International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Machine Learning Model for Optical Character Recognition-Based Food Allergen Detection with Recommendation System for Alternative Food

**Rugved Borade, Arishi Gupta, Anupriya Kathpalia, Tanish Jain, Priti Chakurkar**

**Submitted**: 06/02/2024 **Revised**: 14/03/2024 **Accepted**: 20/03/2024

**Abstract:** In today's diverse and fast-paced food industry, ensuring consumer safety and meeting specific dietary needs is of paramount importance. Food allergen detection and recommendation systems have emerged as crucial tools to address these concerns. This project aims to create an innovative OCR based solution for automating the identification of allergenic ingredients on food packaging labels. By combining Optical Character Recognition (OCR) technology with a comprehensive allergen database, the system will provide real-time allergen information to consumers. Moreover, it will recommend suitable food alternatives for individuals with specific dietary restrictions, enhancing their shopping experience and reducing the risk of allergen-related incidents. The allergen knowledge base is implemented using several machine learning algorithms and will be updated constantly. This holistic approach not only promotes food safety but also empowers consumers to make informed choices, fostering a healthier and more inclusive food environment.

***Keywords:*** *OCR, Regular Expressions, Recommendation, Database, Food packages, Allergies, Cosine Similarity*

## 1. Introduction

One fruit that we are all familiar with is the pear, which contain Pyridoxine Hydrochloride (B6), which can cause allergies. Additionally, the cuisine of a foreign country whose ingredients. Unknown to us, the composition of proteins might result in allergy symptoms. Likewise, a small amount of preservative in an edible safe food packaging may trigger allergies. In certain instances, as well. Preservatives and the contents of the package can interact triggering an allergic response. Food recommendation systems are designed to offer suggestions based on the diet, cooking methods, and personal preferences of the user. These systems are said to be useful in assisting individuals in changing their eating patterns in order to embrace a nutritious diet that meets their preferences. The majority of earlier food recommendation systems have limited their capacity to produce appropriate recommendations by failing to consider the nutrition and health benefits of foods. We recommend an ingredient substitution approach for cooking that is safe for allergy sufferers, based on the food content derived based on resemblance to the recipe. The accuracy of removing allergy-free substances can be increased by taking the meal composition into account. Food ingredients that may trigger an individual's allergy can be found in packaged food using optical character recognition. Many details are provided, often in scientific terms that are hard for the average person to understand, in the image of the Ingredient Index on the package which can be recognized and processed by the system.

*School of Computer Engineering and Technology*
*MITWPU, Pune, India*
*E-mail: flakjacket015@gmail.com, guptaarishi18@gmail.com, anupriya.kathpalia@gmail.com*
*, tanishjain1023@gmail.com, priti.chakurkar@mitwpu.edu.in*

## 2. Motivation

The motivation behind the problem statement is rooted in the increasingly prevalent issue of food allergies and dietary restrictions in today's world. With a growing number of individuals affected by these conditions, there is a pressing need to provide them with a robust and efficient system to make informed and safe food choices. This issue is further compounded by the health and safety concerns associated with food allergies, which can lead to severe health complications, including life-threatening anaphylactic reactions.

Food allergies are said to be a condition that gets worse with time. Although it is widely acknowledged that 2.5% to 5% of the general population suffers from food allergies, prevalence statistics varies greatly, spanning from 1% to 10%. Given that a variety of factors affect the reported incidence of food allergies, accurately determining their prevalence remains one of the main challenges associated with the condition. [1].

According to epidemiological research, about 80% of infants with food allergies will become tolerant by the time they turn five, but 35% of them may later become hypersensitive to other foods. Food allergies are less likely to worsen among those with the highest IgE levels, the most severe clinical symptoms (such as asthma and anaphylaxis), and the broadest co-sensitizations. Children who are allergic to milk and eggs have a better prognosis than those who are allergic to peanuts, tree nuts, and fish. The natural history of food allergies also relies on the specific food sensitivity. [2]. The distribution for different types of allergies is provided in Fig 1.

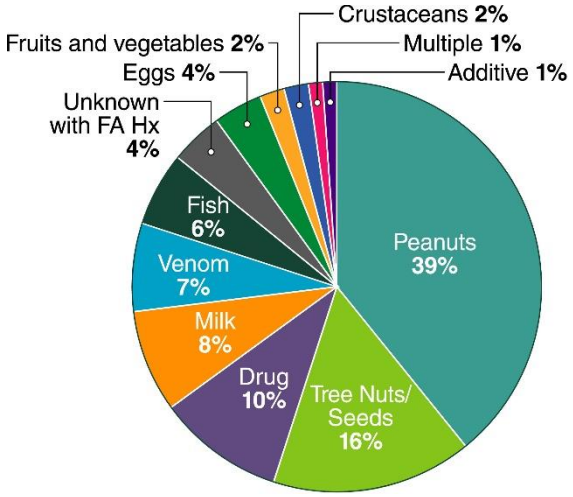


**Figure 1.** According to a study conducted on all admissions for anaphylaxis to pediatric intensive care units in North America (the United States, Canada, and Mexico) between 2010 and 2015 (N = 1989), peanuts were the most often reported trigger. [3]

Existing solutions may fall short in providing a comprehensive and user-friendly approach to this problem, emphasizing the necessity for a more effective solution. The inclusion of Optical Character Recognition (OCR) technology in the proposed system reflects the role of technological advancements in addressing this issue. Moreover, this problem statement underscores the importance of empowering individuals with food allergies to take control of their dietary choices and provides suitable alternative product recommendations, ensuring a more holistic approach to managing food allergies and dietary restrictions. Overall, the motivation for this problem statement lies in improving the lives and safety of those affected by food allergies while embracing technological advances to enhance their food-related decision-making.

Also, individuals with food allergies encounter challenges due to the scarcity of allergen-free options in the market and the inconsistency of stores in stocking such products. This results in a time-consuming process for people with food allergies, who often spend significant time scrutinizing ingredient lists at stores to ensure the safety of the products. However, they remain uncertain whether the store even offers allergen free alternatives. When unable to find suitable products in a particular store, individuals must venture to another store with the same uncertainty. To address this issue, we have developed an application that streamlines the process, allowing users to easily locate allergen-free foods. This innovative solution eliminates the need for individuals to waste time searching for safe food options.

## 3. Literature Review

Food allergies have become a serious global issue, affecting a high number of both children and adults. One of the most common causes of anaphylaxis is food allergies. Food allergies have grown more common in the previous two decades, according to research. Food allergies are recognized as a major public health hazard around the world, including India [4].

Thus, it is critical to assist people in identifying allergic foods and avoiding anaphylactic reactions. Allergic substances are prohibited by law in over 66 countries when included in prepared foods.

Unfortunately, the mandated allergen list is not consistent between countries. Allergy labelling confronts numerous obstacles, including faulty labelling, ambiguous terminology, and varied labelling regulations and compliance, resulting in an inaccurate allergen list [5].

To overcome this problem, many solutions have been proposed which use smartphone-based biosensors for portable food evaluation. These biosensors are portable devices that detect and analyse analytes using biological recognition elements and transducers. They offer convenience and real-time analysis for on-the-spot meal evaluation. They do, however, have drawbacks such as decreased sensitivity, a limited analyte range, and potential accuracy concerns due to ambient influences and user dependence. Cost and standardization issues also exist [6].

Given these drawbacks, an OCR-based allergy detection system is considered as a more viable option. OCR is a subset of image recognition that is used in software to recognize text from scanned documents or photos.

Ref. [7] helped us understand the working of Tesseract OCR and how it can help us extract text from images. Free version of Tesseract OCR has an average error rate of about 67% for a global dataset of images which shows that there is still a lot of scope left to improve the working of Tesseract OCR Engine for English Language but considering it's lightweight and its integration with Artificial Intelligence can help increase its accuracy immensely, it becomes an ideal option for an allergen detection system.

Ref. [8] talked about different benchmarking experiments comparing the performance of Tesseract, Amazon Textract and Google document AI on images of English text. Furthermore, it helped us understand what type of inputs to these tools gave more errors like certain types of "integrated" noise, such as blur and salt and pepper generated more error than "superimposed" noise such as watermarks, scribbles, and ink stains.

Ref. [9] helped us understand the process of developing an Allergen Detection and Recommendation System Using OCR. It gave us a brief idea about system architecture of an OCR based Allergen detection system. They used Pytesseract, an OCR module in python for text extraction and cre-

ated a web-based application which took a picture as an input and returned the result on the screen.

Some of the challenges that may occur when designing such a kind of system are as:

- As different manufacturers may present ingredient information in diverse formats. There is variability in food packaging and labelling practices. The OCR technology should be capable of accurately extracting and interpreting data from a wide range of packaging styles.
- Crafting a system that accommodates various allergen sensitivities and preferences requires robust algorithms for allergen detection. These algorithms must be continually refined to enhance accuracy and keep pace with the dynamic nature of food product formulations.
- The system must be capable of discerning complex ingredient interactions and accurately identifying potential allergens.
- The system must consider individual preferences, dietary restrictions, and allergen sensitivities to offer meaningful and personalized suggestions.

## 4. System Design

A user-friendly website for allergen detection is developed, tailored to each person's requirements, to identify and alert users to the existence of allergens and their nutritional information. It also recommends alternatives using the recommendation system. The architecture overview is provided by Fig 2.

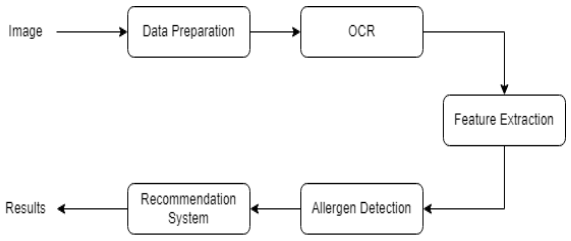


**Figure 2.** Block diagram of the system

### 4.1 Optical Character Recognition

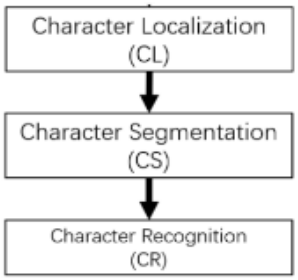


**Figure 3.** OCR Processing Steps

| Feature | Tesseract | OCRopus | Adobe Acrobat OCR | Microsoft OCR(MODI) |
|---|---|---|---|---|
| Open Source | Yes | Yes | No | No |
| Language Support | Extensive | Extensive | Limited | Limited |
| Community Support | Active | Active | Limited | Limited |
| Accuracy | High | Good | Good | Moderate |
| Layout Handling | Good | Varies | Good | Varies |
| Customizable | Yes | Yes | Limited | Limited |
| Integration | CLI, API | CLI | Integrated with acrobat | Integrated with office |
| Platform Compatibility | Linux,Windows,MacOS | Linux,Windows,MacOS | Windows | Windows |
| Cost | Free | Free | Part of acrobat | Part of office |

**Table 1.** Comparison of different OCR Libraries

Python Pytesseract is the OCR tool that is utilised; it features segmentation, rescaling, de-skewing, binarization, and noise reduction. Features Extraction, SVM Classification, and Recognition come next. An OCR model that is remotely deployed is used to identify and give the allergy facts from an image of the ingredient section of packaged food [9].

After careful evaluation as given in Table 1, Pytesseract is chosen because Tesseract is often favored over other OCR tools for several reasons. As an opensource solution backed by a robust community, Tesseract is freely accessible, making it a cost-effective choice for developers and businesses. Its strong language support, adaptability through customization and training, competitive OCR accuracy, and ease of integration into various programming languages contribute to its popularity. Additionally, the continuous community-driven development ensures regular updates and improvements, enhancing Tesseract is versatile for a wide range of applications, including those with multilingual requirements or specific domain-focused language models. It is a crucial step in ensuring accurate and reliable text extraction from images. Proper data preparation helps the OCR engine understand and recognize elements in the image. then comes image scanning which ensures that the source document or image is scanned at an appropriate resolution to capture fine details of characters. Fig 3 shows the different steps in OCR which are elaborated below.

### 4.2 Pre-Processing of OCR

OCR is carried out using the Tesseract OCR library, as was previously indicated. Prior to character recognition, preprocessing operations such as picture rescaling, skew correction, noise reduction, and binarization must be completed. The Tesseract OCR library also has these routines, which are based on the Numpy and OpenCV utilities. After that, the pre-processed image is transmitted to OCR in order to extract the editable text. Numerous superfluous elements, like numerals, special letters, and punctuation, are present in the OCR result. In the subsequent post-processing stage, these will be eliminated [10].

### 4.3 Post-Processing of OCR

Post-OCR processing is typically applied to the results obtained after performing OCR on a document or image, which yields the recognized text. In this case, the provided text describes the preparation steps before OCR, including noise removal, image rescaling, skew correction, and binarization, as well as the intention to perform post-processing to clean the OCR output. To mitigate the impact of noise and enhance the accuracy of ingredient recognition, we employed two postprocessing methods on the OCR output: tokenization and correction/extraction using dictionaries. Tokenization involves breaking down OCR-retrieved strings into word tokens while eliminating irrelevant symbols. Subsequently, the tokens undergo comparison with both an English dictionary and a proprietary dictionary. The English dictionary rectifies potential unrecognized word unit tokens, while the proprietary dictionary, specific to our domain, filters out only relevant words, namely ingredients in our study.

### 4.4 Feature Extraction and Allergen Detection

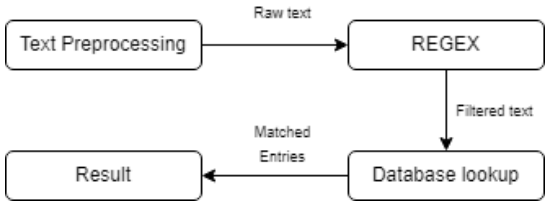


**Figure 4.** Feature extraction output and database used for allergen detection.

As shown by Fig 4, After the OCR input is received, features such as ingredients, protein content, fats content, etc. are extracted from the input text using regular expressions.

A regular expression, also known as a rational expression, is a string of letters that indicates a pattern of matches in text. It can be shortened as regex or regexp. Typically, string searching algorithms use these patterns for input validation or for ”find” or ”find and replace” actions on strings. Formal language theory and theoretical computer science are fields that develop regular expression techniques [11].

Let’s say we have a sentence “In the beginning God created the heaven and the earth”, and our regex pattern is “/h[aeiou]+/g” (the letter h followed by one or more vowels), it will match the “he” in the second, seventh and ninth words along with “hea” in the word heaven[11].

| ID | Name | Category |
|---|---|---|
| 1 | Milk | Lactose Intolerant |
| 2 | Milk Solids | Lactose Intolerant |
| 3 | Cheese | Lactose Intolerant |
| 4 | Cream | Lactose Intolerant |
| 5 | Custard | Lactose Intolerant |

**Table 2.** Different food items in the same allergy category in the database.

Likewise, Specific patterns can be created using regular expressions for detecting protein or other nutrients content using regular expressions. However, Regular expressions are very exact and cannot guarantee 100% robustness. If the OCR input missed a certain keyword which is present in the regex, the regex will not detect the text.

After regex processing is done, the output is provided in JSON format and then the allergen detection module uses these and the database of various food items(sample provided in Table 2) to determine the closest foodstuff it resembles and then the allergies of that foodstuff are also added in the case the allergies detected from ingredients did not provide enough information. All this is done by using SQL queries on a stored database and the matches are then queried for allergies in the database.

### 4.5 Alternate Food item Recommendation using ML

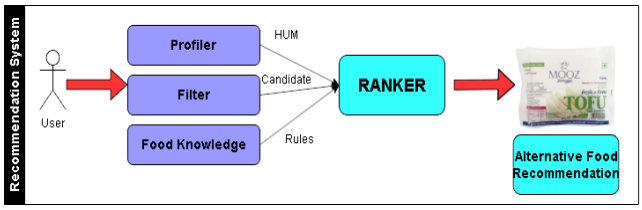


**Figure 5.** Recommendation system flow

The recommendation system is based on NLP using TF-IDF vectorization and finding ingredients similarity using cosine similarity from a database of filtered allergen free items.

The name of the dataset used is "Food Ingredient Lists" which was available on Kaggle (https://www.kaggle.com/datasets/datafiniti/food-ingredientlists). It consists of 10,000 rows and 15 columns. The dataset has columns like: Category, Brand, Ingredients, etc. In this particular project, we have used this dataset to classify the allergies into corresponding groups, Cruciferous, Milk OR lactose intolerance, Fish Allergy, Oral / Skin Allergy. The above identified categories were then transformed into their more generic forms like, Fish: Meat Based, Milk / Lactose: Dairy Based, etc.

The system is fed a list of non-allergic products and the ingredient list of the product that the recommendation needs to be generated for, using TF-IDF and cosine similarity the products are then ranked on how similar they are to the item scanned and the three most similar allergen free alternatives are recommended. The architecture is present in Fig 5.

According to the principle of cosine similarity, the degree of document similarity decreases with the size of the angle formed between two coordinate vector comparison documents. On the other hand, if the cosine similarity level is lower, the degree of similarity will be higher [12].

The TF-IDF vectorization is applied to the 'features.value' column using TfidfVectorizer from scikit-learn. Cosine similarity is calculated between the TF-IDF matrix and itself using linear kernel. The function identifies the index of the new row added to the Data Frame and retrieves the cosine similarity scores for that row. The top 3 similar products are selected based on their similarity scores. Information about these recommended products (name, ingredients, category, and manufacturer) is stored in a list of python dictionaries.

Finally, the system generates a result that indicates whether the food product contains any allergens, and if so, which ones. The result can be displayed to the user on their device, alerting them to the presence of any potential allergens in the food product.

The efficiency of an allergen-free alternative recommendation system depends on the quality and diversity of available data, paired with a sophisticated Natural Language Processing (NLP) machine learning model. This synergy enables the model to accurately identify allergens, understand ingredient relationships, and suggest suitable alternatives based on individual preferences, enhancing the system's precision and adaptability.

### 4.6 Deployment of the model

Python's Flask web framework was utilized to upload the photos to the application. HTML and ReactJS are used to develop a web application that allows users to input photographs and shows the relevant results along with interactions with the backend, which include:

- Allergen details present in the food packet.
- Nutrients of the food items
- Alternatives for the allergens found.

## 5. Results

With respect to detection in [9] the authors analysed a custom-made dataset with images of fruits and vegetables collected along with a custom-made dataset with allergen and nutritional facts, the study applied 7 different CNN models and the best performing models identified were VGG18 and VGG 19 with an average accuracy of 95.9%.

In [14] CNN was used for character recognition, the average accuracy noted was 93.15%. In our study, the average detection accuracy was noticed in the range of 90-95% with Tesseract OCR(Table 3). The detection accuracy of regex is dependent on the OCR library of choice as it does pattern and keyword matching on the OCR output and can give up to 100% accuracy.

The success of any recommendation system depends largely on its ability to represent user's current interests [13]. In [15] the authors proposed a novel model called PFoodReQ which when used without recipe similarity gave an F1 Score of 32.6 and with recipe similarity gave an F1 score of 36.6.

Since a recommendation system is subjective, manual testing was done to determine quality of recommendations. In our methodology, on manually checking for good or bad recommendations it was found that 86 % of recommendations were good recommendations and 14 % of recommendations were bad recommendations (Table 4). Where good recommendations were alternative products that the user was not allergic to, and bad recommendations were alternative products that the user was allergic to (Figure 6).

The novelty in our study is given by the fact that it integrates both detection and recommendation systems.


**Figure 6.** Quality of recommendations provided by the system.

| Ref no. | Accuracy | Method |
|---|---|---|
| [9] | 95.9% | VGG19 with custom dataset |
| [14] | 93.15% | CNN |
| Our Study | 90-95% | Tesseract OCR and Regex |

**Table 3.** Comparative analysis for detection

| Ref no. | Metric | Value | Method |
|---|---|---|---|
| [15] | F1 Score | 32.6 | PFoodReQ |
| [15] | F1 Score | 36.6 | PFoodReQ + Recipe Similarity |
| Our Study | Accuracy | 86% | TfIDF Vectorization |

**Table 4.** Comparative analysis for recommendation

## 6. Conclusion

Food allergies are a significant public health issue that affects millions of people worldwide. Accidental exposure to allergens can lead to serious health issues, including anaphylaxis, which can be life-threatening. Therefore, the development of an Allergen Detection and Recommendation System is crucial to improving the safety of food products and preventing such incidents.

An Allergen Detection and Recommendation System can provide assurance to individuals with food allergies by accurately identifying the presence of allergens in food products. This information allows individuals to make informed choices about what they eat, which can help reduce the risk of an allergic reaction. Moreover, an Allergen Detection and Recommendation System can contribute to enhanced public health outcomes by reducing the risk of allergic reactions and related health issues. By complying with regulations and accurately labelling food products, the food industry can help prevent allergic reactions and improve public health.

The Potential improvements which can be done to this system are a functionality to contribute to database of the system by the user themselves by adding new food products and their information, the system can also factor in cost similarity when recommending products, we can add NLP and machine learning to replace regex pattern matching and allergen detection to get more targeted results.

In conclusion, the successful development and implementation of an Allergen Detection and Recommendation System can have a significant and positive impact on individuals, communities, and society.

## References

[1] Fiocchi A, Fierro V. Food Allergy [Internet]. World Allergy Organization. 2017 [cited 2024 Jan 9]. Available from: https://www.worldallergy.org/education-and-programs/education/allergic-disease-resource-center/professionals/food-allergy

[2] Kattan J. The Prevalence and Natural History of Food Allergy. Current Allergy and Asthma Reports. 2016 Jun 22;16(7).

[3] Lieberman JA, Gupta R, Knibb RC, Haselkorn T, Tilles S, Mack DP, et al. The Global Burden of Illness of Peanut Allergy: A Comprehensive Literature Review. Allergy. 2020 Nov 20;76(5).

[4] Li J, Ogorodova LM, Mahesh PA, Wang MH, Fedorova OS, Leung TF, et al. Comparative Study of Food Allergies in Children from China, India, and Russia: The EuroPrevall-INCO Surveys. The Journal of Allergy and Clinical Immunology: In Practice [Internet]. 2019 Dec; Available from: https://www.sciencedirect.com/science/article/abs/pii/S2213219819310323

[5] Fiocchi A, Risso D, DunnGalvin A, González Díaz SN, Monaci L, Fierro V, et al. Food labeling issues for severe food allergic patients. World Allergy Organization Journal. 2021 Oct;14(10):100598.

[6] Lu Y, Shi Z, Liu Q. Smartphone-based biosensors for portable food evaluation. Current Opinion in Food Science. 2019 Aug; 28:74–81.

[7] Joshi K. Study of Tesseract OCR. GLS KALP – Journal of Multidisciplinary Studies [Internet]. 2021 Mar 28 [cited 2024 Jan 9];1(2):41–51. Available from: https://glskalp.in/index.php/glskalp/article/view/9

[8] Hegghammer T. OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. Journal of Computational Social Science. 2021 Nov 22;

[9] B. Rohini, Divya Madhuri Pavuluri, LS Naresh Kumar, V Soorya, J Aravinth. A Framework to Identify Allergen and Nutrient Content in Fruits and Packaged Food using Deep Learning and OCR. 2021 Mar 19;

[10] Kamis P, Ok Keun Shin. OCR-Based Safety Check System of Packaged Food for Food Inconvenience Patients. Dijiteolkeontencheuhakoenonmunji. 2020 Jun 30;21(6):1025–32.

[11] Wikipedia contributors. Regular expression [Internet]. Wikipedia. Wikipedia, The Free Encyclopedia; 2024 [cited 2024 Jan 9]. Available from: https://en.wikipedia.org/wiki/Regular_expression

[12] Lahitani AR, Permanasari AE, Setiawan NA. Cosine similarity to determine similarity measure: Study case in online essay assessment [Internet]. IEEE Xplore. 2016 [cited 2021 Jun 2]. p. 1–6. Available from: https://ieeexplore.ieee.org/abstract/document/7577578

[13] Shah, Priyanshi Sanghvi, Shikhar. Video Recommender System. 2020

[14] F. De Sousa Ribeiro et al., "An End-to-End Deep Neural Architecture for Optical Character Verification and Recognition in Retail Food Packaging," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 2018, pp. 2376-2380, doi: 10.1109/ICIP.2018.8451555.

[15] Chen, Y., Subburathinam, A., Chen, C.H. and Zaki, M.J., 2021, March. Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (pp. 544-552).