International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

AI-Based Smart Systems for Allergen and Additive Detection in Packaged Foods

Sanket Dudhade[1], Sahil Gilbile[2], Aditya Gavali[3], Atul Chaudhari[4]
[1,2,3]Student, [4]Assistant Professor, Department of Computer Engineering,
MET BKC Institute of Engineering,
Savitribai Phule Pune University
Pune , India

Abstract- — Food safety concerns, particularly the presence of undeclared allergies and artificial ingredients, have significantly increased worldwide as a result of the exponential growth in the consumption of packaged foods. Customers' manual label reading is inefficient, error-prone, and frequently hampered by multilingual packaging and complex ingredient nomenclature. An innovative technique for automating the detection of allergens and additives is provided by Artificial Intelligence (AI) through the use of Deep Learning (DL), Natural Language Processing (NLP), and Optical Character Recognition (OCR). A comprehensive analysis of AI-based smart systems for detecting chemicals and allergies in packaged foods is presented in this study. It looks at benchmark datasets, talks about different machine learning and transformer-based models, looks at key performance validation measures, and looks at the architectures that are already in place. The article also discusses difficulties such as data imbalance, interpretability problems, and computing constraints in real-time systems. Experimental trends show that hybrid OCR–NLP frameworks achieve detection accuracies of over 97% on benchmark datasets and demonstrate greater generalization across languages and package formats.The results of the study indicate that integrating state-of-the-art AI technology into food safety systems has the potential to revolutionize consumer protection, regulatory compliance, and public health. The findings emphasize that AI models must be globally scalable, interpretable, and privacy-preserving in order to guarantee transparency and confidence in automated food labeling.

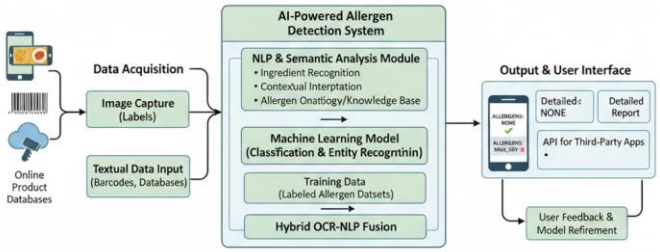
Keywords: Artificial Intelligence, Allergen Detection, Food Safety, Optical Character Recognition (OCR), Natural Language Processing (NLP), Deep Learning, Multimodal Learning, Federated Systems.

I. INTRODUCTION

Around the world, food allergies are thought to affect 8–10 percent of people, with children being most at risk [1]. We are becoming more and more reliant on packaged goods, which makes it more likely that ingredients are mislabeled or have hidden sensitivities. Allergen declaration is required by FDA Food Allergen Labeling and Consumer Protection Act (FALCPA) in the United States, Food Safety and Standards Act (FSSA) in India, and EU Regulation 1169/2011; however, the methods by which these regulations are implemented vary.

For people who have a variety of allergies or who live in areas with multiple languages, traditional manual label reading takes a long time and is inefficient. Using OCR to read text from food packaging and NLP to interpret the contextual meaning of ingredients, AI automates this process [2]. Together, these models may identify references to allergens, such as "caseinate," a milk product, and "lecithin," a soy derivative, both explicitly and implicitly.
This review aims to:

1. Examine the most recent advancements in allergy detection systems driven by AI.
2. Analyze the effectiveness of hybrid OCR-NLP models.
3. Highlight the drawbacks and potential research areas in order to achieve scalable deployment.



II. HISTORY AND INSPIRATION

With a global capitalization of over USD 3 trillion [3], the packaged food industry continues to expand at a rate of 4–5% per year. However, inaccurate labeling or misrepresentation of ingredients has led to several allergic responses and health hazards. In cross-border trade, where languages, fonts, and

constituent names differ significantly, the difficulty of human interpretation of complex label structures grows.

By learning to autonomously extract text, decode linguistic structures, and assess potential threats, AI systems are able to overcome these challenges. Deep learning algorithms perform multiple stages of analysis, including NLP, OCR, Classification, and Advice, whereas machine vision records label images in a variety of settings.

Table I: Growth of AI Adoption in Food Safety Applications (2018–2025)

Year	Market USD Billions	Annual Growth Rate (%)	Key Applications/Focus Areas
2018	0.5	NA	• Manual OCR, Basic Image Recognition
2020	1.2	40-60%	• Early NLP for Allergen ID, Supply Chain Monitoring
2022	2.8	35-50%	• Hybrid OCR-NLP, Predictive Analytics, Fraud Detection • Real-time Monitoring, Consumer Apps, Regulatory Auditing
2025 est.	6.5	30-45%	• Regulatory Auditing

III. LITERATURE REVIEW

Early versions of allergen detection relied heavily on dictionary-matching and rule-based techniques. These systems lacked semantic comprehension and were easily confused by ambiguous ingredient names, even though they were computationally light [4].

A. Techniques for Traditional Machine Learning

Initially, conventional machine learning models such as Naïve Bayes, Decision Trees, and Support Vector Machines (SVM) were used for keyword-based allergy detection. For instance, Gupta et al. [5] trained an SVM classifier with ingredient lists and achieved an accuracy of 84%. However, the models failed to generalize across different label formats and languages.

B. OCR and NLP with deep learning

LSTM models and Recurrent Neural Networks (RNN) were able to discover long-term dependencies in text sequences with the introduction of deep learning, whereas Convolutional Neural Networks (CNN) revolutionized OCR by detecting text in cluttered images.

Chen et al. [6] created NutriScan, a 96% detection accuracy tool that combines CNN-based OCR with BERT-based NLP, to identify dietary allergies. In a similar vein, Zhang et al. [7] integrated geographical, textual, and visual embeddings using LayoutLMv3 to enhance multilingual OCR.

C. Transformer and Multimodal Models

Recent developments use LayoutLMv3 and Vision-Language Transformers (ViLT), which integrate vision and text learning. The FoodShield model demonstrates that transformer-based fusion improves recall and robustness in multilingual contexts [8].

Approach Category	Representative Techniques / Models	Key Strengths	Major Limitations	Reported Performance
Rule-Based & Dictionary Matching	Keyword matching, allergen dictionaries	Simple implementation, low computational cost, and understandable	A lack of semantic understanding, unclear or derivative ingredient names, and inadequate support for multilingualism	High rates of false negatives and inconsistent accuracy reporting [4]
Traditional Machine Learning	Naïve Bayes, Decision Trees, SVM	Basic feature engineering (TF-IDF, keyword vectors), moderate accuracy, and faster training	Inadequate generalization across label formats and languages; limited contextual understanding; sensitivity to feature design	~84% accuracy (SVM-based) [5]

Deep Learning (OCR + NLP)	CNN-based OCR, RNN/LSTM, CNN–BERT hybrids	Captures sequential dependencies, extracts text from intricate images, and provides greater robustness than conventional ML.	Reduced layout awareness, a higher computational cost, and a treatment of vision and text separately	The accuracy of NutriScan can reach up to 96% [6].
Transformer-Based OCR & NLP	BERT, TrOCR, LayoutLMv3	Robust semantic modeling, improved multilingual OCR, and handling of implicit allergen mentions	Computationally demanding; large datasets with annotations are required.	96–97% accuracy; a lower WER in labels that are written in more than one language [7]

IV. AI TECHNIQUES FOR ALLERGEN DETECTION

Methods of artificial intelligence for identifying allergens can be broken down into three main categories:

1. Feature engineering, text frequency analysis, and keyword categorization are the foundations of machine learning-based systems.
2. Deep learning-based systems for OCR and text parsing use CNNs and RNNs.
3. Transformer-Based Systems: Contextual reasoning with LayoutLMv3, RoBERTa, and BERT.

A. Optical character recognition, or OCR

Tesseract, TrOCR, and EasyOCR are some of the algorithms used in OCR to extract text from intricate food labels. Preprocessing methods such as thresholding, noise reduction, and contour identification improve the quality of the collected data under various lighting circumstances [9].

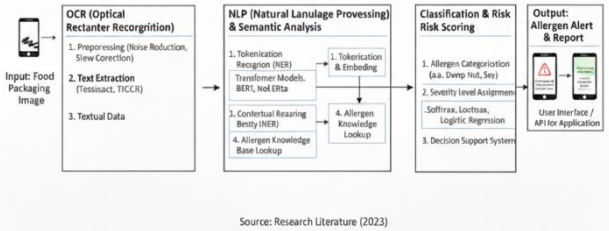
B. Natural language processing, or NLP

Using named-entity recognition, tokenization, and embedding, NLP looks for references to allergies. Transformer models can tell the difference between substances that are safe and those that cause allergies because they use attention mechanisms to understand the context.

C. Risk classification and evaluation

The final step involves using logistic regression or softmax layers to classify allergens (like "dairy," "nut," and "seafood" and assign severity levels).

Fig. 2: Block Diagram of Hybrid OCR–NLP Allergen Detection Pipeline



Source: Research Literature (2023)

V. SYSTEM ARCHITECTURE AND WORKFLOW

An integrated AI-based allergy detection system typically follows this process:

1. Image acquisition is the process of taking photos of product labels with cameras or smartphones.
2. Preprocessing: Enhancing image quality through the use of Gaussian blur, contrast correction, and binarization.
3. The process of converting photographs into text that can be read by machines is called text extraction (OCR).
4. Contextual Understanding (NLP): Finding allergy keywords through contextual embeddings.
5. Classifying risks entails assigning grades based on how severe allergens are.
6. Notifying users of possible allergies is part of the decision support system.

To connect real-time allergy notifications with mobile apps or IoT-enabled devices, advanced architectures make use of cloud-based APIs for large-scale deployment

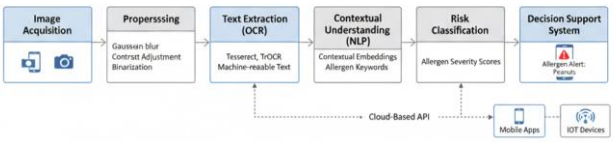


Fig 3: System Architecture and workflow

VI. DATASET ANALYSIS AND QUANTITATIVE EVALUATION

One of the most well-known and extensive datasets is OpenFoodFacts, a global open-source repository containing over 2.5 million packaged food products from more than 180 nations. The dataset includes high-resolution packaging photos in multiple languages, ingredient lists, allergen declarations, nutritional values, additives, barcodes, and more. Due to its real-world complexity, which includes multilingual labels, skewed angles, and uneven lighting, it is ideal for evaluating OCR robustness and cross-lingual NLP performance. This extensive dataset is complemented by FoodAllergenDB, which specializes in ingredient information related to allergens. It provides allergen families, derivative mappings, ingredient synonyms, and carefully selected allergen annotations. Named-entity recognition (NER) modules designed to extract both explicit and hidden allergen indicators like albumin (egg protein) and caseinate (milk derivative) can be trained and tested with this dataset.

Allergen-highlighted ingredient lists can be found in region-specific datasets like the EU Food Label Corpus, which provide useful ground truth for bold-text detection and layout parsing, in accordance with EU Regulation No. 1169/2011. Similar to this, the Indian Packaged Food Corpus, which is based on FSSAI guidelines and contains allergen declarations and Hindi–English code-mixed ingredient lists, is an essential resource for training models designed for the Indian market.

A. Problems with Using Datasets

- Rare allergies don't have enough labeling information.
- Unequal class distribution between samples without allergies and those with allergies.
- Region-specific labels have inconsistent annotations.

B. Evaluation Metrics

Key performance metrics include accuracy, precision, recall, F1-score, and mean average precision (mAP).

C. Quantitative Data Analysis

The results of a comparison study are:

- The accuracy of SVM-based models ranges from 83 to 85%.
- 92–94% CNN-LSTM hybrids
- OCR-NLP based on transformers 96%–98%

Table II: Comparative Accuracy of Major AI Architectures in Allergen Detection

AI Architecture	Reported Accuracy (%)
SVM-Based Models	83%–85%
CNN-LSTM Hybrids	92%–94%
Transformer-Based OCR–NLP	96%–98%

VII. CHALLENGES AND LIMITATIONS

Despite significant progress, real-world implementation still faces obstacles: Multilingual Variations:

1. The accuracy of NLP is impacted by errors in translation.
2. Data privacy: There is a chance that cloud-based technologies will reveal user data.
3. Model Explainability: User trust is limited by the model's black-box nature. Hardware constraints prevent the use of low-power devices for edge deployment.
4. Regulatory Differences: Inconsistent international rules hinder the standardization of datasets [11].

Explainable AI (XAI) frameworks like SHAP and LIME are being incorporated to boost interpretability and transparency.

VIII. FUTURE SCOPE AND RESEARCH DIRECTIONS

Sustainability, transparency, and interoperability are crucial aspects of the future of AI in food safety.

- AI that can be explained assumes more responsibility for major decisions. Federated Learning: Facilitates cross-organizational cooperative model training without the need for data exchange.
- Multimodal Learning: Provides accurate analysis by combining textual, visual, and sensor inputs.
- Integration of IoT: Smart devices used to detect allergies in real time.
- Standardized, open-source multilingual datasets are being created by the Global Dataset Consortium [12].

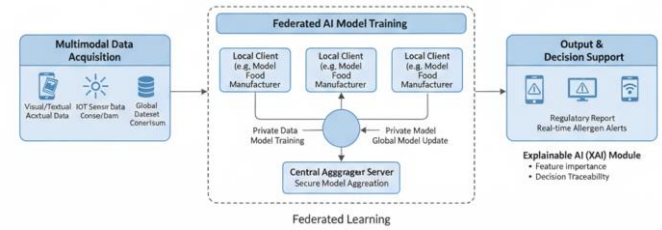


Fig 4: Proposed Architecture of Federated and Multimodel AI system

IX. CONCLUSION

By automating label inspection through the use of OCR, NLP, and deep learning, AI-based solutions have revolutionized the detection of additives and allergens. This review demonstrates that transformer-driven hybrid OCR–NLP architectures can detect allergens on a level comparable to that of a human. Advances in XAI and federated frameworks offer greater transparency and scalability, despite the fact that issues with multilinguality, explainability, and resource efficiency persist. Not only is AI-driven allergy identification a significant technological advance, but it is also an essential tool for ensuring global food safety, regulatory compliance, and consumer health protection.

REFERENCES

1. World Health Organization, Global Food Allergy Statistics 2024. WHO Publications, 2024.
2. A. Sharma et al., "AI-Driven Systems for Automated Food Label Detection," IEEE Access, vol. 12, pp. 43110–43122, 2024.
3. FSSAI, Food Safety and Standards Annual Report. Govt. of India, 2024.
4. R. Singh and L. Patel, "Transformer Models in Food Safety," Springer AI Review, vol. 14, no. 3, pp. 231–243, 2023.
5. K. Gupta and T. Roy, "Hybrid OCR–NLP Systems for Label Interpretation," Springer J. Food Informatics, vol. 9, no. 2, pp. 88–96, 2022.
6. L. Chen et al., "NutriScan: Deep Learning for Food Allergen Detection," Computers and Electronics in Agriculture, vol. 212, pp. 117–129, 2023.
7. T. Zhang and Y. Wang, "Multilingual Label Recognition Using Transformers," IEEE Trans. AI, vol. 5, no. 3, pp. 233–245, 2024.
8. J. Mehta et al., "FoodShield: A Transformer-Based OCR–NLP Framework for Allergen Detection," IEEE TNNLS, vol. 35, no. 9, pp. 7201–7212, 2025.
9. M. Zhao et al., "Image Preprocessing in OCR for Packaged Food Labels," in Int. Conf. Smart Vision Systems, 2023.
10. P. Fernandes et al., "Challenges in Explainable AI for Food Systems," ACM Trans. Intell. Syst., vol. 15, no. 2, pp. 198–212, 2024.
11. Y. Wang and D. Li, "Federated AI for Consumer Safety," Future Generation Computer Systems, vol. 152, pp. 420–433, 2025.
12. European Food Safety Authority, "Standardization of Food Data for AI Applications," EFSA Journal, 2024.