

Automated Multi-Scenario Classification of Food Allergens in Indonesian Culinary Datasets Using Bernoulli Naïve Bayes with Augmentation and Oversampling Methods

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

Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Malang, Indonesia

*Corresponding author, Email: aji.prasetya.ft@um.ac.id

Received 28/10/2025

Revised 06/03/2026

Accepted 01/04/2026

ABSTRACT

Limited awareness of allergen content in traditional Indonesian foods poses a potential health risk because recipe platforms rarely provide explicit allergen information. This study develops an automated allergen classification system for Indonesian recipes using the Bernoulli Naïve Bayes algorithm. The contributions include the construction of a recipe-level allergen dataset, the integration of text preprocessing, augmentation, and resampling techniques, and the introduction of multi-scenario allergen classification to support safer food choices. Recipe data were collected from *Cookpad*, producing 9,921 cleaned recipes after preprocessing. To increase data diversity and address class imbalance, rule-based ingredient augmentation expanded the dataset to 15,030 recipes. The model was evaluated under three classification scenarios: binary allergen detection (0–1), allergen exposure levels (0–3), and detailed allergen counts (0–14). Each scenario was tested using four dataset configurations: raw data, raw + ROS (Random Oversampling), augmented data, and augmented + ROS. Class distributions become increasingly imbalanced as the number of classes grows, particularly in the 15-class scenario. The binary scenario achieved perfect performance, with accuracy, precision, recall, and F1-score all equal to 1.0 across all dataset variants. Scenario 2 achieved the best multi-class performance with an F1-score of 0.9293 using Augmented + ROS data. Performance decreased in the 15-class setting due to increased classification complexity. Although a perfect binary score may raise concerns about overfitting, the result likely reflects the clear separability of the dataset’s allergen features. It should be interpreted cautiously when applied to broader culinary data.

KEYWORDS

Augmentation; Food allergens; Indonesian culinary; Machine learning; Naive bayes

1. INTRODUCTION

Food allergies represent a growing global public health concern, with increasing prevalence across both developed and developing countries [1]. Allergic reactions are often triggered by unintended exposure to allergenic ingredients due to insufficient allergen information, particularly in foods prepared by restaurants and micro, small, and medium enterprises (MSMEs) [2]. In Indonesia, online recipe-sharing platforms such as *Cookpad* have become a major source of culinary information, providing extensive collections of traditional recipes. However, allergen-related information is rarely included, increasing the risk of accidental allergen consumption [3]. This has led to health risks from food allergies.

Existing strategies to reduce allergen-related risks include mandatory allergen labeling, food handler training, public awareness campaigns, and dietary consultation with healthcare professionals [4], [5], [6]. While effective at an individual or regulatory level, these approaches are difficult to scale to large collections of user-generated culinary data [7]. Manual identification of allergenic ingredients becomes impractical when processing thousands of recipes written in informal language, with inconsistent

terminology and spelling variations. This limitation underscores the need for automated systems capable of efficiently and consistently analyzing large-scale recipe data [8]. Consequently, computational approaches are increasingly explored to address the challenges of large-scale culinary data processing and allergen detection.

Recent advances in food informatics demonstrate the potential of food ontology and natural language processing (NLP) techniques for culinary data analysis [9]. Food ontologies enable structured representation of ingredients and their semantic relationships, supporting consistent interpretation of heterogeneous recipe texts and reliable allergen mapping [10]. Meanwhile, NLP-based recipe classification methods have been successfully applied to extract meaningful information from unstructured culinary text [11]. These approaches have proven effective in handling linguistic variability commonly found in user-generated recipes.

Despite these advances, most previous studies on food allergen detection have focused on packaged food labels, commercial ingredient lists, or protein-level allergen prediction [8]. The majority adopt binary classification schemes that only determine allergen presence or absence, without accounting for recipe-level context or varying exposure levels [12]. Furthermore, limited research has addressed traditional or user-generated culinary datasets, particularly within Indonesian or Southeast Asian cuisine [13]. As a result, the scientific position of recipe-based allergen classification in local culinary contexts remains underexplored.

Various machine learning models, including Support Vector Machines, Random Forests, and deep learning-based approaches such as BERT, have been applied in food and allergen classification tasks [14], [15]. Although these models often achieve strong performance, they generally require large annotated datasets, complex feature representations, and substantial computational resources. In contrast, Bernoulli Naïve Bayes is well-suited for recipe-based allergen classification due to its effectiveness on binary ingredient features and sparse textual data [16]. Its computational efficiency and interpretability also make it suitable for scalable, practical allergen-detection systems.

This study adopts the 14 major food allergen categories defined by the European Union as a reference framework [17]. Although Indonesia does not yet enforce a comprehensive national allergen classification standard, the EU framework is internationally recognized and highly relevant to Indonesian cuisine, as most defined allergens are commonly used locally. This research proposes an automated Machine Learning-based allergen classification framework for Indonesian culinary recipes using Bernoulli Naïve Bayes. The contributions include the construction of a recipe-level allergen dataset, the integration of text preprocessing, augmentation, and resampling techniques, and the introduction of multi-scenario allergen classification to support safer food choices.

2. MATERIALS AND METHODS

This study employs a research methodology comprising multiple interconnected phases to achieve its objectives. Each phase was systematically structured to ensure that the research process remains focused and measurable. Figure 1 presents a clearer illustration of these steps. This study carefully designed a research technique to develop and evaluate an allergy categorization model for food recipes. This study began by collecting raw data via online scraping, followed by preprocessing and feature engineering to prepare the data for modelling. To enhance the diversity and scale of the dataset, data augmentation was applied techniques to improve the model's generalization ability [18]. Subsequently, they further processed the prepared data using several labelling techniques, each aligned with a specific experimental setting. Then constructed a classification model using the Bernoulli Naïve Bayes method and trained and tested it across multiple scenarios. In the final stage, they rigorously evaluated the model's performance using k-fold cross-validation. They employed metrics derived from the confusion matrix to assess performance and determine the most effective method for reliably identifying allergies.

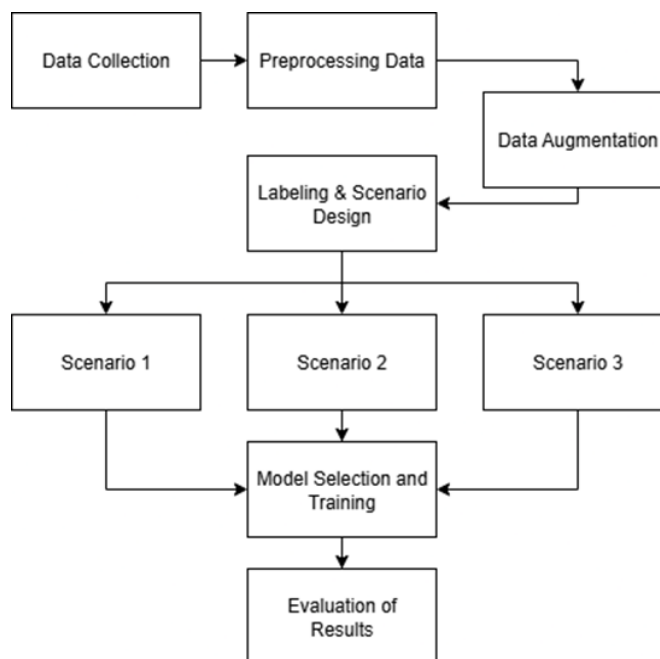


Figure 1. Flowchart of implementation steps.

2.1. Data Collection

Data collection was the initial phase of this study. It was conducted using a structured web-scraping procedure on the *Cookpad* food website. Recipes were retrieved using the keyword “*Masakan Nusantara*” (Indonesian cuisine) to compile a representative dataset of Indonesian culinary data. The extracted attributes included recipe title, ingredient list, cooking instructions, number of likes, and recipe URL. A total of 9,986 recipes were collected and stored in Comma-Separated Values (CSV) format for further processing. To provide a clear overview of the dataset preparation process, Table 1 summarizes the number of recipes at each stage, from the initial web scraping to filtering, data augmentation, and resampling. This summary ensures transparency in data handling and supports the reproducibility of the proposed allergen classification framework.

Table 1. Summary of dataset preparation stages and number of recipes.

Dataset Stage	Description	Number of Recipes
Raw Scraped Data	Recipes collected via web scraping from <i>Cookpad</i> during the data collection period	9,986
Duplicate Removal	Duplicate recipes removed based on identical titles and ingredient lists	65
Filtered Dataset (Raw Dataset)	Recipes after excluding beverages and non-cuisine entries	9,921
Augmented Dataset	Synthetic recipes generated through optional-ingredient augmentation	15,030
Raw Dataset + ROS	Raw Dataset and random oversampling	16,528
Augmented Dataset + ROS	Dataset after augmentation and random oversampling	25,574

As shown in Table 1, the initial scraping process yielded 9,986 recipe entries. A duplicate detection procedure was then applied by comparing combinations of recipe titles, ingredient lists, and URLs to ensure data uniqueness. This verification step confirmed that all retrieved entries were unique at the time of data

collection. The dataset initially included beverage recipes, which were excluded because beverages do not fall under the operational definition of cuisine in this study [19]. After this filtering step, the processed dataset consisted of 9,921 unique Indonesian culinary recipes.

The scraping process was conducted between August and September 2025 to maintain temporal consistency of the collected data. During preprocessing, several forms of textual noise were identified, including typographical variations (e.g., “*telor*” vs. “*telur*”), inconsistent capitalization, special characters, redundant punctuation, informal abbreviations (e.g., “*sdm*”), and optional ingredient indicators such as “*opsional*,” “*atau*,” or “*/*”. These inconsistencies can hinder accurate allergen detection if not addressed properly. Noise handling involved text normalization, lowercasing, removal of non-informative symbols, and standardization of ingredient terminology to ensure reliable ingredient matching and allergen mapping. This systematic cleaning process ensured that the final dataset consisted exclusively of validated and unique Indonesian culinary recipes, making it suitable for subsequent allergen labeling and classification modeling.

2.2. Preprocessing Data

After collecting the data, feature engineering was performed by adding 14 new attributes, each representing a specific type of allergen according to European Union standards, including gluten cereals, shrimp, eggs, fish, peanuts, soybeans, milk, nuts, celery, mustard, sesame seeds, sulfur dioxide & sulfites, lupine, and mollusks [17]. The raw data is presented in Table 2.

Table 2. Raw data with 14 allergen attributes.

Titles	Ingredients	Allergens	Value
<i>Gulai Tumpang (Masakan Minang)</i>	1/2 kg beef tendon, 1/4 kg bamboo shoots, 1.5 liters of coconut milk, 4 tea spoon goat oil, powdered spices, 60 grams of chili, 5 cloves of garlic, 8 cloves of shallots, 1 cm of ginger, 1 cm of galangal, 1 cm of turmeric, aromatic ingredients, 1 stalk of lemongrass, 1 turmeric leaf, 3 lime leaves, 1 bay leaf	Gluten cereals	0
		shrimp	0
		eggs	0
		fish	0
		peanuts	0
		Soybeans	0
		milk	0
		nuts	0
		celery	0
		mustard	0
		Sesame seeds	0
		Sulfur dioxide & sulfites	0
		Lupine	0
		Mollusks	0
<i>Gulai Pisang Telur (Masakan Jadul Yang Endul)</i>	1 comb of banana (<i>tanduk</i> variety) cut diagonally, 5 boiled eggs, 1 packet of Kara coconut milk, gulai seasoning bought at the market, 6 red chilies cut halved, to taste fried shrimp, to taste water, to taste salt, to taste sugar, to taste chicken stock powder, to taste pepper powder, to taste cooking oil	Gluten cereals	1
		shrimp	1
		eggs	1
		fish	0
		peanuts	0
		Soybeans	1
		milk	1
		nuts	0
		celery	0
		mustard	0
		Sesame seeds	0
		Sulfur dioxide & sulfites	0
		Lupine	0
		Mollusks	0

As shown in Table 2, binary values (1 for present and 0 for absent) was assigned to these attributes based on ingredient analysis for each recipe. Each recipe was mapped to the 14 European Union allergen attributes through two stages: automated processing and manual review. In the automated stage, a script scanned the ingredient list and matched terms against the predefined allergen dictionary. The dictionary covered core allergens and common derivatives. For example, soy sauce is included on the list because it is derived from soybeans, a major allergen. For processed or instant products (powdered broth), allergens were identified by reviewing ingredient lists of representative commercial products. The manual review stage corrected issues that the script could not resolve, such as typos, abbreviations, and alternative ingredient names.

2.3. Data Augmentation

Data augmentation aims to increase data quantity and diversity while mitigating class imbalance caused by the limited size of the raw dataset [20], [21]. This study applies a rule-based approach to generate synthetic recipe variations by identifying non-mandatory ingredients marked by keywords such as “optional,” “atau,” and “skip.” When optional ingredients are removed, allergenic components may change, potentially altering the final label across the 14 allergen categories. The number of optional keywords in each recipe determines the number of augmented samples generated, as shown in Table 3.

Table 3. Calculation of augmentation data.

Food Ingredients	Keywords	Status	Number	Augmentation Count
<i>tomat, cabe merah, cabe kecil (optional), kemiri / jinten</i>	<i>optional, /(atau)</i>	original	2	$2^2 = 4$
<i>tomat, cabe merah, cabe kecil, kemiri</i>	-	augmented 1	-	-
<i>tomat, cabe merah, kemiri</i>	-	augmented 2	-	-
<i>tomat, cabe merah, jinten</i>	-	augmented 3	-	-
<i>tomat, cabe merah, cabe kecil, jinten</i>	-	augmented 4	-	-

To evaluate whether augmentation introduced noise or altered the natural data distribution, the original and augmented datasets were compared using distributional analysis. Class proportions and allergen-feature frequencies were assessed using the Kullback–Leibler (KL) divergence [22]. The KL divergence values were 0.001207 for class distributions and 0.002065 for allergen feature frequencies, both well below the accepted threshold ($KL < 0.1$), indicating that augmentation preserves the original data distribution. This statistical approach is consistent with established validation frameworks that utilize information-theoretic metrics to ensure the fidelity of augmented samples and mitigate the risk of model overfitting to synthetic noise [23]. These results confirm that the augmentation strategy maintains the integrity of the dataset while supporting the robustness of the subsequent machine learning analysis.

Figure 2 compares class proportions and allergen feature frequencies before and after augmentation. The augmented dataset closely follows the original distribution, with only minor proportional changes. Augmentation improves class balance and model robustness in moderate-class scenarios [24]. However, in highly granular multi-class settings (Scenario 3), small ingredient variations can disproportionately affect labels, limiting performance gains. This issue is identified as a limitation and a direction for future improvement.

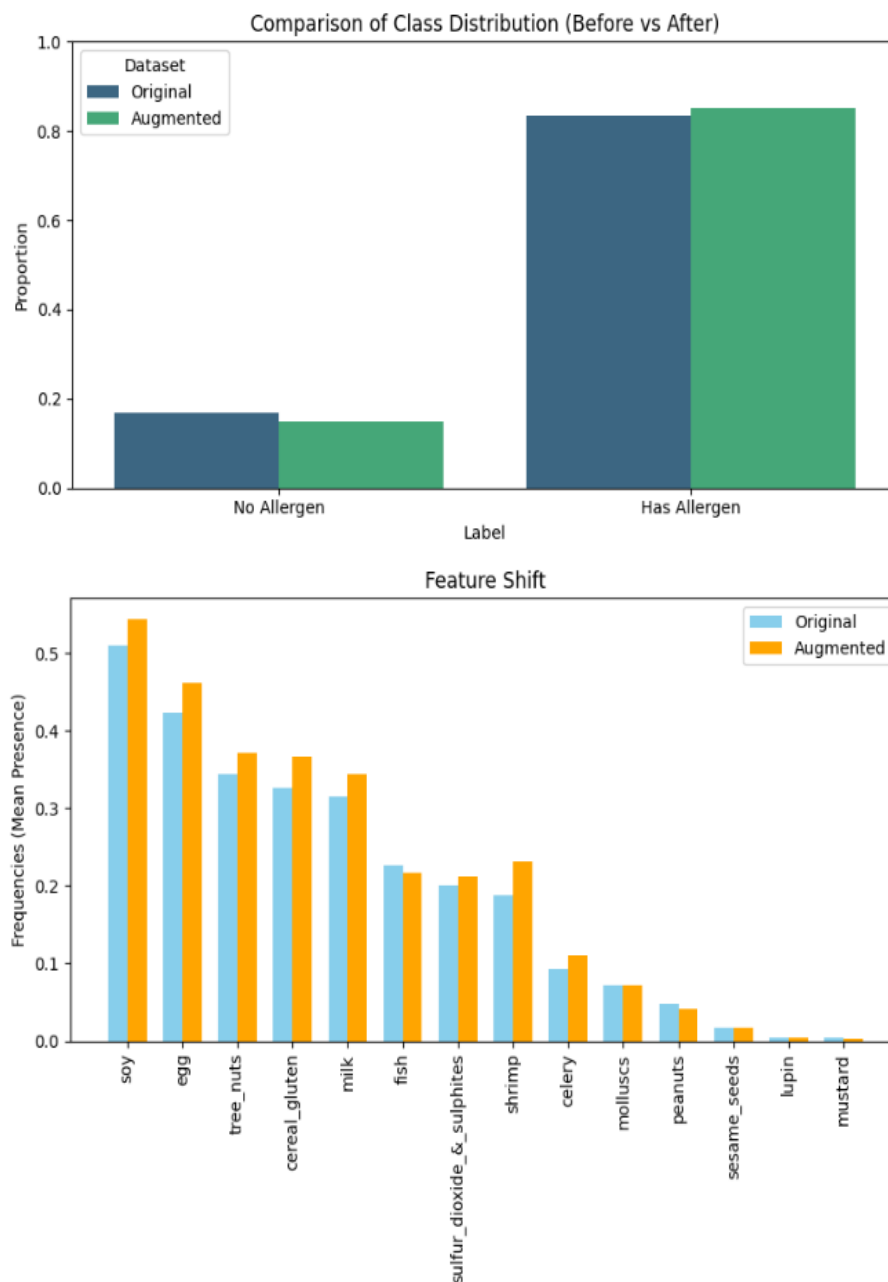


Figure 2. Comparison of class distribution and allergen feature frequencies before and after data augmentation.

2.4. Labeling and Scenario Design

This study defines label determination as a core step because it directly sets the target variables predicted by the classification model [25]. Three labeling scenarios was designed to capture different formulations of allergen exposure in Indonesian recipe data, as summarized in Table 3. This study defines label determination as a core step because it directly sets the target variables predicted by the classification model [25]. Three labeling scenarios was designed to capture different formulations of allergen exposure in Indonesian recipe data, as summarized in Table 4.

Table 4. Scenario design.

Scenario	Description
1	0 = does not contain allergens 1 = contains allergens
2	0 = does not contain allergens (0) 1 = low level (1–4 allergens) 2 = moderate level (5–9 allergens) 3 = high level (10–14 allergens)
3	The category class directly represents the number of allergens, ranging from 0 (does not contain allergens) to 14 (contains all types of allergens)

In Scenario 1, a binary classification was applied scheme that assigns recipes to either allergen-free (Class 0) or allergen-present (Class 1) categories. This approach enables direct detection of allergen presence. In Scenario 2, we model exposure severity by grouping recipes into four classes based on allergen counts. Recipes with 1–4 allergens were assigned to Class 1 (low exposure), 5–9 allergens to Class 2 (moderate exposure), and 10–14 allergens to Class 3 (high exposure), while retaining Class 0 for allergen-free recipes. This design tests the model's ability to distinguish levels of allergen exposure. In Scenario 3, a fine-grained labeling scheme was used, that maps each recipe directly to the exact number of allergens present, resulting in a 15-class classification task (0–14). This setting evaluates detailed allergen discrimination. To mitigate class imbalance, three dataset variants were constructed: (1) raw data with Random Oversampling, (2) augmented data, and (3) augmented data combined with Random Oversampling. Random Oversampling balances class distributions by duplicating minority-class samples [26] and is widely used to improve performance on underrepresented classes [14].

The label consistency was continuously verified after resampling and augmentation to avoid label distortion. To improve label validity, all dataset (raw and augmentation) were reviewed by a single independent culinary expert with domain knowledge of Indonesian cuisine and food ingredients. The expert examined the mapped allergen labels to ensure that they were consistent with established culinary knowledge and common ingredient usage. In addition, the accuracy of the automated labeling process was evaluated by manually verifying a random sample of the generated outputs. This validation step ensured that allergen assignments were correct and aligned with domain knowledge, particularly for processed ingredients where product variation may influence allergen presence. Overall, this two-stage verification process helped ensure the reliability and consistency of the allergen labeling.

2.5. Model Selection and Training

This study employs the Bernoulli Naïve Bayes classifier due to its computational efficiency and strong performance on text-based or binary-feature data, making it well-suited for allergen representation using presence-absence indicators [27]. The model applies Bayes' theorem and assumes conditional independence among features [28]. For a given Class c and feature vector $x = (x_1, x_2, x_n)$, the posterior probability is computed as shown in equation (1).

$$P(X) = \frac{P(x) \prod_{i=1}^n P(x_i|c)}{P(x)} \quad (1)$$

In Bernoulli Naïve Bayes, each feature x_i is modeled as a binary variable, where a value of 1 indicates the presence of a word or attribute (e.g., an allergenic ingredient) in a recipe, and 0 denotes its absence [29]. This formulation enables the model to classify recipes based on allergen presence efficiently. This study uses Bernoulli Naïve Bayes without additional parameter tuning and relies on the library's default settings. By doing so, experimental was maintained simplicity and evaluate the model's baseline performance without hyperparameter optimization. In the 0–14 labeling scenario, which reflects the exact number of allergens per recipe, class imbalance was explicitly addressed. Certain classes contained only a

single sample, which does not satisfy the minimum requirement for inclusion in cross-validation procedures [30]. To ensure valid training and evaluation, a uniform data duplication strategy was applied: when at least one class contains only one instance, all samples were duplicated across classes. This strategy preserves cross-validation consistency and ensures adequate class representation.

2.6. Evaluation and Analysis of Results

Model performance was evaluated using a Confusion Matrix with 10-fold cross-validation ($k=10$). 10-fold cross-validation divides the dataset into 10 parts: 9 for training and 1 for testing, and the process is repeated 10 times [31]. This ensures that the evaluation results are more stable and not biased towards a particular data division [32]. The confusion matrix is a table that summarizes the model's predictions [33]. Using this evaluation framework allows a comprehensive assessment of the model's classification performance across multiple validation iterations. For Scenario 1, the confusion matrix has four main components, 1) True Positive (TP): Positive data that is predicted correctly. 2) True Negative (TN): Negative data that is predicted correctly. 3) False Positive (FP): Negative data incorrectly predicted as positive. 4) False Negative (FN): Positive data that is incorrectly predicted as negative. From these components, four main evaluation matrices are calculated, which are found in equations (2), (3), (4), and (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 - Score = \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Next, after presenting the evaluation metrics including accuracy, precision, recall, and F1-score, a deeper analysis was conducted using the three visualization outputs shown above: the confusion matrix, ROC curve, and F1-score per class with macro and weighted averages. These visual analyses were generated using the out-of-fold evaluation approach. This method allows the model's predictions to be assessed on validation folds that were not used during training, providing a more reliable interpretation of the model's classification behavior across different scenarios. Through this analysis, this research contributes to better understand the performance characteristics of the Bernoulli Naïve Bayes model, particularly its ability to distinguish classes and maintain balanced performance across both majority and minority categories.

3. RESULTS AND DISCUSSION

The following research results describe the classification model's performance in identifying allergens in Indonesian cuisine across the four available data types. The results are presented systematically, starting with Scenarios 1, 2, and 3 in Table 5, Table 6, and Table 7. The classification performance of each scenario is further examined through confusion matrices, ROC curves, and class-level F1-scores to provide deeper insight into prediction behavior and class discrimination. The comparison with previous studies, presented in Table 8, then provides further context for evaluating the performance of the proposed approach. This is followed by a discussion of the methodological and practical implications of the classification scenarios, including classification complexity, class imbalance, augmentation, and Random Oversampling (ROS), as summarized in Table 9 and Table 10. Finally, the potential applications of the proposed allergen classification system and directions for future research are discussed to assess its broader applicability in food safety and culinary information systems.

Table 5. Evaluation results for scenario 1.

Data	Accuracy	Precision	Recall	F1 Score
Raw	1.000	1.000	1.000	1.000
Augmented	1.000	1.000	1.000	1.000
Raw + ROS	1.000	1.000	1.000	1.000
Augmented + ROS	1.000	1.000	1.000	1.000

3.1. Scenario 1: Binary Allergen Classification

The evaluation results in Table 5 show that the model achieved excellent performance across all test scenarios, with accuracy, precision, recall, and F1 score each reaching 1.0. Collectively, these excellent scores indicate that the classifier made no prediction errors in the evaluation dataset. This result implies that allergenic and non-allergenic recipes are fully separable under the applied feature representation and modeling approach. Several factors can influence excellent performance. Among these are dataset characteristics such as strong class separation, limited variability, or low noise levels. Therefore, additional analysis is needed to assess the model's robustness and generalization.

To provide deeper insight, the next analysis examined the dataset configuration using the F1 score as the primary comparison measure. This evaluation was complemented by visual diagnostics, including a confusion matrix, ROC curve, and F1 scores per class (macro and weighted average), to understand the prediction behavior and class balance better. The first scenario evaluates the binary classification task using the raw dataset as the representative dataset, since all data variations produced identical performance values of 1.0 across all evaluation metrics. Figure 3 presents the confusion matrix obtained from 10-fold cross-validation. Figure 4 presents the ROC curve, and Figure 5 shows the F1-score per class, along with the macro and weighted averages.

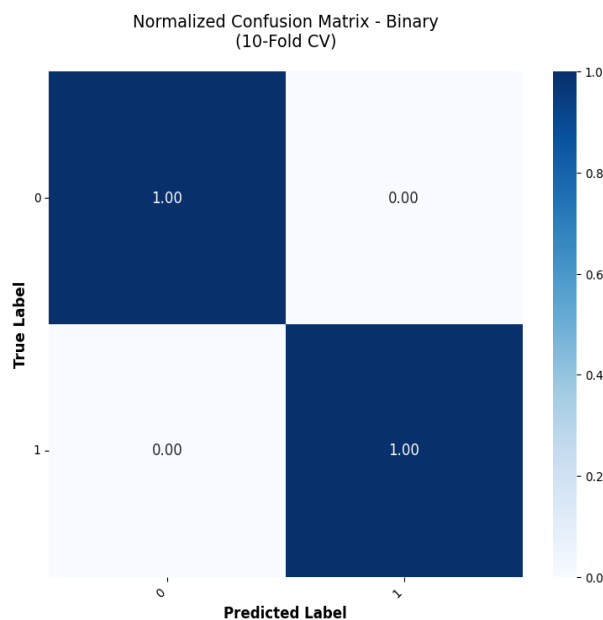


Figure 3. Confusion matrix for scenario 1 (raw dataset).

The confusion matrix demonstrates perfect classification performance. All samples belonging to Class 0 (non-allergen) were correctly predicted as Class 0, and all samples belonging to Class 1 (allergen) were correctly predicted as Class 1. The diagonal values of 1.00 indicate maximum True Positive and True Negative rates, while the off-diagonal values of 0.00 confirm the absence of False Positives and False Negatives. Although such perfect results may raise concerns about extreme overfitting, in this case, the outcome is better explained by the simplicity of the binary task. The distinction between the presence and

absence of allergens is highly explicit in the feature representation, making the two classes clearly separable. Furthermore, the use of 10-fold cross-validation reduces the likelihood that the model merely memorized specific samples.

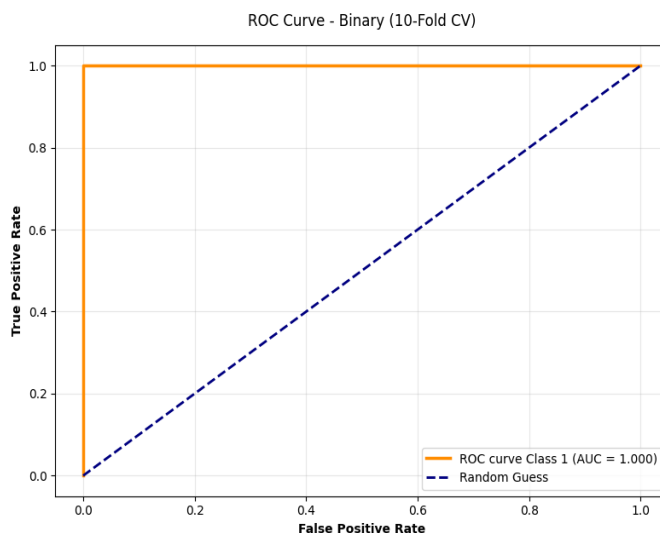


Figure 4. ROC curve for scenario 1 (raw dataset).

The curve reaches the top-left corner of the ROC space, resulting in an Area Under the Curve (AUC) value of 1.000. This indicates perfect discriminative ability, with a True Positive Rate of 1.0 and a False Positive Rate of 0.0 across thresholds. Rather than signaling overfitting, this result indicates that the binary labeling scheme produces a strong, direct relationship between input features and class labels. In other words, the model is not exploiting noise but effectively capturing a deterministic pattern inherent in the dataset.

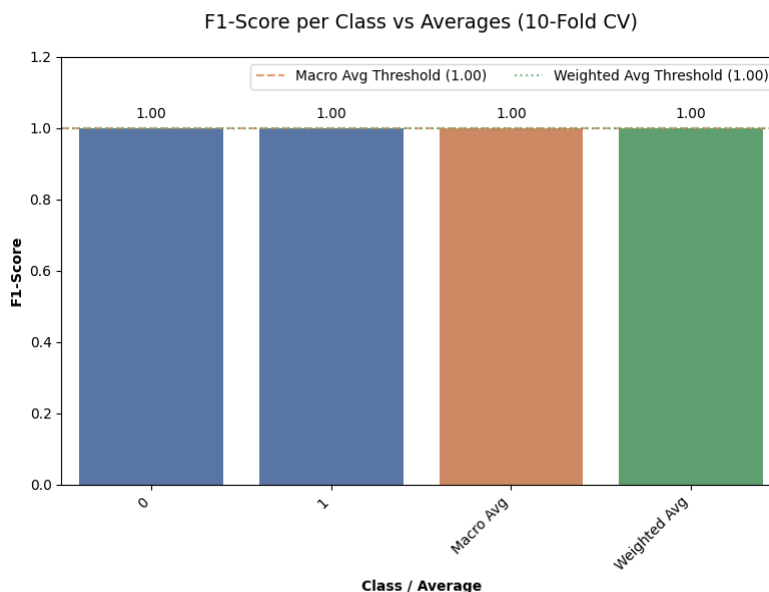


Figure 5. Macro vs weighted average for scenario 1 (raw data).

In Figure 5, both Class 0 and Class 1 achieve an F1-score of 1.00, and consequently, both the macro and weighted averages are also 1.00. This indicates a perfect balance between precision and recall across classes. The identical macro and weighted averages further confirm that performance consistency is

maintained regardless of class distribution. The uniformity of these results supports the interpretation that the binary classification boundary is inherently simple and well-defined. Although achieving a score of 1.0 across all evaluation metrics could suggest extreme overfitting, the evidence in Scenario 1 indicates otherwise. The binary task of distinguishing between the presence and absence of allergens is structurally straightforward and highly separable, given the engineered binary allergen features. Therefore, the perfect performance is more plausibly attributed to the inherent simplicity and determinism of the classification problem rather than to overfitting.

3.2. Scenario 2: Multi-Level Allergen Classification

Table 6 presents the results of allergen-level classification into four categories (Classes 0–3). Overall, the model demonstrated strong performance. The average accuracy, precision, recall, and F1-score exceeded 0.9 across all dataset configurations. Accuracy values ranged from 0.90 to 0.91, indicating high overall correctness [44]. A significant improvement was observed in the augmented configuration, which achieved the highest accuracy (0.9192). However, when ROS was applied, the result was slightly lower (0.9093). The highest precision value was achieved with the augmented configuration, reaching 0.9454. Similar to accuracy, precision decreased with the ROS application (0.9012). This pattern indicates that augmentation more effectively reduces the likelihood of false positives [45]. Recall showed the opposite trend from precision and accuracy. Precision reached its highest value with ROS, at 0.9356. This indicates improved minority class detection after resampling. The F1-score indicates that Augmented + ROS (0.9293) is the most balanced configuration. This result reflects the combined benefits of increased representation and class balancing.

Table 6. Evaluation results for scenario 2.

Data	Accuracy	Precision	Recall	F1 Score
Raw	0.9157	0.9364	0.8945	0.9060
Augmented	0.9192	0.9454	0.9107	0.9219
Raw + ROS	0.9093	0.9012	0.9356	0.9128
Augmented + ROS	0.9138	0.9338	0.9309	0.9293

Building upon the quantitative evaluation presented in Table 6, where the Augmented + ROS configuration achieves the highest F1-score by effectively balancing precision and recall, the subsequent analysis focuses on a more detailed examination of this best-performing setup. Accordingly, the second scenario evaluates the multi-class level-based classification task (four Classes: 0–3) using the Augmented + Random Over-Sampling (ROS) dataset, which demonstrates the most robust overall performance among all sub-scenarios. To further substantiate these numerical findings, the discussion is complemented with visual performance analyses, beginning with the normalized confusion matrix obtained from 10-fold cross-validation, as presented in Figure 6.

The confusion matrix shows strong overall classification performance, with high diagonal values across all classes. Class 0 is perfectly classified (1.00), indicating that recipes without allergens are clearly distinguishable from those with allergens. Class 1 achieves 0.88 correct predictions, with 0.12 misclassified as Class 2. Similarly, Class 2 achieves 0.93 correct predictions, with 0.07 misclassified as Class 1. Class 3 shows 0.91 correct predictions and 0.09 misclassifications as Class 2. These patterns indicate that most misclassifications occur between adjacent levels, particularly between Classes 1 and 2, and between Classes 2 and 3. This suggests that the model struggles slightly when distinguishing borderline allergen levels, where the numerical differences between classes are relatively small. However, the absence of large cross-level misclassifications confirms that the model maintains logical ordering consistency. To further evaluate the model's discriminative capability beyond class-wise accuracy, Figure 7 presents the ROC curves, providing insight into its separability performance across classes.

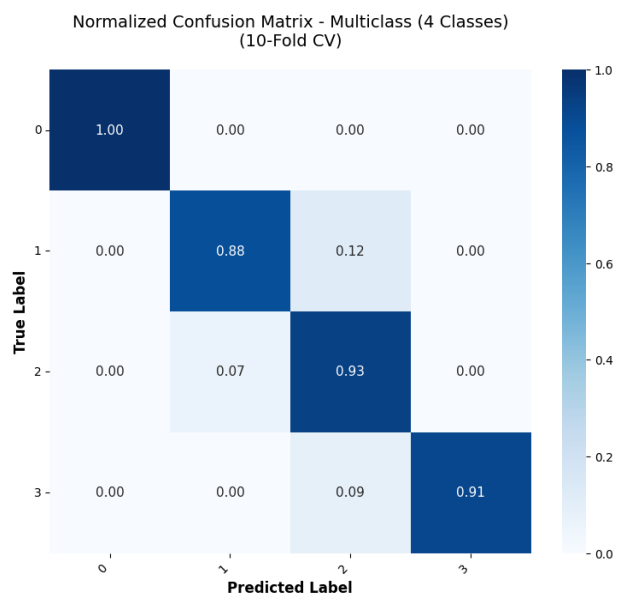


Figure 6. Confusion matrix for scenario 2 (augmented + ROS).

The AUC values are high across all classes: Class 0 (1.000), Class 1 (0.962), Class 2 (0.986), and Class 3 (1.000). These results indicate excellent discriminative capability. Classes 0 and 3 exhibit perfect separability, while Classes 1 and 2 show slightly lower but still strong AUC values. The lower AUC for Class 1 aligns with the confusion matrix findings, which showed minor overlap with Class 2. Overall, the ROC analysis confirms that the model maintains robust discrimination even under increased class complexity. To further complement this finding, Figure 8 presents the F1-score for each class along with the corresponding macro and weighted averages, providing a more detailed view of class-wise performance balance.

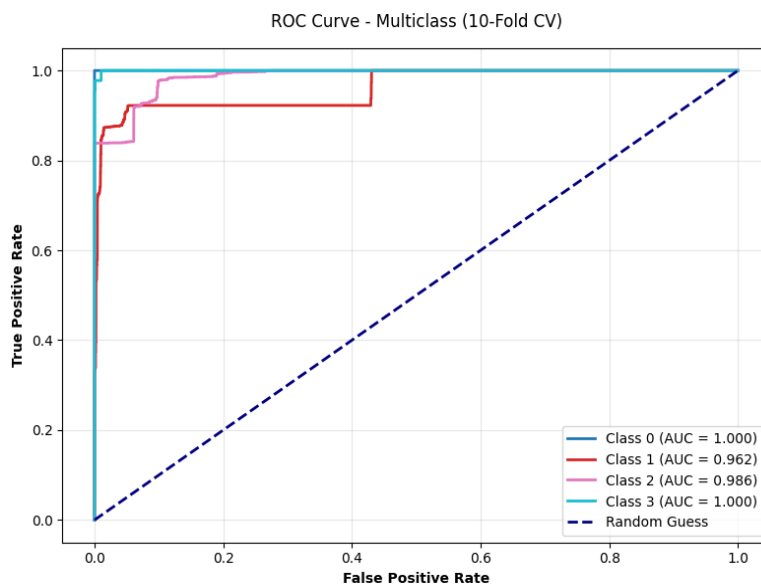


Figure 7. ROC curve for scenario 2 (augmented + ROS).

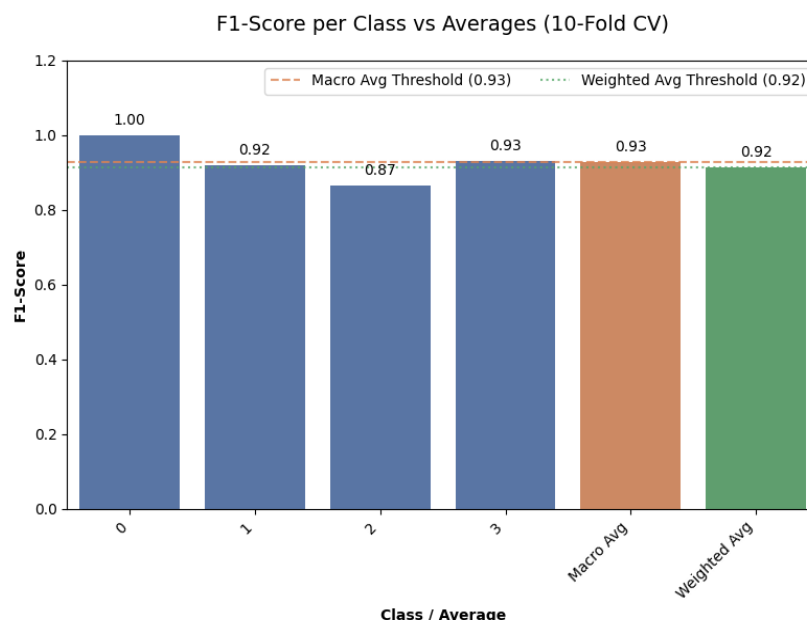


Figure 8. Macro vs weighted average for scenario 2 (augmented + ROS).

In Figure 8, Class 0 achieves an F1-score of 1.00, Class 1 scores 0.92, Class 2 scores 0.87, and Class 3 scores 0.93. The macro average is 0.93, while the weighted average is 0.92. The slightly lower F1-score for Class 2 reflects the bidirectional confusion with adjacent levels observed earlier. The proximity between macro and weighted averages indicates that class imbalance has been effectively mitigated by the Augmented + ROS strategy. This confirms that the combination of augmentation and resampling helps maintain balanced predictive performance across classes. Scenario 2 demonstrates high, stable performance at the multi-class level in allergen classification. While minor confusion occurs between adjacent allergen levels, overall classification remains strong, with macro and weighted F1-scores above 0.90. The Augmented + ROS dataset effectively balances class distribution and improves recall without significantly sacrificing precision. Compared to the binary scenario, this level-based task introduces moderate complexity. Yet, the model maintains reliable discriminative performance across all evaluation metrics.

3.3. Scenario 3: Fine-Grained Allergen Classification

Table 7 presents classification results for 14 allergen categories, ranging from 0 to 14, with overall model performance remaining relatively low. Across the four-evaluation metrics, namely accuracy, precision, recall, and F1-score, the average performance remains below 0.7. This result indicates that the model has substantial difficulty with highly granular multi-class classification. In terms of accuracy, the variation across data strategies reflects the model's sensitivity to preprocessing techniques. The raw dataset yields an accuracy of 0.6174, which indicates a high number of combined false positives and false negatives. The Raw + ROS strategy increases accuracy to 0.6542 and shows that class balancing reduces total misclassification errors. However, the combination of augmentation with ROS does not yield further improvement, suggesting limited interaction between the two methods. Precision follows a similar pattern. The raw dataset yields a precision of 0.4621, indicating that many positive predictions are incorrect due to a high false-positive rate [46]. After applying ROS, the precision increases to 0.5666. This improvement shows that balanced class representation helps the model construct clearer decision boundaries and reduce incorrect positive predictions. Recall also improves after applying ROS. The raw dataset yields a recall of 0.4679, which indicates that the model fails to detect many actual positive cases. This limitation results from a high number of false negatives.

Table 7. Evaluation results for scenario 3.

Data	Accuracy	Precision	Recall	F1 Score
Raw	0.6174	0.4621	0.4679	0.4405
Augmented	0.6356	0.4826	0.4882	0.4560
Raw + ROS	0.6542	0.5666	0.6083	0.5677
Augmented + ROS	0.6357	0.5449	0.6020	0.5474

With ROS, recall increased to 0.6083, which demonstrates improved sensitivity toward minority classes. The F1-score confirms the same trend. The raw dataset achieves the lowest F1 Score of 0.4405, indicating weak precision and recall. The Raw + ROS strategy achieves the highest F1-score of 0.5677, indicating a more balanced reduction in false positives and false negatives. Overall, ROS provides the most consistent improvement in Scenario 3. However, performance remains constrained by severe class imbalance and high label granularity. Building on the quantitative findings summarized in Table 6, which show that the Raw + ROS configuration achieves the highest overall F1-score by offering the most effective balance between precision and recall, the subsequent analysis focuses on a deeper examination of this best-performing setup. Accordingly, the third scenario evaluates fine-grained multi-class classification based on the exact number of allergens (15 classes: 0–14) using the Raw + Random Over-Sampling (ROS) dataset. To further elucidate the factors underlying the observed metric trends, the discussion is complemented with visual performance analyses, beginning with the normalized confusion matrix obtained from 10-fold cross-validation, as shown in Figure 9. Followed by Figure 10, which presents the ROC curve, and Figure 11 presents the F1-score per class along with the macro and weighted averages.

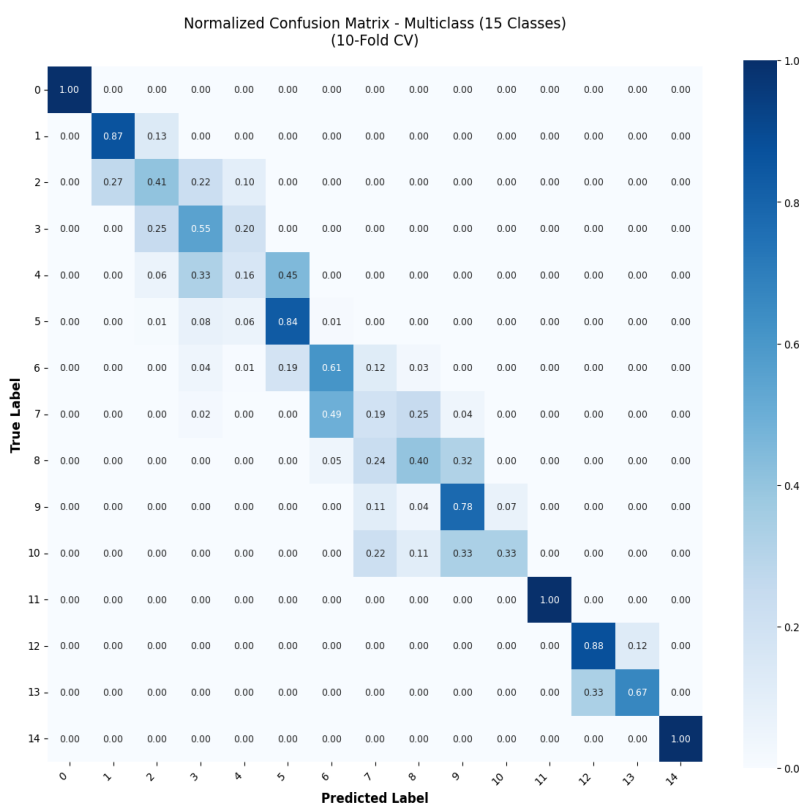


Figure 9. Confusion matrix for scenario 3 (raw + ROS).

The confusion matrix reveals that classification difficulty increases substantially compared to the previous scenarios. Perfect classification is observed for Classes 0, 11, and 14 (1.00), indicating that extreme categories are clearly distinguishable. However, substantial confusion appears among intermediate

classes, particularly between neighboring allergen counts. For example, Class 2 is frequently misclassified as Classes 1 or 3, Class 4 overlaps with Classes 3 and 5, and Class 7 shows dispersion toward adjacent levels. This pattern indicates that most errors occur between numerically adjacent classes, reflecting the intrinsic similarity of recipes with close allergen counts. The diagonal dominance remains visible but weakens for mid-range classes, confirming that fine-grained discrimination is considerably more complex.

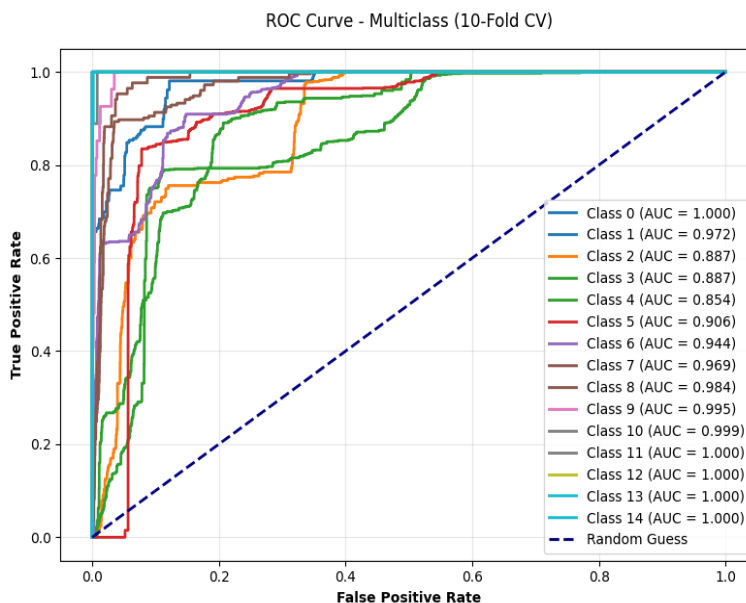


Figure 10. ROC curve for scenario 3 (raw + ROS).

Despite the confusion observed in the matrix, the AUC values remain relatively high across many classes. Several classes achieve near-perfect discrimination, such as Class 0 (1.000), Class 11 (1.000), Class 12 (1.000), Class 13 (1.000), and Class 14 (1.000). Other classes maintain strong AUC values above 0.85, although some mid-range classes show comparatively lower separability. The ROC curves indicate that, from a ranking perspective, the model remains capable of effectively distinguishing classes. However, threshold-based discrete predictions become more challenging when boundaries between adjacent counts are subtle.

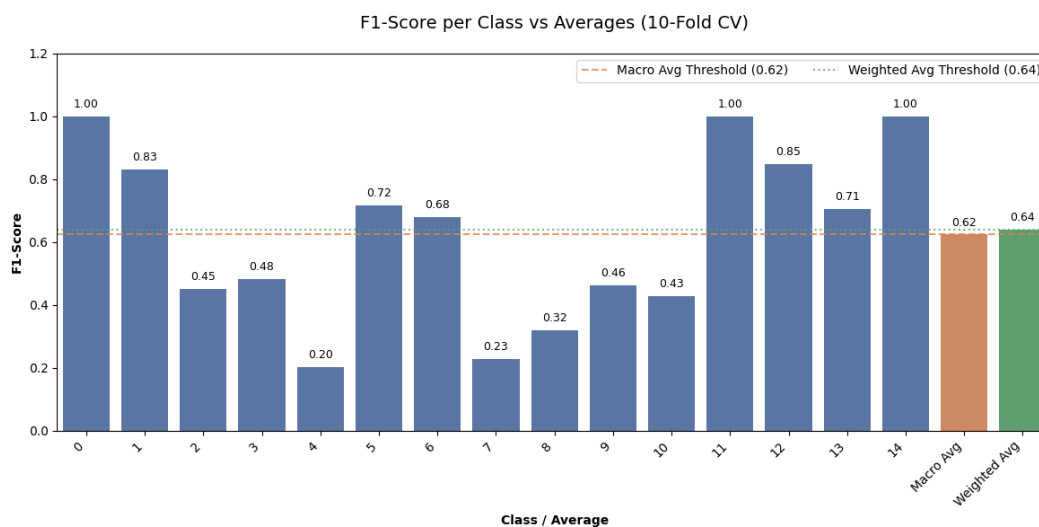


Figure 11. Macro vs weighted average for scenario 3 (raw + ROS).

Figure 11 shows that the macro average is 0.62, while the weighted average is 0.64, indicating moderate overall performance. Performance variability across classes is substantial: some classes achieve high F1-scores (Classes 0, 11, and 14 with 1.00), whereas others, particularly in the middle range (Classes 4 and 7), exhibit significantly lower F1-scores. The small gap between macro and weighted averages suggests that ROS effectively mitigated severe imbalance. However, the inherent difficulty of distinguishing 15 closely related categories limits overall performance. The lower macro average reflects instability across minority and overlapping classes. Scenario 3 demonstrates that increasing the number of classes dramatically raises classification complexity. While Raw + ROS improves class balance and maintains reasonable discriminative power, fine-grained allergen-count prediction remains challenging due to high inter-class similarity among adjacent categories. The model maintains strong ranking performance, as indicated by ROC analysis, yet discrete multi-class accuracy declines, resulting in moderate macro and weighted F1 Scores. This scenario highlights the trade-off between classification granularity and predictive stability.

3.4. Comparison with Previous Studies

To further contextualize the performance of the proposed model, the results of this study were compared with those of several previous studies that applied similar algorithms to comparable classification tasks. The comparison focuses specifically on Scenario 1 (allergen vs. non-allergen) because this scenario is directly comparable to binary classification settings reported in related work. The comparison results are presented in Table 8. As shown in Table 8, the proposed approach achieves the highest performance across all reported metrics. Previous studies reported accuracy values ranging from 0.6612 to 0.9500. In contrast, another study achieved an accuracy of 0.8720 with additional evaluation metrics including precision of 0.8213, recall of 1.0000, and an F1-score of 0.9019. In contrast, the proposed method in this study achieved perfect performance, with accuracy, precision, recall, and F1-score all equal to 1.0000 in the binary allergen classification scenario. These results suggest that the proposed data quality improvement strategy, including refined preprocessing and expert-validated allergen mapping, may yield clearer feature-label relationships and improved classification performance compared to previous work.

Table 8. Performance comparison with previous works.

Previous work	Accuracy	Precision	Recall	F1-Score
Shaukat et al., 2024 [29]	0.6612	-	-	-
Kina et al., 2025 [34]	0.9500	-	-	-
Farooq et al., 2025 [35]	0.6613	-	-	-
Li et al., 2025 [36]	0.8720	0.8213	1.0000	0.9019
This research	1.0000	1.0000	1.0000	1.0000

3.5. Methodological and Practical Implications

Overall, this analysis not only compares the effectiveness of each scenario from a technical perspective but also reviews the methodological implications and classification challenges arising from the increase in the number of classes. A qualitative comparison of allergen classification scenarios is shown in Table 9 and Table 10. Table 9 indicates that increasing classification complexity corresponds with reduced model performance. In Scenario 1 (0 and 1), the Bernoulli Naïve Bayes model achieved perfect scores across all evaluation metrics, reflecting clear class separation and highly distinguishable allergen patterns. The binary label structure minimizes ambiguity and simplifies the learning process, allowing the model to perform optimally without additional tuning. These results align with established theories regarding the efficiency of Naïve Bayes in handling independent binary feature spaces [37]. Consequently, binary allergen classification proves to be methodologically robust and suitable for early-stage detection systems that require only the identification of allergen presence or absence.

Scenario 2 (0–3) introduces greater complexity because the model must classify four allergen-level categories. This increase in label variability leads to a moderate decline in performance. However, the model maintains a strong F1-score of approximately 0.92, indicating balanced precision and recall. The

integration of data augmentation and Random Oversampling (ROS) effectively improves class representation and stabilizes performance across categories. This finding is consistent with prior research on imbalanced classification, which suggests that resampling techniques are vital for maintaining model reliability in multi-level tasks [38]. These results suggest that the model remains reliable for multi-level allergen risk assessment. However, performance sensitivity increases with task complexity.

Table 9. Qualitative comparison of methodological aspects between allergen classification scenarios.

Aspect	Scenario 1 (0 and 1)	Scenario 2 (0–3)	Scenario 3 (0–14)
Classification Purpose	To distinguish recipes that contain and do not contain allergens.	Group recipes by allergen level (low, medium, high).	Classify recipes based on the exact number of allergens (0-14 types).
Model Complexity	Low. Simple class structure and clear boundaries between classes.	Moderate. Differences between classes begin to overlap.	High. Many classes have small variations between categories.
Data Distribution	Relatively balanced because the two main classes are easily distinguishable.	Tends to be unbalanced between levels.	Very unbalanced, with some classes containing only a few data points.
Classification Difficulty Level	Very low. The model easily learns basic patterns.	Medium. Handling of minority data is required.	Very high. High-class overlap and increased noise.
Model Performance (F1 Score)	Perfect (1.00) across all data variations.	High (around 0.92 at best).	Low (highest 0.57 on Raw + ROS).
Effect of Augmentation	Not significant because the data is already balanced and clear.	Provides a moderate improvement in accuracy and precision.	Provides a small improvement, sometimes adding noise.
Effect of ROS (Resampling)	Not significantly affected.	Improves recall by suppressing the minority class.	Most effective at improving recall and F1 compared to augmentation alone.

Scenario 3 (0–14) presents the highest classification complexity and the most substantial performance decline, with a maximum F1-score of 0.57 under the Raw + ROS configuration. The expanded class space exacerbates data imbalance and increases inter-class overlap, limiting the model's capacity to learn discriminative patterns for minority categories. These results indicate that Bernoulli Naïve Bayes has limited scalability when dealing with highly granular label distributions. Nevertheless, this scenario remains analytically important because it enables fine-grained quantification of allergens. Performance improvements may require more expressive models or advanced resampling techniques, such as SMOTE (synthetic minority over-sampling technique) or ADASYN (adaptive synthetic sampling), to better address severe imbalance and class heterogeneity.

In terms of the classification process, the first scenario requires the least computational time. It has the highest interpretability because the results are binary and easy to understand. The second scenario requires additional processing to handle minority classes, but remains efficient for practical application. Meanwhile, the third scenario demands greater computing power and a more complex pipeline design due to the large number of classes. In the context of implementation, the more classes used, the greater the risk of classification errors and the need for strict validation strategies to ensure reliable results.

Table 10. Qualitative comparison of practical and implementation aspects between allergen classification scenarios.

Aspect	Scenario 1 (0 and 1)	Scenario 2 (0–3)	Scenario 3 (0–14)
Computational Requirements	Low, fast process.	Moderate, requires longer processing.	High school requires more resources due to the large number of classes.
Interpretability of Results	Very easy to understand (yes/no).	Fairly clear, representing the level of risk.	Difficult to interpret due to too many numerical categories.
Main Constraints	Potential overfitting due to overly distinct class differences.	Difficulty maintaining balance between levels.	Extreme imbalance and ambiguity between classes.
Accuracy on New Data	High and stable.	Relatively good if the data distribution is balanced.	Unstable, prone to decline if the data domain changes.
Potential for Practical Implementation	Very high, easy to implement for basic automatic detection systems (presence/absence of allergens).	Fairly high, suitable for risk warning systems based on allergen levels.	Limited, more suitable for research analysis or in-depth risk mapping.
Key Advantages	Most stable, fast, and accurate.	Provides more detailed risk information.	Provides granular analysis of allergen distribution.
Key Disadvantages	The information is too simple and does not show risk levels.	There is still overlap between classes and data imbalance.	High complexity, prone to classification errors.
Improvement Strategy	Test on external datasets to ensure generalization.	Use advanced balancing methods such as SMOTE or ADASYN.	Apply models with more complex architectures (Random Forest, ANN (artificial neural network), or BERT (bidirectional encoder representations from transformers)).

Table 10 highlights distinct functional implications across the three scenarios. Scenario 1 supports rapid binary screening and is suitable for consumer-level allergen filtering due to its simplicity and high reliability. Scenario 2 enables graded risk assessment across multiple allergen levels, making it more appropriate for recommendation systems that communicate varying exposure risks. Scenario 3 allows fine-grained allergen quantification, which is relevant for research analysis and industrial risk evaluation. However, it requires greater computational capacity and careful validation. Thus, scenario selection should align with the intended application and available computational resources.

Overall, the comparison results show that increasing the number of classes increases classification complexity, reduces model stability, and reinforces the need for appropriate preprocessing strategies. Augmentation and ROS have been shown to help maintain data balance, but their effectiveness decreases with overly granular label structures. For further research, it is recommended that data handling methods be improved using representation learning-based synthetic sampling techniques and advanced oversampling approaches that better preserve minority class distributions [39] and cross-domain evaluation so that models are not only accurate in one dataset but also able to adapt to recipe variations and ingredient contexts from various sources.

3.6. Application Potential and Future Research Directions

The implementation of machine learning-based allergen classification systems holds considerable potential for application across sectors, from the general public to food businesses and health agencies [15]. Within a global community, this approach can serve as a vital tool to enhance awareness and ensure food safety. The incorporation of these tools into online recipe applications or supermarket shopping platforms enables users to quickly identify allergenic items that may raise concerns [40]. This phenomenon is particularly important in the digital era, as many individuals look for and share recipes on internet platforms without following strict requirements for ingredient labeling [41]. This automated approach enables users from various geographic regions to obtain reliable, precise information on allergen content, thereby markedly reducing the risk of allergic reactions.

In the food industry, specifically in restaurant, catering, and large-scale food production sectors, the implementation of this system can be an invaluable tool for the early detection of potential problems. This method facilitates the anticipatory identification of issues before product release, hence guaranteeing the quality and safety of the final product. Manufacturers may use allergen categorization models during production. This would enable them to validate the safety of substances and formulations before commercialization. This not only improves compliance with international requirements, such as the EU Food Allergen Labeling Regulation, but also strengthens customer trust in trademarks [42]. The method can serve as a guide for micro-businesses to modify recipes, ensuring they are suitable for consumers with specific sensitivities and eliminating the need for manual verification. This invention creates new business opportunities for enterprises, including the creation of digitally verified “allergen-free” products.

The creation of an allergy classification system in healthcare can provide strategic benefits to medical professionals and nutritionists [43]. This technology can enhance allergy diagnosis by utilizing patient consumption patterns. The examination of past dietary records enables healthcare practitioners to swiftly and accurately identify allergenic substances. Moreover, hospitals and nutrition centers can benefit from this system, as it facilitates the creation of secure meal plans for patients with complex allergies, including multiple allergies or specific food intolerances. The application’s capabilities can be enhanced to promote public education on allergy management via digital health platforms. From a public policy perspective, implementing this technology can support food regulatory and government agencies in improving the national food ingredient labeling system. By adopting an automated classification system, data from various online recipe sources or commercial products can be analyzed to monitor trends in the use of allergenic ingredients in the community. This information can serve as a basis for policy-making in public health risk control and for improving the quality of nutrition services. Thus, this research not only contributes academically but also has great potential in shaping a safer, more transparent, and sustainable food ecosystem at the global level.

4. CONCLUSION

Scientifically, this research contributes a structured multi-scenario classification framework for food allergen detection, provides empirical evaluation of augmentation and oversampling strategies in multi-class imbalanced culinary datasets, and establishes methodological insights into the strengths and limitations of Bernoulli Naïve Bayes for binary allergen feature representation validated by domain experts. From a policy and practical perspective, the findings highlight the strong potential of automated allergen classification systems to support digital food safety governance. The binary classification model offers immediate applicability for large-scale monitoring and early-detection systems on online recipe platforms and food databases, while the risk-level model can inform consumer-facing warning systems and help MSMEs, restaurants, and food producers improve allergen transparency. At a broader level, integrating machine learning-based allergen detection into regulatory frameworks can strengthen evidence-based policymaking, enhance compliance with allergen labeling standards, and improve national food safety surveillance. Therefore, this research not only advances academic discourse in food safety informatics but also provides a scalable technological foundation for developing safer and more transparent culinary ecosystems in Indonesia.

AUTHOR CONTRIBUTION

Aji Prasetya Wibawa: Supervision, concept qualification, writing-review and editing, methodology. **David Satria Alamsyah:** Data Collection, writing-original draft, resources. **Aldy Rahmat Yulianto:** Formal Analysis, writing original draft, data collection. **Ahmad ‘Ammar Musyaffa:** Data Collection, writing-original draft, visualization. **Adil Zakaria:** Data Collection, writing the original draft, and investigation. **Agung Bella Putra Utama:** Supervision, writing review, and editing.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGMENT

The authors wish to express their heartfelt appreciation to culinary expert Dr. Mazarina Devi for her invaluable time and expertise in confirming the data used for allergy categorization in this study. Her profound understanding of ingredients and allergenic properties provided critical insights during the validation process. The helpful criticism she provided significantly improved the dependability of the categorization results. Her expert supervision guaranteed that the analysis was precise and scientifically valid. The authors sincerely value her ongoing assistance and contributions to enhancing the overall quality of this research.

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, 2024, <https://doi.org/10.1007/s11882-023-01120-y>.
- [2] N. Roy, “Food allergy knowledge, attitudes, and practices among restaurant’s staff in Bangladesh: A structural modeling approach,” *Food Sci. Nutr.*, vol. 13, no. 8, p. e70754, 2025, <https://doi.org/10.1002/fsn3.70754>.
- [3] A. A. Takrouni, “Knowledge gaps in food allergy among the general public in Jeddah, Saudi Arabia: Insights based on the Chicago food allergy research survey,” *Front. Allergy*, vol. 3, p. 1002694, 2022, <https://doi.org/10.3389/falgy.2022.1002694>.
- [4] F. Chang, L. Eng, and C. Chang, “Food allergy labeling laws: International guidelines for residents and travelers,” *Clin. Rev. Allergy Immunol.*, vol. 65, no. 2, pp. 148–165, 2023, <https://doi.org/10.1007/s12016-023-08960-6>.
- [5] G. N. Konstantinou, O. Pampoukidou, D. Sergelidis, and M. Fotoulaki, “Managing food allergies in dining establishments: Challenges and innovative solutions,” *Nutrients*, vol. 17, no. 10, p. 1737, 2025, <https://doi.org/10.3390/nu17101737>.
- [6] S. La Vieille, J. O. Hourihane, and J. L. Baumert, “Precautionary allergen labeling: What advice is available for health care professionals, allergists, and allergic consumers?,” *J. Allergy Clin. Immunol. Pract.*, vol. 11, no. 4, pp. 977–985, 2023, <https://doi.org/10.1016/j.jaip.2022.12.042>.
- [7] S. Xiong, W. Tian, H. Si, G. Zhang, and L. Shi, “A survey of the applications of text mining for the food domain,” *Algorithms*, vol. 17, no. 5, p. 176, 2024, <https://doi.org/10.3390/a17050176>.
- [8] W. Liu, “Quantitative food allergen risk assessment: Evolving concepts, modern approaches, and industry implications,” *Compr. Rev. Food Sci. Food Saf.*, vol. 24, no. 2, p. e70132, 2025, <https://doi.org/10.1111/1541-4337.70132>.
- [9] D. D. Purwanto, A. P. Wibawa, and S. Patmanthara, “Philosophy of culinary arts: Reading truth through language and taste,” *Bull. Culin. Art Hosp.*, vol. 5, no. 2, 2025, <https://doi.org/10.17977/um069v5i22025p62-72>.
- [10] D. Dooley, “Food process ontology requirements,” *Semant. Web*, vol. 15, no. 4, pp. 1133–1164, 2024, <https://doi.org/10.3233/SW-223096>.
- [11] A. F. U. R. Khilji, “Multimodal recipe recommendation system using deep learning and rule-based approach,” *SN Comput. Sci.*, vol. 4, no. 4, p. 421, 2023, <https://doi.org/10.1007/s42979-023-01870-6>.

- [12] H. Zhou, X. Chen, and X. Li, “Food allergenicity evaluation methods: Classification, principle, and applications,” *Foods*, vol. 14, no. 12, p. 2005, 2025, <https://doi.org/10.3390/foods14122005>.
- [13] F. K. K. Putra, M. K. Putra, and S. Novianti, “Taste of Asean: Traditional food images from Southeast Asian countries,” *J. Ethn. Foods*, vol. 10, no. 1, p. 20, 2023, <https://doi.org/10.1186/s42779-023-00189-0>.
- [14] Y. Zhang, “Deep learning in food category recognition,” *Inf. Fusion*, vol. 98, p. 101859, 2023, <https://doi.org/10.1016/j.inffus.2023.101859>.
- [15] A. Kumar and P. S. Rana, “A deep learning based ensemble approach for protein allergen classification,” *PeerJ Comput. Sci.*, vol. 9, p. e1622, 2023, <https://doi.org/10.7717/peerj-cs.1622>.
- [16] A. Roither, M. Kurz, and E. Sonnleitner, “The chef’s choice: System for allergen and style classification in recipes,” *Appl. Sci.*, vol. 12, no. 5, p. 2590, 2022, <https://doi.org/10.3390/app12052590>.
- [17] P. J. Turner, “Time to ACT-UP: Update on precautionary allergen labelling (PAL),” *World Allergy Organ. J.*, vol. 17, no. 10, p. 100972, 2024, <https://doi.org/10.1016/j.waojou.2024.100972>.
- [18] Y. Zhong, W. Zhou, and Z. Wang, “A survey of data augmentation in domain generalization,” *Neural Process. Lett.*, vol. 57, no. 2, p. 34, 2025, <https://doi.org/10.1007/s11063-025-11747-9>.
- [19] S. Jeong and J. Lee, “Effects of cultural background on consumer perception and acceptability of foods and drinks: A review of latest cross-cultural studies,” *Curr. Opin. Food Sci.*, vol. 42, pp. 248–256, 2021, <https://doi.org/10.1016/j.cofs.2021.07.004>.
- [20] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022, <https://doi.org/10.1016/j.gltp.2022.04.020>.
- [21] M. Lango and J. Stefanowski, “What makes multi-class imbalanced problems difficult? An experimental study,” *Expert Syst. Appl.*, vol. 199, p. 116962, 2022, <https://doi.org/10.1016/j.eswa.2022.116962>.
- [22] J. F. Kurian and M. Allali, “Detecting drifts in data streams using Kullback-Leibler (KL) divergence measure for data engineering applications,” *J. Data, Inf. Manag.*, vol. 6, no. 3, pp. 207–216, 2024, <https://doi.org/10.1007/s42488-024-00119-y>.
- [23] A. Figueira and B. Vaz, “Survey on synthetic data generation, evaluation methods and GANs,” *Mathematics*, vol. 10, no. 15, p. 2733, 2022, <https://doi.org/10.3390/math10152733>.
- [24] F. Gurcan and A. Soylu, “Synthetic boosted resampling using deep generative adversarial networks: A novel approach to improve cancer prediction from imbalanced datasets,” *Cancers (Basel)*, vol. 16, no. 23, p. 4046, 2024, <https://doi.org/10.3390/cancers16234046>.
- [25] E. Sarmas, E. Spiliotis, V. Marinakis, T. Koutselis, and H. Doukas, “A meta-learning classification model for supporting decisions on energy efficiency investments,” *Energy Build.*, vol. 258, p. 111836, 2022, <https://doi.org/10.1016/j.enbuild.2022.111836>.
- [26] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, “A comprehensive evaluation of oversampling techniques for enhancing text classification performance,” *Sci. Rep.*, vol. 15, no. 1, p. 21631, 2025, <https://doi.org/10.1038/s41598-025-05791-7>.
- [27] M. Scheschenja, M. B. Bastian, J. Wessendorf, A. D. Owczarek, A. M. König, S. Viniol, and A. H. Mahnken, “ChatGPT: Evaluating answers on contrast media related questions and finetuning by providing the model with the ESUR guideline on contrast agents,” *Curr. Probl. Diagn. Radiol.*, vol. 53, no. 4, pp. 488–493, 2024, <https://doi.org/10.1067/j.cpradiol.2024.04.005>.
- [28] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, “Variable selection for naïve bayes classification,” *Comput. Oper. Res.*, vol. 135, p. 105456, 2021, <https://doi.org/10.1016/j.cor.2021.105456>.
- [29] J. Li, A. Wu, L. Liu, A. Qu, C. Xu, H. Kuang, and L. Xu, “Analysis of food safety based on machine learning: A comprehensive review and future prospects,” *Food Chem.*, vol. 490, p. 145170, 2025, <https://doi.org/10.1016/j.foodchem.2025.145170>.
- [30] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model selection: A review with examples from ecology,” *Ecol. Monogr.*, vol. 93, no. 1, p. e1557, 2023,

- <https://doi.org/10.1002/ecm.1557>.
- [31] S. M. Malakouti, “Babysitting hyperparameter optimization and 10-fold-cross-validation to enhance the performance of ML methods in predicting wind speed and energy generation,” *Intell. Syst. with Appl.*, vol. 19, p. 200248, 2023, <https://doi.org/10.1016/j.iswa.2023.200248>.
 - [32] T. R. Mahesh, O. Geman, M. Margala, and M. Guduri, “The stratified k-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification,” *Healthc. Anal.*, vol. 4, p. 100247, 2023, <https://doi.org/10.1016/j.health.2023.100247>.
 - [33] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci. Rep.*, vol. 14, no. 1, p. 6086, 2024, <https://doi.org/10.1038/s41598-024-56706-x>.
 - [34] E. Kına, “Real-time food allergen detection using OCR-enhanced machine learning techniques,” *PeerJ Comput. Sci.*, vol. 11, p. e3338, 2025, <https://doi.org/10.7717/peerj-cs.3338>.
 - [35] Zeeshan Ahmad, Anum Farooq, Muhammad Fuzail, Yasir Aziz, and Naeem Aslam, “Food allergy detection using machine learning approach,” *Kashf J. Multidiscip. Res.*, vol. 2, no. 04, pp. 116–127, 2025, <https://doi.org/10.71146/kjmr402>.
 - [36] S. Brahimi, “AI-powered dining: text information extraction and machine learning for personalized menu recommendations and food allergy management,” *Int. J. Inf. Technol.*, vol. 17, no. 4, pp. 2107–2115, 2025, <https://doi.org/10.1007/s41870-024-02154-9>.
 - [37] G. Charizanos, H. Demirhan, and D. İçen, “Binary classification with fuzzy logistic regression under class imbalance and complete separation in clinical studies,” *BMC Med. Res. Methodol.*, vol. 24, no. 1, p. 145, 2024, <https://doi.org/10.1186/s12874-024-02270-x>.
 - [38] T. Wongvorachan, S. He, and O. Bulut, “A Comparison of under sampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining,” *Information*, vol. 14, no. 1, p. 54, 2023, <https://doi.org/10.3390/info14010054>.
 - [39] Y. Li, Y. Yang, P. Song, L. Duan, and R. Ren, “An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space,” *Sci. Rep.*, vol. 15, no. 1, p. 23521, 2025, <https://doi.org/10.1038/s41598-025-09506-w>.
 - [40] M. K. Vadiveloo, “Exploring facilitators and barriers for personalized dietary incentives among online shoppers at cardiovascular risk and key informants to inform an automated shopping platform,” *J. Nutr. Educ. Behav.*, 2025, <https://doi.org/10.1016/j.jneb.2025.07.010>
 - [41] K.-S. Lee and C.-W. W. Tao, “Culinary knowledge sharing on social media: Case of the 2019 Malaysian World Pastry Champion Wei Loon Tan,” *J. Hosp. Tour. Manag.*, vol. 52, pp. 52–64, 2022, <https://doi.org/10.1016/j.jhtm.2022.06.006>.
 - [42] W. M. Blom, “Allergen labelling: Current practice and improvement from a communication perspective,” *Clin. Exp. allergy*, vol. 51, no. 4, pp. 574–584, 2021, <https://doi.org/10.1111/cea.13830>.
 - [43] S. Barker, “Allergy education and training for physicians,” *World Allergy Organ. J.*, vol. 14, no. 10, p. 100589, 2021, <https://doi.org/10.1016/j.waojou.2021.100589>.
 - [44] M. Shahab, J. Xiao, J. Wang, and Z. Huang, “Molecular basis of BACE1 modulation revealed by machine learning, molecular simulations, and experimental validation,” *Int. J. Biol. Macromol.*, p. 152409, 2026, <https://doi.org/10.1016/j.ijbiomac.2026.152409>.
 - [45] F. Gerz and M. Jelali, “Towards reliable data augmentation in machine learning: Practices to prevent data leakage,” *Smart Agric. Technol.*, vol. 15, p. 102462, 2026, <https://doi.org/10.1016/j.atech.2026.102462>.
 - [46] Y. Rimal and N. Sharma, “Ensemble machine learning prediction accuracy: local vs. global precision and recall for multiclass grade performance of engineering students,” *Front. Educ.*, vol. 10, 2025, <https://doi.org/10.3389/feduc.2025.1571133>.