This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

Robust Multimodal Learning Framework For Intake Gesture Detection Using Contactless Radar and Wearable IMU Sensors

Chunzhuo Wang, *Member, IEEE*, Hans Hallez, *Member, IEEE*, and Bart Vanrumste, *Senior Member, IEEE*

***Abstract*— Automated food intake gesture detection is a critical task in the domain of dietary monitoring, with the potential to objectively and continuously record individuals' eating patterns, thereby improving the quality of life (QoL) for diverse populations. Wrist-worn inertial measurement units (IMUs), have been extensively explored for this task and have demonstrated promising performance. Recently, ambient-based contactless radar sensor, has also shown feasibility for intake gesture detection. This study aims to investigate whether the complementary features of wearable and contactless sensors can be effectively leveraged through multimodal learning to further enhance detection performance. Additionally, this study addresses a key challenge in multimodal learning, the reduced robustness when handling missing modalities. To this end, we propose a robust multimodal temporal convolutional network with cross-modal attention (MM-TCN-CMA) framework, designed to efficiently integrate information from IMU and radar sensors, improve detection performance, and maintain effectiveness when modality-incomplete data are encountered during the inference phase. A dataset comprising 52 meal sessions (3,050 eating gestures and 797 drinking gestures) collected from 52 participants is developed and made publicly available for validation. Experimental results demonstrate that the proposed fusion framework achieves a segmental F1-score improvement of 4.3% and 5.2% over unimodal-Radar and unimodal-IMU baselines, respectively. Under missing modality conditions, the framework still yields performance gains of 1.3% and 2.4% in the missing-Radar and missing-IMU scenarios, respectively. To the best of our knowledge, this is the first study to explore a robust multimodal learning framework combining IMU and radar. The proposed robust radar-IMU fusion framework holds potential for broader applications in other continuous, fine-grained human activity recognition (HAR) tasks.**

***Index Terms*— Food intake gesture detection, robust multimodal learning, human activity recognition, FMCW radar, IMU sensor.**

This work is funded in part by the Horizon Europe Research and Innovation Program under Grant GA.No. 101083388, in part by the Flanders AI Research Program, and in part by the Leuven.AI Institute.

Chunzhuo Wang, and Bart Vanrumste are with the e-Media Research Lab, and also with the ESAT-STADIUS Division, KU Leuven, 3000 Leuven, Belgium (e-mail: chunzhuo.wang@kuleuven.be; bart.vanrumste@kuleuven.be).

Hans Hallez is with the M-Group, DistriNet, Department of Computer Science, KU Leuven, 8200 Sint-Michiels, Belgium (e-mail: hans.hallez@kuleuven.be).

I. Introduction

AUTOMATED intake gesture detection is an important task in food intake monitoring domain, which has drawn lots of attention due to its potential application in personalized nutrition and the intervention of dietary-related health problems [1], [2]. Beyond its importance in the context of food intake, intake gesture detection also presents a challenging problem within the broader field of human activity recognition, owing to several unique characteristics. First, intake gestures represent continuous activities embedded within real-world scenarios (e.g., a meal session or full-day monitoring), and are often interwoven seamlessly with other daily activities. Second, intake gestures exhibit considerable variability due to differences in utensils, eating habits, and factors such as age or social context (eating alone versus eating in a group). Third, the duration of intake gestures varies substantially across instances. Based on prior studies [3], [4], the intake gesture (eating and drinking) is defined as a series of movements from raising the hand to the mouth with cutleries/liquid container until the hand is moved away from the mouth.

To enable the automatic and efficient detection of intake gestures, a variety of sensors combined with machine learning techniques have been explored over the past decade. These efforts include wearable-based inertial measurement unit (IMU) [3], [5], [6], ambient-based camera [7], and portable-based smart food plate [8] solutions. Recently, our studies [9], [10] demonstrated that contactless millimeter-wave frequency-modulated continuous wave (FMCW) radar can also be effectively utilized for intake gesture detection in real-world meal scenarios. Along with findings from other studies [11]–[13], radar sensing has emerged as a promising technique in this domain. With the growing prevalence of diverse sensors in real-life settings (e.g., smart homes), it has become increasingly common for multiple sensors to coexist. Consequently, researchers have begun leveraging multimodal sensor fusion to compensate for the limitations of individual modalities and to enhance detection performance. Examples include combinations such as IMU-flex sensors [14], IMU-proximity sensors [15], and IMU-camera systems [16].

However, despite the demonstrated potential of radar as a sensing modality for intake gesture detection, the fusion of IMU sensors and contactless radar sensors has yet to

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

be investigated. We argue that fusing radar and IMU represents a promising and practical combination in real-world applications. IMU sensors are already widely integrated into consumer devices such as smartwatches and fitness trackers, while radar sensors are gaining traction in indoor environments due to their advantages of low privacy concerns, independence from lighting conditions, and contactless sensing capabilities [17]. These modalities offer complementary sensing perspectives: wrist-worn IMUs provide fine-grained motion and orientation data from an egocentric viewpoint, capturing the nuanced dynamics of hand and arm gestures, whereas FMCW radar sensors enable contactless sensing, capturing global spatial and velocity information without requiring direct attachment to the body, offering an exocentric view.

While maximizing the information collected from multiple modalities holds great promise for improving human activity recognition (HAR) performance, a common limitation of multimodal systems is their inefficiency in handling missing modality problems. Specifically, multimodal models trained with complete modalities often experience performance degradation compared to unimodal models when incomplete modality information is encountered during the testing phase [18]. Consequently, a more challenging problem arises: how to design multimodal models that are robust and capable of effectively processing both modality-complete and modality-incomplete data during inference, as illustrated in Fig. 1.

This challenge is particularly significant in multimodal learning involving contactless radar and wearable IMU sensors. These two modalities are typically deployed in different spatial contexts, wearable devices are attached to the body, while radar sensors are installed in the environment. On one hand, they offer strong complementary sensing capabilities; on the other hand, the likelihood of both modalities being simultaneously available is lower compared to systems which are integrated into a single device. To date, robust multimodal learning methods capable of addressing missing modalities in the HAR domain have not been widely explored.

In this paper, we propose a robust multimodal temporal convolutional network combined with cross-modal attention (MM-TCN-CMA) framework for efficiently fusing radar and IMU sensors to achieve fine-grained intake gesture detection during meal sessions. The main contributions of this research can be summarized as follows:

- We propose a sequence-to-sequence (seq2seq) multimodal intake gesture detection framework that fuses radar and IMU sensing for fine-grained recognition during continuous meal sessions. Unlike conventional window-based fusion methods, our approach performs sample-wise prediction and supports temporal segmentation and gesture counting. This design effectively exploits the complementary strengths of radar and IMU data, yielding statistically significant performance improvements in both detection accuracy and temporal boundary localization.
- We introduce a robust mechanism that enables the model to handle missing modalities at inference time. This allows a model trained with radar and IMU data to maintain high performance even when one modality is unavailable for entire session, enhancing system robustness and generalization. To the best of our knowledge, this is the first work to investigate robust multimodal learning for continuous HAR based on IMU–radar fusion.
- We present and publicly release the Radar-IMU multimodal dataset has been collected as part of this study[1]. The dataset comprises 52 continuous meal sessions collected from 52 participants and represents the first public multimodal HAR dataset that combines wrist-worn IMU data with contactless radar sensing.

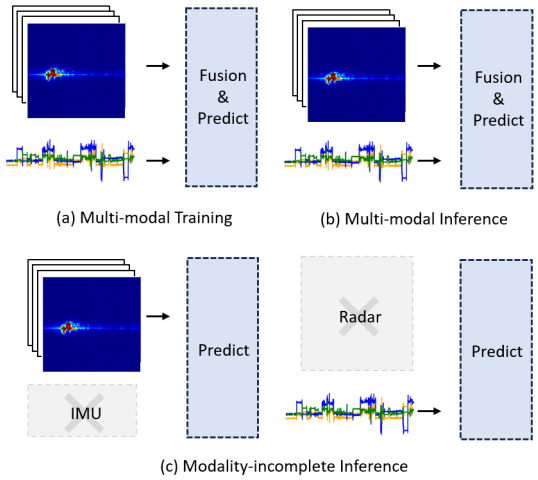


Fig. 1. Training and inference procedures in robust multimodal learning. (a) Training: both radar and IMU modalities are utilized for model training; (b) Inference: both radar and IMU modalities are available during testing; (c) Inference: only the radar modality is available during testing, or only the IMU modality is available during testing.

II. Related Work

In this section, we first review existing studies on food intake gesture detection, covering both unimodal and multimodal approaches. We then introduce research efforts focused on multimodal learning in HAR. Subsequently, we discuss specific studies on radar–IMU fusion for HAR tasks. Finally, we introduce robust multimodal learning approaches designed to address missing modality challenges in other domains.

A. Food Intake Gesture Detection

1) Unimodal Approach: Various sensors have been employed for unimodal intake gesture detection, which can be categorized into three groups based on their deployment position: wearable, portable, and ambient systems. Wearable systems primarily utilize IMU sensors embedded in wristbands [3], [19], small-sized cameras [20], and acoustic sensors [21] configured as necklaces for intake gesture detection. Additionally, photoplethysmography (PPG) sensors [22] and electromyography (EMG) sensors [23] have also been explored for bite detection. Portable solutions include devices such as smart plates [8] and snacking boxes [24] equipped with pressure sensors. In terms of ambient solutions, cameras have been the most common choice [20]. More recently, radar has emerged as a promising alternative for intake gesture detection [9].

[1]https://rdr.kuleuven.be/dataset.xhtml?persistentId=doi:10.48804/LKT5AC

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

2) Multimodal Approach: Several studies have investigated the fusion of multiple modalities for intake gesture monitoring to further enhance detection performance. Bedri et al. [15] proposed FitByte, a wearable system consisting of eyeglasses integrated with a camera, a proximity sensor, and six IMUs, detect intake gestures. Wu et al. [25] developed an ambient-based multimodal system incorporating an RGB camera, a depth camera, and a radar sensor. Notably, these solutions typically integrate multiple sensors into a single device to create either a fully wearable or a fully ambient-based system. Alternatively, Heydarian et al. [16] explored the fusion of wearable and ambient sensors. Specifically, an IMU-based wristband and an exocentric camera, for intake gesture detection. Their fused approach achieved a higher F1-score compared to using either the inertial or video modality alone. Although these multimodal approaches have demonstrated improvements in intake gesture detection performance, it remains an open question whether fusing wearable IMU sensors with ambient radar sensors can further enhance performance.

B. Multimodal Fusion in HAR

1) Basic Fusion Technique: Sensor-based multimodal learning has been extensively studied across various tasks, including daily activity recognition [26], [27] and gait analysis [28]. Three primary approaches are commonly employed for fusing data from different modalities: data-level fusion, feature-level fusion, and decision-level fusion. In data-level fusion, raw data from multiple sensors are concatenated along the feature dimension to form a unified input for subsequent processing [28]. In feature-level fusion, sensor data are first processed, either through deep learning models or traditional feature engineering methods, to extract higher-level representations, which are then combined and used for prediction [26], [27]. In decision-level fusion, data from each sensor are independently processed to generate separate predictions, which are subsequently processed (e.g., through majority voting or averaging) to produce a final decision.

2) Radar-IMU Fusion in HAR: Although no existing studies have specifically explored IMU–Radar fusion for intake gesture detection, several studies have investigated the integration of IMU and radar sensors in HAR. Li et al. [29] proposed a hybrid fusion approach based on Bi-LSTM networks to classify human activities, including fall detection, using IMU and FMCW radar data. Their experiment involved 16 participants and generated 288 recorded events. Recently, Yu et al. [30] combined data from IMU, radar, and a universal software radio peripheral (USRP) to classify two activities: sit-down and stand-up, using manually selected features and a discrete Hopfield neural network (DHNN). While these approaches demonstrated promising performance using handcrafted features, end-to-end learning methods without manual feature engineering have yet to be thoroughly explored. More recently, Shi et al. [31] investigated radar–IMU fusion for classifying walking, running, and stair-climbing activities. Their model employed a convolutional neural network (CNN) module to process IMU data and a vision transformer (ViT) module for radar inputs, with channel-attention mechanisms and concatenation used for multimodal feature fusion. A dataset comprising 1,200 events was collected to validate their approach. However, these prior studies were conducted in highly controlled environments with scripted activity protocols. In contrast, intake gesture detection during meal sessions constitutes a fine-grained, continuous recognition problem, characterized by substantial inter- and intra-subject variability in eating behavior. Moreover, none of the existing works have examined the impact of missing modalities under Radar-IMU sensing conditions. Therefore, this study aims to investigate the potential of Radar-IMU fusion for detecting more realistic and challenging continuous activities.

3) Robust Multimodal Learning with Missing Modality: Given the inherent difficulty of obtaining modality-complete data consistently, the design of robust multimodal learning models capable of handling missing modalities has emerged as a critical area of research. Existing studies demonstrate that standard multimodal models experience a sharp decline in performance when confronted with incomplete modality data, often performing worse than unimodal models [18]. Current research primarily focuses on modality incompleteness across various tasks, including semantic understanding [18], [32], medical imaging segmentation [33], emotion recognition [34], and sleep staging [35]. These solutions can be categorized into two main approaches: 1) missing data reconstruction and 2) joint representation learning. The missing data reconstruction approach aims to recover the missing modality based on the available modalities, often utilizing generative models or autoencoders. However, the overall performance is heavily dependent on the quality of the reconstructed modalities. On the other hand, joint representation learning seeks to extract shared latent features from either fully observed multimodal data or a subset of available modalities. The central premise of this approach is that different modalities share common features. While this method shows promise, it is generally more effective when applied to datasets containing more than two modalities and typically requires larger datasets for effective training [34], [36].

In contrast to the aforementioned tasks, the study of robust multimodal learning for sensor-based HAR remains relatively underexplored. In the Centaur study [37], the authors proposed a convolutional layer-based autoencoder to reconstruct missing data, followed by convolutional and attention layers for multimodal fusion and activity detection. The modalities investigated included IMU data (collected from smartwatches and smartphones) and Electrocardiogram (ECG) signals (acquired from a chest sensor). However, this study primarily addresses scenarios involving partial data loss over short durations relative to the entire session. To date, no HAR studies have specifically tackled situations where data is missing for an entire session, such as a complete meal session. Furthermore, prior research typically assumes multimodal data with similar or identical dimensionalities, leaving a gap in addressing the challenges posed by multimodal data of heterogeneous dimensions. In this study, we propose a feature reconstruction-based approach to overcome the limitations identified in previous HAR research and multimodal food intake gesture detection studies. To the best of our knowledge, no existing work

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

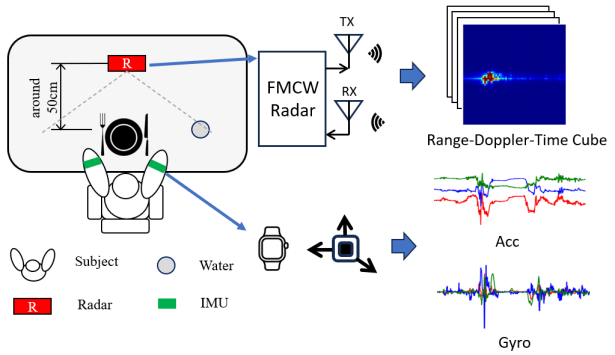


Fig. 2. Data experiment scene.

has developed a robust multimodal learning framework for sensor-based HAR tasks that leverages feature reconstruction, particularly by incorporating both Radar and IMU data.

TABLE I
MEAL DATA STATISTICS

Parameter	Values
Participants	52
Female : male	17:35
Duration ratio of other : eating : drinking	9.44:2.21:1
Mean meal duration	(17.19±5.44) min
Total duration	894 min

TABLE II
EATING AND DRINKING GESTURE DETAILS

Parameter	Values	
	Eating	Drinking
Number of gestures	3,050	797
Duration range of gestures	1.00-17.08 s	2.36-24.56 s
Median duration of gestures	2.72 s	4.68 s
Mean ± std of gestures	3.07 ± 1.42 s	5.32 ± 2.42 s

III. Data Collection and Preprocessing

A. Sensors

An FMCW 76GHz IWR1843 radar from Texas Instrument (TI)[2] and two Shimmer3 IMU sensors[3] were utilized for data collection. The radar system parameters were configured as described in [9]. The radar generates 25 frames of raw data per second (fps), with each frame having a dimension of 128×128. Two IMU sensors were attached to the participants' wrists, each operating at a sampling rate of 64 Hz. Each IMU sensor contains a 3-axis accelerometer and a 3-axis gyroscope, providing six degrees of freedom (DoF). In addition to the radar and IMU sensors, a camera was deployed to record the experimental process for annotation and temporal alignment purposes. The experimental setup is depicted in Fig. 2.

B. Data Collection and Annotation

This study was approved by the KU Leuven ethical committee (Reference Number: G-2021-4025-R4), and written informed consent was obtained from all participants prior to data collection. A total of 52 participants (17 females and 35 males) were recruited, each participating in a single meal session. Data collection was conducted across seven different rooms with varying layouts at the KU Leuven University Campus Group T, including one staff office, three classrooms, two meeting rooms, and one study room. The total duration of the collected dataset amounts to 894 minutes, with an average meal duration of 17.19 ± 5.44 minutes. During the experiment, participants were seated directly in front of the radar while eating, with the distance between the radar and the food plate ranging from 0.3 m to 0.6 m, as illustrated in Fig. 2. Detailed information regarding the meals is provided in Table I. It is noteworthy that a portion of the data was collected concurrently with our previous Eat-Radar project [9]. Participants were instructed to eat in their natural manner, which inherently included non-feeding movements during the meal sessions, such as touching their nose, combing their hair with their hands, adjusting their eyeglasses, using smartphones.

The ELAN tool [15] was utilized to annotate the video recordings and to align the timestamps of the radar and IMU signals. To enable accurate synchronization, participants were instructed to perform a quick up-and-down motion with both hands as a motion indicator before starting their meal. Video recordings served as the reference for aligning IMU and radar data using this motion cue. Once each modality was aligned with the video, the IMU and radar signals were inherently synchronized. The annotated data were categorized into three classes: eating, drinking, and other activities. It is important to note that food intake gestures (both eating and drinking) exhibit substantial variability between participants, and even within a single meal session for the same individual. Therefore, the endpoint of a food intake gesture was defined as the moment the hand moves away from the mouth, rather than strictly when it returns to the plate or table. The annotation was performed independently by two annotators following the same instructions: one being the first author and the other a research assistant not otherwise involved in the study. The task was divided by sequence, and the first author subsequently reviewed and reconciled any discrepancies to minimize ambiguity. In total, 3,050 eating gestures and 797 drinking gestures were recorded. Detailed statistics of the dataset are summarized in Table II.

C. Data preprocessing

1) Radar Data Preprocessing: The range-Doppler-time (RDT) cube was used as the input representation for the radar modality. During each meal session, the radar continuously generated raw data at 25 fps. Several preprocessing steps were applied to the raw data, including Fast Fourier Transform (FFT) for range measurement (Range-FFT), per-frame static object removal, FFT for velocity measurement (Doppler-FFT), superposition, and cropping, following the procedure described in detail in [9]. It is worth noting that

[2]https://www.ti.com/tool/IWR1843BOOST
[3]https://shimmersensing.com/product/shimmer3-imu-unit/

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

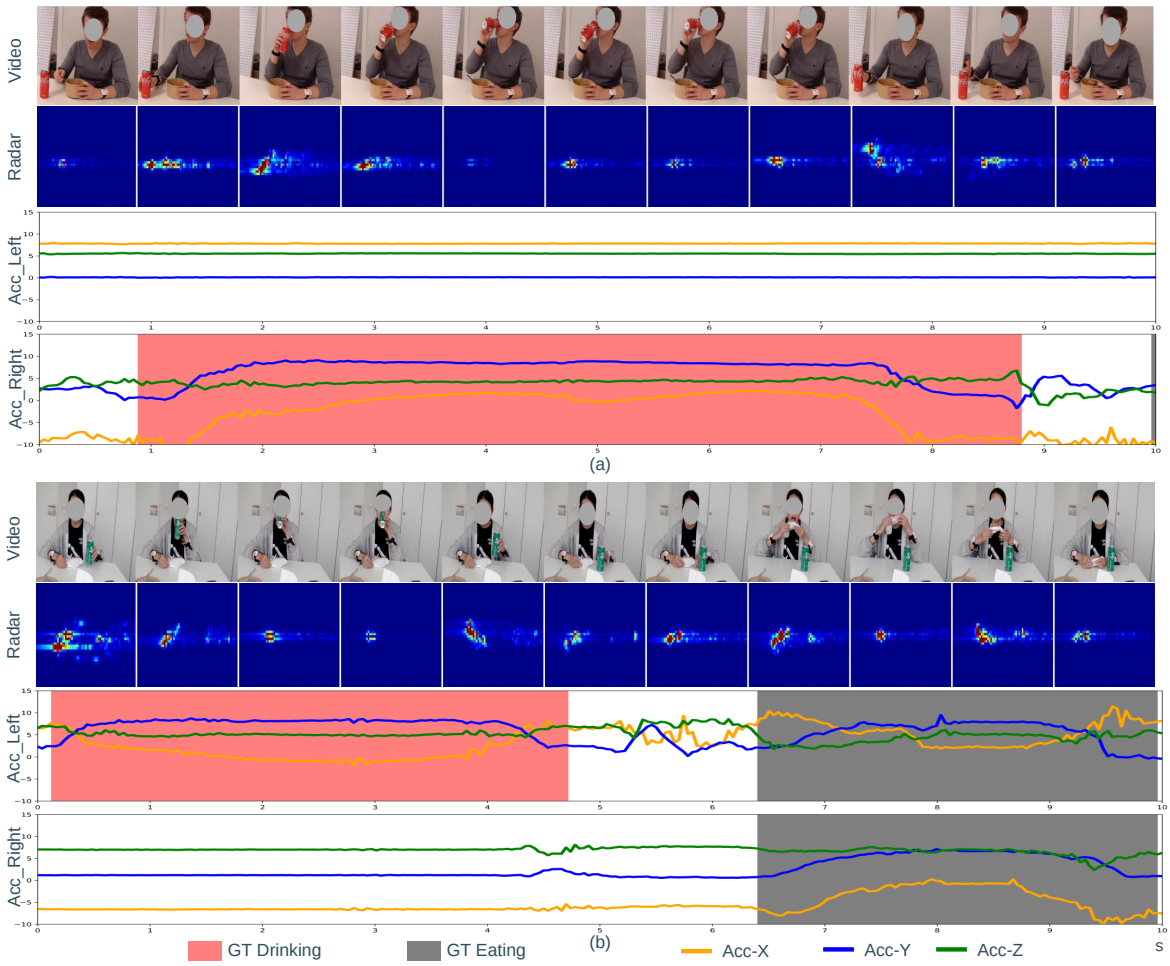


Fig. 3. Examples of radar range-doppler-time data and 3-axis accelerometer data from dual-hand IMU sensors for intake gestures. In (a), the participant uses the right hand to drink while the left hand remains static. In (b), the participant initially uses the left hand to drink, followed by a bimanual action to eat a sandwich.

the static object removal is performed independently on each frame by subtracting the slow-time mean across chirps, which effectively suppresses stationary clutter without relying on any pre-estimated background template or room-specific calibration. After preprocessing, the RDT cube was structured with dimensions $N_r \times N_d \times N = 32 \times 64 \times N$, where N_r denotes the number of range bins (maximum range: 1.28 m), N_d the number of Doppler bins (maximum velocity: $\pm$ 1.28 m/s), and N the number of frames in a meal session.

2) IMU Data Preprocessing: Two wrist-worn IMU sensors were utilized in this study, each providing 6 DoF data. The IMU signals from both hands were concatenated along the feature channel, resulting in a 12-dimensional feature set. The IMU data for one meal session can thus be represented as a matrix of size $N_i \times N = 12 \times N$, where N_i denotes the number of features and N the number of sample points in the session. To achieve sample-wise alignment between the radar and IMU signals, the IMU data were downsampled from 64 Hz to 25 Hz. Our previous studies in [38] have indicated that IMU signals sampled above 16 Hz do not substantially impact model performance. Examples of RD frames and IMU signals aligned with eating movements are illustrated in Fig. 3.

IV. Robust Multimodal Framework

A. Problem Definition

To simplify the problem without loss of generality, we consider a modality-complete dataset comprising two modalities: the RDT Cube (x_R) and IMU data (x_I). Each sample can be denoted as x_R, x_I, y, where y represents the corresponding label. Similarly, modality-incomplete datasets are represented as x_R, y (missing IMU data) or x_I, y (missing radar data). The objective is to leverage the modality-complete dataset to train a multimodal learning model capable of efficiently processing both modality-complete and modality-incomplete data during the inference phase, with an emphasis on achieving higher performance compared to unimodal approaches.

B. Model Architecture

The proposed robust multimodal learning framework comprises four key components: modality-specific feature extraction encoders (MSFE) for extracting representative features from each modality, modality adaptation encoders (MAE) for generating target modality features from source modality data,

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

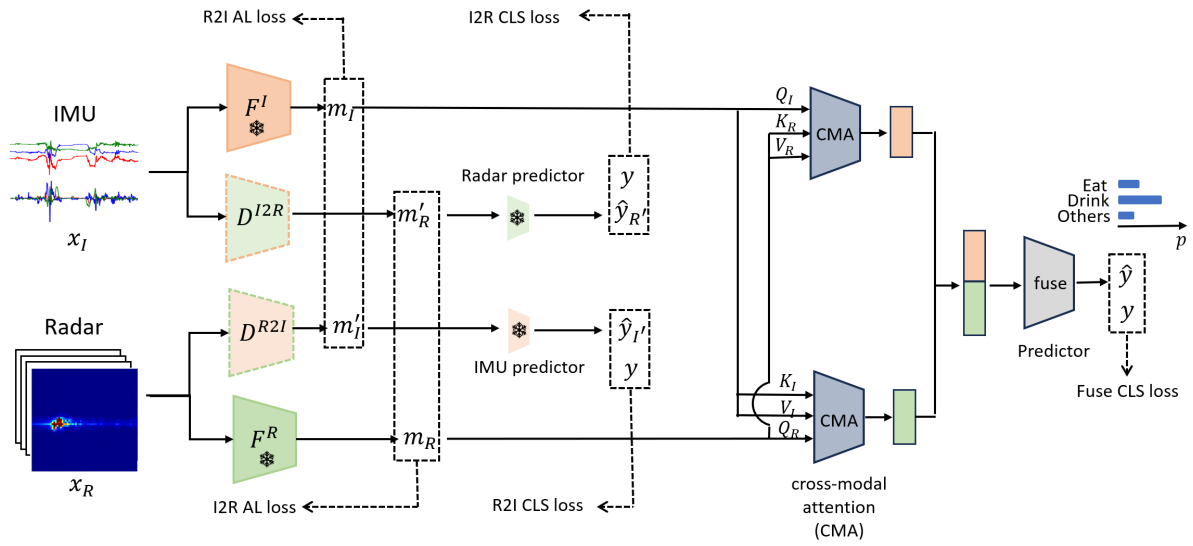


Fig. 4. Overall architecture of the proposed robust multimodal learning framework. Input data from the IMU and Radar modalities, denoted as $\boldsymbol{x_I}$ and $\boldsymbol{x_R}$, are processed by the modality specific feature extraction encoders (MSFE), $\mathbb{F}^{\boldsymbol{I}}$ and $\mathbb{F}^{\boldsymbol{R}}$, to obtain modality-specific intermediate features $\boldsymbol{m_I}$ and $\boldsymbol{m_R}$. Simultaneously, $\boldsymbol{x_I}$ and $\boldsymbol{x_R}$ are fed into the modality adaption encoders (MAE), $\mathbb{D}^{\boldsymbol{I2R}}$ and $\mathbb{D}^{\boldsymbol{R2I}}$, to generate adapted features $\boldsymbol{m'_R}$ and $\boldsymbol{m'_I}$. The cross-modal attention (CMA) modules are then employed to fuse the extracted features. The snow symbol represents that the module with frozen weights.

a feature fusion module (FFM) for integrating multimodal information, and a final classifier for prediction (Predictor).

1) Modality-specific Feature Extraction (MSFE): The feature extraction encoders are designed to extract modality-specific features from each input modality. For the IMU modality, the encoder is denoted as $\mathbb{F}^I : x_I \rightarrow m_I$, where $x_I \in \mathbb{R}^{12\times N}$ represents the input IMU data, and $m_I \in \mathbb{R}^{64\times N}$ denotes the extracted intermediate IMU features. The radar feature extraction encoder is denoted as $\mathbb{F}^R : x_R \rightarrow m_R$, where $x_R \in \mathbb{R}^{32\times 64\times N}$ corresponds to the input RDT radar data, and $m_R \in \mathbb{R}^{64\times N}$ represents the extracted intermediate radar features.

2) Modality Adaption Encoders (MAE): To address the issue of missing modalities, inspired by the IMU2Doppler framework proposed by Bhalla et al. [39], we designed a modality adaptation encoder (MAE) to reconstruct the features of a missing modality using information from the available modality. The module responsible for generating radar features from IMU data (IMU-to-Radar) is denoted as $\mathbb{D}^{I2R} : x_I \rightarrow m'_R$, where $m'_R \in \mathbb{R}^{64\times N}$ represents the reconstructed intermediate radar features. Conversely, for reconstructing IMU features from radar data (Radar-to-IMU), the corresponding module is denoted as $\mathbb{D}^{R2I} : x_R \rightarrow m'_I$, where $m'_I \in \mathbb{R}^{64\times N}$ represents the reconstructed intermediate IMU features.

3) Feature Fusion Module (FFM): The extracted intermediate features m_R and m_I are fused through a feature fusion module (FFM). To implement the FFM, we investigated four fusion techniques: (1) element-wise addition, (2) channel-wise concatenation, (3) decision-level fusion, and (4) cross-modal attention (CMA) [18]:

1) Element-wise addition: This method directly combines the features by adding m_R and m_I element-wise:

$$m_{\text{fused}} = m_R + m_I \tag{1}$$

2) Channel-wise concatenation: The intermediate features m_R and m_I are concatenated along the channel dimension:

$$m_{\text{fused}} = \text{concat}(m_R, m_I) \tag{2}$$

3) Decision-level fusion: Here, the prediction probabilities from the IMU and radar modalities are first computed separately. These probabilities are then summed and averaged to obtain the final fused prediction.

4) Cross-modal attention (CMA): The CMA, derived from self-attention mechanisms, is utilized to integrate complementary information between modalities. Unlike self-attention, where the queries (Q), keys (K), and values (V) are from the same modality, CMA uses queries from one modality and keys and values from the other. Specifically, the cross-attention from IMU to radar is formulated as:

$$\text{crs-att}_{I\rightarrow R} = \text{softmax}\left(\frac{Q_R K_I^T}{\sqrt{d_k}}\right) V_I \tag{3}$$

To fully leverage information from both modalities, a symmetric cross-modal attention mechanism is employed, where features are fused bidirectionally:

$$m_{\text{fused}} = \text{concat}(\text{crs-att}_{I\rightarrow R}, \text{crs-att}_{R\rightarrow I}) \tag{4}$$

Based on experimental results, the CMA method was ultimately selected for feature fusion. A detailed discussion and comparison of the different fusion strategies are provided in Section VI-C.

C. Loss Function

1) Sample-wise Classification Loss Function: A classification loss and a smoothing loss are combined to form the sample-wise loss function. First, a cross-entropy loss is used

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

to represent the classification loss:

$$\mathcal{L}_{ce} = \frac{1}{N}\sum_{t,c} -y_{t,c}\log(\hat{y}_{t,c}) \tag{5}$$

where $y_{t,c}$ is the ground truth label, and $\hat{y}_{t,c}$ is the predicted probability for class c at time t.

Second, to further improve prediction quality, a truncated mean squared error (T-MSE) over sample-wise log-probabilities is applied as a smoothing loss:

$$\mathcal{L}_{T-MSE} = \frac{1}{NC}\sum_{t,c} \widetilde{\triangle}^2{}_{t,c} \tag{6}$$

$$\widetilde{\triangle}_{t,c} = \begin{cases} \triangle_{t,c} : \triangle_{t,c} \leq \tau \\ \tau : otherwise \end{cases} \tag{7}$$

$$\triangle_{t,c} = |\log(\hat{y}_{t,c}) - \log(\hat{y}_{t-1,c})| \tag{8}$$

where N is the data length, C is the number of classes, $\hat{y}_{t,c}$ is the probability of class c at time t, and τ is a threshold to truncate the mean squared error. The smoothing function is used to reduce the over-segmentation errors by calculating the class log probability deviation between adjacent prediction points [40]. The loss function is obtained to combine the mentioned losses :

$$\mathcal{L}_{cls-fuse} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{T-MSE} \tag{9}$$

where λ is a parameter to determine the contribution of the two losses. The τ and λ are set to 4 and 0.15 from [40].

2) Modality Adaption Loss Function: The feature alignment loss function is employed to encourage the MAE to generate intermediate features from the source modality that closely match the features extracted by the MSFE of the target modality. For MAE $\mathbb{D}^{R2I}$, the mean square error (MSE) is adopted as the alignment loss, defined as follows:

$$\mathcal{L}_{al-R2I} = \frac{1}{n}\|m'_I - m_I\|_2^2 \tag{10}$$

where $\mathcal{L}_{al-R2I}$ denotes the alignment loss between the original IMU features m_I from MSFE $\mathbb{F}^I$ and the reconstructed IMU feature m'_I from MAE $\mathbb{D}^{R2I}$, n denotes the total number of elements $64 \times N$. To further enhance the classification performance of the MAE, a classification loss, denoted as $\mathcal{L}_{cls-R2I}$, is also applied to the reconstructed features m'_I using an IMU predictor. The classification loss is consistent with the formulation described in Section IV-C.1. Consequently, the total modality adaptation loss for the R2I direction is defined as the sum of the feature alignment loss and the classification loss:

$$\mathcal{L}_{R2I} = \mathcal{L}_{al-R2I} + \beta\mathcal{L}_{cls-R2I} \tag{11}$$

where β is a parameter to determine the contribution of the two losses and is set to 0.35 according to [39]. Similarly, $\mathcal{L}_{\text{I2R}}$ is defined for the reverse direction, adapting IMU features to radar features.

3) Overall Loss Function: The overall loss function combines the primary classification loss and the auxiliary modality adaptation losses, and is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls-fuse}} + \mathcal{L}_{\text{R2I}} + \mathcal{L}_{\text{I2R}} \tag{12}$$

D. Training and Testing

A two-step training strategy is employed during the training phase. In the first step, unimodal models for the IMU and radar modalities are independently trained. Each unimodal model consists of two components: a MSFE block to extract intermediate representations, and a Predictor to generate class probabilities from these intermediate features. In the second step, the trained MSFEs and Predictors, with their parameters frozen, are integrated into the overall multimodal learning framework. Subsequently, the remaining modules, such as the MAE, CMA, and multimodal predictor, are trained within this framework, as illustrated in Fig. 4. This strategy ensures that $\mathbb{D}^{I2R}$ learns to mimic $\mathbb{F}^I$ in a one-way manner, thereby preventing the undesirable scenario where $\mathbb{F}^I$ adapts to $\mathbb{D}^{I2R}$, which could result in suboptimal performance. It is important to note that the Radar Predictor and IMU Predictor are utilized exclusively during the training phase.

During the testing phase, under a modality-complete scenario, the outputs of $\mathbb{F}^I$ and $\mathbb{F}^R$, denoted as m_I and m_R, respectively, are input into the cross-modal attention module for feature fusion. In modality-incomplete scenarios, such as when the IMU modality is missing, the outputs of $\mathbb{D}^{R2I}$ and $\mathbb{F}^R$, denoted as m'_I and m_R, respectively, are used as inputs to the cross-modal attention module for feature fusion.

E. Implementation

Building upon our previous study [4], a 1D-TCN is employed as the modality-specific feature extractor (MSFE) for the Uni-IMU branch. For the radar modality, a 3D-TCN, as introduced in [9], is adopted as the MSFE for the Uni-Radar branch. TCNs are chosen for their efficiency in capturing long-range temporal dependencies through stacked dilated convolutional layers, which also facilitate multi-scale temporal modeling of motion dynamics, which has been validated in [41]. To simplify the overall framework complexity, the MAE module for each modality shares the same architecture as the corresponding MSFE, e.g., $\mathbb{D}^{I2R}$ and $\mathbb{F}^I$. For FFM in multimodal integration, we employ cross-modal fusion, which allows the model to explicitly align and transfer complementary information between radar and IMU features. This approach was selected because it enhances robustness under missing modalities and improves feature compatibility, enabling effective reconstruction or adaptation when one modality is unavailable. In the CMA module, eight attention heads are utilized, with each head having a dimensionality of 8. Model training is performed using the Adam optimizer with a learning rate of 0.0005. The number of training epochs is set to 100. The input window length is configured to 40 seconds, and the batch size is set to 4. All training, validation, and testing experiments were conducted using two NVIDIA Tesla A100-SXM2-80GB GPUs provided by the Vlaams Supercomputer Centrum (VSC) [4].

V. Evaluation and Validation Scheme

[4]See https://www.vscentrum.be/

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

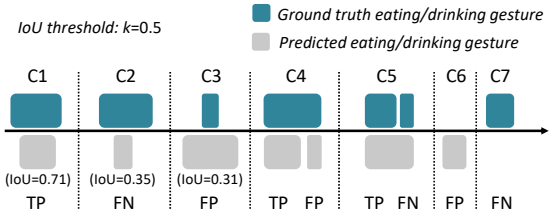


Fig. 5. Segment-wise evaluation examples. If the target is drinking class, eating gestures are categorized as other, and vice versa.

A. Evaluation Scheme and Index

The proposed seq2seq model generates multi-class predictions on a sample-wise basis. Since bite-related datasets tend to be imbalanced, we opt to use Cohen's Kappa index to assess the performance of sample-wise classifications.

While sample-wise results provide insight into the model's direct classification performance, it's important to recognize that the goal of bite detection is to accurately count bites, which sample-wise metrics do not directly reflect. To address this limitation, we apply a segment-wise evaluation method, previously employed in related work [9], [42], to assess bite detection. Examples of this evaluation are shown in Fig. 5. This approach involves two main steps. First, we compute the intersection over union (IoU) between each predicted bite and its corresponding ground truth bite, as illustrated in Fig. 5 C1-C3. Then, the calculated IoU is compared against a predefined threshold, k, to categorize segment-wise true positives (TP), false negatives (FN), and false positives (FP). Based on these values, we calculate the segmental F1-score for each class (eating and drinking):

$$F1_{c=\{E,D\}} = \frac{2TP_c}{2TP_c + FP_c + FN_c} \tag{13}$$

where c represents the class, E, D represents eating and drinking class, respectively.

The segment-wise evaluation accommodates minor temporal misalignments between the ground truth and predictions, which may arise due to variability in annotation. It also provides clear information on the number of bites detected. By varying the threshold k, we are able to assess not only detection accuracy but also segmentation performance. In our evaluation, we used two threshold values, 0.1 and 0.5. Detailed information on this segment-wise evaluation has been introduced in [43].

B. Models for Comparison

To further evaluate the effectiveness of the proposed framework, various models from existing related studies were selected as backbone modules for the MSFE and MAE components. It is important to note that, to the best of our knowledge, no existing work has addressed the exact same scenario investigated in this study. Therefore, directly comparable benchmark models are not available. Instead, several representative models commonly used in related research areas were adopted for comparative analysis.

1) CNN-(Bi)LSTM: The Convolutional Neural Network with Long Short-Term Memory Network (CNN-LSTM) in MSFE-IMU module comprises three 1D convolutional (1D-CNN) layers followed by a single LSTM layer. The CNN-BiLSTM combines the CNN with a bidirectional LSTM layer (BiLSTM). The CNN section includes three layers, each with 64 kernels of size 3, and utilizes the ReLU activation function. The output from the CNN block is flattened and subsequently fed into the LSTM layer, which contains 128 hidden units. The MSFE-Radar module is designed with four 3D-CNN layers, configured with 16, 32, 32, and 32 kernels, respectively, each employing a kernel size of $3 \times 3 \times 3$. ReLU activation is applied after each convolutional layer. Additionally, each 3D-CNN layer is followed by a max-pooling layer with a pooling size of $2 \times 2 \times 1$.

2) ResNet-(Bi)LSTM: The ResNet-LSTM architecture within the MSFE-IMU module integrates a one-dimensional Residual Network (1D-ResNet-10) backbone with a subsequent LSTM layer, as described in [7]. ReLU activation is employed after each convolutional layer. The output from the ResNet backbone is flattened and subsequently input into the LSTM layer, which contains 128 hidden units. The MSFE-Radar module, a 3D-ResNet-18 backbone is utilized, following the design proposed in [44].

3) CNN-MHA: The CNN-MHA model employed in the MSFE-IMU module integrates a three-layer 1D-CNN with a multi-head attention layer (MHA layer), as described in [45]. The 1D-CNN layers utilize 64, 128, and 128 filters, respectively. The MHA layer comprises eight attention heads with an embedding dimension of 128. The MSFE-Radar module adopts a similar architectural design, replacing the 1D-CNN layers with 3D-CNN. This design choice is also applied to the MSFE-Radar modules in the subsequent models.

4) DCNN-LSTM-MHA: The DCNN-LSTM-MHA within the MSFE-IMU module integrates a three-layer 1D-CNN with 32,64,64 filters, three LSTM layers with hidden sizes of 64, respectively, and two 8-head attention modules with dimension of 64 as described in [46].

5) TCCSNet: The temporal-channel convolution with self-attention network (TCCSNet) architecture [47] within the MSFE-IMU module integrates two parallel branches: a time-wise branch and a channel-wise branch. Each branch comprises a two-layer 1D-CNN with 32 and 64 filters, respectively, followed by two layers of 8-head self-attention modules with an embedding dimension of 64.

TABLE III
SAMPLE-WISE KAPPA SCORE RESULTS USING SINGLE MODALITY AND MULTIPLE MODALITIES

Model	Uni-IMU	Uni-Radar	IMU+Radar
CNN-LSTM	0.672	0.574	0.701
CNN-BiLSTM	0.718	0.651	0.750
CNN-MHA	0.631	0.531	0.708
DCNN-LSTM-MHA	0.639	0.561	0.718
ResNet-LSTM	0.666	0.645	0.715
ResNet-BiLSTM	0.742	0.637	0.725
TCCSNet	0.699	0.550	0.734
MM-TCN-CMA	0.747	0.729	0.777

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

TABLE IV
SEGMENT-WISE F1-SCORES RESULTS USING SINGLE MODALITY AND MULTIPLE MODALITIES

Model	k	Uni-IMU		Uni-Radar		IMU+Radar	
		Eating	Drinking	Eating	Drinking	Eating	Drinking
CNN-LSTM	0.1	0.833	0.874	0.797	0.771	0.869	0.883
CNN-BiLSTM		0.852	0.870	0.837	0.853	0.891	0.910
CNN-MHA		0.837	0.873	0.798	0.750	0.867	0.908
DCNN-LSTM-MHA		0.852	0.877	0.770	0.765	0.886	0.882
ResNet-LSTM		0.824	0.834	0.839	0.838	0.875	0.890
ResNet-BiLSTM		0.868	0.906	0.837	0.862	0.886	0.910
TCCSNet		0.864	0.854	0.817	0.756	0.889	0.901
MM-TCN-CMA		0.880	0.893	0.892	0.880	0.922	0.920
CNN-LSTM	0.5	0.764	0.816	0.659	0.655	0.803	0.821
CNN-BiLSTM		0.793	0.821	0.735	0.782	0.833	0.863
CNN-MHA		0.771	0.818	0.669	0.642	0.809	0.858
DCNN-LSTM-MHA		0.785	0.830	0.634	0.564	0.829	0.842
ResNet-LSTM		0.748	0.782	0.717	0.752	0.796	0.836
ResNet-BiLSTM		0.812	0.861	0.707	0.783	0.816	0.854
TCCSNet		0.755	0.792	0.742	0.653	0.833	0.845
MM-TCN-CMA		0.832	0.862	0.841	0.838	0.884	0.887

TABLE V
SEGMENT-WSIE F1-SCORES RESULTS USING MM-TCN-CMA MODEL TRAINED ON TWO MODALITIES, BUT TESTED ON SINGLE MODALITY

Approach	$k = 0.1$		$k = 0.5$	
	Eating	Drinking	Eating	Drinking
Uni-IMU	0.880	0.893	0.832	0.862
Uni-Radar	0.892	0.880	0.841	0.838
Fusion	0.922	0.920	0.884	0.887
Fusion-missing IMU	0.906	0.907	0.865	0.879
Fusion-missing Radar	0.886	0.920	0.845	0.894

C. Validation Strategy

To evaluate the model's performance on unseen data, a 5-fold cross-validation strategy is employed. In each fold, 42 out of 52 meals are assigned to the training set, while the remaining 10 meals are used as the test set (with 11 meals allocated to the test set in the fourth and fifth folds). This design ensures comprehensive evaluation across the entire dataset throughout the 5-fold experiments. The specific meal_id lists corresponding to each fold are provided along with the dataset.

VI. EXPERIMENTAL RESULTS

A. Modality-complete Experiments

We first conducted experiments to validate the efficacy of the fusion of radar and imu data for intake gesture detection on both sample-wise and segment-wise. In this experiment, different models are used as MSFE module, as mentioned in Section V-B. The sample-wise results are shown in Table III. The improved kappa scores have been observed from all models when fusing imu and radar data, with at least 2.8% improvement except ResNet-BiLSTM model. The MM-TCN-CMA based approach obtained highest kappa score of 0.777, followed by the CNN-BilSM model with kappa score 0.750.

In the view of segment-wise evaluation scheme, results in Table IV show that the fusion of imu and radar sensors obtained higher eating and drinking gesture detection performance compared to Uni-IMU or Uni-Radar for all models. When $k = 0.1$, in terms of single modality, the highest F1-score for eating gesture detection is 0.892 from Uni-Radar on MM-TCN-CMA, for drinking gesture detection is 0.906 from Uni-IMU on ResNet-BiLSTM. After fusing IMU and Radar data, the eating and drinking F1-scores improved to 0.922 and 0.920, respectively, using the MM-TCN-CMA approach. When applying a stricter threshold by increasing k to 0.5, a general performance decline was observed across all models. Nevertheless, the fusion of IMU and radar data still outperformed uni-modality models, achieving eating and drinking F1-scores of 0.884 and 0.887, respectively. Furthermore, it was noted that drinking gesture detection performance using the Uni-IMU was generally higher than that of the Uni-Radar modality, a trend that became more pronounced when $k = 0.5$. A detailed individual performance analysis is presented in Fig. 6(a). Based on the Wilcoxon signed-rank test, for MM-TCN-CMA at $k = 0.5$, the fusion of IMU and radar data significantly improved eating gesture detection compared to using Uni-IMU or Uni-Radar modalities ($p < 0.001$). For drinking gesture detection, the fusion approach also demonstrated significantly higher performance than the Uni-Radar modality ($p < 0.001$); however, no significant difference was observed when compared to the Uni-IMU modality. The p values are adjusted using the Benjamini–Hochberg (BH) correction for multiple comparisons. For all comparisons with $p < 0.05$, the corresponding effect sizes exceed 0.5.

B. Modality-incomplete Experiments

After validating the effectiveness of fusing IMU and radar data, we further designed experiments to assess the robustness of the proposed method under modality-incomplete conditions during the inference phase. Specifically, the models were trained on a modality-complete training set; however, during testing, one modality was intentionally dropped. For the missing radar experiment, the radar data were removed from all meal sessions in the test set, leaving only IMU data as input. Similarly, in the missing IMU experiment, only radar data were provided during testing. Table V presents the segment-wise results on MM-TCN-CMA. When $k = 0.1$, eating F1-score under the missing IMU scenario (0.906) surpasses that of Uni-Radar (0.892), although lower than the IMU+Radar fusion

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

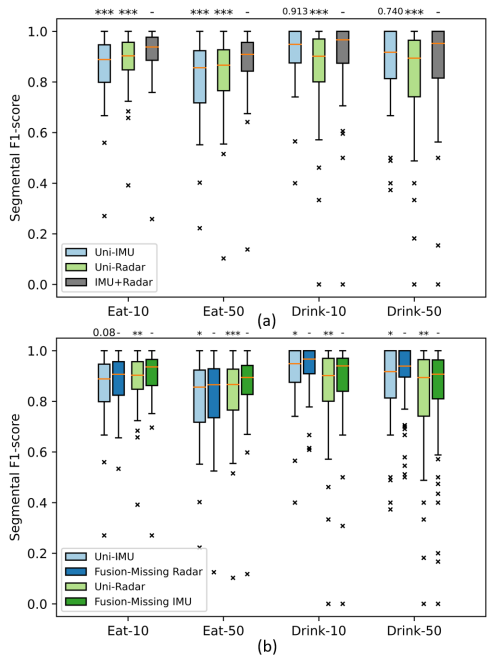


Fig. 6. Boxplots of individual intake gesture detection performance. (a) Performance of unimodal IMU-based, unimodal radar-based, and multimodal fusion approaches using the MM-TCN-CMA model. (b) Performance under missing-modality conditions (missing radar and missing IMU). The y-axis represent the segmental F1-score. In x-axis, Eat-10 and Eat-50 represent eating gesture performance at $\boldsymbol{k = 0.1}$ and $\boldsymbol{k = 0.5}$, and similar for Drink-10 and Drink-50. The $\boldsymbol{p}$ values after Benjamini–Hochberg (BH) multiple comparison correction are shown on top of the box plots. $\boldsymbol{p < 0.05}$: *,$\boldsymbol{p < 0.01}$: **,$\boldsymbol{p < 0.001}$: ***.

(0.922). A similar trend is observed under the missing Radar condition. Notably, the drinking F1-score remains unaffected under the missing Radar scenario, which remained at 0.920. When $k = 0.5$, under the missing Radar scenario, the eating and drinking F1-scores improve by 1.3% and 3.2%, respectively, over Uni-IMU. Specifically, the drinking performance is higher than IMU+Radar fusion (0.894 vs. 0.887). Similarly, under the missing IMU condition, eating and drinking F1-scores increase by 2.4% and 4.1%, respectively, compared to Uni-Radar. Fig. 6(b) shows that the detection performance of intake gestures using fusion with missing modalities is still higher than that achieved by Uni-IMU or Uni-Radar modalities with statistic significance ($p < 0.05$), with the exception of eating gesture detection using Uni-IMU when $k = 0.1$ ($p = 0.08$). For significant comparisons, effect sizes exceed 0.49, except for eating gesture detection using Uni-IMU at $k = 0.5$ ($p < 0.05$, effect size=0.32).

C. Ablation Study

1) Joint Operation: At the feature fusion stage, various methods were investigated, including addition, concatenation, decision-level fusion, and cross-modal attention, as summarized in Table VI. Among these methods, the CMA approach achieved the highest performance.

2) Training Strategy: The proposed robust multimodal learning framework consists of several sub-modules, including MSFE, MAE, and FFM. In the MSFE module, the feature extraction encoders are pretrained and their weights are frozen during training, following a two-step training strategy. Alternatively, the entire framework can be trained from scratch in an end-to-end manner. In this experiment, both training setups were evaluated to assess their impact on performance, as shown in Table VII. The two-step training strategy yielded superior performance, supporting the hypothesis presented in Section IV-D: joint training of all modules may lead to suboptimal results due to the risk of the undesirable scenario in which $\mathbb{F}^I$ adapts to $\mathbb{D}^{R2I}$.

TABLE VI
SEGMENTAL F1-SCORES ($\boldsymbol{k = 0.5}$) ON DIFFERENT FEATURE FUSION METHODS

Fusion Method	IMU+Radar		Missing Radar		Missing IMU	
	Eating	Drinking	Eating	Drinking	Eating	Drinking
Add	0.871	0.877	0.839	0.890	0.859	0.865
Contatenation	0.874	0.888	0.839	0.888	0.864	0.855
Decision-level	0.881	0.886	0.846	0.891	0.854	0.857
CMA	0.884	0.887	0.845	0.894	0.865	0.879

TABLE VII
SEGMENTAL F1-SCORES ($\boldsymbol{k = 0.5}$) ON DIFFERENT TRAINING STRATEGY

Training Method	IMU+Radar		Missing Radar		Missing IMU	
	Eating	Drinking	Eating	Drinking	Eating	Drinking
End-to-end	0.864	0.885	0.828	0.872	0.839	0.856
Two-step	0.884	0.887	0.845	0.894	0.865	0.879

D. Radar + One-hand IMU Experiemnts

The previous experiments assumed IMUs were worn on both wrists. However, in more practical scenarios, participants may only wear a single wrist-mounted IMU. This experiment investigates the performance of fusing radar with IMU exclusively from the dominant hand, defined as the hand with a higher frequency of eating gestures. The overall data processing pipeline remains unchanged, except that the IMU input dimension is reduced from 12 to 6. Importantly, ground truth labels are still based on both hands to ensure the detection of all intake gestures, rather than only those from the dominant hand. Table VIII presents the results. In the Uni-IMU setting, compared to the two-hand IMU configuration (Table V), the eating F1-score decreased by 1.7% at $k = 0.5$, while the drinking F1-score exhibited a notable drop of 26.9%. In the fusion mode, combining radar and one-hand IMU data yielded improvements in both eating and drinking F1-scores relative to the Uni-IMU and Uni-Radar baselines.

VII. DISCUSSION

This study first investigated the efficacy of fusing radar and IMU sensors for hand-based intake gesture detection during meal sessions. As shown in Table IV, the combination of contactless radar and wearable IMU data achieved superior performance compared to using a single modality. The food intake dataset collected in this study contains four different eating styles: fork & knife, spoon, chopsticks, and hand. Another objective was to investigate which eating style contributes most to the performance improvements. As illustrated in Fig.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

TABLE VIII
SEGMENT-WSIE F1-SCORES RESULTS USING MM-TCN-CMA MODEL ON RADAR AND ONE-HAND IMU

Approach	$k = 0.1$		$k = 0.5$	
	Eating	Drinking	Eating	Drinking
Uni-IMU(one-hand)	0.854	0.647	0.815	0.593
Uni-Radar	0.892	0.880	0.841	0.838
Fusion	0.909	0.898	0.871	0.861
Fusion-missing IMU(one-hand)	0.897	0.890	0.851	0.853
Fusion-missing Radar	0.864	0.653	0.824	0.614

TABLE IX
ERROR ANALYSIS ACROSS DIFFERENT EATING STYLES ON MM-TCN-CMA APPROACH AT $k = 0.5$.

Type	Total GT[a]	#FP			#FN		
		Uni-IMU	Uni-Radar	Fusion	Uni-IMU	Uni-Radar	Fusion
Fork & Knife	781	175	151	78	234	82	110
Spoon	634	75	99	34	57	68	56
Chopsticks	1225	107	108	65	109	242	146
Hand	410	78	115	57	140	81	100

[a] Total GT denotes the total number of ground-truth gestures for each style.

TABLE X
DETAILED EATING GESTURE DETECTION PERFORMANCE ON MM-TCN-CMA APPROACH

Approach	$k = 0.1$			$k = 0.5$		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Uni-IMU	0.891	0.869	0.880	0.847	0.817	0.832
Uni-Radar	0.879	0.905	0.892	0.841	0.841	0.841
Fusion	0.942	0.902	0.922	0.913	0.857	0.884
Fusion-missing IMU	0.915	0.898	0.906	0.883	0.847	0.865
Fusion-missing Radar	0.917	0.857	0.886	0.878	0.815	0.845

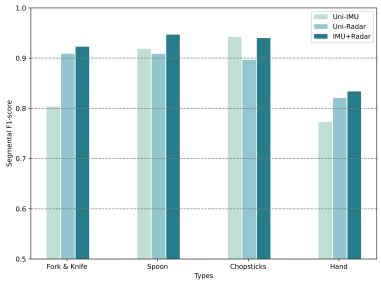


Fig. 7. Bar plots of the eating gesture detection performance across different eating styles, achieved by unimodal IMU-based, unimodal Radar-based, and multimodal fusion approaches using MM-TCN-CMA.

7, performance improvements were observed for the fork & knife, spoon, and hand eating styles. However, for eating with chopsticks, the performance of the fusion method was comparable to that of the Uni-IMU approach. This suggests that the Uni-IMU-based method may have already reached optimal performance in detecting chopstick-based eating gestures, leaving limited room for further enhancement. A detailed error analysis across different eating styles is presented in Table IX. Compared to the number of false negatives (#FN), the reduction in false positives (#FP) was more pronounced across all eating styles when employing multimodal fusion. When comparing performance across eating styles, the relatively low F1-score (Fig. 7), precision (0.845) and recall (0.756) (derived from Table IX) for hand-based eating indicate that this category poses the greatest challenge for the detection model. A detailed examination of the prediction errors and recorded video reveals several contributing factors. First, participants often consumed items such as sandwiches with minimal arm movement, keeping the arm near the mouth and relying mainly on subtle hand or finger motions. These small, localized movements lead to missed detections. Second, the model sometimes misclassified similar but contextually different actions (e.g., wiping the mouth with a napkin or touching the face) as eating with the hand, suggesting a limitation in the model's ability to capture fine-grained contextual cues. Finally, the relative scarcity of hand-eating samples may reduce the model's ability to generalize well for this specific category. Together, these factors highlight the need for improved temporal modeling to enhance robustness in detecting hand-based eating behaviors.

When comparing our work to prior studies utilizing multimodal sensors for intake gesture detection, the size of our dataset is comparable to those used in similar research. In these studies, different sensors are typically integrated into a wearable system, or an ambient system. Only one study [16] investigates the combination of data from wearable IMU and ambient-based camera devices. When comparing our work to prior studies on Radar-IMU combined HAR [29], [30], our dataset size is significantly larger. Moreover, the experimental scenario in our study involves continuous meal sessions.

Our two-step training strategy, which freezes the MSFE and Predictor modules before training the MAE and CMA components, was chosen to prevent destabilizing updates and preserve the integrity of pre-learned representations. However this approach limits the potential for joint, end-to-end optimization, which could further improve feature compatibility and fusion performance. The end-to-end refinement can be a promising direction for future work, particularly in light of recent advances in self-supervised multimodal learning.

This study also addressed the missing modality issue, a challenging and meaningful research problem in HAR, within the context of robust multimodal learning that involves heterogeneous wearable and ambient modalities. By leveraging modality adaption, the model achieves higher performance when tested with incomplete modalities compared to a unimodal approach. This suggests that the MAE can provide additional complementary features from the perspective of the target modality when processing the source modality. Table V shows that the drinking F1-score with missing Radar (0.894) is slightly higher than with fusion (0.887) at k=0.5. After investigating the video, ground-truth labels, and prediction outcomes, we found this phenomenon can partly be attributed to two facts: 1) The drinking gesture is a larger and more dynamic activity compared to eating. Radar signals are more sensitive to occlusions, table surfaces, and surrounding hand or utensil movements, which can occasionally introduce noisy reflections. When fused, these imperfect radar features may slightly degrade the joint representation compared to the $\mathbb{D}^{I2R}$ module generated features by IMU signal alone. 2) The fusion mechanism learns modality importance implicitly. In scenarios where radar contains more noise, the model may not fully down-weight its contribution, leading to suboptimal fusion.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

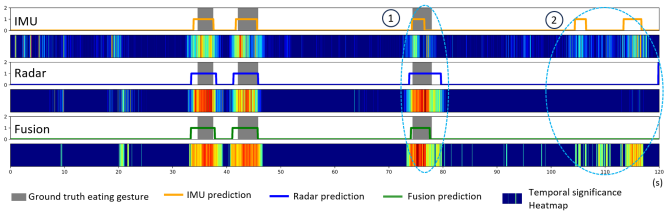


Fig. 8. Examples of gesture prediction results obtained from unimodal IMU-based, unimodal Radar-based, and multimodal fusion approaches, accompanied by the temporal significance heatmaps generated by Grad-CAM [48]. Circled case (1) illustrates enhancement in temporal segmentation precision for detected gestures when fusing IMU+Radar, and circled cases (2) illustrate improvements in detection by removing FPs.

TABLE XI
COMPUTATIONAL COMPLEXITY ANALYSIS OF THE PROPOSED MODEL.

Model	#Params (M)	#FLOPs (G)	Memory (GB)	Lat-full[a] (s)	Lat-I (s)	Lat-R (s)
CNN-LSTM	0.43	0.04	0.32	0.66	0.01	1.34
CNN-BiLSTM	1.17	0.10	0.33	0.66	0.02	1.36
CNN-MHA	0.79	0.66	0.42	0.82	0.08	1.40
DCNN-LSTM-MHA	0.87	0.69	0.83	0.81	0.08	1.69
ResNet-LSTM	101	3873	5.03	10.8	0.03	22.7
ResNet-BiLSTM	103	3874	5.03	10.9	0.04	23.3
TCCSNet	0.49	0.52	0.42	0.70	0.07	1.24
MM-TCN-CMA	1.00	176	2.32	1.08	0.09	2.18

[a] Lat-full denotes the latency for processing 40 s data using the IMU–Radar fusion mode; Similarly, Lat-I and Lat-R represent the latency under missing-radar and missing-IMU conditions, respectively.

When radar is excluded, the model relies on the more stable IMU features, resulting in a small apparent improvement.

The proposed approach achieved significant improvement in intake gesture detection accuracy. Because accurate gesture detection serves as a fundamental component for downstream tasks such as meal session localization and eating speed estimation, these improvements can lead to more precise meal segmentation, more reliable assessment of eating pace, and consequently more accurate monitoring of dietary behaviors in both research and clinical settings. Enhanced detection performance may also increase confidence and adoption among clinicians and researchers when applying such techniques in real-world dietary assessment and intervention scenarios.

This study focused on the scenario where an entire session's modality, such as radar or IMU data, may be missing. In this context, MAE-based feature reconstruction is effective, as it explicitly predicts the missing modality's features, enabling robust inference even under complete modality dropout. However, scenarios with partial modality absence are also worthwhile to explore. In such cases, alternative strategies, including probabilistic fusion, uncertainty-aware models, and late-fusion with modality gating, could be applied to dynamically down-weight missing or unreliable modalities and improve overall robustness. Investigating these approaches in partial-missing settings represents an important direction for future work.

To evaluate the computational complexity of the proposed models, Table XI summarizes the number of trainable parameters, floating-point operations (FLOPs), memory usage, and inference latency for each architecture. Despite achieving competitive performance, it maintains a moderate parameter size (1.0 M) and a manageable computational load (176 GFLOPs) compared to heavier architectures such as ResNet-BiLSTM (101 M parameters, 3873 GFLOPs). Although its memory footprint (2.3 GB) is nontrivial, it remains within the capacity of modern high-end edge processors (e.g., NVIDIA Jetson series). Inference latency was further evaluated using a 40 s input sequence on a laptop equipped with an Intel Core i7-10750H CPU @ 2.6 GHz (6 cores, no GPU). The MM–TCN–MHA model required 1.08 s to generate predictions in the Radar + IMU fusion mode. In the missing-radar configuration, latency decreased markedly to 0.09 s, while in the missing-IMU mode, latency increased to 2.18 s, primarily due to the higher computational cost associated with 3D radar data processing. These results indicate that (near-) real-time inference is feasible, particularly with optimization techniques such as mixed-precision inference, model pruning, and quantization.

Wearable and ambient sensing represent two main paradigms for automated intake gesture detection. IMU sensors are widely used in wearable systems, while radar offers a privacy-preserving ambient sensing alternative. Integrating these complementary modalities enables more robust, long-term dietary behavior monitoring, particularly for older adults in daily living environments. The proposed framework can be extended to include additional sensors, such as smart food plates [8], AIM-2 system [14], or continuous glucose monitors (CGMs), to further enrich dietary insights. Beyond intake gesture detection, the radar–IMU fusion paradigm can also generalize to broader HAR applications, including elderly care, smart home monitoring, and in-vehicle behavior analysis.

An analysis of the detailed performance results presented in Table X reveals that the primary improvement stems from a reduction in the number of FPs, which in turn leads to an increase in precision. This finding suggests that the proposed multimodal framework effectively leverages the complementary information from IMU and radar sensors, helping to suppress ambiguous or noisy inputs that would otherwise result in false positives when relying on a single modality. Examples of this effect are illustrated in Fig. 8(2). Additionally, it is observed that as the IoU threshold k increases from 0.1 to 0.5, the decreases in precision and recall for the multimodal framework (2.9% and 4.5%) are smaller compared to those for the unimodal baselines (IMU: 4.4% and 5.2%, Radar: 3.8% and 6.4%). These results suggest that the integration of radar motion profiles and IMU body motion signals enables the model to produce more accurate and tighter gesture boundaries, resulting in higher overlap ratios, as demonstrated in Fig. 8(1), when using GRAD-CAM heatmap [48] to analysis temporal significance, the fusion model achieves more accurate segmentation compared to the radar modality alone. Although the Grad-CAM heatmap highlights a similar temporal region as the radar's, this indicates that the network still considers that region important. With the aid of the CMA fusion module, which integrates complementary IMU information, the model effectively suppresses the radar's overconfident but incorrect activations. This suggests that the fusion does not merely replace radar or IMU cues but refines their interpretation. Through cross-modal information exchange, the CMA provides additional temporal context from IMU dynamics to

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

disambiguate noisy radar responses. A similar mechanism is observed in Fig. 8(2), where the fusion helps mitigate FPs. Overall, these observations indicate that the proposed multimodal framework not only enhances classification performance but also improves segmentation quality. Furthermore, these benefits are maintained even when one modality is absent during the inference phase, underscoring the framework's robustness to partial sensor availability.

Apart from the annotations performed by the two annotators, we also applied an IoU-threshold-based, segment-wise evaluation method, which is robust to minor temporal shifts in gesture boundaries. While these procedures help mitigate the effects of annotation inconsistencies, we acknowledge that formally quantifying inter-annotator reliability would provide a more rigorous assessment to better characterize annotation reliability and minimize potential bias.

Regarding practical deployment, variations in environmental conditions, such as furniture layout, room geometry, and user movement patterns, pose challenges for real-world implementation, particularly for the radar modality. To assess cross-environment generalization, we conducted a leave-one-room-out (LORO) analysis. The radar modality achieved an average kappa score of 0.717 ± 0.063, and the eating F1-score for $k = 0.1$ remained high at 0.890 ± 0.044 across rooms. These results indicate that the per-frame static object removal supports reasonable robustness when the system is deployed in unseen rooms. Although the LORO performance is also influenced by factors such as the uneven distribution of sessions and variability in user behavior, the overall consistency across rooms indicates stable generalization. The remaining performance variability suggests that environmental characteristics can still affect radar reflections. Future work will investigate more environment-invariant representations and adaptation strategies to further enhance robustness.

Despite the contributions of this research, there are still some limitations. First, the training phase relies on modality-complete data, which inherently restricts the size of the available training set. Although the model can handle missing modalities during inference, future work will explore strategies for training with incomplete data, such as incorporating modality dropout, imputation-based learning, or generative reconstruction methods, to enhance the robustness and practicality of the framework in imperfect sensing conditions. Secondly, the integration of the modality adaptation module introduces additional parameters, which increases the computational overhead during both the training and inference phases. Third, the current study primarily investigates a single-person scenario, which simplifies the problem setting. In more realistic environments where multiple individuals equipped with IMUs may be present within the radar's field of view, it becomes necessary to accurately associate each radar signal with the corresponding participant's IMU data before sensor fusion. Addressing such multi-person scenarios will require advanced strategies, including radar-based target separation, adaptive beamforming for spatial signal isolation, identity tracking, and multi-sensor fusion methods, to achieve reliable and person-specific signal association. Finally, while this study employs MAE-based feature reconstruction for handling missing modalities, other robust multimodal learning strategies could be considered. These include joint representation learning, which learns a shared embedding across modalities; generative imputation methods, which reconstruct missing data using learned generative models; and cross-modal contrastive learning, which enforces alignment between modalities.

VIII. Conclusion

In this study, we propose a robust multimodal learning framework for intake gesture detection, utilizing contactless radar and wearable IMU sensors, with the capability to process data with missing modalities. The results obtained from the collected dataset validate the efficacy of the model. Specifically, the fusion of wearable and ambient sensor signals enhances the performance of intake gesture detection. Additionally, the multimodal model demonstrates robustness when either radar or IMU data is missing during the inference phase. The proposed framework has significant potential for application in other HAR studies.

Acknowledgment

The authors would like to thank the participants who participated in the experiments for their efforts and time. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation Flanders (FWO) and the Flemish Government – department EWI.

References

[1] N. Alshurafa, S. Zhang, C. Romano, H. Zhang, A. F. Pfammatter, and A. W. Lin, "Association of number of bites and eating speed with energy intake: Wearable technology results under free-living conditions," *Appetite*, vol. 167, pp. 1–7, 2021.

[2] P. A. Neves *et al.*, "Thought on food: A systematic review of current approaches and challenges for food intake detection," *Sensors*, vol. 22, no. 17, pp. 1–21, 2022.

[3] K. Kyritsis, C. Diou, and A. Delopoulos, "A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 1, pp. 22–34, 2020.

[4] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Eating speed measurement using wrist-worn imu sensors towards free-living environments," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2024.

[5] Y. Shen, J. Salley, E. Muth, and A. Hoover, "Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 3, pp. 599–606, 2017.

[6] P. V. Rouast, H. Heydarian, M. T. P. Adam, and M. E. Rollo, "Oreba: A dataset for objectively recognizing eating behavior and associated intake," *IEEE Access*, vol. 8, pp. 181 955–181 963, 2020.

[7] P. V. Rouast and M. T. P. Adam, "Single-stage intake gesture detection using ctc loss and extended prefix beam search," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 7, pp. 2733–2743, 2021.

[8] G. Mertes, L. Ding, W. Chen, H. Hallez, J. Jia, and B. Vanrumste, "Measuring and localizing individual bites using a sensor augmented plate during unrestricted eating for the aging population," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 5, pp. 1509–1518, 2020.

[9] C. Wang, T. S. Kumar, W. D. Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Eat-radar: Continuous fine-grained intake gesture detection using fmcw radar and 3d temporal convolutional network with attention," *IEEE J. Biomed. Heal. Informatics*, vol. 28, no. 2, pp. 1000–1011, 2024.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3667308

[10] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Evaluation of different radar placements for food intake monitoring using deep learning," in *2023 IEEE Radar Conference (RadarConf23)*, 2023, pp. 1–6.

[11] Y. Xie, R. Jiang, X. Guo, Y. Wang, J. Cheng, and Y. Chen, "mmeat: Millimeter wave-enabled environment-invariant eating behavior monitoring," *Smart Heal.*, vol. 23, no. 100236, pp. 1–8, 2022.

[12] Y. Wu, Y. Chen, S. Shirmohammadi, and C.-H. Hsu, "Ai-assisted food intake activity recognition using 3d mmwave radars," in *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management (MADiMa '22)*, 2022, pp. 81–89.

[13] H.-C. Chiang, Y.-H. Wu, S. Shirmohammadi, and C.-H. Hsu, "Memory-efficient high-accuracy food intake activity recognition with 3d mmwave radars," in *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, ser. MADiMa '23, 2023, p. 33–41.

[14] A. Doulah, T. Ghosh, D. Hossain, M. H. Imtiaz, and E. Sazonov, "'automatic ingestion monitor version 2' - a novel wearable device for automatic food intake detection and passive capture of food images," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 2, pp. 568–576, 2021.

[15] A. Bedri, D. Li, R. Khurana, K. Bhuwalka, and M. Goel, "Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses," in *CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[16] H. Heydarian, M. T. P. Adam, T. Burrows, and M. E. Rollo, "Exploring score-level and decision-level fusion of inertial and video data for intake gesture detection," *IEEE Access*, vol. 13, pp. 643–655, 2025.

[17] A. Li, E. Bodanese, S. Poslad, P. Chen, J. Wang, Y. Fan, and T. Hou, "A contactless health monitoring system for vital signs monitoring, human activity recognition and tracking," *IEEE Internet of Things Journal*, vol. 11, no. 18, pp. 29 275–29 286, 2024.

[18] S. Woo, S. Lee, Y. Park, M. A. Nugroho, and C. Kim, "Towards good practices for missing modality robust action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2776–2784.

[19] C. Wang, G. Mertes, K. Kyritsis, L. Ding, W. De Raedt, G. Camps, T. Sunil Kumar, W. Chen, J. Jia, H. Hallez, and B. Vanrumste, "Detection of eating gestures in older persons using imu sensors with multistage temporal convolutional network," *IEEE Sensors Journal*, vol. 24, no. 21, pp. 35 231–35 244, 2024.

[20] J. Qiu, F. P.-W. Lo, S. Jiang, Y.-Y. Tsai, Y. Sun, and B. Lo, "Counting bites and recognizing consumed food from videos for passive dietary monitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1471–1482, 2021.

[21] K.-S. Lee, "Food intake detection using ultrasonic doppler sonar," *IEEE Sensors Journal*, vol. 17, no. 18, pp. 6056–6068, 2017.

[22] V. Papapanagiotou, C. Diou, L. Zhou, J. van den Boer, M. Mars, and A. Delopoulos, "A novel approach for chewing detection based on a wearable ppg sensor," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 6485–6488.

[23] R. Zhang, S. Bernhart, and O. Amft, "Diet eyeglasses: Recognising food chewing using emg and smart eyeglasses," in *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2016, pp. 7–12.

[24] F. J. de Gooijer, A. van Kraaij, J. Fabius, S. Hermsen, E. J. Feskens, and G. Camps, "Assessing snacking and drinking behavior in real-life settings: Validation of the snackbox technology," *Food Quality and Preference*, vol. 112, p. 105002, 2023.

[25] Y.-H. Wu, H.-C. Chiang, S. Shirmohammadi, and C.-H. Hsu, "A dataset of food intake activities using sensors with heterogeneous privacy sensitivity levels," in *Proceedings of the 14th ACM Multimedia Systems Conference*, ser. MMSys '23, 2023, p. 416–422.

[26] J. Kang, J. Shin, J. Shin, D. Lee, and A. Choi, "Robust human activity recognition by integrating image and accelerometer sensor data using deep fusion network," *Sensors*, vol. 22, no. 1, 2022.

[27] X. Yang, Z. Xu, H. Liu, P. B. Shull, S. Redmond, G. Liu, and C. Wang, "A multiscale cross-modal interactive fusion network for human activity recognition using wearable sensors and smartphones," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 27 139–27 152, 2024.

[28] P.-K. Yang, B. Filtjens, P. Ginis, M. Goris, A. Nieuwboer, M. Gilat, P. Slaets, and B. Vanrumste, "Multimodal freezing of gait detection: Analyzing the benefits of physiological data," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 33, pp. 956–965, 2025.

[29] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-lstm network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2020.

[30] Z. Yu, A. Zahid, A. Taha, W. Taylor, J. L. Kernec, H. Heidari, M. A. Imran, and Q. H. Abbasi, "An intelligent implementation of multi-sensing data fusion with neuromorphic computing for human activity recognition," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1124–1133, 2023.

[31] J. Shi, Y. Zhu, J. He, Z. Xu, L. Chu, R. Braun, and Q. Shi, "Human activity recognition based on feature fusion of millimeter wave radar and inertial navigation," *IEEE Journal of Microwaves*, vol. 5, no. 2, pp. 409–424, 2025.

[32] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, jun 2022, pp. 18 156–18 165.

[33] Y. Chen, Y. Pan, Y. Xia, and Y. Yuan, "Disentangle first, then distill: A unified framework for missing modality imputation and alzheimer's disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3566–3578, 2023.

[34] H. Zuo, R. Liu, J. Zhao, G. Gao, and H. Li, "Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[35] K. Kontras, C. Chatzichristos, H. Phan, J. Suykens, and M. De Vos, "Core-sleep: A multimodal fusion framework for time series robust to imperfect modalities," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 840–849, 2024.

[36] T. Zhou, "Feature fusion and latent feature learning guided brain tumor segmentation and missing modality recovery network," *Pattern Recognition*, vol. 141, p. 109665, 2023.

[37] S. Xaviar, X. Yang, and O. Ardakanian, "Centaur: Robust multimodal fusion for human activity recognition," *IEEE Sensors Journal*, vol. 24, no. 11, pp. 18 578–18 591, 2024.

[38] C. Wang, J. Kong, Y. Cai, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Intake gesture detection with imu sensor in free-living environments: The effects of measuring two-hand intake and down-sampling," in *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, 2023, pp. 1–4.

[39] S. Bhalla, M. Goel, and R. Khurana, "Imu2doppler: Cross-modal domain adaptation for doppler-based activity recognition using imu data," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 4, dec 2022.

[40] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proc. 32th IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, 2019, pp. 3570–3579.

[41] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1003–1012.

[42] M. Shang, L. Dedeyne, J. Dupont, L. Vercauteren, N. Amini, L. Lapauw, E. Gielen, S. Verschueren, C. Varon, W. De Raedt, and B. Vanrumste, "Otago exercises monitoring for older adults by a single imu and hierarchical machine learning models," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 462–471, 2024.

[43] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Evaluation metrics for food intake activity recognition using segment-wise iou," in *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2024, pp. 1–6.

[44] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[45] T.-H. Tan, Y.-L. Chang, J.-R. Wu, Y.-F. Chen, and M. Alkhaleefah, "Convolutional neural network with multihead attention for human activity recognition," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 3032–3043, 2024.

[46] M. A. Khatun, M. A. Yousuf, S. Ahmed, M. Z. Uddin, S. A. Alyami, S. Al-Ashhab, H. F. Akhdar, A. Khan, A. Azad, and M. A. Moni, "Deep cnn-lstm with self-attention model for human activity recognition using wearable sensor," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–16, 2022.

[47] E. Essa and I. R. Abdelmaksoud, "Temporal-channel convolution with self-attention network for human activity recognition using wearable sensors," *Knowledge-Based Systems*, vol. 278, p. 110867, 2023.

[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on February 25,2026 at 08:51:37 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.