NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

Food Identification and Adequacy Assessment using Computer Vision and ChatGPT in Inflammatory Bowel Diseases

Cláudia Zanatty Pombal Couceiro

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Information Management

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Food Identification and Risk Assessment using Computer Vision and ChatGPT in Inflammatory Bowel Diseases

by

Cláudia Zanatty Pombal Couceiro

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence

Supervised by

Márcia Lourenço Baptista, Phd, NOVA Information Management School

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, July 2025

Cláudia Couceiro

ABSTRACT

Inflammatory Bowel Diseases (IBD), such as Crohn's Disease and Ulcerative Colitis, require careful dietary management to reduce symptoms and improve quality of life. Recent advances in artificial intelligence have enabled new approaches for automating dietary assessment through computer vision and large language models (LLMs). This study compares the performance of a Convolutional Neural Network (CNN) and a multimodal LLM (GPT-4o) in food recognition and dietary suitability classification for individuals with IBD. A subset of the UECFOOD256 dataset was used, focusing on 34 food categories relevant to Western diets. The CNN model achieved the highest accuracy in both food classification (78.1%) and IBD suitability assessment (73.6%), outperforming GPT-4o, which achieved significantly lower results, particularly when images were pre-processed. The findings suggest that, in their current state, LLMs do not surpass CNNs in either recognition accuracy or descriptive capability. This research highlights the importance of robust visual models for dietary analysis and suggests that multimodal approaches still face challenges in clinical dietary applications.

KEYWORDS

Inflammatory Bowel Diseases; Computer Vision; ChatGPT; Dietary Analysis; Food Recognition; IBD Flare-up Prevention

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CD	Crohn's Disease (CD)
CNN	Convolutional Neural Network
COT	Chain of Thought
EEN	Exclusive Enteral Nutrition
FMT	Fecal Microbiota Transplantation
IBD	Inflammatory Bowel Disease
LLM	Large Language Model
PEN	Partial Enteral Nutrition
ROT	Reflect Observe Think
UC	Ulcerative Colitis
UCED	Diet and the UC Exclusion Diet

1. INTRODUCTION

Nutrition is fundamental to overall health and well-being. A well-balanced diet provides the body with essential macronutrients (proteins, carbohydrates, and fats) and micronutrients (vitamins and minerals), supporting optimal physiological and cognitive function. Nutrition assessment and diagnosis are especially important for patients who need to daily manage inflammatory bowel diseases (IBD), such as Crohn's disease or ulcerative colitis. As referred by (Casanova et al., 2017), patients with IBD have a high prevalence of malnutrition and nutrient deficiencies compared to the general population. Proper dietary management is essential to reducing inflammation, preventing flare-ups, and improving overall quality of life (Limketkai et al., 2023).

Technological advances are transforming the way we analyze and recognize food, leading to more precise, personalized, and efficient dietary management (Papastratis et al., 2024). Innovations in artificial intelligence (AI), machine learning, and computer vision enable real-time food recognition, nutrient analysis, and dietary tracking (Tsolakidis et al., 2024), benefiting both general consumers and patients with specific nutritional needs such as IBD.

This thesis is organized as follows:

- Literature Review: including the relevance of diet in IBD management and recent advancements in computer vision and large language models for dietary analysis.
- Methodology: describing the dataset selected, used preprocessing techniques, model architecture, and evaluation procedures.
- Results and discussion: where is presented the results obtained from both models, CNN and ChatGPT.
- Conclusion: conclusion of the thesis and suggestions for future research.

2. LITERATURE REVIEW

This review is organized into five main sections:

- Inflammatory Bowel Diseases (IBD): Introduces IBD by defining its symptoms, impact on quality of life, and the important role of diet in symptom management.
- Computer Vision in Dietary Analysis: Examines the use of computer vision in nutrition, including convolutional neural networks (CNNs) for food recognition, language models such as ChatGPT, and a comparison between traditional and multimodal methods.
- Dataset: Discusses the datasets used to train food recognition systems, their limitations, and their relevance in clinical contexts.
- Conclusion and Research Gaps: Summarizes the key findings, identifies limitations, and highlights areas for future research, emphasizing the relevance of this study.

The main goal of this review is to examine new technological methods for dietary management, focusing on the case of IBD. It explores advancements in techniques such as convolutional neural networks and multimodal models. This literature review explores and maps existing work on a) computer vision and b) multimodal models in dietary management.

2.1. INFLAMMATORY BOWEL DISEASES (IBD)

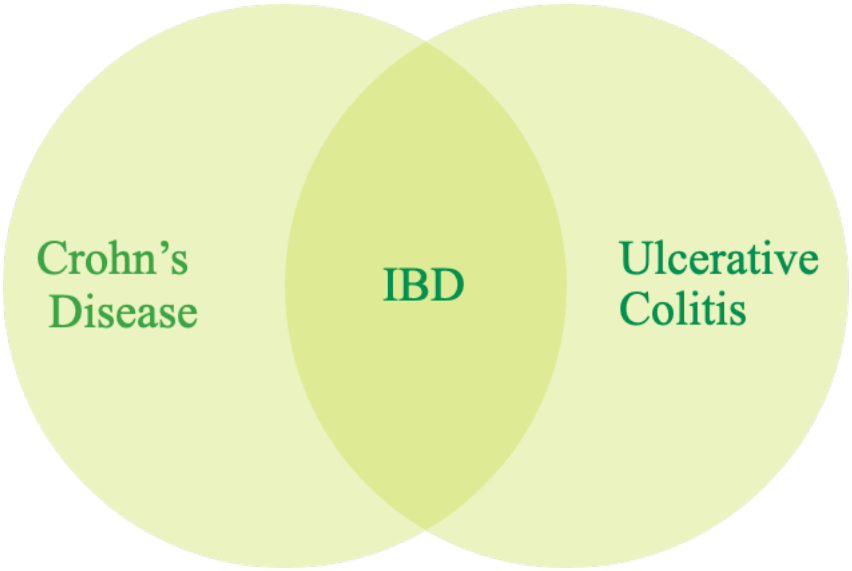

Figure 1 Inflammatory Bowel Diseases

Inflammatory Bowel Diseases (IBD), which include Crohn's Disease (CD) and Ulcerative Colitis (UC), are chronic conditions characterized by inflammation of the gastrointestinal tract (Halmos et al., 2024). The main symptoms include abdominal pain, persistent diarrhea, fatigue, and weight loss (O'Sullivan & O'Morain, 2006).

The impact of IBD on patients' quality of life is significant, including physical, emotional, and social consequences. Complications such as malnutrition, loss of muscle mass, and micronutrient deficiencies, including vitamin B12, calcium, and vitamin D, are frequently observed (Saha & Patel, 2023). In addition, chronic inflammation can lead to reduced bone mineral density and an increased risk of osteoporosis, particularly in patients treated with corticosteroids (O'Sullivan & O'Morain, 2006).

2.1.1. Role of diet in IBD Management

Although we do not know enough and more evidence is needed to fully understand the mechanisms linking diet to intestinal inflammation (Saha & Patel, 2023), studies such as the work of Reddavide et al., (2018), point that diet plays a crucial role in modulating the gut microbiome, influencing both the development and progression of the disease.

Westernized diets, characterized by high levels of saturated fats, processed meats, and ultra-processed foods, are associated with an increased risk of developing IBD. Conversely, a diet rich in fruits, vegetables, fiber, and omega-3 fatty acids may offer benefits, particularly in the prevention of Crohn's Disease (Halmos et al., 2024).

Specific dietary interventions have shown effectiveness in inducing and maintaining remission (Limketkai et al., 2023). In the context of IBD, remission refers to a phase in which the disease is inactive or significantly reduced in severity. Exclusive Enteral Nutrition (EEN), which replaces regular food with liquid supplements for 6 to 8 weeks, is widely used to induce remission in the specific case of Chrohn's Disease. Alternatives such as Partial Enteral Nutrition (PEN) and the Crohn's Disease Exclusion Diet (CDED) have been well-tolerated and show promising results. For Ulcerative Colitis, emerging approaches such as the Fecal Microbiota Transplantation Diet (FMT) and the UC Exclusion Diet (UCED) show potential (Halmos et al., 2024).

Additionally, dietary patterns like the Mediterranean Diet, rich in fiber, antioxidants, and healthy fats, have been associated with lower relapse rates in observational studies (Reddavide et al., 2018).

In summary, diet is an essential factor in the management of IBD, with the potential to influence symptoms and disease progression. The quality of life of patients can be improved with a personalized approach (De Castro et al., 2021).

2.2. Computer Vision and Dietary Analysis

2.2.1. Computer Vision for Food Recognition: How CNNs and Related Methods are Used

Computer vision, especially using Convolutional Neural Networks (CNNs), is increasingly used for food recognition and analysis due to its ability to extract complex features from images (Halmos et al., 2024). CNNs are effective at recognizing patterns in images, using multiple convolutional layers to automatically detect and classify foods. A study demonstrated the use of CNNs for food recognition, achieving high accuracy when trained on datasets such as UEC-256 and Food-101 (Yunus et al., 2019). These techniques have been further advanced with recent approaches like FMiFood, which utilized multimodal contrastive learning to improve food classification accuracy, offering a promising solution for dietary tracking (Pan et al., 2024).

Different studies have demonstrated significant accuracy in classifying images of food. For example, the work of Solanki et al., (2020) achieved 78.9% accuracy in distinguishing food categories. These developments have been facilitated by the publication of large datasets that have enabled models to learn complex visual patterns and improve recognition over time.

2.2.2. Large Language Models for Dietary Management and Prompt Engineering

Although computer vision models show considerable performance in classifying food images, Large Language Models (LLMs) like ChatGPT (Guo et al., 2025) could also play an important role in dietary management, particularly in multimodal tasks. A Large Language Model (LLM) is a type of AI model designed to understand, generate, and process human language. These models use deep learning, particularly Transformer architectures (e.g., GPT, BERT, LLaMA), to analyze and generate text.

ChatGPT, can assist in the personalized assessment of dietary habits by generating meal plans, providing nutritional recommendations, and helping users track food consumption (Kasneci et al., 2023). This functionality is particularly relevant in systems that combine image recognition with textual instructions to analyze dietary patterns, thereby enhancing user experience and improving the accuracy of nutritional monitoring.

Prompt engineering is the practice of designing and optimizing input prompts to guide LLMs like ChatGPT, to generate accurate and useful responses. This approach has been explored in research on personalized nutrition systems, which combine generative networks with language models to create meal plans tailored to user preferences and health metrics (Ma et al., 2023).

Prompt engineering has emerged as a critical discipline for maximizing the performance of LLMs, enabling users to extract more coherent and relevant responses without requiring additional training. Techniques such as Chain of Thought (CoT), which encourages the model

to reason step by step, have been shown to significantly improve the performance of models in complex reasoning tasks, such as detailed nutritional analysis (Kojima et al.).

Zero-shot-CoT has proven particularly effective at extracting multi-step reasoning without specific examples, making nutritional assessment more efficient and accessible for patients with IBD. Also, approaches such as Self-Consistency prompting and Reflect Observe Think (ROT) prompting - a strategy that encourages the model to iteratively reflect on its output, observe inconsistencies, and think through corrections - have demonstrated a significant improvement in model compliance with clinical guidelines. This suggests that adapted dietary assessments can be more accurate and evidence-based if these strategies are integrated into clinical nutrition. (L. Wang et al., 2024).

Prompt Engineering Technique	Description	Application	Reference
Chain of Thought (CoT)	Encourages the model to reason step by step before producing a final answer.	Enhances detailed food analysis, nutritional reasoning, and justification of dietary choices.	Kojima et al.
Zero-shot-CoT	A CoT variant that requires no specific examples - only a prompt like “Let’s think step by step.”	Enables complex dietary reasoning without prior user data, improving accessibility for new or unseen scenarios.	Kojima et al.
Self-Consistency Prompting	Generates multiple CoT outputs and selects the most consistent or frequent one.	Increases reliability of dietary recommendations by reducing variability and uncertainty in model responses.	L. Wang et al., 2024
Reflect-Observe-Think (ROT)	An iterative strategy where the model reflects on its output, observes inconsistencies, and reasons through corrections.	Aligns model outputs with clinical guidelines by encouraging internal validation and refinement of dietary suggestions.	L. Wang et al., 2024

Table 1 - Prompt Engineering Techiniques

2.2.3. Computer Vision-Based Methods vs Language-Image Models: Strenghts, Limitations and Applications

Computer vision-based methods and multimodal LLM models like ChatGPT offer distinct advantages and limitations. CNN-based methods are highly effective for direct image recognition and classification, making them ideal for systems focused exclusively on food image analysis. However, their performance can be limited by the complexity of visual signals, particularly in cases of foods with similar appearances (J. Wang et al., 2023).

On the other hand, language-image models leverage both visual and textual data, enabling them to understand the context of food images in a more holistic manner. This multimodal approach helps mitigate some of the limitations of purely image-based models by incorporating language to refine the classification process.

The strength of language-image models lies in their ability to unify visual and textual data, enhancing food recognition and dietary analysis. However, these models can be more computationally expensive and require large, well-labeled datasets to function effectively (Papastratis et al., 2024).

Regarding the limitations, one of the challenges identified in the literature is the reduced ability to recognize foods from different cultural contexts. The Food-500 Cap study evaluated nine vision-language models, including CLIP and text-to-image generation models, demonstrating that the accuracy of these systems is significantly lower in the food domain compared to other areas (Ma et al., 2023). The analysis revealed that these models show a clear bias, classifying Western foods with higher accuracy while struggling to correctly identify Asian and Latin American foods.

Additionally, the Food-500 Cap study highlighted that current computer vision models tend to struggle when they need to provide detailed and precise descriptions of foods. Generative image models, such as Stable Diffusion and minDALL-E, demonstrated difficulties in capturing crucial details, such as ingredients, textures, and colors, directly impacting their ability to provide accurate nutritional information in practical applications (Ma et al., 2023).

2.3. Dataset for Food Recognition

2.3.1. Mains Datasets and Limitations

Food recognition systems rely heavily on large datasets. These datasets not only provide the foundation for training deep learning models but also serve as benchmarks for evaluating model performance.

The Food-101 dataset, for instance, contains over 100,000 images across 101 food categories and serves as a standard benchmark for evaluating food classification models (Shukor et al.,

2024). Despite its popularity, it is limited in representing global cuisines, which can affect the generalization capabilities of the trained models.

The UECFOOD-256 dataset is another popular dataset and very used for food image classification, particularly in the context of Japanese cuisine. It contains 256 food categories and over 32,000 images, many of which are annotated with bounding boxes, enabling both classification and localization tasks (Kawano & Yanai, 2015). A key strength of this dataset lies in its fine-grained categorization and the inclusion of multiple food items per image, which reflects real-world dining scenarios. However, it presents several challenges: the dataset is culturally biased towards Japanese dishes, which limits generalizability to global diets.

Other datasets, such as UEC-256, have been used for food/non-food classification tasks but also face challenges related to variability in food presentation, image quality, and the need for extensive data augmentation to address these issues (Yunus et al., 2019). Additionally, datasets that include food labels and calorie information often suffer from labeling inconsistencies, which can affect the accuracy of classification and dietary estimation tasks (Kasneci et al., 2023).

Moreover, while models like CLIP benefit from large datasets that include both visual and textual information, the requirement for these models to process diverse foods and ingredients introduces new challenges in dataset curation. Ensuring that datasets are large and diverse enough to capture the variability of foods remains a key limitation for these systems (Ma et al., 2023).

Dataset	Description	Strengths	Limitations	Reference
Food-101	101 food categories with over 100.000 images.	Large and diverse dataset, widely adopted for model evaluation.	Lacks representation of global cuisines, limiting generalisation to non-Western foods.	Shukor et al., 2024
UECFOOD-256	256 food categories and over 32.000 images with bounding boxes.	Fine-grained categorisation, supports classification and localisation	Culturally biased towards Japanese dishes, variable image conditions,	Kawano & Yanai, 2015
UEC-256	Dataser user for food/non-food classification task.	Useful for binary classification scenarios and foundational tasks.	Inconsistent food presentation, low image quality, high dependency	Yunus et al., 2019

			on data augmentation.	

Table 2 – Datasets

2.4. RESEARCH GAP

2.4.1. KEY FINDINGS FROM THE LITERATURE: SUMMARY OF WHAT IS KNOWN AND ESTBABILISHED

From this review, it is possible to conclude that LLMs like ChatGPT have potential for dietary management, offering possibilities for personalized recommendations(Ponzo et al., 2024). However, to date, there have been no direct comparisons between computer vision approaches and language models in dietary assessment, especially in the context of patients with IBD.

The existing literature shows that both computer vision-based approaches and language models, such as ChatGPT, have been widely studied in the field of food recognition and assessment. Convolutional Neural Networks (CNNs) have excelled in food image classification due to their ability to process large volumes of visual data and achieve high accuracy in identifying food items. Recent studies highlight that, models like CLIP, hold significant potential for food recognition by leveraging images and textual descriptions to improve classification accuracy and relevance.

Furthermore, language models like ChatGPT have been applied to assess the nutritional value of foods, offering potential for personalized diet planning (Ponzo et al., 2024). However, to date, there have been no direct comparisons between computer vision approaches and language models in dietary assessment, especially in the context of patients with IBD.

2.4.2. LIMITATIONS AND AREAS FOR FURTHER INVESTIGATION

Although CNN and LLM models have demonstrated significant advancements, several limitations remain to be addressed. Firstly, the accuracy of these models can be affected by the quality and diversity of food images, which poses a challenge in real-world environments where variety and lighting conditions vary significantly. Additionally, many studies on nutritional assessment using ChatGPT fail to account for the nuances of individual patient biology, particularly in complex conditions such as IBD, which limits their applicability in clinical contexts.

Another major challenge is the lack of sufficiently large and representative datasets that include a variety of foods specific to individuals with IBD. This limitation hinders the models' ability to identify foods that may trigger symptoms in patients, such as certain types of fats or high-fiber foods.

2.4.3. Relevance to Current Study: How this Study Addresses the Gaps

The proposed study aims to address existing gaps by comparing the food recognition capabilities of convolutional neural networks (CNNs) with the contextual intelligence of language models (LLMs), such as ChatGPT. Focusing particularly on foods that may trigger symptoms in patients with IBD, the study seeks to compare the accuracy and effectiveness of these approaches within the specific context of IBD, a field that still lacks in-depth investigations.

To guide this study, it is driven by the following research question: **Do multimodal language models outperform classical computer vision tools in descriptive capability and achieve comparable recognition accuracy?**

By identifying foods and evaluating their impact on digestive health, this thesis contributes to the advancement of artificial intelligence techniques applied to nutrition and introduces a new paradigm for managing conditions such as IBD.

3. METHODOLOGY

In this section, we described the procedures adopted to assess the capability of Large Language Models (LLMs) compared to Convolutional Neural Networks (CNNs) in food recognition and their suitability verification for individuals with Inflammatory Bowel Disease (IBD).

This study falls within applied research and employs an explanatory design, in which the performance of two computational approaches was analyzed using the UECFOOD256 dataset, which contained images of food.

The methodology was structured into two main phases:

- Food recognition, where a trained CNN and GPT-4o were used for image classification.
- Evaluation of dietary suitability for IBD, where the models' ability to provide accurate recommendations was compared.

A visual diagram illustrating the overall flow, is presented below:

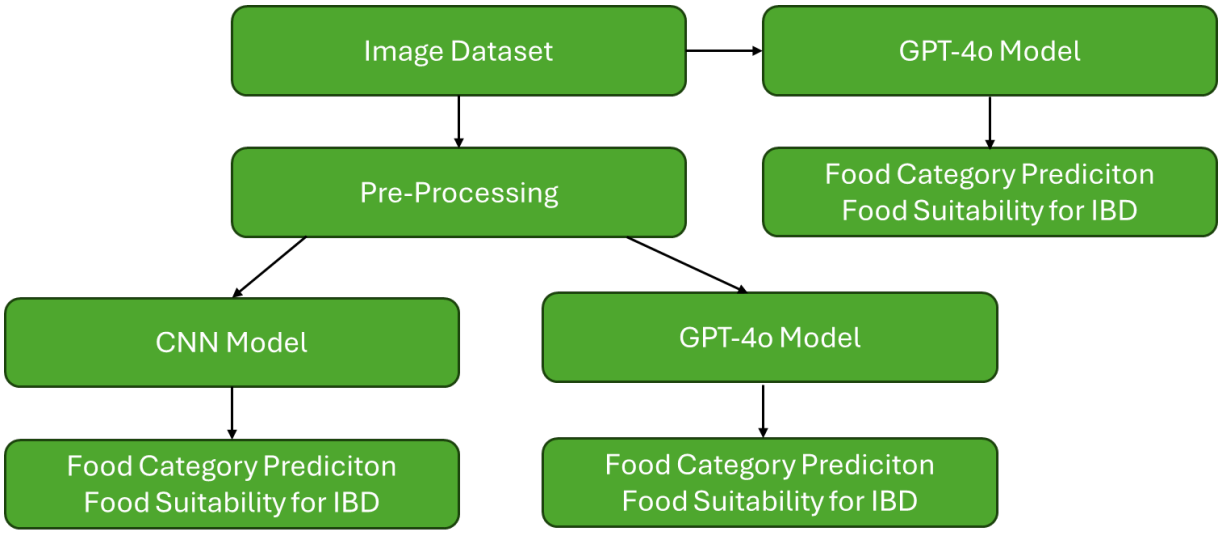

Figure 2 - Methodology Flow Diagram

The choice of this approach was driven by the need to compare the performance of a CNN model and the ChatGPT API (GPT-4o) in food identification, assessing which one demonstrated higher accuracy in visual classification and which provided more precise recommendations for IBD cases.

Three distinct model were evaluated in this study:

1. A multi-task Convolutional Neural Network (CNN), trained to classify food categories and assess dietary suitability for IBD.

2. GPT-4o with no preprocessing, where the image was only cropped using bounding box coordinates and converted to base64 format before being submitted to the model.

3. GPT-4o with full preprocessing, involving resizing, center cropping, and normalization of the image prior to base64 conversion and submission.

In both GPT-4o configurations, the model received the image along with a structured prompt instructing it to return the food category and a binary suitability label ("1" for suitable, "0" for unsuitable).
These two configurations enabled an analysis of how different levels of preprocessing affect GPT-4o's performance in food recognition and dietary assessment tasks.

3.1. USED DATASET

For the development of this thesis, a subset of the UECFOOD256 dataset was used. This dataset has been widely employed in food recognition studies. It comprised 31,395 annotated images, belonging to 256 food categories, which were captured in natural environments, such as restaurants and cafeterias. Additionally, the dataset images included bounding boxes, which indicated the exact location of the food.

However, to ensure relevance in relation to the objectives of this study, since this dataset was predominantly based on an Oriental diet, only certain categories were selected that could also be found in a Western diet, namely: baked salmon, boiled chicken and vegetables, boiled fish, chicken rice, clear soup, croissant, fried chicken, fried rice, green salad, grilled eggplant, grilled salmon, hamburger, mango pudding, mushroom risotto, oatmeal, omelet, omelet with fried rice, pizza, pork cutlet on rice, potage, potato salad, raisin bread, rice, rice ball, rice gruel, roast chicken, roast duck, roll bread, sandwiches, sautéed spinach, steamed egg potch, sushi, teriyaki grilled fish, and toast.

Dataset	Total Images	Total Categories	Selected Categories
UECFOOD-256	31.295	256	34

Table 3 - Selected Dataset Summary

3.2. DATA PREPROCESSING

In addition to selecting only specific categories, a manual classification of foods was conducted regarding their suitability for consumption by individuals with IBD, as there are no specific datasets for this type of study.

Following this classification, the images were preprocessed to ensure uniformity in size and format, as well as to remove non-essential elements that could compromise the quality of the

analysis. The objective of this preprocessing was to standardize the images and highlight the region of interest.

Initially, the images were loaded in PIL format from a directory structured into folders, where each folder represents a specific food category. Inside of each folder, there was a text file containing the bounding box coordinates, formatted as “image_name x1 y1 x2 y2,” which define the region where the food is present, isolating it from the background and reducing visual noise. This cut, based on bounding boxes, was essential to focus on the relevant area of the image.

Subsequently, to standardize dimensions, all images were resized to 256×256 pixels and then subjected to a center crop, extracting a 224×224 pixel sub-image while preserving the central region. This sequence of operations not only standardized dimensions but also minimized undesirable variations resulting from differences in framing or food positioning.

Finally, the images were converted into tensors using the ToTensor() function and normalized with the RGB channel means and standard deviations ([0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively). This normalization was particularly relevant, as the pre-trained model used ResNet50—originally trained on the ImageNet dataset—had been fine-tuned with data normalized according to these same parameters, ensuring stability and efficiency in training by leveraging prior knowledge.

With the introduction of GPT-4o, which directly analyzes images, the focus of pre- processing changed. The new objective was to ensure that the images sent to the API were well-defined and optimized for visual analysis. To achieve this, each image was cutted based on its bounding box and converted into base64 format, ensuring that the multimodal model could process it without requiring an intermediate text conversion step.

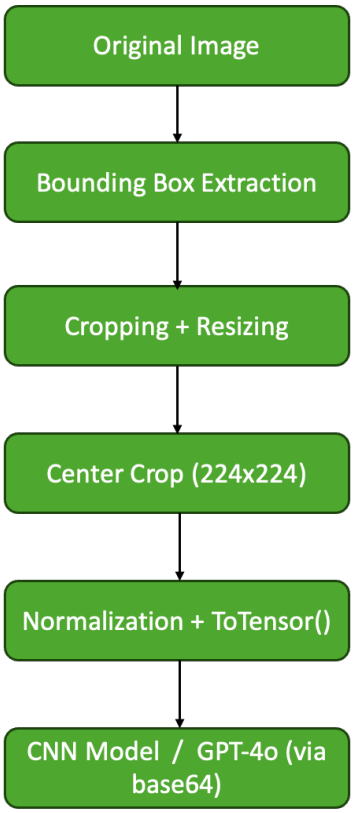

Figure 3 - Pre-Processing Diagram

3.3. MODEL ARQUITECTURE

For the classification of foods and the assessment of their suitability for IBD, a hybrid pipeline was implemented, integrating two distinct models.

The first model consisted of a multi-task convolutional neural network (CNN), with a ResNet50 backbone pre-trained on ImageNet, responsible for extracting visual features and classifying food categories. In addition to identifying the food category, a second output was incorporated into the CNN, designed to perform a binary classification of suitability for IBD. Thus, the CNN generated two outputs:

- One for food classification, identifying the detected food item.
- and Another for suitability assessment, assigning “1” for suitable foods and “0” for unsuitable ones.

The first output was structured as a linear layer followed by a softmax activation, with an output dimension corresponding to the number of classes in the dataset, allowing the identification of the food category (e.g., "pizza", "salad", etc.).

The second output was designed as another linear layer with two neurons, corresponding to binary classification, where “1” was assigned to foods considered suitable and “0” to those

unsuitable for individuals with IBD. This output also used softmax to provide a probabilistic distribution between the two classes.

This multi-task architecture enabled the network to learn shared visual representations, optimizing both tasks and improving the model's generalization ability by leveraging prior knowledge acquired during training.

The second and third models were based on GPT-4o, a multimodal model, capable of processing both images and text in an integrated manner. In the no pre-processing configuration, the GPT-4o received the image cropped by bounding box and converted to base64 format. The other with pre-processing configuration, the image was resized, cropped, normalized, and then encoded before submission. In both cases, a specialized prompt was formulated to guide the model in identifying the food item and determining its suitability for IBD. The response was generated in a fixed two-line format:

- the name of the identified food item.
- "1" for suitable or "0" for unsuitable.

By combining these two models, the approach leveraged both the visual analysis capability of the CNN and the specialized recognition of ChatGPT, providing a more robust and precise food assessment.

Model	Input	Output
CNN	Image (224x224)	Food Category and IBD Suitability
GPT-4o (without and with pre-processing)	Image (base24)	Food Category and IBD Suitability

Table 4 - Model Arquitecture

3.4. TRAINING PROCEDURES

The training procedure began with image preprocessing, as previously described, followed by splitting the dataset into training (70%), validation (15%), and test (15%) sets to ensure a robust evaluation and a reliable estimate of the model's generalization capacity.

During training, the preprocessed images - already cutted based on bounding boxes, resized, center-cropped, converted to tensors, and normalized - were processed in batches of size 32 using a DataLoader. In each iteration, the images, and their corresponding labels (both category and suitability) were transferred to the available computational device (CPU).

The model was set to training mode using the model.train() method. For each batch, two outputs were generated during the forward pass:

- one for food category classification.
- other for suitability prediction (determining whether the food was suitable or unsuitable for individuals with IBD).

Individual losses were computed for both tasks using the Cross-Entropy Loss function, and the total batch loss was obtained by summing the individual losses. Then, the backward pass was performed to propagate the error, and the model's parameters were updated using the Adam optimizer, configured with a learning rate of 0.001.

This training cycle was repeated for 5 epochs, during which performance metrics - such as loss and accuracy for both tasks - were monitored on both the training and validation sets.

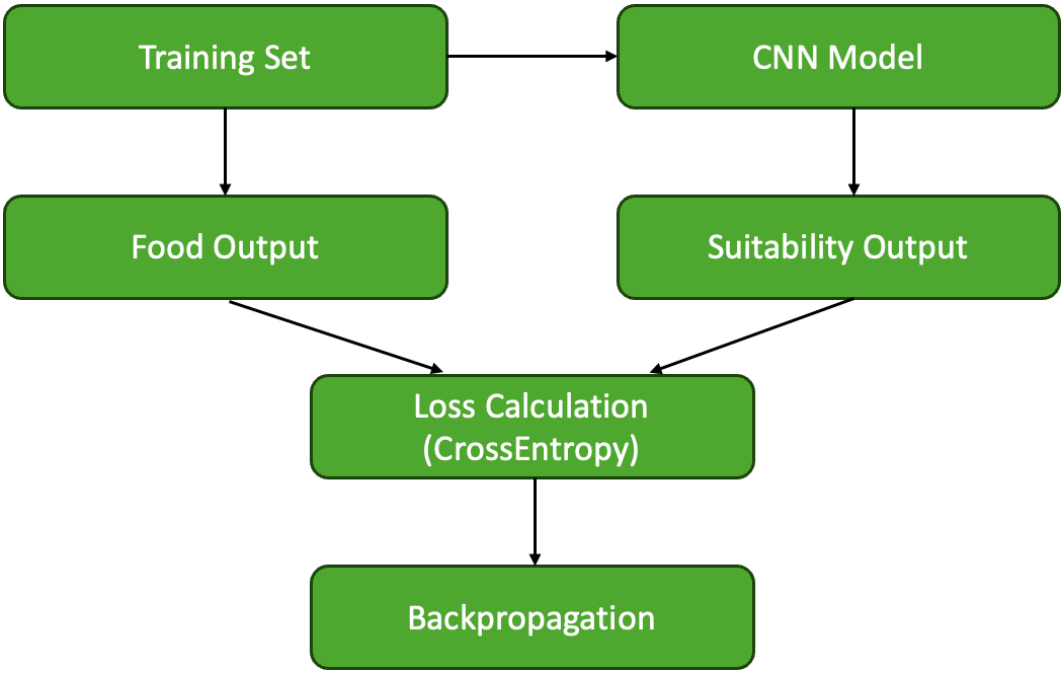

Figure 4 - CNN Model Training Procedure

For the ChatGPT models, no training procedure was performed. The GPT-4o model used in this study was already a large-scale language model pre-trained on an extensive corpus of data, possessing broad and robust knowledge, which was directly leveraged for analysis.

The model simply received the image in base64 format, along with a detailed prompt, and responded by determining the food category and its suitability for IBD. The temperature parameter was set to 0, ensuring deterministic and consistent responses.

The message structure included:

- a system message, instructing the model to act as an IBD expert.
- an user message, containing the detailed prompt.

Thus, ChatGPT functioned solely as an analysis component, returning a binary classification regarding suitability for IBD, without any internal parameter adjustments during the process.

3.5. Evaluation Metrics

In evaluating the proposed models, various performance metrics were employed to provide a comprehensive analysis of both overall accuracy and the specific behavior of each class.

For the multi-task CNN model, which simultaneously performed food category classification and binary suitability assessment for IBD, the following metrics were calculated:

Metric	Purpose
Accuracy	Provided a general overview of the model's performance.
Precision	Considered particularly important to minimize false positives in identifying foods suitable for IBD.
Recall	Measured the model's ability to identify all relevant positive instances.
F1-Score	Offered a balanced evaluation by harmonizing precision and recall.

Table 5 - Evaluation Metrics

Additionally, classification reports and confusion matrices were generated to detail the performance per class, allowing for the identification of potential classification challenges or category overlaps.

The GPT-4o model produced a CSV file as output, containing the identified food category and the corresponding binary suitability classification for IBD. These files were then analyzed and compared against the predictions generated by the CNN, enabling a direct performance evaluation of both models.

This comparative approach enhanced transparency and objectivity, facilitating the identification of patterns, discrepancies, and potential limitations of the models.

4. RESULTS AND DISCUSSION

This section presents the results obtained from both computer vision and large language models (LLMs) in the context of food identification and assessment of its suitability for individuals with Inflammatory Bowel Disease (IBD).

Three models were analyzed: a Convolutional Neural Network (CNN), the ChatGPT-4o model with pre-processing, and ChatGPT-4o without pre-processing.

A total of 5124 food images from the UECFOOD256 dataset were evaluated.

Two main tasks were assessed: the identification of the food category and the evaluation of its suitability for individuals with IBD.

The three models exhibited distinct performances in terms of recognition and suitability analysis.

The CNN model achieved the highest accuracy in food category recognition, reaching 78.1%, confirming its effectiveness in pure visual classification tasks. However, the ChatGPT-4o model also showed promising capabilities, achieving 53.2% accuracy with preprocessing and 46.5% without preprocessing. While it did not surpass the CNN in this task, these results show the potential of multimodal models in food classification.

Model	Task	Accuracy
ChatGPT-4o (no prep.)	Category	0.465
ChatGPT-4o (with prep.)	Category	0.532
CNN	Category	0.781

Table 6 - Comparative performance of models (food recognition)

In terms of assessing dietary suitability for IBD, the CNN model again showed the best performance, with an accuracy of 73.6% and an F1-score of 0.74. However, ChatGPT-4o without preprocessing reached 65.1% accuracy, and the preprocessed version achieved 61.9%. These results are encouraging, especially given that the LLM was not specifically trained for clinical dietary analysis, yet still managed to provide plausible outputs with relatively strong performance.

Model	Task	Accuracy	Precision	Recall	F1-Score
ChatGPT-4o (no prep.)	Suitability	0.651	0.608	0.548	0.576

ChatGPT-4o (with prep.)	Suitability	0.619	0.570	0.315	0.406
CNN	Suitability	0.736	0.744	0.750	0.740

Table 7 - Comparative performance of models (suitability)

This thesis sought to answer the following research question: Do multimodal language models outperform classical computer vision tools in descriptive capability and achieve comparable recognition accuracy?

Based on the current results, it is concluded that LLMs (in their current state) do not outperform the CNN in recognition accuracy.

The lower performance of the pre-processed GPT-4o model indicates that the applied techniques may have removed relevant contextual elements critical for the multimodal reasoning capabilities of the language model. In contrast, the version without preprocessing, which preserved more of the original image characteristics, achieved better results, suggesting that GPT-4o benefits from richer, less constrained visual inputs.

The results answer the research question: LLMs did not outperform classical computer vision methods in either descriptive capability or classification performance.

The CNN model demonstrated a high performance in both food recognition and suitability evaluation, proving to be the most reliable option for integration into dietary decision-support systems in the context of IBD.

But, while the CNN remains the most accurate and consistent model for both tasks, large language models (LLMs) such as GPT-4o offer distinct advantages that make them valuable in complementary roles and in contexts requiring more flexibility and reasoning.

Some of the advantages are:

- One-shot learning: GPT-4o does not require task-specific training and can generalize from limited input, making it scalable and adaptable to new scenarios or datasets.
- Context awareness: unlike CNNs, which operate on isolated visual data, LLMs incorporate semantic understanding, allowing them to respond to dietary restrictions, cultural nuances, and user-specific queries.
- **Multimodal flexibility:** GPT-4o can integrate both image and text inputs, supporting more comprehensive dietary analysis and user interaction, particularly in digital health applications.

However, some limitations were also observed during this study:

- Contextual image dependence: the lower performance of the preprocessed version suggests that GPT-4o relies in complete visual inputs. Cropping, resizing, and

normalization may strip away critical contextual, such as background, texture, and color.

- Lack of clinical fine-tuning: GPT-4o has not been trained on domain-specific clinical data, which limits its capacity to make accurate dietary recommendations without manual prompting.
- Prompt sensitivity: output quality is strongly influenced by the phrasing and structure of the prompt, requiring careful design and expertise in prompt engineering to ensure consistent, evidence-aligned outputs.

5. CONCLUSIONS AND FUTURE RESEARCH

This study evaluated and compared the capabilities of Convolutional Neural Networks (CNNs) and multimodal Large Language Models (LLMs), specifically GPT-4o, in food recognition and the assessment of dietary suitability for individuals with Inflammatory Bowel Disease (IBD). The results showed an advantage of CNNs over LLMs in both recognition accuracy and reliability of dietary recommendations. While GPT-4o demonstrated potential in handling multimodal inputs, its performance was limited.

The findings answer the research question: multimodal language models do not outperform classical computer vision models in descriptive capability or in recognition accuracy, within the scope and constraints of this study.

As future work, it is recommended that this comparative analysis be extended to a larger and more diverse dataset, incorporating foods from various cultural backgrounds to evaluate generalization capabilities. Additionally, the integration of nutritional value estimation and macronutrient analysis could further enhance the relevance of such study, contributing to the development of personalized dietary support tools for individuals with IBD and other chronic conditions.

BIBLIOGRAPHICAL REFERENCES

Casanova, M. J., Chaparro, M., Molina, B., Merino, O., Batanero, R., Dueñas-Sadornil, C., Robledo, P., Garcia-Albert, A. M., Gómez-Sánchez, M. B., Calvet, X., Trallero, M. D. R., Montoro, M., Vázquez, I., Charro, M., Barragán, A., Martínez-Cerezo, F., Megias-Rangil, I., Huguet, J. M., Marti-Bonmati, E., … Gisbert, J. P. (2017). Prevalence of Malnutrition and Nutritional Characteristics of Patients With Inflammatory Bowel Disease. *Journal of Crohn's and Colitis*, *11*(12), 1430–1439. https://doi.org/10.1093/ecco-jcc/jjx102

De Castro, M. M., Pascoal, L. B., Steigleder, K. M., Siqueira, B. P., Corona, L. P., Ayrizono, M. D. L. S., Milanski, M., & Leal, R. F. (2021). Role of diet and nutrition in inflammatory bowel disease. *World Journal of Experimental Medicine*, *11*(1), 1–16. https://doi.org/10.5493/wjem.v11.i1.1

Guo, P., Liu, G., Xiang, X., & An, R. (2025). From AI to the Table: A Systematic Review of ChatGPT's Potential and Performance in Meal Planning and Dietary Recommendations. *Dietetics*, *4*(1), 7. https://doi.org/10.3390/dietetics4010007

Halmos, E. P., Godny, L., Vanderstappen, J., Sarbagili-Shabat, C., & Svolos, V. (2024). Role of diet in prevention versus treatment of Crohn's disease and ulcerative colitis. *Frontline Gastroenterology*, flgastro-2023-102417. https://doi.org/10.1136/flgastro-2023-102417

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kawano, Y., & Yanai, K. (2015). Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation. Em L. Agapito, M. M. Bronstein, & C. Rother (Eds.), *Computer Vision—ECCV 2014 Workshops* (Vol. 8927, pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-16199-0_1

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (sem data). *Large Language Models are Zero-Shot Reasoners*.

Limketkai, B. N., Godoy-Brewer, G., Parian, A. M., Noorian, S., Krishna, M., Shah, N. D., White, J., & Mullin, G. E. (2023). Dietary Interventions for the Treatment of Inflammatory Bowel Diseases: An Updated Systematic Review and Meta-analysis. *Clinical Gastroenterology and Hepatology*, *21*(10), 2508-2525.e10. https://doi.org/10.1016/j.cgh.2022.11.026

Liu, Y., Pu, H., & Sun, D.-W. (2021). Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. *Trends in Food Science & Technology*, *113*, 193–204. https://doi.org/10.1016/j.tifs.2021.04.042

Ma, Z., Pan, M., Wu, W., Cheng, K., Zhang, J., Huang, S., & Chen, J. (2023). Food-500 Cap: A Fine-Grained Food Caption Benchmark for Evaluating Vision-Language Models. *Proceedings of the 31st ACM International Conference on Multimedia*, 5674–5685. https://doi.org/10.1145/3581783.3611994

O'Sullivan, M., & O'Morain, C. (2006). Nutrition in inflammatory bowel disease. *Best Practice & Research Clinical Gastroenterology*, *20*(3), 561–573. https://doi.org/10.1016/j.bpg.2006.03.001

Pan, X., He, J., & Zhu, F. (2024). *FMiFood: Multi-modal Contrastive Learning for Food Image Classification* (No. arXiv:2408.03922). arXiv. http://arxiv.org/abs/2408.03922

Papastratis, I., Konstantinidis, D., Daras, P., & Dimitropoulos, K. (2024). AI nutrition recommendation using a deep generative model and ChatGPT. *Scientific Reports*, *14*(1), 14620. https://doi.org/10.1038/s41598-024-65438-x

Ponzo, V., Goitre, I., Favaro, E., Merlo, F. D., Mancino, M. V., Riso, S., & Bo, S. (2024). Is ChatGPT an Effective Tool for Providing Dietary Advice? *Nutrients*, *16*(4), 469. https://doi.org/10.3390/nu16040469

Reddavide, R., Rotolo, O., Caruso, M. G., Stasi, E., Notarnicola, M., Miraglia, C., Nouvenne, A., Meschi, T., de' Angelis, G. L., Di Mario, F., & Leandro, G. (2018). The role of diet in the prevention and treatment of Inflammatory Bowel Diseases. *Acta Bio Medica Atenei Parmensis*, *89*(9-S), 60–75. https://doi.org/10.23750/abm.v89i9-S.7952

Saha, S., & Patel, N. (2023). What Should I Eat? Dietary Recommendations for Patients with Inflammatory Bowel Disease. *Nutrients*, *15*(4), 896. https://doi.org/10.3390/nu15040896

Shukor, M., Thome, N., & Cord, M. (2024). Vision and Structured-Language Pretraining for Cross-Modal Food Retrieval. *Computer Vision and Image Understanding*, *247*, 104071. https://doi.org/10.1016/j.cviu.2024.104071

Solanki, D., Anurag, A., Goel, D. A., Bahl, V., & Sengar, N. (2020). *Detection and Classification of Food Consumption Using Convolutional Neural Networks*. *07*(11).

Tsolakidis, D., Gymnopoulos, L. P., & Dimitropoulos, K. (2024). Artificial Intelligence and Machine Learning Technologies for Personalized Nutrition: A Review. *Informatics*, *11*(3), 62. https://doi.org/10.3390/informatics11030062

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., Shi, E., Pan, Y., Zhang, T., Zhu, D., Li, X., Jiang, X., Ge, B., Yuan, Y., Shen, D., … Zhang, S. (2023).

Review of large vision models and visual prompt engineering. *Meta-Radiology*, *1*(3), 100047. https://doi.org/10.1016/j.metrad.2023.100047

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *Npj Digital Medicine*, *7*(1), 41. https://doi.org/10.1038/s41746-024-01029-4

Yunus, R., Arif, O., Afzal, H., Amjad, M. F., Abbas, H., Bokhari, H. N., Haider, S. T., Zafar, N., & Nawaz, R. (2019). A Framework to Estimate the Nutritional Value of Food in Real Time Using Deep Learning Techniques. *IEEE Access*, *7*, 2643–2652. https://doi.org/10.1109/ACCESS.2018.2879117

APPENDIX A

Project No.: **DSCI2025-3-193792**

Project Title: **Food Identification and Risk Assessment of Foods using Computer Vision and ChatGPT in Inflammatory Bowel Diseases**

Principal Researcher: **Cláudia Couceiro**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 08/04/2025.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of

- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa