

Integrity: AI-Powered Food Label Analysis System

Harshvardhan Patil
Dept. of Computer Engineering
Gangamai College of
Engineering Dhule, India

Vedant Patil
Dept. of Computer Engineering
Gangamai College of
Engineering Dhule, India

Ritik Jagtap
Dept. of Computer Engineering
Gangamai College of
Engineering Dhule, India

Abstract – Integrity is an AI-powered web application designed to help consumers understand the health implications of packaged food products. The system allows users to capture or upload images of food product labels, which are then analyzed using a multimodal large language model combined with a curated harmful ingredients database. Detected ingredients are matched against a custom database and evaluated using a weighted risk scoring algorithm to generate a consumer-friendly health report. A Firebase-backed caching layer significantly reduces redundant AI calls, improving response time from up to sixty seconds to approximately one second for previously analyzed products. The system provides ingredient identification, additive and preservative detection, risk classification, and actionable health insights, making food safety information more accessible to everyday consumers.

Keywords – food label analysis; ingredient detection; risk scoring; Gemini API; Firebase caching

I. INTRODUCTION

Packaged food products often contain a complex mix of additives, preservatives, and artificial sweeteners whose health implications are not immediately apparent to the average consumer. While nutritional information is legally mandated on most packaging, interpreting ingredient lists requires domain knowledge that most consumers lack. Existing digital solutions primarily present raw data without contextual health analysis, leaving users without actionable insights.

Integrity addresses this gap by providing an end-to-end image-based food label analysis pipeline. Users submit an image of a packaged food product along with the brand and product name. The system leverages the Gemini 2.5 Flash multimodal model to extract structured ingredient data, matches it against a curated harmful ingredients database, and computes a quantitative risk score. Results are presented in a consumer-friendly report highlighting key health concerns, additive categories, and suggestions.

The primary contributions of this work are:

- Structured JSON extraction from food label images using a multimodal AI model.

- A hybrid analysis pipeline combining AI extraction with a curated ingredient risk database.
- A cache-assisted architecture using Firebase Firestore with fuzzy string matching to eliminate redundant API calls.
- An explainable, weighted risk scoring algorithm.
- A consumer-friendly health interpretation layers with color-coded risk classification.

II. RELATED WORK

Several prior works have explored automated food label analysis. [1] proposed a smart scanner system for ingredient categorization and nutritional composition identification in packaged food items. NutriScan [2] demonstrated AI-based ingredient detection and evaluation using mobile interfaces. Work by [3] explored the integration of machine learning with health informatics for food safety applications. [4] examined intelligent systems for food ingredient analysis using smart devices. Additionally, [5] proposed computer-vision-based approaches for food label recognition using deep learning frameworks.

While these systems demonstrate the feasibility of automated ingredient recognition, they are largely limited to structured data inputs or fixed ingredient databases. Integrity distinguishes itself through its multimodal image input, fuzzy cache-matching, and an explainable weighted scoring model.

III. SYSTEM ARCHITECTURE

A. Technology Stack

Integrity is built using React for the frontend, with a Node.js and Express backend. Firebase Firestore serves as the database for both user analysis history and the caching layer. The AI backbone is Google's Gemini 2.5 Flash model, accessed via API for multimodal image and text processing.

B. Workflow

The system workflow proceeds as follows:

1. The user uploads or captures an image of a food product label and provides the brand name and product name.
2. The image is converted to Base64 format for transmission.
3. The system queries the Firebase cache using a fuzzy matching algorithm to check for a prior analysis of the same or similar product.
4. On a cache hit, the stored analysis result is returned directly, achieving response times approximately one second.
5. On a cache miss, the image, brand name, product name, custom prompt, and a structured JSON schema are sent to the Gemini API.
6. The model returns a structured JSON response containing identified ingredients and relevant metadata.
7. Extracted ingredients are matched against the custom harmful ingredients database.
8. The risk scoring algorithm computes a final score.
9. The user receives a comprehensive health report.

IV. CACHING MECHANISM

To reduce latency and token usage, Integrity implements a Firebase Firestore-backed caching system. Cache keys are generated via a deterministic hash of the normalized brand and product name strings. On lookup, the system retrieves the 50 most recent cache entries and computes string similarity scores using the Levenshtein distance algorithm for both brand and product name fields.

A strong cache hit is triggered when brand similarity $\geq 80\%$ and product similarity $\geq 90\%$, returning the cached result immediately. A fuzzy cache hit is triggered when the combined weighted score (brand * 0.3 + product * 0.7) ≥ 0.70 . This design tolerates minor spelling variations and abbreviations in product names.

Performance comparison:

- Without cache: 10 seconds to 1 minute per analysis
- With cache: Approximately 1 second per analysis

V. INGREDIENT RISK SCORING

A. Harmful Ingredients Database

A curated database of over 60 harmful or potentially harmful ingredients was constructed, covering categories including synthetic sweeteners (e.g., Aspartame, High Fructose Corn Syrup), preservatives (e.g., Sodium Nitrate, BHA, BHT), artificial colorants (e.g., Red 40, Yellow 5), and industrial additives (e.g., Partially Hydrogenated Oils, Potassium Bromate). Each entry includes common alternative names and E-numbers, a numeric risk score (0-100), category classification, and a plain-language reason for concern.

B. Scoring Algorithm

The final risk score is computed as:

$$S = 0.45T + 0.35W + D + C \quad (1)$$

where:

- S = Final Score
- T = Top Score
- W = Weighted Average
- D = Density Penalty
- C = Count Penalty

The Top Score represents the highest risk value among all detected ingredients and is assigned the greatest weight to ensure that highly concerning ingredients significantly influence the final assessment. The Weighted Average reflects the overall risk profile of all identified ingredients and contributes to a balanced evaluation. Density Penalty and Count Penalty are incorporated to account for the concentration and quantity of potentially harmful ingredients present in the product.

The resulting score is mapped to predefined risk categories, enabling users to quickly interpret the overall health impact of a food product and make informed dietary decisions.

C. Risk Classification

Final scores are mapped to a three-tier color-coded classification:

TABLE I. RISK CLASSIFICATION BASED ON FINAL SCORES

Score Range	Classification
0-30	Red
31-60	Yellow
61-100	Green

VI. RESULTS AND OUTPUT

The system generates a structured health report for each analyzed product, including the overall risk score and color classification, a full list of detected ingredients, flagged additives, preservatives, and sweeteners, key health findings with explanations, and consumer-friendly suggestions. Token usage is approximately 1 token per analysis when served from cache.

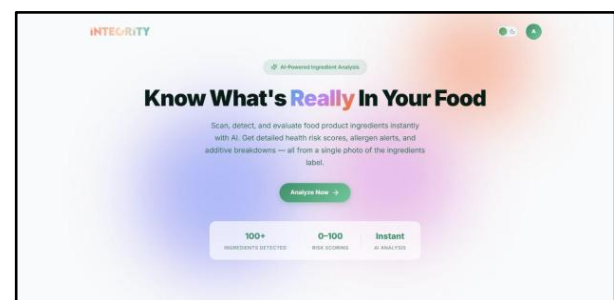


Figure 1. Landing

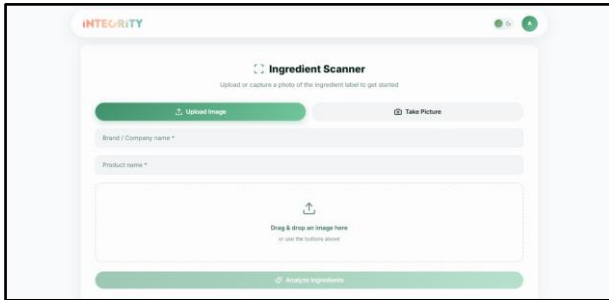


Figure 2. Dashboard



Figure 5. History

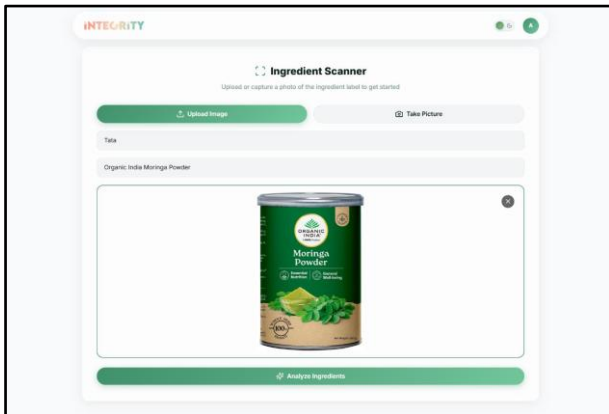


Figure 3. Ingredient Scanner

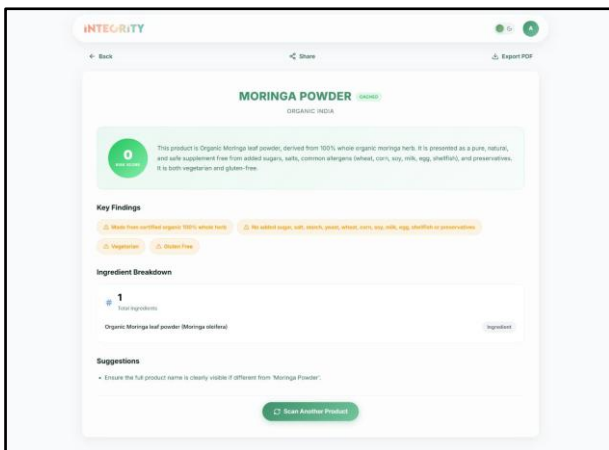


Figure 4. Result

VII. CONCLUSION

Integrity presents a practical and scalable solution for AI-assisted food label analysis. By combining multimodal AI extraction, a curated ingredient risk database, and a fuzzy cache layer, the system delivers fast, explainable, and consumer-friendly health evaluations of packaged food products. Future work may include expansion of the ingredient database, multilingual label support, and integration with barcode-based product lookup APIs.

REFERENCES

- [1] S. Bhatlawande, S. Shilaskar, and A. Surana, "A smart scanner system for ingredient categorization and identification of nutritional composition in packaged food items," *Journal of Integrated Science and Technology*, vol. 13, 2024, doi: 10.62110/sciencein.jist.2025.v13.1008.
- [2] S. Guru, K. D. Bamane, A. Patankar, C. Chandwani, A. Katre, and D. Vyavahare, "Implementation of health impact assessment of packaged foods through nutritional label recognition using OCR," *Frontiers in Health Informatics*, vol. 13, no. 3, pp. 4783–4793, Nov. 2024.
- [3] S. Lodha, S. Shinde, A. Anand, P. Dalvi, and J. Nalavade, "NutriScan: AI-based ingredient detection and evaluation," *International Journal of Engineering Research & Technology (IJERT)*, vol. 14, no. 5, May 2025.
- [4] R. Borade, A. Gupta, A. Kathpalia, T. Jain, and P. Chakurkar, "Machine learning model for optical character recognition-based food allergen detection with recommendation system for alternative food," *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, vol. 12, no. 21s, pp. 1869–1875, Mar. 2024.
- [5] A. Banerjee, P. Bansal, and K. T. Thomas, "Food detection and recognition using deep learning – A review," in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India, 2022, pp. 1221–1225, doi: 10.1109/ICAC3N56670.2022.10074297.