International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

An Enhanced K-Nearest Neighbors (KNN) Algorithm for Predicting Child Malnutrition in The City of Manila

Raizen Joyce R. Daguplo

Bachelor of Science in Computer Science, Computer Science Department, Pamantasan ng Lungsod ng Maynila

Abstract

Malnutrition among children in third-world countries like the Philippines remains a critical issue addressed by the UN's Zero Hunger goal. Traditional methods such as K-Nearest Neighbor face limitations in accuracy due to bias. This study aims to enhance K-Nearest Neighbor's predictive accuracy (KNN) by incorporating Gaussian kernel similarity and normalization techniques to predict malnutrition and BMI categories. Results show that Enhanced KNN consistently outperforms Basic KNN across various K values, achieving an average of 94.33% for bi-class (Malnutrition/Normal) and 92.67% for multi-class (BMI category) compared to Basic KNN's accuracy results: 87.43% for bi-class and 82.51% for multi-class. The Enhanced algorithm performs better across all metrics, averaging an improvement of 8%. Findings revealed that the use of Gaussian kernel enhanced the accuracy of K-Nearest Neighbor because it prioritizes nearer neighbors over distant ones, which reduces the influence of farther points. This enhancement proved that it is effective in malnutrition prediction even without anthropometric measurements, highlighting its potential to address malnutrition in constrained settings.

Keywords: KNN, Machine Learning, Prediction, Algorithm, Malnutrition, Manila, Philippines

1. Introduction

1.1 Background of the Study

Over the years, malnutrition is still prevalent across the world, particularly in third-world countries where different factors affect the nutritional and health status of a child. Based on the World Health Organization, 144 million (21.3%) children that are under the age of five were stunted or too short for their age, while 38 million (5.6%) were overweight.

According to the World Health Organization, undernutrition and overnutrition are the two classes of malnutrition. Undernutrition includes stunting (low height for age), wasting (low weight for height), and underweight (low weight for age). Overnutrition, on the other hand, includes overweight and obesity. According to the National Institute of Health, stunting is the most prevalent type of malnutrition, having 38.5% of stunted Filipino children. Child malnutrition is more common in poorer households than it is in the wealthiest household (Pulok et al., 2016). Based on the 2018 Program for International Student Assessment or PISA, the Philippines ranked the lowest among 79 countries in the field of mathematics, science, and reading. This demonstrates how malnutrition can impact students' academic performance. Malnutrition is covered by the Zero Hunger goal, which is one of the 17 Sustainable Development Goals.

Its objectives are to "end hunger, achieve food security and improved nutrition, and promote sustainable agriculture." This is expected to be achieved by 2030 as malnutrition is a huge threat to the health, growth, and development of a child.

One of the most widely used non-parametric machine learning techniques for regression and classification is the K-Nearest Neighbor (KNN) algorithm. The algorithm predicts the class of a new input data point based on the majority class of its nearest neighbors. It is commonly used in image recognition, recommendation systems, intrusion detection, financial market forecasting, and among other applications. Despite KNN's use in machine learning and prediction, it has its notable challenges such as achieving high accuracy and handling imbalanced datasets. KNN's reliance in majority voting introduces it to these challenges and has become a pain point for the algorithm thus its limitations in performance especially in large imbalanced datasets such as malnutrition datasets. Addressing KNN's challenges can help in scenarios such as malnutrition because of the underrepresentation of malnutrition compared to normal weight which suggests that metrics like precision, recall, and F1 score should also be taken into account.

1.2 Statement of the Problem

In this study, the proponents would like to solve the following:

1.2.1 The K-Nearest Neighbors (KNN) algorithm is prone to be biased in predicting because of its majority voting which affects its accuracy.

1.3 Objective of the Study

1.3.1 The general objective of this study is to improve the K-Nearest Neighbor algorithm's performance across the malnutrition dataset by addressing limited accuracy to improve the algorithm's performance.

1.4 Significance of the Study

This research will benefit the following entities:

Manila Government (LGU)

Local government officials, policymakers, and public health authorities can use the insights gained from study to develop targeted policies and programs aimed at reducing malnutrition rates and improving overall child health in Manila.

Non-Governmental Organizations (NGOs)

NGOs focused on child welfare and nutrition can utilize the study's findings to implement projects such as feeding programs, nutritional counseling, and outreach programs for vulnerable communities.

Department of Health (DOH)

The study can serve as useful information to the DOH to make decisions about policy development and strategies, and budget allocation. The DOH can more effectively target areas with the highest prevalence of malnutrition, ensuring that limited resources are allocated where they are most needed.

Future Researchers

Researchers in the fields of machine learning, public health, and social sciences can benefit from the study's methodology and insights, contributing to further research and academic discourse on predictive methods. Future researchers might build upon and modify the study's approach to solve related problems in various contexts.

1.5 Scope and Limitations

This study focuses on the challenges that are present in the Traditional K-Nearest Neighbor Algorithm. Particularly in the areas of accuracy, performance, scalability, and confidence. This study will explore techniques and methods that will address these issues and come up with an enhancement of the existing algorithm. Specifically, it will address issues with limited accuracy, imbalanced classes, and overfitting.
The researchers will design, choose, and test appropriate techniques that will improve the algorithm.
The effectiveness of the techniques that will be used will be focused on the data that the study will use. It will only focus on the challenges that are mentioned in the objectives and may not be covering all the challenges that may be present in the traditional algorithm. Lastly, external factors that are out of scope are to be considered in the recommendations as it may not be considered by the present study.

2. Review of Related Literature

This chapter contains a review of related literature and studies on an enhanced K-Nearest Neighbor (KNN) algorithm for predicting child malnutrition in Intramuros, Manila, which were collected from various sources pertaining to the study.
According to Ayyad et al. (2022), the classification is done based on the k-nearest neighbor or smallest distances, where k is the number of nearest neighbors. In this study it says that it has many challenges such as slow execution, sensitivity to large data sets, and sensitivity to k. There are existing weighted KNN and according to Huang et. al 2018, it which weights are given correspondingly. The k-nearest neighbor algorithm is useful for assigning missing values and resampling data (Bansal et al., 2022). With the given or labeled dataset, the algorithm predicts the relationship between the unseen data and the known data. Based on that prediction, it assigns the new data to the class that best matches it. It sorts the new data point or figure according to the arrangements of its neighbors. The k in the algorithm represents the number of neighbors of the new data point (Shafi, 2023). The majority class chooses the class label for a new data point from its k nearest neighbors. Therefore, the k-nearest neighbor algorithm can directly classify the query using the information provided by the training set (Pan et al., 2020).
Many academics presented different approaches and methods to address the k-nearest neighbor algorithm's problems. Choosing the optimal value of k is important in determining the number of neighbors in the KNN algorithm. Pan and Wang (2017) suggested an innovative multi-local mean-based k-harmonic nearest neighbor (MLM-KHNN) classifier to improve the classification performance of the KNN algorithm. The proposed KNN-based classifier uses harmonic mean distance as a similarity measure. To reduce the sensitivity of the neighborhood size, the proposed approach uses up to k multi-local means for each class rather than a single local mean. The study produced a decreased classification error rate and was shown to be less sensitive to the parameter k.
The traditional k-nearest neighbor algorithm considers all k nearest neighbors equally when assigning a class label to the query sample using majority vote, which can be problematic. This approach may result in biased classifications. Mateos-García et al. (2016) proposed EvoNN, a modified and optimized voting mechanism for the k closest neighbor method. This distance-weighted voting system assigns bigger weights to the closest neighbors.
To address the negative impact of outliers in the k-nearest neighbor algorithm, Yi et al. (2015) proposed the local mean-based k-nearest neighbor (LMKNN). This approach uses the local mean vector of k nearest neighbors from each class to classify the patterns. Another KNN-based classifier that was developed is the pseudo nearest neighbor algorithm, which is based on distance weighted k-nearest neighbor. This

method finds the pseudo nearest neighbor in each class and classifies the class that is closest to the query point. The LMKNN and PNN classifiers outperform the classic k-nearest neighbor classifier in terms of classification performance and outlier detection (Zhan et al., 2014).

Another instance that k-nearest neighbor has been enhanced is by cross validation and validation set methods in choosing for k. According to Rahim et al. (2022), KNN with this set of enhancements has the potential to be used in the medical field. According to the University of Wisconsin Madison Department of Statistics, Gaussian probability utilizes standard deviation and the square of it or the variance, which is highly used in probability and statistics.

A further study on improving the KNN algorithm is the local mean-based pseudo-nearest neighbor (LMPNN). Gou et al. (2014) introduced a hybrid algorithm KNN-based classifier with the goal of improving the classification performance of the standard method by combining the strengths of the local mean-based k-nearest neighbor algorithm and the pseudo-nearest neighbor. The modified algorithm detects each class's k nearest neighbors, and then computes the corresponding local mean vectors, which serve as class prototypes.

One of the difficulties with the k-nearest neighbor algorithm is its population size. Song et al. (2017) proposed decreasing the size of the training set for the k-nearest neighbor regression technique (DISKR). In this technique, the researchers remove outlier instances that have an impact on the regressor's performance before sorting the remaining instances by the difference in output between instances and their nearest neighbors. Following the sorted order, instances with the lowest contribution to the regressor will be deleted one by one. Because the eliminated instance has an impact on evaluating the contributions of the remaining instances, it is necessary to reassess them. As a result, DISKR provides a simple and effective approach for determining which occurrences have the least and most negative effect on the k-nearest neighbor regressor. The proposed method also removes duplicate instances to improve prediction execution speed.

Rodger's (2014) study on predicting natural gas demand employed nearest neighbors as one of the approaches used by the researcher in cost-saving systems to anticipate natural gas demand. Jiang et al. (2015) developed an improved k-nearest neighbor algorithm for text categorization. The standard KNN algorithm is less efficient when it comes to text categorization, which is why it has been improved. The suggested improved approach combines the single-pass clustering algorithm with the standard k-nearest neighbor algorithm. The study demonstrates that the modified algorithm reduces text similarity and outperforms the Naive Bayes and Support Vector Machine classifiers.

Ramkumar et al. (2022) used a k-nearest neighbor (KNN) classification system to predict recovered and deceased cases of COVID-19. During data preparation, the researchers eliminate missing and outlier data values from the dataset, leaving just the patient's important attributes. The KNN prediction feature is based on the dataset's neighbor data values, with no assumptions made about the dataset. The study's findings revealed that the prediction accuracy of the KNN algorithm was 80.4%. It outperforms the other machine learning algorithms with the lowest error rate of 0.19. The researchers hypothesized that disease may be predicted based on symptoms.

Park et al. (2015) proposed a fast collaborative filtering approach based on a k-nearest neighbor graph for the recommender system. KNN was used to estimate users' preferences for unrated products. This study reverses the process of determining the k neighbors, with the suggested model identifying the k-nearest neighbors of rated items. Meanwhile, Cai et al. (2020) offered another investigation using k-nearest

neighbors. It is called the k-reciprocal nearest neighbor algorithm (k-RNN) and is used in recommender systems. The modified KNN decides whether the two datasets are similar.

One of the challenges of the typical k-nearest neighbor approach is removing noise from data, which improves data quality. Noises can be created by a variety of sources, including entry errors, missing information, label issues, and more (Leevy, 2021). To address this issue, Zhang and Liu (2016) proposed a mutual k-nearest neighbor algorithm based on the k-nearest neighbor framework. The method removes recognized anomalies from the database.

K-nearest neighbor has been used across various fields such as health care. In a study done by Florimbi (2018), KNN has been used to classify human brain tumors. The researchers used a search window as a way to optimize the neighbor selection. According to Taneja et. al KNN's accuracy can be enhanced using a blend of classification and clustering techniques which arguably performed better than the conventional KNN algorithm.

3. Methodology

3.1 Research Design

This study follows a quantitative research design to enhance the performance of K-Nearest Neighbor applied in children malnutrition. This research also evaluates both the Basic KNN and Enhanced KNN. The design focuses on modifying KNN to address issues with accuracy, as well as other metrics such as precision, recall, and F1 score.

The research was designed in experimental approach wherein both algorithms are tested on similar dataset ensuring that the environment is the same. It follows an approach where preliminary requirements such as problems, objectives, literature are first gathered, followed by the process or the methodology which consists of data collection, preprocessing, algorithm enhancement and validation. Lastly, analysis of the results and final conclusions are derived from the algorithm validation. Through numerical metrics, the study provides insights into the effectiveness of the enhanced KNN.

3.2.1 Pseudocode of Traditional KNN Algorithm

K Nearest Neighbor Pseudocode

Input: Training data, Test data

Output: Class of the test data

Steps:

1. Load the training data and test data
2. Choose the value of K
3. For each point in test data:
    1) Find the Euclidean distance to all training data points
    2) Store the Euclidean distances in a list and sort it from nearest to farthest.
    3) Choose the first k points
    4) Assign a class to the test point based on the majority of classes present in the chosen points
4. End

3.2.2 Pseudocode of Proposed Enhanced KNN Algorithm

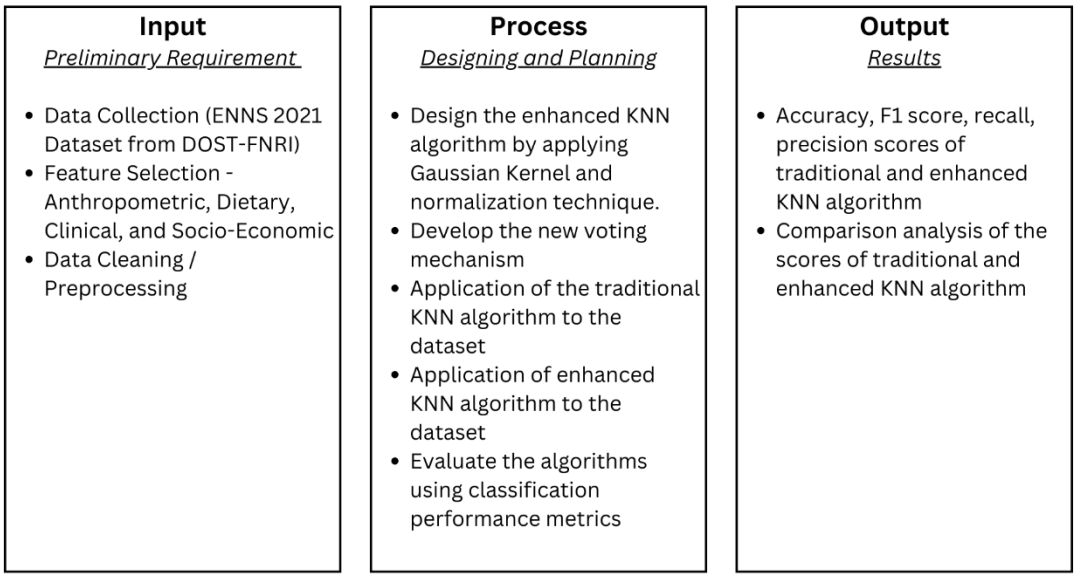
Enhanced KNN Algorithm

Input: Training data (X_train, y_train), Test data (X_test), Number of neighbors (K), Gaussian kernel parameter (sigma)
Output: Predicted class and probability for each test instance
Steps:

1. Load the training data (X_train, y_train) and test data (X_test).
2. Choose the number of neighbors (K) and Gaussian kernel parameter (sigma).
3. For each point x in X_test:
   - Calculate the Euclidean distance between x and each point in X_train.
   - Compute the similarity value for each training instance using the Gaussian kernel formula:
   - Normalize the similarity scores of all training instances to make them proportional.
   - Select the top K neighbors based on the highest similarity scores.
   - Combine the similarity scores of the K neighbors by grouping them by class.
   - Normalize the combined scores to convert them into probabilities, ensuring they sum to 1.
   - Compare the probabilities for each class.
5. Return the class label with the highest probability as the predicted class.
6. Output the predicted class and its corresponding probability.

3.4 Methods and Tools

Figure 1: IPO Model of an enhanced K-Nearest Neighbor algorithm



KNN Algorithm suffers from imbalanced datasets, overfitting, and selecting the right ‘k’. To solve these problems, the researchers chose techniques such as Gaussian Kernel Formula, Normalization, and Comparison.

The methods used in this research can be divided into three parts using the Input-Process-Output model. The input phase requires the Data Collection which consists of gathering data from the ENNS Dataset from DOST-FNRI, followed by feature selection process, and lastly data cleaning/processing for missing values or duplicated entries. After gathering all the requirements, it will proceed to the second part which is the designing and planning or the ‘process’ which includes the designing and application of the enhanced

dataset, Lastly, the output phase is where the prediction of malnutrition rate and the software system that will be integrating the enhanced KNN algorithm.

3.4.1 Data Collection

The data was compiled from the Expanded National Survey produced by the Department of Science and Technology – Food and Nutrition Research Institute via the e-Nutrition website public use files. The ENNS covers various topics such as Anthropometric, Dietary, Clinical, and Socio-Economic Surveys. For this study, the dataset was filtered to include children aged 5 to 10 years old situated in the City of Manila. Features are selected to curate the dataset, and variables are chosen according to the study's scope.

The dataset includes the following variables: weight, height, age of child, sex of child; household status such as type of dwelling (single house, duplex, multi-unit residential, commercial/industrial/agricultural, institutional living quarters, or other housing units), tenure status of the dwelling unit (owned/amortized/owner-like possession, rented, rent-free with consent, rent-free without consent), main roofing material (salvaged/makeshift material, mixed but predominantly salvaged materials, light materials such as cogon/nipa/anahaw, mixed but predominantly light materials, mixed but predominantly strong materials, and strong materials), main material used for the wall (salvaged/makeshift material, mixed but predominantly salvaged materials, light materials such as cogon/nipa/anahaw, mixed but predominantly light materials, mixed but predominantly strong materials, and strong materials), main flooring material (earth/sand, wood planks, coco lumber/bamboo, marble, parquet or polished, ceramic tiles, vinyl or asphalt strips, plain cement), number of bedrooms, main type of fuel used for cooking (not cooking, electricity, liquified petroleum gas, natural gas, kerosene, charcoal, wood, agricultural crop, animal dung, or others), main source of drinking water, type of toilet facility, household wealth quantile, and lastly, head of household's age, sex, civil status code, highest educational attainment, and current work. These variables are carefully merged from various datasets from the ENNS 2018,2019, and 2021 face-to-face sample survey data released in 2023. The data were thoroughly cleaned and processed, ensuring that they are properly joined and have no missing values.

3.4.2 Enhanced KNN Algorithm Methods

The following discusses how the enhanced KNN algorithm was modified.

3.4.2.1 Euclidean Distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (x\_i - x'\_i)\text{^}2}$$

This study utilizes Euclidean Distance as the distance metric for both the traditional and enhanced KNN Algorithms. Euclidean Distance was chosen due to its effectiveness in measuring distance between two data points. In KNN, Euclidean distance is used to get the distance between the testing data and its neighbors. The p and q symbolize the test data and its neighbor. Getting the distance between the two points is a key step in the KNN algorithm as it is the reference to know the nearest neighbor.

3.4.2.2 Similarity Value using Gaussian Kernel Formula

$$K(x,x') = exp(-\frac{\|x - x'\|^2}{2\sigma^2})$$

To give the neighbors a similarity value, the use of Gaussian Kernel Function is utilized. Gaussian Kernel Function was utilized to address the issue of majority voting that leads to bias. This function was used to capture the relationship of two points and convert it to a value that says how close the neighbor is to the

reference test data. The function's basis is the distance, and it projects it into an inverse relationship with similarity. It suggests that the similarity value decreases as the distance between the points increases. The similarity value from this function ranges from 0 to 1 which is the normalization part which helps in reducing the impact of noise and outliers that decreases the accuracy of the algorithm. It can be interpreted that points that are higher in similarity most likely belong to a similar class and the lower similarity suggests the opposite.

3.4.2.3 Combining Similarity values and normalizing to become percentage

$$Normalized\ Score_i = \frac{Score_i}{\sum_{\{j=1\}}^{k} Score_j}$$

After getting the Similarity values and its normalized value, the next step is to combine all of the K neighbors' values by summation and normalizing it to all be equal to 1 which is done by dividing each score by the total sum of scores. This means that all distances are summated, and the degree of its influence is turned into a percentage of probability. Normalizing the score helps in assessing confidence in prediction and generates a probabilistic output.

3.4.3 Algorithm Validation

Both Traditional and Enhanced KNN will be performed on the same dataset across different k and different testing training ratios. It will then be cross validated to compare the accuracy scores across different parameters. This will determine how the proposed algorithm compares to the accuracy, precision, recall, and F1-score.

7. Results and Discussion

4.1 Enhancing Accuracy through Algorithm Enhancement and Performance Evaluation

Table 1: KNN Metrics: Malnutrition as Target (Bi-Class) for K = 1 to 11

Metric	Enhanced KNN Algorithm (K=1)	Basic KNN (K=1)	Enhanced KNN Algorithm (K=3)	Basic KNN (K=3)	Enhanced KNN Algorithm (K=5)	Basic KNN (K=5)
Accuracy	93.14%	93.14%	93.14%	89.60%	92.67%	85.34%
Precision	93.09%	93.09%	93.09%	89.64%	92.61%	85.35%
Recall	93.14%	93.14%	93.14%	89.60%	92.67%	85.34%
F1 Score	93.10%	93.10%	93.09%	89.25%	92.59%	84.53%
Metric	**Enhanced KNN Algorithm (K=7)**	**Basic KNN (K=7)**	**Enhanced KNN Algorithm (K=9)**	**Basic KNN (K=9)**	**Enhanced KNN Algorithm (K=11)**	**Basic KNN (K=11)**
Accuracy	92.43%	84.40%	92.20%	82.27%	92.43%	82.27%
Precision	92.41%	84.94%	92.18%	82.49%	92.45%	83.20%
Recall	92.43%	84.40%	92.20%	82.27%	92.43%	82.27%
F1 Score	92.30%	83.12%	92.05%	80.70%	92.28%	80.33%

Tables 1 shows the performance of the Enhanced KNN Algorithm and the Basic KNN Algorithm when applied to the Malnutrition dataset. The target variable of this classification is Malnutrition status which

is a bi-class classification problem (e.g., Malnourished or Not Malnourished). The results highlight the effectiveness of both algorithms across various K values (1, 3, 5, 7, and 11) using metrics such as Accuracy, Precision, Recall, and F1 Score.

Table 1.3 Average Metrics Scores of Enhanced KNN and Basic KNN

Metric	Enhanced KNN Algorithm (Average)	Basic KNN (Average)
Accuracy	92.67%	86.17%
Precision	92.64%	86.45%
Recall	92.67%	86.17%
F1 Score	92.57%	85.17%

The scores show that Enhanced KNN outperforms Basic KNN in bi-class classification for Malnutrition Status, with an average of 6.50% for accuracy, 6.19% for precision, 6.50% for recall, and 7.40% for F1. Table 1.3 also shows the performance of Enhanced KNN as superior to Basic KNN. On average, the Enhanced KNN achieves an Accuracy of 92.67%, compared to 86.17% for Basic KNN. Similarly, the Enhanced KNN delivers Precision of 92.64% and Recall of 92.67%, outperforming Basic KNN's respective scores of 86.45% and 86.17%. The Enhanced KNN also excels in F1 Score, achieving an average of 92.57%, compared to 85.17% for Basic KNN.

Figure 2: Performance Metrics of Enhanced KNN and Basic KNN for K = 1 to K = 11

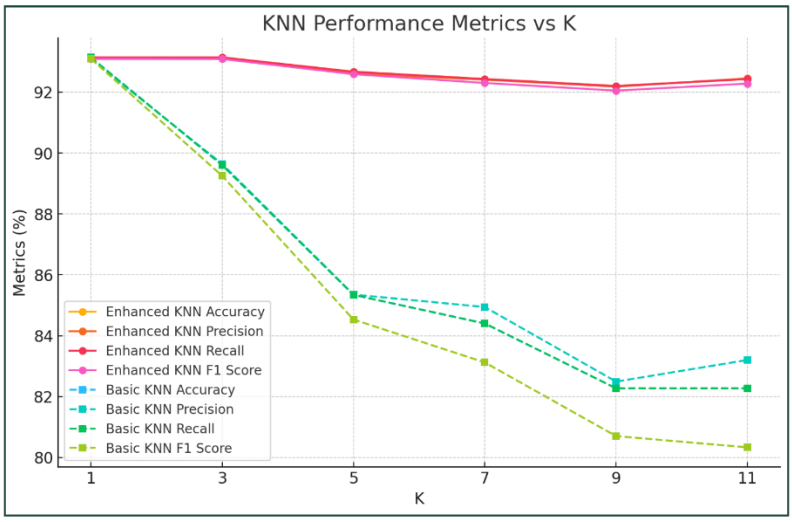


Figure 2 highlights the performance of Enhanced KNN and Basic KNN using a line graph. This figure illustrates the stability of Enhanced KNN while Basic KNN exhibits a steady decline for K > 3.

Enhanced KNN algorithm shows consistency in high performance with all metrics above 92% across all K values while Basic KNN suffers with in decreasing performance across all metrics showing a steep drop while the K increases falling as low as 80.33% by K = 11.

Another approach is by calculating the performance of both algorithms without key direct features such as height and weight. This emphasizes the capability of the algorithm to predict Malnutrition in diverse

and resource-constrained settings such as only using socio-economic or clinical data. The following are the results:

Table 2: KNN Metrics: Malnutrition as Target without Height and Weight for K = 1 to 5

Metric	Enhanced KNN Algorithm (K=1)	Basic KNN (K=1)	Enhanced KNN Algorithm (K=3)	Basic KNN (K=3)	Enhanced KNN Algorithm (K=5)	Basic KNN (K=5)
Accuracy	83.69%	83.69%	84.63%	80.85%	84.16%	78.72%
Precision	83.58%	83.58%	84.47%	80.38%	84.00%	77.76%
Recall	83.69%	83.69%	84.63%	80.85%	84.16%	78.72%
F1 Score	83.63%	83.63%	84.54%	80.53%	84.07%	77.70%

Table 2.1: KNN Metrics: Malnutrition as Target without Height and Weight for K = 7 to 11

Metric	Enhanced KNN Algorithm (K=7)	Basic KNN (K=7)	Enhanced KNN Algorithm (K=9)	Basic KNN (K=9)	Enhanced KNN Algorithm (K=11)	Basic KNN (K=11)
Accuracy	84.63%	78.01%	84.40%	75.89%	84.63%	75.18%
Precision	84.38%	76.96%	84.08%	74.47%	84.28%	73.52%
Recall	84.63%	78.01%	84.40%	75.89%	84.63%	75.18%
F1 Score	84.46%	76.37%	84.16%	73.31%	84.33%	72.46%

Tables 2 and 2.1 show the metrics score for both Enhanced KNN and Basic KNN algorithm for classifying malnutrition status without the height and weight features. The table compares the performance of both algorithms highlighting the difference in favor of Enhanced KNN across all metrics (Accuracy, Precision, Recall, and F1 Score).

Table 2.3 Average Metrics Scores of Enhanced KNN and Basic KNN		
Metric	**Enhanced KNN Algorithm (Average)**	**Basic KNN (Average)**
Accuracy	84.36%	78.72%
Precision	84.13%	77.78%
Recall	84.36%	78.72%
F1 Score	84.20%	77.34%

Table 2.3 reveals that there is a drop in performance in all metrics after removing height and weight as one of the features in classifying malnutrition status. The drop in performance indicates that height and weight contribute highly to the algorithm's capability in classifying. Despite the drop, Enhanced KNN Algorithm still performs better than Basic KNN across all metrics averaging 5.63% in accuracy, 6.35% in precision, 5.63% in recall, and 6.86% in F1 Score. The difference is still significant, suggesting that the Enhanced KNN is a better choice in situations where direct features are unavailable.

Figure 3: Performance Metrics of Enhanced KNN and Basic KNN for K = 1 to K = 11 Malnutrition as target without Height and Weight

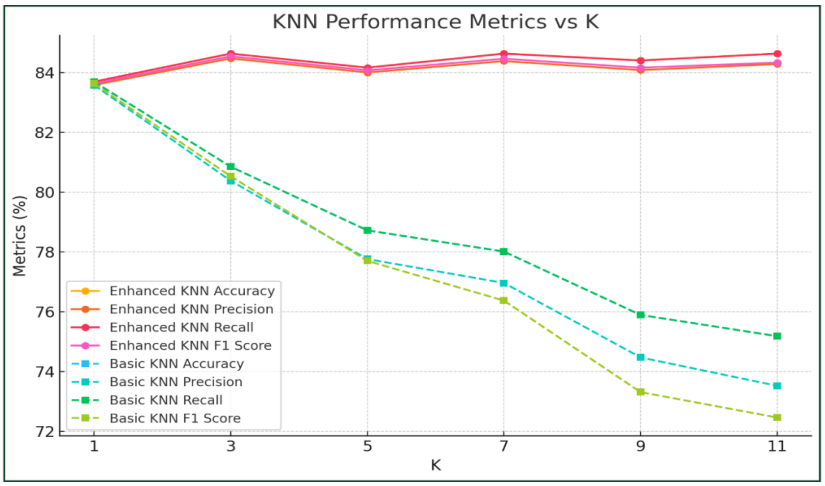


Figure 3 illustrates the comparison of Enhanced KNN and Basic KNN in predicting malnutrition without height and weight as features. This shows that it still follows the trend of Figure 2 wherein Enhanced KNN shows stability and robustness across all K's while Basic KNN decreases in all metrics while K increases, showing a huge gap in all metrics.

The same trend can be seen in applying algorithms in multi-class classification problems such as setting BMI categories as targets.

Table 3: KNN Metrics: BMI Category as Target for K = 1 to 11						
Metric	**Enhanced KNN Algorithm (K=1)**	**Basic KNN (K=1)**	**Enhanced KNN Algorithm (K=3)**	**Basic KNN (K=3)**	**Enhanced KNN Algorithm (K=5)**	**Basic KNN (K=5)**
Accuracy	92.91%	92.91%	92.91%	87.47%	92.43%	83.22%

Precision	92.82%	92.82%	92.79%	87.22%	92.33%	82.69%
Recall	92.91%	92.91%	92.91%	87.47%	92.43%	83.22%
F1 Score	92.79%	92.79%	92.76%	86.62%	92.23%	81.24%
Metric	**Enhanced KNN Algorithm (K=7)**	**Basic KNN (K=7)**	**Enhanced KNN Algorithm (K=9)**	**Basic KNN (K=9)**	**Enhanced KNN Algorithm (K=11)**	**Basic KNN (K=11)**
Accuracy	92.20%	82.03%	92.20%	78.96%	92.43%	78.96%
Precision	92.09%	82.18%	92.12%	75.91%	92.45%	76.17%
Recall	92.20%	82.03%	92.20%	78.96%	92.43%	78.96%
F1 Score	91.88%	79.24%	91.84%	75.42%	92.07%	74.68%

Table 3.1 Average Metrics Scores of Enhanced KNN and Basic KNN for BMI Category as Targets

Metric	Enhanced KNN Algorithm (Average)	Basic KNN (Average)
Accuracy	92.51%	83.93%
Precision	92.43%	82.83%
Recall	92.51%	83.93%
F1 Score	92.26%	81.67%

Table 3 shows the result of testing the Enhanced KNN and Basic KNN in classifying for BMI Category (Underweight, Healthy Weight, Overweight, and Obese). This can be classified as multiclass problem. Table 3.1 demonstrates that Enhanced KNN outperforms Basic KNN in all metrics consistently with the Enhanced KNN reaching 92.51% in accuracy while Basic KNN only reaching 83.93% in the same metric. The gap between the enhanced and basic averages is 8 to 10% which shows significant difference in multi-class classifications. The same trend can be seen in testing the BMI category without the direct features such as height and weight.

Table 4: Average Metrics Scores of Enhanced KNN and Basic KNN for BMI Category as Targets without height and weight

Metric	Enhanced KNN Algorithm (Average)	Basic KNN (Average)

Accuracy	83.26%	75.53%
Precision	83.01%	72.41%
Recall	83.26%	75.53%
F1 Score	82.99%	71.82%

Figure 4: KNN Performance Metrics of Enhanced KNN and Basic KNN with BMI as target

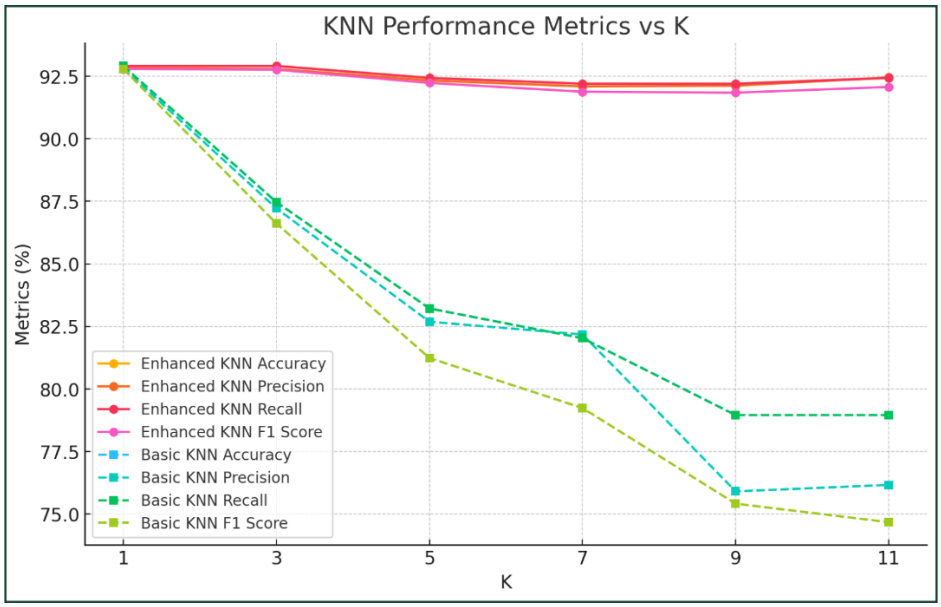
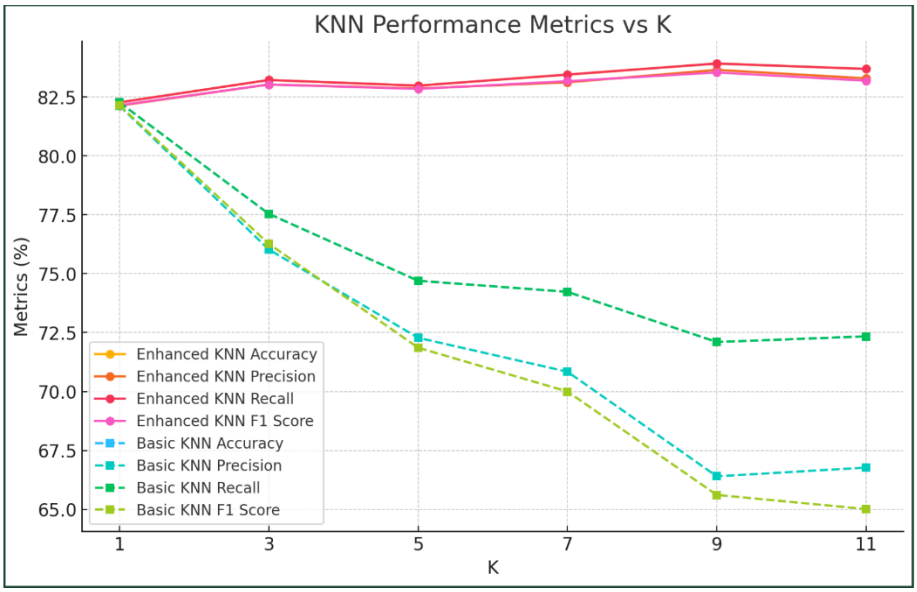


Figure 5: KNN Performance Metrics of Enhanced KNN and Basic KNN with BMI as target without height and weight



Figures 4 and 5 exhibit the same pattern wherein the Enhanced KNN showed consistent performance in metrics and across all Ks while Basic KNN showed declining performance when K becomes larger. Noticeably, both algorithms drop in performance datasets where direct features are excluded such as height and weight. In this scenario, both algorithms dropped 10% in all metrics. In matter of robustness and performance, Enhanced KNN proves to be a better choice in classifying multiclass targets such as BMI category.

The results gathered from the algorithm evaluation of both Enhanced KNN and Basic KNN prove that the Enhanced KNN is a competitive variant of KNN when it comes to malnutrition prediction such as bi-class and multi-class problems such as malnutrition as target and BMI category as target. The scores reveal that the enhancement that was applied to the basic KNN made it score higher in metrics and show consistent performance compared to the traditional algorithm.

Gaussian Kernel function contributes to the distance weighing attribute of the algorithm which makes its overall effectiveness and minimizes errors in classification because of its characteristic that prioritizes nearer neighbors over distant once. It reduces the influence of father, less relevant points making it an ideal enhancement in terms of increasing the ability of the algorithm to classify classes such as malnutrition. Gaussian Kernel in KNN is ideal showing on the results mentioned above because of its benefits such as Distance-Weighted Voting contributed by its formula which consist of weight, distance, and sigma parameter (used for smoothing). The same approach was observed in KDF-KNN or Kernel Difference Weighted KNN which explored the limitations of traditional distance-weighted KNN and enhanced it by applying a kernel which increased its ability to handle nonlinear structures (He et al., 2007) Moreover, RK-KNN also proved the use of kernel functions, random feature selection, and bootstrapping as an effective method in reducing errors and improving prediction performance, thereby enhancing the overall effectiveness of KNN as an algorithm (Wang et al., 2023)

4.1.2 Feature Importance and Key Points in Malnutrition

Figure 6 Feature Importance of Enhanced KNN

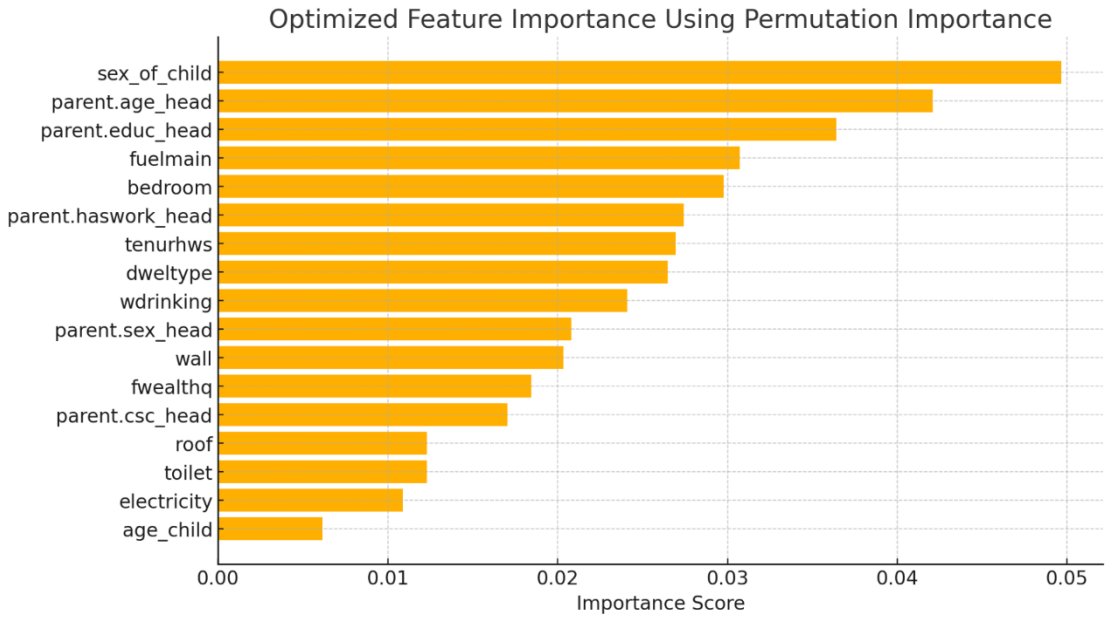


Figure 6 highlights the key features that significantly influence the Enhanced KNN algorithm's predictions of malnutrition. These features include the child's sex, which reveals gender-related differences in malnutrition susceptibility; the head of household's age and education level, both of which impact the household's capacity to provide adequate nutrition; the type of fuel used for cooking, which reflects living standards and household resources; and the number of bedrooms, an indicator of household wealth and living conditions.

Trends reveal that the primary type of cooking fuel plays a role in malnutrition, with "not cooking" and

natural gas usage being significant in specific malnutrition classes, while clean fuels are associated with normal weight. An inverse relationship exists between wealth and malnutrition, as higher wealth levels correspond to lower rates of malnutrition. Toilet type also shows a correlation, with households using water-sealed toilets and other depositories exclusively for their use being linked to better nutrition outcomes, while inadequate sanitation facilities are associated with higher malnutrition rates. Drinking water sources further highlight disparities, as households that predominantly rely on refilling stations for water supply exhibit better nutritional statuses compared to those with less reliable sources. Parental factors also influence malnutrition, with younger parents having a higher prevalence of malnourished children. Gender analysis reveals that females are more prone to malnutrition compared to males, suggesting a need for targeted interventions for girls. The current work status of the head of the household, particularly being unemployed or a student, has the strongest influence on malnutrition. Finally, malnutrition is more prevalent when the head of the household has a low level of education, having only completed or not finished elementary school.

8. Conclusion and Recommendations

The study successfully enhanced the k-nearest neighbor algorithm's performance and accuracy by implementing a gaussian kernel-based weighting mechanism and normalization techniques. The gaussian kernel successfully improved the algorithm's prediction by emphasizing closer neighbors, while normalization enhanced its stability by addressing feature scaling issues. The enhanced algorithm outperformed the traditional KNN algorithm in all metric platforms, making it suitable for large imbalanced datasets such as the dataset for child malnutrition. This study showed that malnutrition and classification of BMI can be predicted by selecting features such as household features and head of household's characteristics. However, the algorithm's performance depends on hyperparameter tuning, and its computational cost increased. Future work should work on optimizing the hyperparameters and improving their efficiency.

References

1. Alcantara, N. D. M., Arenque, L. A. B., & Sicat, J. Z. (2023). PE-M: Prediction of malnutrition of Barangay 35 Tondo Manila using random forest algorithm. *World Journal of Advanced Research and Reviews, 18*(3), 561–567. https://doi.org/10.30574/wjarr.2023.18.3.1130
2. Almomany, A., Ayyad, W. R., & Jarrah, A. (2022). Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study. *Journal of King Saud University - Computer and Information Sciences, 34*(6), 3815–3827. https://doi.org/10.1016/j.jksuci.2022.04.006
3. Classification of malnutrition. (n.d.). *ResearchGate*. https://www.researchgate.net/figure/Classification-of-Malnutrition_tbl1_256447083
4. Florimbi, G., Fabelo, H., Torti, E., Lazcano, R., Madroñal, D., Ortega, S., Salvador, R., Leporati, F., Danese, G., Báez-Quevedo, A., Callicó, G. M., Juárez, E., Sanz, C., & Sarmiento, R. (2018). Accelerating the K-Nearest Neighbors filtering algorithm to optimize the real-time classification of human brain tumor in hyperspectral images. *Sensors, 18*(7), 2314. https://doi.org/10.3390/s18072314
5. Gaupholm, J., Dodd, W., Papadopoulos, A., & Little, M. (2023). Exploring the double burden of malnutrition at the household level in the Philippines: Analysis of National Nutrition Survey data. *PLOS ONE, 18*(7), e0288402. https://doi.org/10.1371/journal.pone.0288402

6. He, X., Zhang, S., & Yu, J. (2007). Kernel difference-weighted KNN classification. *Pattern Analysis and Applications, 10*(3), 239–246. https://doi.org/10.1007/s10044-007-0100-z
7. Karabulut, B., Arslan, G., & Ünver, H. M. (2019). A weighted similarity measure for K-Nearest Neighbors algorithm. *Celal Bayar Üniversitesi Fen Bilimleri Dergisi, 15*(4), 393–400. https://doi.org/10.18466/cbayarfbe.618964
8. Li, S., Du, Q., & Sun, H. (2009). An improved kernel-based KNN for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing, 47*(8), 2088–2098. https://doi.org/10.1109/TGRS.2009.2019280
9. Lonang, Syahrani, Yudhana, Anton, & Biddinika, Muhammad. (2023). Performance analysis for classification of malnourished toddlers using K-Nearest Neighbor. *Scientific Journal of Informatics, 10*, 313–322. https://doi.org/10.15294/sji.v10i3.45196
10. Melek, M., Melek, N., & Kayikcioglu, T. (2017). A novel simple method to select optimal k in K-Nearest Neighbor classifier. *International Journal of Computer Science and Information Security, 15*, 464–469.
11. Meng, D., & Li, Y. (2022). An imbalanced learning method by combining SMOTE with center offset factor. *Applied Soft Computing, 120*, 108618. https://doi.org/10.1016/j.asoc.2022.108618
12. Park, S. E., Kim, S., Ouma, C., Loha, M., Wierzba, T., & Beck, N. (2012). Community management of acute malnutrition in the developing world. *Pediatric Gastroenterology, Hepatology & Nutrition, 15*, 210–219. https://doi.org/10.5223/pghn.2012.15.4.210
13. Rahim, R., & Ahmar, A. S. (2022). Cross-validation and validation set methods for choosing K in KNN algorithm for healthcare case study. *Journal of Information and Visualization, 3*(1), 57–61. https://doi.org/10.35877/454ri.jinav1557
14. Syaliman, K., Nababan, E., & Sitompul, O. (2018). Improving the accuracy of K-Nearest Neighbors using local mean based and distance weight. *Journal of Physics: Conference Series, 978*, 012047. https://doi.org/10.1088/1742-6596/978/1/012047
15. Taneja, S., Gupta, C., Goyal, K., & Gureja, D. (2014). An enhanced K-Nearest Neighbor algorithm using information gain and clustering. *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, 325–329. https://doi.org/10.1109/ACCT.2014.22
16. Wang, Y., Zhang, T., & Chen, H. (2023). Enhancing regression performance with Random Kernel KNN. *Frontiers in Big Data, 5*(2), 1402384. https://doi.org/10.3389/fdata.2024.1402384
17. World Health Organization (WHO). (2024, March 1). Malnutrition. https://www.who.int/news-room/fact-sheets/detail/malnutrition
18. Zhang, A., Yu, H., Zhou, S., Huan, Z., & Yang, X. (2022). Instance weighted SMOTE by indirectly exploring the data distribution. *Knowledge-Based Systems, 249*, 108919. https://doi.org/10.1016/j.knosys.2022.108919
19. Zhang, S. (2021). Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering, 34*, 4663–4675.