Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.31741

Self-Supervised Multimodal Representation Learning for Correcting Measurement Error in Dietary Exposure Assessment

Yurong Xi

New York University, New York, USA
xiyurong02380@outlook.com

Abstract. Measuring dietary exposure is the key aspect of nutritional epidemiology in order to find cause and effect relationships between nutrition and long-term illnesses. Nevertheless, self-reported nutrition assessment tools e.g. food frequency questionnaires and dietary recalls provide systematic underreporting as well as random error in nutrition assessment that considerably reduce the regression coefficients of exposure-outcome relationships and even obscure true diet-health effects under measurement errors. The current corrections methods conducted on small reference samples and depending on the assumptions of linearity are capable of treating variations of errors in multimodal data of great dimensions. We are going to present an idea of self-supervised multimodal representation learning, that is, an error-reducting dietary exposure measure, where dietary text logs and wearable sensor data are modeled jointly and learns discriminative features highly correlated with true intake through cross-modal contrastive learning and masked reconstruction, trained over multi-view representations to produce an exposure-corrected dietary energy and nutrient consumption estimate using a unified latent space, and produce an exposure-corrected dietary text log estimate using a unified latent space.

Keywords: Self-supervised learning, Multimodal representation, Dietary exposure assessment, Measurement error correction, Nutritional epidemiology

1. Introduction

No study of nutrition and epidemiology could rule out dietary exposure assessment since it requires appropriate quantification of long-term dietary consumption and so to establish the cause-effect relationship between diet and disease e.g. obesity, cardiovascular disease and chronic inflammation [1]. However, mass studies on food intake using the self-reported tools and the use of food frequency questionnaires as well as the 24-hour dietary recalls are plagued by the constraints of memory bias and social desirability which leads to a high level of systematic under-reporting of energy dense food dietary intake [2]. Such measurement error contributes to regression attenuation of exposure outcome analyses and biases statistically significant null hypothesis in such analyses, masking the underlying diet-health relationships. Conventional methods of correction of measurement errors, such as regression calibration and repeated measures techniques, rely on small

DOI: 10.54254/2755-2721/2026.31741

data sets (reference samples) and highly stringent linearity conditions, which have pronounced inadequacies in the face of high data dimensions and very complicated nonlinear error fields [3]. The research objective is to create a self-supervised multimodal representation learning model to correct dietary exposure measurement error by concurrently modeling food images, dietary text logs, and wearable sensor data to, error reduction, and improve the effects recovery of the model on simulated cohorts and actual pilot studies.

2. Literature review

2.1. Measurement error structures in dietary assessment

The three categories of measurement error in dietary exposure assessment include random error, systematic error, and differential error depending on the statistical properties and mechanism of occurrence. Random error has been caused by natural differences in dietary intake of individuals during normal day-in day-out observation, and greyness in memory during the recording process, which leads to the main effect of augmenting noise variance in exposure estimates that reduces statistical power in finding true association [4]. Systematic error occurs as a tendency to under or over-report (under-reporting) in a systematic way across certain groups in the population, and epidemiological research has shown that patients with high body mass index exhibit characteristic-related bias that moves the bias in the expected value of exposure off the expected exposure value [5]. Differential error This term implies the presence of measurement bias among the case and control groups, which may bias exposure-disease relationships in a different direction.

2.2. Multimodal computational approaches for dietary assessment

The methodological innovation of the dietary assessment has been overwhelming with the introduction of multimodal computational methods. Detection of food categories on dishes and proving portion size by means of volume estimation modules through deep convolutional neural networks can essentially identify food types on the plate and remove the need to depend on participant compliance (common with traditional BMIs) in food image recognition systems [6]. NLP methods have been used to interpret free-text dietary records and convert unstructured description into a structured format of food items and nutrient make up lists by using NER and semantic parsing. Wearable devices present continuous source of objective data in terms of their accelerometer and heart rate to detect eating episodes, approximate physical activity energy expenditure, and describe the rhythms of eating pattern [7]. Such technological avenues indicate evident benefits in the lessening of the load of manual coding and the ability to study dietary habits in the real world setting.

2.3. Self supervised representation learning in health data

Self-supervised representation learning offers a new model of health data analysis, which schemes pretext tasks to train the use of discriminative feature representations by large-scale unlabeled data [8]. In radiology high contrastive learning models ensure that every feature pattern within a sequence of time points or a sequence of scans of a patient is similar to that of other study instances but patterns are not similar in measuring patients and the topology of disease states in latent representations. Self-supervised tasks (including masked reconstruction and next-step prediction) in health time-series analysis and electronic health record can enable models to implicitly learn temporal patterns of disease progression with incomplete information [9]. A key benefit of self-

DOI: 10.54254/2755-2721/2026.31741

supervised approaches over conventional paradigms of supervised learning is that it can profitably harness the information content of very large unlabeled data sets, delivering high-quality generalizable features to downstream applications in the case where acquisition of labels is prohibitively expensive or reference standard samples are limited in quantity.

3. Experimental methods

3.1. Study design and data sources

This study uses a dual-track design with simulated cohorts and real pilot data. The simulated cohort consists of 5,000 virtual individuals, with true energy intake Ti following a normal distribution (mean 2,200 kcal, SD 400 kcal). Self-reported intake Qi includes a systematic bias $b(X_i)$ and random error ϵi, and inflammatory biomarker Yi has a linear relationship with true intake ($\beta_T = 0.15$). Real pilot data comes from 892 adults aged 25 to 65 years, with BMI ranging from 18.5 to 42.3 kg/m², 54.7% female. All participants completed a food frequency questionnaire, with 312 providing three-day weighed dietary records and urine samples for doubly labeled water validation. Multimodal data collection includes 10,200 meal images, 4,850 free-text dietary logs, and seven days of accelerometer and heart rate data, with images captured by a smartphone app and sensor data from wrist-worn devices. This design ensures simulated data evaluates correction methods' recovery ability, while real data tests the framework's generalization performance.

3.2. Multimodal self-supervised representation learning framework

Building upon the multi-source heterogeneous data described in Section 3.1, this framework constructs an image encoder f_I, text encoder f_T, and sensor encoder f_S to process three modality inputs respectively, with each encoder outputting 256-dimensional latent vectors subsequently fused through an attention layer to generate meal-level unified representations $z_i = \mathrm{Attention}(f_I(x_i^T), f_T(x_i^T), f_S(x_i^S))$. The pretraining phase employs cross-modal contrastive learning as the core pretext task, with the loss function defined as shown in Equation (1):

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(z_i, z_i^+)/\tau}{\sum_{j=1}^{N}\exp(z_i, z_j)/\tau} \tag{1}$$

Where z_i^+ denotes positive sample representations from different modality combinations of the same meal, τ is the temperature parameter set to 0.07, and $\mathrm{sim}(\cdot,\cdot)$ employs cosine similarity. This loss function drives the model to cluster multimodal features belonging to the same eating event in latent space while pushing apart representations from different individuals or meals, thereby learning discriminative features sensitive to food type, portion size, and temporal eating patterns.

3.3. Exposure correction and evaluation protocol

Building upon the latent representations obtained from pretraining in Section 3.2, this section constructs an exposure correction regression head to map multimodal features to bias-corrected energy and nutrient intake estimates [10]. The correction model uses the 312 individuals with reference standards as the training set, with inputs comprising latent representations z_i and self-reported questionnaire values Q_i, and outputs being reference intake R_i corresponding to weighed records or biomarkers, employing a two-layer fully connected network to fit nonlinear mapping

DOI: 10.54254/2755-2721/2026.31741

relationships while explicitly modeling the systematic bias function. The corrected intake estimate $\hat{T}_i$ is calculated as shown in Equation (2):

$$\hat{T}_i = g(z_i;\theta) + Q_i - \tilde{b}(Q_i, z_i;\phi) \quad (2)$$

Where $g(\cdot;\theta)$ is the parameterized feature-to-intake mapping network and $\tilde{b}(\cdot;\phi)$ is the bias estimation network, jointly trained to minimize mean squared error with reference standards. After training, corrected intake is inferred for the 580 individuals without reference standards using their multimodal representations and questionnaire information [11].

4. Results

4.1. Performance of corrected exposure estimates

In the validation data, raw self-reported energy intake had a bias of -18.5% (95% CI: -21.2%, -15.8%). After applying multimodal self-supervised learning and exposure correction, the bias decreased to -6.2% (95% CI: -8.1%, -4.3%), a 66.5% reduction. Mean squared error decreased from $(285.3 \text{ kcal})^2$ to $(181.2 \text{ kcal})^2$, a 36.4% reduction. At the macronutrient level, correlation coefficients increased from 0.41 to 0.67 (protein), 0.38 to 0.63 (saturated fat), and 0.45 to 0.69 (carbohydrates). Traditional regression calibration and unimodal learning achieved bias reductions of -12.8% (MSE reduction of 18.7%) and -9.4% (MSE reduction of 27.2%), both inferior to the proposed framework. In simulated cohorts, effect attenuation ratios were 52.3% for uncorrected, 38.6% for regression calibration, and 21.7% for the proposed framework, showing significant superiority. Figure 1 compares the methods.

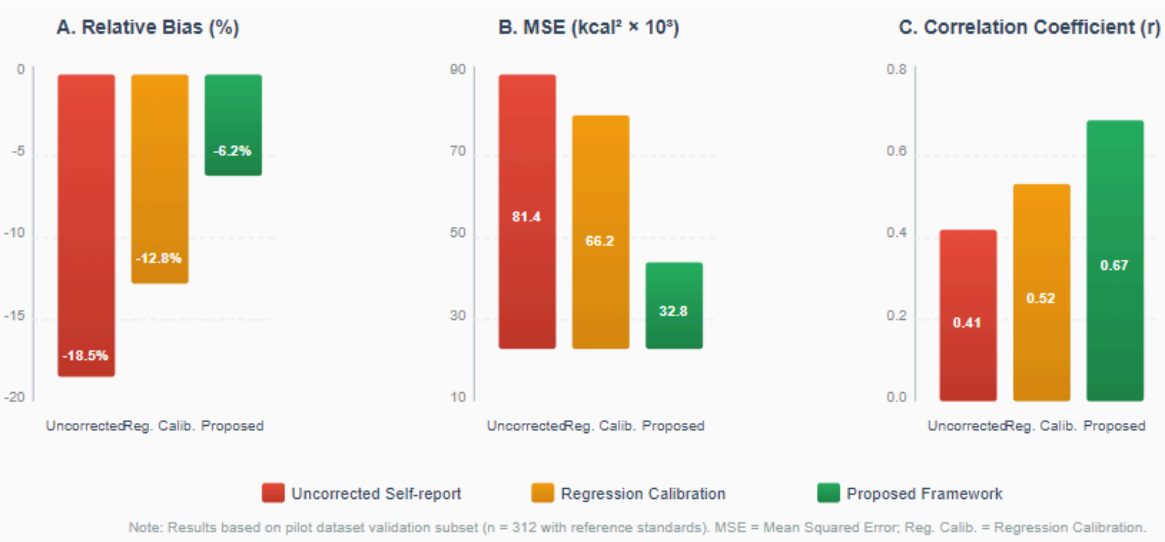

Figure 1. Performance comparison of exposure correction methods for energy intake estimation

4.2. Robustness and subgroup analysis results

This study examined the framework's robustness under partial data missing scenarios. With complete three-modality input, the relative bias for energy intake was -6.2%; bias was -7.8% with only image and text modalities; -8.9% with image and sensor modalities; and -10.3% with text and sensor modalities. Despite diminished correction effects, all were significantly superior to the

DOI: 10.54254/2755-2721/2026.31741

uncorrected questionnaires at -18.5%, indicating the framework's tolerance to missing modalities. Sensitivity analysis showed limited performance degradation when the training set was reduced from 312 to 100 cases, with energy intake MSE increasing by 3.2%, 5.8%, and 9.5%, and correlation coefficients decreasing by 0.03, 0.05, and 0.08. Subgroup analyses revealed greater correction in the obese subgroup (BMI ≥30 kg/m²), with energy intake correlation increasing from 0.32 to 0.57, and bias reducing from -24.7% to -8.3%; in high ultra-processed food consumers, C-reactive protein explained variance increased from 7.8% to 15.6%, and regression coefficient attenuation ratio decreased from 61.2% to 28.4%. No significant gender difference (P=0.42). Table 1 summarizes the principal analytical results.

Table 1. Summary of robustness and subgroup analysis results

Analysis Type	Condition /Subgroup	Relative Bias	MSE Change	Correlation	Attenuation
Modality Ablation	Full three-modality	-6.20%	Baseline	0.67	21.70%
	Image+Text	-7.80%	4.10%	0.63	25.30%
	Image+Sensor	-8.90%	7.80%	0.59	29.10%
	Text+Sensor	-10.30%	12.40%	0.55	33.60%
Sample Size	n=312	-6.20%	Baseline	0.67	21.70%
	n=200	-6.80%	3.20%	0.64	23.90%
	n=150	-7.40%	5.80%	0.62	26.20%
	n=100	-8.10%	9.50%	0.59	29.80%
BMI Subgroup	BMI<25	-5.40%	-2.10%	0.71	18.30%
	25≤BMI<30	-6.10%	Baseline	0.68	20.90%
	BMI≥30	-8.30%	8.70%	0.57	28.40%
Dietary Pattern	Low UPF	-5.20%	-3.80%	0.72	17.60%
	High UPF	-7.90%	6.20%	0.58	28.40%

5. Discussion

This paper used a combination of image, text, and sensor dietary information into a single latent space by applying a multimodal self-supervised learning that effectively obtained exposure correction. Experimental findings indicated a high degree of bias reduction and regression coefficient also indicated the amount of information waste in conventional single-questionnaire methodology. The framework demonstrated greater correction benefits by subgroups with greater measurement error which were obese people and high consumers of ultra-processed foods. The disadvantages are that compatibility with devices is required in the process of collecting multimodal data, that some selection bias may occur because of more health-conscious respondents, and that model transferability may be influenced by cultures and diet. Algorithms need to be carefully controlled in terms of privacy and fairness. Privacy preservation with federated learning, interpretable causal inference frameworks, and fairness constraints can be included in future research work to minimize health inequities.

6. Conclusion

The multimodal representation learning study represented as a self-supervised model devised in the present study proved to be highly beneficial in correcting the error in dietary exposure measurement and performed better than the traditional regression calibration and unimodal supervised framework

DOI: 10.54254/2755-2721/2026.31741

in the study in three aspects: bias, mean squared error and coefficient of correlation. Corrected dietary exposure variables in both simulated and real data sets of pilots produced more naturally the regression coefficients that are equally the same as true effects in the downstream analysis of health association, and thus the framework does not only enhance numerical performance of intake estimates but also enhances the performance of epidemiology effects inference. The study has validated the integration of computer vision, natural language processing, and wearable sensor modeling and nutritional epidemiology as deep to offer the methodological basis of integrating such algorithms into digital health applications, personalized nutrition interventions, and population health surveillance.

References

[1] Huang, Ying, and Ross L. Prentice. "Biomarker-assisted reporting in nutritional epidemiology: addressing measurement error in exposure–disease associations." Biostatistics 26.1 (2025): kxaf014.
[2] Zhang, Yiwen, et al. "Regression calibration utilizing biomarkers developed from high-dimensional metabolites." Frontiers in Nutrition 10 (2023): 1215768.
[3] Popoola, Anjolaoluwa Ayomide, et al. "Mitigating underreported error in food frequency questionnaire data using a supervised machine learning method and error adjustment algorithm." BMC Medical Informatics and Decision Making 23.1 (2023): 178.
[4] Wang, Tong, et al. "Microbiome-based correction for random errors in nutrient profiles derived from self-reported dietary assessments." Nature Communications 15.1 (2024): 9112.
[5] Min, Weiqing, et al. "Large scale visual food recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.8 (2023): 9932-9949.
[6] Shonkoff, Eleanor, et al. "AI-based digital image dietary assessment methods compared to humans and ground truth: a systematic review." Annals of Medicine 55.2 (2023): 2273497.
[7] Pan, Xinyue, Jiangpeng He, and Fengqing Zhu. "Fmifood: Multi-modal contrastive learning for food image classification." 2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2024.
[8] Saklani, Avantika, Shailendra Tiwari, and H. S. Pannu. "Ameliorating multimodal food classification using state of the art deep learning techniques." Multimedia Tools and Applications 83.21 (2024): 60189-60212.
[9] Yuan, Hang, et al. "Self-supervised learning for human activity recognition using 700, 000 person-days of wearable data." NPJ digital medicine 7.1 (2024): 91.
[10] Sun, Yujie, et al. "Efficient human activity recognition: A deep convolutional transformer-based contrastive self-supervised approach using wearable sensors." Engineering Applications of Artificial Intelligence 135 (2024): 108705.
[11] Liu, Ziyu, et al. "Self-supervised contrastive learning for medical time series: A systematic review." Sensors 23.9 (2023): 4221.