OCR and NLP based Personalized Allergen Notifying System

Sahana Mahesh, Shrivarshan N K, Anbuchelvan A P
School of Computer Science Engineering and Information Systems, SCORE
Vellore Institute of Technology
Vellore, Tamil Nadu, India

Abstract-This project focuses on developing a personalized allergen notification system to help individuals identify potential allergens in food products based on their dietary preferences and known allergens. A key feature of this system is Optical Character Recognition (OCR), which enables users to scan ingredient lists directly from product packaging. The OCR technology extracts text from images of ingredients, which is then processed using natural language processing (NLP) and machine learning algorithms to compare the extracted text against the user's allergen profile. The system provides real-time notifications of potential allergens, offering an efficient and portable solution for allergen detection and food safety.

Keywords: Food Allergy Detection, OCR, NLP, Personalized Allergen Notification, Machine Learning

I. INTRODUCTION

Food allergies and intolerances pose significant health risks to individuals, requiring meticulous monitoring of food ingredients to avoid potential allergens. Studies indicate that even trace amounts of allergens in food products can trigger severe allergic reactions, highlighting the critical importance of accurate allergen detection systems. Research shows the prevalence of food allergies is increasing, particularly among children, with reported rates ranging from 5-8% in children and 1-2% in adults. These conditions may result in severe allergic reactions, including life-threatening anaphylaxis, underscoring the need for effective allergen detection and management tools.

Traditional methods of manually reviewing ingredient lists are time-consuming and prone to human error, especially given the variations in ingredient nomenclature across brands and regions. Advancements in Optical Character Recognition (OCR), Natural Language Processing (NLP), and machine learning technologies provide promising solutions for improving allergen detection. OCR facilitates automated text extraction from images, enabling real-time scanning of food product labels. When combined with NLP techniques like Named Entity Recognition (NER) and fuzzy matching algorithms, these systems can identify ingredients and match them against personalized allergen profiles with high accuracy.

This paper presents the design, implementation, and evaluation of a personalized allergen notification system that leverages OCR, NLP, and machine learning technologies to provide real-time alerts about potential allergens in food products. The system addresses the limitations of manual allergen identification while enhancing portability and user convenience.

II. LITERARY REVIEW

Advancements in OCR technology have enabled automated text extraction with high accuracy. Smith (2007) highlighted OCR’s potential in document analysis, while Jiang and Stroud (2018) emphasized its application under challenging image conditions. NLP techniques, such as those described by Jurafsky and Martin (2021), enable ingredient categorization and matching to allergen profiles.

Recent advancements also emphasize the importance of large-scale allergen datasets. Lee et al. (2023) demonstrated the integration of analytical methods with allergen management tools, improving both accuracy and usability in diverse food environments. Similarly, innovative strategies, such as combining OCR with real-time mobile applications (Li et al., 2019), show potential for seamless implementation, making allergen detection accessible to everyday users. This convergence of OCR, NLP, and machine learning forms the foundation for the proposed system, ensuring robust allergen identification while enhancing user convenience.

In the context of allergen detection, fuzzy matching ensures reliable identification of ingredients even when manufacturers use varying nomenclature or regional terminology. Recent studies, such as Lee et al. (2023), focus on analytical methods for allergen control in food processing, demonstrating how algorithms like RapidFuzz can detect and flag potential allergens in diverse datasets. Thereby improving user safety and confidence.

III. METHODOLOGY

The system architecture from Fig 1. consists of the following components. In total the proposed strategy follows 5 main phases to provide the notification. These steps are vital for the system to work with a higher level of accuracy.

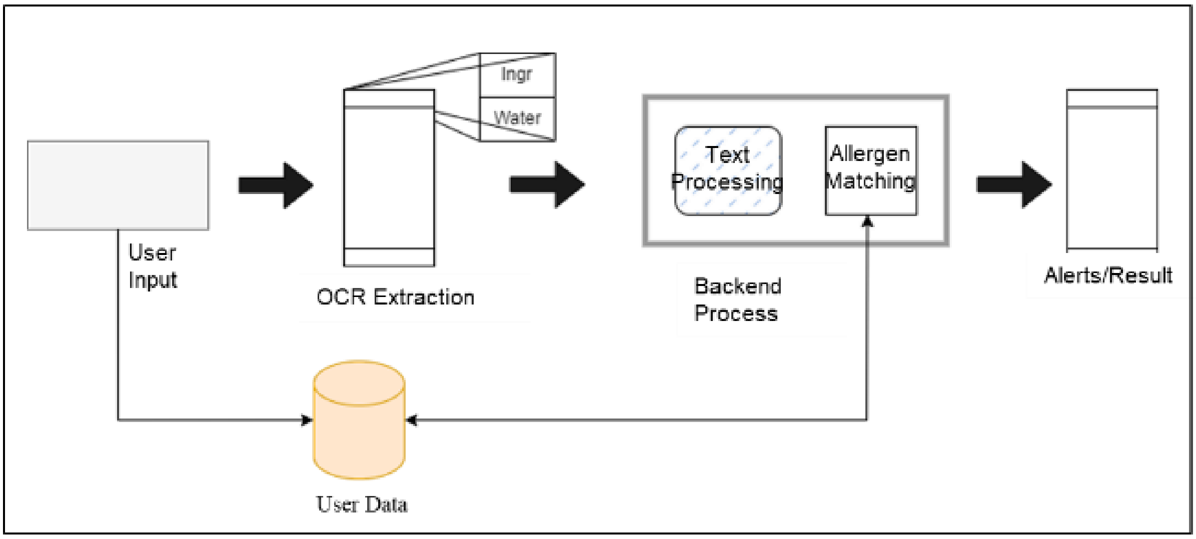


Fig 1. Proposed Architecture

A. Input Phase: User Profile Setup

Users create or update allergen profiles by entering known allergens (e.g., peanuts, gluten, dairy) into the system. Profiles are securely stored in a backend database and form the basis for ingredient analysis.

B. Ingredient Scanning Phase: OCR Extraction

Users capture an image of the ingredient list on food product packaging using a smartphone camera. OCR algorithms, such as EasyOCR, extracts text from the image, converting it into machine-readable format. Preprocessing steps include noise removal, lowercasing, and tokenization to standardize the text.

C. Ingredient Parsing Phase: NLP and Matching

NLP techniques like Named Entity Recognition (NER) identify and categorize ingredients. Fuzzy matching algorithms (e.g., RapidFuzz) match extracted ingredients against the user's allergen profile, considering variations in nomenclature. Cross-reactivity is accounted for by flagging related allergens (e.g., tree nuts for peanut allergies).

D. Allergen Detection Phase: Analysis and Notification

Detected matches are flagged, and confidence scores are computed using Levenshtein distance or minimum edit distance. Users receive real-time alerts via smartphone or smartwatch notifications.

E. *User Interaction Phase: Response and Updates*

Users acknowledge notifications and decide whether to avoid the product. Newly discovered allergens can be added to the profile for future detection.

IV. IMPLEMENTATION

This section provides a detailed explanation of creating a viable dataset for the project, the functioning and integration of OCR technology, and the role of matching algorithms for allergen detection from a local database.

A. Dataset Creation

The success of the system relies heavily on a robust and comprehensive dataset that serves as the backbone for allergen detection. There are few steps that have been followed to build this dataset. These steps to create the dataset are provided in the following Fig 2.

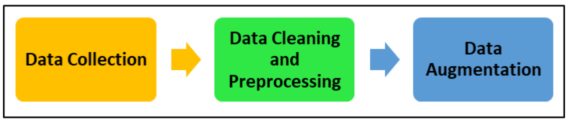


Fig 2. Data Creation workflow

a. *Data Collection:*
   Public allergen datasets, such as those provided by the FDA or FARRP, are used as a foundation. Manufacturers' ingredient lists and food labels are scraped and preprocessed. User-generated data (e.g., user-flagged allergens) is incorporated to enhance personalization.
b. *Data Cleaning and Preprocessing:*
   Noise and duplicates are removed. Synonyms and regional variations of ingredient names are normalized (e.g., "milk powder" vs. "powdered milk").
c. *Data Augmentation:*
   Incorporated multilingual variations (e.g., "leche" for "milk" in Spanish). Cross-reactivity is expanded with references from medical allergen studies.

B. Allergen Dataset

The dataset is structured with the following necessary fields:

a. *Ingredient Name:* Standardized names of ingredients (e.g., "peanuts," "gluten," "milk protein").
b. *Allergen Category:* Classification of allergens (e.g., tree nuts, dairy, gluten).
c. *Cross-Reactivity Data:* Information on related allergens that could trigger reactions (e.g., tree nuts for peanut allergies).
d. *Ingredient Variations:* Alternate names or regional terminologies (e.g., "casein" for "milk protein").
e. *Confidence Scores*: Weightage or severity levels associated with allergens.

C. Optical Character Recognition (OCR)

OCR technology is integral to the system, enabling the extraction of text from food product labels captured by the user. Following diagram, Fig 3 explains EasyOCR's working and application in the project.

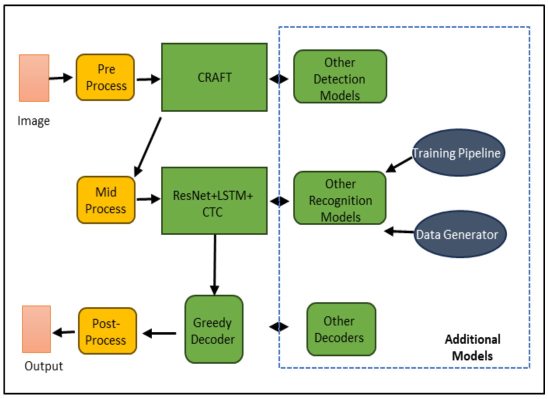


Fig 3. EasyOCR Framework

a. *Image Preprocessing:*
The input image undergoes resizing, binarization, and denoising to improve recognition accuracy. Techniques like Gaussian blur are used to reduce noise, while adaptive thresholding handles varying lighting conditions.

b. *Text Detection:*
OCR engines like EasyOCR identify regions of interest (ROIs) containing text. Connected Component Analysis (CCA) or Machine Learning-based models help isolate text blocks.

c. *Text Recognition:*
The identified regions are converted into character strings using pattern recognition and language modeling. OCR engines employ recurrent neural networks (RNNs) or convolutional neural networks (CNNs) to handle varying fonts and styles.

D. OCR Integration in the Project:

a. *Input Phase:*
Users capture the ingredient list using a smartphone camera.

b. *Processing Phase:*
The captured image is fed to the OCR engine, which extracts text and converts it to machine-readable format.

c. *Post-OCR Cleaning*:
The output text is preprocessed to remove irrelevant characters, correct spelling errors, and tokenize the ingredient list.

E. Matching Algorithms

The extracted text from the OCR process is compared against the local allergen database to identify potential allergens. This section elaborates on the matching process.

a. *Exact Matching:*
A direct comparison is made between the extracted ingredient and entries in the allergen database.

b. *Fuzzy Matching:*
RapidFuzz library is used for approximate string matching to handle minor spelling variations. Similarity scores are calculated using techniques like Levenshtein Distance, and a threshold (e.g., 85%) is set to classify matches. Levenshtein Distance is a string metric used to measure the difference between two texts. It calculates the minimum number of single-character edit operations required to transform one string into another.

$$\operatorname{lev}_{a,b}(i,j)=\begin{cases}\max(i,j) & \text{if } \min(i,j)=0,\\ \min\begin{cases}\operatorname{lev}_{a,b}(i-1,j)+1\\ \operatorname{lev}_{a,b}(i,j-1)+1\\ \operatorname{lev}_{a,b}(i-1,j-1)+1_{(a_i\neq b_j)}\end{cases} & \text{otherwise.}\end{cases}$$

Formula 1. Levenshtein Distance

Formula 1. defines the recursive computation of the Levenshtein distance between two strings a and b. The three main operations are:

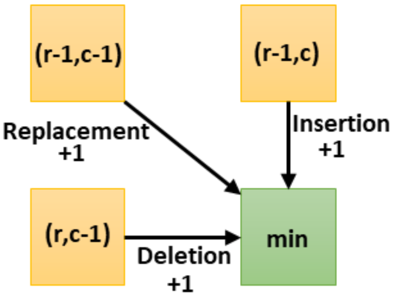


Fig 4. Levenshtein workflow

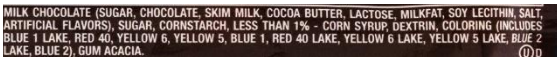
Insertion: $\text{lev}_{a,b}(i,j-1)+1$. Adding a character.
Deletion: $\text{lev}_{a,b}(i-1,j)+1$. Removing a character.
Substitution: $\text{lev}_{a,b}(i-1,j-1)+1$. Replacing one character with another, only if the characters differ.

c. *Cross-Reactivity Handling:*
The allergen database includes mappings of cross-reactive allergens. For instance, if "peanuts" are flagged in the user profile, related allergens like "tree nuts" are also checked and flagged if found.

d. *Confidence Scoring:*
Each matched allergen is assigned a confidence score based on its similarity to the user's allergen profile and its context in the ingredient list.

V. RESULTS

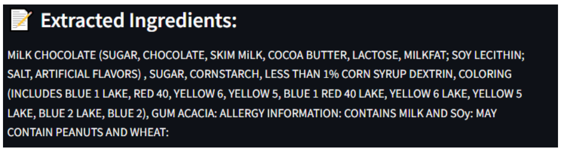
The system was live tested to locally available products. Sample result to demonstrate the system's effectiveness from these products is as followed:



MILK CHOCOLATE (SUGAR, CHOCOLATE, SKIM MILK, COCOA BUTTER, LACTOSE, MILKFAT, SOY LECITHIN, SALT, ARTIFICIAL FLAVORS), SUGAR, CORNSTARCH, LESS THAN 1% - CORN SYRUP, DEXTRIN, COLORING (INCLUDES BLUE 1 LAKE, RED 40, YELLOW 6, YELLOW 5, BLUE 1, RED 40 LAKE, YELLOW 6 LAKE, YELLOW 5 LAKE, *BLUE* 2 LAKE, BLUE 2), GUM ACACIA.

Fig 5. MnM Chocolate ingredient list

Fig 5, shows the sample product MnM being captured for the testing aspect of the system.

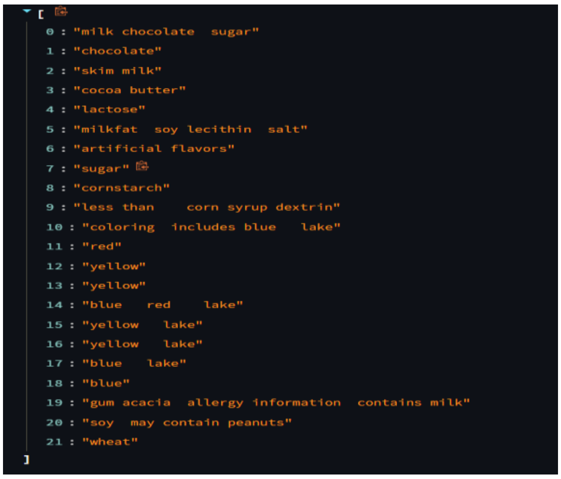


📝 Extracted Ingredients:

MiLK CHOCOLATE (SUGAR, CHOCOLATE, SKIM MiLK, COCOA BUTTER, LACTOSE, MILKFAT; SOY LECITHIN; SALT, ARTIFICIAL FLAVORS) , SUGAR, CORNSTARCH, LESS THAN 1% CORN SYRUP DEXTRIN, COLORING (INCLUDES BLUE 1 LAKE, RED 40, YELLOW 6, YELLOW 5, BLUE 1 RED 40 LAKE, YELLOW 6 LAKE, YELLOW 5 LAKE, BLUE 2 LAKE, BLUE 2), GUM ACACIA: ALLERGY INFORMATION: CONTAINS MILK AND SOy: MAY CONTAIN PEANUTS AND WHEAT:

Fig 6. OCR Extracted text:

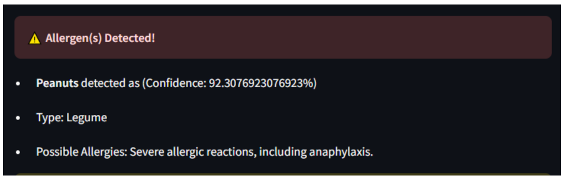
Fig 6, shows the OCR extracted ingredient list. It is clear that the extracted ingredients do have slight mistakes. So, they must go through the Fuzzy Matching to identify the closer yet correct ingredients.



```
[
  0 : "milk chocolate  sugar"
  1 : "chocolate"
  2 : "skim milk"
  3 : "cocoa butter"
  4 : "lactose"
  5 : "milkfat  soy lecithin  salt"
  6 : "artificial flavors"
  7 : "sugar"
  8 : "cornstarch"
  9 : "less than    corn syrup dextrin"
  10 : "coloring  includes blue   lake"
  11 : "red"
  12 : "yellow"
  13 : "yellow"
  14 : "blue   red    lake"
  15 : "yellow   lake"
  16 : "yellow   lake"
  17 : "blue   lake"
  18 : "blue"
  19 : "gum acacia  allergy information  contains milk"
  20 : "soy  may contain peanuts"
  21 : "wheat"
]
```

Fig 7. Text preprocessing

Fig 7 displays the preprocessing function using the RapidFuzz library. The imperfect words are modified to corrected ones



⚠ Allergen(s) Detected!

- **Peanuts** detected as (Confidence: 92.3076923076923%)
- Type: Legume
- Possible Allergies: Severe allergic reactions, including anaphylaxis.

Fig 8. Allergen detection

Fig 8 displays the alert for the given product. This is based on the user allergy, Peanuts.

A. Live Test results

a. *User Allergies*: Peanuts, gluten
b. *Input Image:* Ingredient list of a packaged food product. (MnM Chocolate)
c. *OCR Output:* “Milk chocolate (sugar, cocoa butter, skim milk), peanuts, soy lecithin”
   Parsed Ingredients: ["milk chocolate", "sugar", "cocoa butter", "skim milk", "peanuts", "soy lecithin"]
d. *Allergen Detection:* Peanuts flagged with a confidence score of 95%
e. *Notification:* "Alert: Peanuts detected in the product. Avoid consumption."

The performance of the proposed system was then evaluated based on its accuracy, error rate, and classification performance to a test dataset of about 9000 records containing a variety of ingredients images. Each record contains the allergen found in it. Thus, it was used to test the system performance The confusion matrix and accuracy graph are presented to illustrate the effectiveness of the system.

A. Confusion Matrix Analysis

The confusion matrix from Fig 9 provides a detailed breakdown of the system's classification performance:

a. *True Positives (TP):* 4800 cases
b. *True Negatives (TN):* 4823 cases
c. *False Positives (FP):* 126 cases
d. *False Negatives (FN):* 130 cases

From the confusion matrix, it is evident that the system effectively minimizes false predictions while maintaining a balanced classification of positive and negative cases. The low number of false positives and false negatives further validates the system’s reliability.

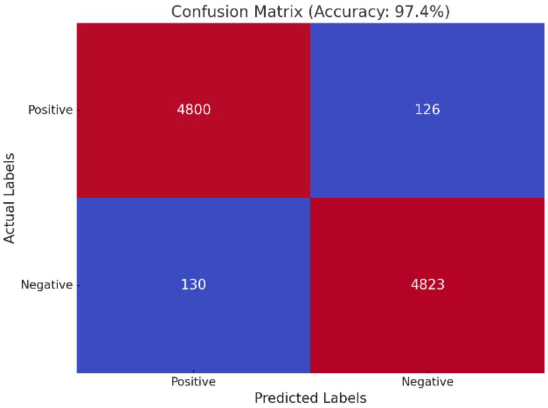


Fig 9. Confusion Matrix

A. Accuracy and Error Rate

The system achieved a high accuracy score of 97.4%, as shown in Fig 10. The error rate was minimal at 2.6%, indicating a strong predictive capability with minimal misclassification. This high accuracy demonstrates the robustness of the implemented algorithms in correctly identifying and matching data.

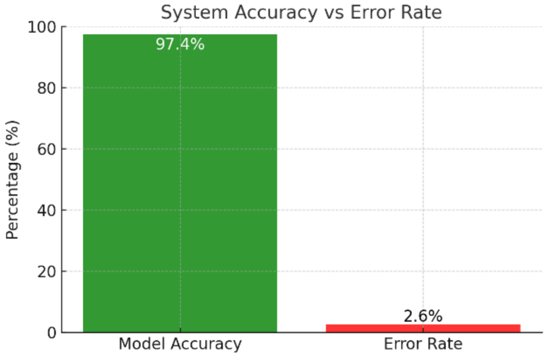


Fig 10. System Accuracy

C. Overall Performance

The combination of a high accuracy rate and a low error percentage confirms the effectiveness of the implemented OCR-based dataset matching approach. These results indicate that the system is well-suited for real-world applications requiring high precision in data extraction and classification.

VI. DISCUSSIONS

The personalized allergen notification system demonstrates significant potential for improving food safety and health management. Its reliance on OCR and NLP technologies ensures portability and accuracy. However, challenges such as OCR inaccuracies under poor lighting conditions and limited allergen datasets must be addressed. Future enhancements could include multilingual support, expanded allergen databases, and improved OCR models for better accuracy.

VII. CONCLUSIONS

The Personalized Allergen Notification System represents a transformative step in food safety technology, leveraging OCR, NLP, and fuzzy matching algorithms to provide real-time alerts about potential allergens in food products. By automating ingredient analysis, the system enables users to make informed dietary choices, reducing the risk of allergic reactions. The high accuracy and efficiency demonstrated in this research validate the effectiveness of AI-driven solutions in addressing critical health concerns. The system's ability to function locally without requiring constant internet connectivity makes it a practical and accessible tool for individuals with dietary restrictions.

Future advancements could further enhance the system's applicability and impact. Expanding dataset integration would improve recognition accuracy for a wider variety of food products, while multilingual capabilities could make the system accessible to a global audience. Additionally, incorporating real-time feedback mechanisms would allow users to refine and personalize their allergen profiles continuously. These enhancements, combined with ongoing improvements in AI and natural language processing, have the potential to make the Personalized Allergen Notification System an indispensable tool for consumers worldwide, fostering safer food consumption practices and empowering individuals to manage their dietary health with confidence.

VIII. REFERENCES

[1] Sicherer, S. H., & Sampson, H. A. (2018). Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. *Journal of Allergy and Clinical Immunology*, 141(1), 41-58.

[2] Smith, R. (2007). An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition* (Vol. 2, pp. 629-633).

[3] Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing*. Pearson.

[4] Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1), 31-88.

[5] Lee, N. A., Lopata, A. L., & Colgrave, M. L. (2023). Analytical methods for allergen control in food processing. *Foods*, 12(7), 1439.