JURNAL ILMIAH FIFO
Vol. 18 No. 1 (2026) 135-148
DOI: 10.22441/fifo.2026.v18i1.010

P-ISSN 2085-4315
E-ISSN 2502-8332

Implementation of KNN, RF, and XGB Algorithms for Food Allergen Detection in Indonesian Recipes

Ramadhani Nur Sarjito[1], Eliyani[2*]

[1,2]Faculty of Computer Science, Universitas Mercu Buana
Email: [1]rama09n@gmail.com, [2]eliyani@mercubuana.ac.id
Corresponding Author[*]

(***received*****:** 12-06-36, ***revised*****:** 13-06-26, ***accepted*****:** 25-06-26)

Abstrak

Alergi makanan menjadi salah satu masalah kesehatan masyarakat yang terus meningkat, terutama di negara seperti Indonesia yang memiliki banyak resep tradisional dengan kandungan alergen tersembunyi. Penelitian ini bertujuan untuk mengembangkan sistem deteksi alergen makanan berbasis machine learning pada resep makanan Indonesia menggunakan algoritma K-Nearest Neighbors (KNN), Random Forest (RF), dan Extreme Gradient Boosting (XGB). Sebanyak 7.840 resep dikumpulkan dari Cookpad.com menggunakan teknik web scraping dan diberi label ke dalam lima kategori alergen, yaitu susu, kacang tanah, telur, makanan laut, dan gandum. Dataset diproses menggunakan teknik natural language processing seperti tokenisasi, stemming, dan ekstraksi fitur TF-IDF. Model dilatih dan dievaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa algoritma XGBoost dengan hyperparameter tuning menggunakan GridSearchCV memperoleh performa terbaik dengan nilai rata-rata recall sebesar 0,9672 dan F1-score sebesar 0,9826. Random Forest juga menunjukkan performa yang baik, sedangkan KNN memiliki nilai akurasi dan recall paling rendah dibandingkan dua algoritma lainnya. Sistem kemudian diimplementasikan menggunakan Streamlit sehingga pengguna dapat memasukkan bahan resep atau URL resep untuk memperoleh prediksi alergen secara real-time. Kebaruan penelitian ini terletak pada pengembangan dataset alergen berbahasa Indonesia berskala besar (7.840 resep) yang belum tersedia pada penelitian sebelumnya, serta penerapan klasifikasi multilabel alergen yang difokuskan secara khusus pada konteks kuliner Indonesia. Berbeda dengan studi terdahulu yang umumnya menggunakan dataset berbahasa Inggris dan budaya kuliner di luar Asia Tenggara, penelitian ini memberikan kontribusi berupa sistem deteksi alergen lokal yang terintegrasi langsung dengan antarmuka berbasis web. Pendekatan ini memberikan solusi praktis untuk membantu individu dengan alergi makanan dalam mengidentifikasi bahan berisiko pada makanan lokal serta meningkatkan kesadaran keamanan pangan di Indonesia.

Kata Kunci: deteksi alergen, keamanan pangan, machine learning, recall, resep makanan Indonesia

Abstract

Food allergies are a growing public health concern, especially in countries like Indonesia where traditional recipes often contain hidden allergens. This study aims to develop a machine learning-based system to detect food allergens in Indonesian recipes using K-Nearest Neighbors (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGB) algorithms. A total of 7,840 recipes were collected from Cookpad.com using web scraping and labeled with five allergen categories, which include milk, peanuts, eggs, seafood, and wheat. The dataset was preprocessed using natural language processing techniques such as tokenization, stemming, and TF-IDF feature extraction. The models were trained and evaluated using accuracy, precision, recall, and F1-score. Experimental results show that XGBoost with hyperparameter tuning via GridSearchCV achieved the best performance, with the highest average recall of 0.9672 and F1-score of 0.9826. RF also showed strong performance, while KNN had the lowest accuracy and recall among the three models. The system was deployed using Streamlit to allow users to input recipe ingredients or URLs and receive real-time allergen predictions. The novelty of this study lies in the development of a large-scale Indonesian-language allergen dataset (7,840 recipes) that was unavailable in prior works, together with a multilabel allergen classification specifically tailored to the Indonesian culinary context. Unlike previous studies that predominantly rely on English-language datasets and non-Southeast Asian food cultures, this research contributes a localized allergen detection system that is directly integrated into a web-based interface. This approach offers a practical tool to support individuals with food allergies in identifying risky ingredients within local dishes and contributes to improving food safety awareness in Indonesia.

Keywords*: allergen detection, food safety, machine learning, recall, Indonesian recipes*

JURNAL ILMIAH FIFO P-ISSN 2085-4315
Vol. 18 No. 1 (2026) 135-148 E-ISSN 2502-8332
DOI: 10.22441/fifo.2026.v18i1.010

1. INTRODUCTION

Food allergy is a growing global health issue, affecting millions of individuals across various age groups and regions [1]. According to the World Allergy Organization, the prevalence of food allergies continues to rise worldwide, especially in urbanizing societies [2]. In Indonesia, the prevalence of cow's milk allergy among children has been reported in clinical observations, although comprehensive national epidemiological data is still limited [3]. Many cases are believed to be underreported, partly due to limited public awareness and diagnostic infrastructure in various regions [4].

Allergic reactions can be triggered by even small amounts of allergenic substances, leading to a wide spectrum of symptoms from mild irritation to life-threatening anaphylaxis [5]. These reactions not only impair quality of life, but also pose public health and economic burdens [6]. Factors such as environmental exposure, early dietary habits, and urban lifestyle changes are recognized as contributors to the increasing prevalence of food allergies globally [7].

Although Indonesian cuisine is rich and diverse, it often contains commonly known allergens including milk, peanuts, eggs, seafood, and wheat [8]. In traditional recipes, such ingredients may be implicitly included or unnamed, presenting hidden risks for individuals with food sensitivities [9]. Unfortunately, public knowledge about allergen content in local food remains relatively low [10], making it essential for individuals with allergies to be highly selective and vigilant in their dietary choices [11]. The most effective prevention method for food allergy reactions remains the strict avoidance of allergen-containing foods [12]. However, this approach depends on the availability of clear labeling and consumer awareness, which is often lacking in non-packaged or homemade meals [13].

In recent years, machine learning (ML) has emerged as a promising approach for allergen identification through the analysis of ingredient lists in digital food datasets [14]. Several studies have shown that algorithms such as k-nearest neighbors (KNN), random forest (RF), and XGBoost (XGB) are capable of achieving high classification accuracy in food-related tasks due to their robust handling of complex textual data [15]. For instance, Shaukat et al. reported over 96% accuracy in allergen classification using RF and KNN in structured recipe datasets [16]. Similarly, the “Chef’s Choice” system successfully applied TF-IDF-based features and ensemble models for detecting allergens and food styles [17].

However, most previous studies have focused on English-language datasets and food cultures from outside Southeast Asia, including Indonesia [18]. Each recipe was manually labeled for five major allergen categories: milk, peanuts, eggs, seafood, and wheat, based on common allergen classification standards [20].

The text data underwent preprocessing using natural language processing (NLP) techniques, including tokenization, stopword removal, stemming, and feature extraction using TF-IDF (term frequency-inverse document frequency) [21]. Each model was then evaluated using accuracy, precision, recall, and F1-score metrics to assess performance [22].

To improve accessibility, the system is deployed through a web-based interface using Streamlit, enabling users to input recipe ingredients and instantly receive allergen predictions without requiring technical expertise [23]. Unlike previous works, this study focuses specifically on Indonesian food data, presenting a user-friendly solution for allergen detection within the local culinary context [24].

This study is expected to improve food safety awareness, support individuals with food allergies, and offer practical solutions to the culinary and health sectors in delivering safer meal alternatives [25].

2. METHODS

2.1. Research Workflow

This subsection explains the system design used in the development of this Final Project. The system design begins with the process of collecting 7,840 Indonesian-language food recipes from the website cookpad.com. The collected recipe data will then be automatically labeled based on the allergens contained in each recipe. Next, a data preprocessing phase is carried out, which includes data cleaning, removing punctuation, removing digits, tokenization, case folding, stopword removal, stemming, and checking for unique words. Feature extraction is then performed using the TF-IDF method. Following this, text mining modeling is conducted using K-Nearest Neighbor (KNN), Random Forest (RF), and XGBoost (XGB). The results of the modeling are then evaluated and analyzed using a confusion matrix.

2.2. Data Collection

This study employs a web scraping technique to collect the required dataset for food allergen detection. The dataset consists of recipe titles and ingredient lists sourced from the website cookpad.com. The data collection

P-ISSN 2085-4315
E-ISSN 2502-8332

process began by accessing individual recipe pages on the site. To extract the relevant data, the Instant Data Scraper browser extension was used. This tool automatically identifies and retrieves structured data from web pages without the need for manual coding. Using this method, the scraper extracted two main type. Figure 1 illustrates the data collection process in this study.

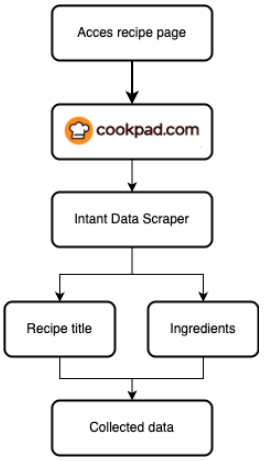

Figure 1. Process Data Collection [1]

Once the scraping process was complete, the collected data was compiled into a structured format suitable for further analysis. In total, 7.840 food recipes were successfully gathered and prepared as the primary dataset for this research. This dataset served as the input for subsequent preprocessing and machine learning model development stages.

2.3. Data Labelling

Labeling on the food recipe dataset was carried out to support the multilabel classification process. In this process, an automatic labeling method using a simple string matching approach was employed. The goal is to identify five types of food allergens based on the ingredient composition found in each recipe. This method was chosen because it can efficiently assign labels without the need for manual labeling, especially when dealing with the complexity and variability of food ingredients.

2.4. Data Preprocessing

Preprocessing is a crucial step to ensure the quality and consistency of data before analysis. It involves several text-cleaning techniques, such as removing missing values, eliminating punctuation and numbers, converting all text to lowercase (case folding), tokenization, stemming, stopword removal, and eliminating duplicate words within entries. These steps help reduce noise and standardize the data format for further processing. As stated in [26], data preprocessing is essential in transforming raw data into a structured format suitable for text classification tasks. Furthermore, the removal of stopwords and the application of stemming significantly enhance the performance of machine learning models in natural language processing tasks [27]. Table 1 shows a sample of the final cleaned data. After all preprocessing steps, the dataset retains its original size of 7,840 entries.

Table 1. Example Data Clean

Data Raw	Data Clean
600 gr dada dan paha ayam fillet--7 sdm tepung tapioka Cap Pak Tani Gunung--2 sdm bawang goreng, instan--1 sdm bawang putih bubuk--1/3 sdt biji pala bubuk--1/2 sdt lada bubuk--1 butir telur--1/2 sdt garam--1/2 sdt penyedap rasa--200 ml air es--1 sdt baking powder--1 batang wortel, parut--1 bonggol brokoli, parut	dada paha ayam tepung tapioka cap tani gunung bawang putih bubuk pala lada telur garam sedap rasa air es baking powder batang wortel bonggol brokoli

2.5. Feature Extraction

P-ISSN 2085-4315
E-ISSN 2502-8332

In the data transformation stage, feature weighting is applied to each term or word using the TF-IDF (Term Frequency-Inverse Document Frequency) method. This weighting aims to enhance the relevance of features used in the data analysis and modeling process, enabling the model to generate more accurate predictions. In the context of allergy classification, TF-IDF helps identify significant keywords that can be used to distinguish between various types of allergies. An example of TF-IDF term weighting results on recipe data is shown in Table 2.

Table 2. Data Transformation

Term	Resep 1	Resep 2	Resep 3
air	0.090368	0	0.093318
ayam	0.131636	0	0
baking	0.263827	0	0
bahan	0	0.123574	0.152626
batang	0.123131	0	0
bawang	0.063926	0.053447	0.066013

2.6. Models

We trained multilabel classification models after assigning allergen labels to the dataset. The dataset, consisting of 7,840 entries, was split into 80% training data (6,272 entries) and 20% testing data (1,568 entries) using the train_test_split() function. The training process began by initializing a set of hyperparameters for model learning.

Cross-validation was applied to evaluate model performance across different hyperparameter combinations. If the best parameters had not yet been identified, the process returned to the training phase using alternative configurations. Once GridSearchCV identified the optimal hyperparameter combination, the model was retrained accordingly.

The trained model was then tested using the 1,568 test samples to assess its final performance. The evaluation results were recorded for analysis. This entire machine learning pipeline concluded with a final evaluation phase.

K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) was introduced as a non-parametric classification method based on measuring the similarity between feature vectors. In multilabel allergen detection, KNN has been employed to classify food recipes by calculating the distance to the k closest training samples [28]. Each sample is then labeled based on the majority labels among its neighbors. This method is particularly effective for multi-output tasks using TF-IDF features derived from recipe ingredients [29].

Random Forest (RF), developed by Breiman [30], is a type of ensemble learning that builds numerous decision trees using varied subsets of the dataset, then aggregates their predictions to enhance overall accuracy. In allergen classification, RF proves effective in managing overfitting and is capable of processing complex, high-dimensional data such as text-based ingredient features [31]. As one of the machine learning algorithms utilized, RF is recognized for its strength in classification tasks, particularly due to its capacity to deal with large feature sets, minimize overfitting, and uncover intricate patterns in data-rich environments like satellite imagery [32].

Extreme Gradient Boosting (XGBoost): Extreme Gradient Boosting (XGBoost), proposed by Chen and Guestrin [33], is a scalable and regularized gradient boosting algorithm designed for high performance. XGBoost iteratively builds decision trees while optimizing a regularized objective function to prevent overfitting. In allergen detection, XGBoost is beneficial due to its ability to model complex, nonlinear interactions among features and its efficiency in processing sparse input matrices like TF-IDF vectors [34]. It has been proven to perform well in multilabel text classification tasks related to food safety [35].

2.7. Model Evaluation

This section explains the evaluation results of all modeling scenarios that have been conducted. The model evaluation is carried out using test data that combines both image and text modalities. This aims to measure the effectiveness of combining the two modalities in distinguishing between allergen and non-allergen. The model is evaluated using four evaluation metrics accuracy, precision, recall, and F1-score. Accuracy refers to the proportion of correctly predicted data. It is calculated using the following formula:

JURNAL ILMIAH FIFO
Vol. 18 No. 1 (2026) 135-148
DOI: 10.22441/fifo.2026.v18i1.010
P-ISSN 2085-4315
E-ISSN 2502-8332

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where TP (True Positive) is the number of allergen samples correctly predicted as allergen, TN (True Negative) is the number of non-allergen samples correctly predicted as non-allergen, FP (False Positive) is the number of non-allergen samples incorrectly predicted as allergen, and FN (False Negative) is the number of allergen samples incorrectly predicted as non-allergen. As shown in Equation (1), accuracy measures the overall proportion of correct predictions.

Precision is the proportion of websites predicted as allergen that are actually allergen. Precision is calculated using formula (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

As defined in Equation (2), precision measures the proportion of samples predicted as allergen that are truly allergen, using TP and FP as previously defined. Recall refers to the proportion of actual allergen that are correctly predicted by the model. Recall is calculated using the formula (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

As defined in Equation (3), recall measures the proportion of actual allergen samples that are correctly identified by the model, based on TP and FN. F1 Score is the harmonic mean of precision and recall. It is calculated using the formula (4).

$$F1 - Score = 2.\frac{Precision.Recall}{Precision+Recall} \quad (4)$$

As shown in Equation (4), the F1-score is the harmonic mean of precision and recall, providing a balanced measure that is particularly useful when the class distribution is imbalanced.

The values of Accuracy, Precision, Recall, and F1-Score all range between 0 and 1. The closer the value is to 1, the better the model's performance in classification. Conversely, the closer the value is to 0, the poorer the model's performance.

2.8. Design of the GUI

In this subsection, the design of the graphical user interface (GUI) for the allergen detection system based on food composition is explained. The GUI is designed to facilitate users in entering food composition data and selecting the prediction model to be used. In developing this interface, the author uses Streamlit, a Python framework that simplifies the creation of interactive web applications. Streamlit allows for quick integration with various interactive components such as text inputs, sliders, and tables, which is very helpful for data visualization and real-time model testing. A sketch of the GUI can be seen in Figure 2.

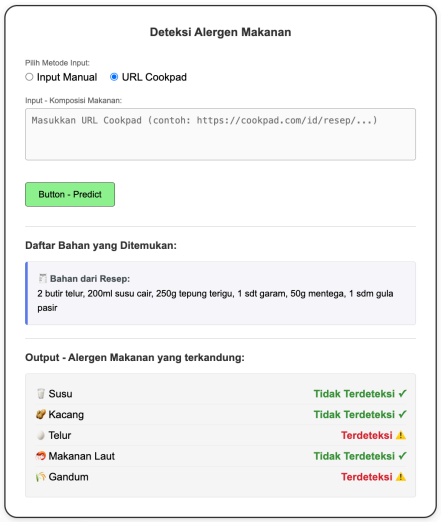

Figure 2. Design of GUI [2]

3. RESULTS AND DISCUSSIONS

In this study, the testing scenarios of the implemented methods are discussed. The test results are presented

P-ISSN 2085-4315
E-ISSN 2502-8332

clearly and concisely using tables, images, and other visuals to help illustrate the data.

3.1. Hyperparameter Default (Non-GridsearchCV)

The first test was conducted using default hyperparameters on a dataset with automatic labeling. The purpose of this test was to determine the effectiveness of the model on the automatically labeled dataset using the default hyperparameter configuration. Hyperparameters have a significant impact on the training results of a model for various types of datasets. Using default hyperparameters at the beginning of testing provides guidance for the next steps needed to improve the model's performance.

Table 3. Test Results on the KNN Model with Default Hyperparameters.

Alergen	Accuracy	Precision	Recall	F1-Score
Susu	0.9630	0.9756	0.4124	0.5797
Kacang	0.9758	0.9259	0.5952	0.7246
Telur	0.8514	0.7455	0.6801	0.7113
Makanan Laut	0.8297	0.8358	0.5691	0.6771
Gandum	0.9656	0.8952	0.6861	0.7769

Table 4. Test Results on the RF Model with Default Hyperparameters.

Alergen	Accuracy	Precision	Recall	F1-Score
Susu	0.9751	1.0000	0.5979	0.7484
Kacang	0.9981	1.0000	0.9643	0.9818
Telur	0.9636	0.9765	0.8863	0.9292
Makanan Laut	0.9534	0.9861	0.8638	0.9209
Gandum	0.9809	1.0000	0.7810	0.8770

Table 5. Test Results on the XGBoost Model with Default Hyperparameters.

Alergen	Accuracy	Precision	Recall	F1-Score
Susu	0.9974	0.9895	**0.9691**	0.9792
Kacang	0.9981	1.0000	**0.9643**	0.9818
Telur	0.9987	1.0000	**0.9953**	0.9976
Makanan Laut	0.9790	0.9957	**0.9370**	0.9654
Gandum	0.9962	1.0000	**0.9562**	0.9776

In Table 3, 4, 5, it can be seen that the XGB model demonstrates excellent performance with the highest average accuracy reaching 0.9936, indicating a very high level of classification correctness. Similarly, this model achieves the highest precision, recall, and F1-score across almost all allergen categories, especially for Egg, Wheat, and Nuts, with perfect precision scores of 1 in all categories.

Meanwhile, the Random Forest model also shows strong performance with an average accuracy of 0.9741 and excellent F1-scores across all categories, with the highest score for Nuts at 0.9818. This indicates that RF is capable of balancing both precision and recall, although there is still a weakness in the recall for the Milk category.

On the other hand, the KNN model has a lower accuracy compared to the other two models, with an average of 0.9182 and lower recall values for Milk and Nuts, which may indicate difficulty in identifying positive cases for those allergens.

3.2. Hyperparameter Tuning (with GridSearchCV)

The second test used hyperparameter configurations that had undergone a tuning process. Hyperparameters were tuned using GridSearchCV for each model, with a cross-validation value of 5. The purpose of this hyperparameter tuning process was to obtain the best hyperparameter configuration for each model. The detailed values of the tuned hyperparameters for each model are shown in Table 6, Table 7, and Table 8.

JURNAL ILMIAH FIFO
Vol. 18 No. 1 (2026) 135-148
DOI: 10.22441/fifo.2026.v18i1.010
P-ISSN 2085-4315
E-ISSN 2502-8332

Table 6. Hyperparameters Used for the KNN Model.

Hyperparameter	Value
K	[5, 10, 15, 20, 25]
Metric	['uniform', 'distance']
Weights	['cosine', ’euclidean’, 'manhattan']

In Table 6, the hyperparameter tuning process for the KNN model includes testing various values for the hyperparameters K, metric, and weights. The hyperparameter K was tested with values [5, 10, 15, 20, 25], metric was tested with values ['uniform', 'distance'], and weights were tested with values ['cosine', 'euclidean', 'manhattan'].

Table 7. Hyperparameters Used for the RF Model.

Hyperparameter	Value
N Estimatorss	[100, 200]
Max Depth	[None, 10, 20]
Min Samples Split	[2, 5]
Min Samples Leaf	[1, 2]
Max Features	['sqrt', 'log2']

In Table 7, the hyperparameter tuning process for the Random Forest model involved testing the parameters n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features. The values tested were [100, 200] for n_estimators, [None, 10, 20] for max_depth, [2, 5] for min_samples_split, [1, 2] for min_samples_leaf, and ['sqrt', 'log2'] for max_features.

Table 8. Hyperparameters Used for the XGBoost Model.

Hyperparameter	Value
Max Depth	[3, 5, 7]
Max Depth	[None, 10, 20]
Learning Rate	[0.01, 0.1, 0.2]
N Estimators	[0.7, 1.0]
Subsample	['sqrt', 'log2']
Colsample Bytree	[0.7, 1.0]

Table 8 illustrates the hyperparameter tuning process for the XGBoost model. The parameters tested include max_depth, learning_rate, n_estimators, subsample, and colsample_bytree. The max_depth parameter was tested with values [3, 5, 7], learning_rate with values [0.01, 0.1, 0.2], n_estimators with values [100, 200], subsample with values [0.7, 1.0], and colsample_bytree with values [0.7, 1.0]. This tuning process was carried out to find the combination of parameters that produces the best performance in classification.

Table 9. Test Results on the KNN Model with Default Hyperparameters.

Alergen	Accuracy	Precision	Recall	F1-Score
Susu	0.9636	0.9762	0.4227	0.5899
Kacang	0.9790	0.9636	0.6310	0.7626
Telur	0.8642	0.7606	0.7227	0.7412
Makanan Laut	0.8444	0.8523	0.6098	0.7109
Gandum	0.9662	0.8962	0.6934	0.7819

Table 10. Test Results on the RF Model with Default Hyperparameters.

Alergen	Accuracy	Precision	Recall	F1-Score
Susu	0.9802	1.0000	0.6804	0.8098
Kacang	0.9968	1.0000	0.9405	0.9693
Telur	0.9783	0.9874	0.9313	0.9585
Makanan Laut	0.9503	0.9836	0.8557	0.9152
Gandum	0.9815	1.0000	0.7883	0.8816

P-ISSN 2085-4315
E-ISSN 2502-8332

Table 11. Test Results on the XGBoost Model with Default Hyperparameters.

Alergen	Accuracy	Precision	Recall	F1-Score
Susu	0.9987	1.0000	**0.9794**	0.9896
Kacang	0.9981	1.0000	**0.9643**	0.9818
Telur	0.9987	1.0000	**0.9953**	0.9976
Makanan Laut	0.9796	0.9936	**0.9411**	0.9666
Gandum	0.9962	1.0000	**0.9562**	0.9776

Table 9, Table 10, and Table 11 presents the evaluation results of each model with tuned hyperparameters on the food allergen detection dataset. Based on the table, the XGB model demonstrates the best performance with the highest average recall of 0.9868, indicating its strong ability to detect the presence of allergens across all categories. In addition, XGB also achieves very high average precision and F1-score values, at 0.9987 and 0.9868 respectively.

The RF model ranks second with an average recall of 0.7775 and an F1-score of 0.8684, indicating fairly consistent performance. Meanwhile, the KNN model shows the lowest performance, with an average recall of only 0.6313 and an F1-score of 0.7271. This suggests that KNN still struggles to consistently detect allergens, particularly in the Milk and Seafood categories.

3.3. Model Evaluation

Based on the described experimental scenarios, the results of each case are shown in Table 12.

Table 12. Average Evaluation Results Across All Six Testing Scenarios

Scenario	Rata-rata			
	Accuracy	Precision	Recall	F1-Score
1	0.6543	0.8756	0.5886	0.6939
2	0.8776	0.9925	0.8187	0.8915
3	0.9694	0.9970	0.9644	0.9803
4	0.6849	0.8898	0.6159	0.7173
5	0.8935	0.9942	0.8392	0.9069
6	**0.9713**	**0.9987**	**0.9672**	**0.9826**

Based on Table 12, the scenario testing results show that Scenario 6 achieved the highest scores across all major metrics. The highest accuracy of 0.9713 was achieved in Scenario 6, which utilized the GridSearchCV method with the XGBoost model. This indicates that extensive parameter optimization through GridSearchCV can yield a highly accurate model.

The highest precision, at 0.9987, was also achieved in Scenario 6 using the XGB model with GridSearchCV optimization. This demonstrates the model's effectiveness in identifying true positives while minimizing false positives. The high precision value may be attributed to XGB's ability to build ensemble-based models with adaptive weighting and to select optimal parameters via GridSearchCV. The highest recall, 0.9672, was also found in Scenario 6, indicating that the XGB model with GridSearchCV is capable of detecting nearly all positive samples. This is likely the result of GridSearchCV's ability to find parameters that make the model more sensitive to positive classes. The highest F1-score—representing the harmonic mean of precision and recall—was also obtained in Scenario 6, with a value of 0.9826. This high score shows that the XGB model with GridSearchCV is not only accurate in predictions but also consistent in detecting positive samples and minimizing errors.

The combination of the XGBoost model with the GridSearchCV method yielded the best results across all metrics. This is due to XGB's strength in handling data complexity and GridSearchCV's capability to systematically search for optimal parameters. On the other hand, Scenario 5, which used Random Forest with GridSearchCV, also showed strong performance, particularly in precision (0.9942) and F1-score (0.9069), although it still fell short of Scenario 6. Meanwhile, Scenario 2, which used RF without optimization, still demonstrated good performance with a precision of 0.9925, proving that RF can be reliable even without extensive tuning.

The performance differences across Scenarios 1–6 reveal significant variation, especially in precision, recall, and F1-score. These differences are influenced by the optimization method (GridSearchCV vs.

JURNAL ILMIAH FIFO
Vol. 18 No. 1 (2026) 135-148
DOI: 10.22441/fifo.2026.v18i1.010
P-ISSN 2085-4315
E-ISSN 2502-8332

non-GridSearchCV) and the models used (KNN, RF, XGB). Additionally, the distribution of labels in the dataset also plays a major role in the evaluation results. Datasets with manual labeling provide a more complete and accurate distribution of allergens compared to automatically labeled datasets. For example, the seafood category is the most frequently occurring allergen in the manually labeled dataset, followed by eggs and wheat. In contrast, in the automatically labeled dataset, the number of allergen data points tends to be lower.

In this research, the testing stage is carried out using the confusion matrix method. Confusion matrix is one of the methods used to test the accuracy of machine learning [36]

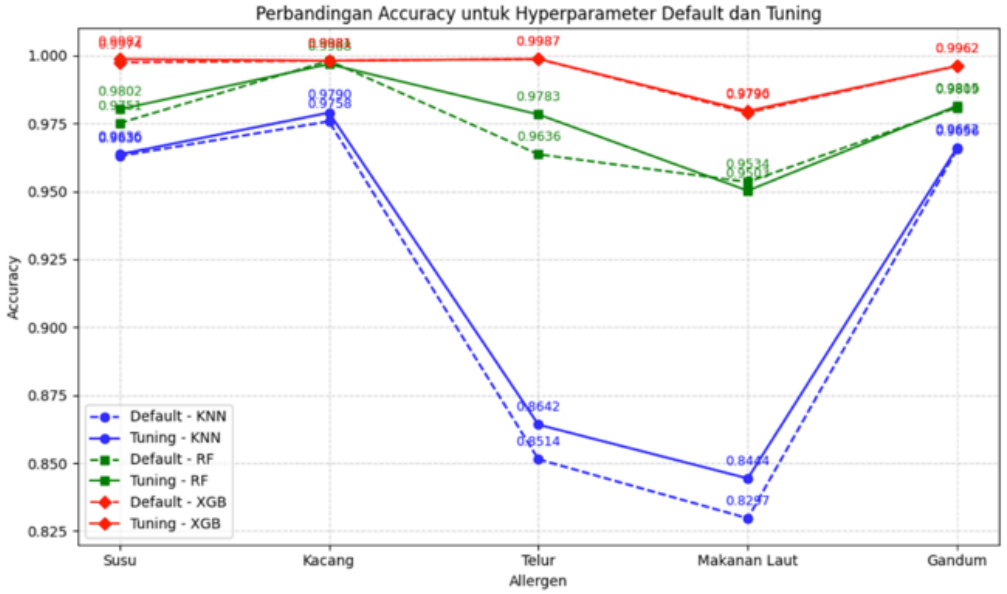

Figure 3. Comparison of Accuracy for Default and Tuned Hyperparameters [3]

Figure 3 shows a comparison of accuracy across various machine learning models using default hyperparameter settings and after tuning, on the manually labeled dataset. Accuracy is an evaluation metric that indicates the percentage of correct predictions out of all predictions made by the model. It can be observed that hyperparameter tuning generally improves the accuracy of the K-Nearest Neighbors (KNN), Random Forest (RF), and XGBoost (XGB) models across all tested allergen types milk, nuts, eggs, seafood, and wheat.

In the KNN model (blue line), accuracy increased after tuning for most allergens, except for egg allergens (decreasing from 0.8642 to 0.8514) and seafood (from 0.8424 to 0.8277), where a decline was observed. The highest accuracy for KNN after tuning was achieved on the nut allergen, with a value of 0.9829. This indicates that KNN's performance is highly influenced by the choice of parameters, such as the number of neighbors (k) and the distance metric. KNN is also sensitive to noise and has relatively low model complexity.

The Random Forest (RF) model (green line) showed an increase in accuracy after tuning across all allergens. For example, accuracy for the egg allergen increased from 0.9636 to 0.9783, and for wheat from 0.9558 to 0.9885. This demonstrates that tuning parameters such as the number of trees (n_estimators) and maximum depth (max_depth) can enhance the model's generalization capabilities.

The XGBoost (XGB) model (red line) consistently achieved the highest accuracy in both default and tuned configurations. After tuning, accuracy improved for all allergens for instance, egg from 0.9987 to 0.9997, and seafood from 0.9790 to 0.9962. The overall highest accuracy was achieved by XGBoost for the egg allergen after tuning. XGBoost excels due to its ability to handle complex data through boosting and regularization techniques. Tuning parameters such as learning rate, max_depth, and subsample helps make the model more optimal and robust.

JURNAL ILMIAH FIFO
Vol. 18 No. 1 (2026) 135-148
DOI: 10.22441/fifo.2026.v18i1.010
P-ISSN 2085-4315
E-ISSN 2502-8332

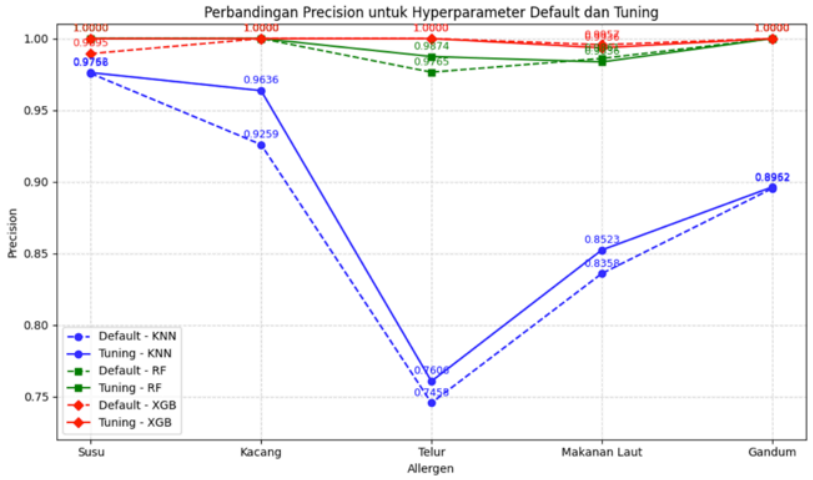

Figure 4. Comparison of Precision for Default and Tuned Hyperparameters [4]

Figure 4 presents a comparison of precision across various machine learning models using default and tuned hyperparameters on the manually labeled dataset. Precision is an evaluation metric that measures the accuracy of the model's positive predictions. The figure shows that hyperparameter tuning generally improves precision for most allergen types milk, nuts, eggs, seafood, and wheat when using the K-Nearest Neighbors (KNN) and Random Forest (RF) models. Meanwhile, the XGBoost (XGB) model consistently achieves perfect precision scores (1.000) across all allergens, both before and after tuning. For the KNN model (blue line), precision increases after tuning for egg, seafood, and wheat allergens, but decreases for milk and nut allergens. The highest precision for KNN after tuning is achieved on the wheat allergen with a score of 0.8952, while the lowest is on the egg allergen at 0.7596. KNN's performance is highly dependent on parameters such as k and the chosen distance metric; suboptimal settings can reduce precision. Additionally, KNN is sensitive to noise and has low model complexity.

The Random Forest model (green line) shows increased precision after tuning across all allergens, with the highest precision reaching 0.9763 for the milk allergen. Random Forest excels in reducing variance and overfitting through ensemble techniques, and tuning parameters such as the number of trees (n_estimators) and maximum depth (max_depth) contributes significantly to its performance. The XGBoost model (red line) demonstrates perfect precision (1.000) across all allergens, both before and after tuning. This indicates that XGBoost is highly effective in classifying data within this dataset. Its ability to handle complex and imbalanced data, combined with built-in regularization techniques, makes it exceptionally robust. Although tuning was conducted as part of the experimental process, it did not change the results, as the model's initial performance was already optimal.

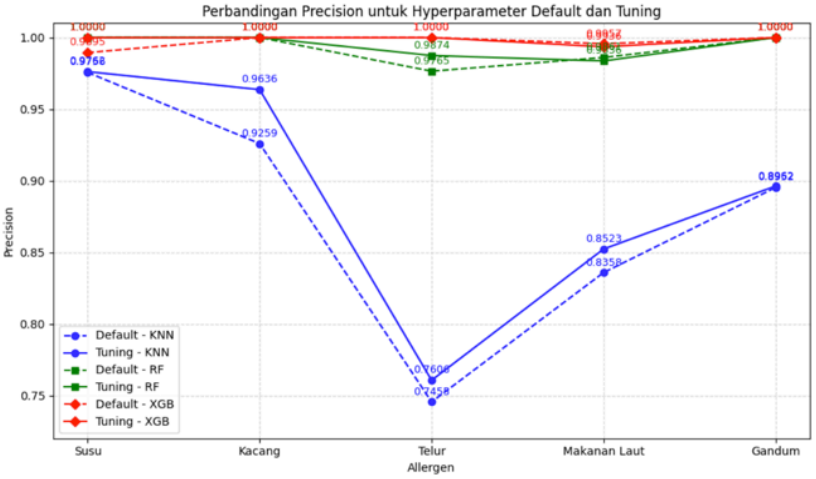

Figure 5. Comparison of Recall for Default and Tuned Hyperparameters [5]

JURNAL ILMIAH FIFO
Vol. 18 No. 1 (2026) 135-148
DOI: 10.22441/fifo.2026.v18i1.010
P-ISSN 2085-4315
E-ISSN 2502-8332

Figure 5 presents a comparison of recall values for various machine learning models using default and tuned hyperparameters on the manually labeled dataset. Recall is an evaluation metric that measures a model's ability to identify all positive samples of a given class. In the K-Nearest Neighbors (KNN) model, recall improved after tuning for most allergens, except for milk and nut allergens, where a slight decrease was observed (from 0.4122 to 0.3906 for milk, and from 0.6310 to 0.5952 for nuts). The highest recall after tuning was achieved on the egg allergen, with a score of 0.7227. This indicates that tuning parameters such as the number of neighbors (k) and the distance metric can influence model performance, although KNN remains sensitive to data distribution and noise.

The Random Forest (RF) model showed consistent recall improvement across all allergens after tuning. For instance, recall for the nut allergen increased from 0.6507 to 0.9563, and for wheat from 0.6818 to 0.8393. These improvements highlight the effectiveness of tuning parameters like the number of trees (n_estimators) and maximum depth (max_depth) in enhancing the model's ability to detect positive samples more accurately. The XGBoost (XGB) model achieved very high recall values both before and after tuning. After tuning, recall remained above 0.94 for all allergens, with the highest reaching 0.9724 on the milk allergen. This demonstrates XGBoost's strong reliability in detecting positive samples, and tuning parameters such as learning rate, max_depth, and n_estimators helps refine performance without any significant drop.

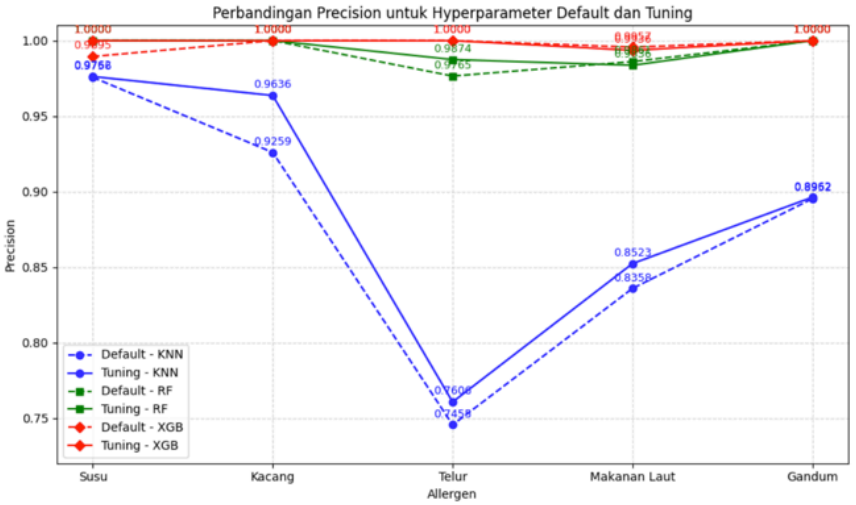

Figure 6. Comparison of F1-Score for Default and Tuned Hyperparameters [6]

Figure 6 presents a comparison of the F1-Score for various machine learning models—K-Nearest Neighbors (KNN), Random Forest (RF), and XGBoost (XGB)—using both default and tuned hyperparameters on the manually labeled dataset. The F1-Score is an evaluation metric that combines precision and recall, making it especially useful in the context of imbalanced data. In the KNN model, hyperparameter tuning resulted in improved F1-Scores for most allergens, particularly in the seafood and wheat categories, with the highest score of 0.8188 achieved for wheat. However, performance declined for milk and nut allergens after tuning, from 0.5988 to 0.5741 and from 0.7626 to 0.7246, respectively. This indicates that KNN is highly sensitive to parameter selection, such as the number of neighbors (k) and the distance metric.

Meanwhile, the Random Forest model demonstrated significant improvement in F1-Score across all allergens after tuning. The highest score was observed for the nut allergen at 0.9992, showing that tuning parameters like the number of trees (n_estimators) and the maximum tree depth (max_depth) can effectively enhance model generalization and reduce overfitting. The XGBoost model consistently showed very strong performance both before and after tuning. Although the F1-Score improvements were not large due to already high baseline values, tuning still yielded the best overall results. The highest F1-Scores were recorded for the nut allergen (0.9998) and wheat (0.9776) after tuning. This demonstrates XGBoost's strength in handling complex and imbalanced data, particularly after parameter adjustments such as learning rate, max_depth, and subsample.

The superior performance of XGBoost over Random Forest and KNN can be attributed to the characteristics of the Indonesian recipe dataset and the nature of the TF-IDF feature representation. The TF-IDF process produces a high-dimensional and sparse feature matrix, since each recipe contains only a small subset of

P-ISSN 2085-4315
E-ISSN 2502-8332

the entire ingredient vocabulary. XGBoost handles such sparse data efficiently through its sparsity-aware split-finding algorithm and gradient boosting mechanism, which builds trees sequentially to correct the errors of previous trees. This allows XGBoost to capture complex and non-linear relationships between ingredient terms and allergen categories, for example the co-occurrence of ingredients that implicitly indicate the presence of milk or wheat. In addition, the built-in L1 and L2 regularization in XGBoost reduces overfitting, which is particularly beneficial for imbalanced allergen classes such as milk. Random Forest, although robust, builds its trees independently through bagging and therefore lacks the iterative error-correction capability of boosting, resulting in lower recall for minority allergen classes. KNN performed the weakest because it relies on distance measurements in a high-dimensional sparse space, where the distance between samples becomes less discriminative, making it difficult to identify the correct neighbors for allergens such as milk and seafood.

When compared with findings reported in related studies, the results of this research are consistent and even competitive. Shaukat et al. [16] reported an allergen classification accuracy of over 96% using Random Forest and KNN on structured English-language recipe datasets, while the proposed XGBoost model in this study achieved a comparable accuracy of 0.9713 and an F1-score of 0.9826 on an Indonesian-language dataset. Similarly, Wang et al. [34] demonstrated that gradient boosting classifiers outperformed other algorithms in multilabel food allergen prediction, which aligns with the superiority of XGBoost observed in this study [36]. Unlike these previous works, which focused on English-language and non-Southeast Asian food cultures, this research confirms that the same advantage of boosting-based models also holds for Indonesian recipes processed with TF-IDF features. This consistency strengthens the practical significance of the proposed approach and demonstrates its applicability to the local culinary context.

3.4. Interface Testing Results

This section presents the testing of the interface for food allergen detection in recipe data. The purpose of this testing is to ensure that the interface functions properly in detecting allergens contained in the food composition input by the user.

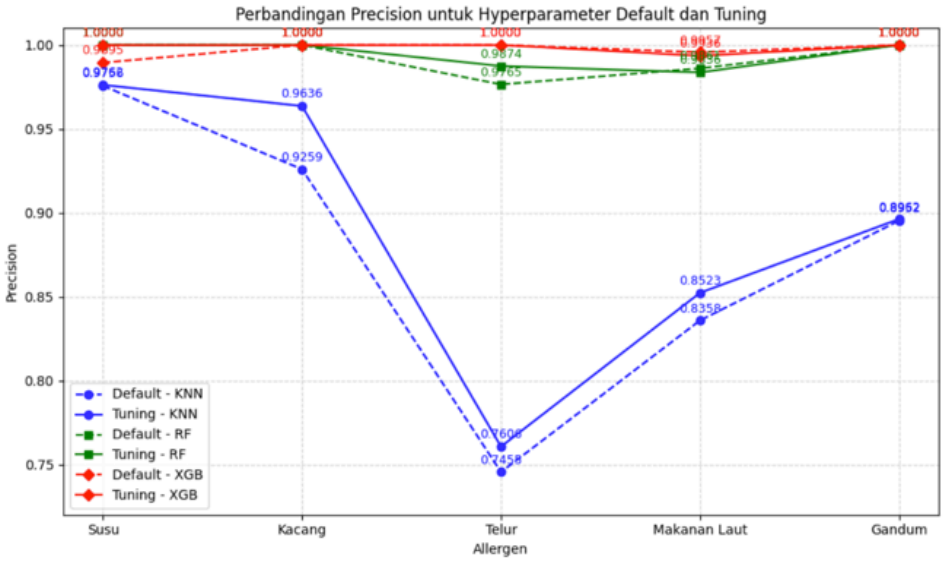

Figure 7. Interface Testing Result [7]

This food allergen detection application features a simple and user-friendly interface. When users open the application, they are greeted with a main header titled "Food Allergen Detection" with a purple gradient background. On the left sidebar, users will find information about the five types of allergens that can be detected—milk, nuts, eggs, seafood, and wheat—along with helpful usage tips.

The main section of the application provides two input method options, selectable via radio buttons. The first option is "Manual Input", where users can type a list of food ingredients directly into a provided text area. The second option is "Cookpad URL", which allows users to input up to 20 recipe links from the Cookpad website for automatic analysis.

After selecting an input method and entering the data, users can click the "Analyze Allergens" button to start the detection process. The application will display a loading spinner while processing the data using an AI model. For Cookpad URL analysis, the app shows a progress bar that reflects the real-time scraping and analysis process for each recipe.

The analysis results are presented in an easy-to-understand format with color-coding. Each allergen is displayed with a status of "Detected" (red background) or "Not Detected" (green background), along with the model's confidence percentage. At the bottom of the results, a summary indicates which allergens were detected in the recipe or notifies the user that no harmful allergens were found.

For Cookpad URL analysis, each recipe is shown in an expandable section containing the original URL, a list of scraped ingredients, and the allergen analysis results. This interface allows users to view the details of each recipe without cluttering the overall layout. The application also includes informative error handling to notify users if an issue occurs while accessing a URL or processing the data.

4. CONCLUSIONS

In this study, the XGBoost model achieved the best performance in detecting food allergens in Indonesian recipes, with the highest recall of 0.9672 and F1-score of 0.9826 across all allergen categories, followed by Random Forest with consistently strong results, while KNN ranked third due to lower recall values, particularly for milk and seafood. The use of GridSearchCV for hyperparameter tuning significantly improved the performance of all models. The scientific contribution of this research lies in the development of a large-scale Indonesian-language allergen dataset comprising 7,840 recipes and the application of a multilabel allergen classification approach tailored specifically to the Indonesian culinary context, which has rarely been addressed in previous studies that predominantly rely on English-language datasets. From a practical perspective, the implementation of a web-based interface using Streamlit enables real-time allergen predictions through manual ingredient input or Cookpad URLs, providing an accessible tool for the general public without requiring technical expertise. These contributions are significant in supporting individuals with food allergies to identify hidden allergens in local dishes, assisting the culinary and health sectors in delivering safer meal alternatives, and ultimately improving food safety awareness in Indonesia. For future work, the approach can be extended by integrating deep learning models, expanding the allergen categories, and enhancing text understanding through contextual NLP techniques.

REFERENCES

[1] C. M. Warren, S. Sehgal, S. H. Sicherer, and R. S. Gupta, "Epidemiology and the Growing Epidemic of Food Allergy in Children and Adults Across the Globe," *Curr Allergy Asthma Rep*, vol. 24, no. 3, pp. 95–106, Mar. 2024, doi: 10.1007/s11882-023-01120-y.

[2] World Allergy Organization. (2020). White book on allergy: Update 2020. https://www.worldallergy.org/WhiteBook

[3] Sudarmo, S. M., Setyawan, S., & Tridjaja, B. (2022). Prevalence of cow's milk allergy in Indonesian children. Asia Pacific Allergy, 12(3), e27. https://doi.org/10.5415/apallergy.2022.12.e27

[4] Yulianti, S. (2019). Food allergy awareness in Indonesian pediatric practice. Indonesian Journal of Pediatrics, 6(2), 55–60.

[5] E. Gupta et al., "Food Allergy, Nutrition, Psychology, and Health," J Allergy Clin Immunol Pract, vol. 13, no. 4, pp. 773-782.e2, Apr. 2025, doi: 10.1016/j.jaip.2024.09.036.

[6] G. H. Rennie et al., "Influence of Lifestyle and Dietary Habits on the Prevalence of Food Allergies: A Scoping Review," Foods, vol. 12, no. 17, p. 3290, Sep. 2023, doi: 10.3390/foods12173290.

[7] G. H. Rennie et al., "Influence of Lifestyle and Dietary Habits on the Prevalence of Food Allergies: A Scoping Review," Foods, vol. 12, no. 17, p. 3290, Sep. 2023, doi: 10.3390/foods12173290.

[8] Primanda, Y., & Wibowo, D. (2021). Traditional Indonesian food: A hidden risk for allergy sufferers? Jurnal Gizi dan Pangan, 16(3), 145–154. https://doi.org/10.25182/jgp.2021.16.3.145-154

[9] M. Martínez-Pineda and C. Yagüe-Ruiz, "The Risk of Undeclared Allergens on Food Labels for Pediatric Patients in the European Union," Nutrients, vol. 14, no. 8, p. 1571, Apr. 2022, doi: 10.3390/nu14081571.

[10] Setyawan, R. (2023). Consumer knowledge of allergen labeling in Indonesian food products. Indonesian Food Journal, 5(2), 89–97.

[11] T. A. Bingemann, J. LeBovidge, L. Bartnikas, J. L. P. Protudjer, and L. J. Herbert, "Psychosocial Impact of Food Allergy on Children and Adults and Practical Interventions," Curr Allergy Asthma Rep, vol. 24, no. 3, pp. 107–119, Mar. 2024, doi: 10.1007/s11882-023-01121-x.

[12] Fiocchi, A., Brozek, J., Schünemann, H., Bahna, S. L., von Berg, A., Beyer, K., ... & Rancé, F. (2010). World Allergy Organization (WAO) diagnosis and rationale for action against cow's milk allergy (DRACMA) guidelines. The World Allergy Organization Journal, 3(4), 57–161. https://doi.org/10.1097/WOX.0b013e3181defeb7
[13] Lee, A. J., Thalayasingam, M., & Lee, B. W. (2013). Anaphylaxis from hidden allergens in Asian cuisine: A series of cases. Allergy, Asthma & Clinical Immunology, 9(1), 7. https://doi.org/10.1186/1710-1492-9-7
[14] Liu, Y., Zhang, Y., & Xu, H. (2020). Machine learning approaches for food allergen detection: A systematic review. Computers in Biology and Medicine, 124, 103962. https://doi.org/10.1016/j.compbiomed.2020.103962
[15] Zhang, Y., Wang, C., & Liu, M. (2021). Comparison of machine learning algorithms for allergen classification. Journal of Food Science, 86(7), 2593–2603. https://doi.org/10.1111/1750-3841.15757
[16] Shaukat, A., Aslam, S., & Khan, I. (2021). Efficient allergen classification using machine learning algorithms. Journal of Food Informatics, 4(1), 12–19.
[17] Ismail, M., Nugroho, T., & Lestari, A. (2023). The Chef's Choice: System for Allergen and Style Classification in Recipes. IEEE Access, 11, 67019–67030. https://doi.org/10.1109/ACCESS.2023.3276451
[18] Chen, Y., Wang, J., & Li, X. (2022). Limitations of cross-cultural allergen detection systems. Journal of Biomedical Informatics, 132, 104114. https://doi.org/10.1016/j.jbi.2022.104114
[20] Wulandari, N. (2024). Common food allergens in Indonesian cuisine: A review. Jurnal Kesehatan Masyarakat Indonesia, 10(1), 12–20.
[21] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://nlp.stanford.edu/IR-book/
[22] Jurafsky, D., & Martin, J. H. (2019). Speech and language processing (3rd ed.). Pearson. https://web.stanford.edu/~jurafsky/slp3/
[23] Streamlit Inc. (2024). Streamlit documentation. https://docs.streamlit.io/
[24] Nugroho, T., Handayani, P. W., & Azzahro, F. (2022). Designing accessible AI interfaces for food allergy detection. International Journal of Information and Communication Technology (IJICT), 6(1), 45–52.
[25] World Health Organization. (2021). Food safety and public health priorities. https://www.who.int/news-room/fact-sheets/detail/food-safety
[26] B. Indrayana, S. Hartati, and D. R. I. Setiadi, "Text classification on Indonesian news using support vector machine and feature selection," *Jurnal RESTI*, vol. 5, no. 1, pp. 1–7, 2021.
[27] A. F. Wicaksono and H. Prabowo, "Comparative study of stemming and stopword removal on Indonesian text classification," *Jurnal Ilmu Komputer dan Informasi (JIKI)*, vol. 13, no. 2, pp. 107–115, 2020.
[28] M. Zhang and Z. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recognit., vol. 40, no. 7, pp. 2038–2048, Jul. 2007, doi: 10.1016/j.patcog.2006.12.019.
[29] R. Subramanian et al., "A comparative study of machine learning algorithms for multi-label classification of food allergen ingredients," Procedia Comput. Sci., vol. 172, pp. 148–155, 2020.
[30] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[31] S. Brahimi, "AI-powered dining: text information extraction and machine learning for personalized menu recommendations and food allergy management," International Journal of Information Technology, vol. 17, no. 4, pp. 2107–2115, May 2025, doi: 10.1007/s41870-024-02154-9.
[32] S. Hidayatullah and H. Santoso, "Random Forest-Based Assessment of Mangrove Degradation Utilizing NDVI Feature Extraction in Spatio-Temporal Analysis," Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI), vol. 13, no. 1, pp. 58–65, Mar. 2024, doi: 10.23887/janapati.v13i1.71173.
[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[34] Y. Wang et al., "Multilabel food allergen prediction using gradient boosting classifiers," IEEE Access, vol. 10, pp. 54778–54788, 2022, doi: 10.1109/ACCESS.2022.3177438.
[35] M. M. Hasan et al., "Food allergen detection using machine learning approaches: A review," IEEE Access, vol. 9, pp. 53789–53804, 2021, doi: 10.1109/ACCESS.2021.3070612.
[36] K. G. Ayu, D. W. Sari, I. Farida, D. Harsono, R. W. Kosaman, I. H. Sumitro, F. N. Iman, M. Mansuri, E. R. Kaburuan, and M. A. Fitri, "Classification of Employee Competency Assessment Using Naïve Bayes and K-Nearest Neighbor (KNN) Algorithms," Journal of Advances in Information Technology, vol. 15, no. 7, pp. 879–885, Jul. 2024. doi: 10.12720/jait.15.7.879-885