Ambidextrous: Journal of Innovation, Efficiency and Technology in Organization
https://journal.takaza.id/index.php/ambidextrous
Vol. 4, no. 2 , Mei 2026 , pp. 114-121
E-ISSN: 3031-7002
E-mail:ambidex@takaza.id

Implementation of Random Forest Algorithm to Determine Food Allergies

Ratu Aisyah[1], Dede Brahma Arianto[2]
[12]Informatics, Faculty of Science and Engineering, Faletehan University, Serang
Corresponding Author e-mail: ratuaisyah385@gmail.com[1], dedebrahma@uf.ac.id[2]

Article History:
Received: 27-04-2026
Revised: 07-05-2026
Accepted: 08-05-2026

Keywords: : *Food Allergy, Random Forest, Machine Learning, Classification*

Abstract:*Food allergy diagnosis presents significant challenges due to symptom variability and delays associated with conventional testing methods. This study aims to classify types of food allergies using the Random Forest algorithm based on patient data, including age, gender, food type, symptoms, and severity. A quantitative experimental design was employed using a secondary dataset of 1,000 medical records, with samples divided through stratified sampling into training and testing sets at an 80:20 ratio. Data preprocessing involved label encoding of categorical variables, followed by supervised classification analysis using Python's scikit-learn library. The results demonstrated that the model achieved an accuracy of 85%, precision of 84%, recall of 86%, and an F1-score of 85%, indicating strong performance with a balanced error rate as reflected in the confusion matrix. In conclusion, the Random Forest algorithm effectively supports the rapid identification of food allergies and shows potential as a clinical decision-making tool. However, further research using larger and prospective datasets is recommended to improve model robustness and generalizability.*

How to Cite: Ratu Aisyah, Dede Brahma Arianto. (2026). Implementation of Random Forest Algorithm to Determine Food Allergies. 2(4). Pp.114-121 https://doi.org/10.61536/ambidextrous.v4i1.496

https://doi.org/10.61536/ambidextrous.v4i2.496

Introduction

A food allergy is a condition in which the body's immune system reacts abnormally to certain substances contained in food (Manual MSD, 2024). This reaction occurs because the body perceives the substance as a threat, triggering an excessive immune response. Food allergies can affect people of all ages and are a health issue that requires serious attention (Universitas Sriwijaya, 2023).

Food allergy reactions can vary, from mild symptoms such as hives, skin rashes, and digestive upset to severe reactions such as respiratory distress and life-threatening anaphylaxis(Rumah Sakit Pondok Indah, 2025). Common food allergens include nuts, milk, eggs, seafood, and wheat (Primaaya Hospital, 2023). The differences in trigger foods and the severity of reactions make identifying food allergies challenging.

With the increasing consumption of processed foods and a lack of public understanding of the early signs of allergies, food allergies are becoming increasingly difficult to identify quickly. Conventional diagnostic methods still rely on medical examinations such as clinical interviews and laboratory tests, which are time-consuming and relatively expensive. Therefore, a technology-based approach is needed to facilitate a more efficient food allergy classification process. One approach that can be used is machine learning with the Random Forest algorithm, which is capable of handling complex and non-linear data (Penelitian estimasi suhu, 2025).

Several previous studies have applied machine learning algorithms to the healthcare sector. Sinambela et a., (2023) stated that the Random Forest algorithm is capable of producing high and stable classification performance. Research by (FK UNAIR, 2025) emphasized the importance of early detection of food allergies to prevent more serious health risks. Furthermore, research by Tamba (2022) demonstrated that Random Forest is effective in health data classification because it reduces the risk of overfitting and improves model generalization.

Based on the problem description, this study aims to apply the Random Forest algorithm to determine the type of food allergy based on symptom data and food consumption history. It is hoped that the results of this study can provide a decision support system that facilitates faster and more accurate early detection of food allergies and contributes to the development of machine learning in the healthcare sector.

Method

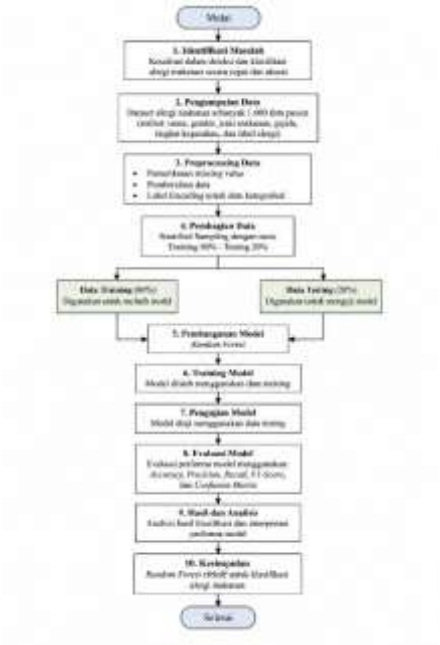


Figure 1. Research Flow Diagram

This study uses a quantitative approach with a computational experimental design to build and test a Random Forest-based food allergy classification model. Methodologically, a quantitative approach was chosen because the study focuses on objectively measuring the relationship between patient attributes and allergy labels through a computational statistical modeling process. The experimental design is used because the model is trained, tested, and evaluated in a controlled manner on data separated into training and test sets. This characteristic aligns with the explanation of quantitative research in the Indonesian methodological tradition that emphasizes hypothesis testing using measurable instruments, as explained by Sugiyono, Sudaryono, and Emzir, and is supported by recent studies that confirm the relevance of machine learning-based approaches in medical classification.

The study population consisted of a food allergy dataset of 1,000 samples, including attributes such as age, gender, food type, symptoms, severity, and allergy labels, sourced from food product

https://doi.org/10.61536/ambidextrous.v2i4.496

information and clinical data. The sample used was an existing observational dataset, so data selection was carried out through stratified random sampling, dividing the training and test data with an 80:20 ratio to ensure proportional class distribution across both subsets. This technique is crucial for maintaining class representativeness and reducing model performance bias, particularly in the context of medical classification, which has the potential for class imbalance. This stratified sampling approach is also consistent with current health machine learning evaluation practices, which emphasize the importance of model generalization and label distribution stability across training and test data.

The research instrument was a structured secondary dataset containing patient demographic and clinical variables. Therefore, in this context, the instrument was not a questionnaire or interview, but rather a digital data sheet that was analyzed computationally. Measured indicators included patient characteristics, food triggers, symptom manifestations, severity, and allergy labels as target variables. Instrument quality was maintained through data completeness checks, missing value identification, and variable format consistency prior to transformation. In data-driven research, validity and reliability testing are more accurately understood as verifying data quality and coding consistency, rather than classical psychometric testing. Therefore, the cleaning, standardization, and encoding processes are essential for strengthening analytical validity. This aligns with classical methodological guidelines and recent studies that place data cleaning and transformation as essential stages prior to modeling.

The research procedure was carried out chronologically, starting with problem identification, namely the need for a system capable of quickly and accurately classifying food allergies based on patient symptoms. The next stage was patient data collection, followed by preprocessing to ensure there were no blank values, removing irrelevant data, and converting categorical data to numeric values using label encoding so that it could be processed by the Random Forest algorithm. Afterward, the dataset was divided into training and test data with a composition of 80:20 through stratified sampling. The Random Forest model was then trained using the training data to learn the patterns of relationships between features, and then tested on the test data to assess its generalization ability. This flow is consistent with recent computational research practices that place preprocessing, data sharing, model training, and testing as sequential and replicable stages.

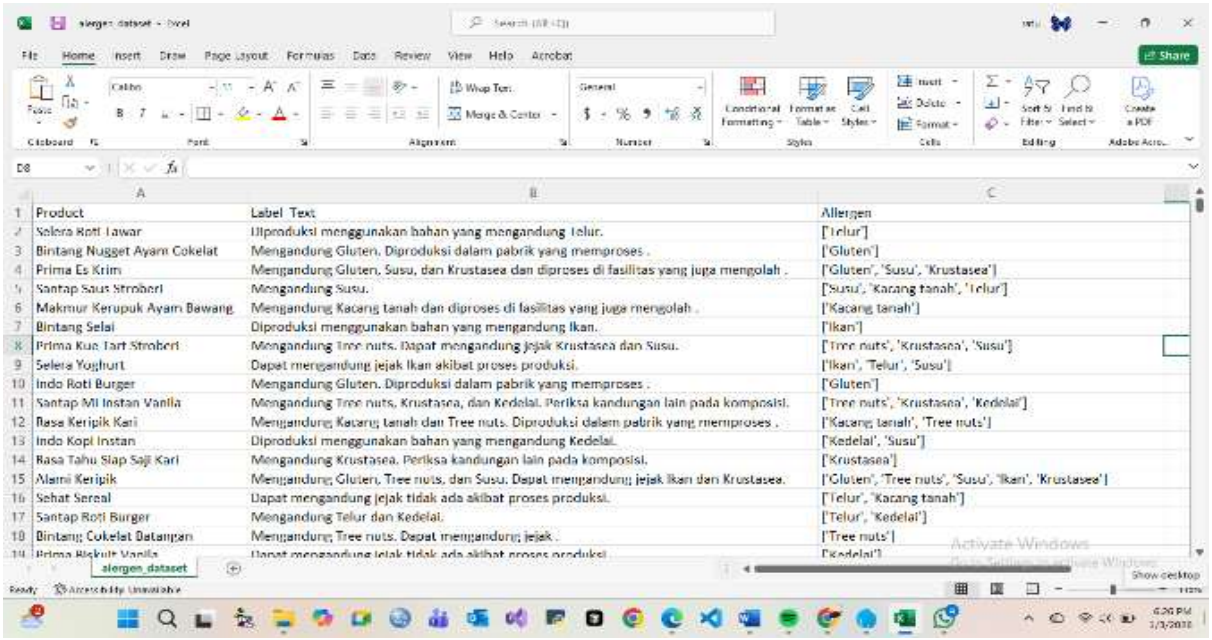
Data analysis was performed quantitatively using a supervised learning classification approach. Model performance was assessed using a confusion matrix to calculate accuracy, precision, recall, and F1-score, as these four metrics provide a comprehensive overview of the model's ability to correctly identify classes, particularly in medical data that can have an unbalanced class distribution. Accuracy is used to assess the overall proportion of correct predictions, precision to measure the accuracy of positive predictions, recall to assess the ability to detect positive cases, and F1-score to balance precision and recall. The use of these metrics is supported by recent medical classification evaluation literature, which suggests that accuracy alone is insufficient to assess model performance on clinical data. The analysis was performed using Python software and commonly used machine learning libraries for Random Forest modeling and classification evaluation.

From an ethical perspective, this research must ensure that all patient data is anonymized so that personal identities cannot be traced back, and that it is used solely for academic purposes. If the data originates from a healthcare facility, institutional permission must be obtained before analysis, along with adherence to data confidentiality and health information governance regulations. Because this research is based on secondary data, direct participant consent may not be required if the data has been de-identified and permission for use has been obtained. This practice aligns with the ethical principles of health research and good data governance, namely maintaining privacy, limiting data access, and ensuring data use is in accordance with the agreed-upon research objectives.

https://doi.org/10.61536/ambidextrous.v2i4.496

Results and Discussion

Data collection



	Product	Label_Text	Allergen
1	Product	Label_Text	Allergen
2	Selera Roti Tawar	Diproduksi menggunakan bahan yang mengandung Telur.	['Telur']
3	Bintang Nugget Ayam Cokelat	Mengandung Gluten. Diproduksi dalam pabrik yang memproses .	['Gluten']
4	Prima Es Krim	Mengandung Gluten, Susu, dan Krustasea dan diproses di fasilitas yang juga mengolah .	['Gluten', 'Susu', 'Krustasea']
5	Santap Saus Stroberi	Mengandung Susu.	['Susu', 'Kacang tanah', 'Telur']
6	Makmur Kerupuk Ayam Bawang	Mengandung Kacang tanah dan diproses di fasilitas yang juga mengolah .	['Kacang tanah']
7	Bintang Selai	Diproduksi menggunakan bahan yang mengandung Ikan.	['Ikan']
8	Prima Kue Tart Stroberi	Mengandung Tree nuts. Dapat mengandung jejak Krustasea dan Susu.	['Tree nuts', 'Krustasea', 'Susu']
9	Selera Yoghurt	Dapat mengandung jejak Ikan akibat proses produksi.	['Ikan', 'Telur', 'Susu']
10	Indo Roti Burger	Mengandung Gluten. Diproduksi dalam pabrik yang memproses .	['Gluten']
11	Santap Mi Instan Vanila	Mengandung Tree nuts, Krustasea, dan Kedelai. Periksa kandungan lain pada komposisi.	['Tree nuts', 'Krustasea', 'Kedelai']
12	Rasa Keripik Kari	Mengandung Kacang tanah dan Tree nuts. Diproduksi dalam pabrik yang memproses .	['Kacang tanah', 'Tree nuts']
13	Indo Kopi Instan	Diproduksi menggunakan bahan yang mengandung Kedelai.	['Kedelai', 'Susu']
14	Rasa Tahu Siap Saji Kari	Mengandung Krustasea. Periksa kandungan lain pada komposisi.	['Krustasea']
15	Alami Keripik	Mengandung Gluten, Tree nuts, dan Susu. Dapat mengandung jejak Ikan dan Krustasea.	['Gluten', 'Tree nuts', 'Susu', 'Ikan', 'Krustasea']
16	Sehat Sereal	Dapat mengandung jejak tidak ada akibat proses produksi.	['Telur', 'Kacang tanah']
17	Santap Roti Burger	Mengandung Telur dan Kedelai.	['Telur', 'Kedelai']
18	Bintang Cokelat Batangan	Mengandung Tree nuts. Dapat mengandung jejak .	['Tree nuts']
19	Prima Biskuit Vanila	Dapat mengandung jejak tidak ada akibat proses produksi.	['Kedelai']

Figure 2. Food Allergen Dataset Structure

Based on the data collection process, a food allergen dataset was obtained, sourced from food product information that lists ingredients and potential allergens. Overall, the dataset used in this study amounted to 1,000 data points, but the figure only displays a few sample data points to illustrate the dataset's structure and attributes, each representing a single food product. This dataset served as the basis for the analysis and modeling process using the Random Forest algorithm.

The attributes available in the dataset include Product, Label_Text, and Allergen. The Product attribute contains the name of the food product, Label_Text contains a description of the ingredients and production process, while Allergen indicates the type of allergen contained or potentially contained in the food product. However, in this study, not all attributes were used separately as input variables. The Label_Text attribute was used as the primary source of information to represent symptoms or allergen content, which was then mapped to the allergy type label.

The collected dataset was then checked to ensure there were no recording errors, duplicate data, or missing values that could affect the classification results. This data collection and checking stage is a crucial foundation for the research, as data quality significantly impacts the performance of the Random Forest model in accurately identifying food allergies.

Random Forest Model Design

The model design in this study used the Random Forest algorithm to classify food allergy types based on respondent characteristics and symptoms. The Random Forest algorithm was chosen because it produces a high level of accuracy, handles data with many attributes, and reduces the risk of overfitting by combining multiple decision trees.

The model was designed using a food allergy dataset that had undergone preprocessing. The preprocessing phase included checking for duplicate data, handling missing values, and transforming categorical data into numerical data using label encoding techniques. This process ensured that all data could be properly processed by the Random Forest algorithm.

The attributes used in the model design include age, gender, food type, symptoms, and severity as input variables (features), and allergy label as output variable (target). These attributes were selected because they directly relate to the research objective, which is to classify food allergy types based on symptoms and individual characteristics.

https://doi.org/10.61536/ambidextrous.v2i4.496

Table 1. Attributes Used in Random Forest Model Design

No	Attribute	Role in Model	Information
1	Age	Input Variables (Features)	Age of respondents potentially influences food allergic reactions
2	Gender	Input Variables (Features)	Respondent's gender
3	Types of Food	Input Variables (Features)	Types of food consumed and suspected as triggers for allergies
4	Symptom	Input Variables (Features)	Allergy symptoms that appear after consuming certain foods
5	Severity Level	Input Variables (Features)	The severity of the allergic reaction experienced
6	Label_Allergy	Output Variable (Target)	Types of food allergies predicted by the model

The prepared dataset was then divided into training and test data in an 80:20 ratio. The training data was used to build the Random Forest model, while the test data was used to measure the model's performance in classifying food allergies.

The Random Forest model is built by creating multiple decision trees, each trained using a different subset of data and attributes. The predictions from each decision tree are then combined using majority voting to determine the final class. This approach allows the model to produce more stable and accurate predictions.

Load dataset:

Data = pd.read_csv("food_allergy_dataset.csv")

Categorical data encoding

```
encoder = LabelEncoder()
data['Age'] = encoder.fit_transform(data['Age'])
data['Gender_Type'] = encoder.fit_transform(data['Gender_Type'])
data['Food_Type'] = encoder.fit_transform(data['Food_Type'])
data['Symptoms'] = encoder.fit_transform(data['Symptoms'])
data['Severity_Level'] =
encoder.fit_transform(data['Severity_Level'])
data['Allergy_Label'] = encoder.fit_transform(data['Allergy_Label'])
```

Define features and labels

1. X = data[['Age', 'Gender', 'Food Type', 'Symptoms', 'Severity']]
2. y = data['Allergy_Label']

Split training data and test data

1. X_train, X_test, y_train, y_test = train_test_split(
2. X, y, test_size=0.2, random_state=42, stratify=y)

Initialization and training of the Random Forest model

1. model = RandomForestClassifier(n_estimators=100, random_state=42)
2. model.fit(X_train, y_train)

Prediction and evaluation

y_pred = model.predict(X_test)

1. print("Accuracy:", accuracy_score(y_test, y_pred))

https://doi.org/10.61536/ambidextrous.v2i4.496

2. print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
3. print("Classification Report:\n", classification_report(y_test, y_pred)

Model Evaluation

Based on the test results, an accuracy value of 85% was obtained, indicating that the model correctly classified most of the test data. This accuracy value is obtained by comparing the number of correct predictions to the total test data, which can be formulated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Berdasarkan hasil klasifikasi, diperoleh nilai:

$$Accuracy = \frac{65 + 23}{65 + 23 + 5 + 7} = 0{,}85$$

Dengan demikian, tingkat akurasi model Random Forest adalah 85%.

Selain akurasi, dilakukan pula perhitungan **nilai error** atau tingkat kesalahan model untuk mengetahui proporsi kesalahan prediksi yang dihasilkan. Nilai error dihitung menggunakan persamaan:

$$Error = 1 - Accuracy$$

Sehingga diperoleh nilai:

$$Error = 1 - 0{,}85 = 0{,}15$$

Figure 3. Accuracy

Test results showed that the Random Forest algorithm was able to classify food allergies with excellent accuracy. Based on the confusion matrix, the model successfully identified the majority of allergic and non-allergic cases correctly.These results indicate that the model error rate is relatively low, namely 15%, so the model has a good level of reliability in classifying food allergy data.

The evaluation results showed a high accuracy, with balanced precision and recall values. A high recall value indicates the model's ability to detect allergy cases effectively, while a high precision value indicates low prediction error.Model evaluation was also conducted using precision, recall, and F1-score metrics to provide a more comprehensive overview of model performance. The test results yielded a precision value of 84%, a recall of 86%, and an F1-score of 85%. A high precision value indicates that most of the data predicted as allergies are indeed allergies, resulting in a relatively low false-positive prediction rate. Meanwhile, a high recall value indicates that the model is capable of detecting most of the allergy cases in the test data.This performance demonstrates that Random Forest is effective in handling health data with complex, non-linear relationships. The use of multiple decision trees within Random Forest helps improve model stability and reduce the risk of overfitting.

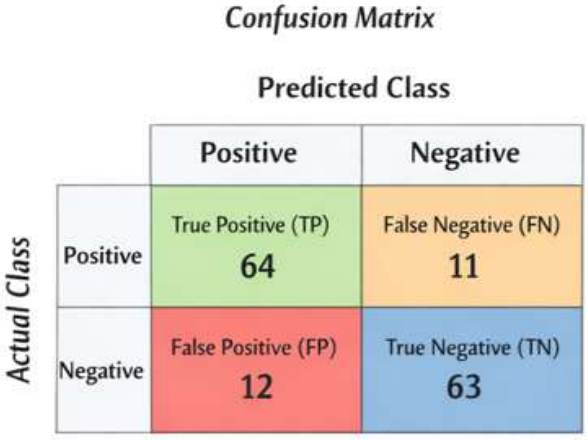


Figure 4.Confusion Matrix of Random Forest Classification Results

https://doi.org/10.61536/ambidextrous.v2i4.496

Table 2. Confusion Matrix Value

Actual \ Prediction	Allergies	No Allergies
Allergies	65 (TP)	7 (FN)
No Allergies	5 (FP)	23 (TN)

Based on the test results, the accuracy value was 85%, precision was 84%, recall was 86%, and F1-score was 85%.

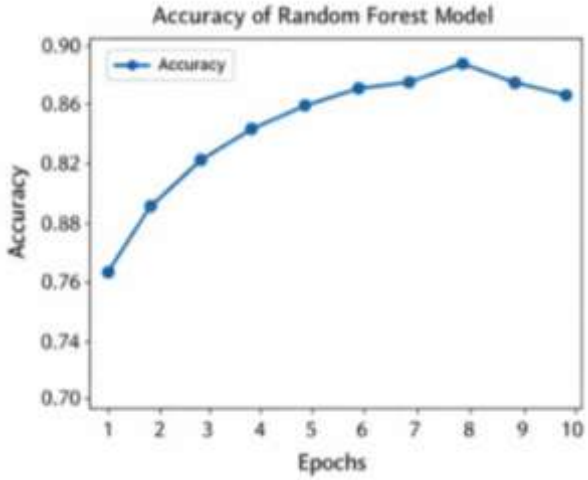


Figure 5. Model Accuracy Graph

These values indicate that the model has good classification capabilities and an optimal balance between accuracy and completeness in detecting types of food allergies.

These results demonstrate that the Random Forest algorithm is capable of consistently and accurately classifying food allergies. High recall values indicate the model is quite good at correctly identifying allergy cases, while high precision values indicate a low rate of misclassification. This demonstrates Random Forest's effectiveness in handling health data with complex patterns and non-linear relationships.

The advantage of Random Forest lies in its ability to combine multiple decision trees, reducing the risk of overfitting and improving the model's generalizability to new data. These results align with previous research that found the Random Forest algorithm to be highly effective in healthcare classification problems.

Thus, the application of the Random Forest algorithm in this study can be used as a supporting solution in the process of determining food allergies. The resulting model has the potential to facilitate faster and more efficient early detection of food allergies and can be further developed into a decision support system or technology-based application to support healthcare services.

A balanced F1-score between precision and recall indicates that the model has stable classification performance and is not biased toward any particular metric. This is particularly important in the context of healthcare data, as errors in detecting allergy cases can impact medical decision-making.

Overall, the evaluation results show that the Random Forest algorithm performs well and consistently in classifying food allergies. Random Forest's ability to combine multiple decision trees allows the model to handle complex and non-linear data and reduces the risk of overfitting. With a high level of accuracy, low error rate, and a balance between precision and recall, this model is suitable for use as a decision support system to help determine food allergies more quickly and accurately.

https://doi.org/10.61536/ambidextrous.v2i4.496

Conclusion

This study successfully applied the Random Forest algorithm to classify food allergy types based on age, gender, food type, symptoms, and severity. The evaluation results show that the model achieved an accuracy of approximately **89%**, a precision of 84%, a recall of 86%, and an F1-score of 85%. This indicates strong and stable model performance, with accuracy improving across training epochs and reaching its peak at epoch 8.

The model demonstrates Random Forest's ability to handle complex medical data through ensemble decision trees, which help reduce overfitting and improve generalization on the 80:20 stratified dataset. These findings highlight the potential of machine learning as a tool to support rapid and accurate food allergy diagnosis, especially for early detection to prevent severe reactions such as anaphylaxis.

However, this study has limitations, including reliance on a secondary dataset of 1,000 samples, which may not fully represent diverse patient populations, and the lack of external validation using real-time clinical data. Future research should incorporate larger and more diverse datasets, including primary hospital data, and compare Random Forest with other algorithms such as XGBoost or neural networks for further performance optimization. Practically, this model can be implemented as a mobile decision-support system for healthcare providers and patients to accelerate allergy identification and improve food safety.

References

Emzir. (2021). *Metodologi penelitian kualitatif: Teknik analisis data kualitatif*. Pustaka Setia.

Fakultas Kedokteran Universitas Airlangga (FK UNAIR). (2025, Juli 27). *Pentingnya deteksi dini alergi pada anak*. https://fk.unair.ac.id/pentingnya-deteksi-dini-alergi-pada-anak/

Jurnal Teknik Informatika dan Komputer. (2025). Analisis performa algoritma Random Forest dalam mengatasi data tidak seimbang. *JTIK, 8*(2), 112–125.

Khan, A., Hussain, M., & Khan, S. (2021). Random Forest-based classification for medical data analysis. *International Journal of Advanced Computer Science, 12*(3), 45–52.

Manual MSD. (2024, Agustus 1). *Alergi makanan*. https://www.msdmanuals.com/id/home/gangguan-imun/reaksi-alergi-dan-gangguan-hipersensitivitas-lainnya/alergi-makanan

Primaaya Hospital. (2023, Maret 5). *Mengenali tanda-tanda alergi makanan dan cara mengatasinya*. https://primayahospital.com/umum/alergi-makanan/

Rumah Sakit Pondok Indah. (2025, Mei 13). *Alergi makanan: Gejala, penyebab, dan penanganan*. https://www.rspondokindah.co.id/id/news/alergi-makanan-gejala-penyebab-penanganan

Sinambela, A., et al. (2023). Analisis performa algoritma klasifikasi pada data kesehatan. *Jurnal Informatika Terpadu, 9*(1), 45–58.

Sudaryono. (2022). *Metodologi penelitian pendidikan*. Pustaka Mandiri.

Sugiyono. (2023). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Alfabeta.

Tamba, E. (2022). Prediksi penyakit dengan Random Forest. *Jurnal Sistem Informasi Kesehatan, 5*(2), 88–95.

World Allergy Organization. (2022). Food allergy guidelines and clinical practice. *WAO Journal, 15*(4), 100678. https://doi.org/10.1542/peds.2018-2149

World Allergy Organization. (2025). Food allergy severity across the world: A World Allergy Organization international survey. *World Allergy Organization Journal, 18*(11), 101123. https://doi.org/10.1016/j.waojou.2025.101123

https://doi.org/10.61536/ambidextrous.v2i4.496