NutriEar: Robust Nutrition-Aware Food Classification from In-Ear Acoustic Signals

Zoey Xiaochen Tan
University of Cambridge
Cambridge, United Kingdom
xt229@cam.ac.uk

Yang Liu
Florida State University
Tallahassee, United States
yl25r@fsu.edu

Kayla-Jade Butkow
University of Cambridge
Cambridge, United Kingdom
kjb85@cam.ac.uk

Cecilia Mascolo
University of Cambridge
Cambridge, United Kingdom
cm542@cam.ac.uk

Abstract

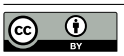
Convenient tracking of food intake is essential for linking diet to health, enabling personalised nutrition guidance, early metabolic risk detection, and prevention of chronic disease. Recent wearable sensing advances have begun to automate eating monitoring. However, these systems largely focus on detecting *when* users eat and only weakly address *what* they eat. In particular, state-of-the-art solutions typically cover only a narrow range of foods or textures and rely on strong assumptions about individual eating behaviour. Moreover, they overlook the nutritional implications most relevant to end users, limiting their usefulness for real-world dietary guidance. In this paper, we present *NutriEar*, an in-ear audio sensing system for nutrition-aware classification of food intake from chewing sounds. Rather than recognising arbitrary food types, *NutriEar* maps in-ear acoustics to an eight-class nutrition-texture taxonomy grounded in food science, capturing both dominant macronutrient role and mechanical texture. *NutriEar* records in-ear audio during eating, segments chewing events, and derives a hybrid representation combining engineered acoustic features with learned embeddings from supervised contrastive learning, enabling a compact nutrition-aware classification pipeline. Evaluation on a dataset collected from 15 users consuming over 30 food types under varied eating conditions shows that *NutriEar* achieves 80.18% average leave-one-subject-out (LOSO) accuracy and outperforms state-of-the-art baselines. These results highlight the untapped potential of earable audio sensing as a practical pathway toward everyday dietary monitoring with meaningful nutritional insights.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**.

Keywords

Earable Sensing, Dietary Monitoring, In-Ear Acoustic Sensing



SenSys '26, Saint Malo, France

ACM ISBN 979-8-4007-2309-4/2026/05
https://doi.org/10.1145/3774906.3802763

ACM Reference Format:
Zoey Xiaochen Tan, Yang Liu, Kayla-Jade Butkow, and Cecilia Mascolo. 2026. *NutriEar*: Robust Nutrition-Aware Food Classification from In-Ear Acoustic Signals. In *ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems (SenSys '26), May 11–14, 2026, Saint Malo, France.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3774906.3802763

1 Introduction

Accurate and continuous dietary tracking is fundamental to understanding how eating behaviour and nutrient intake shape long-term health [66, 76]. Consistently imbalanced macronutrient intake is associated with obesity [13], cardiovascular disease [50], and type 2 diabetes [2], underscoring the need for reliable, day-to-day measurements beyond sporadic clinical assessments. Yet in practice, diet assessment still relies heavily on self-report tools such as food diaries and 24-hour recalls, which suffer from recall bias, low compliance, and subjective interpretation [28, 55]. These limitations impede both personalised nutrition guidance and large-scale, data-driven interventions. There is a growing need for automated dietary sensing methods that can infer what people eat in everyday life, with sufficient granularity and reliability to capture nutrition-relevant patterns rather than coarse, user-reported proxies, serving as building blocks toward more comprehensive dietary assessment.

Existing automated approaches in wearable sensing have opened promising directions for unobtrusive dietary monitoring using cameras [32, 33, 40, 46, 62, 63, 77], microphones [3, 4, 9, 35, 38, 39, 41, 67, 68, 78], inertial sensors [20, 21, 70], and physiological signals [67, 68], but they only partially address this need. Most existing systems focus on detecting *when* eating occurs, providing only binary or coarse-grained episode detection that is informative about timing but largely agnostic to nutritional content. Robust inference of *what* is being consumed and thus its nutritional implications remains underexplored. Specifically, prior work typically (i) evaluates on a narrow and system-specific set of foods [3, 4], (ii) adopts heterogeneous and ad-hoc label spaces (e.g., brand names or visually distinct items) that are poorly aligned with nutritional meaning [38, 39, 78], and (iii) relies on strong assumptions about structured bite–chew–swallow cycles or controlled eating protocols that do not hold in everyday settings [59]. Consequently, existing systems provide limited and non-standardised resolution on food type and composition, making it difficult to deliver consistent,

nutrition-aware feedback that reflects macronutrient balance or textural quality of the diet.

To address this gap, we explore earphones as a practical and socially acceptable platform for regular dietary monitoring, given their comfort, proximity to chewing activity [15], frequent use during meals, and growing role in human bio-signal monitoring [12, 30, 44]. In this paper, we present *NutriEar*, an in-ear audio sensing system that infers nutrition-aware food types from chewing sounds. Rather than estimating exact food identities or quantities, *NutriEar* functions as a nutrition-aware sensing primitive, mapping bite-level chewing acoustics to an interpretable set of texture-nutrition classes encoding dominant macronutrient roles and mechanical texture properties.

Specifically, *NutriEar* is grounded in food-science principles: we model how macronutrient composition shapes food microstructure and texture, and how these textures generate characteristic chewing acoustics. Accordingly, we formulate in-ear chewing analysis as a classification problem over a novel, interpretable taxonomy (Section 2), rather than unconstrained food labels: bite-level predictions can be temporally aggregated into coarse macronutrient exposure profiles (e.g., distributions of carbohydrate-, fat-, and protein-dominant bites) for longitudinal dietary pattern analysis. Precise nutritional intake estimation, however, would additionally require bite-weight or portion-size modelling, which we leave for future work.

Challenges. Designing a robust, nutrition-aware food tracking system from in-ear audio presents three main challenges. **First**, eating behaviour varies widely across individuals and contexts, so methods that assume fixed bite–chew–swallow patterns or tightly controlled conditions often fail to generalise. **Second**, chewing acoustics are shaped by both food properties and user-specific factors such as jaw strength and oral geometry, meaning similar foods can sound different, while nutritionally distinct foods may sound similar. **Third**, texture and nutrition are related but not equivalent: foods with different macronutrient profiles can share the same coarse texture, so texture-only classification cannot reliably capture nutrition-relevant differences.

Our approach. *NutriEar* addresses these challenges with a sensing and learning pipeline that maps in-ear audio signals to nutrition-relevant food categories. **First**, to accommodate diverse and unconstrained eating patterns, we design a lightweight signal processing pipeline on in-ear audio to automatically detect chewing-related segments, without assuming fixed bite–chew–swallow cycles or controlled eating protocols. **Second**, to handle cross-individual variability, we employ a supervised contrastive learning framework combining engineered acoustic features that encourages embeddings of samples from the same texture–nutrition class to cluster together while separating embeddings from different classes. This structure-aware objective improves robustness across users, sessions, and recording conditions. **Third**, to reliably capture nutrition-relevant differences beyond coarse texture, we define an eight-class taxonomy grounded in food science [74] that couples mechanical texture characteristics with dominant macronutrient profiles (e.g., Crunchy-Fat, Soft-Protein). This taxonomy provides a compact, physically motivated, and interpretable target space for classification, aligning *NutriEar*'s outputs with macronutrient balance and textural diversity rather than unconstrained food identities.

Leveraging this taxonomy with both hand-crafted features and contrastively refined deep representations, *NutriEar* achieves 80.18% LOSO accuracy across eight nutrition-relevant classes. Evaluated on a self-collected in-ear audio dataset under varied eating conditions, *NutriEar* outperforms state-of-the-art baselines and maintains robustness under realistic variability, demonstrating the feasibility of earable audio sensing for practical nutrition-aware monitoring.

The main contributions of this paper are:

- We define a unified, nutrition-aware taxonomy that links mechanical texture properties with dominant macronutrient composition, providing a principled label space for nutrition-relevant food recognition from in-ear audio.
- We design *NutriEar*, a hybrid representation learning framework that combines engineered acoustic features with contrastively fine-tuned deep features to capture subtle, nutrition-related differences in chewing acoustics.
- We build and evaluate *NutriEar* on a self-collected in-ear audio dataset with over 30 food types from 15 users and diverse eating conditions, demonstrating 80.18% LOSO accuracy across eight nutrition-relevant classes and outperforming state-of-the-art baselines. These results establish the feasibility of using in-ear audio to produce structured, bite-level nutrition-aware predictions that can serve as primitives for future macronutrient-aware dietary monitoring systems.

2 Preliminaries

This section establishes the physical and nutritional foundations underlying *NutriEar*, drawing on principles from food science to link macronutrient composition, food texture, and chewing acoustics, and to motivate our nutrition-aware taxonomy.

2.1 Nutritional Foundations and Motivation

Macronutrients as sensing targets. Food intake supplies both energy and structural components that sustain bodily functions. Among all nutrients, **proteins**, **fats**, and **carbohydrates**, collectively termed **macronutrients**, are primary determinants of energetic balance, tissue maintenance, and metabolic health [23, 79]. Proteins support tissue repair and enzyme and hormone synthesis [19]; fats provide energy-dense storage and form cell membranes [6, 18]; carbohydrates are the principal immediate energy source [31]. Together, these macronutrients characterise the dominant functional role of a food (e.g., protein-dominant vs. fat-dominant), making macronutrient-aware categorisation a natural target for dietary sensing.

Need for composition-aware feedback. For everyday dietary management, the *balance* among macronutrients is often more critical than precise quantities [37, 72]. Chronic excess or deficiency (e.g., sustained high sugar intake, prolonged high-fat consumption, insufficient protein) disrupts metabolic homeostasis and contributes to obesity, cardiovascular disease, and type 2 diabetes [2, 13, 50]. These imbalances are frequently unintentional, driven by limited awareness of the nutritional profile of routine foods. Thus, sensing systems that provide interpretable, composition-aware feedback about *what kind* of foods are consumed, rather than only total intake or eating frequency, are important for preventive health and motivate our focus on macronutrient-informed food categorisation.

Table 1: Illustrative relationship between dominant nutrient, water content, and typical texture regimes.

Dominant Nutrient	Water Content	Typical Texture	Associated Class Name
Fat	Low	Hard, brittle	Crunchy-Fat (CF)
Fat	Medium	Firm, cohesive	Firm-Fat (FF)
Fat	High	Soft, creamy	Soft-Fat (SF)
Protein	Low	Firm, elastic	Firm-Protein (FP)
Protein	High	Soft, cohesive	Soft-Protein (SP)
Carbohydrate	Low	Hard, brittle	Crunchy-Carb (CC)
Carbohydrate	High	Soft, tender	Soft-Carb (SC)

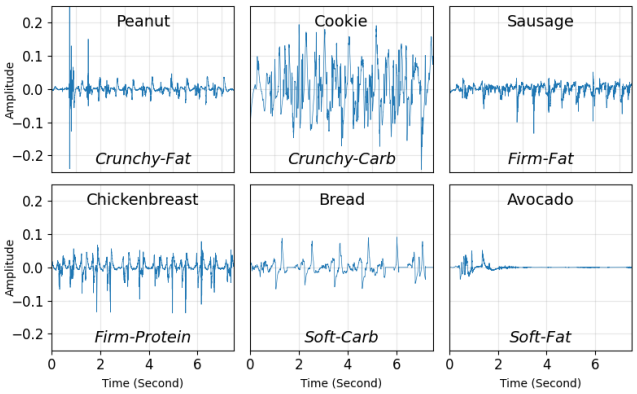


Figure 1: Illustrative in-ear acoustic signals for six foods.

Texture as a co-benefit. Beyond composition, the mechanical diversity of foods, especially their textures, also has implications for oral and systemic health [5]. Harder or chewier foods (e.g., nuts, raw vegetables) increase masticatory effort, stimulate saliva, support periodontal tissues, and improve bolus formation, whereas prolonged consumption of exclusively soft foods can weaken masticatory muscles and impair oral function [27, 52, 75]. A healthy diet therefore benefits from both nutritional balance and textural diversity. These dual aspects motivate us to model macronutrient composition and texture jointly as coupled, health-relevant dimensions in our sensing design.

2.2 Nutrition–Texture Links in Food Structure

Macronutrient composition fundamentally shapes food microstructure and thereby its mechanical texture. Textural attributes such as **hardness**, **brittleness**, **cohesiveness**, and **elasticity** emerge from the interplay between proteins, fats, carbohydrates, and water [11]. In this work, following common usage in food science [74], we use *hardness* to describe the force required to bite or deform a food, *brittleness* to describe the tendency to fracture with little deformation (e.g., crackers), *cohesiveness* to describe how well the food mass holds together during chewing (e.g., sausage, tofu), and *elasticity* to describe the extent to which a food deforms and then partially springs back (e.g., certain cheeses or lean meats). Table 1 provides an illustrative (non-exhaustive) mapping from dominant nutrient and water content to typical combinations of these texture attributes; we use this mapping as a conceptual guide when designing our texture–nutrition taxonomy.

Fat-rich foods. Low-moisture, fat-rich products (e.g., fried snacks) form porous, brittle matrices that yield hard, crunchy textures [7, 10, 53]. High-moisture fat-rich foods (e.g., avocado, soft cheeses) behave as lubricated emulsions where fat reduces internal friction, producing soft, creamy textures [22, 57, 64]. Medium-moisture systems such as sausages lie in between, with fat plasticising the matrix and creating firm yet cohesive textures [16, 49, 56].

Protein-rich foods. Protein-dominant foods form crosslinked fibrillar networks whose density, hydration, and denaturation level control firmness and elasticity. Lean meats and aged cheeses, with extensive protein–protein interactions and limited mobile water, are typically firm and elastic [26, 45, 69]. Higher-moisture protein systems such as fresh tofu or scrambled eggs exhibit softer, cohesive textures due to fewer strong interactions and more mobile water [29].

Carbohydrate-rich foods. Carbohydrate-dominant foods derive texture from starch gelatinisation and retrogradation. Dehydrated or aerated matrices (e.g., crackers) are hard and brittle, whereas hydrated starch gels (e.g., fresh bread, cooked rice) are soft and tender [25, 65].

In summary, dominant macronutrient type together with water content leads to recurring texture regimes. These regularities provide a mechanistic basis for grouping foods into a small number of interpretable texture–nutrition categories, which we subsequently use as target classes for *NutriEar*'s nutrition-aware classification.

2.3 Texture–Acoustics Links During Chewing

Building on the nutrition–texture relationships in Section 2.2, we next examine how these nutrition–texture regimes manifest in chewing acoustics. We conducted a preliminary measurement in which a participant consumed six representative foods from distinct texture–nutrition regimes while wearing in-ear microphones; Fig. 1 shows example waveform segments.

When foods are bitten and chewed, their mechanical properties shape the acoustic signatures captured by the in-ear microphone. Among the texture dimensions introduced earlier, **hardness**, **brittleness**, **cohesiveness**, and **elasticity** exhibit clear qualitative acoustic correlates that guide our feature design:

Hardness. Hard, low-moisture foods require higher bite force and produce pronounced impact sounds with sharp onsets and larger amplitudes, as seen for peanut and cookie in Fig. 1. Softer foods produce lower-amplitude, smoother fluctuations (e.g., bread in Fig. 1).

Brittleness. Brittle, porous structures fracture abruptly, generating multiple high-amplitude, broadband transients (cracks and snaps), visible as irregular, spiky waveforms, exemplified by the cookie signal in Fig. 1.

Cohesiveness and elasticity. Cohesive and elastic foods deform gradually and resist fragmentation, resulting in more periodic, lower-amplitude chewing sounds with smoother envelopes and fewer sharp fracture peaks. Sausage, chickenbreast and bread

exhibit such patterns, while very soft, high-moisture foods like avocado show quickly decaying, low-energy signals with minimal fracture events.

Because these patterns stem from the structural mechanisms described in Section 2.2, foods from different texture–nutrition regimes tend to occupy distinct regions in acoustic space. Even when two foods share a coarse label such as "crunchy," differences in fat, protein, or carbohydrate matrices produce subtle yet learnable variations in chewing sounds. For example, although both peanuts and cookies fracture abruptly, the higher fat content in peanuts leads to less brittle fracture and less erratic acoustic patterns than cookies. This coupling between nutrition, texture, and acoustics directly motivates *NutriEar*'s design and differentiates it from existing studies: *we treat in-ear chewing audio as evidence for a compact, physically grounded set of texture–nutrition classes aligned with macronutrient balance and textural diversity, rather than unconstrained food identities.*

2.4 Our Taxonomy

Taken together, these observations suggest a hierarchical relationship: changes in macronutrient composition shape food microstructure and texture, and these texture properties in turn govern the chewing acoustics captured by the in-ear microphone.

To balance interpretability, robustness, and acoustic separability, *NutriEar* formulates dietary sensing as a *macronutrient-aware classification* problem over an eight-class taxonomy that jointly encodes dominant macronutrient role and texture: *Crunchy-Fat*, *Crunchy-Carb*, *Firm-Fat*, *Firm-Protein*, *Soft-Fat*, *Soft-Carb*, *Soft-Protein*, and *Mixed*. We focus on macronutrients because they primarily shape food structure, directly influencing mechanical properties and chewing acoustics. In contrast, micronutrients (e.g., sodium or trace minerals) mainly affect flavour or biochemical properties and do not produce clear mechanical signatures in bite-level sounds. These classes are derived from the conceptual mapping in Table 1 and the acoustic patterns as showcased in Fig. 1:

- Dehydrated, fat- or carbohydrate-rich foods with brittle matrices naturally fall into *Crunchy-Fat* or *Crunchy-Carb*, characterised by high-energy, transient signals.
- Dense, cohesive matrices dominated by fat or protein populate *Firm-Fat* and *Firm-Protein*, exhibiting more periodic, mid-amplitude chewing sounds.
- Hydrated, predominantly fat-, carb-, or protein-based foods map to *Soft-Fat*, *Soft-Carb*, and *Soft-Protein*, with smoother envelopes and lower amplitudes.
- The *Mixed* class captures composite foods and heterogeneous bites that combine multiple structures and nutrients, reflecting realistic eating patterns and avoiding unreliable forced assignments.

This taxonomy defines decision boundaries grounded in established links between macronutrient composition, microstructure, texture, and chewing acoustics, targeting the nutritional dimensions most directly inferable from mastication. Each class corresponds to a coarse-grained nutritional profile (e.g., high-fat crunchy snacks vs. soft protein sources) and a distinct, sound-distinguishable texture regime. *NutriEar* leverages this structured label space to perform nutrition-aware classification from in-ear audio, producing outputs that are both learnable from the signals and directly interpretable for real-world dietary monitoring.

3 System Design

NutriEar uses a modular pipeline to convert raw in-ear chewing audio into nutrition-aware bite-level predictions (Fig. 2(Left)). It first applies adaptive preprocessing to suppress environmental noise and device artefacts, then detects biting events and extracts short post-bite chewing segments. From each segment, *NutriEar* derives a hybrid representation that combines texture- and motion-related features with embeddings from a contrastively fine-tuned CLAP encoder [14, 80]. These features are fed into a lightweight neural classifier, and segment-level outputs are aggregated through weighted voting. This design enables robust inference of nutrition-texture categories across users and natural eating conditions.

3.1 Data Pre-processing

After acquiring the in-ear acoustic signals generated from food consumption, we first perform signal pre-processing to filter the noise. This enables subsequent feature extraction to focus only on meaningful portions of the signals.

Low-pass filter based on chewing sound frequency. Human mastication produces rhythmic bursts with fundamental frequencies around 1-2 Hz [61], and prior acoustic analyses show that most eating-related sound energy is concentrated below 5-6 kHz, with high-frequency transients above 5 kHz arising primarily from crispy or brittle foods [73]. This can also be observed in Fig.3 as most sounds are distributed below 6 kHz. We therefore filter the raw signal, sampled at 44.1 kHz, using a tenth-order Butterworth low-pass filter with a cut-off at 6 kHz, to retain informative chewing and fracture components while discarding irrelevant high-frequency environmental noise.

Adaptive noise filtering. We observed harmonic noise generated by the hardware prototypes, manifested as regular horizontal lines in the spectrogram. To eliminate the impact of these hardware noises which varies across experiment subjects due to the environment and wearing of the device, we use Median-filtering Harmonic Percussive Source Separation (HPSS) filters and utilise an automatic kernel-size selector to balance noise suppression and transient preservation. We compute the *horizontal residual energy*

$$R_h = \frac{1}{F}\sum_f \text{median}_t\, S_p(f,t),$$

which measures the time-stationary background in the percussive output, and the *percussive preservation ratio*

$$P_p = \frac{\Phi(S_p)}{\Phi(S)}, \quad \Phi(S) = \tfrac{1}{T-1}\sum_t \sqrt{\sum_f (\Delta_t^+ S(f,t))^2},$$

where $\Phi(S)$ is the spectral flux capturing vertical transients and $\Delta_t^+ S(f,t) = \max(S(f,t) - S(f,t-1), 0)$ retains only positive spectral changes corresponding to chewing onsets. The optimal kernel minimises $J(k_t, k_f) = R_h/(P_p + \epsilon)$, yielding $(k_t^*, k_f^*) = \arg\min J$ that balances tonal-noise removal against transient retention for robust adaptive filtering.

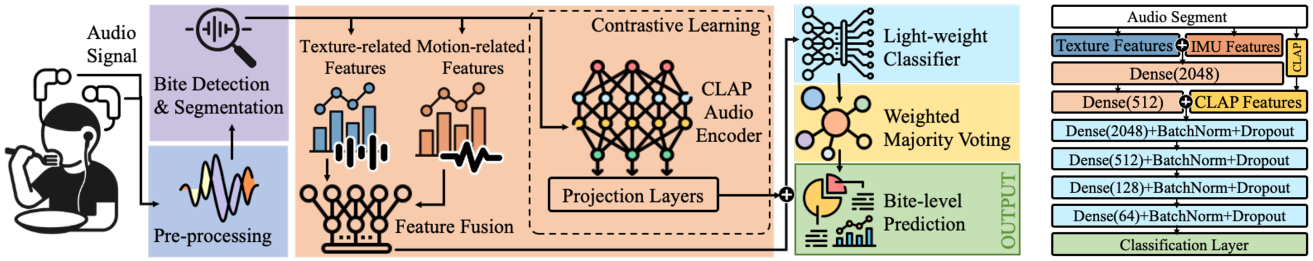


Figure 2: Left: System Overview. Right: Model Architecture.

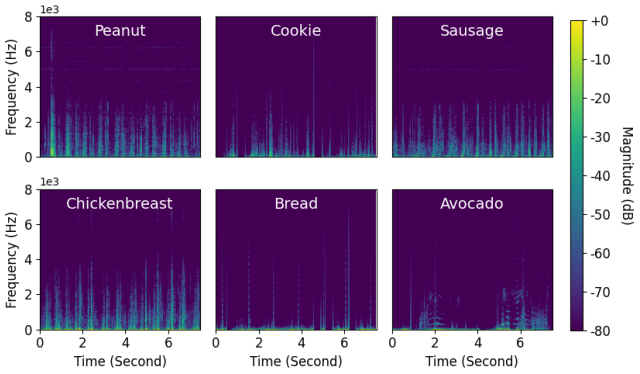


Figure 3: Acoustic Signals from 6 Different Foods.

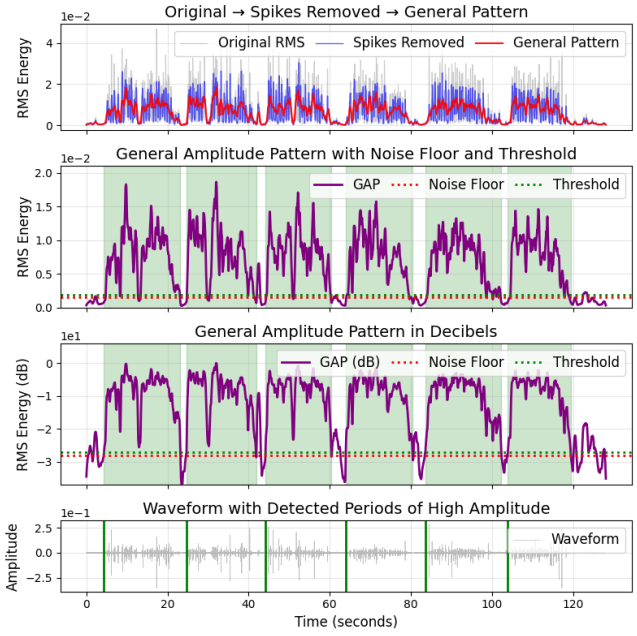


Figure 4: Example Plots of Biting Events Detection.

3.2 Bite Detection and Segmentation

To accommodate diverse and unconstrained eating patterns, we first detect biting events directly from the audio signal without making any assumptions about how individuals bite, chew, swallow or their orders. We focus on the first few chewing cycles after each bite, as they contain the most distinctive, unattenuated fracture sounds before sounds become muffled as chewing progresses.

Bites are identified using a *General Amplitude Pattern* (GAP) detector that analyses the smoothed amplitude envelope to accommodate inter-subject variability. Unlike onset-based methods relying on spectral peaks, GAP isolates sustained above-noise energy segments while suppressing transient spikes, enabling robustness under varying gains and environmental conditions.

Energy framing and smoothing. While chewing produces rapid rhythmic fluctuations in signal amplitude, each bite involves an initial mouth opening followed by a relatively longer low-energy period before the onset of the chewing cycles.

To emphasise these slower bite-level transitions and suppress high-frequency chewing rhythms, we compute the short-time root-mean-square (RMS) energy of the waveform y as

$$e_t = \sqrt{\frac{1}{L}\sum_{n=1}^{L} y_t[n]^2},$$

where we used a frame length L of 2048 and hop size of 512. The energy sequence e_t is then median-filtered over K_{spike} frames (approximately 200 ms) to suppress fast oscillations caused by consecutive chewing strokes (Fig.4(a)), and then smoothed with a moving-average filter of window $K_{\text{gap}} = 3K_{\text{spike}}$ to yield the envelope g_t that highlights the broader amplitude variations associated with bite initiation and mouth opening, illustrated in Fig.4(b).

Noise floor estimation. To robustly estimate the ambient sound level and adapt to varying recording environments, we derive a noise floor directly from the smoothed amplitude envelope. The envelope g_t is first converted to a relative decibel scale $g_t^{(\text{dB})} = 10\log_{10}(g_t^2/g_{\text{max}}^2)$, where $g_{\text{max}} = \max_t g_t$, as shown in Fig.4(c). The logarithmic compression aligns with human auditory perception, where equal ratios rather than equal differences in amplitude correspond to perceptually uniform changes in loudness. This makes thresholding and percentile-based statistics more stable across recordings with different absolute energy levels.

The noise floor is then defined as the 10th percentile of $g_t^{(\text{dB})}$. This percentile-based estimate provides a stable measure of the

environmental background level while remaining insensitive to short impulsive peaks in the signal.

Adaptive thresholding and period detection. Frames exceeding an adaptive threshold $T_{\text{dB}} = N_{\text{dB}} + \Delta_{\text{dB}}$ are marked as active, where Δ_{dB} represents the required signal-to-noise margin. Contiguous active regions shorter than a minimum duration $D_{\min} = 3\,\text{s}$ are discarded to suppress short spurious bursts. The remaining intervals $\{[t_i^{\text{start}}, t_i^{\text{end}}]\}$ correspond to sustained chewing or bite periods, with the start of each biting event marked in Fig.4(d).

The GAP detector thus identifies biting events from raw waveform in a robust, session-adaptive manner. Its median and uniform filters remove impulsive artefacts, while percentile-based thresholding provides consistent sensitivity across recording conditions. For each detected bite, we extract a 5-second post-bite segment and split it into overlapping 3-second windows (0.5-second stride) for feature extraction and classification. Parameters are tuned via 5-fold cross-validation using manually labelled bite timestamps.

Importantly, this event-driven design also limits processing to short chewing-related intervals rather than continuous audio streams for better privacy. Non-eating segments are filtered out by the bite detector and need not be stored. As a result, the system reduces exposure of background conversations or environmental sounds, retaining only structured bite-level representations.

3.3 Feature engineering

To transform raw chewing audio into representations that highlight the signal characteristics most relevant for distinguishing nutrition-texture classes, we perform feature extraction before classification. We extract texture- and motion-related features through engineered methods and improve cross-individual generalisability using CLAP-based features.

3.3.1 *Texture-related Features.*

To capture texture-dependent acoustic characteristics, we extract a set of time-, frequency-, and cepstral-domain audio features.

Time-domain features such as RMS energy and zero-crossing rate provide coarse indicators of texture, since high RMS and high zero-crossing rates typically arise from hard foods whose fracture events create strong, rapidly oscillating waveforms, while soft foods exhibit lower energy and smoother oscillations.

Frequency-domain descriptors reveal finer-grained texture cues. Spectral band energy ratios and high-frequency energy variance (2–6 kHz) are used reveal the presence of crisp fracture bursts from crunchy foods, whereas low-frequency dominance indicates soft and continuous mastication. Conversely, spectral harmony tends to be higher for soft and cohesive foods whose chewing sounds contain more tonal or smoothly varying components.

Cepstral-domain features also reflect texture, as Mel-Frequency Cepstral Coefficients (MFCCs) are commonly used to capture differences in spectral envelope shape, with high MFCC variability indicating crunchy textures. Finally, higher-order spectral statistics, including the skewness, kurtosis and variance of spectral bins, measure the distributional shape of the frequency content. Together, these features provide a rich characterisation of acoustic signatures linked to food texture.

3.3.2 *Motion-related Features.*

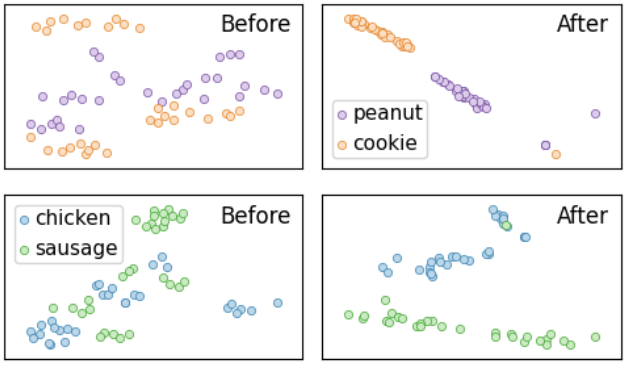


Figure 5: CLAP Feature Embedding Space Before and After Contrastive Fine-Tuning.

We also extract a set of motion-related features from the in-ear audio signal to approximate chewing dynamics that are traditionally measured by motion sensors.

Chewing dynamics. We infer chewing frequency by computing the amplitude-envelope peak rate, which counts the periodic rises in the envelope that correspond to individual chew cycles. To estimate chewing intensity, we compute per-chew RMS energy and the peak-to-RMS ratio since stronger or more forceful chews produce higher overall energy and sharper transient peaks. In addition, chew force reflects how abruptly the jaw closes and how strongly the food fractures. We capture this using spectral centroid shift, since harder bites and fracture events generate increased high-frequency content and therefore shift the centroid upward.

Jaw Movements. We estimate jaw-movement smoothness using the modulation spectrum, where smooth motions concentrate energy in low modulation frequencies and irregular motions spread energy into higher modulation bands. Additionally, we compute the RMS amplitude, variance, and kurtosis of the envelope itself, to capture the fact that chewing on hard or crunchy foods typically yield large-amplitude, high-kurtosis envelopes with sharp transients, whereas chewing on soft or cohesive foods exhibit low-energy, low-kurtosis, and more uniform envelopes.

Together, these engineered physically grounded descriptors provide interpretable cues enhancing robustness across users, bite sizes, and recording sessions. We concatenate these features with the texture-related features described in the previous section for later classification.

3.3.3 *Contrastively Fined-tuned CLAP Audio Encoder.*

To improve generalisability under limited labelled data, we incorporate the CLAP audio encoder to obtain complementary embeddings. Trained on millions of audio–text pairs, CLAP produces semantically rich and transferable representations capturing acoustic events, timbre, texture, and temporal structure. However, its embedding space is optimised for general audio semantics rather than fine-grained chewing distinctions. We therefore apply **supervised contrastive learning**, adding two trainable dense layers while freezing the encoder. This pulls same-class embeddings closer and pushes different classes apart, sharpening decision boundaries in the latent space.

Given a batch of N samples and their corresponding labels $y_i \in \{1, \ldots, K\}$, the supervised contrastive loss [36] is defined as

$$\mathcal{L}_{\text{SupCon}} = -\sum_{i=1}^{N} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_p)/\tau)}{\sum_{a=1}^{N} \mathbf{1}_{[a \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_a)/\tau)},$$

where $\mathbf{z}$ denotes generated embeddings, τ is a temperature hyper-parameter, and $\mathcal{P}(i)$ is the set of indices sharing the same class label as sample i.

As shown in Fig. 5, pre-trained embeddings are loosely scattered with substantial inter-class overlap (e.g., peanut–cookie, chicken breast–sausage), indicating weak separation. After contrastive training, same-class samples form tight clusters and different classes separate with clear margins.

The resulting embeddings capture both the general acoustic structure inherited from CLAP's large-scale pre-training and the fine-grained distinctions among the eight texture/nutrition categories. Under the LOSO setting, one model is trained for each subject. We use the fine-tuned CLAP encoder to obtain 512-dimensional embeddings for each audio segment by generating a 64-bin mel-scaled spectrogram of the audio segment as input to the encoder.

3.4 Classification

With texture-focused features, engineered Motion-related features and CLAP embeddings extracted from each 3-second chewing segment, *NutriEar* performs classification to make texture-nutrition predictions on the audio segments. Fig.2(Right) illustrates the overall model structure. Given the relatively high dimensionality of the non-CLAP features, we first compress the features using two dense layers to encode them into a 512-dimensional embedding. Next, all embeddings generated are fused by concatenation into a 1024-dimensional feature vector. We use five dense layers as the classification backbone. To prevent overfitting, each dense layer is followed by a batch normalisation layer and a dropout layer with a dropout rate of 0.2.

3.5 Postprocessing and Bite-level Classification

To enhance model performance, *NutriEar* aggregates segment-level predictions using a weighted majority voting scheme since each 5-second chewing sequence after a bite consists of multiple overlapping 3-second analysis segments that may exhibit slightly different predictions as the food structure evolves during mastication. We use the weighting to reflect the physical observation that chewing sounds gradually become more muffled as the food bolus softens and mixes with saliva; consequently, the acoustic signature immediately following the bite onset carries the most distinctive texture information.

Formally, let $p_{i,k}$ denote the softmax probability of class k for the i-th window within a 3-second post-bite segment. The final bite-level score for class k is computed as $S_k = \sum_{i=1}^{M} w_i \, p_{i,k}$,where M is the total number of overlapping segments and w_i is a decaying weight that decreases with the temporal offset from the bite onset. In our implementation, we use an exponential decay $w_i = \exp(-\lambda \Delta t_i)$ with $\lambda = 0.4\ \text{s}^{-1}$, assigning the largest weight to the first analysis window immediately after the detected bite with the value of λ chosen through cross-validation. The final class label for the bite event is then given by $\hat{y} = \arg\max_k S_k$.

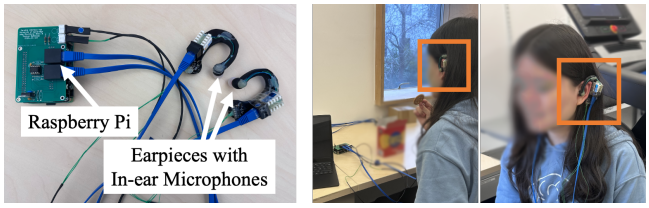


Figure 6: Left: Hardware design. Right: Experiment scenario.

Table 2: Detailed Description of Experimental Conditions.

Name	Detailed Description
FD	20-second interval between bites. Swallow after chewing.
FE	Following subjects' personal eating patterns.
FE-H	FE with head movements of increasing frequency.
FE-A	FE with background noise mimicking public cafes.
FE-M	FE with music played at subjects' self-selected loudness

Table 3: Participant-food coverage table.

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15
Wholemeal Bread	✓		✓	✓										✓	✓
Apple		✓	✓	✓			✓			✓			✓		
Cookies		✓	✓			✓		✓				✓			
Chicken Breast	✓				✓				✓					✓	✓
Boiled Egg	✓			✓			✓		✓	✓					
Beef Jerky	✓			✓		✓					✓	✓			
Avocado					✓			✓		✓			✓		✓
Peanuts		✓				✓	✓				✓	✓	✓		
Chocolate		✓	✓					✓	✓	✓				✓	
Pork Sausage		✓						✓	✓		✓	✓	✓		
Apple Pie					✓		✓								✓
Beef/Bean Burger					✓	✓								✓	✓

This postprocessing step effectively emphasises the early-phase chewing features while down-weighting later windows dominated by low-energy, muffled sounds. The weighted voting strategy improves bite-level classification stability by smoothing short-term fluctuations and aligning the decision with the physical progression of mastication.

4 Experiment Setup and Data Collection

Experiment Setup. Fifteen healthy adult volunteers (aged 18–35 years) participated in the study. Each participant consumed five out of twelve representative food items selected to cover the three primary macronutrient categories, carbohydrates, proteins and fats, as well as a mixed group representing composite foods commonly encountered in real-life diets. The experiments were conducted after approval from the department's ethics committee.

The selected items included wholemeal bread, apple and chocolate cookies (carbohydrate-rich); chicken breast, boiled egg, and beef jerky (protein-rich); avocado, peanuts, chocolate, and pork sausage (fat-rich); beef burger and apple pie (mixed composition) with the details shown in Table 3. The nutritional information of all food items was obtained from their respective product packaging.

The volunteers consumed each food item with four different patterns: Fixed Duration (FD), Free Eating (FE), Free Eating with

Head Movements (FE-H), Free Eating with Ambient Noise (FE-A) and Free Eating with Music (FE-M). The detailed information of these experiments are shown in Table 2. The FD and FE experiments each lasted two minutes, while the other three experiments lasted one minute each.

To assess system extendability, we also collected acoustic signals from other 20 different foods, each consumed by a single user, following the same experimental procedure. This extended dataset, excluded from training, enables evaluation of model performance on unseen foods in a more realistic setting.

Data Collection. As no commercial product currently provides open access to the raw in-ear audio stream, we used a research prototype to record in-ear acoustic signals during eating activities.

As illustrated in Fig. 6(Left), the research prototype consists of a 3D-printed ear-mounted frame with inward-facing Knowles SPU141-0LR5HQB microphones [1] positioned near the ear canal openings to capture body-conducted chewing sounds while attenuating ambient noise. This configuration mirrors microphone placement already used in many commercial earbuds for active noise cancellation and voice enhancement. A custom PCB connects microphones in both ears to a HiFiBerry DAC+ADC Pro audio HAT on a Raspberry Pi, powered by a 5 V power bank for portable operation. In commercial deployments, dual-channel in-ear audio could be transmitted via standard stereo Bluetooth profiles such as A2DP, which support synchronised left–right streaming in modern TWS earbuds. Audio was recorded at 44.1 kHz with synchronised dual-channel acquisition.

Fig.6(Right) depicts the experimental setup: during data collection sessions, participants are instructed to wear the custom ear-mounted devices, inserting them into the ear canal so that the in-ear microphones can capture inner-ear acoustic signals.

5 Results and Evaluation

5.1 Overall Performance

NutriEar achieved an averaged LOSO accuracy of 80.18%, with the normalised LOSO confusion matrix across eight nutrition-texture classes shown in Fig. 7. The matrix exhibits a strong diagonal, indicating that most classes can be reliably recognised even when the user is new to the system.

Some of the confusions shown in this matrix reflect the inherent difficulty of separating very similar textures, under the diverse chewing patterns exhibited by different individuals. These confusions highlight the representational limits of a static classifier when texture and macronutrient cues partially overlap across categories. For instance, confusion between *Crunchy-Fat* and *Crunchy-Carb* (approximately 20–25% in both directions) illustrates that acoustically similar brittle structures can dominate the learned representation, even when macronutrient roles differ.

Likewise, the mutual misclassification between *Firm-Fat* and *Firm-Protein* reflects overlap among dense, protein–fat rich items where overlapping structural properties reduce separability at the bite level. However, if we group the classes further by only macronutrients (Fig. 8(Left)), all classes achieve accuracy of above 80%, which still indicates a relatively reliable cross-individual prediction of the

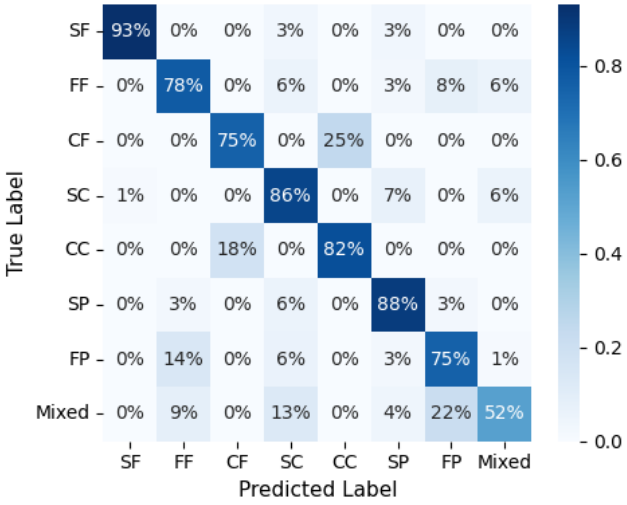


Figure 7: Overall confusion matrix showing classification performance across all nutrition-texture groups. Higher diagonal values indicate stronger bite-level discrimination.

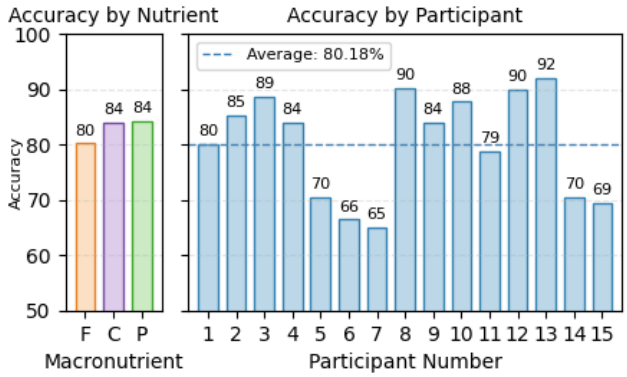


Figure 8: Left: Classification accuracies summed over macronutrient groups. Right: Classification accuracies by participant under controlled indoor conditions.

food nutrition when texture information important to oral functions (as explained in Section 2.1) are no longer considered.

Per-participant performance. Fig. 8 (Right) presents the classification accuracy for each of the 15 participants in LOSO settings. Most participants achieve accuracies above 80%, demonstrating consistent generalisation across individuals. For the top-performing subjects we exceed 89% accuracy, indicating effective bite-segmentation and feature representations. However, a subset of participants (5, 6, 7, 14, 15) exhibit markedly lower accuracy, ranging from 65–71%. The reduced accuracy is largely driven by the inclusion of *Mixed* foods (as shown in Table 3), which produce composite acoustic patterns not well captured by single-structure prototypes. Multiple texture and macronutrient cues co-occur within a single window, disrupting clear amplitude and spectral signatures. As a result,

Table 4: Baseline Comparisons.

System Name	LOSO	80/20 Random Split
NutriEar	80.18%	94.61%
AutoDietary	54.73%	81.10%
iHearken	66.02%	98.23%

mixed-food bites project into intermediate feature regions (Fig. 7), increasing cross-class confusion. These findings highlight composite foods as an open challenge and suggest the need for user-specific adaptation to refine decision boundaries over time.

5.2 Baseline Selection and Model Comparison

It is challenging to identify directly comparable baselines for *NutriEar*, as prior acoustic food recognition studies vary widely in sensor placement, fusion strategy, dataset scale, environmental control, and number of food classes, as discussed in Section 6. Among them, *AutoDietary* [9] and *iHearken* [35] are the closest to our in-ear acoustic sensing setting and represent strong baseline systems.

AutoDietary uses a neck-worn microphone to capture chewing and swallowing sounds and combines MFCCs and time-domain statistics with a hybrid decision tree and SVM classifier. *iHearken* uses an ear-mounted microphone and a Bi-LSTM model to capture temporal dependencies in chewing acoustics, reporting 97% accuracy over 20 food items. We re-implemented both methods following their original descriptions and evaluated them under both LOSO and 80/20 settings for comparison with *NutriEar*.

As shown in Table 4, although *iHearken* achieves the highest 80/20 accuracy (98.23%), its performance drops sharply to 66.02% under LOSO, indicating limited cross-user generalisation and likely overfitting to user-specific chewing patterns. *AutoDietary*, which uses handcrafted MFCC and time-domain features with an SVM–decision-tree ensemble, shows more balanced but overall lower performance, suggesting limited capacity for this more complex classification task. In contrast, *NutriEar* achieves 94.61% accuracy under 80/20 and maintains 80.18% under LOSO, demonstrating stronger robustness across individuals. This gain likely comes from its in-ear sensing design, which reduces external interference, together with contrastive feature learning that promotes user-invariant chewing and texture representations.

5.3 Impact of External Factors

Impact of Eating Patterns. We compare system performance under free eating patterns to fixed metronomic eating intervals, to evaluate the effectiveness of the system in less-controlled settings. Each participant chose their own bite timing, sequence, and portion size and were allowed to sip beverages (water/tea) at will. We recorded two minutes per participant per food, maintaining the same food set across conditions for each individual. During free eating, participants naturally varied their paces, introducing realistic fluctuations in bite spacing and short non-chewing acoustic events. Non-chewing events (e.g., breathing, utensil sounds) occur naturally during social meals. Due to in-ear microphones capturing body-conducted jaw vibrations, airborne speech is attenuated relative to chewing. Additionally, people pause chewing while speaking,

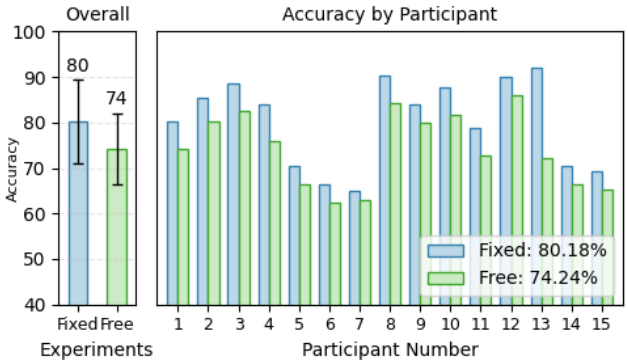


Figure 9: Left: Overall accuracy and variance with fixed duration and free eating. Right: Per-participant accuracy.

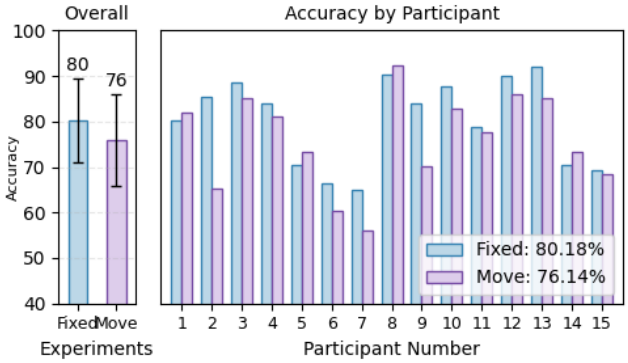


Figure 10: Left: Overall accuracy and variance with fixed duration and head movement. Right: Per-participant accuracy with fixed duration and head movement.

creating distinct patterns that can be separated from mastication events. Fig. 9 summarises the results, showing that while free eating introduces modest degradations, classification stability remain comparable to the fixed-interval baseline. These findings indicate that *NutriEar* tolerates unconstrained bite timing and intermittent drinking, sustaining reliable performance even under everyday eating behaviour.

While most participants saw a 4–8% accuracy drop, Participant 13 experienced a larger decline due to a habit of taking new bites before fully swallowing previous ones, creating acoustic overlap between successive chewing sequences. This behaviour blurs bite boundaries and alters the short-term amplitude and rhythmic cues used for segmentation, causing the bite detector to merge or fragment events unpredictably. This sensitivity is amplified by the limited behavioural diversity captured in our dataset; a larger and more behaviourally diverse dataset would help the system better learn such patterns and mitigate these effects in the future.

Impact of Head Movement. We also evaluated robustness under free head movement to assess the effect of sensor displacement and motion-induced artefacts. The experiment was conducted indoors without added noise, and participants were instructed to eat

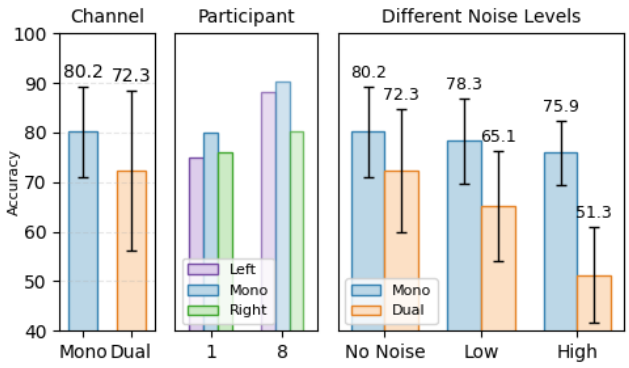


Figure 11: Left: Overall accuracy and variance in mono and dual cases. Middle: Accuracy by participant of the left-only, right-only and mono cases. Right: Overall accuracy and variance with different noise levels.

freely while deliberately moving their heads in a natural but exaggerated manner, including tilting, nodding, and lateral rotations. Each participant completed a two-minute trial for each food, with self-paced bites to preserve realistic eating behaviour.

As shown in Fig. 10 (Left), average accuracy drops only slightly from 80.2% to 76.14%, indicating that the system remains reliable even under exaggerated motion. Larger decreases for Participants 2 and 9 were associated with more substantial earable displacement, likely altering the occlusion condition at the ear canal. Since our study uses academic prototype earables, we expect this issue to be reduced with better-fitting eartips in commercial earbuds.

Impact of Integrated versus Dual-Channel Input. In *NutriEar*, the left- and right-channel acoustic signals are merged into a *monaural signal* (Mono) to suppress spatial cues that may amplify cross-subject variability in eating sounds. To examine whether preserving channel separation offers more information, we re-ran the training pipeline using the same model architecture but concatenated features from both channels (Dual). As shown in Fig. 11(Left), the dual-channel configuration yields a lower mean accuracy of 72.3% and exhibits substantially higher variance. This degradation likely arises because spatial differences between channels introduce irrelevant variability, making the classification task more difficult rather than more informative.

To better understand the increased variance in the dual-channel setting, we examined per-participant performance and found that chewing is often unilaterally dominant. For instance, unlike Participant 1, who shows relatively balanced accuracy between the left-only and right-only channels (Fig. 11(Middle)), Participant 8 performs substantially better when only the left channel is used. Observations during the experiment confirm that this participant predominantly chews on the left side of the mouth, resulting in a stronger and more distinctive acoustic signal on that channel.

Impact of Different Noise Levels. We also evaluated the system performance under different levels of environmental noise to assess robustness in realistic conditions. The experiments were conducted indoor with the same setup and food types as in the baseline condition, while varying the background sound environment. Each participant completed two one-minute sessions corresponding to (i) a *low-noise* condition simulating typical ambient sounds in a shared workspace or café, and (ii) a *high-noise* condition in which participants listened to music through speakers at a volume level of their own daily preference.

Together with the original experiments with no noise, this design captures everyday variability in environmental acoustics, ranging from silent individual meals to social or leisure eating contexts. Fig. 11(Right) shows the results, indicating a gradual decrease in classification accuracy with increasing background noise—80.18% in silence, 78.3% in low noise, and 75.9% in high noise. The relatively small decline confirms that the system remains robust under elevated noise levels, since the inward-facing microphone predominantly captures body-conducted vibrations within the occluded ear canal, attenuating airborne environmental speech and ambient noise. In addition, we examined whether separating the signal channels would help in reducing the noise, but across all scenarios the performance is worse (shown as Dual in Fig. 11(Right)), potentially due to the further differences created by the external noise, which adds spatial bias to the signals collected. While we did not simulate conversational or restaurant noise, the in-ear sensing system inherently suppresses airborne disturbances [47, 48, 81]. These results suggest that system robustness stems not only from algorithmic processing but also from the physical sensing modality itself.

Table 5: Ablation Study.

Component Removed	Result
N/A (Full System)	80.18%
Bite Detection	61.22% (-18.96)
Texture-related Features	76.58% (-3.60)
Motion-related Features	73.14% (-7.04)
CLAP Embedding	75.05% (-5.13)
Contrastive Learning	65.28% (-14.90)

5.4 Ablation Study

We conducted an ablation study to quantify the contribution of each major module in *NutriEar*, as summarised in Table 5. Removing the *bite detection* module and instead using fixed, non-adaptive segments reduced accuracy to 61.22% (−18.96%), confirming the importance of precise temporal segmentation for isolating chewing intervals from non-eating and transitional sounds. Eliminating *texture-related features* lowered accuracy to 76.58% (−3.60%), showing that time–frequency cues related to food structure help distinguish items with similar macronutrient composition. Removing *motion-related features* further reduced accuracy to 73.14% (−7.04%), suggesting their value in capturing chewing intensity and rhythmic variation. Disabling the *CLAP embedding* caused a 5.13% drop to 75.05%, highlighting the benefit of general acoustic priors from large-scale pre-training. Finally, removing the *contrastive learning* objective led to the largest decline, to 65.28% (−14.90%), indicating its key role in improving inter-class separation, intra-class compactness, and cross-user generalisation. Overall, the results show that

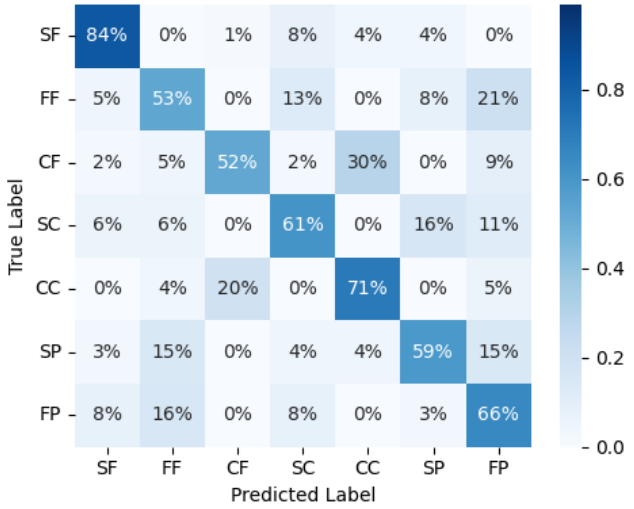


Figure 12: Overall confusion matrix showing classification performance with unseen foods.

all modules contribute synergistically, with bite detection and feature extraction providing the foundation, and CLAP and contrastive learning strengthening discriminative power.

5.5 System Performance on Device

To evaluate the computational feasibility of real-world deployment, we tested the performance of the *NutriEar* model on a Samsung Galaxy S24 smartphone. We evaluate its runtime latency and energy consumption. For a 2-minute input audio, it takes 2.278 s to produce the classification result of all detected bites. Although this duration is slightly longer than ideal for real-time operation, dietary tracking only requires near real-time responsiveness which our model can achieve. According to measurements obtained using a battery monitor, continuous inference with screen-off uses up to 79 mAh/hour (0.02 mAh/second), which is negligible as compared to the device's 4,000 mAh battery.

While inference on mobile devices is computationally feasible, the current system uses a custom research prototype rather than a consumer-grade earable. Transitioning to an integrated consumer-grade system is therefore one potential direction of our future work, particularly for stable dual-channel acquisition and low-latency wireless transmission.

5.6 Generalisation to Unseen Foods

Our system performs well with unseen users as shown in the first part of our evaluation. In this section, we evaluate robustness under a food-level distribution shift, where novel items are excluded during training. We conducted an extension experiment in which food items were held out from training and only evaluated at test time. The label space remained the same (texture–nutrition categories), but the specific foods were new and structurally diverse: bacon crisps, potato crisps and pork rinds (CF); salami, beef patty and bacon (FF); cheese cake and mousse (SF); cooked prawn, roast beef and roast lamb (FP); tofu, poached fish and canned tuna (SP); crackers, cornflakes and pretzels (CC); boiled peas, fries and mashed potato (SC). Figure 7 reports accuracy on the in-distribution set, while Figure 12 shows the confusion matrix for the unseen foods. A total of 20 additional foods were introduced to approximate a more open-world deployment setting, where a large and evolving set of food items may be encountered.

Texture-driven confusions. The most prominent new off-diagonals align with texture similarity: (i) *Crunchy-Fat→Crunchy-Carb* at 30% (and vice versa at 20%), reflecting that high-frequency crackle spectra and transient rates are dominated by brittleness; (ii) *Firm-Fat↔Firm-Protein* (21%/ 16%), consistent with dense, low-moisture bites producing similar attack slopes and envelope dynamics; (iii) *Soft-Carb→ Soft-Fat* at 16%, suggesting that soft, moist foods with low percussiveness share envelope smoothness and low spectral flatness, blurring category boundaries. In all cases, the errors predominantly align with texture similarity, indicating that under significant distribution shift, acoustic structure may dominate over macronutrient cues in the learned representation.

Why performance drops on unseen foods. Chewing acoustics are sensitive to microstructure and preparation (e.g., crust thickness, hydration, oil/fat content, brand-specific processing). Unseen foods often shift the distribution of high-frequency energy ratios and roll-off, amplitude envelopes and attack/decay slopes, and modulation spectra. When these texture cues move, the classifier tends to map novel items to the nearest *texture prototype* seen during training, leading to increased cross-category confusions when novel structural patterns appear.

Implications and mitigation. These results highlight the limitation of static training under real-world dietary variability. To address this, models must adapt over time to new foods and user-specific patterns. Continual learning offers a natural solution as our future direction of study, enabling incremental updates without full retraining. Techniques such as few-shot contrastive refinement or memory-based replay could help incorporate novel texture–nutrition relationships while preserving prior knowledge.

6 Related Work

Eating Behaviour Analysis via Non-Acoustic Sensors. Early chewing detection research focused on motion, strain, and sensor fusion. Wrist-worn inertial systems detect pre-bite wrist-roll motions, using approaches such as template matching [21], regression calibration [70], and adaptive thresholds in commercial devices like Bite Counter [20]. More recently, *IMChew* explored chewing analysis using earphone inertial measurement units [34]. However, these systems primarily capture body motion rather than oral interaction or texture cues, limiting their value for nutrition inference.

Other sensors have also been used to infer chewing. Magnetic-field [43] and optical [82] necklace systems detect chewing-related distance changes and classify gesture patterns, but require specialised neck-worn accessories with limited comfort and social acceptability. Piezoelectric strain gauges near the jaw can capture chewing through mechanical deformation [24, 58], but depend on stable placement and strong skin contact, which can cause discomfort and calibration drift. Sensor-fusion approaches have further enabled detection of internal ingestion events such as swallowing by combining acoustic, respiratory, and accelerometric signals [67, 68].

Although effective in controlled settings, these systems require multiple synchronised sensors and have higher power demands, limiting their practicality for long-term unobtrusive use.

Overall, these non-acoustic systems infer behavioural cues but they depend heavily on sensor placement, individual anatomy, and assumed eating patterns, limiting their reliability in everyday use. More importantly, the motion-focused nature of the captured signals makes it difficult to reveal the structural or textural properties needed for nutrition-aware inference.

Food-Type Recognition via Non-Acoustic Sensors. Food-type recognition has been primarily addressed using camera-based and physical-sensing systems. Camera methods extract visual features such as colour histograms, texture, and convolutional embeddings to classify dishes, achieving between 82–95% accuracy on datasets including UEC-100 and UEC-256 [32, 33, 40, 46, 62, 63, 77]. While highly accurate, vision-based methods require line-of-sight, consistent lighting, and capture detailed visual context which raises significant privacy concerns. In contrast, in-ear acoustic sensing relies on body-conducted chewing vibrations, recording significantly less identifiable scene information.

Utensil-based sensors have also been explored for liquid and semi-liquid classification. Optical, ion-selective, and conductivity sensors can differentiate beverages [42], and LIDS integrates ultrasonic, colour, temperature, and accelerometer data for fluid recognition [60]. While enabling quantitative analysis, these methods require food contact, frequent calibration, and are impractical for long-term use. Earable and eyewear inertial systems are extending towards unobtrusive food recognition. *BiteSense* [17], for instance, uses earable IMUs and transformer-based modelling to classify foods by texture and type, outperforming prior systems like FitByte [8] and MyDJ [71]. However, its performance degrades with head motion when jaw signals overlap with non-eating gestures, and limited cross-user validation raises concerns about generalisability. While less obtrusive than vision-based methods, such approaches remain vulnerable to motion artefacts and user variability.

Eating Behaviour Analysis via Acoustic Sensors. Acoustic sensing offers a direct means to capture chewing and swallowing events via bone-conducted or air-propagated sounds. Sazonov *et al.* [68] first modelled swallowing dynamics using Mel-scale spectral features and hidden Markov models, achieving about 85% accuracy for event-level detection under controlled laboratory settings. Their earlier system relied on frequency-energy envelopes and adaptive thresholding for real-time segmentation of chewing and swallowing [67], but performance degraded in the presence of ambient noise or speech. Makeyev *et al.* [51] extended this framework with wavelet-packet energy features and support vector machine classification, reporting 84.7% weighted accuracy. Nishimura *et al.* later showed that the quality of captured signals can be improved by positioning miniature microphones in the ear canal to isolate bone-conducted jaw vibrations [54], while Lee *et al.* employed ultrasonic Doppler sensing to track jaw and throat motion using neural networks [41]. Although these early acoustic systems reliably detect when eating occurs, they rely on stereotyped chewing and swallowing patterns and do not model the variability of real-world eating behaviour, leaving them unsuitable for the nutrition-aware, unconstrained food inference targeted in our study.

Food-Type Recognition via Acoustic Sensors. Food-type recognition through acoustic sensing exploits variations in chewing rhythm and spectral texture to infer what a user eats. The *AutoDietary* system [9] was the first to employ a neck-mounted microphone for daily-life food monitoring. It extracted MFCCs and time-domain features, then applied decision-tree and support-vector classifiers to distinguish among seven food types with 84.9% accuracy. However, because the transducer was fixed to the throat, it suffered from comfort issues and degraded performance under ambient noise or loose placement, limiting robustness in real-world use. Turan and Erzin [78] later advanced this by applying convolutional neural networks directly to spectrograms of throat-microphone recordings, improving noise tolerance and yielding an F1 score of 0.90 for food-intake event detection. Yet, the model was evaluated on a small, laboratory-controlled dataset without cross-user evaluation, restricting generalisation. To better handle natural acoustic variability, Kondo *et al.* [38, 39] optimised SVM kernels and applied data augmentation to eating-sound recordings collected in restaurants, achieving F1 > 0.85 under environmental noise; nonetheless, their system remained sensitive to microphone orientation and background clatter. More recently, Khan *et al.* [35] proposed *iHearken*, a headphone-like earable using a Bi-LSTM softmax network to classify chewing sounds from 20 food items with 97% accuracy. Together, these approaches demonstrate the feasibility of acoustic food recognition but remain constrained by rigid sensor placement, limited datasets and models generalisability. In contrast, *NutriEar* leverages common earable devices and explicitly models how chewing acoustics relate to food nutritional properties, enabling the first nutrition-aware, in-ear system that sustains reliable performance across users in natural eating conditions.

7 Conclusion

We presented *NutriEar*, an in-ear acoustic system designed as a nutrition-aware sensing primitive that maps real-world chewing audio to structured texture–nutrition categories at the bite level. Motivated by the need for practical and composition-aware food monitoring, *NutriEar* models how macronutrient-driven food structures shape chewing acoustics and uses this relationship to infer texture–nutrition cues from bite-level segments. Through a combination of bite detection, interpretable feature engineering, and contrastively fine-tuned CLAP embeddings, the system achieves robust performance across users and eating conditions. Our results demonstrate that earables can serve as a practical platform for capturing structured, bite-level nutrition signals, which can be aggregated into coarse macronutrient profiles for longitudinal dietary analysis and potentially for quantitative diet monitoring when combined with bite-weight predictions.

Acknowledgments

We thank the shepherd and reviewers for their insightful comments. This work was supported by EPSRC grant EP/Z53447X/1. The first author was financially supported by the China Scholarship Council and the Cambridge Trust.

References

[1] 2024. SPU1410LR5H-QB Knowles | Audio Products | DigiKey. https://www.digikey.co.uk/en/products/detail/knowles/SPU1410LR5H-QB/3621629

[2] Amani Alhazmi, Elizabeth Stojanovski, Mark McEvoy, Wendy Brown, and Manohar L. Garg. 2014. Diet quality score is a predictor of type 2 diabetes risk in women: The Australian Longitudinal Study on Women's Health. *British Journal of Nutrition* 112, 6 (2014), 945–951. doi:10.1017/S0007114514001688

[3] Oliver Amft. 2010. A wearable earpad sensor for chewing monitoring. In *SENSORS, 2010 IEEE*. 222–227. doi:10.1109/ICSENS.2010.5690449

[4] Oliver Amft, Martin Kusserow, and Gerhard Tröster. 2009. Bite Weight Prediction From Acoustic Recognition of Chewing. *IEEE transactions on bio-medical engineering* 56 (04 2009), 1663–72. doi:10.1109/TBME.2009.2015873

[5] Milagros Arnal, Lucía Salcedo, Pau Talens, and Susana Ribes. 2024. Role of Food Texture, Oral Processing Responses, Bolus Properties, and Digestive Conditions on the Nutrient Bioaccessibility of Al Dente and Soft-Cooked Red Lentil Pasta. *Foods* 13 (2024). https://api.semanticscholar.org/CorpusID:271489616

[6] Atif B. Awad and Jyoti P. Chattopadhyay. 1983. Effect of Dietary Fats on the Lipid Composition and Enzyme Activities of Rat Cardiac Sarcolemma. *The Journal of Nutrition* 113, 9 (1983), 1878–1883. doi:10.1093/jn/113.9.1878

[7] Maryam Bahram-Parvar, Toktam Mohammadi Moghaddam, and Seyed Razavi. 2014. Effect of deep-fat frying on sensory and textural attributes of pellet snacks. *Journal of Food Science and Technology* 51 (12 2014). doi:10.1007/s13197-012-0914-6

[8] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. 2020. FitByte: Automatic Diet Monitoring in Unconstrained Situations Using Multimodal Sensing on Eyeglasses. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376869

[9] Y. Bi, M. Lv, C. Song, W. Xu, N. Guan, and W. Yi. 2016. AutoDietary: A Wearable Acoustic Sensor System for Food Intake Recognition in Daily Life. *IEEE Sensors Journal* 16, 3 (2016), 806–816. doi:10.1109/JSEN.2015.2469095

[10] Pascal Bouchon. 2009. Understanding oil absorption during deep-fat frying. *Advances in Food and Nutrition Research* 57 (2009), 209–234. doi:10.1016/S1043-4526(09)57005-2

[11] Malcolm Bourne. 2002. *Food texture and viscosity: concept and measurement*. Elsevier.

[12] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, Yang Liu, and Cecilia Mascolo. 2024. An evaluation of heart rate monitoring with in-ear microphones under motion. *Pervasive Mob. Comput.* 100, C (May 2024), 15 pages. doi:10.1016/j.pmcj.2024.101913

[13] Ya-Jing Cao, Hui-Jun Wang, Bing Zhang, Su-Fen Qi, Ying-Jun Mi, Xing-Bing Pan, Chao Wang, and Qing-Bao Tian. 2020. Associations of fat and carbohydrate intake with becoming overweight and obese: an 11-year longitudinal cohort study. *British Journal of Nutrition* 124, 7 (2020), 715–728. doi:10.1017/S0007114520001579

[14] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

[15] Jin-Young Choi, Seonghee Jeon, Hana Kim, Jaeyoung Ha, Gyeong suk Jeon, Jeong Lee, and Sung il Cho. 2022. Health-Related Indicators Measured Using Earable Devices: Systematic Review. *JMIR mHealth and uHealth* 10, 11 (2022). doi:10.2196/36696

[16] Yun-Sang Choi, Hyun-Wook Kim, Ko Eun Hwang, Dong Heon Song, Tae-Jun Jeong, Young-Boong Kim, Ki-Hong Jeon, and Cheon-Jei Kim. 2015. Effects of fat levels and rice bran fiber on the chemical, textural, and sensory properties of frankfurters. *Food Science and Biotechnology* 24 (04 2015), 489–495. doi:10.1007/s10068-015-0064-5

[17] Garvit Chugh, Indrajeet Ghosh, Sandip Chakraborty, and Suchetana Chakraborty. 2025. BiteSense: Earable-Based Inertial Sensing for Eating Behaviour Assessment. 110–120. doi:10.1109/PerCom64205.2025.00030

[18] Carla de Carvalho and Maria Caramujo. 2018. The Various Roles of Fatty Acids. *Molecules* 23 (10 2018), 2583. doi:10.3390/molecules23102583

[19] Robert Demling. 2009. Nutrition, Anabolism, and the Wound Healing Process: An Overview. *Eplasty* 9 (02 2009), e9.

[20] J. Desendorf, D. R. Bassett, H. A. Raynor, and D. P. Coe. 2014. Validity of the Bite Counter device in a controlled laboratory setting. *Eating Behaviors* 15 (2014), 502–504. doi:10.1016/j.eatbeh.2014.06.013

[21] Y. Dong, A. Hoover, and E. Muth. 2009. A Device for Detecting and Counting Bites of Food Taken by a Person during Eating. In *Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*. 265–268. doi:10.1109/BIBM.2009.29

[22] Raid El-Metwally, Reham El-Menawy, and Magdy Ismail. 2022. Correlation between free fatty acids content and textural properties of Gouda cheese supplemented with denatured whey protein paste. *Journal of Food Science and Technology* 60 (12 2022). doi:10.1007/s13197-022-05643-6

[23] Santiago Espinosa-Salas and Mauricio Gonzalez Arias. 2023. Nutrition, Macronutrient Intake, Imbalances, and Interventions. (08 2023).

[24] M. Farooq and E. Sazonov. 2016. Automatic Measurement of Chew Count and Chewing Rate during Food Intake. *Electronics* 5, 4 (2016), 62. doi:10.3390/electronics5040062

[25] Ciaran Forde and Dieuwerke Bolhuis. 2022. Interrelations Between Food Form, Texture, and Matrix Influence Energy Intake and Metabolic Responses. *Current Nutrition Reports* 11 (06 2022), 1–9. doi:10.1007/s13668-022-00413-4

[26] Patrick F. Fox, Timothy P. Guinee, Timothy M. Cogan, and Paul L. H. McSweeney. 2017. *Cheese: Structure, Rheology and Texture*. Springer US, Boston, MA, 475–532. doi:10.1007/978-1-4899-7681-9_14

[27] A. Fujishita, Y. Koga, D. Utsumi, A. Nakamura, T. Yoshimi, and N. Yoshida. 2015. Effects of feeding a soft diet and subsequent rehabilitation on the development of the masticatory function. *J. Oral Rehabil.* 42, 4 (2015), 266–274. doi:10.1111/joor.12248

[28] Rosalind Gibson, Ute Charrondiere, and Winnie Bell. 2017. Measurement Errors in Dietary Assessment Using Self-Reported 24-Hour Recalls in Low-Income Countries and Strategies for Their Prevention. *Advances in Nutrition* 8 (11 2017), 980–991. doi:10.3945/an.117.016980

[29] Xiangfei Guan, Xuequn Zhong, Yuhao Lu, Xin Du, Rui Jia, Hansheng Li, and Minlian Zhang. 2021. Changes of soybean protein during tofu processing. *Foods* 10, 7 (2021), 1594. doi:10.3390/foods10071594

[30] Changshuo Hu, Thivya Kandappu, Yang Liu, Cecilia Mascolo, and Dong Ma. 2024. BreathPro: Monitoring Breathing Mode during Running with Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 71 (May 2024), 25 pages. doi:10.1145/3659607

[31] Eric Jéquier. 1994. Carbohydrates as a source of energy. *The American journal of clinical nutrition* 59 3 Suppl (1994), 682S–685S. https://api.semanticscholar.org/CorpusID:23382750

[32] Landu Jiang, Bojia Qiu, Xue Liu, Chenxi Huang, and Kunhui Lin. 2020. DeepFood: Food Image Analysis and Dietary Assessment via Deep Model. *IEEE Access* 8 (2020), 47477–47489. doi:10.1109/ACCESS.2020.2973625

[33] Yoshiyuki Kawano and Keiji Yanai. 2013. Real-Time Mobile Food Recognition System. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1–7. doi:10.1109/CVPRW.2013.5

[34] Tamisa Ketmalasiri, Yu Yvonne Wu, Kayla-Jade Butkow, Cecilia Mascolo, and Yang Liu. 2024. IMChew: Chewing Analysis using Earphone Inertial Measurement Units. In *Proceedings of the Workshop on Body-Centric Computing Systems* (Minatoku, Tokyo, Japan) *(BodySys '24)*. Association for Computing Machinery, New York, NY, USA, 29–34. doi:10.1145/3662009.3662022

[35] M. I. Khan, B. Acharya, and R. K. Chaurasiya. 2022. iHearken: Chewing sound signal analysis based food intake recognition system using Bi-LSTM softmax network. *Computer Methods and Programs in Biomedicine* 221 (2022), 106843. doi:10.1016/j.cmpb.2022.106843

[36] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[37] Nicholas Koemel, Alistair Senior, Nasser Laouali, David Celermajer, Amanda Grech, Helen Parker, Stephen Simpson, David Raubenheimer, Tim Gill, and Michael Skilton. 2024. Associations between dietary macronutrient composition and cardiometabolic health: data from NHANES 1999–2014. *European Journal of Nutrition* 64 (12 2024). doi:10.1007/s00394-024-03523-7

[38] Takumi Kondo, Haruka Kamachi, Shun Ishii, Anna Yokokubo, and Guillaume Lopez. 2019. Robust classification of eating sound collected in natural meal environment. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 105–108. doi:10.1145/3341162.3343780

[39] Takumi Kondo, Hidekazu Shiro, Anna Yokokubo, and Guillaume Lopez. 2019. Optimized Classification Model for Efficient Recognition of Meal-Related Activities in Daily Life Meal Environment. In *2019 Joint 8th International Conference on Informatics, Electronics Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision Pattern Recognition (icIVPR)*. 146–151. doi:10.1109/ICIEV.2019.8858526

[40] Junghyo Lee, Ayan Banerjee, and Sandeep K. S. Gupta. 2016. MT-diet demo: Demonstration of automated smartphone based diet assessment system. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 1–3. doi:10.1109/PERCOMW.2016.7457078

[41] K. Lee. 2017. Food Intake Detection Using Ultrasonic Doppler Sonar. *IEEE Sensors Journal* 17, 20 (2017), 6056–6068. doi:10.1109/JSEN.2017.2734688

[42] Jonathan Lester, Desney Tan, Shwetak Patel, and A.J. Bernheim Brush. 2010. Automatic classification of daily fluid intake. doi:10.4108/ICST.PERVASIVEHEALTH2010.8906

[43] C. Li, Y. Bai, W. Jia, and M. Sun. 2013. Eating Event Detection by Magnetic Proximity Sensing. In *Proc. 39th Annual Northeast Bioengineering Conference (NEBEC)*. 15–16. doi:10.1109/NEBEC.2013.85

[44] Jiao Li, Yang Liu, Tao Sun, Ziheng Zhou, and Jin Zhang. 2026. Earable-based Continuous Blood Pressure Monitoring via a Single-Point Flexible Sensor. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Finland) *(UbiComp Companion '25)*. Association for Computing

Machinery, New York, NY, USA, 828–833. doi:10.1145/3714394.3757255
[45] Anne Listrat, Beatriz Lebret, Irène Louveau, Thierry Astruc, Nicolas Bonhomme, Loïc Lefaucheur, Blandine Picard, Ian Richardson, Magali Gautier, and Tarek Sayd. 2016. How Muscle Structure and Composition Influence Meat and Flesh Quality. *BioMed Research International* 2016 (2016), 3182746. doi:10.1155/2016/3182746
[46] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, Ma Yunsheng, Songqing Chen, and Peng Hou. 2018. A New Deep Learning-Based Food Recognition System for Dietary Assessment on An Edge Computing Service Infrastructure. *IEEE Transactions on Services Computing* 11, 2 (2018), 249–261. doi:10.1109/TSC.2017.2662008
[47] Yang Liu, Kayla-Jade Butkow, Jake Stuchbury-Wass, Adam Pullin, Dong Ma, and Cecilia Mascolo. 2025. RespEar: Earable-Based Robust Respiratory Rate Monitoring. In *2025 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 67–77. doi:10.1109/PerCom64205.2025.00026
[48] Yang Liu, Qiang Yang, Kayla-Jade Butkow, Jake Stuchbury-Wass, Dong Ma, and Cecilia Mascolo. 2025. EarMeter: Continuous Respiration Volume Monitoring with Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 198 (Dec. 2025), 29 pages. doi:10.1145/3770655
[49] José M. Lorenzo and Daniel Franco. 2012. Fat effect on physico-chemical, microbial and textural changes through the manufactured of dry-cured foal sausage Lipolysis, proteolysis and sensory properties. *Meat Science* 92, 4 (2012), 704–714. doi:10.1016/j.meatsci.2012.06.026
[50] Yibin Ma, Zekun Zheng, Litao Zhuang, Huiting Wang, Anni Li, Liangkai Chen, and Liegang Liu. 2024. Dietary Macronutrient Intake and Cardiovascular Disease Risk and Mortality: A Systematic Review and Dose-Response Meta-Analysis of Prospective Cohort Studies. *Nutrients* 16, 1 (2024). doi:10.3390/nu16010152
[51] Oleksandr Makeyev, Paulo Lopez-Meyer, Stephanie Schuckers, Walter Besio, and Edward Sazonov. 2012. Automatic food intake detection based on swallowing sounds. *Biomedical Signal Processing and Control* 7, 6 (2012), 649–656. doi:10.1016/j.bspc.2012.03.005 Biomedical Image Restoration and Enhancement.
[52] L. Mioche, P. Bourdiol, and M.-A. Peyron. 2004. Influence of age on mastication: effects on eating behaviour. *Nutrition Research Reviews* 17 (2004), 43–54. doi:10.1079/NRR200375
[53] Pedro C. Moyano and Franco Pedreschi. 2006. Kinetics of oil uptake during frying of potato slices:: Effect of pre-treatments. *LWT - Food Science and Technology* 39, 3 (2006), 285–291. doi:10.1016/j.lwt.2005.01.010
[54] J. Nishimura and T. Kuroda. 2008. Eating habits monitoring using wireless wearable in-ear microphone. In *Proc. 2008 3rd Int. Symp. on Wireless Pervasive Computing (ISWPC)*. 130–132. doi:10.1109/ISWPC.2008.4556181
[55] Michele Novaes Ravelli and Dale Schoeller. 2020. Traditional Self-Reported Dietary Instruments Are Prone to Inaccuracies and New Approaches Are Needed. *Frontiers in Nutrition* 7 (07 2020), 90. doi:10.3389/fnut.2020.00090
[56] Alicia Olivares, José L. Navarro, Ana Salvador, and Mónica Flores. 2010. Sensory acceptability of slow fermented sausages based on fat content and ripening time. *Meat Science* 86, 2 (2010), 251–257. doi:10.1016/j.meatsci.2010.04.005
[57] Bhavesh Panchal, Tuyen Truong, Sangeeta Prakash, Nidhi Bansal, and Bhesh Bhandari. 2021. Influence of fat globule size, emulsifiers, and cream-aging on microstructure and physical properties of butter. *International Dairy Journal* 117 (2021), 105003. doi:10.1016/j.idairyj.2021.105003
[58] S. Päßler and W.-J. Fischer. 2014. Food Intake Monitoring: Automated Chew Event Detection in Chewing Sounds. *IEEE Journal of Biomedical and Health Informatics* 18, 1 (2014), 278–289. doi:10.1109/JBHI.2013.2268663
[59] Sebastian Päßler, Matthias Wolff, and Wolf-Joachim Fischer. 2012. Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food. *Physiological Measurement* 33 (2012), 1073 – 1093. https://api.semanticscholar.org/CorpusID:13255378
[60] Mahdi Pedram, Seyed Iman Mirzadeh, Seyed Ali Rokni, Ramin Fallahzadeh, Diane Myung-Kyung Woodbridge, Sunghoon Ivan Lee, and Hassan Ghasemzadeh. 2021. LIDS: Mobile System to Monitor Type and Volume of Liquid Intake. *IEEE Sensors Journal* 21, 18 (2021), 20750–20763. doi:10.1109/JSEN.2021.3081012
[61] J.M.C. Po, Jules Kieser, Luigi Gallo, A.J. Tésenyi, P Herbison, and Mauro Farella. 2011. Time-Frequency Analysis of Chewing Activity in the Natural Environment. *Journal of dental research* 90 (08 2011), 1206–10. doi:10.1177/0022034511416669
[62] Parisa Pouladzadeh, Shervin Shirmohammadi, and Rana Al-Maghrabi. 2014. Measuring Calorie and Nutrition From Food Image. *IEEE Transactions on Instrumentation and Measurement* 63, 8 (2014), 1947–1956. doi:10.1109/TIM.2014.2303533
[63] J R Rajayogi, G Manjunath, and G Shobha. 2019. Indian Food Image Classification with Transfer Learning. In *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*. 1–4. doi:10.1109/CSITSS47250.2019.9031051
[64] N.R. Rogers, D.J. McMahon, C.R. Daubert, T.K. Berry, and E.A. Foegeding. 2010. Rheological properties and microstructure of Cheddar cheese made with different fat contents. *Journal of Dairy Science* 93, 10 (2010), 4565–4576. doi:10.3168/jds.2010-3494
[65] Laura Roman and Mario M. Martinez. 2019. Structural Basis of Resistant Starch (RS) in Bread: Natural and Commercial Alternatives. *Foods* 8 (2019). https://api.semanticscholar.org/CorpusID:198170063
[66] Rafaela Rosário, Tuyen Duong, and Inês Fronteira. 2023. *Dietary Intake, Eating Behavior and Health Outcomes*. doi:10.3389/978-2-83251-877-9
[67] E. Sazonov, S. Schuckers, P. Lopez-Meyer, O. Makeyev, N. Sazonova, E. L. Melanson, and M. Neuman. 2008. Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior. *Physiological Measurement* 29, 5 (2008), 525–541. doi:10.1088/0967-3334/29/5/001
[68] E. S. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. L. Melanson, and M. R. Neuman. 2010. Automatic Detection of Swallowing Events by Acoustical Means for Applications of Monitoring of Ingestive Behavior. *IEEE Transactions on Biomedical Engineering* 57, 3 (2010), 626–633. doi:10.1109/TBME.2009.2033037
[69] Floor K.G. Schreuders, Miek Schlangen, Konstantina Kyriakopoulou, Remko M. Boom, and Atze Jan van der Goot. 2021. Texture methods for evaluating meat and meat analogue structures: A review. *Food Control* 127 (2021), 108103. doi:10.1016/j.foodcont.2021.108103
[70] Y. Shen, J. Salley, E. Muth, and A. Hoover. 2017. Assessing the Accuracy of a Wrist Motion Tracking Method for Counting Bites Across Demographic and Food Variables. *IEEE Journal of Biomedical and Health Informatics* 21, 3 (2017), 599–606. doi:10.1109/JBHI.2017.2698523
[71] Jaemin Shin, Seungjoo Lee, Taesik Gong, Hyungjun Yoon, Hyunchul Roh, Andrea Bianchi, and Sung-Ju Lee. 2022. MyDJ: Sensing Food Intakes with Attachable on Your Eyeglass Frame. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 341, 17 pages. doi:10.1145/3491102.3502041
[72] Stephen Simpson, David Couteur, and David Raubenheimer. 2015. Putting the Balance Back in Diet. *Cell* 161 (03 2015). doi:10.1016/j.cell.2015.02.033
[73] Charles Spence. 2015. Eating with our ears: assessing the importance of the sounds of consumption on our perception and enjoyment of multisensory flavour experiences. *Flavour* 4, 3 (2015), 1–12. doi:10.1186/2044-7248-4-3
[74] ALINA SURMACKA SZCZESNIAK. 1963. Classification of Textural Characteristics. *Journal of Food Science* 28, 4 (1963), 385–389. arXiv:https://ift.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2621.1963.tb00215.x doi:10.1111/j.1365-2621.1963.tb00215.x
[75] Akio Tada and Hiroko Miura. 2018. Association of mastication and factors affecting masticatory function with obesity in adults: A systematic review. *BMC Oral Health* 18 (05 2018). doi:10.1186/s12903-018-0525-3
[76] Anne-Julie Tessier, Fenglei Wang, Andres Korat, A. Eliassen, Jorge Chavarro, Francine Grodstein, Jun Li, Liming Liang, Walter Willett, Qi Sun, Meir Stampfer, Yao Hu, and Marta Guasch-Ferré. 2025. Optimal dietary patterns for healthy aging. *Nature Medicine* 31 (03 2025), 1644–1652. doi:10.1038/s41591-025-03570-5
[77] Ukrit Tiankaew, Peerapon Chunpongthong, and Vacharapat Mettanant. 2018. A Food Photography App with Image Recognition for Thai Food. In *2018 Seventh ICT International Student Project Conference (ICT-ISPC)*. 1–6. doi:10.1109/ICT-ISPC.2018.8523925
[78] M.A. Tuğtekin Turan and Engin Erzin. 2018. Detection of Food Intake Events From Throat Microphone Recordings Using Convolutional Neural Networks. In *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*. 1–6. doi:10.1109/ICMEW.2018.8551492
[79] Jibran Wali, Samantha Solon-Biet, Therese Freire, and Amanda Brandon. 2021. Macronutrient Determinants of Obesity, Insulin Resistance and Metabolic Health. *Biology* 10 (04 2021), 336. doi:10.3390/biology10040336
[80] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.
[81] Qiang Yang, Yang Liu, Jake Stuchbury-Wass, Mathias Ciliberto, Tobias Röddiger, Kayla-Jade Butkow, Adam Luke Pullin, Emeli Panariti, Dong Ma, and Cecilia Mascolo. 2025. HearForce: Force Estimation for Manual Toothbrushing with Earables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 232 (Dec. 2025), 22 pages. doi:10.1145/3770688
[82] S. Zhang, Y. Zhao, D. T. Nguyen, R. Xu, S. Sen, J. Hester, and N. Alshurafa. 2020. NeckSense: A Multi-Sensor Necklace for Detecting Eating Activities in Free-Living Conditions. *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol. (IMWUT)* 4, 2 (2020). doi:10.1145/3397313