# Improving nutrient prediction models with polynomial and ratio features and mRMR selection

**Fatma Indriani[1], Irwan Budiman[1], Dwi Kartini[1], Lilies Handayani[2,3]**

[1]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, Banjarmasin, Indonesia
[2]Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan
[3]Department of Statistics, Tadulako University, Palu, Indonesia

**Article Info**

***Article history:***

Received Aug 23, 2024
Revised Oct 28, 2024
Accepted Nov 19, 2024

***Keywords:***

Classification
Feature engineering
Machine learning
Minimum redundancy maximum relevance feature selection
Nutrient prediction

## ABSTRACT

Due to limited space and regulations, food labels often lack information on micronutrients, i.e., vitamins and minerals. Accurately predicting missing these micronutrient data is essential yet challenging. This study explores the feasibility of using machine learning to predict these missing nutrients based on a limited reported nutrient (protein and carbs). Using the *Tabel Komposisi Pangan Indonesia* (TKPI) dataset, we evaluated the performance of 12 diverse classifiers to predict binary classes ("low" or "high") for 13 target micronutrients. Random forest emerged as the best performing classifier with an average accuracy of 0.7421 across all target nutrients. Additionally, we introduced feature engineering techniques by incorporating polynomial and ratio features to enhance model performance. Minimum redundancy maximum relevance (mRMR) feature selection was then applied to identify the most informative features. This approach boosted the average accuracy of the random forest classifier to 0.7591. These findings highlight the efficacy of feature engineering and selection in enhancing nutrient prediction models, demonstrating the potential to improve consumer knowledge about unknown nutrients in food.

***Corresponding Author:***

Fatma Indriani
Department of Computer Science, Faculty of Mathematics and Natural Sciences
Lambung Mangkurat University
A. Yani Street Km 36, Banjarbaru, Banjarmasin, Indonesia
Email: f.indriani@ulm.ac.id

## 1. INTRODUCTION

Complete nutritional information on food labels is crucial for consumers to make informed dietary choices [1], [2]. Food labels provide essential data about the macronutrients (such as carbohydrates, proteins, and fats) and micronutrients (such as vitamins and minerals) contained in food products. This information helps consumers maintain balanced diets, manage specific health conditions, and adhere to dietary guidelines. For example, individuals with conditions such as diabetes, hypertension, or cardiovascular disease rely on accurate nutritional labels to manage their intake of sugar, sodium, and fat [3], [4]. In addition, personalized food recommender systems can utilize this nutritional information to suggest healthier meal options tailored to individual dietary needs [5].

However, current food labels often lack comprehensive data, particularly for micronutrients. Regulatory guidelines typically mandate the inclusion of macronutrient information and a select few micronutrients, leaving many vitamins and minerals unreported. As a case in point, the Indonesian regulation "*Peraturan Badan Pengawas Obat Dan Makanan Nomor 26 Tahun 2021 Tentang Informasi Nilai Gizi Pada*

*Label Pangan Olahan* (2021 Regulation of the National Agency of Drug and Food Control Number 26 on Nutritional Information on Processed Food Labels)" [6] specifies that the mandatory nutrients that must be reported on food labels are total energy, total fat, saturated fat, protein, total carbohydrates, sugars, and salt (sodium). As a result, many food companies only fulfill these minimum requirements. Furthermore, companies may also omit many nutrients due to space constraints on packaging and the cost of extensive nutritional testing. Consequently, consumers may not have access to crucial information needed for optimal nutrition.

The challenge of incomplete or missing nutrient data on food labels is further compounded when only a limited number of input features are available. Previously, Razavi and Xue [7] experimented with regression and classification models to predict vitamins and minerals from 14 other nutrients commonly available on food labels. However, that study was conducted in the USA, where there are more mandatory nutrient requirements on labels. There is a need for a study that is interested in models that can work with less input features.

Existing research efforts aimed at predicting nutrient content have explored various approaches. For instance, some methods use computer vision techniques to analyze food images and predict nutrient composition based on visual features [8]-[13]. Other approaches involve parsing recipe ingredients and quantities to calculate nutritional values [14]-[18]. While these methods can be effective, they require specific types of input data that are not always readily available or practical for all food products.

Given these constraints, our study aims to develop models capable of predicting unreported micronutrients using a minimal set of input features: energy, protein, fat, carbohydrate, fiber, and natrium (sodium). By focusing on these commonly available nutrients, we aim to provide a practical solution for improving customer knowledge about the nutritional content of food. This approach allows consumers to gain additional and more useful information regarding the nutrients in their food without the need for extensive and costly testing. To achieve this, it is important to compare various machine learning classifiers to determine which performs best with the limited input features. Additionally, feature engineering techniques and feature selection can enhance the model's predictive capabilities [19]. This study makes two key contributions: i) an evaluation of 12 machine learning classifiers for nutrient prediction from limited input features, with random forest emerging as the best performer and ii) an improvement in prediction accuracy through the addition of ratio and second-degree polynomial features, combined with minimum redundancy maximum relevance (mRMR) feature selection.

In the following sections, we present a comprehensive examination of our approach to nutrient prediction. Section 2 outlines the method, including the dataset description, feature engineering techniques, and the application of mRMR feature selection. We then detail the experimental setup and classifier comparison in section 3. Section 4 presents the results and discussion, providing a comparison of classifier performance and the impact of feature engineering on predictive accuracy. Finally, section 5 offers conclusions, as well as the implications of our findings and potential avenues for future research.

## 2. METHOD

### 2.1. Dataset

The data for this study is based on the *Tabel Komposisi Pangan Indonesia* (TKPI) 2017 [20], a comprehensive database that provides detailed nutritional information about various foods commonly consumed in Indonesia. It contains 1,146 food items categorized into 14 groups. Each entry in the dataset includes the name of the food item, portion size, and amounts of various nutrients per 100 grams or per serving.

For this study, we use 6 features as input variables (energy, protein, fat, carbohydrate, fiber, and sodium) and 13 features as target variables (calcium, phosphorus, iron, potassium, copper, zinc, retinol, beta-carotene, total carotenoids, thiamin, riboflavin, niacin, and vitamin C). Many of the variables have missing values and most variables exhibit a highly right-skewed distribution. The statistics of these variables are shown in Table 1.

### 2.2. Polynomial features

Polynomial features are a feature engineering technique used to capture non-linear relationships between input variables and the target variable. This method creates new features by combining existing features through multiplication, allowing the model to learn more complex patterns in the data [21], [22].

In general, for a set of n features [$x_1$, $x_2$, ..., $x_n$], second-degree polynomial features include:

- The original features: $x_1$, $x_2$, ..., $x_n$
- Squared terms: $x_1^2$, $x_2^2$, ..., $x_n^2$
- Interaction terms: $x_1x_2$, $x_1x_3$, ..., $x_1x_n$, $x_2x_3$, ..., $x_2x_n$, ..., $x_{n-1}x_n$

Table 1. Data description of all features in the dataset

| Feature names | Type of features | Empty | Mean | Standard deviation | Minimum | 25% | 50% | 75% | Maximum |
|---|---|---|---|---|---|---|---|---|---|
| Energy (cal) | input | 0 | 198.45 | 162.39 | 8 | 74 | 142.5 | 320 | 902 |
| Protein (g) | input | 1 | 9.67 | 11.51 | 0 | 1.7 | 4.8 | 14.5 | 74.3 |
| Fat (g) | input | 1 | 7.40 | 13.70 | 0 | 0.5 | 2 | 8.1 | 100 |
| Carbohydrate (g) | input | 0 | 23.76 | 25.00 | 0 | 4.7 | 13 | 35.575 | 94 |
| Fiber (g) | input | 221 | 2.75 | 4.55 | 0 | 0.1 | 1.4 | 3.4 | 46.5 |
| Sodium (mg) | input | 40 | 152.41 | 338.66 | 0 | 20 | 60 | 153 | 4608 |
| Calcium (mg) | target | 40 | 166.18 | 217.29 | 0 | 43 | 97 | 202.75 | 1976 |
| Phosphorus (mg) | target | 43 | 3.38 | 5.78 | 0 | 0.8 | 1.9 | 3.8 | 96.4 |
| Iron (mg) | target | 282 | 214.70 | 713.30 | 0 | 7 | 34.5 | 164 | 11998 |
| Potassium (mg) | target | 323 | 341.31 | 625.95 | 0 | 91 | 227 | 404.35 | 11890 |
| Copper (mg) | target | 351 | 11.07 | 160.13 | 0 | 0.1 | 0.2 | 0.4 | 4303.45 |
| Zinc (mg) | target | 347 | 1.42 | 3.26 | 0 | 0.3 | 0.7 | 1.5 | 60 |
| Retinol (mcg) | target | 605 | 246.63 | 1575.52 | 0 | 0 | 0 | 39 | 24242 |
| Beta-carotene (mcg) | target | 467 | 634.94 | 2083.39 | 0 | 0 | 11 | 168.5 | 30200 |
| Total caretenoids (mcg) | target | 539 | 1225.21 | 2778.66 | 0 | 0 | 87 | 936 | 28000 |
| Thiamin (mg) | target | 88 | 0.40 | 6.15 | 0 | 0.05 | 0.1 | 0.27 | 200 |
| Riboflavin (mg) | target | 350 | 0.21 | 1.08 | 0 | 0.03 | 0.1 | 0.2 | 29.47 |
| Niacin (mg) | target | 339 | 1.99 | 5.63 | 0 | 0.2 | 0.8 | 2.1 | 136.5 |
| Vitamin C (mg) | target | 289 | 11.63 | 26.15 | 0 | 0 | 0 | 11 | 230 |

The general formula for creating second-degree polynomial features is:

$$P_2(x_1, x_2, \dots, x_{\mathrm{n}}) = [x_1, x_2, \dots, x_{\mathrm{n}}, {x_1}^2, {x_2}^2, \dots, {x_{\mathrm{n}}}^2, x_1x_2, x_1x_3, \dots, x_1x_{\mathrm{n}}, x_2x_3, \dots, x_2x_{\mathrm{n}}, \dots, x_{\mathrm{n}-1}x_{\mathrm{n}}] \quad (1)$$

In this study, we apply second-degree polynomial features to our dataset of nutritional information. With our six input features (energy, protein, fat, carbohydrate, fiber, and sodium), we generate the following new features:

- Squared terms: energy², protein², fat², carbohydrate², fiber², sodium².
- Interaction terms: energy×protein, energy×fat, energy×carbohydrate, energy×fiber, energy×sodium, protein×fat, protein×carbohydrate, protein×fiber, protein×sodium, fat×carbohydrate, fat×fiber, fat×sodium, carbohydrate×fiber, carbohydrate×sodium, fiber×sodium.

These polynomial features expand our original set of 6 features with an additional 21 features (6 squared+15 interaction terms). This expansion allows our models to capture more complex relationships in the data. For instance, it might reveal that the effect of protein on a certain micronutrient depends on the level of fat, which would be captured by the protein×fat interaction term.

### 2.3. Ratio features

Ratio features are calculated to capture the proportional relationships between different variables [21]. These features help in identifying and emphasizing the relative importance of one variable with respect to another, providing deeper insights into the data. In this study, ratio features were engineered using the six input variables: energy, protein, fat, carbohydrate, fiber, and sodium. Ratio features used in this study are shown in Table 2. These features aim to enhance the model's ability to predict the target nutrients by providing additional context about the relationships between these nutrients. With the addition of ratio features, our dataset was expanded into a total of 39 features (6 original features+21 polynomial features+12 ratio features). Note that in this study, all new features are calculated using the standardized values of the original features.

### 2.4. Minimum redundancy maximum relevance feature selection

The mRMR algorithm was employed for feature selection in this study. mRMR is an efficient and effective method that aims to select a subset of features that are highly relevant to the target variable while minimizing redundancy among the selected features [23]. The mRMR algorithm operates on two key principles: maximizing the mutual information between selected features and the target variable (maximum relevance), and minimizing the mutual information between pairs of selected features (minimum redundancy). This approach helps identify features that are both informative and non-redundant [24].

The selection process begins by calculating the mutual information between each feature and the target variable. The feature with the highest mutual information is selected first. Subsequent features are chosen based on a trade-off between their relevance to the target and their redundancy with already selected features. This trade-off is quantified using the mutual information quotient (MIQ) criterion:

$$\max \left[ \frac{I(x_i;y)}{\frac{1}{S}\sum_{x_j \in S} I(x_i;x_j)} \right] \quad (2)$$

where *I(*$x_i$*; y)* represents the mutual information between feature $x_i$ and the target *y*, and *I(*$x_i$*;* $x_j$*)* is the mutual information between features $x_i$ and $x_j$. *S* is the set of already selected features.

By selecting features that provide unique and relevant information, mRMR helps reduce dimensionality, mitigate overfitting, and improve model interpretability while maintaining predictive power. The Python package *pyrmr* [25] is used to do the feature selection in this study.

Table 2. Generated ratio features in this study

| Ratio features | Formula |
|---|---|
| Protein to fat ratio | $\frac{Protein}{Fat}$ |
| Carbohydrate to protein ratio | $\frac{Carbohydrate}{Protein}$ |
| Carbohydrate to fat ratio | $\frac{Carbohydrate}{Fat}$ |
| Energy to protein ratio | $\frac{Energy}{Protein}$ |
| Energy to fat ratio | $\frac{Energy}{Fat}$ |
| Energy to carbohydrate ratio | $\frac{Energy}{Carbohydrate}$ |
| Carbohydrate to fiber ratio | $\frac{Carbohydrate}{Fiber}$ |
| Energy to fiber ratio | $\frac{Energy}{Fiber}$ |
| Protein to carb and fat ratio | $\frac{Protein}{Carbohydrate + Fat}$ |
| Protein to total nutrients ratio | $\frac{Protein}{Energy + Fat + Carbohydrate}$ |
| Fat to total nutrients ratio | $\frac{Fat}{Energy + Fat + Carbohydrate}$ |
| Carbohydrate to total nutrients ratio | $\frac{Carbohydrate}{Energy + Fat + Carbohydrate}$ |

### 2.5. Experimental setup

The experimental setup for this study (Figure 1) involved two distinct phases: i) classifier comparison and ii) feature engineering combined with feature selection. The first phase aimed to evaluate the performance of various classifiers using the original features from the dataset, while the second phase focused on enhancing the classifier performance through the introduction of new features and feature selection techniques. Initially, the data was prepared individually for each target variable, as a separate classification model was developed for each nutrient (i.e., iron, calcium, and phosphorus). The data preprocessing involved handling missing values using median imputation for input variables and removing instances with missing values in the target variable. Additionally, data standardization was performed using z-score normalization [26] to ensure that all features contributed equally to the classification process. Each target variable, originally numerical, was transformed into a binary classification problem by applying a threshold based on the median value, dividing the data into "high" and "low" categories.

In the first phase, twelve different classifiers were trained and evaluated to determine the most effective model for predicting unreported nutrients. The classifiers included logistic regression, decision tree, random forest, gradient boosting, linear support vector machines (SVM), radial basis function SVM (RBF SVM), K-nearest neighbors, Naive Bayes, AdaBoost, extra trees, XGBoost, and LightGBM. Hyperparameter tuning was performed for each classifier using a grid search on a subset of the data. This ensured that the optimal settings for each model were identified before evaluating their performance on the entire dataset. The classifiers' performance was assessed using accuracy as the main metric, with the process repeated five times using different seeds for train-test splits. The average accuracy from these iterations was reported as the final performance measure for each target-classifier pair. The random forest classifier emerged as the best performer with the highest average accuracy across all target nutrients.

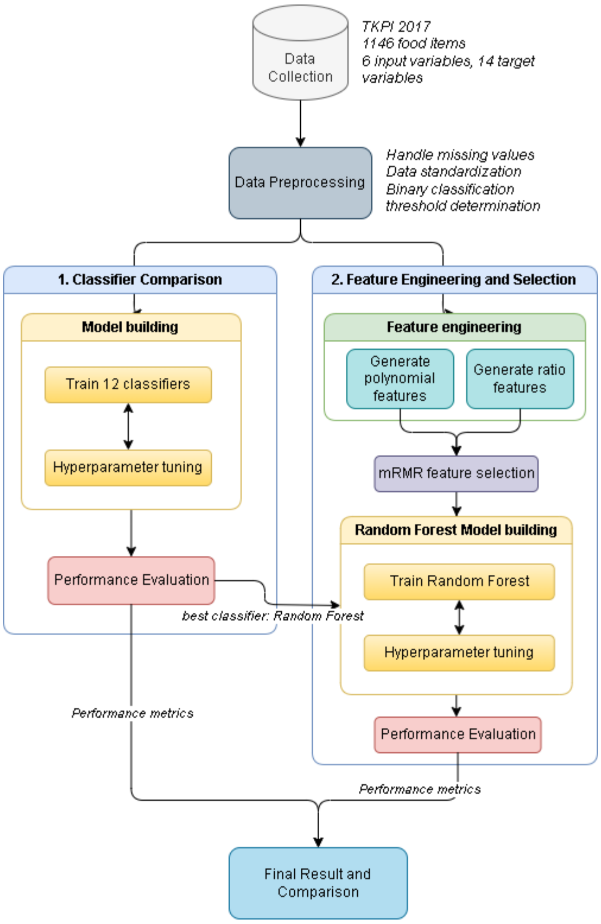


Figure 1. Flowchart of the research method

The second phase of the experiments involved feature engineering and feature selection, focusing solely on the random forest classifier identified as the best model in the first phase. Second degree polynomial features were created by squaring each original feature, while multiplying pairs of original features. Additionally, various ratio features were derived, such as the protein-to-fat ratio, carbohydrate-to-protein ratio, and energy-to-fiber ratio (Table 2). These new features aimed to capture more complex relationships between the original input variables, potentially improving the model's ability to predict the target nutrients.

Given the high dimensionality resulting from the feature engineering process, not all newly generated features were beneficial for the classification tasks. Therefore, a feature selection step was necessary to identify the most informative features while discarding redundant or irrelevant ones. The mRMR technique was employed for this purpose. mRMR is computationally efficient, making it suitable for handling the increased number of features generated during the feature engineering process.

After selecting the most informative features using mRMR, the random forest classifier was retrained and subjected to hyperparameter tuning, similar to the first phase. This ensured that the classifier was optimized for the new feature set. The performance of the classifier with the engineered and selected features was then evaluated using the same methodology as in the first phase, with accuracy averaged over five train-test splits with different seeds.

## 3. RESULTS AND DISCUSSION

In this section, we present the outcomes of our experiments, focusing on the performance of various classifiers and the impact of feature engineering and selection techniques. As mentioned in the previous section, we conducted in two phases of experiments: i) the first phase involved evaluating twelve different machine learning classifiers using the original features and ii) the second phase focused on improving the best-performing classifier, random forest, by incorporating polynomial and ratio features followed by mRMR feature selection.

### 3.1. Performance of 12 classifiers

The results of our comparison across 12 different classifiers reveal significant insights into their performance for predicting micronutrient content in foods. Notably, the random forest algorithm emerged as the top performer, achieving the highest average accuracy of 0.7420 across all micronutrients (Figure 2). This performance was closely followed by other ensemble methods, with extra trees and LightGBM securing the second and third positions with average accuracies of 0.7393 and 0.7372, respectively. In contrast, naive Bayes demonstrated the lowest average accuracy of 0.6402, significantly underperforming compared to the ensemble methods. This significant difference in performance, with the best classifier outperforming the worst by approximately 10 percentage points, underscores the importance of algorithm selection in this domain.

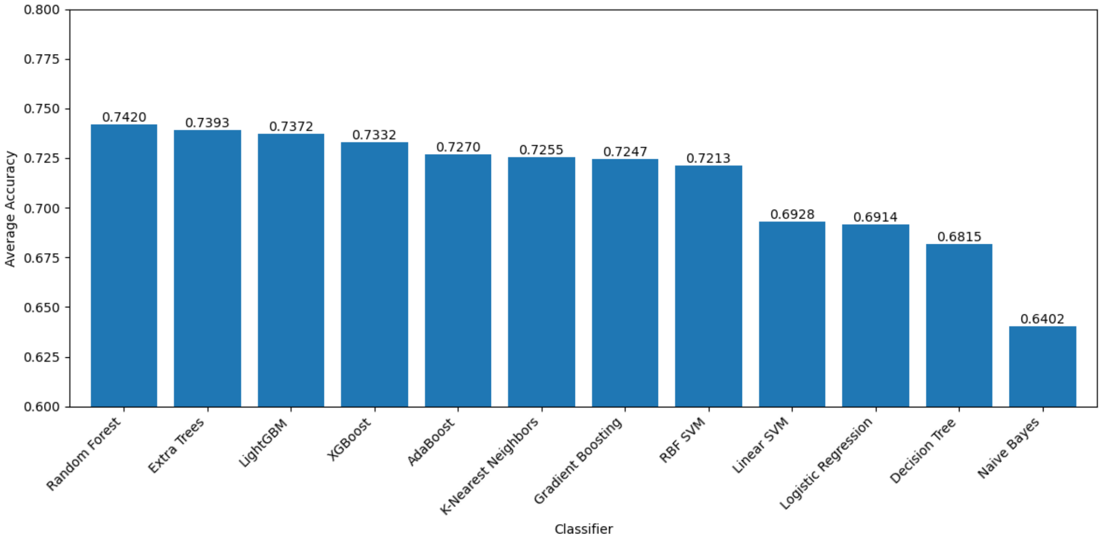
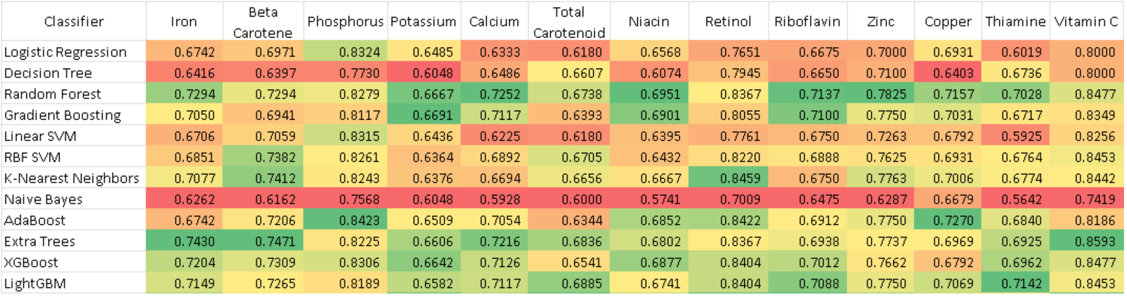


Figure 2. Average accuracy by classifier

Ensemble methods, namely random forest, extra trees, LightGBM, XGBoost, AdaBoost, and gradient boosting, generally perform better in this study than single classifiers. Extra trees, random forest, and LightGBM perform consistently across most micronutrients (Figure 3). This suggests that their ability to combine multiple weak learners into a strong predictor is well-suited to the complex relationships inherent in nutritional data. The consistent performance further reinforces the robustness of these ensemble approaches for this specific prediction task.



| Classifier | Iron | Beta Carotene | Phosphorus | Potassium | Calcium | Total Carotenoid | Niacin | Retinol | Riboflavin | Zinc | Copper | Thiamine | Vitamin C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.6742 | 0.6971 | 0.8324 | 0.6485 | 0.6333 | 0.6180 | 0.6568 | 0.7651 | 0.6675 | 0.7000 | 0.6931 | 0.6019 | 0.8000 |
| Decision Tree | 0.6416 | 0.6397 | 0.7730 | 0.6048 | 0.6486 | 0.6607 | 0.6074 | 0.7945 | 0.6650 | 0.7100 | 0.6403 | 0.6736 | 0.8000 |
| Random Forest | 0.7294 | 0.7294 | 0.8279 | 0.6667 | 0.7252 | 0.6738 | 0.6951 | 0.8367 | 0.7137 | 0.7825 | 0.7157 | 0.7028 | 0.8477 |
| Gradient Boosting | 0.7050 | 0.6941 | 0.8117 | 0.6691 | 0.7117 | 0.6393 | 0.6901 | 0.8055 | 0.7100 | 0.7750 | 0.7031 | 0.6717 | 0.8349 |
| Linear SVM | 0.6706 | 0.7059 | 0.8315 | 0.6436 | 0.6225 | 0.6180 | 0.6395 | 0.7761 | 0.6750 | 0.7263 | 0.6792 | 0.5925 | 0.8256 |
| RBF SVM | 0.6851 | 0.7382 | 0.8261 | 0.6364 | 0.6892 | 0.6705 | 0.6432 | 0.8220 | 0.6888 | 0.7625 | 0.6931 | 0.6764 | 0.8453 |
| K-Nearest Neighbors | 0.7077 | 0.7412 | 0.8243 | 0.6376 | 0.6694 | 0.6656 | 0.6667 | 0.8459 | 0.6750 | 0.7763 | 0.7006 | 0.6774 | 0.8442 |
| Naive Bayes | 0.6262 | 0.6162 | 0.7568 | 0.6048 | 0.5928 | 0.6000 | 0.5741 | 0.7009 | 0.6475 | 0.6287 | 0.6679 | 0.5642 | 0.7419 |
| AdaBoost | 0.6742 | 0.7206 | 0.8423 | 0.6509 | 0.7054 | 0.6344 | 0.6852 | 0.8422 | 0.6912 | 0.7750 | 0.7270 | 0.6840 | 0.8186 |
| Extra Trees | 0.7430 | 0.7471 | 0.8225 | 0.6606 | 0.7216 | 0.6836 | 0.6802 | 0.8367 | 0.6938 | 0.7737 | 0.6969 | 0.6925 | 0.8593 |
| XGBoost | 0.7204 | 0.7309 | 0.8306 | 0.6642 | 0.7126 | 0.6541 | 0.6877 | 0.8404 | 0.7012 | 0.7662 | 0.6792 | 0.6962 | 0.8477 |
| LightGBM | 0.7149 | 0.7265 | 0.8189 | 0.6582 | 0.7117 | 0.6885 | 0.6741 | 0.8404 | 0.7088 | 0.7750 | 0.7069 | 0.7142 | 0.8453 |

Figure 3. Accuracy of individual nutrients from different classifiers

Our results also highlight considerable variation in prediction accuracy across different micronutrients. For instance, vitamin C and phosphorus generally exhibited higher prediction accuracies across all classifiers, while Thiamin and Niacin proved more challenging to predict accurately. This variation suggests that certain micronutrients may have stronger or more linear relationships with the input features, making them easier to predict, while others may have more complex or weaker relationships, leading to greater challenges for prediction.

### 3.2. Performance of feature engineering and minimum redundancy maximum relevance selection

The second phase of our experiment is focusing on enhancing the random forest classifier through feature engineering and mRMR selection. The combined approach led to a notable increase in average accuracy from 0.7421 to 0.7591, representing an improvement of 1.7 percentage points (Figure 4). This enhancement is notable given that feature engineering alone, without mRMR selection, actually resulted in a slight decrease in performance to 0.7372.

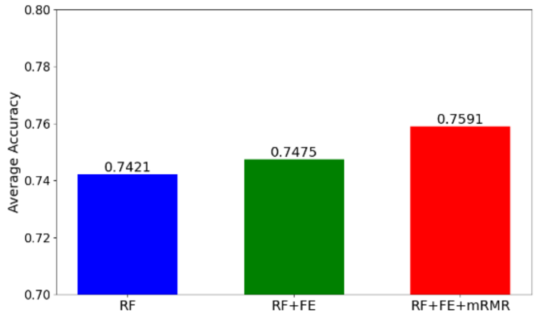


Figure 4. Comparison of average accuracy with and without FE+mRMR

Upon applying feature engineering, modest improvements were observed in the prediction accuracy of several micronutrients, including beta carotene, potassium, calcium, and zinc (Figure 5). However, the impact was not uniformly positive, with slight decreases noted for iron, phosphorus, and vitamin C. This suggests that feature engineering alone may not consistently enhance predictive performance across all micronutrients.

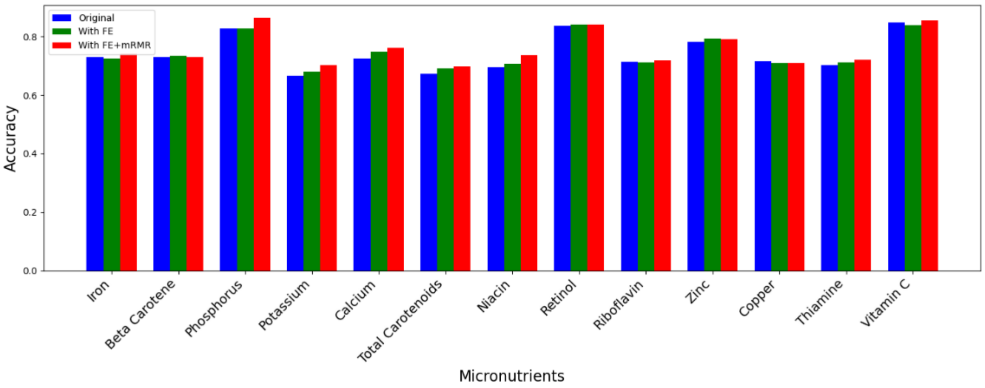


Figure 5. Comparison of average accuracy with and without FE+mRMR across micronutrients

The subsequent application of mRMR feature selection in conjunction with feature engineering yielded more pronounced improvements. Notably, phosphorus, potassium, and total carotenoids exhibited substantial gains in prediction accuracy. This indicates that the mRMR algorithm effectively identified the most relevant features from the expanded feature set, mitigating potential noise introduced by feature engineering.

Interestingly, some micronutrients, such as beta carotene and copper, showed no change or slight decreases in accuracy following the mRMR process. This underscores the complex nature of nutrient interactions and the challenges in developing a universally effective prediction model. Vitamin C consistently maintained the highest prediction accuracy across all methods, suggesting a robust relationship between this nutrient and the input features.

### 3.3. Analysis of selected features

The total number of features available for selection was 39, yet most nutrients required fewer than this maximum for optimal prediction (Table 3). This indicates efficient feature selection tailored to each nutrient's needs. Polynomial features, with up to 20 selected for nutrients like retinol, significantly contribute to capturing complex, nonlinear relationships.

Table 3. Types of features selected based on nutrient target

| Nutrient target | Features selected | | | |
|---|---|---|---|---|
| | Total | Original features | Polynomial features | Ratio features |
| Calcium | 22 | 2 | 11 | 9 |
| Phosphorus | 7 | 0 | 3 | 4 |
| Iron | 27 | 3 | 15 | 9 |
| Potanssium | 9 | 0 | 5 | 4 |
| Copper | 33 | 6 | 19 | 8 |
| Zinc | 22 | 3 | 11 | 8 |
| Retinol | 36 | 5 | 20 | 11 |
| Beta Carotene | 14 | 2 | 5 | 7 |
| Total Carotenoids | 26 | 5 | 12 | 9 |
| Thiamine | 8 | 1 | 3 | 4 |
| Riboflavin | 11 | 1 | 5 | 5 |
| Niacin | 31 | 4 | 17 | 10 |
| Vitamin C | 10 | 1 | 5 | 4 |

Ratio and polynomial features, which capture non-linear relationships and interactions, are consistently chosen over original features across most nutrients. Notably, the number of selected features varies considerably across nutrients, indicating differing levels of complexity in nutrient prediction. Some nutrients require more features, suggesting more complex relationships with food components, while others can be predicted with fewer features. For instance, retinol and copper required 36 and 33 features respectively, suggesting intricate relationships within these nutrients that demand a broader array of features to model accurately. On the other hand, nutrients like phosphorus and thiamin, which needed only 7 and 8 features respectively, indicate simpler predictive models.

The analysis of the most and least selected features reveals important insights into the feature selection process and its impact on nutrient prediction models. The top 10 most selected features include a balanced mix of polynomial and ratio features, such as "Energy_to_Carb_Ratio" and "FIBER_Squared," highlighting the importance of capturing non-linear relationships and nutrient proportions (Table 4). Protein-related features, like "Proten_x_Fiber" and "Protein_to_Fat_Ratio," are particularly prominent, underscoring the critical role that protein interactions play in accurate nutrient prediction. This balanced contribution suggests that both polynomial and ratio features significantly enhance the model's ability to predict complex nutrient relationships, improving overall accuracy.

Table 4. Ten most selected features

| Feature name | Feature type | Number of times selected |
|---|---|---|
| Energy_to_Carb_Ratio | ratio | 11 |
| Protein_to_Fat_Ratio | ratio | 11 |
| FIBER_Squared | polynomial | 11 |
| Carb_to_Total_Nutrients_Ratio | ratio | 10 |
| PROTEIN_x_FIBER | polynomial | 10 |
| Carb_to_Protein_Ratio | ratio | 10 |
| Protein_to_Carb_to_Fat_Ratio | ratio | 10 |
| ENERGY | original | 9 |
| SODIUM_Squared | polynomial | 9 |
| PROTEIN_Squared | polynomial | 9 |

On the other hand, the least selected features (Table 5), including original features like "FAT" and "FIBER," indicate that engineered features generally provide more valuable information than original data alone. The underutilization of certain polynomial features, such as "CARB_Squared" and "ENERGY_Squared," suggests that not all quadratic transformations contribute equally to model performance. Additionally, the limited selection of some ratio features, like "Energy_to_Fiber_Ratio," implies that specific nutrient ratios may be less relevant in this dataset. These findings emphasize the importance of context-specific feature selection, where certain interactions are less influential across different nutrient predictions, highlighting the need for tailored approaches to feature engineering in nutritional analysis.

### 3.4. Comparison with related work

Our study shares similarities with Razavi and Xue work [7] in predicting unreported nutrients using machine learning, but offers several unique contributions. While both studies aim to enhance nutritional information availability, our approach is tailored to more restrictive labeling requirements, using only six input features based on Indonesia's mandatory labeling. This makes our model more applicable in contexts with limited nutritional data.

Table 5. Ten least selected features

| Feature name | Feature type | Number of times selected |
|---|---|---|
| FAT_x_CARB | polynomial | 5 |
| ENERGY_x_FAT | polynomial | 4 |
| FAT_Squared | polynomial | 4 |
| FIBER | original | 4 |
| Energy_to_Protein_Ratio | ratio | 3 |
| Energy_to_Fiber_Ratio | ratio | 3 |
| CARB_x_FIBER | polynomial | 3 |
| ENERGY_Squared | polynomial | 2 |
| CARB_Squared | polynomial | 2 |
| FAT | original | 1 |

We evaluated a broader range of classifiers (12 vs. 7), providing a more comprehensive comparison of machine learning approaches for this task. Our best-performing model, random forest, achieved an average accuracy of 0.7591 across all target nutrients after feature engineering and mRMR selection. This performance, while slightly lower than Razavi and Xue [7] reported accuracies (>0.80), is notable given our fewer input features and binary classification approach (Table 6).

Table 6. Comparison with related research

| Aspect | This research | Razavi and Xue [7] |
|---|---|---|
| Dataset | TKPI 2017 (Indonesian) | FNDDS (US) |
| Number of input/independent features | 6 | 14 |
| Number of instances (food items) | 1146 | 5624 |
| Target variables | 13 micronutrients | 15 micronutrients |
| Classification approach | Binary (high/low) | Ternary (low/medium/high) |
| Number of classifiers evaluated | 12 | 7 |
| Best performing classifier | Random forest | GBM and random forest |
| Feature engineering | Polynomial and ratio features | None |
| Feature selection | mRMR | None |
| Average accuracy (best model) | 0.76 | Not mentioned |
| Highest accuracy for single nutrient | 0.86 (phosphorus), 0.85 (vitamin C), 0.84 (retinol) | 0.94 (vitamin B12, phosphorus) |
| Lowest accuracy for single nutrient | 0.70 (potassium) | 0.81 (vitamin E) |

The performance difference can be attributed to several factors. Firstly, our use of fewer input features naturally constrains the available information for prediction. Secondly, our binary classification (high/low) versus their three-class approach (high/medium/low) presents a different challenge, potentially affecting accuracy metrics. Lastly, differences in datasets (TKPI vs FNDDS) and target nutrients may contribute to performance variations.

Despite these challenges, our study demonstrates the feasibility of nutrient prediction even with highly limited input data. The introduction of polynomial and ratio features, combined with mRMR selection, offers a novel approach to improving model performance in this context. This method could be particularly valuable in regions where comprehensive nutritional labeling is not mandatory.

## 4. CONCLUSION

This study demonstrated the potential of using machine learning and feature engineering to predict unreported micronutrients from a limited set of input features. Using the TKPI dataset, we successfully evaluated the performance of twelve classifiers and found that random forest provided the highest accuracy. Additionally, the incorporation of polynomial and ratio features, along with mRMR feature selection, significantly enhanced model performance, increasing the average accuracy from 0.7421 to 0.7595.

Despite these promising results, several limitations must be acknowledged. One primary limitation is the dataset composition, which consists mainly of single food ingredients. This restricts the generalizability of the models to more complex food items or commercial packaged foods, which often contain a mix of ingredients and additives not represented in the TKPI dataset. The absence of commercial packaged foods in the dataset further constrains the applicability of the findings to everyday consumer products, where the nutrient composition might differ significantly. Another limitation is the relatively small size of the dataset, with 1,146 food items. Although the models performed well within this context, larger datasets could provide more robust and reliable results. Moreover, the binary classification approach, while simplifying the problem, might overlook subtle variations in nutrient levels that could be captured in a more granular analysis.

Future work should aim to address these limitations by expanding the dataset to include a more diverse range of food items, particularly commercial packaged foods. This would enhance the model's generalizability and relevance to real-world applications. Additionally, further research could explore the integration of more advanced feature engineering techniques and the inclusion of additional contextual information, such as food categories, or cooking methods. These information could provide more robust models, contributing to better-informed dietary choices for consumers. Overall, this study lays a foundation for future research aimed at improving nutrient prediction models.

## ACKNOWLEDGEMENTS

The authors would like to express gratitude to the Data Science Lab, Department of Computer Science, Lambung Mangkurat University, for the computing support and resources provided for this research.

## REFERENCES

[1] K. Włodarska, K. Pawlak-Lemańska, M. Sielicka-Różyńska, and U. Samotyja, "Food labelling system— consumers' perspective," in *Sustainable food. Production and consumption perspectives*, 1st ed., K. Pawlak-Lemańska, B. Borusiak, and E. Sikorska, Eds., Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, 2024, pp. 132–148, doi: 10.18559/978-83-8211-209-2/9.

[2] C. D. Pfledderer *et al.*, "Using the nutrition facts label to make food choices is associated with healthier eating among 8th and 11th-grade students: an analysis of statewide representative data from the 2019–2020 Texas school physical activity and nutrition survey," *Nutrients*, vol. 16, no. 2, p. 311, Jan. 2024, doi: 10.3390/nu16020311.

[3] M. Egnell *et al.*, "Impact of the nutri-score front-of-pack nutrition label on purchasing intentions of individuals with chronic diseases: results of a randomised trial," *BMJ Open*, vol. 12, no. 8, p. e058139, Aug. 2022, doi: 10.1136/bmjopen-2021-058139.

[4] M. Du *et al.*, "Cost-effectiveness analysis of nutrition facts added-sugar labeling and obesity-associated cancer rates in the US," *JAMA Network Open*, vol. 4, no. 4, Apr. 2021, doi: 10.1001/jamanetworkopen.2021.7501.

[5] A. Outfarouin and N. Laaffat, "Towards a new healthy food decision-making system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 2, pp. 1088-1098, Aug. 2023, doi: 10.11591/ijeecs.v31.i2.pp1088-1098.

[6] "2021 regulation of the National Agency of Drug and Food Control Number 26 on nutritional information on processed food labels (in Indonesian: Peraturan Badan Pengawas Obat Dan Makanan Nomor 26 Tahun 2021 tentang informasi nilai gizi pada label pangan olahan," *Badan Pengawas Obat dan Makanan Republik Indonesia*, 2021.

[7] R. Razavi and G. Xue, "Predicting unreported micronutrients from food labels: machine learning approach," *Journal of Medical Internet*, vol. 25, Apr. 2023, doi: 10.2196/45332.

[8] Q. Thames *et al.*, "Nutrition5k: towards automatic nutritional understanding of generic food," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 8899–8907, doi: 10.1109/CVPR46437.2021.00879.

[9] B. Wang, T. Bu, Z. Hu, L. Yang, Y. Zhao and X. Li, "Coarse-to-fine nutrition prediction," *IEEE Transactions on Multimedia*, vol. 26, pp. 3651-3662, 2024, doi: 10.1109/TMM.2023.3313638.

[10] Y. Han, Q. Cheng, W. Wu, and Z. Huang, "DPF-nutrition: food nutrition estimation via depth prediction and fusion," *Foods*, vol. 12, no. 23, Nov. 2023, doi: 10.3390/foods12234293.

[11] M. Keller, C. A. Tai, Y. Chen, P. Xi, and A. Wong, "NutritionVerse-Direct: exploring deep neural networks for multitask nutrition prediction from food images," *arXiv*, 2024, doi: 10.48550/arXiv.2405.07814.

[12] M. Al-Saffar and W. R. Baiee, "Nutrition information estimation from food photos using machine learning based on multiple datasets," *Bulletin of Electrical Engineering and Informatics.*, vol. 11, no. 5, pp. 2922–2929, Oct. 2022, doi: 10.11591/eei.v11i5.4007.

[13] M. A. Rasyidi, Y. S. Mardhiyyah, Z. Nasution, and C. H. Wijaya, "Performance comparison of state-of-the-art deep learning model architectures in Indonesian food image classification," *Bulletin of Electrical Engineering and Informatics.*, vol. 13, no. 5, pp. 3355–3368, Oct. 2024, doi: 10.11591/eei.v13i5.7996.

[14] K. Neha, P. Sanjan, S. Hariharan, S. Namitha, A. Jyoshna, and A. B. Prasad, "Food prediction based on recipe using machine learning algorithms," in *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India: IEEE, Aug. 2023, pp. 411–416, doi: 10.1109/ICAISS58487.2023.10250758.

[15] G. Ispirova, T. Eftimov, and B. K. Seljak, "Domain heuristic fusion of multi-word embeddings for nutrient value prediction," *Mathematics*, vol. 9, no. 16, Aug. 2021, doi: 10.3390/math9161941.

[16] R. Li, P. Ji, and Q. Kong, "DelicacyNet for nutritional evaluation of recipes," *Frontiers in Nutrition*, vol. 10, Sep. 2023, doi: 10.3389/fnut.2023.1247631.

[17] J. Kalra, D. Batra, N. Diwan, and G. Bagler, "Nutritional profile estimation in cooking recipes," in *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*, Dallas, TX, USA: IEEE, Apr. 2020, pp. 82–87, doi: 10.1109/ICDEW49219.2020.000-3.

[18] D. Tao, P. Yang, and H. Feng, "Utilization of text mining as a big data analysis tool for food science and nutrition," *Comprehensive Reviews In Food Science And Food Safety*, vol. 19, no. 2, pp. 875–894, Mar. 2020, doi: 10.1111/1541-4337.12540.

[19] J. V. N. Ramesh, A. Kushwaha, T. Sharma, A. Aranganathan, A. Gupta, and S. K. Jain, "Intelligent feature engineering and feature selection techniques for machine learning evaluation," *International Conference on Mobile Radio Communications & 5G Networks*, Singapore: Springer Nature Singapore, 2024, pp. 753–764, doi: 10.1007/978-981-97-0700-3_56.

[20] Tim Direktorat Gizi Masyarakat, "Indonesian Food Composition Table *2017* (in Indonesian: Tabel Komposisi Pangan Indonesia 2017*)," Kementerian Kesehatan RI*, 2018.

[21] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," *SoutheastCon 2016*, Norfolk, VA, USA, 2016, pp. 1-6, doi: 10.1109/SECON.2016.7506650.

[22] M. Tetelman, "Continuous learning: engineering super features with feature Algebras," *arXiv*, 2013, doi: 10.48550/arXiv.1312.5398.

[23] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 03, no. 02, pp. 185–205, Apr. 2005, doi: 10.1142/S0219720005001004.

[24] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[25] "pymrmr PyPI." [Online]. Available: https://pypi.org/project/pymrmr/. (Accessed: Jul. 25, 2024).

[26] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, Second edition, Beijing [China]; Sebastopol, CA: O'Reilly Media, Inc, 2019.

## BIOGRAPHIES OF AUTHORS

**Fatma Indriani** is a lecturer in the Department of Computer Science at Lambung Mangkurat University. Her research interest is focused on data science. Before becoming a lecturer, she completed her undergraduate program in the Department of Informatics at Bandung Institute of Technology. In 2008, she started working as a lecturer in the Computer Science department at Lambung Mangkurat University. She then completed her master's degree at Monash University, Australia, in 2012. Her latest education is a doctorate in Bioinformatics at Kanazawa University, Japan, which was completed in 2022. The research fields she focuses on are data science and bioinformatics. She can be contacted at email: f.indriani@ulm.ac.id.

**Irwan Budiman** successfully finished his bachelor's degree in the Department of Informatics at the Islamic University of Indonesia. Subsequently, he assumed the role of a lecturer in Computer Science at Universitas Lambung Mangkurat starting in 2008. Additionally, in 2010, he pursued a master's degree in information systems at Diponegoro University. His area of research expertise lies in data science. He can be contacted at email: irwan.budiman@ulm.ac.id.

**Dwi Kartini** received her bachelor's and master's degrees in computer science from the Faculty of Computer Science, Putra Indonesia University "YPTK" Padang, Indonesia. Her research interests include the applications of artificial intelligence and data mining. She is an Assistant Professor of the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University in Banjarbaru, Indonesia. She can be contacted at email: dwikartini@ulm.ac.id.

**Lilies Handayani** is a lecturer in Department of Statistics, Tadulako University. She completed her bachelor's degree in statistics at Hasanuddin University, Makassar in 2011. Then, she completed her master's degree in statistics at IPB University, Bogor in 2014. After that, she pursued her doctoral's degree in bioinformatics at Kanazawa University from 2022. Her current research interests include gene expression, machine learning, WGCNA, and GOEA. She can be contacted at email: lilies.stath@gmail.com.