p-ISSN: 1907-8056
e-ISSN: 2527-5410
AGROINTEK
Jurnal Teknologi Industri Pertanian
journal homepage: journal.trunojoyo.ac.id/agrointek

Allergen classification and ingredient substitution strategy for indonesian cuisine using clustered random forest

Aldy Rahmat Rizky Yulianto*, Aji Prasetya Wibawa, Ahmad 'Ammar Musyaffa', Adil Zakaria, David Satria Alamsyah, Agung Bella Putra Utama

Universitas Negeri Malang, Indonesia

Article history
Received:
3 December 2025
Revised:
25 June 2026
Accepted:
13 July 2026

Keyword
Cookies;
Allergen Detection;
Food Safety Certification;
Machine Learning;
Indonesian Cuisine;
Ingredient Substitution;
MSMEs

ABSTRACT

The absence of standardized recipes in traditional Indonesian cuisine poses a challenge for Micro, Small, and Medium Enterprises (MSMEs) in complying with food safety regulations, particularly allergen labeling requirements. Manual allergen identification is inefficient, susceptible to bias, and costly for routine industrial screening. This study proposes a cost-effective automated computational framework for detecting allergen risks and supporting product reformulation. A dataset comprising 9,921 recipes extracted from the Cookpad platform was analyzed using a semi-supervised approach. K-Means clustering was employed to generate risk labels, while Random Forest classification was used to predict allergen-risk categories mapped to 14 European Union allergen standards. Semantic data augmentation and Random Over Sampling (ROS) were applied to improve ingredient diversity and class representation during model development. The proposed framework achieved consistently high predictive performance across different clustering scenarios, with accuracy reaching 0.999 and recall exceeding 0.997. The combination of semantic augmentation and ROS produced the most stable classification performance and reduced the likelihood of false-negative predictions, which are critical in food-safety applications. Bayesian Optimization achieved predictive performance comparable to Grid Search while requiring approximately 6.3 times less computational time, thereby improving the practicality of large-scale allergen screening and food-safety monitoring. Beyond allergen detection, this study introduces an Ingredient Substitution Advisor based on clustering patterns to support food technologists in formulating hypoallergenic products. The proposed framework contributes to the digitalization of food safety systems by providing an accurate and low-cost decision-support tool for HACCP implementation and certification in the agro-industrial sector.

This work is licensed under a Creative Commons Attribution 4.0 International License.

* Corresponding author
Email : aldy.rahmat.2205356@students.um.ac.id
DOI 10.21107/agrointek.v20i3.32625

INTRODUCTION

Food allergens represent a critical issue in modern food safety management systems due to their potential to trigger severe immunological reactions, including anaphylaxis (Krisnawati et al. 2022). Consequently, global standards and regulations, such as those in the European Union (EU Regulation No. 1169/2011) and Indonesia (BPOM) (Chang et al. 2023), mandate strict allergen labeling on food products. Compliance with these regulations is not merely a legal requirement but also a fundamental aspect of quality assurance in the agro-industry (Dominguez et al. 2022). However, implementing accurate allergen labeling remains a complex operational challenge, particularly for food manufacturers who must ensure consumer safety while managing diverse product lines.

In the context of the Indonesian culinary sector, this challenge is amplified by the unique characteristics of traditional food production. Unlike large-scale manufacturing with fixed and standardized formulations, Indonesian cuisine is known for its rich diversity of ingredients, complex spice mixtures (*bumbu*) (Yayusman et al. 2023), and regional variations in cooking techniques (Cempaka et al. 2023). This variability imposes a substantial burden on Micro, Small, and Medium Enterprises (MSMEs), which constitute the majority of businesses in the local food industry (Dominguez et al. 2022). For these smaller entities, manually identifying potential allergens in non-standardized recipes is inefficient, prone to human error, and inconsistent, creating a barrier to obtaining food safety certification and entering broader markets (Roither et al. 2022).

Laboratory-based methods such as Enzyme-Linked Immunosorbent Assay (ELISA) and Polymerase Chain Reaction (PCR) are established protocols for definitive allergen verification. Within agro-industrial management, these analytical techniques are critical for final quality control (Holzhauser et al. 2020), (Linacero et al. 2020). In addition to these physical tests, data-driven approaches are increasingly necessary to efficiently screen extensive traditional recipe collections and facilitate early-stage product development. An automated system using recipe composition data may serve as a preliminary screening tool, enabling more targeted monitoring and improving the efficiency of food safety auditing in the developing sector (Mu et al. 2024).

This study introduces a semi-supervised machine learning framework for allergen-risk analysis in Indonesian recipes. The framework integrates allergen-oriented feature engineering, semantic recipe augmentation, cluster-based pseudo-label generation, and Random Forest classification to facilitate automated allergen screening. In addition to predictive modeling, the research examines the impact of dataset enhancement and optimization strategies on classification performance and computational efficiency (Lu et al. 2023, Onorato 2024). The resulting clustering patterns are also used to generate ingredient substitution recommendations, thereby supporting hypoallergenic product reformulation and food safety management (Pellegrini et al. 2021).

METHODS

This study developed a semi-supervised machine learning framework for allergen risk identification in Indonesian cuisine. The proposed framework integrates unsupervised clustering and supervised classification to transform unlabeled recipe data into actionable allergen-risk information. As illustrated in Figure 1, the workflow consists of dataset construction, allergen feature engineering, semantic recipe augmentation, semi-supervised risk labeling generation, allergen risk classification, and model evaluation. Each stage contributes to generating a scalable food-safety screening system suitable for large culinary databases.

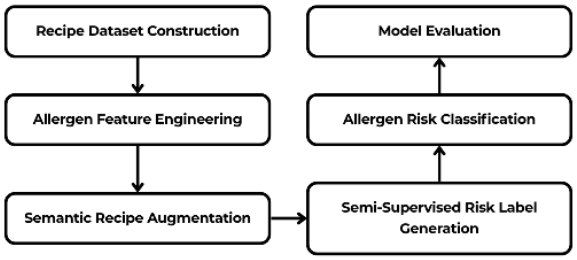


Figure 1 Flowchart of Implementation Steps

Recipe Dataset Construction

The dataset was compiled from publicly accessible Indonesian recipes on the Cookpad platform. A web-scraping mechanism extracted recipe titles, ingredient lists, cooking instructions, ratings, and source links. After data collection, non-food categories, including beverages, were excluded to maintain alignment with the study objective.

After filtering, 9,921 food recipes were retained for analysis. The resulting dataset includes a wide range of regional dishes and ingredient combinations, offering sufficient variability to investigate allergen patterns and support machine learning model development.

Allergen Feature Engineering

Because recipe ingredients were initially stored as unstructured textual descriptions, a feature engineering process was necessary prior to model development. Ingredient names were standardized to address inconsistencies arising from variations in spelling, notation, and alternative naming conventions prevalent in Indonesian recipes. The standardized ingredients were subsequently mapped to the fourteen priority allergen categories established by EU Regulation No. 1169/2011, a widely recognized reference framework for food allergen management (Roither et al. 2022, Chang et al. 2023)

After the mapping process, each recipe was converted into a binary allergen representation, where 1 indicates the presence of a specific allergen category, and 0 denotes its absence. This approach allows heterogeneous ingredient information to be transformed into a structured numerical format appropriate for machine learning applications, while maintaining food safety related characteristics. Table 1 provides examples of allergen feature extraction from Indonesian recipes, demonstrating how ingredient-based recipe descriptions are transformed into standardized allergen attributes.

The allergen profile for each recipe is represented as a binary feature vector, as shown in Equation (1).

$$X_i = [a_{i,1}, a_{i,2}, \ldots, a_{i,14}] \quad (1)$$

Where X_i denotes the allergen feature vector of a recipe, and a_i indicates the presence (1) or absence (0) of the corresponding allergen category. Through this representation, recipes with diverse ingredient compositions can be described using a consistent feature space, enabling subsequent clustering and classification processes to identify allergen-risk patterns more effectively.

Semantic Recipe Augmentation

Data augmentation was introduced to enrich ingredient diversity and improve model generalization. This strategy follows recent findings showing that semantic augmentation can enhance recipe-based learning systems by exposing models to realistic ingredient variations (Song et al. 2024). These patterns frequently appear in Indonesian recipes and often influence allergen composition.

As illustrated in Table 2, a single recipe may generate several valid ingredient configurations depending on whether optional ingredients are included and which alternative ingredients are selected. Such modifications can produce distinct allergen profiles while preserving the culinary identity of the original dish.

Through this strategy, the dataset expanded from 9,921 to 15,031 samples. Unlike conventional synthetic augmentation methods, all generated samples remained semantically valid because they originated from realistic ingredient substitutions already specified within the recipe.

Table 1 Example of Recipe Dataset and Allergen Attributes

Titles		Telur Balado	Kotok Udang Tahu Tempe (masakan Jawa Tengah)
Ingredients		1/2 Kg Telur (kurleb 8 butir), Secukupnya Air (untuk merebus telur), Secukupnya Minyak Goreng…	1 blok tempe, 3 buah tahu, 200 gr udang, Iris Bumbu, 5 siung bawang merah, 2 siung bawang putih, 3 buah cabai merah…
Allergens	**Cereal**	0	0
	Shrimp	0	1
	Fish	1	0
	...	...	...
	Mollusks	0	0

Table 2 Illustration of Semantic Data Augmentation for Food Ingredient Recipes

Original Ingredient List	Keywords	Optional Item	Augmentation Count
2 butir telur (boleh ditambah lagi sesuai kebutuhan), 1/2 ekor ayam (opsional), 1 ekor ikan bandeng (opsional), 6 butir bawang putih, 5 butir bawang merah, 2 buah lombok merah besar, 2 buah lombok rawit, Secukupnya gula merah, Secukupnya Garam, micin, penyedap rasa	optional, / (or)	*1/2 ekor ayam, 1 ekor ikan bandeng*	$(2^2) = 4$
After Augmentation			
Augmented Ingredient List	**Item removed from Original**		**Augmentation Type**
2 butir telur (boleh ditambah lagi sesuai kebutuhan), 1 ekor ikan bandeng, 6 butir bawang putih...	*½ ekor ayam*		Augmentation 1
2 butir telur (boleh ditambah lagi sesuai kebutuhan), 1 ekor ikan bandeng, 6 butir bawang putih, 5 butir bawang merah	*1 ekor ikan bandeng*		Augmentation 2
2 butir telur (boleh ditambah lagi sesuai kebutuhan), 1/2 ekor ayam, 1 ekor ikan bandeng, 6 butir bawang putih...	-		Augmentation 3
2 butir telur (boleh ditambah lagi sesuai kebutuhan), 6 butir bawang putih...	*½ ekor ayam, 1 ekor ikan bandeng*		Augmentation 4

Semi-Supervised Risk Label Generation

The collected recipe dataset did not contain predefined allergen-risk labels, making direct supervised learning infeasible. To address this limitation, a semi-supervised labeling strategy was implemented using K-Means clustering. This approach enables the discovery of latent structures within the data by grouping recipes according to similarities in their allergen compositions rather than relying on manually assigned categories. K-Means was selected because of its effectiveness in partitioning numerical feature spaces and its ability to reveal naturally occurring patterns in unlabeled datasets (Sinaga and Yang 2020).

Clustering was conducted using allergen feature vectors generated in the preceding stage. Cluster centroids were initially assigned and then updated through an iterative optimization procedure. In each iteration, recipes were assigned to the nearest centroid, and centroid positions were recalculated until convergence. The objective of this process is to minimize within-cluster variation, as defined in Equation (2).

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (2)$$

Where $x_i^{(j)}$ represents the feature vector assigned to cluster j, and c_j denotes the centroid of that cluster. The optimization process aims to minimize the distance between data points and their corresponding centroids, thereby producing clusters with similar allergen characteristics.

The number of clusters plays an important role in determining the granularity of risk categorization. Therefore, four clustering configurations were investigated, namely k = 3, 5, 7, and 9. Lower values of *k* generate broader risk groups that provide general segmentation of allergen patterns, whereas larger values allow more detailed differentiation among recipes with similar allergen profiles. Evaluating multiple cluster configurations enables assessment of how label granularity influences subsequent classification performance.

After the clustering stage, each cluster assignment was converted into a pseudo-label representing a relative allergen-risk category. These pseudo-labels transformed the originally unlabeled dataset into a form suitable for supervised learning while preserving the intrinsic structure identified during clustering. Such a strategy is consistent with semi-supervised learning paradigms, where naturally emerging data distributions are exploited to construct predictive models without extensive manual annotation requirements (Song et al. 2022).

Allergen Risk Classification

The pseudo-labels generated from the clustering stage were subsequently used as target

classes for supervised learning. In this study, the Random Forest (RF) algorithm was employed as the primary classifier due to its capability to handle high-dimensional feature spaces, robustness against overfitting, and stable predictive performance across heterogeneous datasets (Dudek 2022). These characteristics make Random Forest particularly suitable for recipe-based data, where allergen patterns may vary considerably across food formulations.

Random Forest operates by constructing multiple decision trees from bootstrap samples of the training data. During model development, each tree independently learns classification rules using randomly selected subsets of observations and features. The final class prediction is obtained through a majority-voting mechanism across all decision trees within the ensemble. Each tree contributes one vote, and the class receiving the highest number of votes is selected as the final output. The voting process can be mathematically expressed as follows:

$$\hat{y} = \arg\max_{k \in \mathcal{K}} \sum_{b=1}^{B} 1(h_b(x) = k) \qquad (3)$$

where $h_b(x)$denotes the prediction generated by the b-th decision tree, Brepresents the total number of trees in the forest, and Kdenotes the set of possible class labels. This ensemble mechanism improves classification robustness by reducing prediction variance and limiting overfitting (Ghosh and Cabrera 2021).

One challenge encountered in allergen-risk classification is the imbalance of class distributions resulting from the clustering process. Certain risk categories contain substantially fewer observations than others, potentially causing the classifier to favor majority classes. To address this issue, Random Over Sampling (ROS) was applied to the training data. ROS increases the representation of minority classes by duplicating existing observations until a more balanced distribution is achieved. This approach aims to improve the model's sensitivity toward underrepresented allergen-risk categories while preserving the original feature characteristics of each sample.

Table 3 Hyperparameter Search Configuration for Model Optimization

No.	Hyperparameters	Type	Search Space
1	Min samples split	Discrete	[2, 4, 8, 16]
2	Max depth	Discrete	[5, 10, 15, 20]
3	N estimators	Discrete	[25, 50, 100, 150]
4	max features	Discrete	['sqrt', 'log2']

Table 4 Classification Performance Using the Original Recipe Dataset

Class Label	Tuning Parameters	ROS	Accuracy	Precision	Recall	F1-Score	Computation Time (ms)
0 - 2	None	No	0.999	0.999	0.999	0.999	7,573
0 – 4			0.999	0.999	0.999	0.999	4,039
0 – 6			0.998	0.998	0.996	0.997	4,467
0 – 8			0.999	0.998	0.997	0.998	4,528
0 - 2	Gridsearch	No	0.998	0.999	0.998	0.998	344,615
0 – 4			0.998	0.997	0.998	0.997	338,996
0 – 6			0.997	0.997	0.996	0.996	358,921
0 – 8			0.999	0.999	0.998	0.999	361,938
0 - 2	Bayesian	No	0.998	0.999	0.998	0.998	45,544
0 – 4			0.998	0.997	0.998	0.997	59,096
0 – 6			0.997	0.996	0.995	0.996	69,113
0 – 8			0.999	0.999	0.998	0.999	60,476

Table 5 Classification Performance Using the Original Dataset with Random Over Sampling

Class Label	Tuning Parameters	ROS	Accuracy	Precision	Recall	F1-Score	Computation Time (ms)
0 - 2	None	Yes	0.999	1	0.999	0.999	17,241
0 – 4			0.999	0.999	0.999	0.999	7,033
0 – 6			0.998	0.997	0.996	0.997	5,424
0 – 8			0.999	0.998	0.997	0.998	6,501
0 - 2	Gridsearch	Yes	0.998	0.999	0.998	0.998	391,433
0 – 4			0.999	0.999	0.999	0.999	467.430
0 – 6			0.997	0.997	0.996	0.996	539,803
0 – 8			0.999	1	0.999	1	687,042
0 - 2	Bayesian	Yes	0.998	0.999	0.998	0.998	54,406
0 – 4			0.999	0.999	0.999	0.999	76,553
0 – 6			0.995	0.994	0.994	0.994	85,997
0 – 8			0.999	1	0.999	1	109,885

To further enhance classification performance, hyperparameter tuning was conducted using three different optimization strategies: default parameter settings, Grid Search, and Bayesian Optimization. The default configuration served as a baseline model, while Grid Search systematically evaluated all parameter combinations within a predefined search space. Although exhaustive, this method often requires considerable computational resources due to the large number of candidate configurations evaluated (Rohman et al. 2025).

As an alternative optimization strategy, Bayesian Optimization was implemented to improve search efficiency. Rather than evaluating all possible combinations, Bayesian Optimization constructs a probabilistic surrogate model to estimate the relationship between hyperparameter values and model performance, allowing the algorithm to focus on promising regions of the search space (Lu et al. 2023). This strategy has been shown to reduce optimization time while maintaining competitive predictive performance (Onorato 2024).

The hyperparameters considered in this study include the number of trees (n_estimators), maximum tree depth (max_depth), minimum samples required for node splitting (min_samples_split), and feature-selection strategy (max_features). The search ranges were adopted from previous studies reporting effective Random Forest configurations for classification tasks (Afuan and Isnanto 2025). A complete description of the explored parameter space is presented in Table 3.

To investigate the combined effects of data balancing and optimization strategies, experiments were conducted under four data scenarios: raw data, raw data with ROS, augmented data, and augmented data with ROS. The resulting models were subsequently evaluated to identify the most effective configuration for allergen-risk prediction.

Model Evaluation

The proposed framework was evaluated using confusion-matrix-based metrics, namely Accuracy, Precision, Recall, and F1-Score. These metrics were selected because they provide complementary information regarding classification performance, ranging from overall predictive capability to the model's ability to correctly identify allergen-risk categories. Given the food-safety context of this study, particular emphasis was placed on Recall, as false-negative predictions may result in undetected allergens and potentially compromise consumer safety (Adedeji et al. 2024).

Accuracy measures the proportion of correctly classified observations among all evaluated samples and is defined by Equation (4).

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision evaluates the reliability of positive predictions by quantifying the proportion of correctly predicted positive samples relative to all

predicted positive samples, as shown in Equation (5).

$$\frac{TP}{TP+FP} \quad (5)$$

Recall, also referred to as Sensitivity, measures the ability of the model to identify all relevant positive instances and is calculated using Equation (6).

$$\frac{TP}{TP+FN} \quad (6)$$

To provide a balanced assessment between Precision and Recall, the F1-Score was also calculated using their harmonic mean, as presented in Equation (7).

$$2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (7)$$

where TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative classifications, respectively.

In addition to predictive performance, computational time was recorded to evaluate the efficiency of each model configuration. This metric enables a comparative assessment of the trade-off between classification effectiveness and computational cost, which is particularly relevant for large-scale deployment in food-safety management systems.

The evaluation procedure was applied to all clustering scenarios (k = 3, 5, 7, and 9), as well as to every combination of data balancing and optimization strategy. The resulting performance metrics were then compared to determine the most accurate, reliable, and computationally efficient model configuration for allergen-risk classification.

RESULT AND DISCUSSION

Result

Table 4 presents the baseline performance obtained from the original recipe dataset without augmentation or class-balancing techniques. Across all clustering scenarios, the Random Forest classifier achieved consistently high predictive performance, with accuracy ranging from 0.998 to 0.999 and recall above 0.996. These results indicate that the allergen-based feature representation successfully captured meaningful distinctions among recipes and enabled reliable allergen-risk classification.

Although performance remained stable across all configurations, a slight decline in Recall and F1-Score was observed as the number of clusters increased from three to nine. This trend is expected because higher clustering configurations yield allergen-risk segmentation, thereby increasing classification complexity. From a computational perspective, the baseline dataset also required the shortest training time among all dataset-treatment scenarios, reflecting the lower processing requirements of the original recipe collection.

Table 6 Classification Performance Using the Augmented Recipe Dataset

Class Label	Tuning Parameters	ROS	Accuracy	Precision	Recall	F1-Score	Computation Time (ms)
0 - 2	None	No	0.999	0.999	0.999	0.999	17,393
0 – 4			0.999	1	0.999	0.999	6,343
0 – 6			0.994	0.995	0.987	0.990	4,244
0 – 8			0.998	0.997	0.996	0.996	5,682
0 - 2	Gridsearch	No	1	1	0.999	1	398,297
0 – 4			1	1	1	1	443,105
0 – 6			0.999	0.999	0.997	0.998	453,842
0 – 8			0.999	0.998	0.997	0.997	470,192
0 - 2	Bayesian	No	1	1	0.999	1	65,732
0 – 4			1	1	1	1	68,904
0 – 6			0.999	0.999	0.998	0.999	75,130
0 – 8			0.999	0.999	0.998	0.998	78,483

Table 7 Classification Performance Using the Augmented Dataset with Random Over Sampling

Class Label	Tuning Parameters	ROS	Accuracy	Precision	Recall	F1-Score	Computation Time (ms)
0 - 2	None	Yes	0.999	0.999	0.999	0.999	9,664
0 – 4			0.999	0.999	0.999	0.999	7,642
0 – 6			0.996	0.994	0.992	0.993	7,103
0 – 8			0.998	0.997	0.996	0.997	9.387
0 - 2	Gridsearch	Yes	1	1	0.999	1	469,960
0 – 4			1	1	1	1	689,732
0 – 6			0.999	0.999	0.998	0.999	733,407
0 – 8			0.999	0.999	0.998	0.998	862,619
0 - 2	Bayesian	Yes	1	1	0.999	1	74,767
0 – 4			1	1	1	1	99.867
0 – 6			0.999	0.999	0.998	0.999	116,739
0 – 8			0.999	0.998	0.997	0.998	138,388

The impact of class balancing was assessed by implementing Random Over Sampling (ROS) on the original dataset, as summarized in Table 5. ROS preserved the high predictive performance observed in the baseline experiments and yielded more consistent Recall values across various clustering scenarios. This improvement was particularly evident in the seven- and nine-class configurations, where recipes were distributed among a greater number of allergen-risk categories and class distributions became more uneven.

Although the increase in Accuracy was modest, ROS enhanced the representation of minority allergen-risk categories and reduced the classifier's bias toward majority classes. This improvement is significant from a food-safety perspective, as it decreases the likelihood of overlooking less-represented allergen-risk categories. The balancing process resulted in a slight increase in computational time due to the larger training set size; however, the additional processing cost was minimal compared to the improvements in classification stability.

Table 6 presents the classification performance after applying semantic recipe augmentation. The dataset was expanded by generating valid ingredient variations from optional and alternative recipe components, thereby preserving culinary consistency. The augmented dataset achieved near-perfect Accuracy and recall across all clustering scenarios, suggesting that the additional samples contributed meaningful allergen information rather than introducing noise.

The advantages of augmentation were most apparent in higher clustering configurations, where the classifier needed to distinguish among a greater number of allergen-risk categories. Increased ingredient diversity enhanced the model's capacity to generalize across heterogeneous recipe formulations and maintained performance despite greater label granularity. As anticipated, the larger dataset resulted in longer computation times for the baseline, due to increased processing demands during model training.

The combined application of semantic augmentation and Random Over Sampling (ROS) produced the most consistent classification performance among all evaluated dataset strategies, as shown in Table 7. Across all clustering scenarios, Accuracy, Precision, Recall, and F1-Score remained close to 1.000, indicating that the classifier maintained both predictive accuracy and detection sensitivity as allergen-risk segmentation structures became increasingly complex.

The complementary nature of augmentation and ROS appears to be responsible for this performance. While augmentation enriched the diversity of ingredient combinations available during training, ROS ensured sufficient representation of minority allergen-risk categories. The combined strategy was particularly beneficial in the seven- and nine-class scenarios, where classification

complexity was highest. Although this configuration required the longest computational time due to the increased dataset size and balancing process, the resulting gains in classification robustness suggest that the additional computational cost is justified.

Collectively, the results presented in Tables 4–7 demonstrate that the proposed semi-supervised framework accurately identifies allergen-risk categories across varying levels of classification complexity. While the baseline model already achieved excellent predictive performance, incorporating augmentation and ROS further improved classification robustness, particularly in higher-clustering configurations. The findings also reveal a trade-off between predictive performance and computational requirements, highlighting the importance of evaluating not only classification accuracy but also the efficiency of different optimization and data-enhancement strategies.

Discussion

While all optimization strategies demonstrated similarly high predictive performance, notable differences in computational efficiency were evident, as presented in Table 8. Grid Search yielded the highest average Accuracy and Recall, but its performance advantage over Bayesian Optimization was minimal.

Conversely, Bayesian Optimization achieved comparable predictive results while requiring approximately 6.3 times less computational time. This outcome aligns with previous research indicating that Bayesian Optimization can efficiently identify high-performing parameter configurations without exhaustive search of the parameter space (Lu et al. 2023, Onorato 2024).

In the agro-industrial context, computational efficiency is especially important for micro, small, and medium-sized enterprises (MSMEs) and food processors that frequently operate with limited technical infrastructure. Accelerated model training and deployment can enable the routine integration of digital allergen-screening tools into food safety management practices. Therefore, maintaining high predictive performance while minimizing computational demands supports the practical adoption of allergen monitoring within prerequisite programs (PRPs) and HACCP-based food safety systems (Hassoun et al. 2023).

Table 9 presents the overall contributions of augmentation and class-balancing strategies to allergen-risk classification.

The findings demonstrate that both techniques enhanced model performance, with each serving complementary functions. Semantic augmentation increased ingredient diversity by introducing realistic recipe variations, while random oversampling (ROS) improved the representation of minority allergen-risk categories during training.

Table 8 Trade-Off Between Predictive Performance and Computational Cost of Hyperparameter Optimization Methods

Optimization Method	Avg Accuracy	Avg Recall	Avg F1	Avg Time (ms)	Computational Efficiency Gain
Default Parameters	0.9983	0.9968	0.9974	7,767	64.5× faster
Grid Search	0.9988	0.9981	0.9984	500,708	Reference
Bayesian Optimization	0.9987	0.9980	0.9984	79,943	6.3× faster

Table 9 Effect of Dataset Enhancement Strategies on Overall Classification Performance

Dataset Strategy	Avg Accuracy	Avg Recall	Avg F1-Score	Avg Time (ms)
Raw	0.9982	0.9973	0.9977	159,159
Raw + ROS	0.9984	0.9978	0.9980	202,150
Augmentation	0.9990	0.9981	0.9985	167,968
Augmentation + ROS	0.9990	0.9985	0.9988	246,838

Table 10 Ingredient Substitution Recommendations Based on Allergen Tendencies and Culinary Function

Example Dish	Risk Ingredient	Risk Category	Suggested Alternative	Reformulation Objective
Pecel	Peanut	High Peanut Risk	Sunflower seeds	Peanut-free product
Gado-gado	Peanut sauce	High Peanut Risk	Yellow peas	Reduced allergen content
Roa chili sauce	Shrimp paste	Crustacean Risk	Mushroom extract	Seafood-free formulation
Tempeh orek	Soybean	Soy Risk	Chickpea tofu	Soy-free alternative
Fried noodles	Wheat flour	Gluten Risk	Rice flour	Gluten-free product

The highest average Recall and F1-Score were achieved when both augmentation and ROS were applied concurrently, suggesting that effective allergen-risk identification relies on both the quantity and balance of training samples. From a food-safety perspective, improved Recall is critical because it reduces the likelihood of false-negative predictions, which could result in allergens remaining undetected in food products (Adedeji et al. 2024). Consequently, the combined strategy provides a more robust foundation for hazard identification activities that support allergen control and preventive food safety management.

Beyond classification performance, the proposed framework offers practical utility for allergen management within HACCP-based food production systems. Its ability to identify recipes associated with specific allergen-risk categories supports food processors in conducting hazard analysis and determining appropriate control measures at Critical Control Points (CCPs). Additionally, the framework can strengthen prerequisite programs (PRPs) by providing a systematic method for allergen screening and ingredient-risk assessment prior to product commercialization.

The ingredient substitution recommendations presented in Table 10 further strengthen the framework's practical relevance.

By identifying alternative ingredients that maintain similar culinary functions while reducing allergen risks, this approach facilitates product reformulation without sacrificing traditional product characteristics. Comparable strategies have been recognized as effective for developing hypoallergenic food products and enhancing consumer accessibility without diminishing product quality (Pellegrini et al. 2021). These features are particularly relevant for micro, small, and medium-sized enterprises (MSMEs) seeking to improve allergen labeling, enhance food safety compliance, and implement HACCP at lower operational costs. Collectively, these findings suggest that the proposed framework functions as both an allergen-risk classification tool and a decision-support system for preventive allergen management, product reformulation, and food-safety assurance in the agro-industrial sector (Mu et al. 2024).

CONCLUSION

This study demonstrates that the proposed semi-supervised framework can effectively support allergen-risk identification in Indonesian cuisine by integrating allergen-based feature engineering, K-Means clustering, and Random Forest classification. Across all experimental configurations, the framework consistently achieved near-perfect predictive performance, with Accuracy reaching 0.999 and Recall exceeding 0.997. These results indicate that allergen-oriented recipe representations provide a reliable foundation for automated food-safety screening, even when explicit allergen-risk labels are unavailable.

The findings further reveal that dataset enhancement strategies play a significant role in improving model robustness. Semantic augmentation enriched ingredient diversity, while Random Over Sampling (ROS) improved the representation of minority classes, resulting in more stable classification performance across increasingly complex clustering scenarios. Among the evaluated optimization strategies, Bayesian Optimization achieved predictive performance comparable to Grid Search while requiring approximately 6.3 times less computational time, making it a more practical solution for large-scale implementation.

Beyond allergen-risk classification, the proposed framework provides practical support for ingredient substitution and product reformulation by identifying alternative ingredients with comparable culinary functions but lower allergen risks.

The generated recommendations provide a systematic basis for identifying alternative ingredients with similar culinary functions but reduced allergen risks. From an agro-industrial perspective, the framework has the potential to support allergen management activities within HACCP-based food-safety systems by assisting hazard identification, strengthening allergen labeling practices, and facilitating product reformulation. These capabilities are particularly relevant for MSMEs seeking cost-effective approaches to improve food-safety compliance while preserving the authenticity of traditional Indonesian food products.

REFERENCE

Adedeji, A. A., P. V Priyesh, and A. A. Odugbemi. 2024. The magnitude and impact of food allergens and the potential of AI-Based Non-Destructive testing methods in their detection and quantification. Foods 13(7):994. https://doi.org/10.3390/foods13070994

Afuan, L., and R. R. Isnanto. 2025. Enhanced Fall Detection using Optimized Random Forest Classifier on Wearable Sensor Data. Journal of Applied Data Sciences 6(1):213–224. https://doi.org/10.47738/jads.v6i1.498

Cempaka, L., A. A. Rizki, N. Asiah, W. David, K. Ramadhan, A. S. Mukaromah, H. Pramastya, F. Husain, and N. Huda. 2023. Characteristics of Soto, an ethnic food that reflects Indonesian diversity: Based on ingredients. Canrea Journal: Food Technology, Nutritions, and Culinary Journal:1–16. https://doi.org/10.20956/canrea.v6i1.680

Chang, F., L. Eng, and C. Chang. 2023. Food Allergy Labeling Laws: International Guidelines for Residents and Travelers. Clinical Reviews in Allergy & Immunology 65(2):148–165. https://doi.org/10.1007/s12016-023-08960-6

Dominguez, S., J. Théolier, B. Povolo, J. Gerdts, and S. B. Godefroy. 2022. Allergen management under a voluntary PAL regulatory framework–A survey of Canadian food processors. Heliyon 8(11). https://doi.org/10.1016/j.heliyon.2022.e11302

Dudek, G. 2022. A comprehensive study of random forest for short-term load forecasting. Energies 15(20):7547. https://doi.org/10.3390/en15207547

Ghosh, D., and J. Cabrera. 2021. Enriched random forest for high dimensional genomic data. IEEE/ACM transactions on computational biology and bioinformatics 19(5):2817–2828. https://doi.org/10.1109/TCBB.2021.3089417

Hassoun, A., S. Jagtap, H. Trollman, G. Garcia-Garcia, N. A. Abdullah, G. Goksen, F. Bader, F. Ozogul, F. J. Barba, and J. Cropotova. 2023. Food processing 4.0: Current and future developments spurred by the fourth industrial revolution. Food Control 145:109507. https://doi.org/10.1016/j.foodcont.2022.109507

Holzhauser, T., P. Johnson, J. P. Hindley, G. O'Connor, C.-H. Chan, J. Costa, C. K. Fæste, B. J. Hirst, F. Lambertini, and M. Miani. 2020. Are current analytical methods suitable to verify VITAL® 2.0/3.0 allergen reference doses for EU allergens in foods? Food and Chemical Toxicology 145:111709. https://doi.org/10.1016/j.fct.2020.111709

Krisnawati, D. I., M. Alimansur, D. S. Atmojo, E. Q. Rahmawati, D. Rahayu, E. Susilowati, and T.-R. Kuo. 2022. Food allergies: immunosensors and management. Applied Sciences 12(5):2393. https://doi.org/10.3390/app12052393

Linacero, R., A. Sanchiz, I. Ballesteros, and C. Cuadrado. 2020. Application of real-time PCR for tree nut allergen detection in processed foods. Critical Reviews in Food Science and Nutrition 60(7):1077–1093. https://doi.org/10.1080/10408398.2018.1557103

Lu, Q., K. D. Polyzos, B. Li, and G. B. Giannakis. 2023. Surrogate modeling for Bayesian optimization beyond a single Gaussian process. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(9):11283–11296. https://doi.org/10.1109/TPAMI.2023.3264741

Mu, W., G. A. Kleter, Y. Bouzembrak, E. Dupouy, L. J. Frewer, F. N. Radwan Al Natour, and H. J. P. Marvin. 2024. Making food systems more resilient to food safety risks by including artificial intelligence, big data,

and internet of things into food safety early warning and emerging risk identification tools. Comprehensive Reviews in Food Science and Food Safety 23(1):e13296. https://doi.org/10.1111/1541-4337.13296

Onorato, G. 2024. Bayesian optimization for hyperparameters tuning in neural networks. arXiv preprint arXiv:2410.21886. https://doi.org/10.48550/arXiv.2410.21886

Pellegrini, C., E. Özsoy, M. Wintergerst, and G. Groh. 2021. Exploiting Food Embeddings for Ingredient Substitution. HEALTHINF 5:67–77. https://doi.org/10.5220/0010202000670077

Rohman, F. N., F. Farikhin, and B. Surarso. 2025. Hyperparameter Tuning of Random Forest Algorithm for Diabetes Classification. International Journal of Current Science Research and Review 8(01). https://doi.org/10.47191/ijcsrr/V8-i1-31

Roither, A., M. Kurz, and E. Sonnleitner. 2022. The chef's choice: system for allergen and style classification in recipes. Applied Sciences 12(5):2590. https://doi.org/10.3390/app12052590

Sinaga, K. P., and M.-S. Yang. 2020. Unsupervised K-means clustering algorithm. IEEE access 8:80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Song, F., B. Zhu, Y. Hao, and S. Wang. 2024. Enhancing recipe retrieval with foundation models: A data augmentation perspective. Pages 111–127 *European Conference on Computer Vision*. Springer. https://doi.org/10.1007/978-3-031-72983-6_7

Song, Z., X. Yang, Z. Xu, and I. King. 2022. Graph-based semi-supervised learning: A comprehensive review. IEEE Transactions on Neural Networks and Learning Systems 34(11):8174–8194. https://doi.org/10.1109/TNNLS.2022.3155478

Yayusman, M. S., U. K. Yaumidin, and P. N. Mulyasari. 2023. On considering Australia: exploring Indonesian restaurants in promoting ethnic foods as an instrument of Indonesian gastrodiplomacy. Journal of Ethnic Foods 10(1):43. https://doi.org/10.1186/s42779-023-00207-1