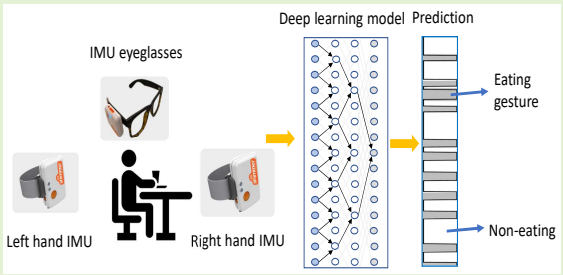
Detection of Eating Gestures in Older Persons Using IMU Sensors with Multi-Stage Temporal Convolutional Network

Chunzhuo Wang, Gert Mertes, Konstantinos Kyritsis, Li Ding, Walter De Raedt, Guido Camps, T. Sunil Kumar, Wei Chen, Jie Jia, Hans Hallez, Bart Vanrumste

***Abstract*— Obesity and malnutrition have psychological and physiological impacts on the health of older adults. Automated eating gesture detection has the potential to advance the assessment of their dietary intake activity. However, detecting eating gestures in older adults poses heightened challenges due to the variability in their eating behaviors. Some individuals exhibit slower eating habits, while others experience limitations in using one hand. Conversely, certain individuals conform to typical eating patterns observed in the adult population. To address this issue, we propose an automated eating gesture detection system using wrist and head mounted inertial measurement unit (IMU) sensors. An end-to-end approach is developed to detect and segment the time interval of eating gestures by employing a multi-stage temporal convolutional network (MS-TCN). Compared to existing eating gesture detection approaches, the present method is able to segment intervals of detected eating gestures more efficiently. We assess our methodology using one self-collected dataset containing 24 older adults, along with three publicly available datasets: FIC, OREBA, and Clemson. The leave-one-subject-out (LOSO) evaluation shows that our method achieves a segmental F1-score of 0.944 on our dataset. Furthermore, results on the FIC, OREBA, and Clemson datasets consistently indicate that our detection approach outperforms existing sliding window-based algorithms that combine convolutional neural networks and recurrent neural networks (CNN-RNNs).**



***Index Terms*— Eating gesture detection, Food intake monitoring, Wearable sensors, Deep learning.**

I. Introduction

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

Gert Mertes is with the Nuffield Department of Population Health, University of Oxford, Oxford, U.K., and with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, U.K. (e-mail: gert.mertes@ndph.ox.ac.uk).

Konstantinos Kyritsis is with the Multimedia Understanding Group, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece (e-mail: kokiritsk@mug.ee.auth.gr).

Li Ding and Jie Jia are with the Department of Rehabilitation Medicine, Huashan Hospital, Fudan University, Shanghai 200040, China (e-mail: leodin@163.com; shannonjj@126.com).

Walter De Raedt and Chunzhuo Wang are with the Life Science Department, IMEC, 3001 Heverlee, Belgium (e-mail: walter.deraedt@gmail.com).

Guido Camps is with the Division of Human Nutrition and Health, Department of Agrotechnology and Food Sciences, Wageningen University and Research, and also with the OnePlanet Research Center, 6708WE Wageningen, The Netherlands (e-mail: guido.camps@wur.nl).

T. Sunil Kumar is with University of Gävle, 801 76 Gävle, Sweden (e-mail: sunilkumar.telagam.setti@hig.se).

Wei Chen is with the School of Biomedical Engineering, The University of Sydney, Darlington, NSW 2006, Australia (e-mail: wei.chenbme@sydney.edu.au).

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

THE World Health Organization (WHO) reports that obesity has become a global epidemic [1]. According to the WHO [2], 1.9 billion adults across the world exhibit overweight conditions, defined by a body mass index (BMI) equal to or exceeding 25, with 34% of this population falling into the category of obesity (BMI $\geq$ 30). Notably, there is a marked increase in the proportion of the overweight population with increasing age. In Europe, 65.6% of adults aged 65 to 74 are overweight [3]. In contrast to obesity, studies also indicate that among adults aged over 65 residing alone in their own homes or within nursing homes, 15% are malnourished and 45% are at risk of malnutrition [4].

Obesity and malnutrition significantly impact the psychological and physiological well-being of older adults, emphasizing the importance of maintaining a well-balanced diet within this demographic [5], which necessitates detecting their dietary intake activities. With recent advancements in sensing technologies and artificial intelligence, researchers are increasingly focusing on developing automated eating gesture detection systems. In general, eating gesture detection systems involve data acquisition using a single sensor or multiple sensors, followed by applying machine learning techniques to analyze the data. These approaches have been shown to be capable of daily monitoring [6].

The eating gesture is defined as a sequence of continuous movements including picking food from the plate, transferring food toward the mouth, and moving the hand away from the mouth. It should be noted that the movement of transferring food toward the mouth can be done not only by moving the hand upward but also by moving the head downward, especially in Asian-style eating gestures. We present an approach for monitoring in-meal eating behavior by detecting fine-grained eating gestures using wrist and head-mounted inertial measurement unit (IMU) sensors. First, we use three IMU sensors to collect data: one on each hand and the last one on the eyeglasses. Then, we apply a sequence-to-sequence (seq2seq) multi-stage temporal convolutional network (MS-TCN) [7], [8] to classify the current motion at the data point level, referred to point-wise classification. Afterward, a sequence of point-wise predictions sharing the same value is interpreted as a segment-wise interval. To evaluate the results, we introduce a modified segment-wise assessment scheme. Finally, the results are evaluated both point-wise and segment-wise, as shown in Fig. 1. The contributions of our work are as follows:

1) A wrist-head combined IMU system is deployed for eating behavior detection. Existing wrist-only IMU systems may fail to capture eating gestures when participants primarily use head movements without significant hand activity. Our system effectively accounts for head movements, enhancing the comprehensiveness of eating gesture detection.
2) We use the seq2seq architecture, the MS-TCN model [7], [8], to detect in-meal eating gestures and segment time intervals of detected eating gestures. In addition to counting the number of eating gestures, the captured intervals of these gestures can serve as a valuable resource for exploring motion characteristics associated with eating, such as the duration of each gesture and the speed of the hand when delivering food from plate to mouth.
3) A segment-wise evaluation scheme is applied to assess the performance based on the overlap ratio between ground truth and predicted eating gestures. The segment-wise evaluation is capable of both detection and segmentation tasks. Notably, this research represents the first evaluation of the segmentation performance of eating gesture detection models.

The rest of this paper is organized as follows. Section II reviews related research in two areas, the relevant studies on food intake monitoring systems and the introduction of the TCN [7] and MS-TCN [8] models and their advantages. Section III describes the detailed pipeline of the proposed approach. In Section IV, we present the evaluation schemes and results of the experiments. Section V discusses the results. The conclusions are presented in Section VI.

II. Related Work

A. Food Intake Monitoring

Automatic food intake detection addresses three core issues: when, what and how much the user has been eating [9].

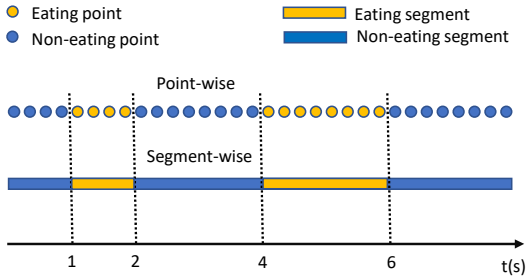


Fig. 1. Explanation of point-wise and segment-wise conception. One eating segment represents an eating gesture that contains multiple eating points. Assume the data sample frequency is 4 Hz, with point-wise conception, the first eating gesture covers 4 points with a time duration of 1 s, and the second eating gesture contains 8 points with a time duration of 2 s. In segment-wise conception, there are 3 non-eating segments and 2 eating segments.

Numerous types of sensors have been applied to solve one or two of the above problems [6], [9]. Strain gauges can measure the weight of consumed food [10]–[12]. Acoustic sensors are used for bite detection and food type recognition by analyzing chewing sounds [13]–[16]. The photoplethysmography (PPG) sensor is used to detect blood variations caused by temporalis muscle movement during chewing [17], [18]. The electromyography (EMG) sensor is used to detect chewing and swallowing movements by sensing the muscle activity around the masticatory muscles [19]–[21]. Cameras are normally used to recognize the food type and to estimate the volume of consumed food [22]–[26]. IMU sensors are used to detect food intake gestures, i.e., hand-to-mouth gestures and chewing activities [27]–[31]. S. Zhang *et al.* [32] developed NeckSense, a necklace that contains a proximity sensor, an ambient light sensor, and an IMU sensor to detect food intake activities. Doulah *et al.* [33] designed the AIM-2 system containing a pair of eyeglasses mounted with a camera, an accelerometer and a flex sensor to monitor food intake activities. The flex sensor was used to detect the contraction of the temporalis muscle during chewing. R. Zhang and Amft [34] designed 3D-print eyeglasses that include bilateral EMG sensors to sense the activity of the temporalis muscles during eating.

Different sensors have been reported for food intake monitoring. The IMU sensor is a popular choice for hand-based gesture recognition [30], [35] due to its easy compatibility with smart devices such as smartwatches and mobile phones, which are highly accepted by users [36]. In the food intake monitoring domain, another popular approach involves using cameras for intake activity recognition and food type classification [23], [24], [26]. Although vision-based approaches can provide richer information, cameras are typically fixed in a specific position. For eating, it is unlikely that participants eat every meal in the same room every day. Additionally, vision-based approaches are more sensitive to lighting conditions. Wearable cameras offer more mobility; however, maintaining continuous video recording for fine-grained eating gesture detection raises concerns regarding battery life and user privacy. In contrast, wearable IMUs have several advantages, including lower power consumption, ease of wear, lack of sensitivity to location, and reduced privacy concerns [37].

Dong *et al.* [38] developed a method for detecting eating gestures by processing gyroscope signals. Using the rotation velocity of wrist motion, they defined the specific eating motion pattern with three features: the roll velocity of the wrist, the time distance between the two wrist rotations of one eating gesture, and the time distance between consecutive eating gestures. S. Zhang *et al.* proposed a solution by detecting two sub-eating gestures [28]. A sliding window was applied to segment the data. A window was defined as an eating gesture only if its overlap ratio with food-to-mouth and back-to-rest segments was higher than the selected threshold. They initially extracted 132 features and selected the best feature subset and an optimal classifier to achieve the best results. Their evaluation metrics were based on the segmented window. A similar approach was presented by Kyritsis *et al.* [30], where the authors modeled an eating gesture as a series of specific micromovements, including actions such as picking food from the plate, hand's upward motion, placing food into the mouth, and hand's downward motion. Their solution consists of two steps. First, a convolutional neural network (CNN) is applied to estimate the probability distribution of micromovements. A long-short-term-memory (LSTM) network is used in the second step to classify eating gestures by analyzing the temporal evolution of micromovements. Another work from Kyritsis *et al.* [39] presented a two-stage approach without classifying micromovements. In the first stage, a convolutional neural network and a long-short-term-memory network combined algorithm (CNN-LSTM) was used to generate prediction series. In the second stage, a local maxima search on the network's predictions was applied to detect eating gestures. Rouast *et al.* [40] further developed a single-stage approach that transfers probabilities into sparse intake detections using the connectionist temporal classification (CTC) loss and an extended prefix beam search algorithm.

The methods mentioned above that use wrist-mounted IMU sensors have shown good performance in the literature. However, they are limited to detecting hand movements exclusively. In real-life scenarios, it is common for individuals to lower their heads while consuming food, with minimal hand movement, as observed in activities such as eating soup or following an Asian-style eating pattern. To address this limitation, we propose a wrist-head mounted IMU system to detect eating behavior.

B. Research on Temporal Convolutional Networks

In existing approaches, a sliding window is applied to divide data into segments before training the model [28], [30], [39], [40]. In the work of S. Zhang *et al.* [28], the window length was 1.5 s. In the micromovement-based work of Kyritsis *et al.* [30], the window length was 3.6 s. In the data-driven approach, the lengths were 5 s and 8 s in [39] and [40], respectively.

The sliding window method has the disadvantage that the window length is fixed while the time duration of an eating gesture varies. If the period of the eating gesture is longer than the window length, the gesture will be divided into two segments or even more. When using the CNN and recurrent neural network (CNN-RNN), the sliding window will limit the model from learning long-range temporal information. Intuitively, it will reduce the model performance when training data with sliding window-based segments since the segment represents only a part of an eating gesture. Another shortcoming lies in the choice of window length. If the window length is too short, one eating gesture will be divided into more windows; if the window length is too long, one window may cover multiple eating gestures. In the case of older people, this problem is even more pronounced because the physical condition of older people varies; hence, the time taken for one eating gesture and the duration gap between two eating gestures vary significantly among older people.

To overcome the shortcomings of the sliding window, Lea *et al.* [7] first proposed a novel single-stage TCN architecture (SS-TCN) for recognizing long-range temporal information in video-based action segmentation by employing the hierarchy of temporal convolutional filters. They presented two different architectures: the encoder-decoder TCN (ED-TCN) and dilated TCN. The dilated TCN applies dilated convolutions and skip connections between convolutional layers, adapted from WaveNet [41]. Recent studies have indicated that TCN outperforms recurrent architectures such as LSTM and the gated recurrent unit (GRU) when processing multiple time-series data [42]. Farha *et al.* [8] expanded upon this by developing a multi-stage TCN (MS-TCN) model, incorporating multiple stages to refine the initial prediction generated by the first stage. TCNs and MS-TCNs have recently been applied to process multivariate time-series data for healthcare and disease diagnosis [43]–[46].

The MS-TCN model has shown its superior ability [42], [43] to capture long-term dependencies in video-based datasets by using dilated convolutions, which allows the network to have a larger effective receptive field without greatly increasing the number of parameters. The superiority of the MS-TCN model in video-based datasets motivated us to leverage its architecture in IMU-based data. By utilizing the characteristics of an MS-TCN, we present an approach that can detect eating gestures with varying durations. Specifically, a non-causal MS-TCN seq2seq model is adopted to process data with a long-range receptive field and classify eating and non-eating gestures. The main difference between the MS-TCN-based approach [8] and the sliding window-based method is that, with the sliding window method, it is common for some eating gestures to be divided across several windows. In the MS-TCN-based approach [8], the longer receptive field of the TCN enables its ability to utilize the temporal information that is further away from the considered sample point, and to determine which class each sample point belongs because of the seq2seq architecture of the TCN.

III. Methodology

A. Wrist-head Combined IMU System

We employ Shimmer3[1] off-the-shelf IMU sensors that include a 3-axis accelerometer unit with 0.1% full-scale (FS) of non-linearity and 125 $\mathrm{ug}/\sqrt{\mathrm{Hz}}$ noise density; a 3-axis

[1]https://shimmersensing.com/product/shimmer3-imu-unit/

TABLE I
HUASHAN DATASET DETAILS

Parameters	Values
Participants (Male : Female)	24 (12 : 12)
Average participant age	63±10
Ratio of total duration/eating gestures	3.03 : 1
Mean meal duration	911±317 s
Number of eating gestures	1,461
Number of chopsticks : spoons : hands	961 : 447 : 53
Duration range of eating gestures	1-28 s
Median duration of eating gestures	3.25 s
Mean ± std of eating gestures	4.23 ± 3.11 s

gyroscope unit with 0.1% FS of non-linearity and 0.015 $\mathrm{dps}/\sqrt{\mathrm{Hz}}$ noise density, resulting in 6 degrees-of-freedom (6 DoF). Two IMU sensors were mounted to wristbands to be worn on the wrist of each hand. The third one was attached to a pair of eyeglasses. The placement of these sensors is illustrated in Fig. 2. The participants need to wear eyeglasses and wristbands during meal sessions. The sampling frequency was 128 Hz. The software Consensys[2] enabled us to export raw data from multiple Shimmer3 IMU sensors to the laptop to facilitate further processing. For each participant, the needed devices were three IMU sensors and a camera for annotation. It's noteworthy that individual data processing did not require a separate computer for each participant. Data from multiple participants were gathered and processed offline using a single laptop.

B. Dataset Description

1) Self-collected Huashan Dataset: The data were collected at the Fudan University Huashan Hospital Jing'An branch in Shanghai, China. The dataset was named 'Huashan' in reference to the collection location. This research was approved by both the Huashan Hospital Institutional Review Board, with reference number KY2013-163, and the ethics committee of KU Leuven with reference number G-2021-3537-R2. Written informed consent from each participant was collected. A total of 24 subjects participated in the experiment. All participants were healthy older adults that could independently eat. Table I shows general information about the dataset. During the data collection process, no restrictions were placed on the participants' eating behavior. They were free to use their preferred utensils, including chopsticks, spoons or their hands, to consume food. They could eat, talk, move their hands or leave their seats at their leisure. A camera was used to record the entire meal session. We used the video to annotate the time-series IMU data. The meals were recorded in the doctor's office, patient's common area, or activity room, based on the participant's preference. Participants were free to choose hospital food or bring food from external restaurants. The types of consumed food included, but were not limited to, rice, vegetables, fish, shrimp, pork, dumplings, eggs, and steamed buns.

To achieve data uniformity, we selected the right hand as the dominant hand. Five of the 24 participants ate with their left hand. We applied the hand-mirroring method introduced in [39], [47] to the data collected from the left-handed participants. In our dataset, participants ate with their dominant hand (either left or right) only, hence there was no eating gestures from non-dominant hand. Therefore, non-dominant hand data were abandoned. It should be noted that the dataset was acquired in a free-living environment. Participants ate their food in their usual way. As a result, the dominant hand IMU data (6 channels) and the head IMU data (6 channels) were used as inputs for the model (12 channels). The data from the two IMUs were synchronized using UNIX timestamps. The data sampling frequency was 128 Hz, resulting in a high level of signal processing redundancy and a heavy computational burden. Therefore, the data were downsampled to 16 Hz.

Fig. 2. Example scenes of food intake monitoring: Participants wore IMU wristbands on both hands and an eyeglass-mounted IMU on the head.

The annotation tool ELAN [48] was used for annotating the data. ELAN can combine video and time-series signals, which is convenient for time-series signal annotation. The annotation work was done by the author. The IMU data were classified into two types, with eating class labeled as 1 and non-eating class labeled as 0. The definition of eating gesture has been introduced in Section I. A non-eating class contains everything else during the meal. It should be noted that this dataset contains only a limited number of drinking gestures (30 drinking gestures in total), and they are labeled as 0. In total, there were 1,461 annotated eating gestures. Given that our dataset was collected under real-life conditions, the time spent on the food intake activity was much shorter than the overall duration consumed on non-eating activities, resulting in an unbalanced dataset. The duration of eating gestures in this dataset varies from 1 s to 28 s, as shown in Table I. Although 70% of these durations are less than 5 s, those slower eating gestures cannot be ignored, especially considering that older adults tend to perform actions more slowly [49].

2) FIC Dataset: The publicly available FIC dataset [30] is used to evaluate our approach. The FIC dataset contains 21 meals from 12 subjects and 1,108 eating gestures. The starting and ending times of each eating gesture were annotated. The sampling frequency was 100 Hz, and we downsampled it to

[2]https://shimmersensing.com/product/consensyspro-software/

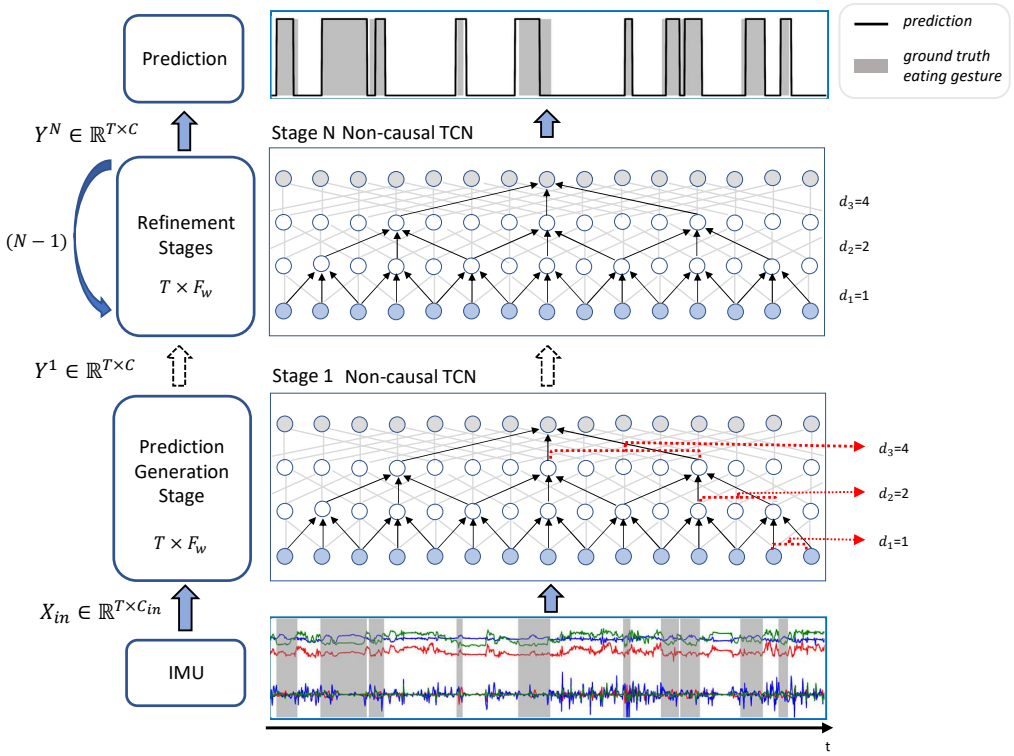


Fig. 3. The example of the pipeline for in-meal eating behavior detection. The MS-TCN model processes the IMU data to generate predictions. The MS-TCN comprises N stages of non-causal SS-TCN. The first stage is dedicated to prediction generation, while the subsequent $N-1$ stages are considered refinement stages. $d_l(l = 1, 2, 3, \ldots, L)$ represents the dilation factor of each layer, X_{in} represents the input IMU data, T is the number of data points, C_{in} represents the number of IMU data channels(12), F_w is the number of convolution filters, Y^1 represents the output of the first stage (prediction generation stage), C indicates the number of classes (2), and Y^N is the output of the final stage. It is worth noting that in the visual representation of each stage, black lines with arrows represent the prediction of a single data point. Gray lines also depict predictions for other data points. The model employs a seq2seq architecture.

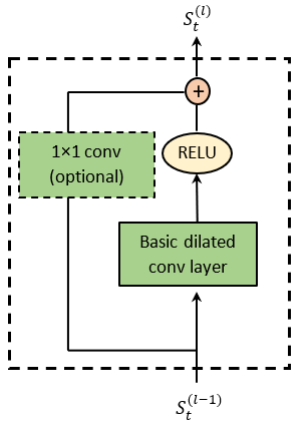


Fig. 4. Example of a dilated residual layer. There is an optional 1×1 convolution applied to the bypass path, which is used in the first layer of the first stage. In that case, the input and output have different sizes, and a 1×1 convolution can adjust them to the same size.

20 Hz for our experiment. The utensils used in the FIC dataset included forks, spoons, and knives. The dataset was collected in the restaurant of the Aristotle University of Thessaloniki, Greece. Leave-one-subject-out (LOSO) validation was used in the experiment. It should be noted that the FIC dataset contains data from only one IMU sensor on the wrist, so we adapted the input dimension of our model in the experiment.

3) OREBA Dataset: The OREBA (OREBA-DIS) dataset [50] contains 100 meal sessions from 100 participants, with 4,790 intake gestures. The dataset was collected in Australia. The utensils used in the OREBA dataset are forks & knives, spoons, and hands. Data were captured from both hands using two IMU wristbands. The hand-mirroring method was also applied to left-handed data. In our experiments, we downsampled the data from the original 64 Hz to 16 Hz. We validated the dataset with the same train/valid/test split (61/20/19).

4) Clemson Dataset: The Clemson dataset [51] contains 488 eating episodes involving 264 participants, with 20,644 intake gestures. The dataset was collected in a cafeteria at Clemson University in the United States. The utensils used in this dataset are forks & knives, spoons, hands, and chopsticks. Data collection was performed using a single IMU wristband mounted on the dominant hand, with a sampling frequency of 15 Hz. Left-handed data were processed using the hand-mirroring method. For validation, the dataset was divided into three sets with the same train/valid/test split (302/93/93).

C. The MS-TCN Model

The MS-TCN consists of multiple single-stage TCNs (SS-TCNs) stacked on top of each other. The skeleton of the MS-TCN used in this study is adopted from [8]. As illustrated in Fig. 3, the first stage, known as the prediction generation

stage, receives preprocessed IMU data $X_{in} \in \mathbb{R}^{T \times C_{in}}$ as input and generates initial predictions $Y^1 \in \mathbb{R}^{T \times C}$, where X_{in} represents the input IMU data, T is the number of data points, C_{in} represents the channels of IMU data (12), Y^1 represents the output of the first stage, and C indicates the number of classes (2). The following stages, referred to as refinement stages, further refine the initial outputs from the preceding stage. The architecture of the first SS-TCN (the prediction generation stage) is the same as the subsequent SS-TCNs (the refinement stages), with the only difference being the input dimension. In the prediction generation stage, the input is $X_{in} \in \mathbb{R}^{T \times C_{in}}$, whereas, in the SS-TCNs of the refinement stages, the input is derived from the output of the previous stage, denoted as $Y^{n-1} \in \mathbb{R}^{T \times C}$, $(2 \leq n \leq N)$, where n is the stage order, and N represents the total number of stages in MS-TCN.

In Fig. 3, each single stage employs a 'non-causal' type of TCN, indicating that the output is determined not only by past data but also by future data. Each stage is composed of a series of L convolutional layers. The number of filters F_w in each layer is the same, enabling us to apply skip connections among different layers. As depicted in Fig. 4, each layer comprises a series of dilated convolutions with ReLU activation and a residual path connection. The dilated convolution denotes the dilation factor is doubled at consecutive layers within a stage, such that $d_l = 2^{l-1}$, $(1 \leq l \leq L)$. This design permits a significant increase in the receptive field without massively increasing the number of parameters. The convolutions are applied over three time steps, $t - d_l$, t, and $t + d_l$. Let $\hat{S}_t{}^{(l)}$ represent the result of the dilated convolution in the $l - th$ layer at time t and $S_t{}^{(l)}$ denote the output after adding the residual connection such that:

$$\hat{S}_t{}^{(l)} = \mathrm{ReLU}(W^{(1)} S_{t-d_l}{}^{(l-1)} + W^{(2)} S_t{}^{(l-1)} + W^{(3)} S_{t+d_l}{}^{(l-1)} + b_d) \tag{1}$$

$$S_t{}^{(l)} = S_t{}^{(l-1)} + V \hat{S}_t{}^{(l)} + b \tag{2}$$

where $W = \{W^{(1)}, W^{(2)}, W^{(3)}\}$ with $W^{(i)} \in \mathbb{R}^{F_w \times F_w}$ are the weights of the dilated convolution filters and $V \in \mathbb{R}^{F_w \times F_w}$ represents a set of weights for the residual. b_d, $b \in \mathbb{R}^{F_w}$ are bias vectors for the dilated convolution and the residual connection, respectively. Notably, the parameters W, b_d, V, and b are different for each layer. Because the dimension of the input data in the first layer of the first stage differs from that of the output, a 1×1 convolution is applied to adjust the size, as shown in Fig. 4.

A softmax activation is applied after the last dilated residual layer to generate the prediction $\hat{y}_t \in [0, 1]^C$ for each time step t, given by:

$$\hat{y}_t = \mathrm{Softmax}(U S_t{}^{(L)} + c) \tag{3}$$

where C represents the number of classes. In this work, there are two classes, $S_t{}^{(L)}$ is the output of the last layer, with weight matrix $U \in \mathbb{R}^{C \times F_w}$ and bias $c \in \mathbb{R}^C$.

The receptive field has a length of $r(L) = (M - 1) \times 2^L - (M - 2)$, where L represents the number of layers in each stage. and M is the number of time steps needed for the convolution. For non-causal type, $M = 3$.

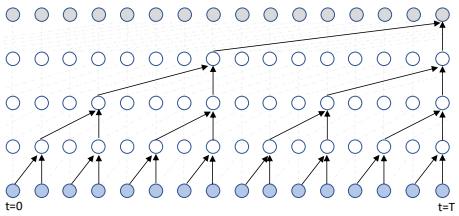


Fig. 5. Architecture of a causal TCN. There are two steps for every convolution in each layer.

Compared to the non-causal model, the output of the causal type at time t depends on data only from time 0 to t and not $t+1$ as shown in Fig. 5. Therefore, convolutions in the causal TCN only involve data over two time steps, data at time t and $t - d_l$, so we modify (1) to the equation below.

$$\hat{S}_t{}^{(l)} = \mathrm{ReLU}(W^{(1)} S_{t-d_l}{}^{(l-1)} + W^{(2)} S_t{}^{(l-1)} + b_d) \tag{4}$$

where the filters are parameterized by $W = \{W^{(1)}, W^{(2)}\}$ with $W^{(i)} \in \mathbb{R}^{F_w \times F_w}$, and b_d is the bias vector for the dilated convolution. The residual connection (2) is also applied.

A classification loss and a smoothing loss are combined to form the MS-TCN model's loss function. First, a cross-entropy loss is used to represent the classification loss:

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

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 point-wise log-probabilities is applied as a smoothing loss:

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

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

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

where T 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 [8]. The loss function for a single stage is obtained to combine the mentioned losses :

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

where n is the stage order, and λ is a parameter to determine the contribution of the two losses. The sum of the losses over all stages (N) is shown as follows:

$$\mathcal{L} = \sum_{n=1}^{N} \mathcal{L}_n \tag{10}$$

We used PyTorch to implement the MS-TCN architecture. Based on the hyperparameter tuning experiment, the non-causal MS-TCN with 2 stages and 9 layers per stage was

applied. There were 128 filters with kernel size 3 in each layer of each stage. Dropout with a probability of 0.3 was applied after each layer. The length of the receptive field was 1,023, equivalent to a duration of 64 s ($1023/16 = 64$). In the loss function, we set $\tau = 4$ and $\lambda = 0.15$ according to [8]. During training, an Adam optimizer was utilized, the learning rate was set to 0.0005, the batch size was 4, and the number of epochs was 100. We conducted all training and testing experiments on a system equipped with an Intel 9-core Xeon Gold 6140 CPUs@2.3 GHz (Skylake) with 5 GB RAM per core, and one piece of NVIDIA P100-SXM2-16 GB GPU from Vlaams Supercomputer Centrum (VSC)[3]. Notably, this was not the minimum requirement for training; the model can be trained with fewer configurations; however, this results in a longer training time. The minimum needed memory for training the model is 600 MB, making a 2 GB RAM GPU recommended for the most basic operational configurations.

IV. Evaluation and Comparison

A. Evaluation Scheme

The output of the MS-TCN model is point-wise prediction. However, the point-wise prediction cannot reflect the number of predicted eating gestures. Hence, the MS-TCN is evaluated both point-wise and segment-wise. Fig. 1 illustrates the difference between point-wise and segment-wise evaluations.

1) Point-wise Evaluation: For point-wise evaluation, the classical confusion metrics are used. Specifically, each point is categorized as true positive (TP), false positive (FP), false negative (FN), or true negative (TN). Subsequently, accuracy, precision, recall, and F1-score are calculated to evaluate the results.

2) Segment-wise Evaluation: Evaluating the point-wise performance alone lacks practical significance. It is more meaningful to understand how many eating gestures are correctly detected, falsely detected, or missed. For research that aims to detect eating gestures or chewing movements, a confusion matrix on segments obtained by a sliding window is another common evaluation scheme. However, this method may not be the best choice for all scenarios. The strict evaluation scheme proposed in [39] is another method used to evaluate the time point detection of eating gestures. Strict evaluation can be used to evaluate the model's performance on the detection task (whether there is an eating gesture). However, it does not consider the segmentation performance (the duration of each eating gesture). The choice between different evaluation schemes depends on the specific research goals. In this research, segment-wise evaluation is included to obtain better insight into the detected eating gestures.

The segmental F1-score, as introduced by [7], serves as a metric for segment-wise evaluation. Before calculating the segmental F1-score, the intersection over union (IoU) is determined for each predicted eating gesture. The IoU is the overlap ratio between the ground truth segment and the prediction segment, also known as the Jaccard index in [39]. The formula is defined as follows:

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

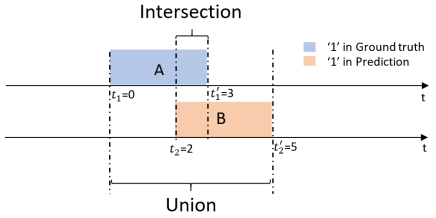


Fig. 6. Example of IoU calculation. Dividing the length of overlap between the ground truth segment and the predicted segment by the length of union yields the IoU. In this example, the intersection is 1 and the union is 5, so IoU = 0.2.

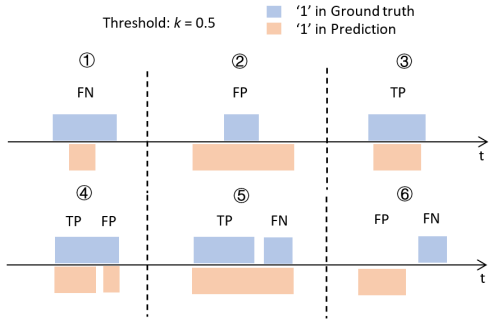


Fig. 7. Several cases of segment-wise evaluation. In the first line, the calculated IoU of examples (1) and (2) is under the threshold 0.5. After comparing the temporal length between the predicted segment and the ground truth segment, example (1) is FN because of underfill, and example (2) is FP because of overfill. Example (3) is a TP because its IoU is above 0.5. In the second line, example (4) with one ground truth segment covers two predictions. There is 1 TP and 1FP since only the first one will be counted. Likewise, example (5) contains 1 TP and 1 FN. For example in (6), there is a predicted segment without a ground truth segment and a ground truth segment without a predicted segment, resulting in 1 FP and 1 FN.

$$IoU = \frac{A \cap B}{A \cup B} = \frac{\min(t_1^{'}, t_2^{'}) - \max(t_1, t_2)}{\max(t_1^{'}, t_2^{'}) - \min(t_1, t_2)} \tag{11}$$

with ground truth segment A and predicted segment B as presented in Fig. 6, where t_1 and $t_1^{'}$ represent the starting and ending times of the ground truth eating segment, respectively, and t_2 and $t_2^{'}$ represent the starting and ending times of the predicted eating segment, respectively. Then, each segment is considered TP if its IoU score is above a pre-defined threshold k; otherwise, it is an FP or FN segment. For each segment, the decision criteria are shown below:

$$Segment = \begin{cases} TP, & IoU \geq k \\ FP, & IoU < k, length_{gt} < length_p \\ FN, & IoU < k, length_{gt} > length_p \end{cases} \tag{12}$$

where $length_{gt}$ and $length_p$ are the temporal lengths of eating gestures in the ground truth and prediction, respectively. In this experiment, we selected three IoU thresholds: 0.25, 0.5, and 0.75. Fig. 7 presents examples of segment-wise evaluation. If one ground truth gesture covered multiple predictions, only one prediction (if IoU $> k$) was marked as a TP, while all others were FPs (Fig. 7 example (4)). Similarly, if there was a single predicted segment that covered several ground truth eating gestures, we counted only one TP (if IoU $> k$), and

all others were considered as FNs (Fig. 7 example (5)). In this study, we defined the FP and FN by comparing not only the IoU threshold, but also the duration between the ground truth and predicted segments. The segment-wise evaluation scheme offers three advantages. First, it punishes overfill errors (Fig. 7 example (2)) and underfill errors (Fig. 7 example (1)) by setting the IoU threshold k; second, it tolerates minor temporal shifts between the ground truth and prediction, which may be induced by annotation variability; third, it evaluates not only the detection performance but also the segmentation performance. It should be noted that our modified evaluation scheme differs from the original scheme in the definition of FP and FN by comparing the lengths of predicted segments to ground truth segments when IoU $< k$ (as illustrated in Fig. 7 examples (1) and (2)).

B. Models for Benchmarking

1) CNN-RNN Approach: To compare the performance of our approach, we also built a CNN-LSTM model. The model consists of three subnetworks: a CNN, an LSTM network, and a fully connected network (FCN). The network's architecture is inspired by Kyritsis *et al.* [39].

Considering that we use data from two IMU sensors, one from the dominant hand and one from the head. In contrast, Kyritsis *et al.* [39] utilized only one IMU sensor, we modified the model to make it suitable for our situation. We solely adopted the CNN-LSTM model from [39] and not the entire algorithm because the post-processing method in [39] is used to detect the time point of the eating gesture, which is different from our objective of time interval detection (Fig. 7). It should also be noted that the annotation strategy in [39] is different from ours. In [39], segments are labeled as 1 only if their right ends are close to the end of ground truth eating gestures. Based on the experiment, the window length is set to 3 s, and the window step is set as 0.5 s to achieve the best performance.

In addition to the existing CNN-LSTM model, three extra CNN-RNN-based models were used for comparison: the CNN combined with bidirectional LSTM model (CNN-BiLSTM), the CNN combined with GRU model (CNN-GRU), and the CNN combined with bidirectional GRU model (CNN-BiGRU). The CNN part of these three models was identical to the CNN-LSTM model. The number of units in the BiLSTM, GRU, and BiGRU layers for the three models was set to (64, 64, 64), which were chosen based on the performance (16; 32; 64; 128). Furthermore, the ResNet-(Bi)LSTM from [40] were also applied for benchmarking. The window length for ResNet-(Bi)LSTM is 8 s without overlap based on experiments.

2) SS-TCN Approach: In this study, we also evaluated the non-causal and causal types of a single-stage TCN. The non-causal SS-TCN is the prediction generation stage in an MS-TCN, as indicated in Fig. 3. The causal SS-TCN contains 9 layers, and the number of filters in each layer is 128, with a kernel size of 2.

C. Validation Method

The LOSO method was employed to assess the model. At each LOSO step, we excluded all the data of a single subject from the dataset in the training process and utilized the omitted data for testing. The results of each iteration were aggregated to calculate the overall performance metrics.

TABLE II
THE F1-SCORES OF MS-TCN MODEL WITH DIFFERENT NUMBERS OF STAGES ON HUASHAN DATASET

MS-TCN Stages	Point-wise F1-score	Segment-wise F1-score		
		$k = 0.25$	$k = 0.5$	$k = 0.75$
1	0.807	0.944	0.915	0.735
2	0.832	0.962	0.944	0.765
3	0.826	0.952	0.926	0.750
4	0.823	0.952	0.930	0.766
5	0.834	0.958	0.936	0.752

TABLE III
POINT-WISE RESULTS WITH DIFFERENT ARCHITECTURES ON HUASHAN DATASET.

Model	Accuracy	Precision	Recall	F1-score
CNN-LSTM (modified from [39])	0.867	0.800	0.797	0.799
CNN-BiLSTM	0.878	0.826	0.802	0.813
ResNet-LSTM	0.864	0.817	0.783	0.800
ResNet-BiLSTM	0.873	0.820	0.812	0.816
CNN-GRU	0.872	0.828	0.776	0.801
CNN-BiGRU	0.880	0.834	0.796	0.815
SS-TCN (causal)	0.855	0.779	0.786	0.783
SS-TCN (non-causal)	0.872	0.809	0.805	0.807
MS-TCN (2 stage)	0.890	0.842	0.822	0.832

D. Experiments and Results

Table II shows the results of MS-TCN model with varying numbers of stages. After comparing both point-wise and segment-wise results, we selected the 2-stage MS-TCN as the final model, as increasing the number of stages did not improve the performance significantly in our case.

1) Point-wise Experiment: Table III illustrates the point-wise performance of different models on the Huashan dataset. For SS-TCN, the non-causal model achieved better results with accuracy and F1-score of 0.872 and 0.807, respectively, compared to 0.855 and 0.783 for the causal model. The 2-stage non-causal MS-TCN obtained the best performance with accuracy and F1-score of 0.890 and 0.832, respectively. The accuracy and F1-score of CNN-BiGRU were 0.880 and 0.815, respectively, which were lower than those of MS-TCN. In addition to the overall point-wise F1-score, statistical analysis using pairwise student t-tests on subject-specific F1-scores indicated a significant difference between the MS-TCN model and other models ($p < 0.01$).

2) Segment-wise Experiment: Table IV provides a performance comparison between TCN models and existing CNN-RNN models using segment-wise evaluation on the Huashan, FIC, OREBA, and Clemson datasets. For the Huashan dataset, when comparing the two types of SS-TCN models, the non-causal TCN model achieved higher segmental F1-scores of 0.944, 0.915, and 0.735 with $k = 0.25$, 0.5, and 0.75, respectively. At $k = 0.5$, the causal model had an F1-score of 0.909. The CNN-BiGRU model yielded F1-scores of 0.938, 0.910, and 0.752 with $k = 0.25$, 0.5, and 0.75, respectively. In the case of MS-TCN, the highest F1-scores of 0.962, 0.944,

TABLE IV

SEGMENT-WISE PERFORMANCE WITH DIFFERENT ARCHITECTURES ON HUASHAN, FIC, OREBA, AND CLEMSON DATASETS. THE F1-SCORES IN THIS TABLE REPRESENT SEGMENTAL F1-SCORES.

Dataset	Model	$k = 0.25$			$k = 0.5$			$k = 0.75$		
		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Huashan	CNN-LSTM (modified from [39])	0.901	0.961	0.930	0.880	0.931	0.905	0.763	0.664	0.710
	CNN-BiLSTM	0.932	0.934	0.933	0.911	0.887	0.899	0.750	0.706	0.728
	ResNet-LSTM	0.922	0.941	0.931	0.888	0.937	0.912	0.753	0.645	0.695
	ResNet-BiLSTM	0.940	0.953	0.947	0.919	0.922	0.921	0.767	0.709	0.737
	CNN-GRU	0.903	0.953	0.927	0.880	0.895	0.888	0.748	0.693	0.719
	CNN-BiGRU	0.935	0.941	0.938	0.917	0.903	0.910	0.776	0.731	0.752
	SS-TCN (causal)	0.916	0.962	0.938	0.896	0.923	0.909	0.762	0.694	0.726
	SS-TCN (non-causal)	0.925	0.963	0.944	0.901	0.929	0.915	0.743	0.726	0.735
	MS-TCN (2 stage)	0.953	0.971	0.962	0.934	0.954	0.944	0.770	0.759	0.765
FIC	CNN-LSTM (modified from [39])	0.827	0.955	0.886	0.803	0.911	0.853	0.555	0.703	0.620
	CNN-BiLSTM	0.887	0.964	0.924	0.819	0.913	0.863	0.557	0.762	0.643
	ResNet-LSTM	0.822	0.973	0.891	0.818	0.921	0.866	0.557	0.712	0.625
	ResNet-BiLSTM	0.830	0.978	0.898	0.826	0.945	0.881	0.633	0.784	0.701
	CNN-GRU	0.825	0.951	0.884	0.774	0.867	0.818	0.550	0.715	0.621
	CNN-BiGRU	0.857	0.974	0.912	0.787	0.927	0.851	0.580	0.790	0.669
	SS-TCN (causal)	0.864	0.921	0.892	0.788	0.892	0.837	0.575	0.710	0.636
	SS-TCN (non-causal)	0.917	0.936	0.927	0.904	0.888	0.896	0.622	0.675	0.647
	MS-TCN (2-stage)	0.926	0.941	0.934	0.918	0.912	0.915	0.669	0.731	0.698
OREBA	CNN-LSTM (modified from [39])	0.809	0.828	0.819	0.788	0.801	0.795	0.692	0.691	0.692
	CNN-BiLSTM	0.841	0.812	0.826	0.819	0.789	0.803	0.723	0.701	0.712
	ResNet-LSTM	0.834	0.796	0.814	0.809	0.758	0.782	0.725	0.657	0.689
	ResNet-BiLSTM	0.858	0.810	0.833	0.837	0.754	0.793	0.738	0.694	0.715
	CNN-GRU	0.781	0.819	0.800	0.756	0.787	0.771	0.647	0.689	0.667
	CNN-BiGRU	0.840	0.791	0.815	0.811	0.757	0.783	0.700	0.674	0.687
	SS-TCN (causal)	0.843	0.791	0.816	0.828	0.752	0.788	0.673	0.729	0.700
	SS-TCN (non-causal)	0.792	0.859	0.824	0.767	0.843	0.803	0.666	0.772	0.715
	MS-TCN (2-stage)	0.844	0.839	0.842	0.831	0.832	0.831	0.740	0.760	0.750
Clemson	CNN-LSTM (modified from [39])	0.808	0.855	0.831	0.742	0.775	0.758	0.562	0.570	0.566
	CNN-BiLSTM	0.840	0.865	0.853	0.780	0.813	0.796	0.588	0.635	0.610
	ResNet-LSTM	0.830	0.849	0.839	0.783	0.777	0.780	0.604	0.579	0.591
	ResNet-BiLSTM	0.855	0.860	0.857	0.811	0.803	0.807	0.609	0.672	0.639
	CNN-GRU	0.820	0.852	0.836	0.754	0.766	0.760	0.564	0.552	0.558
	CNN-BiGRU	0.840	0.870	0.855	0.776	0.818	0.796	0.569	0.634	0.600
	SS-TCN (causal)	0.798	0.868	0.831	0.742	0.792	0.766	0.565	0.567	0.566
	SS-TCN (non-causal)	0.845	0.882	0.863	0.785	0.843	0.813	0.588	0.683	0.632
	MS-TCN (2-stage)	0.882	0.886	0.884	0.831	0.830	0.831	0.674	0.653	0.663

TABLE V

SEGMENT-WISE PERFORMANCE WITH WRIST-ONLY IMU SENSOR FROM THE DOMINANT HAND AND HEAD-ONLY IMU SENSOR BY USING NON-CAUSAL MS-TCN MODEL COMPARED TO WRIST-HEAD COMBINED IMU SENSORS.

Sensor	$k = 0.25$			$k = 0.5$			$k = 0.75$		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Wrist-head	0.953	0.971	0.962	0.934	0.954	0.944	0.770	0.759	0.765
Wrist-only[a]	0.947	0.949	0.948	0.929	0.937	0.933	0.764	0.749	0.756
Head-only	0.845	0.817	0.831	0.751	0.710	0.730	0.587	0.533	0.559

[a] The wrist-only sensor is the IMU sensor from the dominant hand with 6 channels.

and 0.765 were obtained with $k = 0.25$, 0.5, and 0.75, respectively. Furthermore, the F1-score decreased as the IoU threshold increased. When comparing $k = 0.25$ to 0.5, the MS-TCN model experienced the smallest reduction in the F1-score of 1.8%, followed by 1.9% for the ResNet-LSTM, 2.5% for the CNN-LSTM. For the FIC dataset, the MS-TCN model achieved the highest segmental F1-scores of 0.934, 0.915, and 0.698 with $k = 0.25$, 0.5, and 0.75, respectively. For the OREBA and Clemson datasets, the proposed approach also attained the best performance. To evaluate the models under more rigorous conditions, we obtained their performance on the Huashan dataset for each model using different thresholds (k) ranging from 0.1 to 0.9 with increments of 0.1, as shown in Fig. 8. Notably, all models exhibited a sharp decrease in segmental F1-scores when k exceeded 0.7. The statistical analysis of the segmental F1-score under different thresholds ($k = 0.25$, 0.5, and 0.75) was added to Fig. 9. The pairwise student t-test was applied to the F1-score between the MS-TCN model and other models. Fig. 9 indicates that the results had statistical significance ($p < 0.05$) when $k = 0.25$, and 0.5. When $k = 0.75$, there was no significant difference between MS-TCN and CNN-BiGRU ($p = 0.105$).

3) IMU Placements Experiment: Table V shows the performance of the MS-TCN model with IMU sensors at different positions: wrist-head IMU, wrist-only IMU, and head-only IMU. Using the wrist-only sensor (IMU data from the dominant hand, 6 channels) yielded an F1-score of 0.933, while using the head-only sensor (IMU data from eyeglasses, 6 channels) obtained an F1-score of 0.730 ($k = 0.5$). Compared with the result using wrist-head IMUs (data from the IMU wristband and eyeglasses, 12 channels), using the wrist-only sensor resulted in a 1.1% reduction in performance at $k = 0.5$ with statistical significance ($p < 0.05$), and using head-only data obtained a 21.4% drop ($k = 0.5$). At $k = 0.75$, the F1-

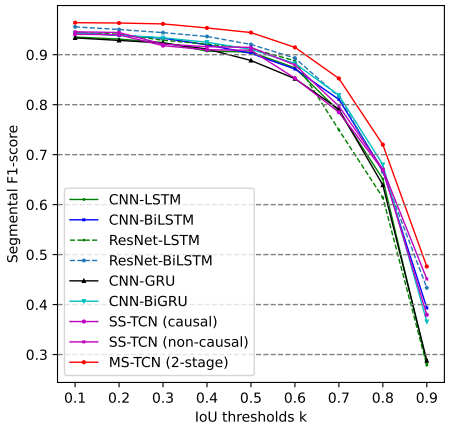


Fig. 8. Segmental F1-scores under different IoU thresholds, ranging from 0.1 to 0.9 with increments of 0.1.

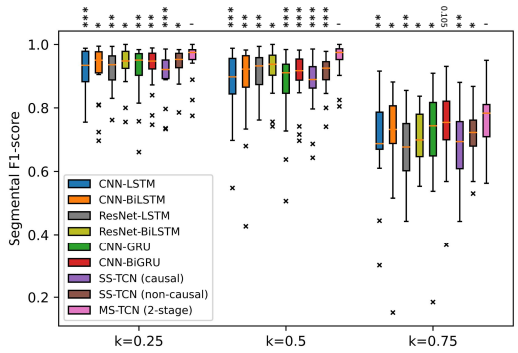


Fig. 9. The box plot shows the distribution of segmental F1-scores for 24 participants using different models. Three IoU thresholds are applied from left to right (k = 0.25, 0.5, 0.75). The statistical analysis is obtained by applying the pairwise t-test between MS-TCN and other models for segmental F1-score. p<0.05: *, p<0.01: **, p<0.001: ***.

score of the wrist-head IMU data was higher than that of the wrist-only data without statistical significance ($p = 0.09$).

V. Discussion

This study compares the MS-TCN model with CNN-RNN models on the Huashan, FIC, OREBA, and Clemson datasets for eating gesture detection. The results show that the applied model outperforms CNN-RNN models. The results in this study are in line with studies in the literature [8], [42], [52]. The superiority of the applied MS-TCN model relies on two factors. First, the MS-TCN processes spatial-temporal features simultaneously, whereas the CNN-RNNs first extract spatial features from the CNN and then feed these features into the RNN module to learn temporal relations. This two-step decoupling paradigm may prevent capturing more nuanced spatial-temporal relationships [53]. Second, given the significant variation in the time taken for one eating gesture among older adults, both short-term and long-term temporal processing capabilities are needed for models. The MS-TCN model incorporates multiple dilation factors by stacking a series of dilated convolution layers to capture temporal information at different temporal resolutions. Subsequently, MS-TCN can capture both short-term and long-term dependencies in the data, while an existing study [54] shows that RNNs cannot process long temporal sequences effectively due to the vanishing/exploding gradient problem.

Compared to existing approaches that aim to detect time points of eating gestures [39], our approach has the advantage of segmenting time intervals for each eating gesture. Consequently, our method provides more comprehensive information, allowing for the estimation of both the duration of the gesture and the time gap between gestures. In existing approaches, the detected point can be anywhere within the interval of the corresponding eating gesture, which introduces more uncertainty when estimating the gap between two eating gestures. Beyond its utility for food intake monitoring, the segmented duration of eating gestures holds the potential for additional practical applications. One potential application of our approach is assessing eating difficulties and exploring quality of life (QoL) challenges in older adults, because the obtained eating gesture segments can provide better insight into the motion characteristics during meals, such as the time taken for each eating gesture and the speed of hand movements when transferring food from plate to mouth [55]–[57].

In this experiment, several preprocessing steps, including smoothing, normalization, and gravity removal, were explored before training deep learning models. However, there was no significant performance difference compared to using raw data directly. This outcome aligns with the conclusions drawn in a related study [50]. Therefore, we chose to train the deep learning model with raw IMU data.

We compared the wrist-head combined IMU placement to the head-only placement and the wrist-only placement. The results indicate that the wrist-head combined IMU placement achieves the best performance. The wrist-only placement yields an F1-score of 0.933 at $k = 0.5$. Upon reviewing the predictions and video, we observed that the wrist-only placement tends to miss eating gestures when participants primarily move their heads downward rather than raising their hands. In contrast, the wrist-head combined IMU system effectively addresses this limitation. The head-only sensor exhibits a significantly lower F1-score. This outcome aligns with previous research conducted within our lab, which suggests that IMU sensors mounted on eyeglasses may be better suited for tasks related to chewing detection [29].

The proposed approach that utilizes the wrist-head combined IMU system can be applied to specific target groups requiring precise recording of eating gestures, such as post-surgical patients in hospitals. The results from the wrist-only sensor are still rather good (0.944→0.933, $k = 0.5$), as shown in Table V, indicating that the proposed data processing pipeline is also applicable when using a singular sensor (e.g., a smartwatch) for eating gesture detection, which is more acceptable for individuals who function independently in real life. However, it's important to note that we apply hand-mirroring for the data of left-handed participants, which requires clear identification of the used hand (left/right) prior to prediction.

Three IoU thresholds were applied for evaluation, with $k = 0.25$ representing the lowest threshold and $k = 0.75$ representing the highest threshold. The number of FN and FP segments at $k = 0.25$ represents the number of eating

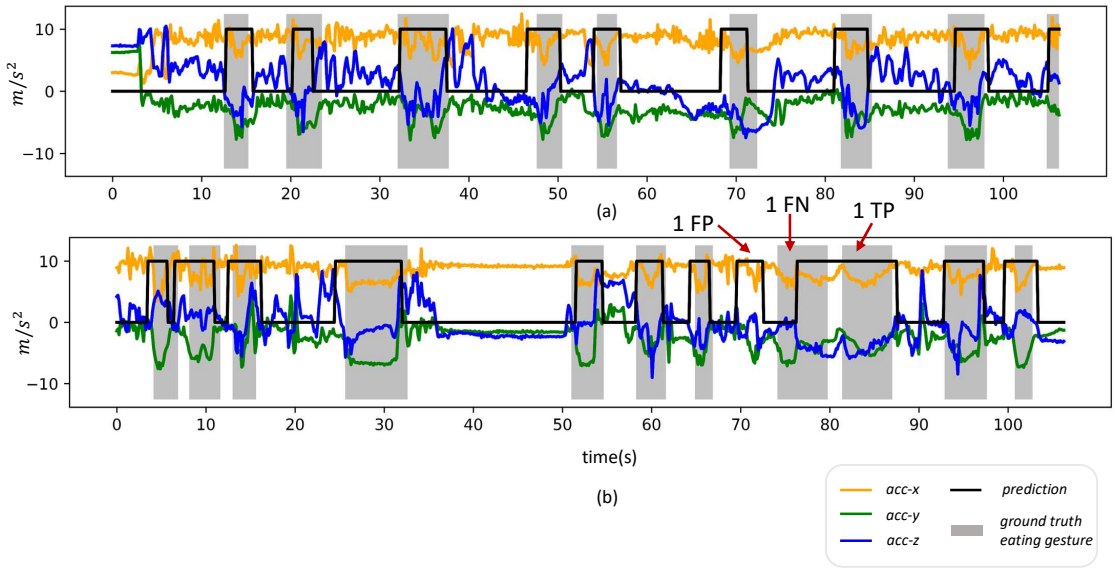


Fig. 10. Two examples of results from test meal sessions. Fig. (a) shows a series of successfully detected eating gestures. Fig. (b) shows another series of eating gestures, which contains 1 FP and 1 FN. The FP segment represents a deceptive movement; the FN segment merges two ground truth eating gestures.

segments that were entirely undetected and segments that were incorrectly predicted (example (6) in Fig. 7). The performance of the MS-TCN at $k = 0.5$ exhibits only a marginal reduction compared to the results obtained at $k = 0.25$, indicating that the robustness of eating gesture segmentation is generally high for a non-causal MS-TCN ($k \leq 0.5$), as illustrated in Fig. 9. The figure showcases the distribution of the segmented F1-score for 24 participants.

Two types of SS-TCN models are presented: causal and non-causal. According to the results in Table IV, the non-causal variant exhibited better performance. However, a non-causal architecture has limitations, such as its inability for real-time prediction. This limitation arises from the fact that a non-causal model utilizes not only previous and current data but also future information. In contrast, a causal model relies solely on previous temporal information and current data. It is worth noting that causal type TCN still performs well compared to CNN-LSTM and CNN-GRU.

Although all training and testing tasks were conducted offline utilizing an NVIDIA GPU, the proposed approach is feasible for deployment on mobile devices, such as smartphones, enabling real-time execution. For predicting a 1-min IMU data segment, the estimated floating point operations (FLOPs) for MS-TCN was 1.13 GFLOPs, which is substantially lower than the number of floating point operations per second (FLOPS) of GPUs in mobile devices [58]. Moreover, the size of the trained model is 1.19 MB, and the minimum memory for running the model is under 100 MB. In practice, wrist-worn IMU sensors stream real-time data to a smartphone via Bluetooth, and the smartphone can run the deployed model to detect eating gestures.

The precision shows a lower value than recall across all models at threshold $k = 0.25$ on the Huashan dataset, as illustrated in Table IV. This discrepancy suggests that the number of FPs is higher than that of FNs, indicating that these models tend to misclassify some other movements as eating gestures. By comparing the output and the annotation video, we found that the model classifies some deceptive movements as eating gestures, such as wiping the mouth with a napkin, which is similar to eating with the hand, as illustrated in Fig. 10. For the two types of SS-TCN models, the causal SS-TCN suffers more from this because its predictions rely solely on current and previous data. This problem in the CNN-LSTM model is more pronounced.

The participants in the Huashan dataset are older Asian individuals, and the utensils used in the dataset include chopsticks, spoons, and bare hands. However, it is essential to emphasize that the proposed method is not confined to this specific context, as evidenced by its segmental F1-scores on the FIC, OREBA, and Clemson datasets, showcasing its ability to detect Western-style eating gestures involving forks and knives. By inspecting the results from MS-TCN, the segmental recall (true positive rate/sensitivity) of using chopsticks, spoons, and hands is 0.971, 0.935, and 0.784, respectively ($k = 0.5$). The segmental recall of using hands was the lowest. One potential explanation for this disparity could be the comparatively smaller quantity of eating gestures with hands compared to those performed with chopsticks or spoons (Table I).

This work utilized MS-TCN to effectively capture long-term dependencies in the data by using dilated convolutions, which expands the effective receptive field without significantly increasing the number of parameters. The results show that the MS-TCN model outperforms other existing models. Meanwhile, adopting of the segment-wise evaluation method enables us to evaluate both detection and segmentation performance. Compared to the wrist-only IMU system, the wrist-head combined IMU system can effectively detect eating gestures.

It is important to note that the proposed approach can also be applied to detect eating gestures in other age groups. The performance of our approach on three public datasets, as shown in Table IV, underscores this capability. We use the term 'older persons' in the title to highlight the unique health challenges faced by this age group and to emphasize the importance of addressing their specific needs. This study focuses on older adults, who are underrepresented in existing datasets like FIC, OREBA and Clemson (none in FIC and OREBA, and fewer in Clemson), making our dataset valuable in bridging this research gap. Additionally, this study applies wearable IMU sensors to detect eating gestures in older adults. Considering the health conditions of older adults, non-wearable contactless sensors, such as depth sensors [59], sonars [60], and radar sensors [61], which have been investigated for other human activity recognition, are also worth exploring for food intake monitoring, as these sensors raise lower privacy concerns and can provide richer information.

The primary output of our approach is eating gesture detection, which involves counting the number of bites. Other potential information can also be obtained, such as the time taken for one eating gesture, and the time duration between two eating gestures, which can be used to estimate the eating speed, a metric directly related to human obesity and diabetes [62], [63]. Such a system can also be extended to full-day monitoring, which is part of our plans. We can estimate the meal time during a day and the number of meals a person has consumed by analyzing the eating gesture distribution.

Despite the superior performance of the approach, several limitations persist in this work. First, the non-causal type MS-TCN introduces a time delay for prediction, equivalent to 1/2 of the receptive field (32 s). Second, the Huashan dataset lacks two-handed eating data. To further explore the potential of the approach, additional experiments are needed, which should include recordings involving participants engaging in two-handed eating activities. Third, the eating gesture detection system cannot identify or quantify actual food intake. Combining it with additional sensors, such as a smart plate [12], has the potential to advance calorie intake estimation in daily life.

VI. Conclusions

In this study, we developed an approach that utilizes wrist-head combined IMU sensors and applies the seq2seq MS-TCN model for eating gesture detection in older adults. Instead of segmenting the input data using a sliding window, the MS-TCN model applies dilated convolutions to expand the receptive field. We also introduced an adapted segment-wise evaluation scheme that can assess both the detection and segmentation performance of eating gestures. The results demonstrated the feasibility of the proposed approach for eating gesture detection and segmentation. In the future, we plan to extend our research to detect food intake behavior over longer periods rather than meal sessions. Additionally, we aim to develop and validate the proposed data processing approach further using a larger dataset with a diverse population and various types of utensils, enabling us to detect a wider range of food intake-related activities.

Acknowledgment

The authors would like to gratefully acknowledge the support and participation of the staff and visitors of Huashan Hospital in Shanghai, China, in the data collection process. This work was supported in part by the National Key R&D Program of China under Grants 2018YFC2002300 and 2018YFC2002301; and in part by China Scholarship Council (CSC grant number: 202007650018).

References

[1] W. H. Organization, *Obesity: Preventing and Managing the Global Epidemic. Report of a WHO consultation.* Geneva: World Health Organization, 2000.
[2] "Obesity and overweight." https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (accessed Jun. 05, 2021).
[3] "Overweight and obesity - BMI statistics." https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Overweight_and_obesity_-_BMI_statistics (accessed Jun. 05, 2021).
[4] L. M. Donini *et al.*, "Malnutrition in elderly: Social and economic determinants," *J. Nutr. Heal. Aging*, vol. 17, no. 1, pp. 9–15, Jan. 2013.
[5] E. M. Mathus-Vliegen, "Obesity and the elderly," *J. Clin. Gastroenterol*, vol. 46, no. 7, pp. 533-544, Aug. 2012.
[6] N. A. Selamat and S. H. M. Ali, "Automatic food intake monitoring based on chewing activity: A survey," *IEEE Access*, vol. 8, pp. 48846–48869, Mar. 2020.
[7] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," In *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, 2017, pp. 1003–1012.
[8] 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.
[9] T. Vu, F. Lin, N. Alshurafa, and W. Xu, "Wearable food intake monitoring technologies: A comprehensive review," *Computers*, vol. 6, no. 1, pp. 1–28, Jan. 2017.
[10] B. Zhou *et al.*, "Smart table surface: A novel approach to pervasive dining monitoring," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2015, pp. 155–162.
[11] R. S. Mattfeld, E. R. Muth, and A. Hoover, "Measuring the consumption of individual solid and liquid bites using a table-embedded scale during unrestricted eating," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 6, pp. 1711-1718, Nov. 2017.
[12] 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, May. 2020.
[13] K. S. Lee, "Joint audio-ultrasound good recognition for noisy environments," *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 5, pp. 1477–1489, May. 2020.
[14] J. Liu *et al.*, "An intelligent food-intake monitoring system using wearable sensors," in *Proc. 9th Int. Conf. Wearable Implant. Body Sensor Netw.*, 2012, pp. 154–160.
[15] K. Lee, "Food intake detection using ultrasonic Doppler sonar," *IEEE Sens. J.*, vol. 17, no. 18, pp. 6056-6068, Sep. 2017.
[16] M. Pedram et al., "LIDS: Mobile system to monitor type and volume of liquid intake," *IEEE Sens. J.*, vol. 21, no. 18, pp. 20750–20763, 2021.
[17] 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 *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2016, pp. 6485-6488.
[18] V. Papapanagiotou, C. Diou, L. Zhou, J. van den Boer, M. Mars, and A. Delopoulos, "A novel chewing detection system based on PPG, audio, and accelerometry," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 3, pp. 607–618, May. 2017.
[19] R. Zhang, S. Bernhart, and O. Amft, "Diet eyeglasses: Recognising food chewing using EMG and smart eyeglasses," in *Proc. 13th Int. Conf. Wearable Implant. Body Sensor Netw.*, 2016, pp. 7–12.
[20] D. Hossain, M. H. Imtiaz, and E. Sazonov, "Comparison of wearable sensors for estimation of chewing strength," *IEEE Sens. J.*, vol. 20, no. 10, pp. 5379–5388, 2020.
[21] T. Ghosh, D. Hossain, and E. Sazonov, "Detection of food intake sensor's wear compliance in free-living," *IEEE Sens. J.*, vol. 21, no. 24, pp. 27728–27735, 2021.

[22] H. He, F. Kong, and J. Tan, "DietCam: Multiview food recognition using a multikernel SVM," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 3, pp. 848–855, May. 2016.

[23] S. Mezgec and B. K. Seljak, "Nutrinet: A deep learning food and drink image recognition system for dietary assessment," *Nutrients*, vol. 9, no. 7, pp. 1–19, Jun. 2017.

[24] J. Qiu, F. P. W. Lo, S. Jiang, C. Tsai, Y. Sun, and B. Lo, "Counting bites and recognizing consumed food from videos for passive dietary monitoring," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 5, pp. 1471-1482, May 2021.

[25] V. B. Raju and E. Sazonov, "A systematic review of sensor-based methodologies for food portion size estimation," *IEEE Sens. J.*, vol. 21, no. 11, pp. 12882–12899, 2021.

[26] V. B. Raju, D. Hossain, and E. Sazonov, "Estimation of plate and bowl dimensions for food portion size assessment in a wearable sensor system," *IEEE Sens. J.*, vol. 23, no. 5, pp. 5391–5400, 2023.

[27] M. Farooq and E. Sazonov, "Accelerometer-based detection of food intake in free-living individuals," *IEEE Sens. J.*, vol. 18, no. 9, pp. 3752–3758, May. 2018.

[28] S. Zhang, R. Alharbi, W. Stogin, M. Pourhomayun, B. Spring, and N. Alshurafa, "Food watch: Detecting and characterizing eating episodes through feeding gestures," in *Proc. 11th EAI Int. Conf. Body Area Netw.*, 2016, pp. 91–96.

[29] G. Mertes, H. Hallez, B. Vanrumste, and T. Croonenborghs, "Detection of chewing motion in the elderly using a glasses mounted accelerometer in a real-life environment," in *Proc. IEEE 39th Annu. Int. Conf. Eng. Med. Biol. Soc.*, Jeju, Korea (South), 2017, pp. 4521-4524.

[30] K. Kyritsis, C. Diou and, A. Delopoulos, "Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 6, pp. 2325-2334, Nov. 2019.

[31] K. C. Liu, C. Y. Hsieh, H. Y. Huang, L. T. Chiu, S. J. P. Hsu, and C. T. Chan, "Drinking event detection and episode identification using 3D-Printed smart cup," *IEEE Sens. J.*, vol. 20, no. 22, pp. 13743–13751, 2020.

[32] S. Zhang *et al.*, "NeckSense: A multi-sensor necklace for detecting eating activities in free-living conditions," in *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 2, 2020, pp. 1–26.

[33] 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, Feb. 2021.

[34] R. Zhang and O. Amft, "Monitoring chewing and eating in free-living using smart eyeglasses," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 1, pp. 23–32, Jan. 2018.

[35] F. Sufyan, S. Sagar, Z. Ashraf, S. Nayel, M. S. Chishti, and A. Banerjee, "A novel and lightweight real-time continuous motion gesture recognition algorithm for smartphones," *IEEE Access*, vol. 11, no. March, pp. 42725–42737, 2023.

[36] M. Dehghani, K. J. Kim, and R. M. Dangelico, "Will smartwatches last? factors contributing to intention to keep using smart wearable technology," *Telemat. Informatics*, vol. 35, no. 2, pp. 480–490, May. 2018.

[37] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 81–94, 2016.

[38] Y. Dong, A. Hoover, J. Scisco, and E. Muth, "A new method for measuring meal intake in humans via automated wrist motion tracking," *Appl. Psychophysiol. Biofeedback*, vol. 37, no. 3, pp. 205–215, Sep. 2012.

[39] 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, Jan. 2021.

[40] 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, Jul. 2021.

[41] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499.

[42] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*. [Online]. Available: http://arxiv.org/abs/1803.01271.

[43] B. Filtjens, P. Ginis, A. Nieuwboer, P. Slaets, and B. Vanrumste, "Automated freezing of gait assessment with marker-based motion capture and multi-stage spatial-temporal graph convolutional neural networks," *J. Neuroeng. Rehabil.*, vol. 19, no. 1, pp. 1–14, May. 2022.

[44] Z. Wang and B. Yao, "Multi-branching temporal convolutional network for sepsis prediction," *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 2, pp. 876-887, Feb. 2022.

[45] B. Zhang *et al.*, "Towards accurate surgical workflow recognition with convolutional networks and transformers," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 10, no. 4, pp. 349–356, 2022.

[46] C. Wang, T. S. Kumar, W. De Raedt, G. Camps, H. Hallez, and B. Vanrumste, "Drinking gesture detection using wrist-worn IMU sensors with multi-stage temporal convolutional network in free-living environments," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2022, pp. 1778–1782.

[47] H. Heydarian, P. V. Rouast, M. T. P. Adam, T. Burrows, C. E. Collins, and M. E. Rollo, "Deep learning for intake gesture detection from wrist-worn inertial sensors: The effects of data preprocessing, sensor modalities, and sensor positions," *IEEE Access*, vol. 8, pp. 164936–164949, Sep. 2020.

[48] H. Sloetjes and P. Wittenburg, "Annotation by category - ELAN and ISO DCR," In *Proc. 6th Int. Conf. Lang. Resour. Eval. Lr.*, 2008, pp. 816–820.

[49] D. G. Lamb *et al.*, "The aging brain: Movement speed and spatial control," *Brain Cogn.*, vol. 109, pp. 105–111, Sep. 2016.

[50] 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. 181955–181963, 2020.

[51] 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.

[52] S. Gopali, F. Abri, S. Siami-Namini, and A. S. Namin, "A comparison of TCN and LSTM models in detecting anomalies in time series data," In *Proc. IEEE Int. Conf. Big Data, Big Data 2021*, 2021, pp. 2415–2420.

[53] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," In *ECCV 2016*, 2016, pp. 47–54.

[54] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1961–1970.

[55] A. Salarian, H. Russmann, C. Wider, P. R. Burkhard, F. J. G. Vingerhoets, and K. Aminian, "Quantification of tremor and bradykinesia in Parkinson's disease using a novel ambulatory monitoring system," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 2, pp. 313-322, Feb. 2007.

[56] D. J. Wile, R. Ranawaya, and Z. H. T. Kiss, "Smart watch accelerometry for analysis and diagnosis of tremor," *J. Neurosci. Methods.*, vol. 230, pp. 1–4, Jun. 2014.

[57] K. Kyritsis *et al.*, "Assessment of real life eating difficulties in Parkinson's disease patients by measuring plate to mouth movement elongation with inertial sensors," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, Jan. 2021.

[58] A. Ignatov *et al.*, "AI benchmark: Running deep neural networks on android smartphones," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 288-314.

[59] Y. Endo and C. Premachandra, "Development of a bathing accident monitoring system using a depth sensor," in *IEEE Sensors Lett.*, vol. 6, no. 2, pp. 1-4, Feb. 2022.

[60] S. Franceschini, M. Ambrosanio, V. Pascazio, and F. Baselice, "Hand gesture signatures acquisition and processing by means of a novel ultrasound system," *Bioengineering*, vol. 10, no. 1, pp. 1–13, 2023.

[61] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks," *IEEE Sens. J.*, vol. 19, no. 8, pp. 3041–3048, 2019.

[62] P. Fagerberg *et al.*, "Fast eating is associated with increased BMI among high-school students," *Nutrients*, vol. 13, no. 3, pp. 1–19, 2021.

[63] A. Kudo *et al.*, "Fast eating is a strong risk factor for new-onset diabetes among the Japanese general population," *Sci. Rep.*, vol. 9, no. 1, pp. 1–8, Dec. 2019.