# NutriLens: An AI-Powered Smart Food Label Scanner for Real-Time Ingredient Transparency and Regulatory Compliance

Atharva Sonar, Gopal Dhurjad, Madhur Dhurjad
*MIT School of Computing, MIT Art, Design and Technology University, Pune 412201*

***Abstract* - NutriLens is an AI-powered web and mobile application that enables consumers to make safer, more informed food choices by analyzing packaged food labels in real time. Leveraging Optical Character Recognition (OCR) and Natural Language Processing (NLP), the system extracts and interprets ingredient data from complex, multilingual labels. Extracted ingredients are cross-referenced against trusted global food safety databases including FSSAI, FDA, and EFSA to detect harmful substances, allergens, and banned additives. Users can scan a product label or upload an image to receive an instant, easy-to-understand safety report categorized into Safe, Moderate, Harmful, and Banned levels. The platform is built with ReactJS, Express.js, and MongoDB, and integrates Large Language Models (LLMs) such as Llama-3 to generate personalized, human-readable insights based on individual health profiles. NutriLens bridges the gap between complex food science and consumer awareness, setting a new standard for intelligent, transparent, and regulation-aware food labeling technology.**

***Index Terms:*** **Artificial Intelligence, OCR, NLP, Food Label Analysis, Ingredient Safety, FSSAI, FDA, EFSA, Allergen Detection, LLM, ReactJS, MongoDB, Consumer Health, Food Transparency, Regulatory Compliance.**

## I. INTRODUCTION

In today's fast-paced world, packaged food consumption has become a daily norm for millions of consumers. However, interpreting food ingredient labels remains a significant challenge, especially when labels are printed in multiple languages or contain complex chemical names unfamiliar to the average consumer. This complexity often prevents individuals from making fully informed decisions about what they eat, leading to potential health risks such as allergic reactions, exposure to banned additives, and long-term dietary imbalances.

Global variations in food safety regulations such as those set by the Food Safety and Standards Authority of India (FSSAI), the U.S. Food and Drug Administration (FDA), and the European Food Safety Authority (EFSA) make it difficult for consumers to assess whether a product complies with international safety norms. Ingredients approved in one country may be restricted or banned in another due to health or environmental concerns. As globalization increases the availability of imported food products, there is a pressing need for a reliable system that helps consumers navigate this complex regulatory landscape.

Despite the abundance of nutrition-tracking and barcode-scanning applications, most existing tools provide only surface-level insights, such as calorie counts or macronutrient breakdowns. They lack the depth to analyze ingredient-level safety, detect harmful additives, or evaluate compliance across multiple jurisdictions. This gap leaves consumers without the necessary clarity to understand potential health impacts or regulatory implications of purchased food products.

To address these challenges, we developed NutriLens an AI-based packaged food ingredient analyzer designed to decode, analyze, and evaluate food labels with high precision. The platform utilizes advanced OCR and NLP technologies to extract ingredients from multilingual labels and interpret their significance in real time. NutriLens cross-references each ingredient with a comprehensive global food safety database to identify harmful substances, allergens, and banned chemicals.
Built with a ReactJS frontend and a Node.js–Express.js backend, NutriLens delivers a seamless user experience for both web and mobile users. The system stores and manages data securely using MongoDB, ensuring scalability and privacy. AI models continuously learn from new data sources, enhancing accuracy of ingredient recognition and risk categorization over time.

## II. PROBLEM STATEMENT

In today's globalized market, packaged food products are increasingly diverse, often featuring multilingual ingredient lists and complex chemical compositions. While consumers have become more health-conscious, they face significant challenges in understanding what goes into the food they consume. Ingredient information printed on packaging is often lengthy, technical, and difficult to interpret without scientific or nutritional knowledge, preventing accurate assessment of safety, nutritional value, or regulatory compliance.

Existing food and nutrition applications are largely limited in functionality, offering only calorie counts, nutritional breakdowns, or barcode-based product information. They do not analyze ingredients for potential health risks, allergens, or banned substances across different countries. The absence of multilingual recognition and lack of regulatory comparison tools make these platforms ineffective for a diverse and global consumer base.

Most current solutions do not leverage OCR to extract text directly from product labels, nor do they use NLP to interpret and classify ingredients intelligently. This leaves users to manually search for unfamiliar ingredient names a time-consuming and unreliable approach. Furthermore, the absence of a unified global database mapping ingredients to regulatory standards (FSSAI, FDA, EFSA) prevents consumers from understanding cross-regional safety compliance.

NutriLens is designed as an intelligent, AI-powered solution that scans packaged food labels, extracts and analyzes ingredients using OCR and NLP, and cross-references them with global food safety databases. By integrating advanced AI technologies with an intuitive interface, NutriLens empowers users to take control of dietary choices while promoting global standards for food safety and compliance.

## III. OBJECTIVES

The primary objective of NutriLens is to develop an intelligent, AI-driven platform that helps consumers understand and evaluate packaged food ingredients accurately, efficiently, and in real time. The specific objectives are enumerated below.

1) Real-Time Ingredient Analysis: Integrate OCR and NLP to extract text from food label images and convert it into structured, human-readable safety insights without requiring technical expertise from the user.
2) Regulatory Compliance Verification: Cross-reference each identified ingredient against verified global databases — FSSAI (India), FDA (USA), and EFSA (EU) to deliver region-specific safety evaluations.
3) Personalized Health Insights: Generate customized safety reports based on individual user profiles, including medical conditions such as diabetes, hypertension, and known allergens.
4) Consumer Education: Provide concise explanations for each ingredient, including its purpose, potential health effects, and safer alternatives where applicable.
5) Data Security and Privacy: Ensure all user data, images, and health-related preferences are processed under strict ethical AI guidelines and modern data protection standards.
6) Scalability and Accessibility: Build the system with modular architecture using ReactJS, Express.js, and MongoDB to support future expansions including mobile apps, multilingual datasets, and API-based regulatory integration.

## IV. LITERATURE SURVEY

### A. OCR for Food Labels

Optical Character Recognition (OCR) is the critical first step in automated label analysis. Tesseract remains a widely adopted open-source OCR engine, with foundational work detailing its architecture, page layout analysis, and character classification strategies. For cloud-hosted options, Google Cloud Vision provides robust text detection including full-page document OCR and dense text extraction, which aids difficult layouts, curved text, and multilingual scripts commonly found on packaged goods. These OCR foundations enable accurate text capture before downstream NLP and compliance checks.

### B. NLP for Ingredient Parsing and Normalization

After text extraction, NLP methods standardize and interpret noisy ingredient strings. Recent studies show that pretrained language models can convert unstructured label text into structured attributes suitable for safety evaluation and retrieval tasks. Typical steps include tokenization, normalization (handling synonyms such as "sucrose" vs. "sugar"), and named-entity recognition to detect additives and allergens. This structured interpretation is crucial for reliable database matching and user-facing explanations.

C. Food Knowledge Graphs and Open Databases
Open Food Facts provides an open database and API with product metadata, ingredient lists, and regular data exports; it is frequently used for prototyping and benchmarking label intelligence systems. FoodKG demonstrates how semantics-driven knowledge graphs can unify heterogeneous food sources and support downstream decision-making such as healthy recommendations and question answering, which is useful for enriching ingredient facts and relations.

D. Regulatory Intelligence and Compliance
Regulatory alignment is essential for consumer safety. In the U.S., the FDA maintains public inventories covering additives, colors, and GRAS notices. India's FSSAI publishes compendia defining permitted additives, usage limits, and lists of prohibited flavoring agents. The EFSA regularly evaluates authorized additives and communicates re-evaluation programs. Practical systems aggregate these sources so that ingredient safety is contextualized by jurisdiction.

E. Allergen Risk and Consumer Health Context
Automated allergen identification is a high-impact use case. Clinical reviews synthesize prevalent allergen classes and mechanisms, informing risk labels and user education within consumer applications. Recent engineering work couples OCR with NLP to build personalized allergen notification pipelines, demonstrating feasibility for real-time consumer feedback once ingredients are normalized and mapped to known allergen lists.

F. LLMs for Safety Explanation and Personalization
Large Language Models (LLMs) improve explanation quality and user comprehension. GPT-4 established strong performance on reasoning and multimodal tasks, indicating suitability for contextualizing ingredients and summarizing health impacts. Meta's Llama-3 delivers competitive open models with multilingual capabilities valuable for global label contexts and on-premise deployments with stricter data controls. Together, these models enable natural-language insights that bridge technical regulations and layperson understanding.

G. Comparative Analysis of Existing Systems

TABLE I: Comparative Analysis of Food Label Scanning Systems

| Aspect | NutriLens | Yuka | Google Lens | CodeCheck |
|---|---|---|---|---|
| Working Method | OCR + AI (NLP & LLMs); full ingredient label; no barcode needed. | Barcode scan to fetch product data and health scores. | Visual recognition only; not designed for food safety. | Barcode scan for ingredient and nutrition info. |
| Analysis Depth | Ingredient-level safety ratings (Safe/Moderate/Harmful/Banned) with regulatory checks. | General product score based on nutrition and additives. | Visual recognition only; no safety analysis. | Basic ingredient info; no AI reasoning. |
| User Focus | User-specific analysis based on health profile (e.g., allergies, diabetes). | Same score for all users; no personalization. | No personalization. | Limited allergen settings. |
| Data Accuracy | Verified, regularly updated databases (FDA, FSSAI, EFSA). | May rely on outdated or limited data. | No food safety data. | Depends on third-party sources; not always current. |

Multilingual and low-quality images (glare, curvature, compression) remain challenging for all existing systems. Tesseract's historical analyses emphasize preprocessing and layout strategies. Cloud OCR with full-text annotations can mitigate complex page structures. For scale, open datasets such as Open Food Facts support continuous evaluation, while regulatory portals must be periodically mirrored to keep compliance logic current.

## V. SYSTEM ARCHITECTURE

The architecture of NutriLens is designed to deliver accurate AI-based food label analysis, ensure smooth user interaction, and maintain scalability and data security. It follows a modular, layered web application architecture where each component frontend, backend, AI module, and database interacts through secure RESTful APIs. The four primary layers are described below.

### A. Frontend Layer (User Interface)

The frontend is developed using ReactJS and styled with Tailwind CSS. It provides the user-facing interface for image upload, profile management, scan history, and result visualization. The frontend communicates with the backend exclusively via HTTPS-secured RESTful APIs, transmitting user-uploaded label images and receiving structured JSON analysis reports.

### B. Backend Layer (Application Logic and APIs)

The backend is powered by Node.js with Express.js. It orchestrates all application logic receiving image uploads from the frontend, routing requests to the AI Processing Layer, querying the MongoDB database for ingredient information, and compiling the final JSON response. User session management is handled via JWT (JSON Web Tokens) for secure and stateless authentication.

### C. AI Processing Layer (OCR + NLP + LLM)

The AI layer is the analytical core of NutriLens. It performs text extraction via Tesseract OCR or Google Cloud Vision API, followed by ingredient normalization via an NLP pipeline (tokenization, named-entity recognition, synonym mapping). A Large Language Model (Llama-3) then receives the normalized ingredient list, the corresponding MongoDB records, and the user's health profile to generate personalized, human-readable safety insights.

### D. Database Layer (Data Storage and Retrieval)

MongoDB serves as the primary data store. It maintains a comprehensive ingredient repository containing ingredient type, potential side effects, health impacts, and regulatory status (banned or restricted) across FSSAI, FDA, and EFSA jurisdictions. User profiles, scan histories, and authentication records are also persisted in MongoDB.

TABLE II: NutriLens Technology Stack

| Component | Technology | Purpose |
|---|---|---|
| Frontend | React.js, Tailwind CSS | User interface, image upload, result visualization |
| Backend | Node.js with Express.js | API handling, user management, AI module integration |
| AI Module | Tesseract OCR, Llama-3 LLM | Text extraction, ingredient interpretation, safety analysis |
| Database | MongoDB | Ingredient data, user profiles, scan history, regulatory references |
| Authentication | JWT (JSON Web Tokens) | Secure access and session management |
| Deployment | Vercel (frontend); AWS/Render (backend) | Hosting and continuous deployment |

## VI. METHODOLOGY

The methodology of NutriLens integrates OCR, NLP, and LLMs in a sequential six-stage pipeline to extract, process, and evaluate food label data accurately. Each stage is described below.

### A. Stage 1: Data Acquisition

The process begins with user interaction through the web application. Users upload an image of a packaged food label or capture it using their smartphone camera. The ReactJS frontend validates the image format and resolution before transmitting it securely to the backend via HTTPS RESTful APIs.

B. Stage 2: OCR Text Extraction

The backend forwards the received image to the OCR engine (Tesseract OCR). The module extracts all textual information from the label including ingredient lists, nutritional data, and additives. Multilingual recognition is supported, enabling processing of labels in various regional and international languages. Extracted text is cleaned and structured into a machine-readable format.

C. Stage 3: NLP Ingredient Processing

The cleaned text is forwarded to the NLP module, which performs: (i) Tokenization breaking text into individual ingredient terms; (ii) Normalization standardizing ingredient names and removing duplicates; and (iii) Named Entity Recognition identifying ingredients and relevant attributes. Models such as spaCy, BERT, or transformer-based language models are used to ensure that variations in ingredient naming (e.g., "sugar," "glucose," "high-fructose syrup") are standardized for consistent database comparison.

D. Stage 4: Database Matching

The normalized ingredient list is cross-referenced with the MongoDB ingredient database. Each ingredient record contains: ingredient type (additive, preservative, natural, or synthetic); possible side effects and health impacts; and regulatory status indicating banned or restricted regions (FDA, FSSAI, EFSA). This stage ensures every recognized ingredient is validated against authoritative data sources.

E. Stage 5: AI-Based Insight Generation (LLM Integration)

The Llama-3 LLM receives three inputs: (i) the extracted and normalized ingredient list; (ii) the corresponding MongoDB database records; and (iii) the user's health profile including medical conditions and allergen sensitivities. The LLM generates a personalized safety analysis categorizing each ingredient into Safe, Moderate-Risk, Harmful, or Banned, and provides human-readable explanations with recommended alternatives where applicable.

F. Stage 6: Result Generation and Visualization

The backend compiles LLM outputs into a structured JSON response containing: a categorized ingredient list, health risk levels with short explanations, relevant regulatory references (FDA, FSSAI, EFSA), and suggested safer alternatives. The ReactJS frontend renders this data through an interactive dashboard using color-coded indicators and expandable information cards.

## VII. RESULTS AND EVALUATION

A. User Authentication and Profile Management

The Login and Sign-Up modules were implemented using ReactJS on the frontend and Node.js–Express.js with JWT authentication on the backend. Users register securely with encrypted credentials stored in MongoDB. Authenticated users can edit their health profile dietary preferences, known allergies, or medical conditions which is subsequently used by the AI module to personalize safety results.

B. Scanner and OCR Performance

The Scanner module allows users to upload a packaged food label image or capture it using their device camera. Upon submission, Tesseract OCR extracts printed text accurately from the label, including multilingual formats. The extracted text is forwarded to the NLP module, which cleans, tokenizes, and normalizes ingredient names. NutriLens was tested across a diverse set of product labels varying in size, lighting, and language. The OCR module achieved an average text extraction accuracy of 90–93% for clear and properly captured images.

C. Ingredient Analysis and Safety Classification

Post OCR and NLP processing, NutriLens performs a database cross check against the global food safety dataset. Regulatory details are retrieved from FSSAI, FDA, and EFSA sources. The Llama 3 LLM then interprets the ingredients alongside user health data to generate personalized safety insights. Each ingredient is classified under one of four risk levels Safe, Moderate, Harmful, or Banned with corresponding health and regulatory details presented in a clear, visual format.

D. System Performance Metrics

Average processing time per scan was measured at approximately 2.8 seconds under standard network conditions. The NLP pipeline demonstrated reliable ingredient recognition, handling synonym variations effectively. LLM-generated insights aligned closely with verified regulatory data, providing contextual and human-readable safety explanations. Performance variability was observed under poor network connectivity and high server load, motivating future edge-based OCR deployment.

## VIII. LIMITATIONS AND CHALLENGES

Despite the successful implementation and promising results of NutriLens, several limitations were identified during system development and testing.

### A. OCR Accuracy and Image Quality

OCR performance depends heavily on input image quality. Labels that are curved, reflective, or captured under poor lighting conditions may result in partial or incorrect text extraction. Preprocessing and denoising techniques improve readability; however, extremely distorted or low-resolution images continue to reduce recognition accuracy.

### B. Multilingual and Complex Label Handling

Although NutriLens supports multilingual text extraction through Tesseract OCR, some regional or less-common languages remain challenging. Ingredient lists written in mixed scripts (e.g., English alongside a regional language) or non-standard font styles can cause parsing errors that impact NLP interpretation.

### C. Dependence on Dataset Coverage

The accuracy of safety analysis is constrained by the breadth and freshness of the ingredient database. Comprehensive ingredient-specific metadata such as dosage thresholds and chronic exposure limits remains limited. Some ingredients may appear as unknown or unclassified if absent from the current dataset.

### D. Model and API Dependency

The analytical layer relies on external AI models and OCR frameworks that may evolve or change API versions over time. LLM-based responses may occasionally produce over-generalized explanations that require post-processing to ensure factual accuracy.

### E. Real-Time Performance Constraints

Performance can vary depending on network latency, server load, or image file size. For mobile environments with slower internet connectivity, response time may increase beyond the average 2.8-second benchmark. Edge-based OCR deployment or model optimization is planned for future releases.

### F. Limited Regulatory Synchronization

Global regulatory bodies frequently update additive lists and safety classifications. If the NutriLens database is not refreshed in real time, temporary inconsistencies may occur between official and in-app data. Automatic synchronization pipelines and API-based regulatory feeds are planned.

### G. Absence of Dedicated Mobile Application

While the web interface is fully functional and responsive, a dedicated mobile application with offline support and camera-based OCR integration is still under development. Current mobile access depends on the browser interface, limiting offline usability and local data caching.

## IX. FUTURE SCOPE

NutriLens represents a significant step toward AI-driven food transparency, and several directions for future development are planned.

### A. Mobile Application Development

A dedicated Android and iOS application is planned with built-in camera OCR, offline scanning capabilities, and cloud synchronization for user profiles and scan history. On-device OCR and caching will enable seamless access in low-connectivity regions.

### B. Real-Time Regulatory Integration

Future versions will incorporate automated regulatory synchronization through APIs connected to FSSAI, FDA, and EFSA, enabling automatic updates to ingredient risk classifications and banned substance lists.

### C. AI Personalization and Health Data Integration

Advanced AI-driven adaptive personalization will allow the system to learn from user behavior, dietary patterns, and scan history to provide predictive dietary recommendations. Integration with wearable health tracking platforms such as Google Fit or Apple Health is envisioned.

### D. Blockchain-Based Ingredient Verification

To enhance data trust and traceability, blockchain technology can be integrated to verify ingredient origins and ensure database authenticity. Each ingredient entry could be linked to a blockchain record containing supplier, certification, and compliance details.

### E. Predictive Analytics and Diet Planning

By leveraging predictive analytics, NutriLens can forecast long-term dietary impacts, identify ingredient patterns linked to potential health issues, and offer AI-powered diet planning suggestions tailored to individual health objectives.

## X. CONCLUSION

The development of NutriLens successfully demonstrates the potential of Artificial Intelligence in enhancing food transparency and consumer health awareness. The system effectively bridges the gap between complex food science and everyday decision-making by transforming detailed ingredient data into clear, actionable insights. Through the integration of OCR, NLP, and Large Language Models, NutriLens automates the extraction, interpretation, and analysis of packaged food labels with high accuracy and personalization.

The platform's architecture built using ReactJS, Node.js, and MongoDB ensures scalability, speed, and secure user interaction. The combination of AI-based ingredient recognition and regulatory data mapping enables users to identify harmful, allergenic, or banned substances in real time, thereby supporting informed and health-conscious food choices.

While certain limitations remain such as dependence on database updates and image quality the project establishes a strong foundation for intelligent food safety applications. NutriLens proves that combining AI, data intelligence, and user-centric design can deliver meaningful solutions to global food safety challenges, setting a new benchmark for intelligent food safety and awareness technologies.

## REFERENCES

[1] C. O. C. Werle, "How a food scanner app influences healthy food choice," *Appetite*, vol. 194, no. 107571, Apr. 2024.

[2] A. A. Adedeji *et al.*, "The magnitude and impact of food allergens," *Nutrients*, vol. 16, no. 3, 2024.

[3] Food Safety and Standards Authority of India (FSSAI), *Food Products Standards and Food Additives Regulations*, 2011 (Compendium, Version XII, 08-09-2020). [Online]. Available: https://www.fssai.gov.in

[4] U.S. Food and Drug Administration (FDA), "Understanding how the FDA regulates food additives and GRAS ingredients," Sept. 26, 2024. [Online]. Available: https://www.fda.gov

[5] World Health Organization (WHO), *Global Food Safety Strategy 2022–2030: Safer Food for Better Health*, 2023. [Online]. Available: https://www.who.int

[6] M. J. Zaki *et al.*, "FoodKG: A semantics-driven knowledge graph for food recommendation," in *Proc. ISWC*, 2019.

[7] G. Hu *et al.*, "Natural language processing and machine learning for food label text," *Am. J. Obstet. Gynecol.*, 2023.

[8] S. B., "OCR and NLP based personalized allergen notifying system," ResearchGate, 2025.

[9] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. ICDAR*, 2007.

[10] P. K. Mariappan, "Artificial intelligence-enabled QR codes in nutrition labelling: A conceptual paper," *Food & Nutrition Journal*, vol. 13, no. 3, 2025.

[11] Open Food Facts, "API documentation," 2025. [Online]. Available: https://openfoodfacts.github.io

[12] D. Sasanski *et al.*, "Food data in the semantic web: A review of nutritional knowledge graphs," arXiv preprint, 2025.

[13] Y. Niu, "Progress and challenge in the risk management of food additives," *Food Control*, 2025.

[14] M. Han, J. Chen, and Z. Zhou, "NutrifyAI: An AI-powered system for real-time food detection, nutritional analysis, and personalized meal recommendations," arXiv:2408.10532, Aug. 2024.

[15] "Global regulations for the use of food additives and processing aids," *Food Compliance International*, 2022.

[16] Food Packaging Forum, "Food safety & regulation – General Standard for Food Additives (GSFA) overview," 2025.