RESEARCH PAPER

AI-Based Food Label Decoder for Disease-Specific Nutritional Risk Analysis

V Padmapriya[1], S Uma[2], Sumathi S[3], Vighnesh[4]

[1] Associate Professor, Department of Computer Applications, B. M. S. College of Engineering, Bengaluru, India
[2] Professor, Department of Computer Applications, B. M. S. College of Engineering, Bengaluru, India
[3]Professor, Department of Electrical and Electronics Engineering, RNS Institute of Technology, Bengaluru, India
[4] PG Student, Department of Computer Applications, B. M. S. College of Engineering, Bengaluru, India
Corresponding author: V Padmapriya, padmapriyav.mca@bmsce.ac.in

ABSTRACT

Food labels provide important information about the nutritional content, ingredients, and serving details of packaged food products, helping consumers make informed dietary choices. However, understanding this information can be challenging. This paper presents an AI-Based Food Label Decoder for Disease-Specific Nutritional Risk Analysis, a system designed to simplify food label interpretation and provide meaningful health insights. The proposed system extracts nutritional information and ingredients from food label images using artificial intelligence and optical character recognition techniques. It then evaluates the nutritional quality of the product, identifies potential health risks associated with diseases such as diabetes, hypertension, heart disease, kidney disease, and obesity, and analyses harmful ingredient combinations. The system also generates a nutrition score and recommends healthier alternatives when appropriate. By combining computer vision, machine learning, and knowledge-based analysis, the proposed approach enables users to better understand packaged food products and make healthier dietary decisions.

Keywords: Artificial Intelligence, Computer Vision, Disease-Specific Nutritional Risk Analysis, Food Label Analysis, Ingredient Interaction Analysis, Knowledge Graph, Machine Learning, Optical Character Recognition (OCR), Random Forest

How to cite this article: Padmapriya V, Uma S, Sumathi S, Vighnesh. AI-Based Food Label Decoder for Disease-Specific Nutritional Risk Analysis. Int J Drug Deliv Technol. 2026;16(68s):1007-1015. DOI: 10.25258/ijddt.16.68s.117

I. INTRODUCTION

Food labels play a vital role in providing consumers with essential information about nutritional composition, ingredients, serving size, allergens, and other health-related details. This information enables individuals to make informed dietary decisions and supports healthier eating habits. As the consumption of packaged and processed foods continues to increase worldwide, food labels have become an important tool for preventing diet-related diseases such as diabetes, hypertension, obesity, cardiovascular disorders, and kidney diseases by helping consumers evaluate the nutritional quality of food products [10], [11].

Recent advancements in Artificial Intelligence (AI), Machine Learning (ML), Computer Vision, Optical Character Recognition (OCR), and Natural Language Processing (NLP) have significantly improved the automation of food label analysis. These technologies enable intelligent systems to extract textual information from food packaging, identify ingredients, categorize food products, assess nutritional quality, detect contaminants, and generate meaningful health-related insights with minimal human intervention [7]– [9], [12].

Despite these technological developments, existing food label analysis systems primarily concentrate on tasks such as text extraction, ingredient recognition, nutritional categorization, or food quality assessment. Most systems provide nutritional information or general recommendations but lack comprehensive disease-specific nutritional risk analysis based on the combined effects of nutritional composition and food additives. Furthermore, limited attention has been given to analyzing ingredient interactions, evaluating additive combinations, and integrating multiple analytical components within a unified framework capable of supporting personalized health-oriented food selection [9], [12], [13].

To address these limitations, this paper proposes an AI-based Food Label Decoder for Disease-Specific Nutritional Risk Analysis. The proposed framework employs an AI-assisted OCR pipeline to extract nutritional facts and ingredient information from food label images. A Random Forest classifier categorizes food products into Healthy, Limit, or Avoid based on their nutritional composition. The system further performs disease-specific nutritional risk assessment for diabetes, hypertension, obesity, heart disease, and kidney disease while analyzing ingredient interactions using a knowledge graph. Additionally, it generates an overall nutrition score and recommends healthier alternatives, providing users with comprehensive nutritional insights for informed food selection.

The proposed framework integrates AI-assisted OCR, machine learning-based nutritional classification, disease-specific health risk prediction, nutrition scoring, and knowledge graph-based ingredient interaction analysis into a single intelligent decision-support system. By combining these components, the system offers comprehensive nutritional evaluation and actionable recommendations that assist consumers in making healthier food choices and contribute to improved dietary awareness and preventive healthcare.

The remainder of this paper is organized as follows. Section II presents the Literature Review. Section III discusses the Research Gap. Section IV describes the Proposed Methodology. Section V explains the System Architecture. Sections VI and VII present the implementation details, machine learning model, knowledge graph-based ingredient interaction analysis, and disease-specific risk assessment, respectively. Section VIII presents the experimental results and discussion.

II. LITERATURE REVIEW

Recent advancements in Artificial Intelligence (AI), Optical Character Recognition (OCR), Computer Vision, and Machine Learning (ML) have significantly improved the automation of food label analysis and nutritional assessment.

Amoggha et al. [1] proposed FoodVisor, an AI-powered food label analysis system that combines OCR, Retrieval-Augmented Generation (RAG), and a fine-tuned LLaMA 2 model to interpret food ingredients and provide personalized dietary recommendations. The system integrates user health profiles, ingredient interpretation, and nutritional logging to generate customized health insights. However, its primary focus is personalized ingredient interpretation rather than disease-specific nutritional risk analysis.

Deepali Ujalambkar et al. [2] introduced FoodX, which integrates OCR, Artificial Neural Networks (ANN), and Natural Language Processing (NLP) to analyze packaged food labels and generate personalized health recommendations. The system extracts nutritional information from food labels and classifies products according to their nutritional characteristics while recommending healthier alternatives. Although the system provides personalized dietary guidance, it offers limited analysis of ingredient interactions and disease-specific nutritional risks.

Chetti [3] presented a food label analysis framework that combines OCR, Convolutional Neural Networks (CNN), and k-Nearest Neighbors (KNN) to extract nutritional information and classify food ingredients. The proposed framework converts complex nutritional information into simple health insights, thereby improving consumer awareness. However, the study primarily emphasizes food label digitization and nutritional interpretation rather than comprehensive disease-oriented health assessment.

Rosyadi et al. [4] developed a smartphone-based food ingredient identification system using PaddleOCR and ChatGPT. The system extracts ingredient information directly from food labels, applies spelling correction to improve OCR accuracy, and retrieves contextual information for individual ingredients. While the approach effectively improves ingredient recognition, it mainly focuses on ingredient-level information retrieval without evaluating nutritional quality or disease-specific health implications.

Roopa et al. [5] proposed an AI and Machine Learning-based framework for analyzing food additives using OCR, Google Vision API, and Convolutional Neural Networks (CNN). The system identifies food additives from packaged food labels and provides dietary recommendations based on detected additives. Although effective in additive detection, the system primarily concentrates on additive analysis and does not incorporate comprehensive nutritional classification or ingredient interaction analysis.

Assiri et al. [6] investigated the use of Large Language Models (LLMs) for extracting nutritional information from bilingual food labels. Their study demonstrated that GPT-4o, supported by appropriate post-processing techniques, achieved high extraction accuracy for multilingual food labels. However, the work mainly focuses on nutritional information extraction and does not extend to nutritional risk prediction or disease-specific dietary analysis.

Collectively, these studies demonstrate the growing role of AI in automating food label understanding through OCR, computer vision, NLP, and machine learning techniques. While significant progress has been made in nutritional information extraction, ingredient interpretation, additive detection, and personalized recommendations, opportunities remain to develop more comprehensive systems capable of integrating nutritional classification, disease-specific health assessment, and ingredient interaction analysis within a unified framework.

III. RESEARCH GAP

Existing food label analysis systems have made considerable progress in automating the extraction and interpretation of nutritional information using OCR, machine learning, computer vision, and large language models. Nevertheless, several research gaps remain.

Most existing studies primarily concentrate on extracting nutritional facts, identifying ingredients, detecting allergens, or providing generalized dietary recommendations. Although these approaches successfully improve food label accessibility, they offer limited support for disease-specific nutritional assessment based on the complete nutritional profile of packaged foods.

Furthermore, current systems generally evaluate ingredients individually and rarely consider the combined effects of multiple additives or ingredient interactions that may influence health risks [2], [4], [5]. As a result, potentially harmful ingredient combinations often remain undetected during food analysis.

Another limitation is the absence of an integrated framework capable of combining intelligent OCR, machine learning-based nutritional classification, disease-specific nutritional risk assessment, ingredient interaction analysis, nutrition scoring, and healthier food recommendation within a single system [1]–[6]. Most existing solutions address only one or two of these components independently, reducing their effectiveness in providing comprehensive dietary guidance.

To bridge these gaps, this research proposes an AI-Based Food Label Decoder for Disease-Specific Nutritional Risk Analysis. The proposed framework integrates AI-assisted OCR for automated nutritional information extraction, machine learning for nutritional classification, disease-specific risk assessment for common lifestyle-related diseases, knowledge graph-based ingredient interaction analysis, nutrition scoring, and healthier food recommendations. By combining these components into a unified framework, the proposed system aims to provide more comprehensive nutritional insights and support informed dietary decision-making.

IV. PROPOSED METHODOLOGY

4.1 OVERVIEW OF THE PROPOSED FRAMEWORK

The proposed AI-Based Food Label Decoder for Disease-Specific Nutritional Risk Analysis is designed as an intelligent decision-support system that automatically extracts, processes, and analyzes nutritional information from packaged food labels. The framework integrates Artificial Intelligence (AI), Computer Vision, Optical Character Recognition (OCR), Machine Learning (ML), and Knowledge Graph reasoning to provide nutritional classification, disease-specific risk assessment, ingredient interaction analysis, nutrition scoring, and healthier food recommendations. The complete workflow of the proposed framework is illustrated in Fig.[1].

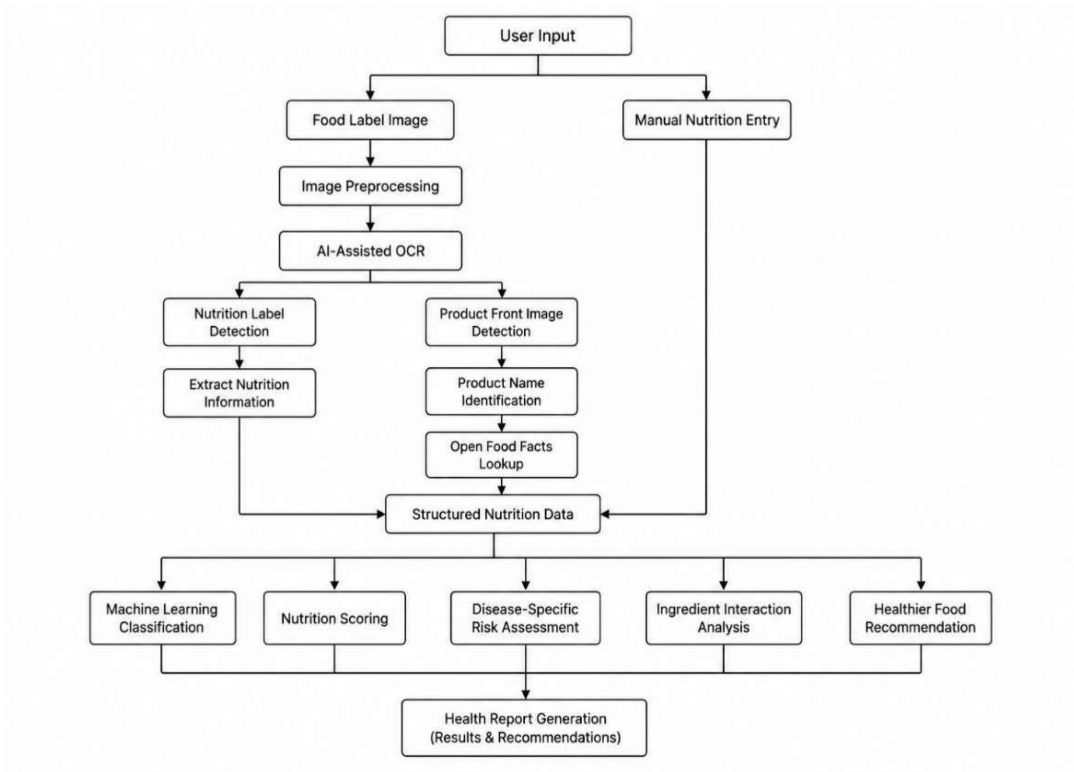

Figure Workflow diagram of the system AI-Based Food Label Decoder for Disease-Specific Nutritional Risk Analysis

As illustrated in Fig.[1], the proposed framework follows a sequential workflow that transforms raw food label information into meaningful health insights. The process begins with data acquisition, where nutritional information is obtained either from a food label image or through manual user input. In the image-based workflow, the input undergoes preprocessing and AI-assisted OCR to extract nutritional information or identify the product for retrieving its nutritional details from the Open Food Facts database. Regardless of the acquisition method, all nutritional information is converted into a unified structured nutrition data format, which serves as the common input for the subsequent analytical modules. This structured data is then used for machine learning-based nutritional classification, nutrition scoring, disease-specific risk assessment, ingredient interaction analysis, and healthier food recommendation, with the outputs finally integrated into a comprehensive health report. The detailed methodology of each stage is presented in the following subsections.

4.2 DATA ACQUISITION

The proposed system supports two independent data acquisition methods to improve usability and ensure compatibility with different types of packaged food products. Users can either upload an image of a food label or manually enter nutritional values through the application interface.

A. Food Label Image Upload

In the image-based workflow, users upload a photograph of a packaged food product captured using a smartphone or other imaging device. The frontend validates the uploaded file type and file size before transmitting the image to the backend server through a REST API. Supported image formats include PNG, JPG, JPEG, BMP, and GIF.

Once received by the server, the uploaded image is securely stored using a sanitized filename to prevent malicious file execution. The image then proceeds to the preprocessing stage before OCR-based nutritional information extraction.

B. Manual Nutritional Information Entry

To support situations where food labels are unavailable or difficult to capture, the system also provides a manual data entry interface. Users can directly enter nutritional values such as calories, carbohydrates, protein, fat, sugars, sodium, dietary fiber, saturated fat, cholesterol, and ingredient information.

Unlike the image-based workflow, manually entered nutritional information bypasses the OCR stage and is directly forwarded to the backend analysis engine after input validation. This approach minimizes processing time while maintaining the same analytical capabilities.

C. Product Information Retrieval

When the uploaded image represents the front side of a packaged product instead of the nutritional facts panel, the AI-assisted vision model first identifies the product name and brand. The identified product is then searched in the Open Food Facts database to retrieve nutritional information and ingredient details. This enables the system to analyze packaged food products even when nutritional labels are not directly visible.

The acquired nutritional information from all three acquisition methods is transformed into a common structured representation before entering the analysis pipeline.

4.3 IMAGE PREPROCESSING

Image quality plays a significant role in the accuracy of Optical Character Recognition. Variations in lighting conditions, shadows, blur, camera angle, and low image resolution may reduce text recognition performance. Therefore, an image preprocessing stage is incorporated before OCR to enhance label readability.

Initially, uploaded images are resized while preserving their aspect ratio to reduce computational overhead during image transmission and AI inference. The resized image is then converted into grayscale to simplify subsequent image processing operations.

Contrast enhancement is performed using Contrast Limited Adaptive Histogram Equalization (CLAHE), which improves the visibility of nutritional tables without excessively amplifying image noise. Compared with conventional histogram equalization, CLAHE enhances local contrast while preventing over-amplification in homogeneous image regions.

To reduce image noise generated during image acquisition, a Gaussian Blur is applied before thresholding. Noise reduction helps improve character continuity and minimizes false text detection. Finally, Adaptive Thresholding converts the grayscale image into a binary image by computing local threshold values for different image regions. Unlike global thresholding, adaptive thresholding effectively handles non-uniform illumination commonly found in photographs captured under varying lighting conditions.

The enhanced image generated after preprocessing is forwarded to the AI-assisted OCR module for nutritional information extraction. By improving text clarity and reducing image artifacts, the preprocessing stage increases OCR reliability and contributes to more accurate nutritional analysis.

4.4 INTELLIGENT OCR FRAMEWORK

The proposed system employs an AI-assisted Optical Character Recognition (OCR) framework to extract nutritional information from packaged food labels accurately. Unlike conventional OCR systems that only recognize text, the proposed framework combines image understanding with text extraction, enabling the system to identify different types of food label images and retrieve structured nutritional information. Initially, the uploaded image is resized and encoded before being processed by the OpenAI Vision model. The system first determines whether the uploaded image represents a nutrition facts label or the front side of a packaged product. This routing mechanism enables the framework to apply different processing strategies depending on the detected image type.

If the image is identified as a nutrition label, the AI-assisted OCR engine extracts nutritional values, including calories, carbohydrates, protein, fat, sugars, sodium, saturated fat, dietary fiber, cholesterol, and ingredient information. The extracted data is returned in a structured JSON format, reducing post-processing complexity and improving extraction consistency.

When the uploaded image corresponds to the front side of a packaged product, the system identifies the product name and brand. The identified product is subsequently searched in the Open Food Facts database to retrieve its nutritional information and ingredient details.

To improve system reliability, a fallback OCR mechanism is incorporated. Whenever the AI-assisted OCR service is unavailable or fails to extract valid information, the system automatically switches to a local OCR pipeline. The fallback module performs grayscale conversion, CLAHE-based contrast enhancement, Gaussian blur denoising, and adaptive thresholding before extracting text using Tesseract OCR. Regular expression-based parsing is then applied to identify nutritional values from the extracted text. [18]

The complete OCR workflow is illustrated in Fig.[2]

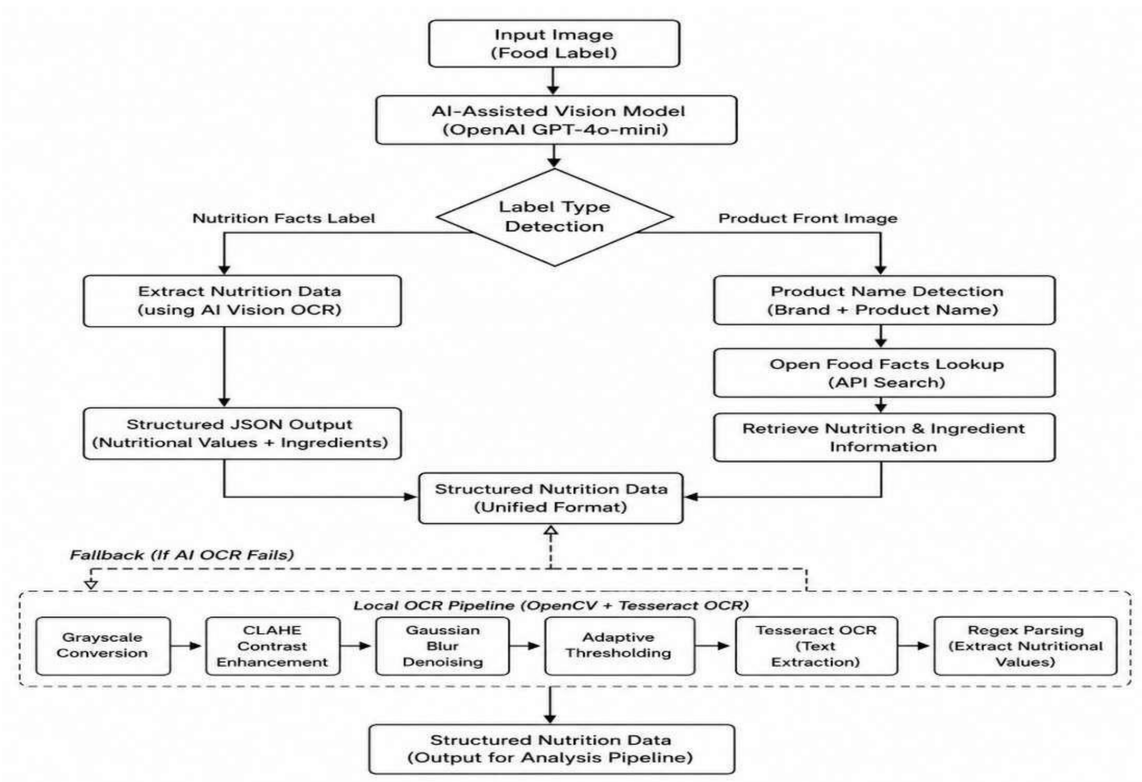

Figure Workflow of the Intelligent OCR framework.

4.5 DATA NORMALIZATION AND FEATURE PREPARATION

The nutritional information obtained through AI-assisted OCR, manual user input, or product database retrieval is transformed into a unified structured representation before analysis. Since data collected from different sources may vary in formatting and units, a normalization process is performed to ensure consistency throughout the analysis pipeline.

Initially, nutritional values are standardized into predefined attributes including calories, carbohydrates, protein, fat, sugars, sodium, saturated fat, dietary fiber, and cholesterol. Missing values are replaced with default values where appropriate, while ingredient names are cleaned by removing unwanted symbols, brackets, and redundant whitespace.

Ingredient standardization is then performed by mapping synonymous ingredient names and E-number codes to their canonical forms. For example, ingredient codes such as E211 are mapped to Sodium Benzoate, while alternative ingredient names are converted into standardized representations. This process improves consistency during ingredient interaction analysis.

The normalized nutritional information is converted into a feature vector consisting of the eight primary nutritional attributes used by the machine learning classifier:

- Caloric Value
- Protein
- Carbohydrates
- Fat
- Sodium
- Sugars
- Saturated Fat
- Dietary Fiber

The resulting feature vector provides a consistent input format for all subsequent analytical modules, irrespective of the original acquisition method.

4.6 MACHINE LEARNING-BASED NUTRITIONAL CLASSIFICATION

Following feature preparation, the nutritional profile of the food product is analyzed using a Random Forest classifier [17]. The objective of the classifier is to categorize packaged food products into one of three nutritional classes: Healthy, Limit, or Avoid.

The Random Forest algorithm was selected due to its ability to model complex non-linear relationships among nutritional attributes while maintaining high prediction accuracy and robustness against overfitting. During model development, multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, and XGBoost, were evaluated. Random Forest demonstrated the highest classification performance and was therefore selected as the final prediction model.

For each food product, the normalized feature vector is supplied to the trained classifier. Each decision tree independently predicts a nutritional category based on different subsets of training data and feature combinations. The final prediction is determined through majority voting among all decision trees, as represented by:

$$H(x) = mode(T_{(1)(x)}, T_{(2)(x)}, \dots, T_{(n)(x)})$$

The final prediction is obtained through majority voting among all decision trees, as shown in the above formula, where denotes the prediction of the decision tree and represents the final predicted nutritional class.

To ensure continuous system availability, a rule-based backup classifier is implemented. If the trained machine learning model is unavailable, the system evaluates nutritional values using predefined health thresholds based on sugar, sodium, saturated fat, dietary fiber, and protein levels. This fallback mechanism guarantees uninterrupted nutritional classification even in the absence of the trained model.

The output generated by the classification module serves as the foundation for nutrition scoring, disease-specific risk assessment, and recommendation generation.

4.7 Nutrition Scoring, Disease-Specific Risk Assessment, and Ingredient Interaction Analysis

After nutritional classification, the extracted nutritional information is further processed to evaluate the overall nutritional quality, identify disease-specific health risks, and analyze interactions among food ingredients. These analytical modules work together to provide users with comprehensive health insights beyond basic nutritional information.

A. Nutrition Scoring

The nutrition scoring module calculates an overall health score ranging from 0 to 100, representing the nutritional quality of the food product. The scoring mechanism considers both beneficial nutrients and risk-associated nutrients. Protein and dietary fiber contribute positively to the score, whereas excessive sodium, sugars, saturated fat, and calorie density reduce the final score.

The calculated score is constrained within the range of 0–100 to ensure consistency.

The overall nutrition score is computed as:

$$Nutrition\ Score = 50 + P + F - S - Su - SF - C$$

where (P) and (F) represent the positive contributions from protein and dietary fiber, while (S), (Su), (SF), and (C) denote the penalties associated with sodium, sugars, saturated fat, and calorie density, respectively. Based on the calculated score, the system provides users with an intuitive indication of the nutritional quality of the selected food product.

B. Disease-Specific Risk Assessment

The disease-specific risk assessment module evaluates the extracted nutritional values against predefined nutritional thresholds to estimate potential health risks associated with common lifestyle-related diseases. The assessment focuses on five major conditions: diabetes, hypertension, heart disease, kidney disease, and obesity.

The system analyzes nutritional parameters such as sugar, sodium, saturated fat, cholesterol, protein, fat, carbohydrates, and calorie content. Whenever a nutritional parameter exceeds its corresponding threshold, an appropriate health warning is generated. For example, high sugar levels trigger diabetes-related alerts, excessive sodium indicates hypertension risk, elevated saturated fat and cholesterol suggest cardiovascular concerns, while high calorie and fat content contribute to obesity-related warnings. Similarly, sodium and protein levels are considered when evaluating kidney disease risk. These disease-specific alerts enable users to better understand the potential health implications of consuming a particular food product.

C. Knowledge Graph-Based Ingredient Interaction Analysis

Unlike conventional food label analyzers that evaluate ingredients individually, the proposed system incorporates a knowledge graph-based ingredient interaction module to analyze relationships among multiple food additives and ingredients.

Initially, the extracted ingredient list is cleaned and standardized by mapping synonymous ingredient names and E-number codes to their corresponding canonical ingredient names. A knowledge graph is then constructed in which ingredient nodes are connected to associated health risks as well as to other ingredients that exhibit known synergistic effects.

During analysis, the system traverses the graph to identify both direct health risks and potentially harmful combinations of ingredients. Examples include interactions between Sodium Benzoate and Tartrazine, Sugar and High Fructose Corn Syrup (HFCS), and Sodium Benzoate and Aspartame. Each detected interaction is assigned a severity weight, and the cumulative interaction score is computed as

$$Interaction\ Score = \sum_{i=1}^{n} Wi$$

where (Wi) represents the severity weight associated with each detected ingredient interaction. A higher interaction score indicates a greater potential health risk arising from combined ingredient effects.

By integrating nutrition scoring, disease-specific evaluation, and ingredient interaction analysis, the proposed framework provides a more comprehensive assessment than conventional food label analysis systems that rely solely on nutritional value extraction.

4.8 HEALTHIER FOOD RECOMMENDATION

Based on the nutritional classification and analytical results, the system recommends healthier alternatives to assist users in making informed dietary choices. If a food product is classified as Avoid, the recommendation engine identifies its product category and retrieves suitable alternatives from a curated food database. For example, highly processed snack foods may be replaced with healthier options such as roasted fox nuts, baked vegetable chips, or air-popped popcorn.

The recommendation process considers the nutritional quality, disease-specific risks, and overall health score of the analyzed product before generating appropriate suggestions. These recommendations serve as practical guidance for users seeking healthier substitutes without significantly changing their dietary preferences.

V. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The proposed system follows a client-server architecture that integrates Artificial Intelligence (AI), Computer Vision, Machine Learning (ML), and Knowledge Graph reasoning to automate food label analysis. The frontend communicates with the backend through RESTful APIs, where the uploaded nutritional information is processed and analyzed before generating a comprehensive health report.

A. Overall System Architecture

The system consists of two primary components: a frontend interface and a backend analysis server. Users can either upload food label images or manually enter nutritional values through the web interface. The backend processes the received data using AI-assisted OCR, data preprocessing, machine learning classification, disease-specific nutritional analysis, and knowledge graph-based ingredient interaction analysis. The analytical results are finally compiled into a structured health report and returned to the user.

B. Frontend Implementation

The frontend is developed using HTML5, CSS3, and Vanilla JavaScript (ES6+) to provide a responsive and user-friendly interface. It supports image upload, manual nutritional data entry, and interactive visualization of the generated results. JavaScript performs client-side validation, asynchronous communication with the backend through REST APIs, and dynamically updates the user interface with nutrition scores, health warnings, ingredient interaction alerts, and healthier food recommendations.

C. Backend Implementation

The backend is implemented using the Flask web framework in Python. It manages incoming requests, validates uploaded files, coordinates communication among analytical modules, and generates structured JSON responses. Flask-CORS is employed to enable secure communication between the frontend and backend, while Werkzeug is used to sanitize uploaded filenames before temporary storage. RESTful APIs are utilized to ensure efficient data exchange between different system components.

D. AI and OCR Implementation

The AI-assisted OCR module extracts nutritional values and ingredient information from uploaded food label images. The primary OCR engine utilizes the OpenAI GPT-4o Vision model to identify image type and extract nutritional information in a structured JSON format. Before OCR, uploaded images undergo preprocessing using OpenCV, including image resizing, grayscale conversion, CLAHE-based contrast enhancement, Gaussian blur denoising, and adaptive thresholding to improve text readability.

When the uploaded image contains only the front side of a packaged product, the system identifies the product name and retrieves nutritional information from the Open Food Facts API. In situations where the AI-assisted OCR service is unavailable, the system automatically switches to a fallback OCR pipeline using PyTesseract and regular expression-based text parsing, thereby ensuring uninterrupted nutritional information extraction [18].

E. Data Processing

The extracted nutritional information is transformed into a standardized format before analysis. This module validates nutritional values, performs unit normalization, handles missing values, and standardizes ingredient names by mapping synonymous ingredients and E-number codes to canonical representations. The processed nutritional information is then converted into a consistent feature vector for subsequent analytical modules.

F. Machine Learning Classification Module

The machine learning classification module employs a pre-trained Random Forest classifier to categorize food products into Healthy, Limit, or Avoid based on their nutritional composition. The classifier utilizes eight nutritional attributes, including caloric value, protein, carbohydrates, fat, sodium, sugars, saturated fat, and dietary fiber, as input features.

During model development, multiple machine learning algorithms, namely Logistic Regression, Decision Tree, XGBoost, and Random Forest, were evaluated. Although XGBoost achieved competitive performance, Random Forest demonstrated the highest overall classification accuracy while providing faster inference, better resistance to overfitting through ensemble bagging, and improved suitability for real-time REST API predictions. Additionally, Random Forest supports feature importance analysis, enabling better interpretation of the nutritional attributes influencing classification decisions. Consequently, Random Forest was selected as the final prediction model for the proposed system.

G. Knowledge Graph Module

The knowledge graph module analyzes relationships among standardized food ingredients to identify potentially harmful additive interactions. Ingredient names are first standardized by mapping synonymous names and E-number codes to canonical representations. A knowledge graph is then constructed using the NetworkX library, where ingredient nodes are connected to known health risks and synergistic ingredient interactions. During analysis, the graph is traversed to detect harmful ingredient combinations, compute an interaction score, and generate corresponding health warnings, thereby providing a more comprehensive ingredient analysis than conventional keyword-based approaches.

H. Disease-Specific Risk Assessment

The disease-specific risk assessment module evaluates the nutritional composition of the analyzed food product against predefined nutritional thresholds associated with diabetes, hypertension, heart disease, kidney disease, and obesity. Nutritional parameters such as sugars, sodium, saturated fat, cholesterol, protein, fat, and calorie content are analyzed to determine potential health risks. Whenever a nutritional parameter exceeds its recommended threshold, the corresponding disease-specific warning is generated and incorporated into the final health report.

I. Result Generation

The outputs generated by the machine learning classifier, nutrition scoring module, disease-specific risk assessment module, and knowledge graph analyzer are consolidated into a unified health report. The final report includes the nutritional classification, nutrition score, disease-specific health warnings, ingredient interaction alerts, and healthier food recommendations. These results are transmitted to the frontend for visualization, enabling users to better understand the nutritional quality and potential health implications of the selected packaged food product.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental evaluation of the proposed AI-Based Food Label Decoder for Disease-Specific Nutritional Risk Analysis. The experiments were conducted to evaluate the effectiveness of the proposed framework in food label extraction, nutritional classification, disease-specific risk assessment, ingredient interaction analysis, and overall system performance.

A. Dataset

The proposed system utilizes two complementary data sources to support both machine learning model development and real-time food product analysis. The machine learning model was developed using a consolidated food nutrition dataset comprising 2,396 food products collected from publicly available food nutrition datasets available on Kaggle. The primary training dataset consists of FOOD-DATA-GROUP1 to FOOD-DATA-GROUP5, which were combined with a supplementary Daily Food Nutrition Dataset containing recommended daily nutritional intake values and standard meal information. These datasets were integrated to create a unified repository suitable for nutritional classification and disease-specific health risk analysis.

The consolidated dataset contains comprehensive nutritional information for each food product, including food name, caloric value, carbohydrates, protein, total fat, saturated fat, dietary fiber, sugars, sodium, cholesterol, vitamins, minerals, and food category metadata. Unlike conventional nutrition datasets that report nutrient values per 100 g, the collected dataset primarily represents nutritional values using standard household serving sizes, making it more representative of real-world food consumption patterns. In addition to the training dataset, the proposed system integrates the Open Food Facts database through its public API during runtime. Open Food Facts is an open, collaborative food product database that provides nutritional facts, ingredient lists, allergen information, food categories, barcode information, and product metadata for packaged foods worldwide. When a user uploads only the front image of a packaged food product, the AI-assisted vision module identifies the product name, and the corresponding nutritional and ingredient information is retrieved from the Open Food Facts database. This enables the system to analyze products even when the nutrition facts panel is not directly visible.

The Kaggle dataset was used to train and evaluate the Random Forest classifier, while the Open Food Facts database was used to retrieve product-specific nutritional information during real-time system execution. Together, these complementary data sources improve both the learning capability of the machine learning model and the practical applicability of the proposed framework. Before model training, the consolidated dataset underwent several preprocessing steps, including duplicate removal, missing value handling, column standardization, and unit normalization. The required nutritional attributes were extracted and converted into a consistent format before feature engineering. Additional features, including macronutrient ratios, protein-to-carbohydrate ratio, nutrient density, sugar percentage, sodium-to-calorie ratio, and binary nutritional risk indicators, were generated to improve the predictive capability of the Random Forest classifier. Since the original datasets did not contain predefined nutritional categories, a bootstrapping strategy was adopted to generate target labels. Expert-defined nutritional rules were applied to classify each food product into one of three categories: Healthy, Limit, or Avoid. Finally, the labeled dataset was randomly divided into 80% training data and 20% testing data for model development and performance evaluation.

B. Experimental Setup

The proposed system was implemented using Python and Flask as the backend framework, while HTML, CSS, and JavaScript were used for frontend development. Image preprocessing was performed using OpenCV, and nutritional information was extracted using the OpenAI GPT-4o Vision model with PyTesseract serving as the fallback OCR engine. The Random Forest classifier was implemented using the Scikit-learn library, whereas NetworkX was employed for knowledge graph construction and ingredient interaction analysis. The experiments were conducted on a standard personal computer running Windows with Python 3.x.

C. Classification Performance

Four machine learning algorithms were evaluated during model development.

Table I. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	F1-Score
Logistic Regression	83	0.82
Decision Tree	88	0.88
XGBoost	93	0.93
Random Forest	94	0.94

Among the evaluated classifiers, the Random Forest model achieved the highest classification accuracy of 94% with an F1-score of 0.94. In addition to its superior predictive performance, the model demonstrated greater robustness against overfitting and lower inference latency, making it suitable for real-time food label analysis.

D. Sample Results

The proposed system successfully extracted nutritional information from food label images, classified food products into Healthy, Limit, or Avoid categories, generated nutrition scores, identified disease-specific health risks, and detected potentially harmful ingredient interactions. The generated health report also recommended healthier food alternatives whenever the analyzed product exhibited poor nutritional quality.

E. Screenshots

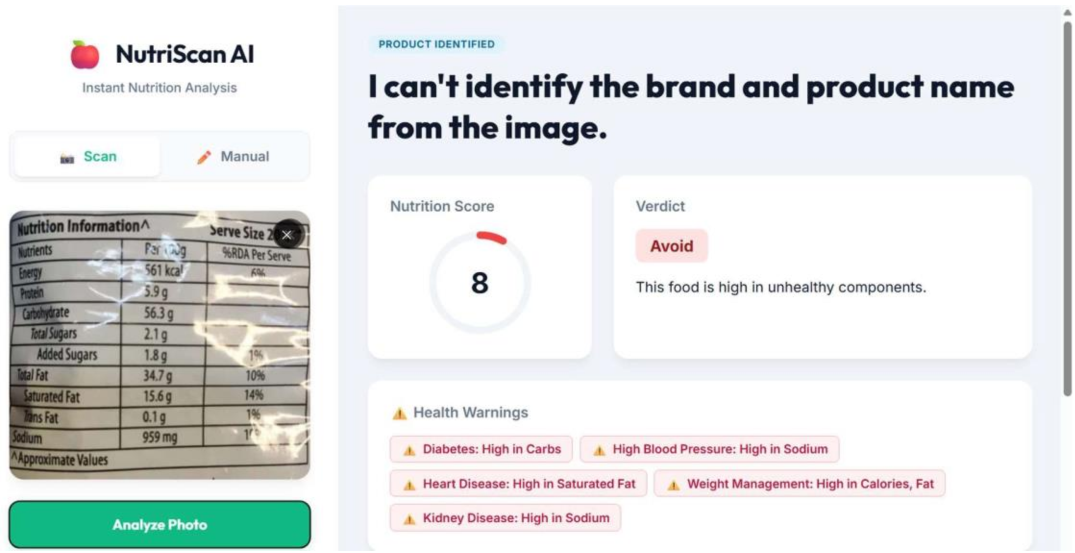

Fig. 3. Example result generated by the proposed system, including nutrition score, verdict, and health warnings.

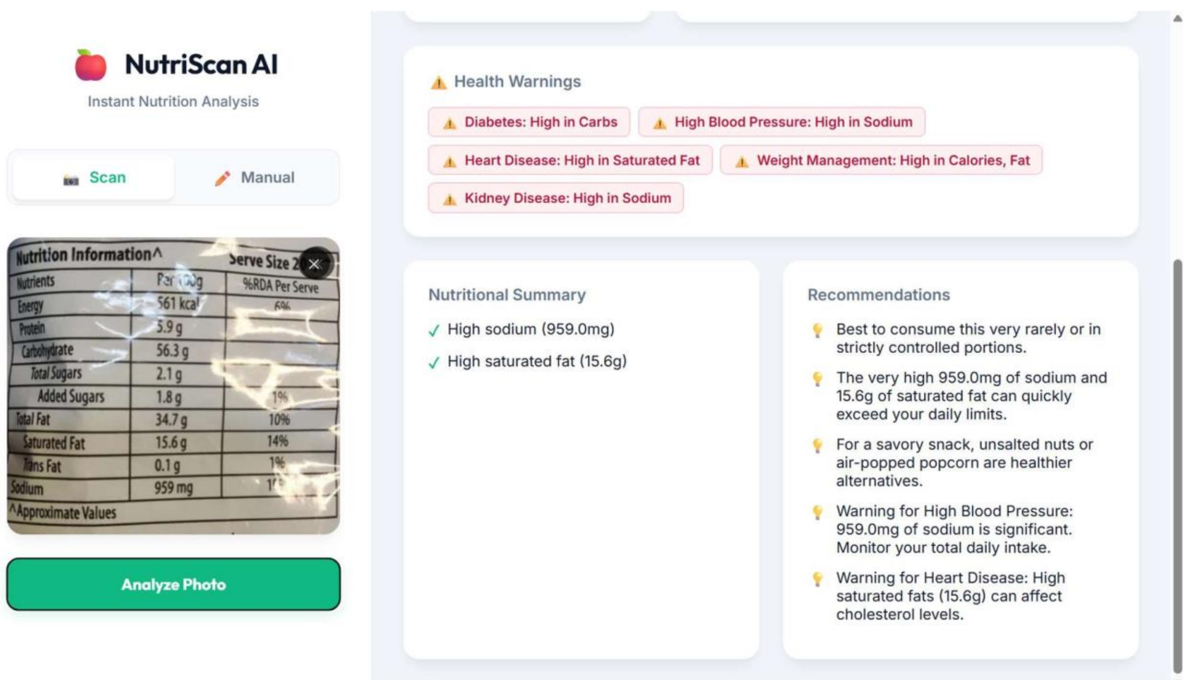

Fig. 4. Output of the proposed system displaying disease-specific health warnings, nutritional summary, and personalized healthier food recommendations

Representative screenshots of the developed system are presented in Fig. 4 and Fig.[5], illustrating the user interface, nutritional classification results, disease-specific warnings, ingredient interaction alerts, and generated health reports.

E. Performance Analysis and Discussion

The experimental results demonstrate that integrating AI-assisted OCR, machine learning, and knowledge graph reasoning significantly improves food label interpretation compared with conventional OCR-based systems. The OpenAI Vision model achieved reliable nutritional information extraction, while the fallback OCR pipeline ensured uninterrupted system operation in situations where AI-assisted extraction was unavailable. The Random Forest classifier effectively distinguished between healthy and unhealthy food products, whereas the knowledge graph module enhanced nutritional analysis by identifying harmful ingredient interactions that cannot be detected through conventional nutritional analysis alone.

Overall, the proposed framework provides an effective decision-support system capable of combining nutritional classification, disease-specific health assessment, ingredient interaction analysis, and healthier food recommendation within a single integrated platform.

VII. CONCLUSION AND FUTURE WORK

This paper presented an AI-Based Food Label Decoder for Disease-Specific Nutritional Risk Analysis, an intelligent decision-support system that automates the interpretation of packaged food labels using Artificial Intelligence, Computer Vision, Machine Learning, and Knowledge Graph reasoning. The proposed framework supports both image-based food label analysis and manual nutritional data entry, enabling users to obtain comprehensive nutritional insights with minimal effort.

The system integrates AI-assisted OCR for extracting nutritional information, a Random Forest classifier for nutritional categorization, a nutrition scoring mechanism, disease-specific risk assessment, and knowledge graph-based ingredient interaction analysis. Furthermore, the framework provides healthier food recommendations, allowing users to make more informed dietary decisions based on the nutritional quality and potential health risks associated with packaged foods.

Experimental evaluation demonstrated that the proposed approach effectively extracts nutritional information, accurately classifies food products into Healthy, Limit, or Avoid categories, identifies disease-specific health risks, and detects potentially harmful ingredient interactions. Among the evaluated machine learning models, the Random Forest classifier achieved the best overall performance, making it well-suited for real-time food label analysis.

In the future, the system can be enhanced by incorporating multilingual food label recognition, real-time barcode scanning, personalized recommendations based on user health profiles, and integration with wearable health devices and electronic health records. Future research may also explore deep learning-based nutritional risk prediction and continuously updated ingredient knowledge graphs to further improve the accuracy and reliability of food label analysis.

VIII. REFERENCES

[1] A. C. H. Amoggha, R. Padmapriya, and A. Ajina, "FoodVisor: *An AI-Powered Food Label Analysis System for Ingredient Interpretation and Personalized Dietary Recommendations*," Research Square, 2025.

[2] D. Ujalambkar, K. Kamat, G. Kulkarni, P. Khedkar, and K. Korpe, "FoodX: An AI-Powered System for Packaged Food Ingredient Analysis and Personalized Health Recommendations," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 13, no. 5, 2025.

[3] G. V. Chennam Chetti, "Food Label Analysis and Breakdown Using Machine Learning and Deep Learning," International Journal of Research Publication and Reviews, vol. 6, no. 11, pp. 3101–3106, 2025.

[4] A. W. Rosyadi, S. Ma'shumah, M. Q. Zaman, and M. R. Fajar, "Ingredients Identification Through Label Scanning Using PaddleOCR and ChatGPT for Information Retrieval," Jurnal RESTI, vol. 8, no. 6, pp. 758–767, 2024.

[5] R. N. K., B. B. Ameena, and A. Siddiqua, "Analysis of Food Additives in Packaged Food Materials Using AI and ML," International Research Journal of Engineering and Technology (IRJET), vol. 11, no. 7, pp. 127–132, 2024.

[6] F. Y. Assiri, M. D. Alahmadi, M. A. Almuashi, and A. M. Almansour, "Extract Nutritional Information from Bilingual Food Labels Using Large Language Models," Journal of Imaging, vol. 11, no. 271, 2025.

[7] Shripad Bhatlawande, Swati Shilaskar, and Anshul Surana. *A Smart Scanner System for Ingredient Categorization and Identification of Nutritional Composition in Packaged Food Items*. Journal of Integrated Science and Technology (JIST), Vol. 13, Issue 1, Article 1008, 2025.

[8] Guanlan Hu, Mavra Ahmed, and Mary R. L'Abbé. Natural Language Processing and Machine Learning Approaches for Food Categorization and Nutrition Quality Prediction Compared with Traditional Methods. The American Journal of Clinical Nutrition, Vol. 117, Issue 3, pp. 553–563, 2023.

[9] Jack Foster and Alexandra Brintrup. Aiding Food Security and Sustainability Efforts Through Graph Neural Network-Based Consumer Food Ingredient Detection and Substitution. Scientific Reports, Vol. 13, Article 18809, 2023.

[10] Sarah Campos, Juliana Doxey, and David Hammond. Nutrition Labels on Pre-Packaged Foods: A Systematic Review. Public Health Nutrition, Vol. 14, Issue 8, pp. 1496–1506, 2011.

[11] Preeti Kamboj, Neena Bhatia, and Gurudayal Singh Toteja. Awareness and Knowledge of Nutrition Labels on Pre-Packaged Food Items Among Urban Adolescents of Delhi, India. The Indian Journal of Home Science, Vol. 34, Issue 1, pp. 154–166, 2022.