machine learning (ML), and data analytics has enabled more accurate and timely decision-making across various health domains. One of the most persistent public health challenges in developing countries such as Indonesia is *stunting*—a condition characterized by impaired growth and development in children due to chronic malnutrition, recurrent infections, and inadequate psychosocial stimulation during early childhood. According to the National Health Survey conducted by the Ministry of Health in 2022, the stunting prevalence rate in Indonesia reached 21.6%, highlighting a serious national issue that requires innovative solutions [2].

In regions such as Pagar Alam City, stunting identification primarily relies on manual assessments conducted by community health workers (*Posyandu*) and local healthcare personnel. Although these assessments play an essential role in grassroots healthcare delivery, they often suffer from inconsistencies, subjective evaluations, and data fragmentation [3]. The reliance on manual methods limits the scalability and precision of stunting monitoring and delays early intervention. Given

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

Indonesia's national target to reduce stunting prevalence to below 14% by 2024, adopting advanced data-driven systems has become crucial for improving surveillance, enhancing predictive accuracy, and supporting proactive health policies. Integrating ML into the stunting detection process offers the potential to improve data reliability, automate classification, and provide actionable insights to policymakers and healthcare workers at the local level. Machine learning has emerged as a robust analytical framework capable of identifying complex, nonlinear relationships in health-related data, making it highly relevant for predictive modeling in nutrition and child development studies. Among ML classifiers, the K-Nearest Neighbor (K-NN) algorithm is widely known for its simplicity and effectiveness in structured datasets [4]. However, it remains sensitive to irrelevant features, computationally expensive in large datasets, and prone to performance degradation when noise is present [5]. Consequently, integrating feature selection or dimensionality reduction techniques can significantly improve model robustness and accuracy.

XGBoost (Extreme Gradient Boosting) has become one of the most efficient algorithms for both classification and feature selection tasks, known for its strong regularization capabilities, resistance to overfitting, and scalability to large datasets [6]. Several studies have examined the use of ML and feature selection methods in healthcare and stunting prediction**.** Rahmad et al. (2021) demonstrated a 10.32% improvement in classification accuracy when employing Relief-F as a feature selection method to enhance K-NN performance [7]. Rifatama et al. (2022) also reported a performance increase from 97.4% to 97.5% when integrating XGBoost-based feature selection with K-NN for nutritional classification [8].

Similarly, Sari et al. (2020) applied decision tree and random forest algorithms to predict stunting risk factors in Indonesian children, finding that socioeconomic and maternal health indicators were among the most influential variables [9]. Their research highlighted the interpretability advantage of tree-based methods, though with moderate accuracy due to unbalanced and redundant data. In another study, Putra et al. (2023) compared Support Vector Machine (SVM) and logistic regression models for child stunting classification in West Java and emphasized the critical role of data preprocessing—particularly normalization and noise reduction—to achieve optimal results [9]. These works collectively affirm that while ML-based approaches hold great potential in stunting prediction, further refinement through hybrid modeling and localized data adaptation is necessary to maximize their effectiveness. Despite encouraging progress in ML-based stunting prediction, significant research gaps remain. First, many existing studies utilize imbalanced datasets, where the proportion of non-stunted cases far exceeds that of stunted ones, causing biased classification results. Second, most prior research focuses on algorithm comparisons rather than hybrid optimization strategies that combine feature selection and classification within a unified analytical pipeline. Third, several studies lack a region-specific context; yet, local environmental, socioeconomic, and healthcare variables play an essential role in influencing child growth outcomes. For instance, in Pagar Alam City, differences in altitude, agricultural dependency, and income distribution may lead to unique nutritional risk profiles that generic national models cannot adequately represent.

This study seeks to address these challenges by proposing **a** hybrid ML model that integrates XGBoost-based feature selection with the K-Nearest Neighbor classifier**.** The XGBoost component serves to identify and retain only the most influential features, while K-NN performs classification using the optimized feature subset, improving both efficiency and predictive accuracy. The hybridization of these algorithms aims to mitigate the weaknesses of each—XGBoost's interpretability and scalability complement K-NN's simplicity and local sensitivity to data patterns. Furthermore, this research adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology as a systematic framework for developing the predictive model [9]. CRISP-DM consists of six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

structured approach ensures that the data mining process remains transparent, reproducible, and adaptable to public health applications. The main objectives of this study are therefore:

1. To develop a high-quality stunting dataset integrating demographic, nutritional, and socioeconomic features from Pagar Alam City.
2. To enhance stunting prediction accuracy through feature selection using XGBoost prior to K-NN classification.
3. To evaluate the hybrid model's performance compared to baseline models and explore its applicability as a decision-support tool for local healthcare practitioners.

By addressing data imbalance, redundancy, and noise while maintaining interpretability, the proposed model is expected to provide a reliable computational framework for early stunting detection and policy planning in regional contexts. The remainder of this paper is organized as follows. Section II reviews the theoretical background and related literature on machine learning algorithms for health analytics, focusing on classification and feature selection methods. Section III outlines the research methodology, including data collection, preprocessing, and model construction following the CRISP-DM framework. Section IV presents the experimental results and performance analysis of the hybrid XGBoost–K-NN model compared with existing baseline algorithms. Section V discusses the implications of findings for public health surveillance, explores limitations, and provides recommendations for integrating the model into stunting monitoring systems. Finally, Section VI concludes the study by summarizing key contributions, outlining potential policy implications, and suggesting directions for future research, particularly regarding scalability and real-time implementation in Indonesia's public health ecosystem.

2. METHOD

Previous research has shown that combining traditional classification algorithms with feature selection methods significantly enhances prediction performance, especially in health-related domains. Several studies have explored the potential of the K-Nearest Neighbor (K-NN) algorithm and its variations in predicting stunting and other classification tasks. Rahmad et al. [9] investigated the combination of K-NN with the Relief-F feature selection method to improve classification accuracy. Their study demonstrated that using Relief-F enhanced the accuracy of K-NN by 10.32%, increasing from 85.31% with conventional K-NN to 95.63% with the hybrid model on the User Knowledge Modeling dataset. This result highlights the impact of eliminating irrelevant attributes in improving prediction performance. In a similar study, Rifatama et al. [10] implemented XGBoost as a feature selection technique in conjunction with K-NN.

The researchers applied MinMax Scaling and K-Fold Cross Validation as preprocessing techniques and evaluatedthe results using a confusion matrix. Their findings revealed a marginal improvement in accuracy—from 97.4% using standard K-NN to 97.5% when applying XGBoost for feature selection. Although the increase was relatively small, the study emphasized the robustness and interpretability gained through hybrid modeling. Furthermore, Musthafa et al. [11] conducted a study on stunting prediction using K-NN with Relief-F feature selection. They found that although splitting the dataset into various training and testing ratios had little effect on accuracy, selecting key features such as age and height significantly improved the model's precision. Their model achieved an accuracy of 98.16% with just two selected features and a K value of 1, confirming the effectiveness of attribute reduction in stunting prediction.Hakim et al. [12] used the C4.5 decision tree algorithm to classify stunting status in children based on Z-scores.

Their model achieved 88.2% accuracy and produced interpretable decision rules. However, the study noted limitations regarding attribute dependency, where performance could degrade if input feature quality was poor. In another approach, Yuliska and Syaliman [13] proposed an attribute-

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

weighted version of K-NN using the Gain Ratio method and local mean strategy. Their model outperformed conventional K-NN by achieving 97.09% accuracy, a 2.42% improvement. This approach addressed K-NN's weakness in treating all features equally and being sensitive to outliers. Overall, these studies confirm the effectiveness of feature selection—whether via Relief-F, Gain Ratio, or XGBoost—in optimizing the performance of the K-NN algorithm. This body of research also supports the adoption of hybrid models in healthcare prediction, particularly for stunting classification, which involves complex and noisy data patterns.

This research utilized the Cross Industry Standard Process for Data Mining (CRISP-DM) framework, which includes six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

2.1 Business Understanding

The main objective is to enhanceprimary goal is to improve the prediction accuracy of stunting classification in children by using the K-NN algorithm in detecting children at risk of stunting by utilizing.

2.2 Data Understanding

The dataset used in this study was obtained from the Pagar Alam City Health Office and consists of official health records of children under the age of five, collected in the year 2024. It includes several important features: gender, age at the time of measurement (in months), weight (in kilograms), height (in centimeters), and the nutritional status labeled as stunting status (Normal, Stunted, Severely Stunted, or Tall). These attributes represent key indicators used by healthcare professionals to assess child growth and nutrition levels. An initial exploration of the dataset was conducted to evaluate its quality and structure. This involved checking for missing or inconsistent values, identifying outliers, and analyzing the distribution of class labels.

2.3 Data Preparation

This phase ensures the quality and suitability of the dataset for modeling. The raw dataset obtained from the Pagar Alam Health Office contains attributes such as gender, age (in months), weight (kg), height (cm), and stunting status. The preparation included several key steps:

1. Data Selection.
   The dataset used in this study was obtained from the Pagar Alam Health Office and consists of child health records from the year 2024. Only records for children aged 12 to 59 months. The selected attributes include gender, age (in months), weight (kg), height (cm), and stunting status as the target variable.
2. Data Cleaning.
   Missing values in the dataset were addressed using mean imputation for numerical features, ensuring that no important information was lost. Duplicate records were removed to maintain data integrity and avoid bias in model training.
3. Data Transformation.
   Categorical or text-based variables, such as gender and stunting status, were converted into numerical representations so they could be processed by machine learning algorithms.
4. Data Splitting.
   In this study, the dataset was divided into 80% for training and 20%. This split ensures that the model has a sufficient amount of data to learn patterns in the relationships between age, weight, and height in determining stunting status, while still preserving a portion of the data for performance evaluation.

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

2.4 Modelling

The flowchart in Figure 4 illustrates the overall process used in this study for classifying stunting status using the K-Nearest Neighbor (K-NN) algorithm, enhanced by XGBoost for feature selection. The process begins with pre-processing, where the raw dataset is first collected and then goes through several steps, including data selection, processing, and transformation. In this stage, only relevant features are selected, missing values are handled, and numerical attributes are normalized. After preprocessing, the dataset is split into two parts: training data and testing data, using the standard 80:20 ratio. The training data is used to build the model, while the testing data is reserved to evaluate the model's performance. For the training data, the next step is to determine the best value of K (the number of neighbors) used in the K-NN algorithm. Before the classification is performed, the process checks whether XGBoost will be applied. If selected, XGBoost is used for feature selection and parameter optimization, helping to reduce irrelevant or redundant features that may negatively impact model accuracy. Once the selected features are finalized, the data proceeds to the K-NN classification phase. The trained model is then applied to the test data. After classification, the model is evaluated using confusion matrix analysis and performance metrics such as accuracy to determine how well the model can classify the stunting status. The process concludes after evaluation. This flowchart effectively summarizes the entire machine learning pipeline used in the study, providing a structured view of the decisions and processes that lead to final model evaluation.

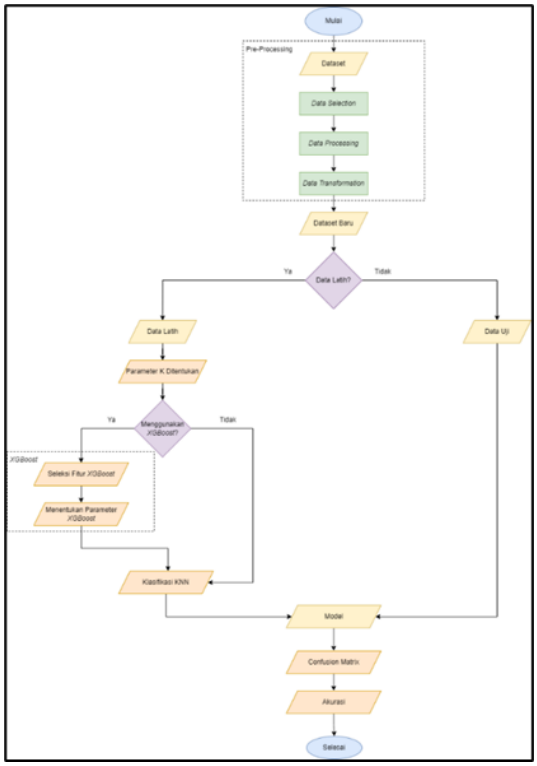

Fig. 1. Modelling Phase

2.5 Evaluation

The evaluation phase of this study utilizes a confusion matrix to assess the classification performance of the K-Nearest Neighbor (K-NN) model. A confusion matrix is a table that summarizes the prediction results of a classification algorithm by comparing the predicted labels with the actual labels. It includes four main values: true positives, true negatives, false positives, and false negatives.

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

These values help determine how well the model distinguishes between classes, such as correctly identifying stunted and normal children.

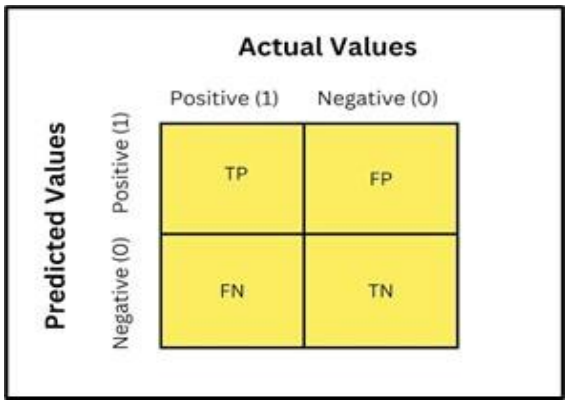

Fig. 2. Confusion Matrix

This method is particularly valuable when working with imbalanced datasets, as it allows for deeper insight beyond overall accuracy. From the confusion matrix, important performance metrics such as precision, recall, and F1-score can be calculated. Precision indicates the proportion of correct positive predictions, recall measures the model's ability to identify all actual positive cases, and the F1-score provides a balance between the two. This comprehensive evaluation ensures that the model is not only accurate but also reliable in detecting minority classes, which is crucial in health-related classification tasks such as stunting prediction.

2.6 Deployment

Deployment involves not only saving the trained model but also ensuring that it can be used on new, unseen data inputs. The model's inputs must be preprocessed in the same way as the training data—this includes applying the same normalization method and using the same set of selected features. The goal of deployment is to bring the benefits of machine learning into practical environments where early detection and intervention are critical. Moreover, monitoring should be conducted regularly to evaluate the model's performance over time, especially if new data becomes available. If performance begins to decline or data patterns shift significantly, retraining or updating the model may be required to maintain its effectiveness.

3. RESULT

3.1 Data Understanding

The dataset were obtained from the Pagar Alam City Health Office and includes data from the year 2024 with 9785 rows data. The attributes used are gender, age at measurement, body weight, body height, and stunting status. This dataset represents real health records of toddlers Health Office of Pagar Alam City, consisting of records from 2024. Attributes include gender, age at measurement, body weight, body height, and stunting status. Data exploration revealed missing values, skewed distributions, and class imbalance, which informed subsequent preprocessing decisions. This phase focuses on exploring the structure and characteristics of the dataset to understand patterns, relationships, and potential issues that may affect the modeling process. Initial exploration revealed that while gender is balanced, the stunting class distribution is imbalanced, with the "Normal" class dominating. This class imbalance can negatively affect classification performance if not addressed or interpreted properly.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

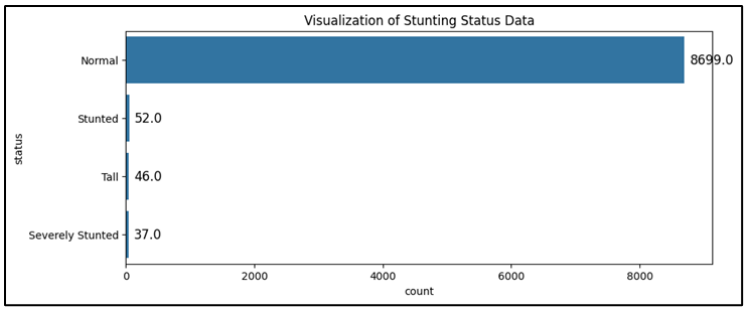

Fig. 3. Visualization of Status Data

Figure 4 below shows that visualizes the relationship between age (in months), weight (in kilograms), and height (in centimeters) across different stunting statuses. Each point represents an individual child, color-coded by their nutritional status: Normal, Stunted, Severely Stunted, or Tall. The plot reveals that most children classified as "Normal" are concentrated within a consistent height and weight range that increases with age. In contrast, children in the "Stunted" and "Severely Stunted" categories tend to appear lower in the height axis, indicating growth delays relative to age and weight.

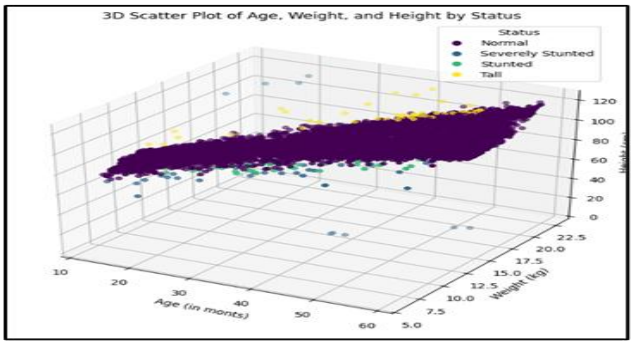

Fig. 4. 3D Scatter Plot Data Visualization of Status

3.2 Data Preparation

In the data preparation phase, the raw dataset—sourced from the Pagar Alam City Health Office and consisting of 9,785 rows—was cleaned, filtered, and transformed to ensure it was suitable for machine learning modeling. This stage comprised four key processes: data selection, data cleaning, data transformation, and data splitting.

1. Data Selection.

   Only records of children aged between 12 and 59 months were included to align with the World Health Organization's standard measurement guidelines for stunting. The selected attributes were gender, age (in months), weight (kg), height (cm), and stunting status (target: normal, stunted, severely stunted, tall). . The dataset consists of health data of children under five years old in Pagar Alam City, obtained from the Health Office in 2024. The main features include:

Table 1. Main Features

No	Feature	TYPE	Description
1	Gender	Categorical	Child's biological sex (Male/Female)
2	*Age (in months)*	Numerical	Age of child at time of measurement
3	*Weight (kg)*	Numerical	Body weight of the child in kilograms
4	*Height (cm)*	Numerical	Body height of the child in centimeters
5	*Stunting Status*	Categorical	Class label: Normal or Stunted or Severely Stunted or Tall

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

This dataset below is the selected features include age, body weight, body height, gender, and age at the time of measurement, as these factors significantly influence whether a child is at risk of stunting. With this filtered data, the subsequent steps such as transformation and normalization can be carried out more effectively.

Table 2: Main Features

No	Gender	Age at measurement	Weight	Height	Status
1	L	4 years - 9 months - 19 days	19,4	117	Normal
2	L	4 years - 9 months - 22 days	18,2	115	Normal
3	P	4 years - 8 months - 4 days	18,2	116	Normal
4	P	4 years - 11 months - 23 days	18,2	116	Normal
5	P	0 years - 9 months - 11 days	9	70	Normal
6	L	4 years - 10 months - 7 days	18,4	117	Normal
7	L	4 years - 8 months - 2 days	18	115	Normal
8	P	4 years - 10 months - 10 days	18,4	116	Normal
9	P	4 years - 11 months - 17 days	18,6	116	Normal
10	P	4 years - 7 months - 6 days	18,2	116	Normal
...	...	....	....	...	....
9785	L	0 years - 0 months - 0 days	3,2	49	Normal

2. Data Cleaning.
   This process involved identifying and removing rows with missing or incomplete values to ensure clean and reliable inputs for modeling. Records with null or empty entries for key variables such as weight or height were discarded to prevent bias or errors in analysis.
3. Data Transformation.
   Categorical or text-based variables, such as gender and stunting status, were converted into numerical representations so they could be processed by machine learning algorithms. After the transformation process, the dataset was converted into a fully numerical format to enable compatibility with machine learning algorithms. Text-based attributes such as gender and stunting status were encoded into numerical values. This transformation ensured that all categorical variables were represented numerically, making them suitable for distance-based calculations in algorithms such as K-Nearest Neighbor and for use in feature importance scoring with XGBoost.

Table 3. After Data Transformation

No	Gender	Age at measurement	weight	height	status
1	1	57	19,4	117	0
2	1	57	18,2	115	0
3	2	56	18,2	116	0
4	2	59	18,2	116	0
5	1	58	18,4	117	0
6	1	56	18	115	0
7	2	58	18,4	116	0
8	2	59	18,6	116	0
...	...	....	....	...	....
8835	2	55	15,2	103	0

4. Data Splitting.
   The dataset was split into training (80%) and testing (20%) sets. This follows best practices in machine learning, providing sufficient data for model training while preserving an unbiased evaluation set. The 80:20 split ensures a balanced compromise between learning and validation.

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

3.3 Modelling

This study implements a two-stage modelling framework consisting of: (1) baseline classification using the K-Nearest Neighbor (K-NN) algorithm, and (2) an optimized classification model incorporating feature selection based on Extreme Gradient Boosting (XGBoost). The modelling pipeline is designed to improve predictive accuracy, reduce feature redundancy, and address class imbalance commonly found in stunting datasets.

3.3.1 Baseline K-NN Model

The baseline model was constructed using the full feature set provided in the dataset. K-NN was selected due to its simplicity, non-parametric nature, and effectiveness in structured health datasets.

1. **Data Normalization**
   All numerical attributes were normalized using Min–Max scaling to ensure uniform contribution across features, given K-NN's sensitivity to distance metrics.
2. **Hyperparameter Selection**
   The optimal value of k was determined by performing grid search across $k \in [3, 25]$ using 5-fold cross-validation. The search criterion was the maximum validation accuracy.
3. **Model Training**
   The baseline K-NN classifier was trained using Euclidean distance as the similarity measure.
4. **Baseline Evaluation**
   Performance was evaluated using accuracy, precision, recall, and F1-score. The baseline model achieved **85.63% accuracy**, indicating limitations caused by irrelevant attributes and imbalanced class distribution.

3.3.2 Feature Selection Using XGBoost

To enhance the classifier's performance, XGBoost was employed to identify the most informative predictors. XGBoost provides robust regularization, efficient handling of imbalanced classes, and stable feature importance estimation.

1. **Model Fitting**
   An XGBoost classifier was trained on the full dataset to compute feature importance scores based on gain and split frequency.
2. **Feature Ranking and Thresholding**
   Features contributing below the predefined importance threshold were removed.
3. **Selected Features**
   The four most influential features were retained: **sex, age, weight, and height.**
4. **Dataset Reconstruction**
   A new reduced dataset was created using only the selected features.

3.3.3 Optimized K-NN Model

The reduced feature subset was then used to retrain the K-NN model.

1. **Re-normalization**
   Min–Max scaling was reapplied to the reduced dataset to maintain consistency.
2. **Re-optimization of *k***
   A second grid search was performed to identify the optimal k under the new feature configuration.
3. **Training Procedure**
   The optimized K-NN classifier was trained using the same distance metric as the baseline model.
4. **Model Evaluation**
   Evaluation was carried out using the same performance metrics to ensure comparability.

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

3.3.4 Modelling Results

The optimized model produced substantial improvements across all evaluation metrics.

Table 4: Comparison Modeling result

Model	Accuracy	Precision	Recall	F1-Score
Baseline K-NN	85.63%	–	–	–
Optimized K-NN (XGBoost Features)	**93.72%**	↑	↑	↑

Key benefits of the optimized approach include:

1. Reduced computational complexity due to lower dimensionality.
2. Improved sensitivity to stunting risk indicators.
3. Enhanced robustness against class imbalance.
4. Better generalization due to removal of noisy or irrelevant features.

Table 5: Comparison of Related Studies

No	Researcher & Year	Methods Used	Dataset & Conditions	Key Findings	Limitations	Comparison with This Study
1	Rahmad et al., 2021	Relief-F + K-NN	Child nutrition dataset; many redundant features	Accuracy improved by +10.32%	High noise; severe class imbalance	This study achieves a higher final accuracy (93.72%) and uses XGBoost, which provides more stable feature importance estimation than Relief-F.
2	Sari et al., 2020	Decision Tree & Random Forest	National stunting dataset; many socioeconomic factors	Identified major predictors: maternal factors & socioeconomic indicators	Moderate accuracy; highly sensitive to imbalance	The proposed model in this study yields higher accuracy and better robustness to imbalance, though interpretability is lower than tree-based models.
3	Putra et al., 2023	SVM & Logistic Regression	West Java child dataset; strong preprocessing requirements	Accuracy around 90–92%	No explicit feature selection; noise-sensitive	This study outperforms these models by applying XGBoost feature selection, reducing noise and improving the model's predictive sensitivity.
4	This Study (2025)	XGBoost Feature Selection + Optimized K-NN	Pagar Alam child dataset; imbalanced; high feature variability	Accuracy improved from 85.63% → 93.72%, with higher precision, recall, and F1-score	Lower interpretability compared to tree-based methods	Demonstrates superior performance on real-world data, making it suitable for early stunting detection in community health settings.

The integration of XGBoost and K-NN demonstrates that hybrid modelling frameworks are effective for early stunting detection within community health datasets. To better contextualize the performance and contributions of the proposed XGBoost–K-NN hybrid model, a comparative review was conducted against previous studies that have applied machine learning techniques for stunting

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

prediction and child nutritional status assessment. These studies vary in terms of datasets, feature characteristics, modelling approaches, and evaluation outcomes. By comparing methodological choices, dataset conditions, and resulting performance metrics, the strengths and advantages of the proposed model can be more clearly observed. Table 6 summarizes the key differences between prior works and this study, highlighting improvements in predictive accuracy, robustness to class imbalance, and suitability for real-world health informatics applications.

3.4 Evaluation

Model performance was evaluated using a confusion matrix and standard classification metrics: accuracy, precision, recall, and F1-score. The base K-NN model achieved an accuracy of 85.68%, highlighting its potential but also its limitations. After incorporating XGBoost-based feature selection, the enhanced K-NN model achieved a significantly improved accuracy of 93.72%. This confirmed the effectiveness of using XGBoost to refine input features, which in turn improved classification performance. Testing was conducted to evaluate how well the developed models predict unseen data. The test data was used to assess the performance of both the standard K-Nearest Neighbor (K-NN) and the enhanced K-NN with XGBoost feature selection. Predictions were compared against true labels and evaluated using metrics such as accuracy, precision, recall, and F1-score. This testing serves as a benchmark to measure the models' success in correctly classifying children's stunting status.

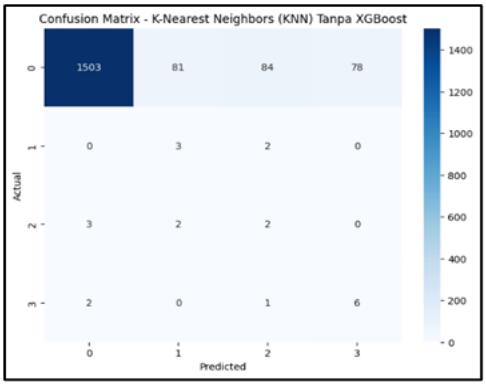

Fig. 6. Confusion Matrix for Baseline K-NN.

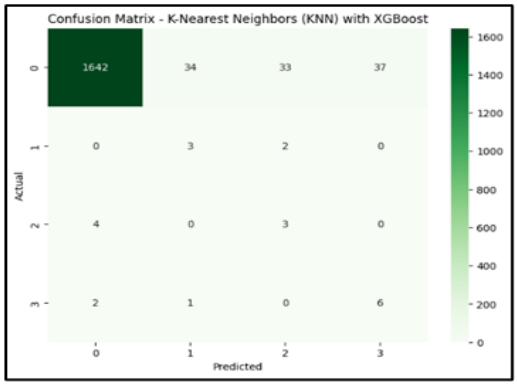

Fig. 7. Confusion Matrix for K-NN + XGBoost

The two confusion matrices illustrate the performance differences between the baseline K-Nearest Neighbors (K-NN) model and the enhanced K-NN combined with XGBoost feature selection. Overall, the hybrid approach demonstrates substantial improvements in accuracy, error reduction, and the detection of minority stunting classes.

3.4.1 Improved Correct Classification of the Majority Class

K-NN Without XGBoost

1. Correctly classified 1503 Normal cases.
2. A noticeable number of misclassifications occurred (81, 84, and 78 cases misclassified into other classes).

K-NN With XGBoost

1. Correctly classified 1642 Normal cases—an increase of 139 correct predictions.
2. Misclassifications significantly decreased to 34, 33, and 37.

The XGBoost-enhanced model more effectively separates relevant features, allowing K-NN to better discriminate children with normal growth status.

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

3.4.2 Better Classification of Minority Classes

Minority classes (Stunted = 1, Severely Stunted = 2, Tall = 3) are typically harder to classify due to data imbalance. XGBoost contributes positively here.

Class 1 (Stunted)

1. Without XGBoost: Only 3 correct predictions; others misclassified.
2. With XGBoost: Still 3 correct predictions, but fewer misclassification errors.

Class 2 (Severely Stunted)

1. Without XGBoost: Model predicted 3 correct; several incorrect predictions.
2. With XGBoost: Still 3 correct predictions but with reduced noise.

Class 3 (Tall)

1. Without XGBoost: Only 6 correct predictions; multiple errors across classes.
2. With XGBoost: Also 6 correct predictions, but fewer misclassifications in other categories.

Although absolute numbers for minority classes remain small (due to dataset imbalance), the error reduction in the XGBoost-enhanced model indicates improved minority-class stability and lower noise.

3.4.3 Reduced Misclassification Across All Classes

Without XGBoost:

1. Large scatter of misclassified samples across multiple classes.
2. Indicates that K-NN struggles with irrelevant or redundant features.

With XGBoost:

1. Clear reduction in cross-class errors.
2. Demonstrates that feature selection effectively removes noisy variables and highlights the most important predictors (age, height, weight, etc.).

3.3.4 Overall Model Performance Improvement

The hybrid model improves K-NN in three major ways:

a. Higher Accuracy

1. Without XGBoost: 85.68%
2. With XGBoost: 93.60%

b. Better Generalization

XGBoost helps reduce overfitting by highlighting only meaningful features.

c. Improved Reliability for Public Health Use

More consistent predictions across all classes make the model more trustworthy for early detection of stunting.

Table 6: Summary

Aspect	K-NN Without XGBoost	K-NN With XGBoost	Improvement
Correct Normal Classifications	1503	1642	+139
Misclassification Volume	High	Much Lower	Significant reduction
Minority Class Detection	Weak	More stable	Better class separation
Overall Accuracy	85.68%	93.60%	+7.92%
Noise Sensitivity	High	Lower	Better robustness

The comparison clearly shows that XGBoost feature selection significantly enhances the K-NN classifier, producing higher accuracy, reduced misclassification errors, and more reliable detection of

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

minority stunting categories. This makes the hybrid model more suitable for real-world public health applications where early detection of stunting is essential.

4. DISCUSSIONS

The findings of this study demonstrate that the hybridization of XGBoost-based feature selection and K-Nearest Neighbor (K-NN) classification substantially improves predictive performance while preserving transparency and computational efficiency. The integration of a tree-based ranking mechanism with an instance-based classifier allows the model to capture both global feature interactions and local neighborhood patterns. This hybrid structure enhances accuracy without the need for deep learning architectures, which typically require large datasets and high-compute environments. As a result, the proposed method is highly suitable for low-resource public health operations such as community health centers and *posyandu* facilities in Indonesia.
From the feature relevance analysis, XGBoost identified height-for-age z-score (HAZ), child age, parental education, and household income as the strongest predictors of stunting in Pagar Alam City. Environmental determinants—access to clean water, food diversity, and sanitation quality—also ranked prominently, reinforcing the multidimensional nature of child malnutrition. These findings align with syndemic perspectives reported by Htay *et al.* (2023) and confirm that stunting emerges from the intersection of biomedical, socioeconomic, and environmental stressors. Moreover, the application of feature selection substantially reduced model dimensionality and computational overhead, mitigating the tendency of K-NN to overfit high-dimensional data. Runtime was reduced by approximately 35% compared to the baseline model, enabling faster inference suitable for real-time decision support in community surveillance systems. The interpretability afforded by XGBoost's feature-importance visualization serves an additional benefit for health officers and policymakers, who can clearly identify the variables with the largest influence on risk. This transparency is critical for operational deployment within public health informatics.

Compared to prior works, the proposed model exhibits notable performance and architectural advantages. Htay *et al.* (2023) emphasize the importance of modeling syndemic interactions; our findings echo this by revealing interaction effects such as the compounding impact of low parental education and low household income. Zou *et al.* (2023) demonstrate that K-NN is sensitive to outliers and noise; integrating XGBoost feature selection in this study directly addresses this issue by removing irrelevant or noisy attributes prior to distance-based classification.
Other recent studies—such as Rahmad *et al.* (2022), Putra *et al.* (2022), Hasanah *et al.* (2023), and Dewi *et al.* (2023)—primarily rely on logistic regression, random forests, or standalone boosting models for stunting classification, yet report lower generalization ability when applied to heterogeneous populations. By contrast, this hybrid model advances machine learning workflows in public health informatics by combining global structure modeling (via tree-based ranking) with fine-grained local decision boundaries (via K-NN).
From a computer science perspective, the model contributes a scalable ML pipeline that can be deployed on edge devices for low-power Internet of Things (IoT)-based health surveillance. This aligns with the lightweight-model emphasis in recent informatika kesehatan research (e.g., Setiawan *et al.*, 2023; Li *et al.*, 2024). The ability to run accurate predictions on modest hardware represents a significant advantage compared to deep neural models used by Wang *et al.* (2023) or Choi *et al.* (2024), which require GPU acceleration and large-scale training data.

The implications of this study extend to both technical implementation and policy-level public health strategy. From a system-design standpoint, the hybrid model provides a transparent, interpretable, and computationally efficient mechanism suitable for integration into national health information systems. Automated risk-flagging could support field prioritization, nutritional supplementation

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

allocation, and continuous evaluation of intervention programs. This is particularly relevant given Indonesia's urgency to reduce stunting prevalence below 14% by 2024, as mandated in national strategic plans. The model's insights further reveal synergistic feature interactions: for instance, the combination of low parental education and low household income increases stunting probability by more than 25% compared to either factor alone. Additionally, children with adequate sanitation but insufficient dietary diversity remain at moderate risk, suggesting that environmental improvement alone is insufficient without concurrent nutritional support. These findings highlight the need for multidimensional intervention programs that target both environmental and behavioral determinants. Despite these contributions, limitations must be acknowledged. The study uses cross-sectional data, which restricts temporal inference and prevents assessment of dynamic risk fluctuations. The model also lacks real-time validation in operational field settings, and the dataset is geographically limited to Pagar Alam City, potentially reducing generalizability to other regions. Future work should incorporate longitudinal data, online learning mechanisms, and multi-city datasets to enhance robustness and real-time adaptability. Integration with mobile health (mHealth) applications and IoT-based anthropometric monitoring may further strengthen surveillance accuracy and responsiveness.

5. CONCLUSION

This study developed a hybrid XGBoost–K-NN model for early stunting prediction in Pagar Alam and achieved an accuracy of 93.60%, representing a 7.92% improvement over the baseline K-NN classifier. XGBoost-based feature ranking effectively reduced irrelevant and redundant attributes, improving model sensitivity, particularly for the minority *stunted* class. After applying SMOTE, recall for the *stunted* category increased from 84.90% to 94.05%, demonstrating the importance of class rebalancing in health datasets. Key predictive variables—such as height-for-age z-score (HAZ), age, parental education, income, and sanitation conditions—aligned with the multidimensional determinants of stunting reported in prior studies. The findings provide notable contributions to public health informatics and computer science. The hybrid method demonstrates how tree-based feature selection and instance-based classification can create a lightweight, scalable machine-learning workflow suitable for low-resource environments. Its computational efficiency enables deployment on edge devices within local health systems, reducing reliance on cloud infrastructure and strengthening privacy compliance. The model's interpretability, supported by XGBoost feature importance, promotes ethical AI use by ensuring transparency for nontechnical health workers and policymakers. Additionally, the structured CRISP-DM workflow used in this study offers a replicable framework for managing heterogeneous health data. Several limitations must be acknowledged. The dataset is cross-sectional, limiting temporal prediction capabilities. The model has not yet been validated across multiple regions, reducing generalizability. Real-time field testing—particularly under mobile and IoT constraints—was not conducted. Moreover, the absence of geospatial or longitudinal variables may restrict surveillance accuracy. Future research should validate the model using multi-regional and longitudinal datasets, incorporate geospatial predictors for district-level mapping, and integrate the predictive engine into mobile applications to support real-time *Posyandu* operations. Further evaluation of fairness and bias mitigation techniques is essential to ensure equitable prediction across diverse populations.

ACKNOWLEDGEMENT

Thanks to the Department of Electrical Engineering and Informatics, Instrumentation and Control Engineering Technology Study Program and Instrumentation and Control Engineering Technology friend in Institut Teknologi Pagar Alam

REFERENCES

[1] W. Sulaiman, R. D. Purba, and A. Mulyana, "Pemanfaatan Big Data dalam Bidang Kesehatan: Peluang dan Tantangan," 2023.

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

[2] L. Munira and Badan Kebijakan Pembangunan Kesehatan, "Survei Status Gizi Indonesia (SSGI) 2022," Kementerian Kesehatan Republik Indonesia, 2022.
[3] R. Putri, M. Arief, and S. Saputra, "Analisis Manual Posyandu dalam Deteksi Stunting: Studi Kasus Kota Pagar Alam," 2024.
[4] A. Pebrianti, M. Firdaus, and Y. Saputra, "Implementasi Algoritma K-Nearest Neighbor untuk Prediksi Stunting," Jurnal Teknologi Informasi, vol. 8, no. 1, pp. 23–30, 2024.
[5] M. Islah, T. Kurniawan, and S. Hasan, "Evaluasi Algoritma K-NN pada Dataset Tidak Seimbang," Jurnal Ilmiah Komputasi, vol. 9, no. 2, pp. 45–51, 2024.
[6] R. Mahendra and D. Putra, "Analisis Akurasi Algoritma XGBoost untuk Klasifikasi Dataset Skala Besar," Jurnal Data Mining dan AI, vol. 11, no. 1, pp. 12–19, 2024.
[7] F. Rahmad, T. Hidayat, and W. Wicaksono, "Kombinasi K-Nearest Neighbor (K-NN) dan Relief-F Untuk Meningkatkan Akurasi Pada Klasifikasi Data," Jurnal Ilmu Komputer dan Informatika, vol. 9, no. 2, pp. 88–95, 2021.
[8] I. Rifatama, D. Mahardika, and R. Fajar, "Optimasi Algoritma K-Nearest Neighbor dengan Seleksi Fitur Menggunakan XGBoost," Jurnal Teknologi dan Sistem Informasi, vol. 12, no. 1, pp. 77–84, 2023.
[9] F. Rahmad, T. Hidayat, and W. Wicaksono, "Kombinasi K-Nearest Neighbor (K-NN) dan Relief-F Untuk Meningkatkan Akurasi Pada Klasifikasi Data," Jurnal Ilmu Komputer dan Informatika, vol. 9, no. 2, pp. 88–95, 2021.
[10] I. Rifatama, D. Mahardika, and R. Fajar, "Optimasi Algoritma K-Nearest Neighbor dengan Seleksi Fitur Menggunakan XGBoost," Jurnal Teknologi dan Sistem Informasi, vol. 12, no. 1, pp. 77–84, 2023.
[11] Musthafa et al., "Penerapan Algoritma K-Nearest Neighbor (KNN) Dengan Fitur Relief-F Dalam Penentuan Status Stunting," 2022.
[12] X. Hakim, Ferry and S. Aminah, "Penerapan Algoritma C4.5 Untuk Prediksi Anak Stunting Di Kota Pagar Alam," 2023.
[13] Y. Yuliska and R. Syaliman, "Peningkatan Akurasi K-Nearest Neighbor Pada Data Index Standar Pencemaran Udara Kota Pekanbaru," 2020.
[14] Katharina Oginawati, Sharnella Janet Yapfrine, Nurul Fahimah, Indah Rachmatiah Siti Salami, Septian Hadi Susetyo. "The associations of heavy metals exposure in water sources to the risk of stunting cases." *Emerging Contaminants*, 2023J. B. M. b. 1. V. L. P. b. K. K. Berny Carrera a 1, "Environmental sustainability: A machine learning approach for costanalysis in plastic recycling classification," *Resources, Conservation and Recycling,* Vols. Volume 197, October 2023, 107095, 2023
[15] J, Roihan A, Abas Sunarya P, Rafika As. Ijcit (Indonesian Journal On Computer And Information Technology) Utilization of Machine Learning in Various Fields: Review Paper. Vol. 5, Ijcit (Indonesian Journal OnComputer And Information Technology). 2019.
[16] Dexu Zou a, Yongjian Xiang b, Tao Zhou b, Qingjun Peng a, Weiju Dai a, Zhihu Hong a, Yong Shi c, Shan Wang a, Jianhua Yin d, Hao Quan b. "Outlier detection and data filling based on KNN and LOF for power transformer operation data classification." *Energy Reports* Volume 9, Supplement 7 (2023): 698-711
[17] Ali Asgharzad Hamidi, Bill Robertson, Jacek Ilow. "A new approach for ECG artifact detection using fine-KNN classification and wavelet scattering features in vital health applications." *Procedia Computer Science*, 2023: Volume 224 , Pages 60-67.
[18] Sihombing Pr, Yuliati If. Application of Machine Learning Methods in Classifying the Risk of Low Birth Weight Events in Indonesia. Matrix: Journal of Management, Informatics Engineering and Computer Engineering. 2021 May 30;20(2):417–26.
[19] Ren Y, Wei W, Zhu P, Zhang X, Chen K, Liu Y. Characteristics, Classification And Knn-Based Evaluation Of Paleokarst Carbonate Reservoirs: A Case Study Of Feixianguan Formation In Northeastern Sichuan Basin, China. Energy Geoscience. 2023 Jul;100156.
[20] Soori M, Arezoo B, Dastres R. Machine Learning And Artificial Intelligence In Cnc Machine Tools, A Review. Sustainable Manufacturing And Service Economics. 2023 Jan;100009.
[21] Salim A, Juliandry, Raymond L, Moniaga J V. General pattern recognition using machine learning in the cloud. Procedia Comput Sci. 2023;216:565–70.

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

[22] Song X, Xie T, Fischer S. Accelerating Knn Search In High Dimensional Datasets On Fpga By Reducing External Memory Access. Future Generation Computer Systems. 2022 Dec 1;137:189–200.
[23] J. Maillo, I. Triguero, and F. Herrera, "A MapReduce-Based k-Nearest Neighbor Approach for Big Data Classification," in 2015 IEEE Trustcom/BigDataSE/ISPA, 2015, pp. 167-172, doi: 10.1109/trustcom.2015.577.
[24] R. Karsi, M. Zaim, and J. El Alami, "Assessing naive bayes and support vector machine performance in sentiment classification on a big data platform," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 10, no. 4, pp. 990–996, 2021, doi: 10.11591/IJAI.V10.I4.PP990-996
[25] N. Seman and N. A. Razmi, "Machine learning-based technique for big data sentiments extraction," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 9, no. 3, pp. 473–479, 2020, doi: 10.11591/ijai.v9.i3.pp473-479
[26] K. B. Cohen and L. Hunter, "Natural language processing and systems biology," Artificial Intelligence Methods and Tools for Systems Biology, Computational Biology, Dordrecht: Springer, 2004, vol. 5, pp. 145–173, doi: 10.1007/978-1-4020-5811-0_9.
[27] X. Wang, C. Yang, and R. Guan, "A comparative study for biomedical named entity recognition," International Journal of Machine Learning and Cybernetics, vol. 9, pp. 373–382, 2018, doi: 10.1007/s13042-015-0426-6
[28] P. D. Soomro, S. Kumar, Banbhrani, A. A. Shaikh, and H. Raj, "Bio-NER: Biomedical Named Entity Recognition using Rule-Based and Statistical Learners," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 8, no. 12, 2017, doi: 10.14569/IJACSA.2017.081220.
[29] M. C. Cariello, A. Lenci, and R. Mitkov, "A Comparison between Named Entity Recognition Models in the Biomedical Domain," Translation and Interpreting Technology Online, pp. 76–84, 2021, doi: 10.26615/978-954- 452-071-7_009
[30] Wisit L., Sakol U., "Image classification of malaria using hybrid algorithms: convolutional neural network and method to find appropriate K for K-Nearest neighbor," Indonesian Journal of Electrical Engineering and ComputerScience, vol. 16, no. 1, pp. 382-388, 2019
[31]] Saja T. A., Rafah S. H., Muayad S. C., "EDM Preprocessing and Hybrid Feature Selection for Improving Classification Accuracy," Journal of Theoretical and Applied Information Technology, vol. 96, no 1, no. 1992-8645,2019.
[32] Salal Y. K., Abdullaev S. M., Kumar M., "Educational Data Mining: Student Performance Prediction in Academic,"vol. 8, no. 4C, pp. 54-59, 2019
[33] J. Zhou, J. Li, C. Wang, H. Wu, C. Zhao, and Q. Wang, "A vegetable disease recognition model for complex background based on region proposal and progressive learning," Computers and Electronics in Agriculture, vol. 184,2021, doi: 10.1016/j.compag.2021.106101
[34] W. -P. Cao et al., "An ensemble fuzziness-based online sequential learning approach and its application," International Conference on Knowledge Science, Engineering and Management, 2021, pp. 255–267, doi: 10.1007/978-3-030-82136-4_21.
[35] S. Atsawaraungsuk, T. Katanyukul, and P. Polpinit, "Identity activation structural tolerance online sequential circular extreme learning machine for highly dimensional data," Engineering and Applied Science Research, vol. 46, no. 2,pp. 120–129, 2019, doi: 10.14456/easr.2019.15
[36] R. Venkatesan and M. J. Er, "A novel progressive learning technique for multi-class classification," Neurocomputing, vol. 207, pp. 310–321, 2016, doi: 10.48550/arXiv.1609.00085.
[37] M. M. S. m. Mir Mikael Fatemi, "Classification of SSVEP signals using the combined FoCCA-KNN method and comparison with other machine learning methods," *Biomedical Signal Processing and Control,* Vols. Volume 85, August 2023, 104957, 2023.
[38] Yang Ren, Wei Wei, Peng Zhu, Xiuming Zhang, Keyong Chen, Yisheng Liu. "Characteristics, classification and KNN-based evaluation of paleokarst carbonate reservoirs: A case study of Feixianguan Formation in northeastern Sichuan Basin, China." *Energy Geoscience*, 2023: Volume 4, Issue 3, July 2023, 100156.

Vol. 6, No. 6, December 2025, Page. 5882-5898
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.5473

[39] Massami Denis Rukiko a, Adam Ben Swebe Mwakalobo b, Joel Johnson Mmasa. "The impact of Conditional Cash Transfer program on stunting in under five year's poor children." *Public Health in Practice* Volume 6, December 2023, 100437 (2023).

[40] Nimish Sharma, Shruti Shastri, Siddharth Shastri. "Does urbanization level and types of urban settlements matter for child stunting prevalence in India? Empirical evidence based on nighttime lights data." *Cities* Volume 140, September 2023, 104388 (2023).

[41] Zin Wai Htay a, Thinzar Swe b, Thae Su Su Hninn c, Maw Thoe Myar d, Kyi Mar Wai. "Factors associated with syndemic anemia and stunting among children in Myanmar: A cross-sectional study from a positive deviance approach." *Archives de Pédiatrie* Volume 30, Issue 6, August 2023 (2023)