Indonesian Journal of Data and Science

Volume 6 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v6i2.257

Research Article

Comparative Analysis of OCR Methods Integrated with Fuzzy Matching for Food Ingredient Detection in Japanese Packaged Products

Muhammad Zaky Rahmatsyah [1], Jevri Tri Ardiansah [2,*], Anik Nur Handayani [3]

[1] Universitas Negeri Malang, Malang, Indonesia, muhammad.zaky.2105356@students.um.ac.id
[2] Universitas Negeri Malang, Malang, Indonesia, jevri.ardiansah.ft@um.ac.id
[3] Universitas Negeri Malang, Malang, Indonesia, aniknur.ft@um.ac.id

Correspondence should be addressed to Jevri Tri Ardiansah; jevri.ardiansah.ft@um.ac.id

Received 18 January 2025; Accepted 29 May 2025; Published 31 July 2025

Abstract:

Advances in digital technology offer a solution to the challenges faced by foreign consumers in understanding ingredient information on Japanese food packaging, especially due to the use of Kanji, Hiragana, and Katakana characters. This study develops and reveals an allergen detection method based on Optical Character Recognition (OCR) and fuzzy match applied to Japanese food packaging. Three OCR methods—Google Vision OCR, PaddleOCR, and Tesseract OCR—were compared and evaluated using Precision, Recall, F1-Score, and Confusion Matrix metrics.The study began with the collection of food product images from bold sources, followed by text extraction using the three OCR methods. The extracted text was then cleaned and normalized before being matched with ground truth data using fuzzy match. Testing was conducted on 10 product image samples with varying quality and lighting conditions. The results showed that Google Vision OCR outperformed the others, achieving an average F1 score of 1.00, followed by PaddleOCR (0.75), and Tesseract OCR (0.30). Google Vision was the most consistent in detecting allergens such as 乳 (milk), 小麦 (wheat), and 卵 (egg). These findings suggest that the integration of OCR and fuzzy matching is effective in detecting allergens, even in the presence of textual variations and recognition errors. This study contributes to the development of automated food recommendation systems for foreign consumers, especially those who have food preferences due to allergies, religious beliefs, or personal preferences.

Keywords: Allergen Detection; Google Vision OCR; Paddle OCR; Tesseract OCR; Fuzzy Matching.
Dataset link: https://drive.google.com/drive/folders/1udtD-T8B5aBvS-3UpNAMzrKQzFF8gyOS?usp=sharing

1. Introduction

Japanese cuisine offers a diverse range of flavors and ingredients, shaped by cultural preferences and sensory experiences [1]. However, for individuals with dietary restrictions due to allergies, religion, or personal preferences, identifying safe food choices can be challenging. The complexity of Japanese food labels, particularly those written in Kanji, often poses difficulties for foreign consumers [2]. Advances in digital technology offer a solution through Optical Character Recognition (OCR) [3] and fuzzy matching. OCR facilitates the extraction of ingredient text from Japanese food packaging, converting Kanji, Hiragana, and Katakana scripts into machine-readable formats [4]. Fuzzy matching enhances identification accuracy by comparing extracted text with a reference database of allergens and ingredients [5]. OCR extracts text from food packaging, including Japanese words as Kanji, Hiragana, and Katakana, enabling digital access to visual information. Fuzzy matching then matches extracted words with a reference list of ingredients and allergens based on similarity, making it more flexible for spelling variations and text noise.

Previous research has applied OCR extensively in document digitization and text recognition [6]. However, OCR implementation on Japanese food labels faces challenges such as image noise, font variations, and non-standard text layouts. Meanwhile, fuzzy matching improves ingredient identification by comparing OCR-extracted text with reference data or ground truth data, though its effectiveness depends on input text quality. This study develops and evaluates an OCR-fuzzy matching pipeline for detecting allergens in Japanese food packaging. The research compares multiple OCR methods, namely Google Vision OCR, PaddleOCR, and Tesseract OCR, along with fuzzy matching techniques to determine the most effective approach. The evaluation is based on Precision, Recall, F1-Score, and Confusion Matrix as performance metrics. The Confusion Matrix is utilized to assess the performance of different OCR methods Google Vision OCR, PaddleOCR, and Tesseract OCR, when combined with fuzzy string matching by comparing their results against the ground truth. This approach provides a comprehensive analysis of classification accuracy, including True Positives, False Positives, True Negatives, and False Negatives [7], enabling a deeper understanding of each method's effectiveness in detecting allergens from Japanese food packaging, however Consumer protection and law enforcement require appropriate analytical techniques to detect allergens in food [8]. The findings aim to support the development of an automated food recommendation system for foreign consumers.

2. Method:

The research process is structured into several key stages, as illustrated in **Figure 1**. It begins with data collection from various open-source sources. Next, OCR processing is carried out using three different OCR models to optimize text extraction performance [9]. The extracted text then undergoes cleaning and normalization to enhance accuracy during the fuzzy matching phase. In this stage, the processed text is compared against reference or ground truth data. Finally, the performance is evaluated in the last stage using several evaluation metrics and confussion matrix to assess the effectiveness of the approach.

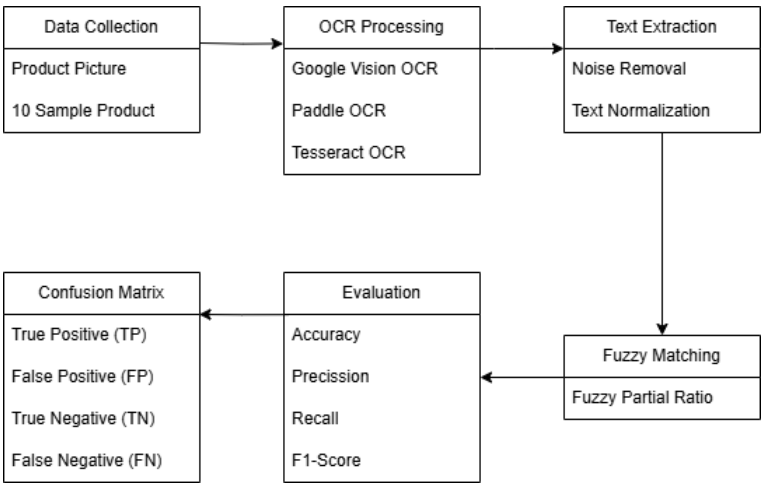

Figure 1. Workflow Method

Data Collection

The initial stage of this research is the collection of Japanese food product packaging images containing food ingredient information in Japanese. This data is obtained from open-source sources such as the halal japan food facebook group and Rakuten marketplace Japan [10], [11].

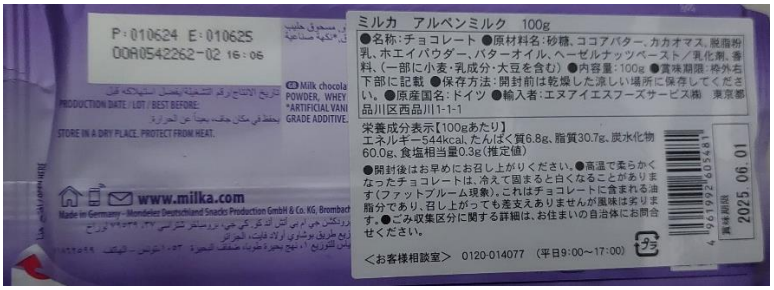

Figure 2. Product Food Product Example

In **Figure 1**. is an example of the Japan Product display, which is used in this study. **Figure 3**. Shows the workflow of data collection, in this study there are 2 processes carried out on the data obtained:

- Product Composition Image Capture, images taken manually with a camera or obtained from online sources that are of high quality, the image format used is JPG/PNG with a resolution that allows OCR to recognize text well, and a total of 10 images were selected to ensure text diversity and different levels of difficulty.
- Ground Truth Data Collection, ground data is collected manually by copying the text of food ingredients from product packaging. This ground truth will be the reference text used to compare the OCR extraction results, and the collection of ground truth text is carried out with repeated reviews to ensure that the reference text is truly accurate and in accordance with the original text on the packaging [12].

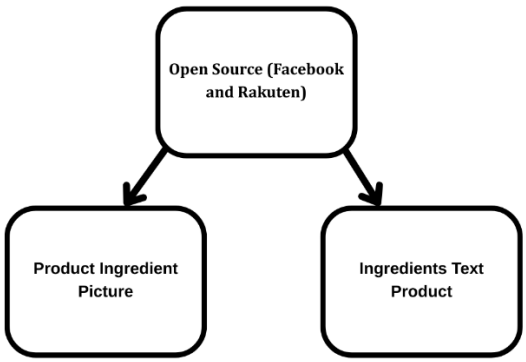

Figure 3. Data Collection WorkFlow

OCR Processing

OCR Processing is the initial stage in the system that is tasked with converting text information contained in images into digital format [13]. The Optical Character Recognition (OCR) processing stage is the core of this research, where the text on Japanese food product packaging is extracted into text using OCR technology. In this study, this stage not only includes the character recognition process through the Google Vision API, but also an important step to clean the extracted text. This cleaning is important to remove noise, irrelevant characters, and to match the text format so that the matching process with ground truth data can later be carried out more accurately. This process aims to extract the text contained in the image automatically. In this study, three different Optical Character Recognition methods were used to compare:

- Google Vision OCR, is a cloud-based service that uses machine learning technology to recognize text in various languages, including Japanese. Google Vision also has good text processing capabilities in various lighting conditions and image tilt angles [14]. The main advantage of this OCR is its good ability to recognize Japanese characters (Kanji, Katakana, and Hiragana) with fairly high accuracy [15].

- Paddle OCR, is a deep learning-based OCR system developed by PaddlePaddle [16] this OCR supports various languages and the main advantage of PaddleOCR is its flexibility in handling multilingual text [17] and its ability to process images with high noise [18].
- Tesseract OCR, is an open source OCR developed by Google [19], Tesseract is known for its adaptability in text recognition tasks [20] and also supports various languages and has a special mode for Japanese, then the disadvantage of this OCR is its dependence on image quality [21], where the extraction results are greatly influenced by factors such as light and image resolution [22].

After the data is obtained, the data collection process, then the image is processed using each OCR method. At this stage, OCR will detect areas in the image that contain text, then the OCR algorithm will try to recognize the characters and words in the text, the text that has been successfully extracted will be added with several functions to modify the extracted data using ocr_processing function which functions to perform several tasks:

- Special Character Replacement, namely replacing Japanese commas ("、") with regular commas or appropriate separators to maintain format consistency.
- Removing Irrelevant Character, using regular expressions (RegEx), this function removes characters other than letters, numbers, and spaces. This helps eliminate noise that can come from OCR recognition errors, such as symbols that shouldn't appear (emad).

Text Extraction and Normalization

Text Extraction and normalization is a method stage that aims to clean and normalize text from OCR processing [23]. This stage focuses on retrieving specific information from the cleaned OCR results, namely the parts of the text containing the list of Japanese food product ingredients are cleaned and rearranged the text extraction results from the OCR system, so that they can be adjusted to the ground truth. After the image is processed by OCR (Google Vision, PaddleOCR, and Tesseract), the extracted text often has noise, inconsistent formatting, or misrecognized characters [24]. Therefore, this stage is very important to improve the quality of the text before comparing it with the reference text using the fuzzy matching method. In this stage, several steps are taken:

- Noise Removal, is a method that is carried out due to several factors that often occur in OCR extraction results, such as foreign or irrelevant characters, for example symbols or numbers that are not supposed to be there, then errors in separating words, for example the word "アレルゲン" can be recognized as "アレルゲン" by OCR with the wrong spacing, and finally similar letter errors, for example the number "1" is recognized as the letter "I" or the letter "O" is recognized as number "0".
- Text Normalization, is a method used to adjust the text format and normalize Kanji, Katakana, or Hiragana letters if there are any discrepancies, and remove furigana marks that may be read as text [25].
- Text Section Extraction, this method is used to search for keywords “原材料”, “成分”, “Ingredients”, “Composition”, “アレルゲン”, and the required words, so that the system can retrieve the text after the keyword, so that it can limit it to a certain line or certain punctuation to ensure that only the ingredients list is retrieved.

Fuzzy Matching

Fuzzy Matching is a string-matching technique used to measure the level of similarity between two text tokens, although not perfectly identical.

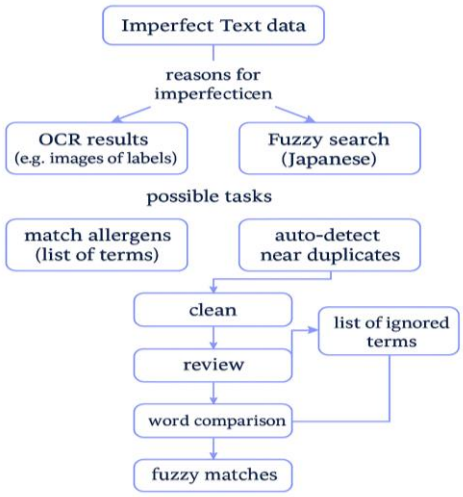

Figure 4. Fuzzy Matching Workflow [26]

In this case, fuzzy matching is used to match the text extraction results from OCR with the previously determined Ground Truth [27], fuzzy matching tolerates minor differences such as spelling errors, spacing variations, or differences in character order [5]. This technique produces a similarity score in the range of 0 to 100 [27], where a higher score indicates a greater level of similarity between the two texts. In this study, Fuzzy Matching is used to help evaluate how accurate the OCR model is in detecting allergens, using a ratio threshold of 5 and 10.

Testing

Testing was carried out to validate system performance starting from OCR Processing, Text Extraction, to Fuzzy Matching in detecting and extracting information on food ingredients and allergens. The test dataset consists of 10 samples of Japanese food product packaging images taken in varying lighting conditions, angles and quality, as well as ground truth data collected manually. The testing procedure includes image processing with OCR, cleaning and extracting text using the clean_ocr_text and extract_ingredients functions, and matching the results with ground truth using fuzzy matching and filter_allergens.

Test parameters were carried out with a fuzzy matching threshold score of 5 and 10 to determine suitability, as well as variations in image types to assess OCR sensitivity. The OCR results and processed text are recorded, with fuzzy matching scores to measure accuracy and identify areas of improvement. This testing ensures that the OCR and text extraction processes produce clean and relevant text, which is then evaluated using metrics such as Accuracy, Precision, Recall, and F1-Score, as well as analysis through the Confusion matrix.

Evaluation

Performance evaluation results from OCR (Optical Character Recognition) and Fuzzy Matching in the form of extract text results for food product composition compared with existing allergen ground truth text, and evaluated based on the confusion matrix. **Table 2** shows the formula for each matrix used.

Table 2. Formula for each Matrix

Attribute	Formula	Description
Precision	$\frac{TP}{TP + FP}$	The ratio of correct positive predictions to all positive predictions made [28].
Recall	$\frac{TP}{TP + FN}$	The ratio of correct positive predictions to the total number of actual positive events [28].
F1 Score	$\frac{2 \times (Precission \times Recall)}{Precission + Recall}$	The harmonic mean between precision and recall that provides balance between the two [28].

Confusion Matrix

Confusion matrix is a representation of the classification performance of a model that shows the number of correct and incorrect predictions for each category. The confusion matrix has the form of a square matrix, where the rows indicate the actual classes, while the columns indicate the predicted classes [29].

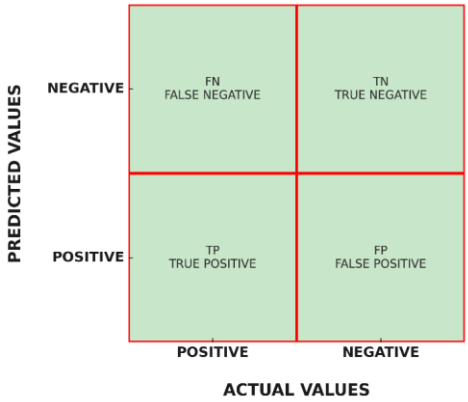

Figure 5. Confusion Matrix

In the context of this study,the four main components of the confusion matrix :

- True Positive (TP): An allergen that is actually present in the product and successfully detected by the model.
- False Positive (FP): An allergen that is not present in the product but detected by the model.
- False Positive (FP): An allergen that is not present in the product but detected by the model.
- True Negative (TN): There is no allergen present and the model also does not detect it which is not used explicitly in this evaluation because the multi-label approach focuses on positive elements.

3. Results and Discussion

In this study, the allergen detection system on Japanese food product packaging was tested using a series of processes ranging from OCR Processing, Text Extraction, to matching with ground truth data through the fuzzy matching method. The evaluation results focused on the Precision, Recall, and F1-Score metrics as well as visual analysis through a confusion matrix to explore the performance of the system [30].

Results

After the OCR method is carried out on the product image, OCR successfully produces raw text which is then cleaned with the clean_ocr_text function to remove special characters such as Japanese commas and irrelevant characters. At the Text Extraction stage, the extract_ingredients function successfully extracts parts of the text that contain food ingredient information by detecting keywords such as "原材料" and "Ingredients". The cleaned extraction results are then compared with ground truth data using fuzzy matching via the match_ingredients function.

From the fuzzy matching results, a product was obtained that matched the similarity level exceeding the score threshold of 10 and 5 so that the experiment would be more specific. Furthermore, allergen detection was carried out using the filter_allergens function, so that a list of allergens detected from the OCR text was obtained. Evaluation was carried out by calculating the Precision, Recall, and F1-Score values for each product described in **Table 3**.

Table 3. Results

No	Product	OCR	Fuzzy Ratio	Ground Truth	Detected Allergen	Precisson	Recall	F1-score
1	Bourbon Slowbar	Google	5	'小麦', '卵', '乳', '大豆'	'卵', '乳', '小麦', '大豆'	1.00	1.00	1.00
			10		'卵', '乳', '小麦', '大豆'	1.00	1.00	1.00
		Tesseract	5		'卵', '乳', '大豆', '乳'	1.00	0.75	0.86
			10		'卵', '乳', '大豆', '乳'	1.00	0.75	0.86
		Paddle	5		'卵', '乳', '小麦', '大豆'	1.00	1.00	1.00
			10		'卵', '乳', '小麦', '大豆'	1.00	1.00	1.00
2	Gogo no Koucha Caramel Tea Latte	Google	5	乳	乳	1.00	1.00	1.00
			10		乳	1.00	1.00	1.00
		Tessract	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
		Paddle	5		乳	1.00	1.00	1.00
			10		乳	1.00	1.00	1.00
3	Sanuki Shisei	Google	5	小麦	-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
		Tessract	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
		Paddle	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
4	Meiji Premium Yougurt	Google	5	乳	乳	1.00	1.00	1.00
			10		乳	1.00	1.00	1.00
		Tessract	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
		Paddle	5		乳	1.00	1.00	1.00
			10		乳	1.00	1.00	1.00
5	Meiji R-1 Yougurt	Google	5	乳	乳	1.00	1.00	1.00
			10		乳	1.00	1.00	1.00
		Tessract	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
		Paddle	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
6	Milka Alpine Milk	Google	5	'小麦', '乳', '大豆'	乳', '小麦', '大豆', '乳'	1.00	1.00	1.00
			10		乳', '小麦', '大豆', '乳'	1.00	1.00	1.00
		Tessract	5		'小麦', '大豆', '乳'	1.00	1.00	1.00
			10		'乳', '小麦', '大豆'	1.00	1.00	1.00
		Paddle	5		'乳', '小麦', '大豆'	1.00	1.00	1.00
			10		'乳', '小麦', '大豆'	1.00	1.00	1.00

7	Nestle KitKat	Google	5	'小麦', '乳', '大豆', '卵'	'卵', '乳', '小麦', '大豆'	1.00	1.00	1.00
			10		'卵', '乳', '小麦', '大豆'	1.00	1.00	1.00
		Tessract	5		-	0.00	0.00	0.00
			10			0.00	0.00	0.00
		Paddle	5		'乳', '小麦', '大豆'	1.00	0.75	0.86
			10		'乳', '小麦', '大豆'	1.00	0.75	0.86
8	Seven and I soba Tempura	Google	5	'えび', '小麦', 'そば', '卵', '乳成分', 'さば', '大豆'	'卵', '乳', '小麦', 'えび', '大豆', '魚', '乳', '乳'	0.67	0.57	0.62
			10		'卵', '乳', '小麦', 'えび', '大豆', '魚', '乳', '乳'	0.67	0.57	0.62
		Tessract	5		'卵', '小麦', 'えび'	1.00	0.43	0.60
			10		'卵', '小麦', 'えび'	1.00	0.43	0.60
		Paddle	5		'えび', '大豆'	1.00	0.29	0.44
			10		'えび', '大豆'	1.00	0.29	0.44
9	Seven Premium Clam Chowder	Google	5	'乳', '小麦'	'乳', '小麦'	1.00	1.00	1.00
			10		'乳', '小麦'	1.00	1.00	1.00
		Tessract	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
		Paddle	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
10	Shimaya Bonito Dashi	Google	5	魚	-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
		Tessract	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00
		Paddle	5		-	0.00	0.00	0.00
			10		-	0.00	0.00	0.00

The results table shows a comparison of the performance of three PCR methods in detecting allergens from 10 products.

Extraction Results and Performance Evaluation of OCR Model

Google Vision OCR demonstrated the highest performance, achieving an average F1-score of 1.00 across nearly all tested products. The model consistently and accurately identified allergens such as **乳** (milk), **小麦** (wheat), and **卵** (egg), even under varying image conditions. Its high precision and recall indicate strong reliability in detecting allergens with minimal prediction errors.

PaddleOCR ranked second, with an average F1-score of approximately 0.75. While its precision remained high, its recall declined, particularly on more complex products like Seven and I Soba Tempura. In these cases, the model failed to identify several allergens, although it still detected some correctly.

Tesseract OCR showed the lowest performance, with an average F1-score of only 0.30. The model struggled to recognize Kanji and Katakana characters, especially on products with small font sizes or low contrast. This resulted in a high number of false negatives, along with some false positives.

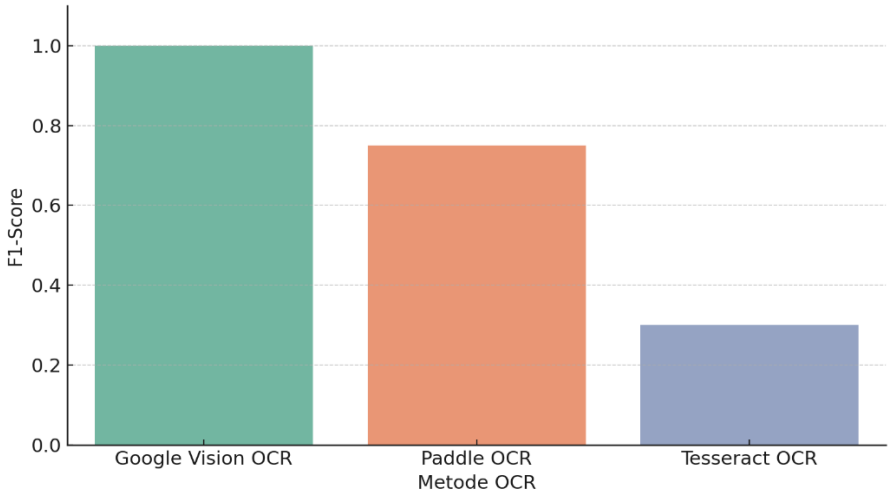

Figure 6. Average F1-Score Models

Figure 6 shows a comparison of the average F1-scores of the three tested OCR methods. Google Vision has the highest accuracy, followed by PaddleOCR, while Tesseract OCR shows much lower performance.

Performance Evaluation and Confusion Matrix

To test the system's sensitivity to variations in character similarity, two fuzzy ratios were used: 5 and 10. The results show that Google Vision OCR remains stable at both ratios, with the Precision and Recall values remaining consistent at 1.00.

In contrast, PaddleOCR saw Recall decrease from 0.70 to 0.60 when the fuzzy ratio increased from 5 to 10, although Precision remained high. This suggests that increasing the ratio makes the system more selective, but loses some relevant results. Tesseract OCR continues to perform poorly at both ratios, with Recalls of only 0.25 and 0.20.

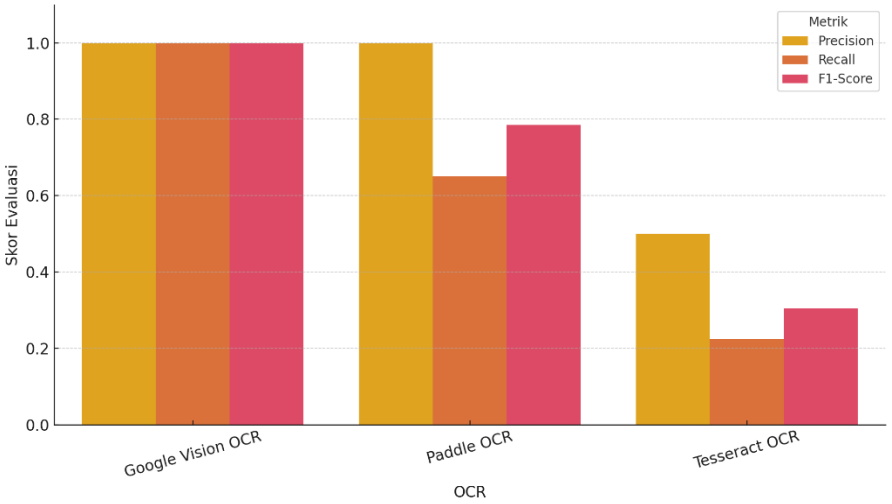

Figure 7. Comparison of Precision, Recall, and F1-Score Based on Fuzzy Ratio

Figure 7 shows the impact of changing the fuzzy ratio (5 and 10) on the three main evaluation metrics for each OCR model. It can be seen that only PaddleOCR experiences a decrease in Recall, while Google Vision remains stable.

Confusion matrix analysis shows that Google Vision OCR has very high TP and low FN, with few detection errors. PaddleOCR shows more FN on complex products, even though FP remains low. Tesseract OCR produces very high FN and more FP than the other two models.

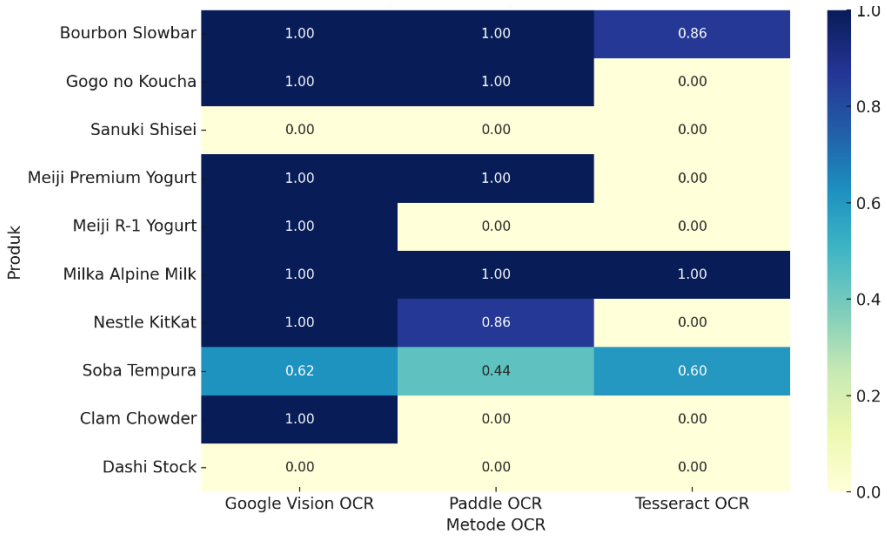

Figure 8. Heatmap F1-Score Models

This heatmap shows the distribution of F1-score values of each OCR method for each product. Google Vision dominates with the darkest color which is the highest value, while Tesseract looks weak in almost the entire product line.

Discussion

The results show that Google Vision OCR is the best model with the highest level of accuracy and stability in detecting allergens from Japanese food packaging. This model excels at recognizing Japanese characters even in less than ideal imaging conditions.

PaddleOCR shows quite good results, especially in detecting allergens with simpler characteristics. However, its inconsistency in some products decreases the overall Recall value.

Tesseract OCR is not recommended in this case because it does not support Japanese characters well and produces many false detections.

These findings support previous literature stating that the integration of OCR and fuzzy matching is effective in handling imperfect text. Fuzzy matching plays an important role in overcoming minor errors that arise during the OCR process, such as one-character differences or spelling variations.

This study makes a significant contribution to the development of an OCR-based automatic allergen detection system that can be used for consumer applications, especially for those with food allergies or certain dietary preferences. This system is also very helpful for foreign consumers who do not understand Japanese

4. Conclusion

The results show that the combination of OCR and fuzzy matching methods can be used effectively to detect allergens in Japanese packaged food products. From the evaluation results of 10 product samples, the Google Vision OCR and PaddleOCR methods showed high performance with precision, recall, and F1-score values reaching 1.00 in most cases, on the contrary, Tesseract OCR showed poor performance in this case and produced low or zero F1-Score in some cases.

This research proves that Google Vision OCR and PaddleOCR are more reliable in real-world conditions with lighting variations and complex text formats. These results support the development of automated food recommendation systems that are safe for consumers with specific preferences.

References:

[1] K. Sasaki, "Diversity of Japanese consumers' requirements, sensory perceptions, and eating preferences for meat," *Anim. Sci. J.*, vol. 93, no. 1, Jan. 2022, doi: 10.1111/ASJ.13705.

[2] K. Toratani, Ed., "The Language of Food in Japanese," *Converging Evid. Lang. Commun. Res.*, vol. 25, Jan. 2022, doi: 10.1075/CELCR.25.

[3] V. Grinkov, G. Grinkova, and S. Grinkov, "V. Hrinkov, G. Hrinkova, S. Hrinkov. Analysis of modern optical character recognition tools for character recognition and text from the image," *Sist. ì Tehnol. zv'âzku, ìnformatizacìï ta kìberbezpeki*, vol. 1, no. 6, pp. 75–84, Dec. 2024, doi: 10.58254/VITI.6.2024.05.75.

[4] S. Kavin and C. P. Shirley, "OCR-Based Extraction of Expiry Dates and Batch Numbers in Medicine Packaging for Error-Free Data Entry," *Proc. Int. Conf. Circuit Power Comput. Technol. ICCPCT 2024*, vol. 4, pp. 278–283, Aug. 2024, doi: 10.1109/ICCPCT61902.2024.10673325.

[5] J. Kalluru, "Enhancing Data Accuracy and Efficiency: An Overview of Fuzzy Matching Techniques," *Int. J. Sci. Res.*, vol. 12, no. 8, pp. 685–690, Aug. 2023, doi: 10.21275/SR23805184140.

[6] S. Kayalvizhi, N. Akash Silas, R. K. Tarunaa, and S. Pothirajan, "OCR-Based Ingredient Recognition for Consumer Well-Being," *Lect. Notes Networks Syst.*, vol. 796, pp. 481–491, Jan. 2024, doi: 10.1007/978-981-99-6906-7_41.

[7] K. Riehl, M. Neunteufel, and M. Hemberg, "Hierarchical confusion matrix for classification performance evaluation," Jun. 2023, doi: 10.1093/jrsssc/qlad057.

[8] C. K. FÆste, H. T. Rønning, U. Christians, and P. E. Granum, "Liquid chromatography and mass spectrometry in food allergen detection.," *J. Food Prot.*, vol. 74, no. 2, pp. 316–345, Feb. 2011, doi: 10.4315/0362-028X.JFP-10-336.

[9] C. Thorat, A. Bhat, P. Sawant, I. Bartakke, and S. Shirsath, "A Detailed Review on Text Extraction Using Optical Character Recognition," *Lect. Notes Networks Syst.*, vol. 314, pp. 719–728, Jan. 2022, doi: 10.1007/978-981-16-5655-2_69.

[10] "(20+) Facebook." Accessed: Apr. 14, 2025. [Online]. Available: https://www.facebook.com/HalalJapanOfficial/

[11] "【楽天市場】Shopping is Entertainment! ： インターネット最大級の通信販売、通販オンラインショッピングコミュニティ." Accessed: Apr. 14, 2025. [Online]. Available: https://www.rakuten.co.jp/

[12] J. R. Fonseca Cacho and K. Taghva, "Aligning Ground Truth Text with OCR Degraded Text," *Adv. Intell. Syst. Comput.*, vol. 997, pp. 815–833, Jul. 2019, doi: 10.1007/978-3-030-22871-2_58.

[13] J. Ghorpade-Aher, S. Gajbhar, A. Sarode, G. Gayake, and P. Daund, "Text Retrieval from Natural and Scanned Images," *Int. J. Comput. Appl.*, vol. 133, no. 8, pp. 10–12, Jan. 2016, doi: 10.5120/IJCA2016907840.

[14] N. P. T. Prakisya, B. T. Kusmanto, and P. Hatta, "Comparative Analysis of Google Vision OCR with Tesseract on Newspaper Text Recognition," *Media Comput. Sci.*, vol. 1, no. 1, pp. 31–46, Jul. 2024, doi: 10.69616/MCS.V1I1.178.

[15] O. Krasynskyi and O. Markovets, "Possibilities of Using OCR Technologies from Google for Recognition and Digitalization of Archive Documents," *Vìsnik Harkìvs'koï deržavnoï Akad. kul'turi*, no. 65, pp. 227–237, Jun. 2024, doi: 10.31516/2410-5333.065.16.

[16] "PaddlePaddle-Parallel Distributed Deep Learning, efficient and extensible deep learning framework." Accessed: Apr. 14, 2025.

[17] P. Sharma, "Advancements in OCR: A Deep Learning Algorithm for Enhanced Text Recognition," *Int. J. Inven. Eng. Sci.*, vol. 10, no. 8, pp. 1–7, Aug. 2023, doi: 10.35940/IJIES.F4263.0810823.

[18] U. K. V. Karanth, A. T. Sujan, T. Y. R. Kumar, S. S. Joshi, A. K. P. Rani, and S. Gowrishankar, "Breaking

Barriers in Text Analysis: Leveraging Lightweight OCR and Innovative Technologies for Efficient Text Analysis," *2nd Int. Conf. Autom. Comput. Renew. Syst. ICACRS 2023 - Proc.*, pp. 359–366, Dec. 2023, doi: 10.1109/ICACRS58579.2023.10404305.

[19] "GitHub - tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository)." Accessed: Apr. 14, 2025. [Online]. Available: https://github.com/tesseract-ocr/tesseract

[20] M. M. Rahman and M. R. Rinty, "Text Information Extraction from Digital Image Documents Using Optical Character Recognition," *Comput. Intell. Image Video Process.*, pp. 1–31, Jan. 2023, doi: 10.1201/9781003218111-1.

[21] V. E. Bugayong, J. Flores Villaverde, and N. B. Linsangan, "Google Tesseract: Optical Character Recognition (OCR) on HDD / SSD Labels Using Machine Vision," *2022 14th Int. Conf. Comput. Autom. Eng. ICCAE 2022*, pp. 56–60, Mar. 2022, doi: 10.1109/ICCAE55086.2022.9762440.

[22] M. K. Audichya, "A Study to Recognize Printed Gujarati Characters Using Tesseract OCR," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. V, no. IX, pp. 1505–1510, Sep. 2017, doi: 10.22214/IJRASET.2017.9219.

[23] Thangam, U. Kumaran, D. Biswas, B. Sneha, S. Nadipalli, and S. Raja, "Text Post-processing on Optical Character Recognition output using Natural Language Processing Methods," *2023 IEEE 3rd Mysore Sub Sect. Int. Conf. MysuruCon 2023*, pp. 1–6, Dec. 2023, doi: 10.1109/MYSURUCON59703.2023.10396964.

[24] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of Post-OCR Processing Approaches," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–37, Jul. 2021, doi: 10.1145/3453476.

[25] Sanjay Kumar Gorai and Shekhar Pradhan, "Bridging the Gap: OCR Techniques for Noisy and Distorted Texts," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 1, pp. 695–703, Jan. 2025, doi: 10.32628/CSEIT2511111.

[26] "R-Vogg-Blog: Fuzzy string matching." Accessed: Apr. 20, 2025

[27] "Character string fuzzy matching method and apparatus," Oct. 12, 2016. Accessed: Apr. 15, 2025.

[28] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification," Jan. 2020, Accessed: Nov. 20, 2024.

[29] O. Caelen, "A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell.*, vol. 81, no. 3–4, pp. 429–450, Dec. 2017, doi: 10.1007/S10472-017-9564-8/METRICS.

[30] "Food package detection method," Jul. 27, 2016. Accessed: Apr. 15, 2025.