Machine Learning (UML501)
Project Report
End-Semester Evaluation

NutriVision
A Computer Vision Project

Submitted by:
Aishwary Srivastava (102317117)

BE Third Year
Group: 3Q15
COPC

Submitted to:
Dr. Anjula Mehto
Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology

Patiala

November 2025

TABLE OF CONTENTS

Introduction

Food plays a central role in everyday life, a fact reflected in the vast volume of food images captured and shared across digital platforms. With the increasing use of mobile photography for dietary monitoring and nutrition tracking, automated food analysis has become a meaningful research problem. Traditional systems rely heavily on nutrition experts or crowdsourced labeling platforms to estimate nutritional attributes, making them slow, costly, and inconsistent. NutriVision aims to address these limitations by designing a machine learning pipeline capable of automatically estimating key nutritional metrics—such as calories, mass, macronutrients, and ingredient-level contributions—directly from visual and structured data. The project leverages the Nutrition5k dataset, which provides dish metadata, ingredient lists, and multimodal imaging (RGB, raw depth, and colorized depth) to build a streamlined and efficient end-to-end solution.

The system integrates metadata preprocessing, image-based feature extraction, and structured nutritional modeling to produce accurate, scalable predictions. After cleaning and aligning dish-level and ingredient-level metadata with available overhead imagery, the dataset is split into robust training and testing partitions using curated dish-ID splits. The model architecture follows a simple yet effective approach: stacked RGB, raw depth, and colorized depth images are processed through a lightweight vision model, while textual metadata is encoded through a compact MLP. These representations are fused using a basic neural layer to estimate nutritional targets. The resulting workflow enables automated prediction and evaluation of nutritional attributes, significantly reducing manual intervention and offering a foundation for practical, real-world food analysis applications.

The NutriVision project presents a simplified yet effective computer-vision-based nutrition estimation pipeline built using overhead RGB, raw-depth, and colorized-depth images paired with structured dish-level metadata. After performing extensive preprocessing—such as dataset cleaning, dish-ID–based filtering, NULL-value handling, and aligned train/test split creation—the system extracts multimodal features by stacking all three overhead image modalities for a basic neural network–based image encoder, while a lightweight MLP processes textual nutritional attributes from metadata. These complementary representations are fused and trained to predict key nutritional properties such as calories, mass, and macronutrients. The workflow includes generating predictions on a held-out test set, saving results to CSV, and computing standard regression metrics to quantitatively assess performance. Designed as a clean, beginner-friendly implementation, NutriVision demonstrates how a practical meal-nutrition estimator can be built without relying on heavy frameworks like PyTorch, while still leveraging multimodal data for improved accuracy and interpretability.

Problem Statement

Accurately estimating the nutritional content of meals from images remains a challenging yet crucial problem across healthcare, fitness, and consumer applications. Traditional dietary assessment methods rely heavily on manual reporting, nutrition experts, or crowd-sourced labeling, all of which are time-consuming, error-prone, and impractical for large-scale or real-time use. As mobile photography becomes ubiquitous and large food-image datasets continue to grow, there is an increasing demand for automated, reliable, and user-friendly systems that can infer calories and macronutrients directly from visual inputs. This creates a significant opportunity to leverage modern computer vision and machine learning techniques for automated nutritional analysis.

However, food images introduce unique difficulties compared to typical object-recognition tasks. Dishes vary widely in appearance, preparation style, lighting conditions, and serving sizes. Even the same ingredient can exhibit diverse visual forms, and visually similar dishes may differ substantially in composition or caloric density. These challenges are amplified when estimating quantities such as mass or portion size, which require depth understanding, segmentation accuracy, and cross-modal reasoning between visual and textual data. The NutriVision project addresses these complexities by combining synchronized RGB, depth, and colorized depth images with structured dish- and ingredient-level metadata, enabling a multimodal learning framework that leverages both visual and tabular information.

The practical objective of NutriVision is to create a simplified, efficient, and deployable pipeline capable of predicting key nutritional attributes—such as calories, protein, fat, carbohydrates, and mass—from overhead and side-angle images of dishes. By integrating image-based feature extraction, tabular MLP processing, and lightweight fusion techniques, the project aims to produce a system that balances accuracy with accessibility, ensuring that the solution can run on standard consumer hardware. Ultimately, this project aims to demonstrate that multimodal learning can significantly reduce dependency on manual dietary assessment, making nutritional estimation faster, more scalable, and more accessible to real-world users.

Dataset Overview

Nutrition5k is a publicly available dataset of visual and nutritional data for ~5k realistic plates of food captured from Google cafeterias using a custom scanning rig. Its key features are :

- Scans data for 5,006 plates of food, each containing:
  - 4 rotating side-angle videos
  - Overhead RGB-D images *(when available)*
  - Fine-grained list of ingredients
  - Per-ingredient mass
  - Total dish mass and calories
  - Fat, protein, and carbohydrate macronutrient masses
- Official train/test split
- Nutrition regression eval scripts

- Dataset Contents:

1. Side-Angle Videos

Video recordings were captured using 4 separate Raspberry Pi cameras (labeled A-D) at alternating 30 degree and 60 degree viewing angles.

2. Overhead RGB-D Images

The imagery/realsense_overhead/ directory contains RGB, raw depth, and colorized depth images organized by dish ID. Raw depth images are encoded as 16-bit integer images with depth units of 10,000 (i.e. 1 meter = 10,000 units).

3. Ingredient Metadata

The ingredient metadata CSV (metadata/ingredient_metadata.csv) contains a list of all ingredients covered in the dataset's dishes, their unique IDs, and per-gram nutritional information sourced from the USDA Food and Nutrient Database.

4. Dish Metadata

The dish metadata CSVs (metadata/dish_metadata_cafe1.csv and metadata/dish_metadata_cafe2.csv) contain all nutrition metadata at the dish-level, as well as per-ingredient mass and macronutrients.

5. Train/Test Splits

Official Train Test split available.

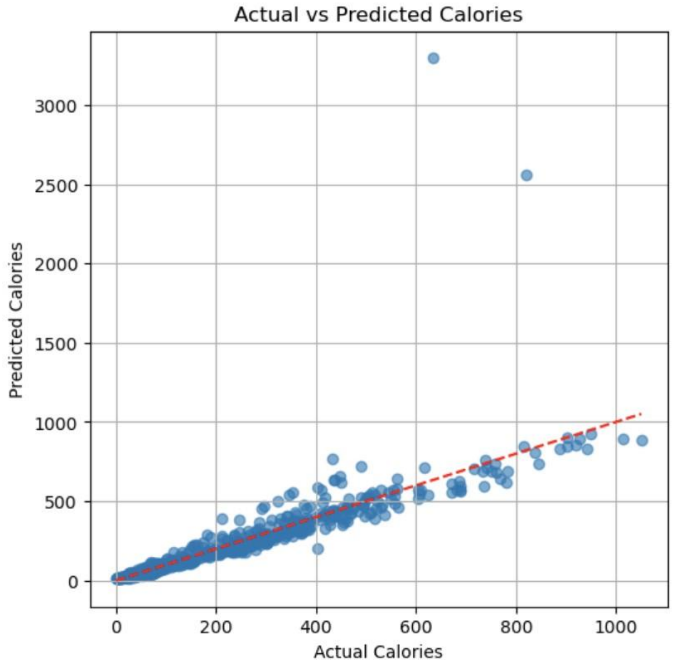
- Dataset Link: https://github.com/google-research-datasets/Nutrition5k?tab=readme-ov-file#download-data

Results

The proposed NutriVision system was evaluated on a held-out test set using standard regression metrics across all target nutritional attributes. The model demonstrated strong predictive capability for several key nutrients. For total calories, the system achieved an MAE of 44.573, RMSE of 151.389, and an R^2 of 0.490, indicating moderate explanatory power and stable estimation performance. Predictions for total mass showed similar behavior, with an MAE of 58.188, RMSE of 114.607, and R^2 of 0.428, reflecting reasonable accuracy given the inherent variability in food quantity and portion size.

Macronutrient predictions also followed encouraging trends. The model predicted total protein with an MAE of 7.306, RMSE of 13.774, and an R^2 of 0.475, and total fat with an MAE of 6.136, RMSE of 10.144, and an R^2 of 0.485, showing appreciable alignment between estimated and ground-truth nutritional values. These results indicate that the architecture effectively captures visual and tabular cues that correlate with protein- and fat-rich food components.

However, performance on total carbohydrates was comparatively weaker, with an MAE of 10.608, RMSE of 18.047, and an R^2 score of –0.073, suggesting that carbohydrate content is more difficult to infer from RGB and depth cues alone. This outcome highlights the need for further refinement—potentially through improved ingredient-level reasoning, enhanced image representations, or incorporating additional metadata. Overall, the results validate the effectiveness of the proposed multimodal pipeline while also identifying avenues for future improvement.



Actual vs Predicted Calories

Conclusion

The NutriVision project set out to design and implement an automated system capable of estimating the nutritional properties of a dish—specifically calories, mass, protein, fat, and carbohydrates—directly from RGB, depth, and colorized depth imagery combined with structured metadata. Through systematic preprocessing, dataset curation, multimodal feature extraction, and deep learning–based regression modeling, the project successfully demonstrated the feasibility of inferring dish-level nutritional metrics without manual intervention. The end-to-end workflow integrated image-based representation learning with tabular ingredient-level features, showing how multimodal fusion can provide richer and more context-aware predictions compared to using vision or metadata alone.

The experimental results indicate that the model captures meaningful relationships across modalities. Metrics for total calories, mass, protein, and fat achieved moderate accuracy with R^2 scores between 0.42 and 0.49, demonstrating that the architecture was able to generalize nutritional patterns from diverse visual and tabular representations. Despite these positive outcomes, carbohydrate estimation proved more challenging, yielding a negative R^2 value (-0.073), suggesting that carbohydrate content may be less visually inferable or require additional features not captured by the current dataset or architecture. These findings collectively highlight both the strengths and limitations of multimodal nutrition estimation and set a clear direction for targeted improvements.

Looking ahead, the project provides fertile ground for enhancing both the dataset and the model. Improvements could include incorporating 3D reconstruction for more accurate portion size estimation, experimenting with more advanced multimodal transformers, or expanding the dataset to include greater variation in lighting, plating styles, and ingredient complexity. Additional metadata, such as cooking methods or ingredient density values, may further improve carbohydrate estimation and overall performance. Deployment-oriented optimizations—such as model compression or on-device inference—could broaden real-world applicability in mobile nutrition tracking or clinical monitoring settings. Ultimately, the NutriVision framework serves as a solid foundation on which more sophisticated, robust, and generalizable nutrition estimation systems can be built in future research and development.