Beyond the Nutrition5k Project: Data Curation and Deep Learning Algorithms to Predict the Nutritional Composition of Dishes from Food Images

Coluccia Sergio[1], Bianco Rachele[2], Marinoni Michela[1], Falcon Alex[3], Fiori Federica[2], Serra Giuseppe[3], Ferraroni Monica[1,4], Parpinel Marial[3], and Edefonti Valeria[1,4]

(1) Branch of Medical Statistics, Biometry and Epidemiology "G. A. Maccacaro", Department of Clinical Sciences and Community Health - Dipartimento di Eccellenza 2023-2027, Università degli Studi di Milano, Milano, Italy.
(2) Department of Medicine - DMED, Università degli Studi di Udine, Udine, Italy.
(3) Department of Mathematics, Computer Science and Physics - DMIF, Università degli Studi di Udine, Udine, Italy.
(4) Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy.
equally contributed to the paper

CORRESPONDING AUTHOR: *Coluccia Sergio, sergio.coluccia@unimi.it*

INTRODUCTION

In recent years, artificial intelligence (AI) has emerged as a powerful tool to overcome limitations of traditional dietary assessment methods such as 24-hour recalls, food frequency questionnaires, and dietary records [1, 2]. Nevertheless, the success of AI models heavily depends on high quality, well-curated data. Pre-processing—handling missing values, outliers, and inconsistencies—is essential to ensure reliable model performance [3, 4]. The Nutrition5k project [5] is the first to adopt Deep Convolutional Neural Networks for the 2D direct prediction of mass and nutritional composition of dishes.

AIMS

We used the US-based Nutrition5k project to evaluate the performance of various deep learning (DL) algorithms, and to compare them in predicting mass, energy, and the macronutrient content from food images. We explored different ground truth configurations (by combining data curation with two country-specific food composition databases—FCDBs) and checked if there were specific dishes consistently mispredicted by most algorithms, and what common features they shared.

METHODS

Within the Nutrition5k project, mass (grams), energy (kcal), protein, fat, and carbohydrates (grams) contents were provided for each of the 5006 dishes as sum of nutritional values of single ingredients derived from the US-FCDB. In a previous publication [6], we have matched the US dishes with their Italian nutritional composition. This gave birth to four versions of the Nutrition5k dataset, specifically obtained as ground truths by crossing country-specific FCDBs with ingredient-mass correction of outlier dishes.

We chose Inception_V3_IMAGENET1K_V1 (IncV3, the updated version of the IncV2 proposed in [5]), ResNet101_IMAGENET1K_V2, ResNet50_IMAGENET1K_V2, ViT_B_16_IMAGENET1K_SWAG_E2E_V1 (ViT-B-16), built in two variants (2+1 and 2+2), and pretrained via the open-source ImageNet. IncV3_2+2 was our benchmark algorithm as in [5]. To ensure reproducibility, we adopted the same pipeline as in the Nutrition5k project for train/test split of dishes, loss function, frame preprocessing, and performance metrics (root mean squared error, mean absolute error – MAE – and its percentage – MAPE).

Dish-specific (raw, absolute) differences between predicted and observed values of the target variables on the test set (n=676) were evaluated across datasets and algorithms (160 predictions per dish), by considering: (1) percentages of perfect, adjacent, and opposite agreement among quartile-based categories, and unweighted Cohen's kappa statistics, and 2) Bland-Altman plots.

We defined "incorrectly predicted dishes" dishes as those that for 7 or 8 DL algorithms (1) exceeded the 95% limits of agreement in the Bland-Altman plots and (2) had the highest 5% of absolute differences across target variables and datasets. Their dish frames were manually inspected and further removed when needed. The "incorrectly predicted dishes" were then grouped based on similarity in content.

DOI: 10.54103/2282-0930/29429

A sensitivity analysis was carried out to study whether energy content should be directly predicted by DL algorithms or deterministically calculated by summing up predicted macronutrients multiplied by the corresponding conversion factor. This led to three scenarios: the 5-task predicted energy content (main analysis), the 5-task computed energy content (energy calculated based on macronutrients predicted together with energy), and the 4-task computed energy content (no energy prediction potentially imporving macronutrient prediction).

RESULTS

The median dish to be predicted on the test set had a mass of 142 g, energy content of 164.5 kcal, 8.3 g of protein, 6.9 g of fat, and 11.3 g of carbohydrates. When dishes showed ingredients with extreme weight or composition, algorithms tended to pull their predictions toward the center of the distribution.

For the same dataset, IncV3s consistently showed the worst percentages of perfect agreement across all target variables. For a given algorithm, perfect agreement was generally higher in the corrected datasets, with the exception of protein. Similarly, Cohen's kappa values were lower for the IncV3s and higher for the corrected datasets.

Globally, mass and energy content had more similar and lower error metrics, followed by protein, carbohydrates, and fat (Figure 1). By dataset, IncV3s generally exhibited the worst performances. Ingredient-mass correction strongly improved performance metrics.

The incorrectly predicted dishes were 80, of which 12 were discarded (7 for discrepancies between ingredient names and images and 5 for image-related issues for all images). Beyond the corrected-portion-size group (5%), Salad-based (44%), Chicken-based (25%), Eggs-based (13%), and the Western-inspired breakfast foods (13%) groups were identified. From this list we removed a median of 60% of the original frames, which led to a slight reduction in MAPE values.

While comparing our three scenarios, we observed a gradient: performance was the highest in the 5-task predicted, then the 5-task computed and finally the 4-task computed energy content scenario, advancing that energy prediction may partially compensate for macronutrient prediction errors, particularly those arising from image grounding issues. The ViT-B-16's showed minimal differences (~<7%) across scenarios.

CONCLUSIONS

We investigated the use of the Nutrition5k dataset for directly predicting the nutritional composition of dishes (including mass) using 2D images. All six selected algorithms outperformed the benchmark IncV3_2+2, as well as the lighter IncV3_2+1. Data curation, especially ingredient-mass correction, is critical in influencing algorithm performance.

REFERENCES

1. Kirk D., Catal C., Tekinerdogan B., Precision Nutrition: A Systematic Literature Review. Comput. Biol. Med. 2021, 133.
2. Boushey C.J., Spoden M., Zhu F.M. et al., New Mobile Methods for Dietary Assessment: Review of Image-Assisted and Image-Based Dietary Assessment Methods. Proc. Nutr. Soc. 2017, 76, 283–294.
3. Aldoseri A., Al-Khalifa K.N., Hamouda A.M, Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. Appl. Sci. 2023, 13.
4. Budach L., Feuerpfeil M., Ihde N., Nathansen A. et al., The Effects of Data Quality on Machine Learning Performance. 2022, 1–40.
5. Thames Q., Karpur A. Norris W., Xia F., et al., Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2021; p. 8899–8907.
6. Bianco R., Marinoni M., Coluccia S., et al. Tailoring the Nutritional Composition of Italian Foods to the US Nutrition5k Dataset for Food Image Recognition: Challenges and a Comparative Analysis: 2024; Nutrients. 2024 Oct 1;16(19):3339.

DOI: 10.54103/2282-0930/29429

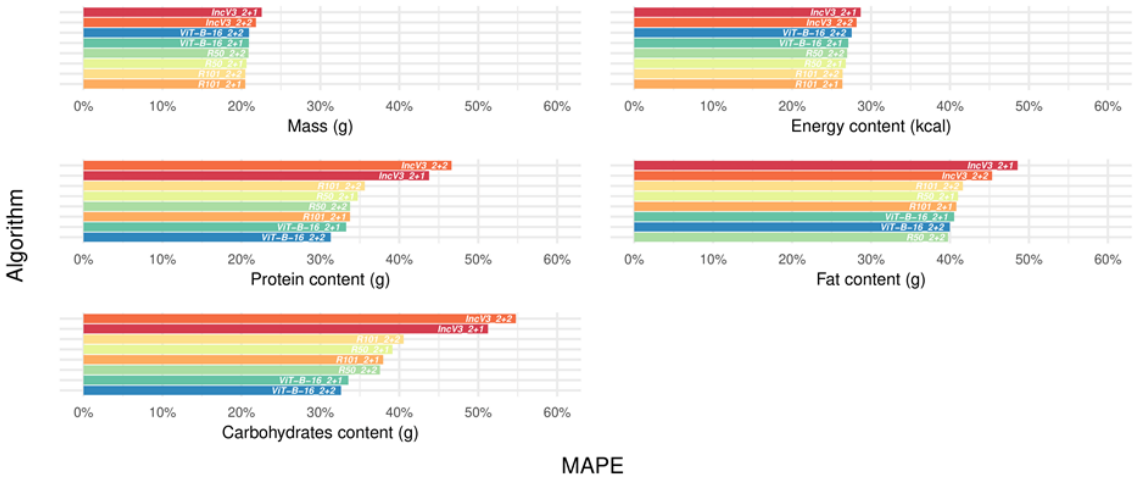

Figure 1. Median error, as measured by mean absolute percentage error, for single target variables and algorithms across datasets, before frame filtering. Abbreviations: MAPE, Mean Absolute Percentage Error.

DOI: 10.54103/2282-0930/29429