Early Detection of Low Cognitive Scores from Dual-task Performance Data Using a Spatio-temporal Graph Convolutional Neural Network

Shuqiong Wu[1], Fumio Okura[2], Yasushi Makihara[1], Kota Aoki[1], Masataka Niwa[1], and Yasushi Yagi[1]

Abstract— **Detecting low cognitive scores at an early stage is important for delaying the progress of dementia. Investigations of early-stage detection have employed automatic assessment using dual-task (i.e., performing two different tasks simultaneously). However, current approaches to dual-task-based detection are based on either simple features or limited motion information, which degrades the detection accuracy. To address this problem, we proposed a framework that uses graph convolutional networks to extract spatio-temporal features from dual-task performance data. Moreover, to make the proposed method robust against data imbalance, we devised a loss function that directly optimizes the summation of the sensitivity and specificity of the detection of low cognitive scores (i.e., score≤ 23 or score≤ 27). Our evaluation is based on 171 subjects from 6 different senior citizens' facilities. Our experimental results demonstrated that the proposed algorithm considerably outperforms the previous standard with respect to both the sensitivity and specificity of the detection of low cognitive scores.**

I. INTRODUCTION

The number of people with dementia has increased substantially in recent years [1], which has created a severe burden on nursing systems. Dementia is a progressive disease, causing declines in such areas as memory, behavioral ability, and language skills. Although dementia is incurable, early intervention could effectively delay its progression [2]. Accordingly, early detection of low cognitive scores, which is an important warning sign of cognitive impairment, facilitates the prevention of dementia.

Traditional approaches to low-cognitive-score detection have employed paper tests, such as the Montreal Cognitive Assessment [3], Mini-Cog test [4], and Mini-Mental State Examination (MMSE) [5]. These examinations are lists of questions for evaluation of cognitive status. However, their questions are fixed and easy to memorize, limiting their use for daily measurements. Moreover, those paper-based exams require on-site questioners and take time to implement [6]. To address these issues, cognitive assessments based on dual-task performance, which combines cognitive and behavioral tasks, was proposed for detection of low cognitive status [7]–[9]. For example, Ahman *et al.* proposed a dual-task that requires subjects to walk while naming different animals or reciting months in reverse order [10].

Machine learning techniques have also been used in dual-task-based detection of low cognitive scores. Boettcher *et al.* applied support vector machine (SVM) [11] to dual-task-based detection of mild cognitive impairment (MCI) [12]. Similarly, Matsuura *et al.* used a dual-task composed of onset stepping and two-digit addition/subtraction. Because the calculation questions are generated randomly, that dual-task is suitable for daily measurements [13]. In that work, a 12-dimensional feature vector was extracted from the dual-task performance data, and then various machine learning methods such as SVM, random forest, and shallow neural networks were applied to detect low cognitive scores [13]. This method achieved high sensitivity and specificity on a database of subjects from three facilities operated by the same company. The three facilities shared the same routines and brain training programs for older adults, leading to similarity of the data between the three facilities. Therefore, the performance of their approach was worse on a larger dataset containing various subjects with different lifestyles from a larger number of facilities [13].

In summary, the previous studies had two major problems. First, the locomotive features were pre-designed, making it difficult to acquire complete motion information from various databases. Second, the traditional machine learning techniques they used have limited learning ability, especially when the training data are imbalanced. To address these issues, we proposed a framework for detecting low cognitive scores based on a spatio-temporal graph convolutional network (ST-GCN) [14] with a loss function specifically designed to handle imbalanced data. The contributions of this paper are threefold:

(1) Dual-task gait analysis using ST-GCN: We applied ST-GCN using joints of the whole 3D skeleton to extract spatio-temporal gait features from dual-task performance data, and then computes high-level features by deep convolution.

(2) Sensitivity+specificity loss for imbalanced data: We designed a loss function that directly optimizes the summation of sensitivity and specificity to address the problem of data imbalance for the proposed framework.

(3) Comprehensive experimental validation using a cross-facility database: Because the proposed approach is data-driven, it had much better performance on the cross-facility database than the method of the previous work [13].

II. RELATED WORK

A. Dual-task-based detection of low cognitive scores

Performing a dual-task (i.e., two tasks simultaneously) imposes a heavier cognitive load than performing a single task,

*We thank the Misasagikai Social Welfare Foundation for giving us the opportunity to collect data. This research was partly supported by JSPS KAKENHI Grant No. JP18H04115.

[1] S. Wu, Y. Makihara, K. Aoki, M. Niwa, and Y. Yagi are with the Department of Intelligent Media, ISIR, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan.

[2] F. Okura is with Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita, Osaka 565-0871, Japan.

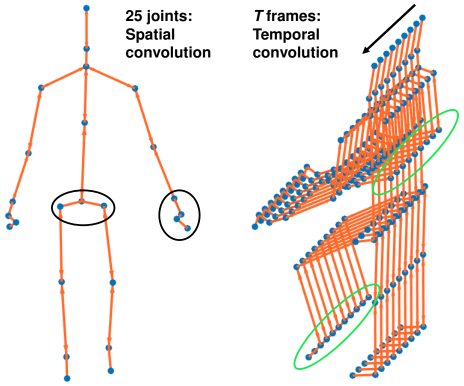


Fig. 1. Spatial and temporal convolution implemented by ST-GCN on both single and sequential data: the green and black ovals show examples of temporal and spatial convolutions, respectively.

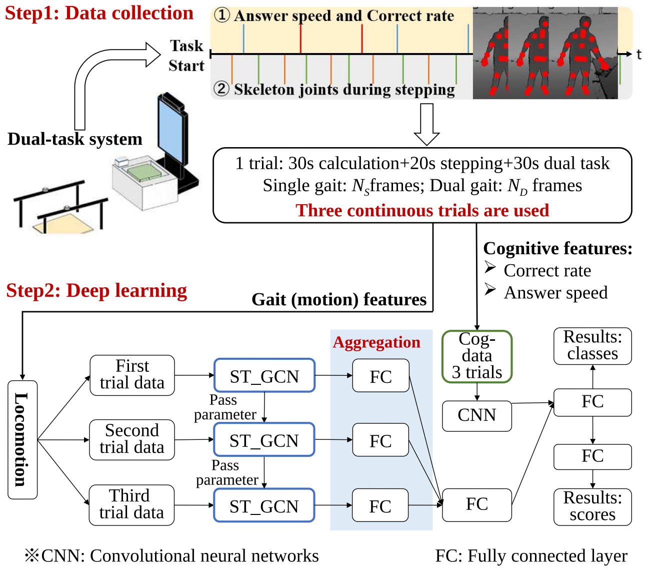


Fig. 2. The proposed framework for the regression of cognitive scores, and the detection of potential dementia (i.e., cognitive score ≤ 23), and potential MCI (i.e., cognitive score ≤ 27).

especially for people with low cognitive scores. Therefore, a dual-task can effectively detect subjects with low cognitive scores [10]. Performance on a dual-task has been shown to reflect trends of dementia and MCI well [10]. Because gait disorder is connected with low cognitive scores [15], it is preferable to involve walking or stepping in real dual-task-based applications [10], [13].

B. Skeleton-based action recognition

Although it has a different purpose, skeleton-based action recognition [16], [17] is relevant to our work, because we also analyze the 3D skeleton sequence during dual-task performance. Before the deep-learning era, methods of action recognition depended on handcrafted features such as motion patterns and trajectories [18]–[21], resulting in limited generalization capability [22]. Thereafter, the ST-GCN architecture was proposed to solve this problem [14]. To use whole skeleton data, it employs a two-directional graph convolution, as shown in Fig. 1. This expresses the relationships between adjacent joints, and the connections among adjacent frames. Because of the high efficiency with which it expresses 3D sequence data, ST-GCN has been widely used in 3D skeleton-based action recognition [23]–[25]. Similarly to action recognition, conventional approaches to the detection of low cognitive scores also suffer from the above-mentioned problems with handcrafted features, and the representation of gait features is particularly difficult [13]. Therefore, in this study, we apply an ST-GCN to the detection of low cognitive scores to improve its generalization capability.

Nevertheless, dual-task-based detection of low cognitive scores is different from action recognition in some aspects. For example, action recognition is based on a sequence of locomotive data, and the features are extracted mainly from the relationships between adjacent frames or adjacent joints. In contrast, dual-task performance data contains several different trials. Not only the relationship between adjacent frames, but also the connections among different trials are required to be considered in cognitive status assessment. This is because people with low cognitive scores tend to perform unstably when they repeat a dual-task several times. In addition, sequential data of cognitive features are important in the detection of low cognitive scores, Therefore, we need a network that represents not only spatio-temporal gait features, but also cognitive and cross-trial features.

C. Loss function for binary classification

Networks for binary classification often have two nodes in the last layer, which encode the probabilities of positive and negative, and the network parameters are trained by minimizing a well-known cross-entropy loss function [26]. Besides the cross-entropy loss function, the contrastive loss function [27] and triplet loss function [28] have been used for pair inputs in several tasks such as biometric authentication. These loss functions reduce the dissimilarity between samples of the same class and increase the dissimilarity between samples of different classes.

Nevertheless, none of the above loss functions can solve the problem of severe data imbalance between positive and negative samples, which is the case in our research as shown in Figs. 5(a) and 5(b). Kobayashi proposed a loss function based on the F-measure to solve the data imbalance problem [29]. However, that loss function has no direct relationship to the specificity measure used in this study.

III. Proposed approach

A. Overview of the proposed approach

We propose a framework for dual-task-based detection of low cognitive scores using the ST-GCN. This framework can also be used to regress cognitive scores. Figure 2 shows an overview of the proposed approach, from data collection to result generation. The dual-task used in this study consisted of stepping and two-digit addition and subtraction, as in [13]. Each subject was requested to step on the yellow mat and answer the questions shown on a front-facing display.

The detection of low cognitive scores based on a single-trial datum is easily influenced by non-cognitive factors such as the subject's emotional and bodily situation. Furthermore, people with low cognitive scores tend to perform unstably when they repeat the same dual-task. Therefore, exploiting the relationships among different trials of dual-task performance data can improve detection ability. In this study, each subject's sample consisted of three continuous trials. During each trial of the dual-task process, the answering speed and correct answer rate were recorded as cognitive data, and the 3D skeleton joints captured by Kinect V2 were used as motion data. We used a convolutional neural networks (CNN) to extract high-level features of the three trial's cognitive data (a 12-dimensional vector). Further, three ST-GCNs, which corresponded with each other by passing parameters, were employed to extract the spatio-temporal features of the three trial's motion data. Finally, we used an aggregation network to combine all of these intermediate outputs.

This framework fuses the features from multiple trials automatically and outputs a score that assesses cognitive status. The proposed approach can be used to detect potential dementia (i.e., the score ≤ 23) or potential MCI (the score ≤ 27) by setting thresholds. Also, it can be directly used to regress cognitive scores, as shown in Fig. 2. In the following section, we will explain the proposed method in more detail.

B. Dual-task performance data sampling

We used a dual-task system proposed in previous works [13], [30] to collect the raw data, as shown in Fig. 2. Compared with MMSE examination, which takes an average of 10 minutes, this dual-task system takes 80 seconds for each measurement trial. Furthermore, it is totally automatic, and its questions are generated randomly to facilitate easy application to daily measurement. In each trial, this system automatically implements a dual-task paradigm composed of a 30-sec calculation, 20 seconds of on-site stepping, and a 30-sec dual-task (calculation while stepping), sequentially. In total, three trials of 240-sec ($(30+20+30)\times 3 = 240$) were sequentially collected for each sample.

As Fig. 2 shows, each sample consists of two types of data. (1) Cognitive data: answering speed and correct answer rate during both single and dual (both 30-sec) calculation tasks for all the three trials; (2) motion data: 3D skeleton joints during both single and dual (20-sec and 30-sec, respectively) stepping tasks. The motion data were captured by Kinect V2, which occasionally has the problem of missing detection. Thus, we simply skipped the frames in which detection was missed. Moreover, the Kinect system in each facility used its own coordinates. To normalize the data from different Kinect systems, we implemented camera calibration by the random sample consensus (RANSAC) [31] algorithm as a pre-processing step. That is, we first detected the floor or roof plane from the depth image by RANSAC, and then computed the rotation matrix between the floor/roof plane and the horizontal plane, and finally implemented the camera calibration based on the rotation matrix.

The subject's MMSE scores were collected every 6 months. With respect to the detection, we focused on potential dementia (i.e., MMSE≤ 23) and potential MCI (i.e., MMSE≤ 27). For the classification of potential dementia and non-dementia, participants with MMSE scores of ≤ 23 and > 23 were labelled as positive and negative, respectively. Similarly, for the distinction between potential MCI and healthy people, participants with MMSE scores of ≤ 27 and > 27 were labelled as positive and negative, respectively. These positive or negative labels are used as the ground-truth labels during training. In contrast, for cognitive score regression, the MMSE scores were directly used as the ground-truth for training.

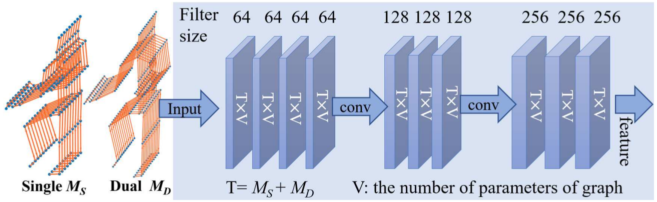


Fig. 3. Outline of layers of ST-GCN

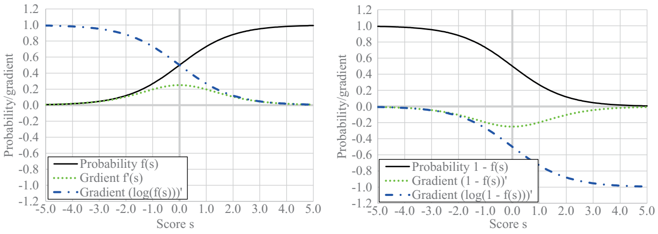
C. Spatio-temporal feature extraction

In this section, we explain how we extracted the spatio-temporal features from the dual-task motion data. Whereas the cognitive features (i.e., correct answer rate and answering speed) could be directly computed, the representation of gait features from the motion data (e.g., frame sequences of 2D images or 3D skeletons) was nontrivial. A possible method of gait representation is to employ predesigned/handcrafted features such as walking speed, step width symmetry angle, step width, and normalized gait speed, which were computed using gait analysis tools [32], [33]. However, the use of such features has some drawbacks: they represent only certain spatio-temporal aspects of gait, and they are not robust against missing frames or outliers among the estimated skeletons. Another method is to represent gait features in a data-driven way. ST-GCN [14] is one such data-driven method, and it has been employed in action recognition research to address the above-mentioned drawbacks of handcrafted features [13]. We therefore employed it in our study.

Figure 3 shows an outline of the layers of the ST-GCN used in this study. From each dual-task trial, we extracted M_S and M_D continuous frames of single and dual stepping, respectively, and used them as the inputs of the ST-GCN. Our network differed from that of the conventional ST-GCN [14] mainly in two ways, as shown in Fig. 2: (1) we used parameter passing and an aggregation network with multiple fully connected layers to fuse the high-level features of the three trials' motion data, and (2) we fused the motion features with cognitive features by another fully connected layer. Finally, for the binary classification task, the network generated a scalar value s as the initial output, and then classified it as positive (i.e., a subject whose ground-truth MMSE ≤ 23 or ≤ 27) if the value is positive, otherwise classified it as negative. For the regression task, the network outputs a predicted MMSE as cognitive score.

D. Sensitivity+specificity loss

In the detection task, subjects with low cognitive scores (i.e., MMSE$\leq$23 or MMSE$\leq$27) should be recognized as



(a) Positive sample

(b) Negative sample

Fig. 4. Probabilities and gradients per sample. The weight and denominators in Eq. 10 are omitted for simplicity. As the probability (i.e., $f(s)$ and $1-f(s)$ for the positive and negative classes, respectively, shown by black solid lines) approaches 0, the gradient with the original sigmoid function (green dotted lines) asymptotically approaches 0 (i.e., vanishes). In contrast, the probability predicted by the logarithmic version (blue dashed lines) asymptotically approaches $1/-1$ (i.e., does not vanish) for the positive and negative classes, respectively.

positive, and vice versa. Given N as the total number of samples in a mini-batch, the accuracy measure for detection of low cognitive scores is defined as

$$P_{\text{accuracy}} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{FP}} + N_{\text{TN}} + N_{\text{FN}}}, \quad (1)$$

where N_{TP} and N_{TN} signify the numbers of correctly classified positive and negative samples, and N_{FP} and N_{FN} denote the numbers of misclassified positive and negative samples.

A possible method of optimizing the network parameters during training of the classifier is to maximize the accuracy. However, there are generally fewer subjects with low cognitive scores (i.e., positive samples) than those with high cognitive scores (i.e., negative samples), as shown in Fig. 5. Thus, the trained classifier is considerably biased. To address this problem caused by data imbalance, we employ both sensitivity and specificity, two important criteria for performance evaluation in the detection of low cognitive scores. Sensitivity and specificity denote the ability to correctly detect positive and negative samples, respectively. Given N_{P} and N_{N} as the numbers of positive and negative samples in a mini-batch, sensitivity and specificity are defined as

$$P_{\text{sensitivity}} = \frac{N_{\text{TP}}}{N_{\text{P}}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \quad (2)$$

$$P_{\text{specificity}} = \frac{N_{\text{TN}}}{N_{\text{N}}} = \frac{N_{\text{TN}}}{N_{\text{FP}} + N_{\text{TN}}}. \quad (3)$$

We then devised a *sensitivity+specificity loss* function, which directly optimizes the summation of the predicted sensitivity $P_{\text{sensitivity}}$ and specificity $P_{\text{specificity}}$ (i.e., $P_{\text{sensitivity}} + P_{\text{specificity}}$).

However, the above-mentioned loss function still has a problem. Classification into positive/negative classes is done by thresholding of the predicted scalar value s that assesses the cognitive status, as described in the previous subsection. Thus, the numbers N_{TP}, N_{FN}, N_{TN}, and N_{FP} change discretely at the boundary $s = 0$, and as a result, the loss function is not differentiable at the boundary $s = 0$. Moreover, the gradient is always zero (i.e., vanishes) except for that at the boundary, which is unfavorable for training with back-propagation. Therefore, we relaxed the sensitivity and specificity into probabilistic ones to make them differentiable. Specifically, we introduced a sigmoid function for value s as

$$f(s) = \frac{1}{1 + \exp(-s)}, \quad (4)$$

and then consider two probabilities: $f(s)$ and $1 - f(s)$ for the positive and negative classes, respectively. This indicates that proportions of the sample equal to $f(s)$ and $1 - f(s)$ belong to the positive and negative class, respectively. We then computed the relaxed version of the sensitivity $\tilde{P}_{\text{sensitivity}}$ and specificity $\tilde{P}_{\text{specificity}}$ for a mini batch with N samples as

$$\tilde{P}_{\text{sensitivity}} = \frac{1}{N_{\text{P}}} \sum_{i \in S_{\text{P}}} f(s_i) \quad (5)$$

$$\tilde{P}_{\text{specificity}} = \frac{1}{N_{\text{N}}} \sum_{i \in S_{\text{N}}} (1 - f(s_i)), \quad (6)$$

where s_i is the score of the i-th sample in the mini-batch, and S_{P} and S_{N} are sets of the indices of positive and negative samples in the mini-batch. The loss function L that is subsequently minimized is subsequently defined as

$$\begin{aligned} L &= -(\tilde{P}_{\text{sensitivity}} + \tilde{P}_{\text{specificity}}) \\ &= -\left(\frac{1}{N_{\text{P}}} \sum_{i \in S_{\text{P}}} f(s_i) + \frac{1}{N_{\text{N}}} \sum_{i \in S_{\text{N}}} (1 - f(s_i)) \right) \quad (7) \end{aligned}$$

Although the loss function in Eq. (7) is differentiable everywhere (see Fig. 4), it still has the vanishing gradient problem at extreme predictions $f(s) \to \{0, 1\}$, which is unfavorable for gradient-based optimization. Specifically, the gradient with respect to the i-th sample is computed as

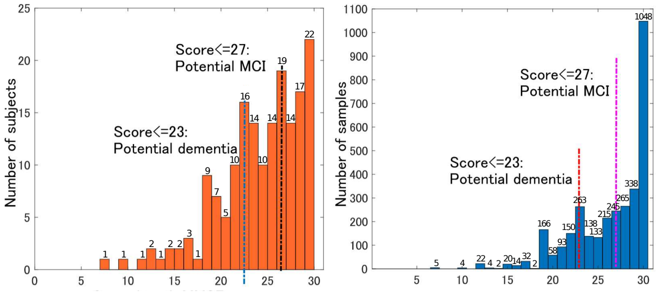
$$\frac{\partial L}{\partial s_i} = \begin{cases} \frac{1}{N_{\text{P}}} f(s_i)(1 - f(s_i)) & (i \in S_{\text{P}}) \\ -\frac{1}{N_{\text{N}}} f(s_i)(1 - f(s_i)) & (i \in S_{\text{N}}) \end{cases}. \quad (8)$$

The gradient asymptotically approaches 0 as the probability of the positive class $f(s_i)$ or that of the negative class $1 - f(s_i)$ approaches 0 (see Fig. 4). We therefore took the logarithm of the sigmoid function, which solves the vanishing gradient problem as well as having a monotonically increasing property. Furthermore, to suppress unstable update within highly imbalanced mini-batches, we introduced a mini-batch-wise weight. Specifically, we used the harmonic mean of the numbers N_{P}, N_{N} of positive and negative samples in the mini-batch, respectively, as the weight: $w(N_{\text{P}}, N_{\text{N}}) = 2/(1/N_{\text{P}} + 1/N_{\text{N}}) = (2N_{\text{P}}N_{\text{N}})/(N_{\text{P}} + N_{\text{N}})$. This is similar to one implementation of a F-measure-based loss function [34]. Given a mini-batch size of $N = N_{\text{P}} + N_{\text{N}}$, the weight is maximized when the ratio of positive to negative samples is completely balanced (i.e., $N_{\text{P}} = N_{\text{N}}$). The weight is minimized (i.e., $w = 0$) when the ratio of positive to negative samples is completely imbalanced (i.e., $N_{\text{P}} = 0$ or $N_{\text{N}} = 0$), which effectively mitigates the influence of imbalanced mini-batches. Consequently, the logarithmic version of the loss function with the mini-batch-wise weight is defined as

$$\tilde{L} = -w(N_{\text{P}}, N_{\text{N}}) \left(\frac{1}{N_{\text{P}}} \sum_{i \in S_{\text{P}}} \log(f(s_i)) + \frac{1}{N_{\text{N}}} \sum_{i \in S_{\text{N}}} \log(1 - f(s_i)) \right). \quad (9)$$

The gradient is then calculated by

$$\frac{\partial \tilde{L}}{\partial s_i} = \begin{cases} \frac{w(N_{\text{P}}N_{\text{N}})}{N_{\text{P}}} (1 - f(s_i)) & (i \in S_{\text{P}}) \\ -\frac{w(N_{\text{P}}N_{\text{N}})}{N_{\text{N}}} f(s_i) & (i \in S_{\text{N}}) \end{cases}. \quad (10)$$



(a) Subject statistics

(b) Sample statistics

Fig. 5. Subject/sample statistics of 6 facilities with different MMSE scores. Subjects/samples with low MMSE scores (score is less than or equal to 23) are fewer than those with high MMSE scores (score is larger than 23).

The gradient does not asymptotically approach 0 (i.e., does not vanish). Instead, it approaches $w(N_{\mathrm{P}}, N_{\mathrm{N}})/N_{\mathrm{P}}$ and $-w(N_{\mathrm{P}}, N_{\mathrm{N}})/N_{\mathrm{N}}$ for positive and negative samples, respectively, as its probability approaches 0 (see Fig. 4). This is favorable for the gradient-based optimization.

IV. Experiments

In this section, we first describe our cross-facility database, and then show the comparison results between the proposed approach and an improved version of the previous method [13] in terms of performance on cognitive scores regression, and the detection of low cognitive scores.

A. Database

This study's data were collected from 6 facilities for older adults, where our dual-task system was installed for early-stage detection of low cognitive scores. Some people in these 6 facilities have brain training programs, but some do not. The people in this database had different life-styles, resulting in larger diversity compared with the database used in [13]. Each person with one MMSE score was regarded as a subject. In total, the data of 171 subjects and 3,217 samples were collected. Each sample contains three continuous trials of dual-task performance data. Notably, one subject may have many samples because they completed the dual-task on a daily basis. Figure 5(a) shows the distribution of subjects who have different MMSE scores, and Fig. 5(b) shows the distribution of samples with different MMSE scores. Fig. 5(a) shows that the number of subjects with high MMSE scores (score$>$ 23) was greater than that with low MMSE scores (score$\leq$ 23), leading to an imbalance between positive and negative data. The distribution in Fig. 5(b) illustrates that the positive-negative data imbalance is more severe for samples.

This study was approved by the Research Ethics Committee of the Institute of Scientific and Industrial Research, Osaka University (Osaka, Japan). All experiments were performed in accordance with the guidelines and regulations. Informed consent was obtained for all subjects.

B. Performance evaluation

In this section, we compare the proposed approach with an improved version of the method by Matsuura *et al.* [13]. Although a shallow neural network was used for detection of low cognitive scores in [13], that method offers no solution to the data imbalance problem. Therefore, we improved that method using the LightGBM algorithm [35] with SMOTE [36]. This yielded notably better results on our cross-facility database with heavy data imbalance. The SMOTE augmented data to ensure that the dataset has the same numbers of samples with each MMSE score, and LightGBM was used to predict MMSE scores from dual-task performance data. Besides, Matsuura *et al.* used a single trial of dual-task performance data for detecting low cognitive scores, but we used three continuous trials for better accuracy. To make a fair comparison, we also implemented the use of the same three continuous trials by the previous method [13]. As in the original version of [13], we searched for the highest sensitivity+specificity value in the receiver operating characteristic (ROC) curves given by the improved implementation of [13] as a criterion for comparison. Because the improved version of [13] was used as the baseline method for evaluating the proposed approach, we denote the improved version with three trials as the "baseline method". The predesigned features of the baseline method are occasionally failed to be extracted because of the instability of Kinect. In total, the baseline method failed in 86 trials. Therefore, we excluded those 86 trials from the evaluation of the baseline method.

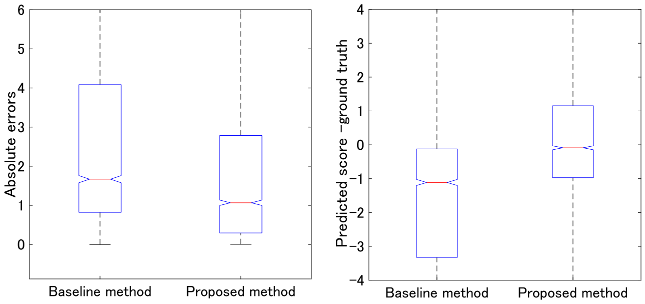
In our experiment, we first evaluated the performance of cognitive score regression based on mean absolute error (MAE) and root mean squared error (RMSE) to show the effectiveness of the proposed ST-GCN-based framework. Second, to demonstrate the effectiveness of the proposed framework as well as the sensitivity+specificity loss function, we evaluated the performance of the detection of low cognitive scores based on binary classification (i.e., the classifications of potential dementia (MMSE$\leq$ 23) / non-dementia and potential MCI (MMSE$\leq$ 27) / healthy person). Here, we used the accuracy, sensitivity, and specificity (defined in Eqs. 1-3) to evaluate detection performance. The training of the proposed framework was implemented on a GPU with 48G of memory. The learning rate of each ST-GCN was set to 0.1, and the dropout value was set to 0.8. In addition, the numbers of frames in the single and dual tasks were set to $M_S = 160$ and $M_D = 260$, respectively, to ensure that the extracted frames contained several gait cycles.

In the first comparison with respect to cognitive score regression, we used the mean square error loss for regressing cognitive scores because either the cross-entropy loss or the proposed sensitivity+specificity loss is suitable for classification, but neither are suitable for regression. Here, we compare the performance of the proposed approach with that of the baseline method [13] using data from one trial and three continuous trials, respectively. Table I shows MAE and RMSE of the compared approaches, illustrating that the baseline method using three continuous trials' data for each sample achieved slightly better results than that using one trial's data for each sample. Thus, in the following experiments, we used the baseline method with three continuous trials' data for each sample to evaluate the proposed approach.

TABLE I

COMPARISON RESULTS OF COGNITIVE SCORE REGRESSION. BOLD FONTS INDICATE THE BEST ACCURACY AMONG METHODS. THIS CONVENTION IS USED THROUGHOUT THIS PAPER.

Method	MAE	RMSE
Baseline method (1 trial)	2.71	3.75
Baseline method (3 trials)	2.67	3.67
Proposed method	**1.81**	**2.88**



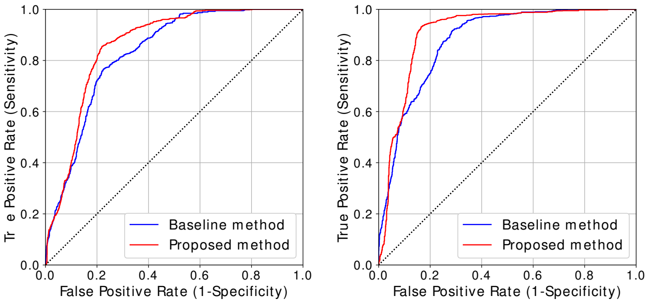
(a) Absolute error

(b) Signed difference

Fig. 6. Significant differences of the absolute errors and signed difference (difference between the predicted and ground-truth MMSE scores) using the baseline and proposed methods ($p < 0.01$).

Figure 6(a) shows that there was a significant difference between the absolute errors of the baseline and proposed methods using three continuous trials' data for each sample. We also computed the signed difference between the predicted and ground-truth MMSEs to determine the bias of the predictions. Figure 6(b) shows that there was a significant difference between the signed differences computed from the compared algorithms under the same conditions as Fig. 6(a). These figures were created using the analysis of variance method (ANOVA [37]). In Fig. 6(b), the scores were severely underestimated by the baseline method, probably because the handcrafted features cannot express the correct motion information. From Table I and Figs. 6(a) and 6(b), we conclude that the proposed method significantly outperforms the baseline method in terms of both the MAE and RMSE. This is consistent with our analysis that the proposed ST-GCN-based framework has more expressive power than the handcrafted-feature-based method. Moreover, the proposed algorithm is totally data-driven, so that it performs well on our cross-facility database.

With respect to the detection of low cognitive scores based on binary classification, we not only compared the proposed approach with the baseline method [13], but also compared the proposed sensitivity+specificity loss and the original cross-entropy loss functions. Table II shows the comparison results of the detection of potential dementia (i.e., MMSE$\leq$ 23) and potential MCI (i.e., MMSE$\leq$ 27), using the baseline and proposed methods. Table II shows that the proposed approach outperforms the baseline method considerably in terms of most aspects of the accuracy, sensitivity, and specificity of both potential dementia and MCI detection. Comparing the results with vs. without the proposed sensitivity+specificity loss, the proposed loss function achieved a better balance between sensitivity and specificity than the original cross-entropy loss.



(a) Potential dementia detection

(b) Potential MCI detection

Fig. 7. Comparison between the ROC curves of the baseline and proposed methods for potential dementia/MCI detection.

Finally, we compare the ROC curves of the baseline and proposed methods to examine the relationship between sensitivity and specificity. Figures 7(a) and 7(b) show the comparison results for the detection of potential dementia (i.e., MMSE$\leq$ 23) and potential MCI (i.e., MMSE$\leq$ 27), respectively. Compared with the baseline method, the proposed approach achieved a higher upper limit of sensitivity+specificity. In summary, the results demonstrated that the proposed approach is more efficient at both cognitive score regression and low-cognitive-score detection than the baseline method. Furthermore, the proposed sensitivity+specificity loss function is more robust against data imbalance than the original cross-entropy loss.

V. DISCUSSION

In this section, we discuss the limitations of the proposed method and directions for future improvement. The proposed approach has two limitations: (1) The graph convolution is only implemented among adjacent joints, as shown in Fig. 1; (2) The use of only parameter passing and an aggregation network is not sufficient to extract the cross-trial features.

Regarding the first limitation, joints that are not adjacent may nevertheless have strong relationships. For example, cognitive impairment may impact people's motor coordination capability, leading to difficulty maintaining good correspondence between the hands and legs. Therefore, a graph convolution considering the relationships between all pairs of joints would be a promising direction for performance improvement. With respect to the second limitation, the parameter passing and aggregation network used to extract the cross-trial feature is not sufficient. For example, the differences between the statistical features of different trials cannot be extracted by the current system. A middle hidden network that connects each of the middle output features between different trial's data would be preferable, as this could exploit strong cross-trial features.

VI. CONCLUSIONS

In this study, we proposed an approach for both cognitive score regression and low-cognitive-score detection based on MMSE. Although the MMSE score is not a definite

TABLE II

Comparison results of cognitive impairment detection. Acc, Sen, and Spec indicates accuracy, sensitivity, and specificity.

Method \Detection target, evaluation items	Potential dementia (i.e., MMSE $\leq$ 23)				Potential MCI(i.e., MMSE $\leq$ 27)			
	Acc	Sen	Spec	Sen+Spec	Acc	Sen	Spec	Sen+Spec
Baseline method	0.78	0.77	0.77	1.54	0.81	**0.92**	0.71	1.63
ST-GCN w/ cross-entropy loss	0.78	**0.91**	0.73	**1.64**	0.86	**0.92**	0.81	1.73
ST-GCN w/ sensitivity+specificity loss (proposed)	**0.80**	0.85	**0.79**	**1.64**	**0.88**	**0.92**	**0.84**	**1.76**

diagnosis, identifying the decline in MMSE score at the earliest possible stage is crucial for early diagnosis. Because the MMSE cannot be used as a daily monitoring tool, we proposed this dual-task-based system for monitoring the cognitive status of older adults in 6 facilities. In contrast to previous studies that used predesigned locomotive features and traditional machine learning methods [13], [32], [33], the proposed approach applied the ST-GCN to extract spatio-temporal locomotive features, fully exploiting the available 3D skeleton information. Furthermore, we proposed a sensitivity+specificity loss function for the proposed networks to directly optimize the summation of sensitivity and specificity of the detection of low cognitive scores. The experimental results demonstrated the proposed method's effectiveness at solving positive-negative data imbalance. In our future work, we will design a graph convolution network that considers both adjacent and non-adjacent joints. We will also focus on connecting dual-task performance with definite diagnoses.

References

[1] World Health Organization. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia, 2019.

[2] G. Livingston, A. Sommerlad, V. Orgeta, et al. Dementia prevention, intervention, and care. *The Lancet*, 390(10113):2673–2734, 2017.

[3] Z. S. Nasreddine, N. A. Phillips, V. Bedirian, et al. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.*, 53(4):695–699, 2005.

[4] D. P. Seitz, C. C. Chan, H. T. Newton, et al. Mini-cog for the diagnosis of alzheimer's disease dementia and other dementias within a primary care setting. *The Cochrane database of systematic reviews*, 2(2):1–40, 2018.

[5] V. C. Pangman, J. Sloan, and L. Guse. An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. *Appl. Nurs. Res.*, 13:209–213, 2000.

[6] H. T. Jung, H. Lee, K. Kim, et al. Estimating mini mental state examination scores using game-specific performance values: A preliminary study. *IEEE EMBC*, pages 1518–1521, 2018.

[7] K. Aoki, T. T. Ngo, I. Mitsugami, et al. Early detection of lower mmse scores in elderly based on dual-task gait. *IEEE Access*, 7:40085–40094, 2019.

[8] G. Mancioppi, L. Fiorini, E. Rovini, et al. How dominant hand and foot dexterity may reveal dementia onset: A motor and cognitive dual-task study. *IEEE EMBC*, pages 5619–5622, 2020.

[9] H. Pashler. Dual-task interference in simple tasks: data and theory. *Psychol Bull*, 116(2):220–244, 1994.

[10] H. B. Ahman, Y. Cedervall, L. Kilander, et al. Dual-task tests discriminate between dementia, mild cognitive impairment, subjective cognitive impairment, and healthy controls-a cross-sectional cohort study. *BMC Geriatr*, 20(258):1–10, 2020.

[11] C. Corinna and V. Vladimir. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[12] L. N. Boettcher, M. Hssayeni, A. Rosenfeld, et al. Dual-task gait assessment and machine learning for early-detection of cognitive decline. *IEEE EMBC*, pages 3204–3207, 2020.

[13] T. Matsuura, K. Sakashita, A. Grushnikov, et al. Statistical analysis of dual-task gait characteristics for cognitive score estimation. *Scientific Reports*, 9(1):401–405, 2019.

[14] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, pages 7444–7452, 2018.

[15] R. Liao, Y. Makihara, D. Muramatsu, et al. A video-based gait disturbance assessment tool for diagnosing idiopathic normal pressure hydrocephalus. *IEEJ Trans. Electr. Electron. Eng.*, 15(3):433–441, 2020.

[16] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. *CVPR*, pages 1110–1118, 2015.

[17] J. Liu, A. Shahroudy, D. Xu, et al. Spatio-temporal lstm with trust gates for 3d human action recognition. *ECCV*, pages 816–833, 2016.

[18] J. F. Hu, W. S. Zheng, J. Lai, et al. Jointly learning heterogeneous features for rgb-d activity recognition. *CVPR*, pages 5344–5352, 2015.

[19] J. Shotton, T. Sharp, A. Kipman, et al. Real-time human pose recognition in parts from single depth images. *CVPR*, pages 1297–1304, 2015.

[20] J. F. Hu, W. S. Zheng, J. Lai, et al. Multi-modal feature fusion for action recognition in rgb-d sequences. *ISCCSP*, pages 1–4, 2014.

[21] J. Sung, C. Ponce, B. Selman, et al. Human activity detection from rgbd images. *AAAI*, pages 47–55, 2011.

[22] L. Wang, D. Q. Huynh, and P. Koniusz. A comparative review of recent kinect-based action recognition algorithms. *IEEE Transactions on Image Processing*, 29:15–28, 2020.

[23] Y. Jiang, K. Song, and J. Wang. Action recognition based on fusion skeleton of two kinect sensors. *IEEE ICCST*, pages 240–244, 2020.

[24] L. Shi, Y. F. Zhang, J. Cheng, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *CVPR*, pages 12026–12035, 2019.

[25] C. L. Yang, A. Setyoko, H. Tampubolon, et al. Pairwise adjacency matrix on spatial temporal graph convolution network for skeleton-based two-person interaction recognition. *ICIP*, pages 2166–2170, 2020.

[26] Z. Zhang and M. R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *NeurIPS*, pages 8792–880, 2018.

[27] R. Hadsell, S. Chopra, and L. LeCun. Dimensionality reduction by learning an invariant mapping. *CVPR*, 2:1735–1742, 2006.

[28] J. Wang, Y. Song, T. Leung, et al. Learning fine-grained image similarity with deep ranking. *CVPR*, pages 1386–1393, 2014.

[29] T. Kobayashi. Spiral-net with f1-based optimization for image based crack detection. *ACCV*, pages 88–104, 2018.

[30] F. Okura, I. Mitsugami, M. Niwa, et al. Automatic collection of dual-task human behavior for analysis of cognitive function. *ITE Trans. Media Technology and Applications*, 6(2):138–150, 2018.

[31] A. Hast, J. Nysjö, and A. Marchetti. Optimal ransac–towards a repeatable algorithm for finding the optimal set. *J. of WSCG.*, 21(1):21–30, 2013.

[32] A. Sabo, S. Mehdizadeh, K. D. Ng, et al. Assessment of parkinsonian gait in older adults with dementia via human pose tracking in video data. *J. NeuroEngineering Rehabil*, 17(97):1–10, 2020.

[33] A. M. DeCock, E. Fransen, S. Perkisas, et al. Gait characteristics under different walking conditions: Association with the presence of cognitive impairment in community-dwelling older people. *PLoS ONE*, 12(6):1–19, 2017.

[34] M. W. P. David. What the f-measure doesn't measure: Features, flaws, fallacies and fixes. *arXiv:1503.06410*, 2015.

[35] G. Ke, Q. Meng, T. Finley, et al. Lightgbm: A highly efficient gradient boosting decision tree. *InAdv. Neural Inf. Process Syst.*, pages 3146–3154, 2017.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, et al. SMOTE: synthetic minority over-sampling technique. *Int. J. Artif. Intell. Res.*, 16:321–357, 2002.

[37] G. E. P. Box. Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *Ann. Math. Statist.*, 25(2):290–302, 1954.