IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 February 2024, accepted 25 February 2024, date of publication 29 February 2024, date of current version 14 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3371517

RESEARCH ARTICLE

PPGCN: Phase-Aligned Periodic Graph Convolutional Network for Dual-Task-Based Cognitive Impairment Detection

ÁKOS GODÓ[1], SHUQIONG WU[1], FUMIO OKURA[2], (Member, IEEE), YASUSHI MAKIHARA[1], MANABU IKEDA[3], SHUNSUKE SATO[3,4], MAKI SUZUKI[5], YUTO SATAKE[3], DAIKI TAOMOTO[3], AND YASUSHI YAGI[1], (Senior Member, IEEE)

[1]SANKEN, Osaka University, Osaka, Ibaraki 565-0871, Japan
[2]Graduate School of Information Science and Technology, Osaka University, Osaka, Suita 565-0871, Japan
[3]Department of Psychiatry, Graduate School of Medicine, Osaka University, Osaka, Suita 565-0871, Japan
[4]Department of Psychiatry, Esaka Hospital, Osaka, Suita 565-0871, Japan
[5]Department of Behavioral Neurology and Neuropsychiatry, United Graduate School of Child Development, Osaka University, Osaka, Suita 565-0871, Japan

Corresponding author: Ákos Godó (godo@am.sanken.osaka-u.ac.jp)

This work was supported in part by the Japan Agency for Medical Research and Development (AMED) under Grant JP23uk1024001, and in part by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP20H00607.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Osaka University Clinical Research Review Committee, Suita, Japan under Application No. 21236.

ABSTRACT Early detection methods for cognitive impairment are crucial for its effective treatment. Dual-task-based pipelines that rely on skeleton sequences can detect cognitive impairment reliably. Although such pipelines achieve state-of-the-art results by analyzing skeleton sequences of periodic stepping motion, we propose that their performance can be improved by decomposing the skeleton sequence into representative phase-aligned periods and focusing on them instead of the entire sequence. We present the phase-aligned periodic graph convolutional network, which is capable of processing phase-aligned periodic skeleton sequences. We trained it with a cross-modality feature fusion loss using a representative dataset of 392 samples annotated by medical professionals. As part of a dual-task cognitive impairment detection pipeline that relies on two-dimensional skeleton sequences extracted from RGB images to improve its general usability, our proposed method outperformed existing approaches and achieved a mean sensitivity of 0.9231 and specificity of 0.9398 in a four-fold cross-validation setup.

INDEX TERMS Cognition, convolutional neural networks, dementia, task analysis.

I. INTRODUCTION

In 2023, the World Health Organization estimated that there were over 55 million people worldwide suffering from dementia, with 10 million new cases projected per year [1]. The early and reliable detection of mild cognitive impairment (MCI) or dementia are paramount for early and effective treatment.

The clinical diagnosis of dementia and MCI are based on clinical criteria [2]. Although clinical symptoms and courses are essential for diagnosis, various auxiliary assessments can support the diagnostic procedure. It is advantageous if the assessment can be performed repeatedly and reliably, which allows the tracking of patients' cognitive decline and adjustment of treatment, if necessary.

It is important to note that although these assessments are helpful, alone, they are not capable of providing a full clinical diagnosis of either MCI or dementia. Given limited examination times, their applicability greatly depends

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

on their average sensitivity (ratio of correct true positive predictions) and specificity (ratio of correct true negative predictions), which indicate how well the assessment method predicts the ground truth diagnoses based on clinical criteria.

The diagnosis of cognitive decline can be assisted by imaging-based assessment, for example, using computer tomography (CT) [3] or magnetic resonance imaging (MRI) [4] to detect biomarkers linked to dementia and MCI in brain blood vessels. MRI has a sensitivity between 78% and 84% and a specificity between 93% and 98% for the diagnosis of dementia [5]. Unfortunately, CT and MRI examinations take a relatively long time (20-60 minutes) [6], may not be comfortable for the subject, and the limited availability of the required CT and MR machines makes it difficult to rely on such methods in the long run.

By contrast, paper- and test-based assessments are significantly faster to perform and may also be used as an indicator of cognitive decline. A popular example is the Mini-Mental State Examination (MMSE) [7], which consists of a set of predetermined questions that subjects answer on a test sheet. The MMSE assigns a score of 0-30. Subjects that score below 28 are classified as potentially having MCI and those that score below 24 are classified as potentially having dementia.

The MMSE's sensitivity, specificity, and time to administer vary based on the version of the test; different language and question set variants exist. The original MMSE [7] takes between 5 and 10 minutes to administer and has a sensitivity of 81% and specificity of 89% for classifying dementia, whereas MCI classification has a sensitivity between 45% and 60% and a specificity of 65%-90%. The Japanese variant MMSE-J [8] has a sensitivity of 86% and specificity of 89%, and takes somewhat longer: between 10 and 15 minutes to administer. It is important to note that the reliability of the MMSE is not perfect because of its literacy requirements and use of preset, memorizable questions that affect the results [9].

The dual-task paradigm is a behavioral assessment that aims to overcome the above limitations. In this assessment, subjects perform a motor task and cognitive task first separately (single task) and then simultaneously (dual task).

The motor task may be walking on a treadmill or in place (stepping). Gait has a strong correlation with cognition [10]; hence, a deterioration in motor task performance caused by increased mental load while performing the cognitive task is an indicator of a subject's cognitive state [11].

The cognitive task may be the recital of the months in reverse order [12] or naming animals shown randomly on a screen, which solves the memorization issue in the cognitive task [13]. The assessment takes only minutes to perform and yields a wealth of data. The sensitivity and specificity of detection rely heavily on the methods used to analyze the collected motor and cognitive task data, and are still being explored.

In this paper, we build on the previous study conducted by Matsuura et al. [14], who presented a dual-task assessment based on stepping (motor task) and a randomly generated mental arithmetic quiz (cognitive task). However, in the pipeline used to process the data collected during the dual-task assessment, the researchers did not consider the periodicity (Fig. 1) of the stepping motion, which we recognize as being highly important for predicting a subject's cognitive state.

Therefore, we propose a novel approach to gait-based dual-task cognitive impairment assessment. By recognizing the periodicity of the stepping motion and focusing on representative gait periods and interperiod (in)consistencies rather than the entire skeleton sequence, we show that the overall performance of the prediction pipeline can be improved.

Given that periodicity is a long-term dependency, it is difficult to capture using convolutional networks unless large convolutional kernels are used. However, the use of such large kernel sizes reduces sensitivity to shorter-term features. We propose a representation that decomposes the temporal dimension of the skeleton sequences into phase-aligned periods. The resulting phase-period representation places the same phases in close proximity in the phase-period space. This decouples the period dimension from the time (phase) dimension, thus overcoming the need for large kernel sizes. Our novel neural network architecture, the phase-aligned periodic graph convolutional network (PPGCN), is capable of processing the extracted phase-period decomposed skeleton sequences.

To process the collected data and predict cognitive impairment, we present a modified dual-task pipeline (Fig. 2). We extract the skeleton sequences from RGB images of subjects performing the stepping task. Then we decompose the skeleton sequences into phase-aligned periods and process them using the proposed PPGCN architecture.

In previous studies [15], [16], task features were reduced to logits and combined to form the output probabilities, which resulted in a loss of feature information. In our study, we modify the presented pipeline so that the binary predictions (healthy vs. MCI or dementia) are based on the combined features of both motor and cognitive tasks.

We summarize the main contributions of this study as follows:

- We propose a periodic approach to skeleton-based cognitive impairment detection and present the PPGCN architecture, which is capable of handling phase-period decomposed skeleton sequences.
- Additionally, we introduce a pairwise feature distance loss alongside cross-modality feature level fusion into our multi-modal pipeline (Fig. 2) to further improve the separation of healthy and cognitively impaired subjects.

II. RELATED WORK

A. PERIODIC SIGNAL EXTRACTION

The motion of joints in the skeleton sequence during the stepping task is quasi-periodic: the individual joints largely follow similar trajectories, but because of fluctuations in both human walking patterns, particularly present during the

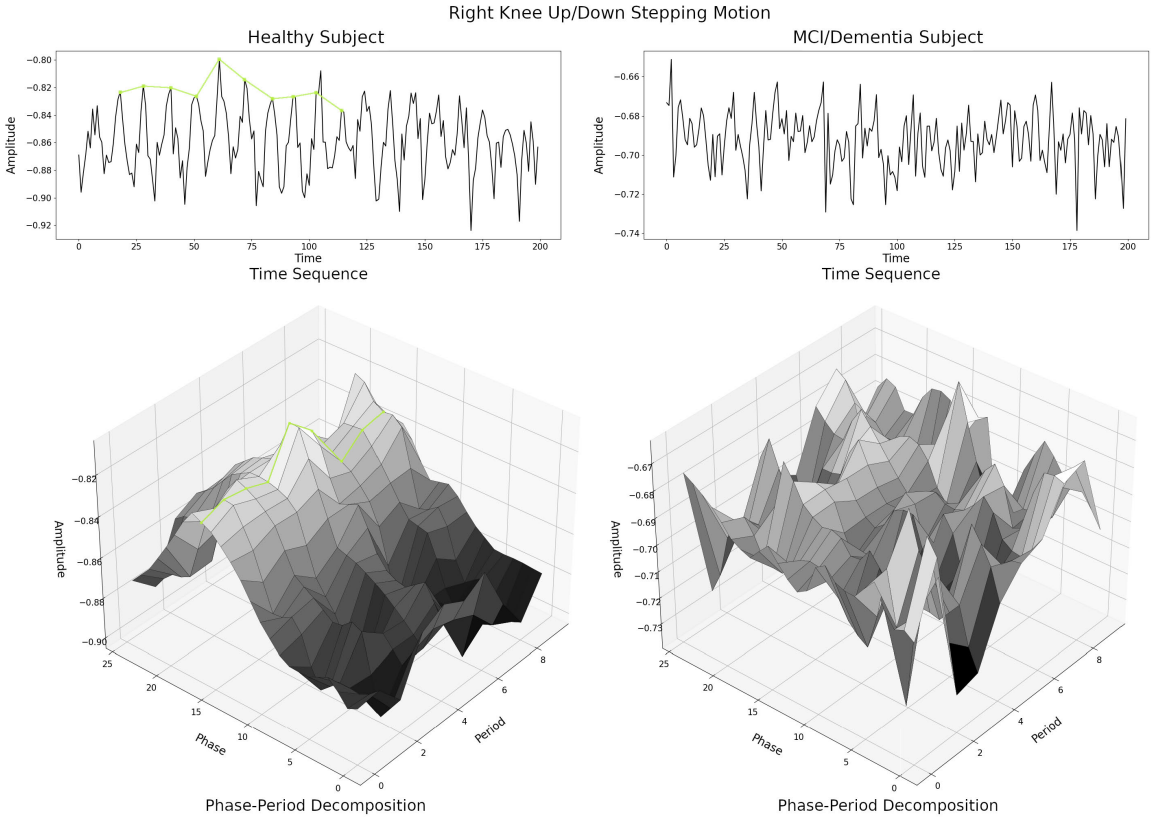


FIGURE 1. Joint motion in the skeleton sequence is periodic; however, this is less pronounced for cognitively impaired subjects (right) than healthy subjects (left). We propose that by recognizing and leveraging this property, gait-based dual-task cognitive impairment detection performance can be improved. Decomposing the temporal sequences (top) into phase-aligned periods (bottom) allows for the analysis of interperiod fluctuations in the same phase. The phase-period representation places the same phases of different periods in close proximity in phase-period space, which allows convolutional networks to capture these long-term dependencies across periods without prohibitively large kernel sizes, unlike in the temporal representation (green line).

dual-task assessment, and the sampling interval (unstable recording speed), the phase and amplitude of the signal alters across periods.

Various [17], [18] phase registration and period extraction methods exist, but require a reference signal, which is not available in our case. We rely on Makihara et al.'s self-dynamic time warping (Self-DTW) [19], which implements phase registration and period extraction from a single quasi-periodic signal without a reference signal.

B. GRAPH CONVOLUTIONAL NETWORKS

Choosing a neural network architecture that can handle the underlying structure of input data is essential for achieving high performance for any application. Similar to convolutional neural networks (CNN) [20] capturing the spatial relationship and recurrent neural networks [21] capturing the temporal structure within their respective inputs, graph convolutional networks (GCN) [22], [23] effectively handle the graph-based structure of the input data.

The graph structure, which consists of nodes and edges, within the input data may represent dependencies or connections (edges) between data points (nodes). GCNs operate by performing the graph convolution operation (1), which transforms node features based on the graph's adjacency matrix:

$$X_{l+1} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X_l W_l) \tag{1}$$

where X_l is the input and X_{l+1} is the output of layer l. X_0 denotes the input graph's node values. $\hat{A} = A + I$, where A is the graph's adjacency matrix and I is the identity matrix. Using $\hat{A}$ instead of A creates a self-connection to at least the same nodes of the previous layer's features, which allows the forward propagation of features. $\hat{D}$ is the diagonal degree matrix, which normalizes the number of connections to keep the feature scale (X_l) normalized. W_l are the learnable parameters of the graph convolutional layer and σ is an activation function.

GCNs generally operate by transforming the graph structure into an array that can be handled by CNNs and apply the above operation. By extracting fixed-size locally connected regions from the graph's nodes and edges (sub-graphs) and applying CNN-style convolutions, PATCHY-SAN [24] and LGCN [25] have achieved state-of-the-art results. Recent applications [26] of GCNs include communication networks [27], COVID-19 diagnosis [28] and computer vision [29], demonstrating the versatility of GCNs.

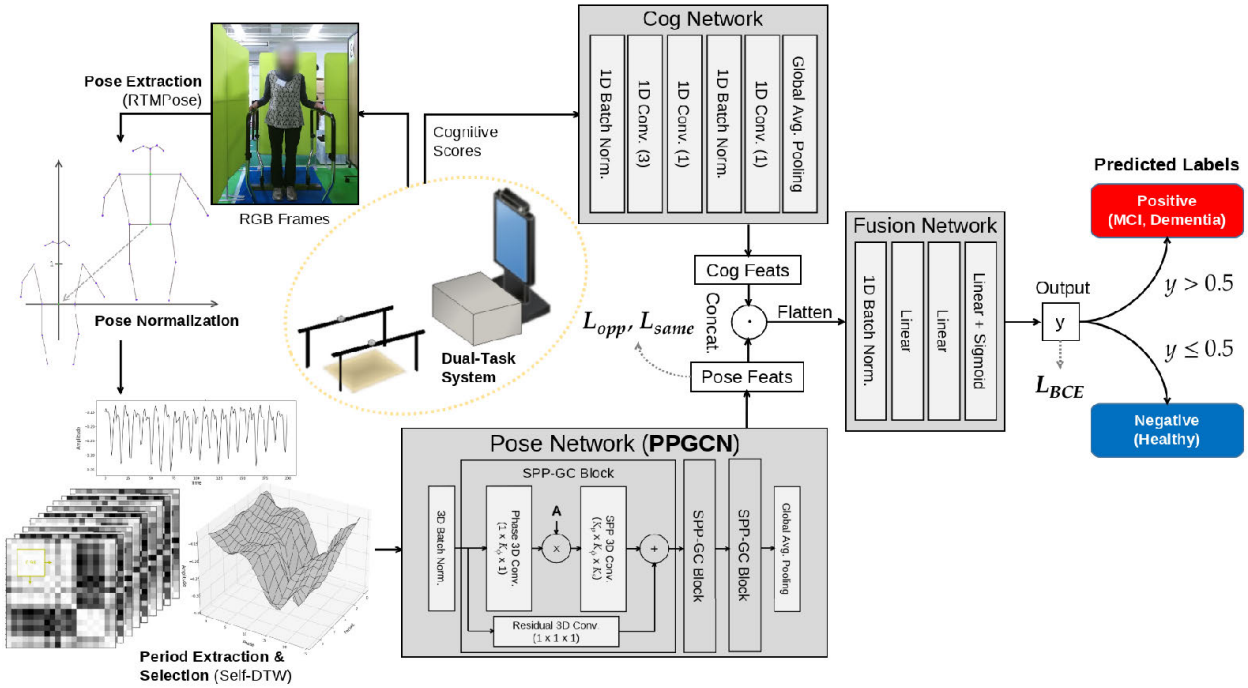


FIGURE 2. The dual-task-based prediction pipeline predicts whether an individual is healthy or cognitively impaired. The inputs are phase-aligned periodic skeleton sequences extracted from RGB images and mental arithmetic scores recorded during a dual-task assessment. The pipeline is separated into modules based on modalities. The Pose Network processes the skeleton sequence, the Cog Network processes the cognitive task scores, and the Fusion Network creates the final output from the combined Pose and Cog Network feature outputs.

In this application, it has been recognized [15] that the human skeleton can be represented by a graph in which the joints are nodes and the bones are edges of the graph. The node values of such a graph are the coordinates of each joint of the skeleton. Because the skeleton pose evolves over time as the subject moves, stacking each skeleton graph in a temporal skeleton sequence results in a spatio-temporal graph representation of motion. In this temporal graph, another dimension of adjacency exists, where each joint or node is connected to the same node in the previous and next spatial graph.

The spatio-temporal GCN (STGCN) [30] effectively handles the spatio-temporal adjacencies of such sequences. Originally designed for traffic forecasting, the generalizability of graph-based methods also allows for its application to our problem setting. The STGCN operates by alternating the graph convolution operation (1) between spatial and temporal dimensions. Wu et al. [15] and Liu et al. [16] applied the STGCN to dual-task-based cognitive decline prediction and achieved high performance.

C. SKELETON-BASED ACTION RECOGNITION

Skeleton-based action recognition methods are related to our study because one of the modalities used to predict cognitive impairment is skeleton sequence information. Skeleton sequences are a strong feature for action recognition [31] that is leveraged by machine learning methods such as naive Bayes classifiers [32] and support vector machines [33], but have become popular with the advent of deep learning graph convolution-based methods.

GCNs such as the aforementioned STGCN are a fit for skeleton-based action recognition because human skeletons have a well-defined graph structure. The STGCN is the progenitor of various architectures and pipelines that perform skeleton-based action recognition.

To handle the long-term dependencies of periods using convolutional networks, they require large kernel sizes, which sacrifice the ability to process short-term features. Although AM-STGCN [34] combines attention models [35] with graph convolutions to also capture global information, it cannot process the granular interperiod (in)consistencies of the skeleton sequence.

We propose an architecture based on the STGCN that is capable of processing an additional period dimension of phase-period decomposed skeleton sequences rather than the entire time series to better focus on differences between periods.

D. DUAL-TASK-BASED COGNITIVE DECLINE DETECTION

Dual-task assessment yields a wealth of data and may be performed sufficiently regularly to collect a dataset suitable for analysis using machine learning methods. Various machine learning approaches, such as support vector machines [36] and clustering algorithms [37], rely on data collected during dual-task assessment [38].

Wu et al. [15] achieved high classification performance using a multi-modal dual-task pipeline, where three-dimensional (3D) pose information was processed with the STGCN alongside cognitive scores, which resulted in 94%

specificity and 89% sensitivity. Liu et al. [16] also used 3D skeleton sequences with the STGCN in a proposed two-stream pipeline and achieved a sensitivity of 96% and specificity of 94%. As shown by the above, using the right approach to analyze the collected information, results rivaling clinical (CT and MRI) diagnoses and paper-based (MMSE) assessments may be achieved.

The above approaches rely on 3D skeleton sequence information extracted from color and depth (RGB+D) images captured by a depth camera, which limits their general usability. We wish to forgo the need for expensive depth-camera hardware and propose a method that works with only color (RGB) images recorded with easily available cameras.

III. PROPOSED METHOD

In this section, we present a novel approach to gait-based dual-task cognitive impairment assessment. We focus on the representative stepping periods of the motor task to improve detection performance. We achieve this by decomposing the temporal dimension of the original skeleton sequences into phase-aligned periods. We present a network architecture that is capable of processing periodic graph information: the PPGCN.

We also describe a dual-task assessment pipeline to demonstrate the use of the PPGCN. We based the pipeline (Fig. 2) on Wu et al.'s previous study [15], with an emphasis on the separation of pipeline modules based on their respective tasks. We designed the pipeline with general usability in mind: it relies on two-dimensional (2D) skeleton data extracted from RGB frames, which can be captured using any camera.

We separate the prediction pipeline into three distinct modules based on their task and modality:

- Pose Network: extracts features from the skeleton sequence inputs;
- Cognitive (Cog) Network: processes and extracts features from the cognitive task scores; and
- Fusion Network: combines the outputs of the Pose and Cog Networks, and renders the final output of the pipeline: the predicted label.

We use the proposed PPGCN architecture as the Pose Network in the above pipeline. Next, we outline the steps required to acquire the skeleton sequence of the stepping motion, extract periods, and process the data to obtain the predicted label.

A. 2D SKELETON EXTRACTION AND PRE-PROCESSING

Our dual-task assessment system [14] records frontal color (RGB) video of the stepping task and logs mental arithmetic scores while subjects perform multiple consecutive trials of the dual-task assessment. We process the recorded RGB frames using RTMPose [39] to extract 2D skeleton sequences (Fig. 2).

For each trial, we extract skeleton sequences from 10 seconds of single-task (stepping without mental arithmetic) and 20 seconds of dual-task (stepping while performing mental arithmetic) videos. We resample the recordings to 10 frames per second, which results in 100 single-task frames and 200 dual-task frames from which to extract skeletons.

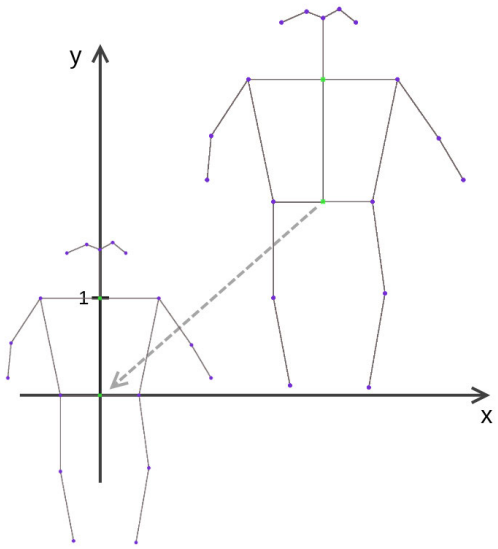


FIGURE 3. The extracted skeletons are normalized by translating and scaling the skeleton so that the hip joint is the origin and the spine is the unit length.

The relative scale and position of the skeletons can encode the location of the assessment. The extracted skeletons' scale and position do not vary significantly at a given location because of the camera setup and the subjects' position relative to the camera being fixed, but do vary between locations. This is an issue because, for example, some locations are biased toward cognitively impaired subjects (e.g., nursing homes and hospitals); hence, it can be seen that certain scales and positions can be mapped to certain distributions of labels. To remove any bias from the scale and position, we normalize the skeleton sequences.

We calculate hip and neck joints, which are not part of the 17-joint skeleton extracted by RTMPose, based on the mean positions of the left/right shoulder and pelvis joints, respectively. We scale and translate the skeletons so that the spine, that is, the bone between the calculated hip and neck joints, is unit length, and the hip joint lies at the origin (Fig. 3).

B. PERIOD EXTRACTION AND SELECTION

A trivial approach to interperiod convolution without sampling the intermediate points would be to include dilated temporal convolutions with a period-length dilation parameter. This is not feasible because period lengths vary as a result of the quasi-periodic nature of the data, and the dilation parameter needs to be constant. Instead, our approach is to decompose the time dimension of the temporally quasi-periodic skeleton sequence into periods and phases, then align and uniformly sample the phase dimension to overcome the issue of variable temporal period lengths.

We rely on Makihara et al.'s Self-DTW [19] framework to decompose the temporal dimension of the skeleton sequence into phase and period dimensions (Fig. 1). Self-DTW aligns

the phases of the extracted periods and samples the phase dimension at a number of points N_{phase}. The result is a list of extracted joint periods P^{joint} of length $M^{joint}_{periods}$. P^{joint}_i is the i-th extracted period of a given joint, which consists of N_{phase} samples of the joint's motion during its P^{joint}_i motion period.

Because neural networks require a fixed input shape, we also have to determine the number of periods to input $N_{periods}$ (Algorithm 1). The number of extracted periods $M^{joint}_{periods}$ present in the time-series data and extracted by Self-DTW differ between the original temporal skeleton sequences; hence, we must devise an approach to select $N_{periods}$ from the extracted $M^{joint}_{periods}$.

Our approach is to select the most coherent $N_{periods}$ long contiguous sequence of periods for all joints. $M^{joint}_{periods}$ is a per joint quantity because some joints move faster than others. If $N_{periods} > M^{joint}_{periods}$, fewer periods are extracted than the desired $N_{periods}$. In this case, the extracted periods are repeated until $N_{periods} = M^{joint}_{periods}$.

Then, we calculate a pairwise absolute correlation matrix for the x and y dimensions of all joint periods: $C^{joint_x}, C^{joint_y} \in [0,1]^{M^{joint}_{periods} \times M^{joint}_{periods}}$. Each element is the absolute Pearson correlation coefficient (2) defined between a given joint's i-th and j-th periods $(P^{joint}_i, P^{joint}_j)$ as

$$C^{joint}_{i,j} = \left| \frac{cov(P^{joint}_i, P^{joint}_j)}{\sigma_{P^{joint}_i} \sigma_{P^{joint}_j}} \right| \tag{2}$$

where $cov(P^{joint}_i, P^{joint}_j)$ is the covariance, and $\sigma_{P^{joint}_i}$ and $\sigma_{P^{joint}_j}$ are the standard deviations of the i-th and j-th periods of a given joint, respectively.

To determine the starting index of k of the most coherent run of $N_{periods}$ gait cycles from the extracted $M^{joint}_{periods}$ periods, we stack and sample all C^{joint} matrices in $N_{periods} \times N_{periods}$-sized sliding windows (Fig. 4). We use the gait periods in the sliding window with the maximal absolute correlation (3) at index k as inputs to the prediction pipeline:

$$k = \arg\max_x \Big(\sum_{i=x}^{x+N_{periods}} \sum_{j=y}^{y+N_{periods}} \sum_{joint=1}^{N_{joints}} C^{joint}_{i,j} \Big) \tag{3}$$

where $x \in [1, M^{joint}_{periods} - N_{periods}]$, $y \in [1, M^{joints}_{periods} - N_{periods}]$

We use the periods $\{P_k, P_{k+1}, \ldots, P_{k+N_{periods}-1}\}$ as inputs to the Pose Network for all joints. The single index convention works because the C^{joint} matrices are symmetric. The Appendix provides a pseudocode representation of the process.

C. PPGCN ARCHITECTURE

We propose the PPGCN architecture (Fig. 6) as the Pose Network module, which is an upgrade of the STGCN [30] that is capable of processing the decomposed phase-period information output by Self-DTW as inputs.

The input data are phase-aligned periodic skeleton sequences. Both the PPGCN and Self-DTW are agnostic to the spatial dimensionality (N_{dims}) of the data and function both with 2D and 3D inputs. Furthermore, if available, multiple skeleton sequences from multiple dual-task trials (N_{trials}) may be stacked to provide more information, which results in an input shape of $(N_{trials} \times N_{dims}) \times N_{periods} \times N_{phase} \times N_{joints}$. We split the multi-trial input into N_{trials} trials and feed them consecutively trial-by-trial into the PPGCN. The PPGCN uses 3D convolutions to process the period, phase, and spatial (joint) dimension (Fig. 5) of the per-trial data of shape $N_{dims} \times N_{periods} \times N_{phase} \times N_{joints}$.

After we apply 3D batch normalization, we input the per-trial data into a stack of three of our novel space-phase-period graph convolution (SPP-GC) blocks. First, each SPP-GC block contains a 3D phase convolution layer with a kernel size of $1 \times K_\phi \times 1$ to allow the preprocessing of the phase dimension independently of other joints and periods.

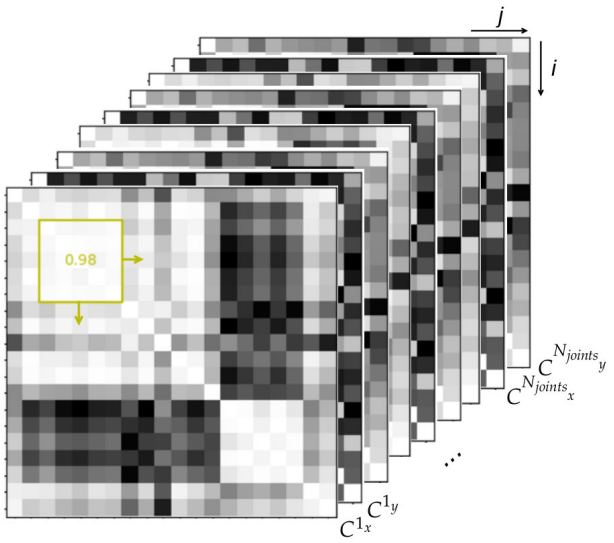


FIGURE 4. The stacked C^{joint} correlation matrices are used in conjunction with a sliding window-based search (in green) to find the most coherent long run of periods: $N_{periods} = 5$ in this example.

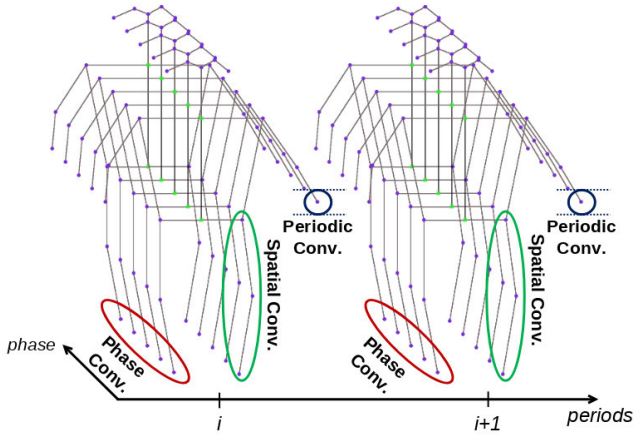


FIGURE 5. The PPGCN performs various graph convolution operations on the phase-aligned periodic skeleton sequence. Spatial convolutions are performed across joints within the same phase-aligned frame. Phase convolutions are calculated along the phase dimension of a given joint within the same period. Periodic convolutions are along the period dimension of a given joint and phase-sampled frame.

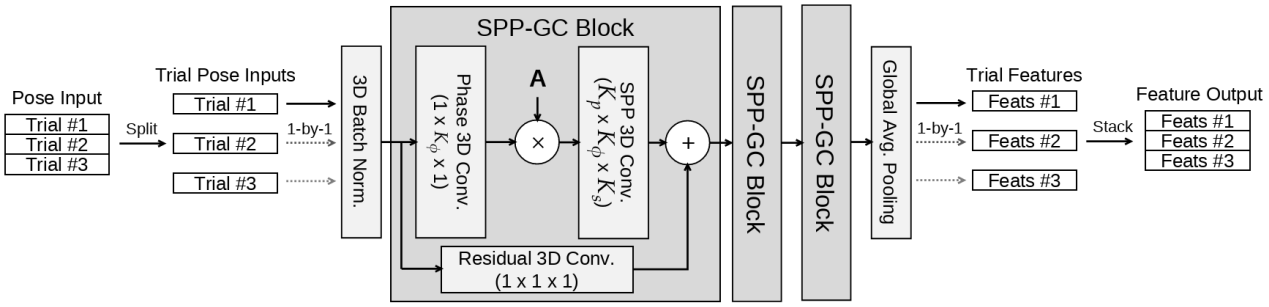


FIGURE 6. The proposed PPGCN architecture consists of three space-phase-period graph convolution (SPP-GC) blocks. SPP-GC blocks perform a phase-only convolution, followed by a simultaneous space, phase, and period convolution. Adherence to the graph structure is enforced by multiplying the phase convolution output with the graph's adjacency matrix (*A*). Phase-aligned skeleton sequence recordings from multiple trials may be stacked as inputs, if available, which the PPGCN processes individually, and their output features are stacked.

Before further processing, to enforce the graph connectivity, we multiply the feature map by the adjacency matrix of graph A.

This is followed by a 3D spatio-temporal-periodic convolution layer simultaneously processing the data over the space, phase, and period dimensions with a kernel size of $K_p \times K_\phi \times K_s$. Finally, there is a residual connection to the input of the SPP-GC block.

The last SPP-GC block is followed by an average pooling layer to produce a 64-long feature vector for each trial. Then we stack the per-trial features to create an $N_{trials} \times 64$ shaped output.

D. CROSS-MODALITY FEATURE FUSION

The other modality in our proposed pipeline is the cognitive task scores obtained from the subjects' performance in the mental arithmetic quiz. We log four data points, the average response time, and accuracy for both the single- and dual-task cases. These form the $N_{trials} \times 4$-shaped inputs to the Cog Network.

In their studies, Wu et al. [15] and Liu et al. [16] opted for 'late-fusion' in which they reduced Pose and Cog Network features to logits as if each had its own predicted label outputs. They processed logits via a stack of fully connected layers to obtain another set of logits that corresponded to the final output. We propose not reducing the Pose and Cog Features to logits because there is no benefit to losing the modality feature information downstream and there is also no supervision for the individual module predictions.

The Cog Network in our pipeline (Fig. 2) is similar to that in Wu et al.'s pipeline [15], with the reduction to logits removed. It consists of a one-dimensional (1D) batch normalization layer and 1D convolution layer with a kernel size of 3, followed by two 1D depth convolutions with another 1D batch normalization in between, and finally, a global average pooling layer. The Cog Feature output is a 160-long feature vector concatenated with the Pose Network's output features to form a combined, cross-modality feature vector.

We input the combined feature vectors into a feature fusion network that consists of a batch normalization layer followed by a sequence of fully connected layers and sigmoid nonlinearity. The output is the predicted probability of the input data having a positive label (subject has MCI or dementia), $y \in [0, 1]$, which is supervised via a binary cross entropy loss L_{BCE}.

E. PAIRWISE FEATURE DISTANCE LOSS

We leverage the availability of per-trial pose features that results from the use of cross-modality feature fusion to introduce a two-component unsupervised feature-level loss. Prior works have shown [40], [41] that high interclass and low intraclass variance of features is beneficial for classification problems. Our main goal is to encourage the Pose Network to produce similar features for inputs with the same label (decreasing intraclass variance) and different features for inputs with opposite labels (increasing interclass variance). Although the networks attempt to achieve this through the task loss L_{BCE}, the effect may be magnified by directly encouraging the networks to separate feature distributions through a feature distance loss.

When the feature distributions of classes are distant, it is easier for a classifier to create decision manifolds with wider margins between the clusters of feature distributions. By having a wide margin between feature distributions, the generalization performance, both between the training and validation datasets, and for possible future out-of-dataset samples, is expected to increase.

The first component L_{same} minimizes the pairwise cosine distance of features that have the same label. We separate each batch of Pose Features F into a positive F^+ and negative F^- sub-batch based on the corresponding ground truth labels. Then we calculate the positive (4) and negative (5) sub-batch losses, L^+_{same} and L^-_{same}, respectively:

$$L^+_{same} = \sum_i \sum_j 1 - \frac{\mathbf{F_i^+ F_j^+}}{\|\mathbf{F_i^+}\| \|\mathbf{F_j^+}\|} \quad \text{where } i, j \in [0, \|F^+\|] \text{ and } i \neq j \tag{4}$$

$$L^-_{same} = \sum_i \sum_j 1 - \frac{\mathbf{F_i^- F_j^-}}{\|\mathbf{F_i^-}\| \|\mathbf{F_j^-}\|} \quad \text{where } i, j \in [0, \|F^-\|] \text{ and } i \neq j \tag{5}$$

TABLE 1. Distribution of samples in the dual-task assessment dataset used to train and validate the proposed method.

Class	Label	Sex	# of Samples	Age Range
MCI, Dementia	Positive	Male Female	50 93	53 to 90 53 to 91
Healthy	Negative	Male Female	107 142	70 to 85 70 to 87

Finally, L_{same} (6) is the weighted sum of the positive and negative components:

$$L_{same} = \frac{1}{\|F^+\|(\|F^+\|-1)} L^+_{same} + \frac{1}{\|F^-\|(\|F^-\|-1)} L^-_{same} \tag{6}$$

The second component L_{opp} minimizes the pairwise cosine similarity (7) for features that have opposite labels:

$$L_{opp} = \frac{1}{\|\mathbf{F}^+\|\|\mathbf{F}^-\|} \sum_i \sum_j \frac{\mathbf{F}^+_\mathbf{i} \mathbf{F}^-_\mathbf{j}}{\|\mathbf{F}^+_\mathbf{i}\|\|\mathbf{F}^-_\mathbf{j}\|} \tag{7}$$

where $i \in [1, \|F^+\|], j \in [1, \|F^-\|]$

We save and reuse features from the $(n-1)$-th batch for the loss calculation in the case in which the n-th batch does not contain either class, which ensures that the loss is always active, even in the case of severe dataset imbalance. Furthermore, we ensure that the initial batch contains at least one positive and negative sample so that there is always a F^+ and F^- saved to avoid errors.

The combined loss function (8) of the network is

$$L = \frac{(\alpha+\beta)L_{BCE} + \alpha L_{same} + \beta L_{opp}}{2\alpha + 2\beta} \tag{8}$$

where α and β are the coefficients of L_{same} and L_{opp}, respectively. By weighting the task loss L_{BCE} with $(\alpha+\beta)$, we ensure that the classification task remains the most impactful of the three, while the $\frac{1}{2\alpha+2\beta}$ term normalizes the total loss magnitude.

IV. EXPERIMENTS AND RESULTS

A. DATASET

To train and validate our proposed method, we used a dataset (Table 1) collected using our dual-task system [14]. The dataset samples were annotated by geriatric psychiatrists at Osaka University Hospital based on subjects' diagnoses. The subtypes of dementia were limited to Alzheimer's dementia with Lewy bodies. Our dataset contained 392 samples of dual-task assessment data. Each sample referred to the three dual-task assessments (recorded consecutively) of skeleton sequences and mental arithmetic scores of a subject. Informed consent was obtained from all subjects prior to the collection of data.

A total of 143 samples were labeled as positive (diagnosed with MCI or dementia), of which 93 were from female subjects aged 53-91 and 50 from male subjects aged 53-90. From the 249 healthy samples, 142 were from females aged 70-87 and 107 from males aged 70-85. We split the dataset into four balanced folds to perform four-fold cross-validation (Table 2). As some subjects contributed multiple samples to the dataset, we took care to ensure that the folds were split over subjects and there were no samples from the same subject in different folds.

TABLE 2. Distribution of samples in the four-fold split of the dataset.

Class	Fold 1	Fold 2	Fold 3	Fold 4
Healthy	62	63	62	62
MCI	14	14	14	15
Dementia	22	21	22	21

TABLE 3. Network and training hyperparameters obtained after hyperparameter optimization.

Network Parameters	Value(s)	Training Parameters	Value(s)
PPGCN Conv. Channels	32, 56, 36	Init. LR	0.00025
Cog Net Conv. Channels	16, 12, 12	Weight Decay	0.0526
Fusion Net Layer Sizes	192, 416	PPGCN Dropout	0.0493
Spatial Kernel Size (K_s)	3	Fusion Net. Dropout	0.0392
Phase Kernel Size (K_p)	3	α	9.0233
Period Kernel Size (K_ϕ)	9	β	8.7723

B. FOUR-FOLD CROSS VALIDATION

In our experiments, we used the 17 skeleton joints extracted by RTMPose [39] from three consecutive dual-task experiments per subject: $N_{dims} = 2$ and $N_{joints} = 17$. Following our previous study [15], we performed three trials per subject to improve stability: $N_{trials} = 3$. As the average $M_{periods}$, that is, the number of periods from the skeleton sequence extracted by Self-DTW, was 10 for single-task and 20 for dual-task skeleton sequences, we chose their minimum value to use as the number of input periods: $N_{periods} = 10$. This used the most available information while minimizing redundancy. Self-DTW sampled the phase-aligned periods in 25 points. We concatenated single- and dual-task phase-aligned periodic skeleton sequences along the phase dimension, which resulted in $N_{phase} = 50$.

We determined the PPGCN and loss hyperparameters by performing hyperparameter search using the Optuna hyperparameter optimization framework [42]. The selected hyperparameters are shown in Table 3.

The performance evaluation metrics are the balanced accuracy (Acc.), sensitivity (Sens., a.k.a. true positive rate), and specificity (Spec., a.k.a. true negative rate) scores. We consider the sum of sensitivity and specificity (Sens. + Spec.) to be the most important descriptor of performance for the task.

Although we performed training and validation in batches, we calculated the metrics for each fold by collecting all predicted labels and then comparing the results with the expected labels of the respective fold. Similarly, we evaluated the combined metrics by aggregating all the above predicted labels and comparing them with all expected labels in the dataset. This ensured that no errors were present as a result of averaging per-batch values when metrics were calculated.

We obtained results after training the models for 100 epochs with the parameters in Table 3. The four-fold cross-validation results (Table 4) and ROC curves in Fig. 8 show balanced performance across all four folds.

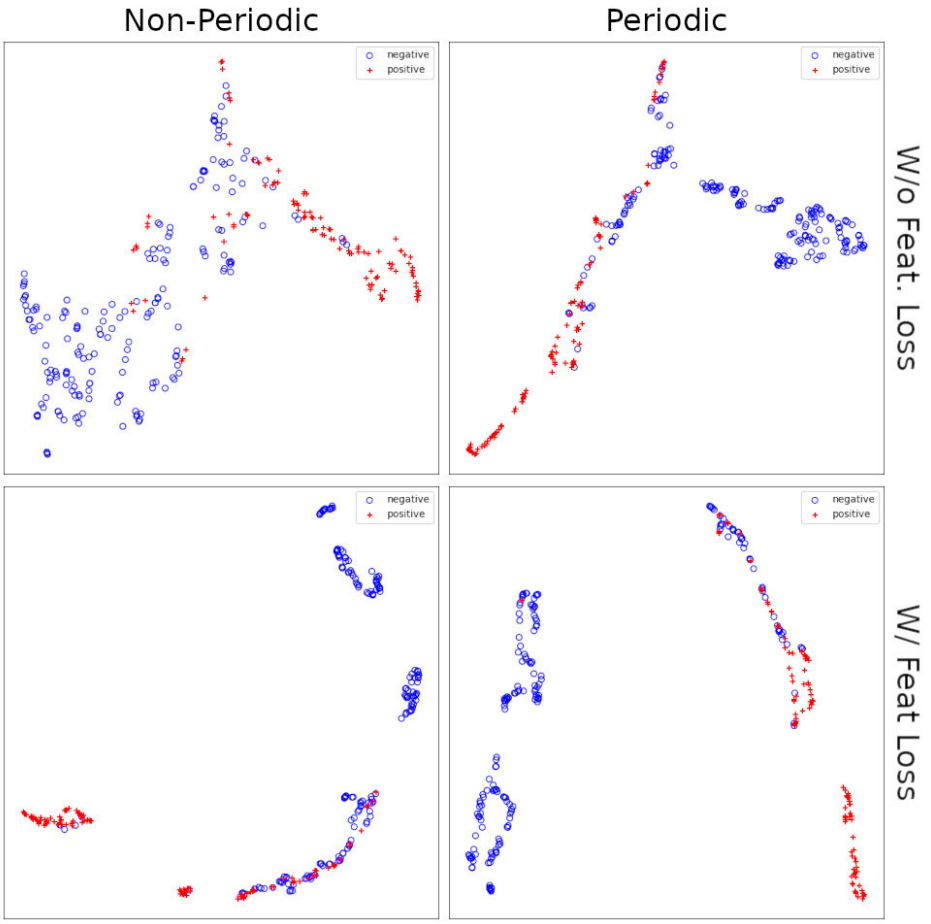


FIGURE 7. UMAP visualization of the embedded trial features without (top) and with (bottom) the Feature Loss active. When the Feature Loss is active, features corresponding to positive labels (red +) are better separated from those with negative labels (blue ∘). The effect is observable both in the non-periodic case for the STGCN (left), and also for our proposed PPGCN (right).

TABLE 4. Four-fold cross-validation results ($N_{trials} = 3$). Combined metrics were obtained by calculating them over the aggregated fold-wise predictions.

Fold #	Acc.	Sens.	Spec.	Sens. + Spec.
Fold 1	0.9364	0.8889	0.9839	1.8728
Fold 2	0.8921	0.9429	0.8413	1.7842
Fold 3	0.9283	0.8889	0.9677	1.8566
Fold 4	0.9699	0.9722	0.9677	1.9399
Combined	0.9315	0.9231	0.9398	1.8629

C. ABLATION STUDIES

The modular approach to the prediction pipeline allows for the easy swapping of modules to observe the individual impact of the proposed enhancements. In ablation studies (Table 5), we compared the performance impact of the following combinations: swapping the PPGCN to STGCN, applying logit fusion instead of feature fusion, and turning the proposed Feature Loss on or off. When using the STGCN, we used the non-periodic skeleton sequences as inputs. For the PPGCN, we used the phase-aligned periodic skeleton sequences.

To observe the effects of logit fusion, we reduced the output features of the STGCN, PPGCN, and Cog Network to logits.

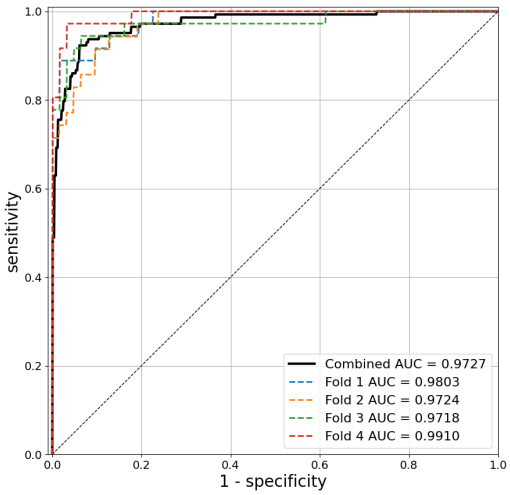


FIGURE 8. Four-fold cross-validation and combined ROC-curves ($N_{trials} = 3$).

The STGCN with the logit fusion configuration is identical to Wu et al.'s [15] model. For feature fusion, the pipeline is presented in Fig. 2, with the networks outputting their extracted feature vectors.

TABLE 5. Mean Sens. + Spec. comparison with individual enhancements turned on and off ($N_{trials} = 3$).

	Non-Periodic (STGCN)	**Periodic (PPGCN)**
Logits Fusion (Baseline)	1.6596	1.7358
+ Feature Fusion	1.6977	1.8249
+ Feature Loss	1.7077	**1.8629**

TABLE 6. Mean performance comparison of our method with existing methods ($N_{trials} = 1$).

Method	Acc.	Sens.	Spec.	Sens. + Spec.
Wu et al. [15]	0.8246	0.9021	0.7470	1.6491
Liu et al. [16]	0.8800	0.8600	0.9000	1.7600
PPGCN (Ours)	**0.9178**	**0.9441**	**0.8916**	**1.8356**

For testing the proposed Feature Distance Loss, we either trained networks only with the task loss L_{BCE} (without applying L_{same} or L_{opp}) or with the full, combined loss function L (including both L_{same} and L_{opp}). The PPGCN with feature fusion and Feature Loss enabled represents the full proposed method in this study.

We observed a clear upgrade in performance when we used our proposed PPGCN versus the STGCN (Table 5). Applying feature fusion instead of the original logit fusion improved performance for both setups, and when training with our proposed Feature Loss, we even observed further performance improvement, although this effect was diminished using the STGCN.

The effect of the Feature Loss was illustrated well when we visualized the embedded features via Uniform Manifold Approximation and Projection (UMAP) [43] (Fig. 7). When the loss was active, the features of subjects with positive labels were better separated from those with negative labels. These observations are in line with the measured results: the effect of the feature separation loss enhanced the performance of both the PPGCN and STGCN (Table 5).

D. COMPARISON OF OUR METHOD WITH EXISTING METHODS

To compare and contrast the results of the proposed PPGCN architecture with other methods, we compared its performance with the pipelines proposed in [15] and [16] trained on our dataset. Because Liu et al.'s method [16] used only a single trial, to enable a closer comparison, we used only the data from the first of three dual-task trials for training and validation ($N_{trials} = 1$).

Comparing the mean results of four-fold cross-validation (Table 6), we observed that the periodic upgrade to the STGCN represented the most significant jump in performance, which was consistent with our observation during the ablation studies. Furthermore, the introduction of feature level fusion and then the feature loss both had a positive effect on performance, even in the $N_{trials} = 1$ case.

Although the final performance of the PPGCN with feature level fusion and the feature loss was slightly lower (1.8628 vs. 1.8356 mean Sens. + Spec.) than in the original $N_{trials} = 3$ case, it still outperformed methods in previous studies.

We observed a drop in the performance of existing methods. This is explained by multiple factors: first, the dataset differed between the existing methods and ours. Both Wu et al. and Liu et al. relied on 3D skeleton sequences, whereas our dataset contained 2D skeleton sequences. Furthermore, our model selection criterion was stricter: we reported the performance of the final training epoch, whereas researchers reported the best validation set performance in their respective studies. This favors methods that stably converge and are not prone to overfitting.

Despite this, our proposed method approached Liu et al.'s originally reported performance of 1.90 Sens. + Spec., which they achieved with 3D skeleton sequences. This shows that stepping-based dual-task cognitive impairment with easily available RGB cameras is not only possible but can also match the performance of methods that rely on expensive, more difficult-to-acquire depth cameras, which increases the general applicability of the paradigm.

E. DISCUSSION

The main application of our proposed method is to assist in predicting diagnoses based on clinical criteria [2]. Most importantly, out of the two metrics "sensitivity" and "specificity", increasing sensitivity is particularly important. This is because high sensitivity reduces false negatives, which is critical when selecting subjects to dedicate further attention to.

A possible approach to increase sensitivity or decrease the imbalance between sensitivity and specificity is to assign a weight to the loss function based on the ratio of positive to negative samples, either in the current mini-batch or the dataset. Various other task loss functions may also be trialed to directly target increasing specificity instead of the binary cross entropy loss used in this study.

Unfortunately, the above approaches may only go so far when the dataset is heavily imbalanced and has a limited number of samples, as in our case. Increasing the dataset using ongoing data collection and reliably annotating the samples would go a long way in increasing performance and generalization capability.

The current dataset is imbalanced (Tables 1 and 2) even for binary classification (2:1 ratio of healthy vs. unhealthy samples). Separating into the dataset three classes would further exacerbate the imbalance (6:1:2 ratio of healthy vs. MCI vs. Dementia) which renders a 3-way classification experiment infeasible. Once sufficient data samples are collected, the task may also be extended to a three-way classification problem (healthy vs. MCI vs. dementia).

Furthermore, although we focused on using phase-aligned periodic skeleton sequences and their processing with convolutional networks in this study, various possible approaches exist for upgrading the pipeline, for example, by modifying the Cog Network or Fusion Network.

V. CONCLUSION

In this study, we proposed the PPGCN architecture as the Pose Network in a dual-task-based cognitive impairment detection pipeline. We modified the STGCN [30] architecture using our novel SPP-GC blocks to support periodic data. We acquired phase-aligned periods using Self-DTW [19] from quasi-periodic 2D skeleton sequences extracted by RTMPose [39] from RGB images.

By focusing on representative motion periods rather than the entire skeleton sequence in conjunction with cross-modality feature fusion and an unsupervised pose feature loss, we achieved a mean 1.8628 sensitivity + specificity in a four-fold cross-validation setup, which outperformed the results in previous studies.

APPENDIX

We provide a pseudocode representation (Algorithm 1) of the period extraction and selection process described in Section III-B, Period Extraction & Selection.

Algorithm 1 Period Extraction and Selection

$pose \leftarrow$ Normalized pose extracted from RGB video
$N_{periods} \leftarrow$ Num. of periods to select

procedure PeriodSelect($pose, N_{periods}$)
  $P := SelfDTW(pose)$ ▷ *Extract periods*

▷ *Create correlation matrices for periods*
  $C \leftarrow$ Corr. matrix list of all joints, coordinates
  for all $P^{joint} \in P$ **do** ▷ *For both x, y coordinates*
    $M^{joint}_{periods} \leftarrow$ Num. of periods from SelfDTW
    $C^{joint} \leftarrow$ Correlation matrix of joint
    if $size(P^{joint}) < N_{periods}$ **then**
      $P^{joint} := repeat(P^{joint})$ ▷ *Repeat until size OK*
    for $i := 0 to M^{joint}_{periods}$ **do**
      for $j := 0 to M^{joint}_{periods}$ **do**
        $C^{joint}[i, j] := |corr(P^{joint}[i], P^{joint}[j])|$
    $C.push(C^{joint})$

▷ *Search for the best starting index of $N_{periods}$ long window*
  $corr_{max} := 0$ ▷ *Best correlation*
  $k := 0$ ▷ *Starting idx for $N_{periods}$*
  for $x := 0 to size(C)[1] - N_{periods}$ **do**
    for $y := 0 to size(C)[2] - N_{periods}$ **do**
      $corr := sum(C[:, x : x + N_{periods}, y : y + N_{periods}])$
      if $corr > corr_{max}$ **then**
        $corr_{max} := corr$
        $k := x$ ▷ *x or y, because C is symmetric*
  return $P[k : k + N_{periods}]$

ACKNOWLEDGMENT

The authors thank Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

REFERENCES

[1] WHO. (Mar. 2023). *Dementia*. Accessed: Oct. 17, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder, M. C. Carrillo, B. Thies, and C. H. Phelps, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimers. Dement.*, vol. 7, pp. 270–279, May 2011.

[3] L. Chouliaras and J. T. O'Brien, "The use of neuroimaging techniques in the early and differential diagnosis of dementia," *Mol. Psychiatry*, vol. 28, no. 10, pp. 4084–4097, Aug. 2023.

[4] E. C. W. van Straaten, P. Scheltens, and F. Barkhof, "MRI and CT in the diagnosis of vascular dementia," *J. Neurological Sci.*, vol. 226, nos. 1–2, pp. 9–12, Nov. 2004.

[5] P. Vemuri, G. Simon, K. Kantarci, J. L. Whitwell, M. L. Senjem, S. A. Przybelski, J. L. Gunter, K. A. Josephs, D. S. Knopman, B. F. Boeve, T. J. Ferman, D. W. Dickson, J. E. Parisi, R. C. Petersen, and C. R. Jack, "Antemortem differential diagnosis of dementia pathology using structural MRI: Differential-STAND," *NeuroImage*, vol. 55, no. 2, pp. 522–531, Mar. 2011.

[6] W. A. Edelstein, M. Mahesh, and J. A. Carrino, "MRI: Time is dose—And money and versatility," *J. Amer. College Radiol.*, vol. 7, no. 8, pp. 650–652, Aug. 2010.

[7] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: A comprehensive review," *J. Amer. Geriatr. Soc.*, vol. 40, pp. 922–935, Sep. 1992.

[8] A. Morita, R. O'Caoimh, H. Murayama, D. Molloy, S. Inoue, Y. Shobugawa, and T. Fujiwara, "Validity of the Japanese version of the quick mild cognitive impairment screen," *Int. J. Environ. Res. Public Health*, vol. 16, no. 6, p. 917, Mar. 2019.

[9] C. Carnero-Pardo, "Should the mini-mental state examination be retired?" *Neurología English Ed.*, vol. 29, pp. 473–481, Oct. 2014.

[10] R. Morris, S. Lord, J. Bunce, D. Burn, and L. Rochester, "Gait and cognition: Mapping the global and discrete relationships in ageing and neurodegenerative disease," *Neurosci. Biobehavioral Rev.*, vol. 64, pp. 326–345, May 2016.

[11] A. Bishnoi and M. E. Hernandez, "Dual task walking costs in older adults with mild cognitive impairment: A systematic review and meta-analysis," *Aging Mental Health*, vol. 25, no. 9, pp. 1618–1629, Sep. 2021.

[12] H. B. Åhman, Y. Cedervall, L. Kilander, V. Giedraitis, L. Berglund, K. J. McKee, E. Rosendahl, M. Ingelsson, and A. C. Åberg, "Dual-task tests discriminate between dementia, mild cognitive impairment, subjective cognitive impairment, and healthy controls—A cross-sectional cohort study," *BMC Geriatrics*, vol. 20, no. 1, pp. 1–10, Dec. 2020.

[13] Y. Wang, Q. Yang, C. Tian, J. Zeng, M. Yang, J. Li, and J. Mao, "A dual-task gait test detects mild cognitive impairment with a specificity of 91.2%," *Frontiers Neurosci.*, vol. 16, Feb. 2023, Art. no. 1100642.

[14] T. Matsuura, K. Sakashita, A. Grushnikov, F. Okura, I. Mitsugami, and Y. Yagi, "Statistical analysis of dual-task gait characteristics for cognitive score estimation," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Dec. 2019.

[15] S. Wu, F. Okura, Y. Makihara, K. Aoki, M. Niwa, and Y. Yagi, "Early detection of low cognitive scores from dual-task performance data using a spatio-temporal graph convolutional neural network," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1895–1901.

[16] J. Liu, S. Wu, F. Okura, Y. Makihara, and Y. Yagi, "Two-stream graph convolutional networks with task-specific loss for dual-task gait analysis," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 24–27.

[17] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*. Berlin, Germany: Springer, 2007, pp. 69–84.

[18] Z. Zhou, R. I. Damper, and A. Prugel-Bennett, "Model selection within a Bayesian approach to extraction of Walker motion," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2006, pp. 17–22.

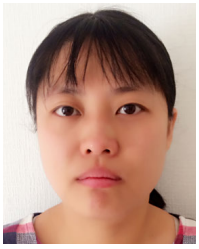
[19] Y. Makihara, M. R. Aqmar, N. T. Trung, H. Nagahara, R. Sagawa, Y. Mukaigawa, and Y. Yagi, "Phase estimation of a single quasi-periodic signal," *IEEE Trans. Signal Process.*, vol. 62, no. 8, pp. 2066–2079, Apr. 2014.

[20] J. Wu, "Introduction to convolutional neural networks," *Nat. Key Lab Novel Softw. Technol.*, vol. 5, no. 23, p. 495, 2017.

[21] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design Appl.*, vol. 5, nos. 64–67, p. 2, 2001.
[22] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: A comprehensive review," *Comput. Social Netw.*, vol. 6, no. 1, pp. 1–23, Dec. 2019.
[23] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl
[24] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 2014–2023.
[25] H. Gao, Z. Wang, and S. Ji, "Large-scale learnable graph convolutional networks," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Jul. 2018, pp. 1416–1424.
[26] U. A. Bhatti, H. Tang, G. Wu, S. Marjan, and A. Hussain, "Deep learning with graph convolutional networks: An overview and latest applications in computational intelligence," *Int. J. Intell. Syst.*, vol. 2023, pp. 1–28, Feb. 2023.
[27] Z. Yan, J. Ge, Y. Wu, L. Li, and T. Li, "Automatic virtual network embedding: A deep reinforcement learning approach with graph convolutional networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 6, pp. 1040–1057, Jun. 2020.
[28] X. Song, H. Li, W. Gao, Y. Chen, T. Wang, G. Ma, and B. Lei, "Augmented multicenter graph convolutional network for COVID-19 diagnosis," *IEEE Trans. Ind. Informat.*, vol. 17, no. 9, pp. 6499–6509, Sep. 2021.
[29] Y. Shen, H. Fu, Z. Du, X. Chen, E. Burnaev, D. Zorin, K. Zhou, and Y. Zheng, "GCN-Denoiser: Mesh denoising with graph convolutional networks," *ACM Trans. Graph.*, vol. 41, no. 1, pp. 1–14, Feb. 2022.
[30] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3634–3640.
[31] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. (2013). *Towards Understanding Action Recognition*. Accessed: Oct. 27, 2023. [Online]. Available: https://www.cv-foundation.org/openaccess/content_iccv_2013/html/Jhuang_Towards_Understanding_Action_2013_ICCV_paper.html
[32] X. Yang and Y. L. Tian, "EigenJoints-based action recognition using Naïve-Bayes-nearest-neighbor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 14–19.
[33] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Ambient Assisted Living and Home Care*. Berlin, Germany: Springer, 2012, pp. 216–223.
[34] H. Wu, Y. Han, M. Zhang, B. D. Abebe, M. B. Legesse, and R. Jin, "Identifying unsafe behavior of construction workers: A dynamic approach combining skeleton information and spatiotemporal features," *J. Construct. Eng. Manage.*, vol. 149, no. 11, Nov. 2023, Art. no. 04023115.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[36] L. N. Boettcher, M. Hssayeni, A. Rosenfeld, M. I. Tolea, J. E. Galvin, and B. Ghoraani, "Dual-task gait assessment and machine learning for early-detection of cognitive decline," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 3204–3207.
[37] C. Ricciardi, M. Amboni, C. De Santis, G. Ricciardelli, G. Improta, G. D'Addio, S. Cuoco, M. Picillo, P. Barone, and M. Cesarelli, "Machine learning can detect the presence of mild cognitive impairment in patients affected by Parkinson's disease," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2020, pp. 1–6.
[38] B. Ghoraani, L. N. Boettcher, M. D. Hssayeni, A. Rosenfeld, M. I. Tolea, and J. E. Galvin, "Detection of mild cognitive impairment and Alzheimer's disease using dual-task gait assessments and machine learning," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102249.
[39] (Oct. 2023). *RTMPose: Real-Time Multi-Person Pose Estimation Based on MMPose*. Accessed: Oct. 18, 2023. [Online]. Available: https://github.com/open-mmlab/mmpose/tree/1.x/projects/rtmpose
[40] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.
[41] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[42] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 2623–2631.
[43] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *J. Open Source Softw.*, vol. 3, no. 29, p. 861, Sep. 2018.

ÁKOS GODÓ received the B.Sc. degree in molecular bionics engineering and the M.Sc. degree in info-bionics engineering from Pázmány Péter Catholic University, Budapest, Hungary, in 2016 and 2018, respectively, and the Ph.D. degree from Osaka University, Japan, in 2023.

He is currently with the Institute of Scientific and Industrial Research (SANKEN), Osaka University. His current research topics are dual-task-based cognitive impairment detection and the development of protein structure segmentation algorithms for X-ray diffraction and cryo-EM. His research interests include interdisciplinary research topics inspired by or that assist in biology-related fields, machine learning, neural network architecture design, image processing, and computer vision.



SHUQIONG WU was born in Shanxi, China, in 1985. She received the B.E. and M.E. degrees from Beihang University, Beijing, China, in 2011, and the Ph.D. degree from Tokyo Institute of Technology, Tokyo, Japan, in 2015.

Her major during the doctoral period was computational intelligence and systems science. From 2015 to 2020, she was a Research Fellow with the Graduate School of Informatics, Kyoto University. Since 2020, she has been an Assistant Professor with the Institute of Scientific and Industrial Research (SANKEN), Osaka University. Her current research interests include dual-task-based cognitive impairment detection, cognitive status monitoring, 3D CT reconstruction, biomedical signal processing, image processing, 3D reconstruction, pattern recognition, and machine learning.

FUMIO OKURA (Member, IEEE) received the M.S. and Ph.D. degrees in engineering from Nara Institute of Science and Technology, in 2011 and 2014, respectively.

He was an Assistant Professor with the Institute of Scientific and Industrial Research (SANKEN), Osaka University, until 2020. He is currently an Associate Professor with the Computer Vision Laboratory (Matsushita Lab), Graduate School of Information Science and Technology, Osaka University. His research interests include the boundary domain between computer vision and computer graphics. He is a member of IEICE, IPSJ, and VRSJ.

YASUSHI MAKIHARA received the B.S., M.S., and Ph.D. degrees in engineering from Osaka University, in 2001, 2002, and 2005, respectively.

He is currently a Professor with the Institute of Scientific and Industrial Research (SANKEN), Osaka University. His research interests are computer vision, pattern recognition, and image processing, including gait recognition, pedestrian detection, morphing, and temporal super-resolution. He is a member of IPSJ, IEICE, RSJ, and JSME. He has received several honors and awards, including the Second International Workshop on Biometrics and Forensics (IWBF), in 2014,

the IAPR Best Paper Award, the Ninth IAPR International Conference on Biometrics (ICB), in 2016, and the Honorable Mention Paper Award. He was the Program Co-Chair of the Fourth Asian Conference on Pattern Recognition (ACPR), in 2017.

MANABU IKEDA received the M.D. and Ph.D. degrees from Osaka University, Suita, Japan, in 1988 and 1993, respectively.

He has been the Chair and a Professor with the Department of Psychiatry, Osaka University, since 2016. He has engaged with anthropology, general psychiatry, neuropsychology, neuropathology, and old age psychiatry with Tokyo University, Osaka University, the University of Cambridge, and the Ehime University School of Medicine. He was the Chair and a Professor with the Department of Neuropsychiatry, Faculty of Life Sciences, Kumamoto University, from 2007 to 2016. His current research interests include the neural basis of BPSD, late onset psychosis, and trials for pharmacotherapy and non-pharmacological interventions. He serves as a Board Member for many academic societies, such as the Neuropsychology Association of Japan (President), from 2019 to 2023, Japanese Psychogeriatric Society (President), since 2018, Japanese Society of Psychiatry and Neurology, Asian Society Against Dementia, and the International Psychogeriatric Association (President), since 2021.

SHUNSUKE SATO was born in Tokushima, Japan, in 1987. He received the M.D. degree from Tokushima University, in 2012, and the Ph.D. degree from Osaka University, in 2018. From 2021 to 2023, he was an Assistant Professor with the Department of Psychiatry, Osaka University. Since 2023, he has been a Guest Faculty Member of the Department of Psychiatry, Osaka University, and a Medical Staff with the Esaka Hospital. His research interests include dual-task-based cognitive impairment detection and behavioral symptoms in frontotemporal dementia.

MAKI SUZUKI received the M.S. and Ph.D. degrees in disability medicine from Tohoku University, Japan, in 2001 and 2004, respectively. She was a Postdoctoral Fellow and a Researcher with Tohoku University, Kyoto Sangyo University, Kumamoto University, Kyoto University, Japan, and the University of California at Irvine, USA. Since 2017, she has been an Endowed Chair Associate Professor with the Department of Behavioral Neurology and Neuropsychiatry, United Graduate School of Child Development, Osaka University, and a Neuropsychologist with the Department of Psychiatry, Osaka University Hospital, Japan. She has used neuropsychological assessment and neuroimaging techniques to investigate the cognitive nature of human behavior in both healthy adults and patients with neurodegenerative disease. Her research interest includes brain-behavior relationships in cognitive neuroscience.

YUTO SATAKE was born in Ashiya, Japan, in 1990. He received the M.D. and Ph.D. degrees from Osaka University, Suita, Japan, in 2015 and 2023, respectively. He is currently a Neuropsychiatrist with Osaka University Medical Hospital and a specially appointed Assistant Professor with the Department of Psychiatry, Osaka University. He has published more than 30 articles, mainly related to geriatric psychiatry. His research interests include the etiology and treatment of delusions and hallucinations, neurodegenerative imaging and fluid biomarkers, and loneliness and isolation in older people.

DAIKI TAOMOTO was born in Fukuoka, Japan, in 1990. He received the degree from the Faculty of Medicine, Osaka University, in 2016. He is currently pursuing the degree in neuropsychology with the Graduate School of Medicine, Osaka University. His research interests include frontotemporal dementia, prodromal dementia with Lewy bodies, and hoarding symptoms in older people.

YASUSHI YAGI (Senior Member, IEEE) received the Ph.D. degree from Osaka University, in 1991.

In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he was involved in robotics and inspections. He was a Research Associate with Osaka University, in 1990, where he was a Lecturer, an Associate Professor, and a Professor, in 1993, 1996, and 2003, respectively. He was the Director of the Institute of Scientific and Industrial Research (SANKEN), Osaka University, from 2012 to 2015. He was an Executive Vice President of Osaka University, from 2015 to 2019. His research interests include computer vision, pattern recognition, biometrics, human sensing, medical engineering, and robotics.

Dr. Yagi is a fellow of IPSJ and a member of IEICE and RSJ. He has served as a Board Member of the Asian Federation of Computer Vision Societies (the Vice President, from 2014 to 2022, and the Financial Chair, from 2010 to 2014). Currently, he is a member of the AFCV Administrative Committee. He received the ACM VRST2003 Honorable Mention Award, the IEEE ROBIO2006 Finalist of the T. J. Tan Best Paper in Robotics, the IEEE ICRA2008 Finalist for the Best Vision Paper, the PSIVT2010 Best Paper Award, the MIRU2008 Nagao Award, the IEEE ICCP2013 Honorable Mention Award, the MVA2013 Best Poster Award, the IWBF2014 IAPR Best Paper Award, and the *IPSJ Transactions on Computer Vision and Applications* Outstanding Paper Award, in 2011 and 2013. He served as the Chair for international conferences, including ROBIO2006 (PC), ACCV (2007PC and 2009GC), PSVIT2009 (FC), and ACPR (2011PC, 2013GC, 2021GC, and 2023GC). He has also served as an Editor for the IEEE ICRA Conference Editorial Board, in 2008 and 2011. He is a member of the Editorial Board of the *International Journal of Computer Vision*. He was the Editor-in-Chief of *IPSJ Transactions on Computer Vision and Applications*.