Detecting Lower MMSE Scores in Older Adults Using Cross-Trial Features From a Dual-Task With Gait and Arithmetic

SHUQIONG WU\textsuperscript{1}, TAKU MATSUURA\textsuperscript{1}, FUMIO OKURA\textsuperscript{2}, (Member, IEEE), YASUSHI MAKIHARA\textsuperscript{1}, CHENGJU ZHOU\textsuperscript{3}, KOTA AOKI\textsuperscript{1}, IKUHISA MITSUGAMI\textsuperscript{4}, (Member, IEEE), AND YASUSHI YAGI\textsuperscript{1}, (Senior Member, IEEE)

\textsuperscript{1}Department of Intelligent Media, SANKEN (The Institute of Scientific and Industrial Research), Osaka University, Suita, Osaka 565-0871, Japan
\textsuperscript{2}Department of Multimedia Engineering, Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan
\textsuperscript{3}School of Software, South China University of Technology, Guangzhou 510006, China
\textsuperscript{4}Graduate School of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan

Corresponding author: Shuqiong Wu (wu@am.sanken.osaka-u.ac.jp)

This work was supported in part by the Japan Agency for Medical Research and Development (AMED) under Grant J215101010.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of SANKEN (The Institute of Scientific and Industrial Research), Osaka University, Osaka, Japan, under Approval No. H29-10.

ABSTRACT The Mini-Mental State Examination (MMSE) is widely used in clinics to screen for low cognitive status. However, it is limited in that it requires examiners to be present; and has fixed questions that constrain its repeated use. Thus, the MMSE cannot be used as a daily assessment to facilitate early detection of cognitive impairment. To address this issue, we developed an automated system to detect older adults with lower MMSE scores by analyzing performance during a dual task involving stepping and calculation, which can be used repeatedly because its questions were randomly created. Leveraging this advantage, this paper proposes a learning-based method to detect subjects with lower MMSE scores using multiple trials with the dual-task system. We investigated various patterns for effectively combining the features acquired during multiple continuous trials, and analyzed the sensitivity of the number $N$ of trials on detection performance to find the optimal $N$ via experiments. We compared our approach with previous methods and demonstrated the superiority of our strategy. Using the cross-trial feature, our approach achieved an overall performance (sensitivity + specificity) as high as 1.79 for detecting older adults whose MMSE score is equal to or less than 23 (indicate a relatively high probability of dementia), and 1.75 for detecting older adults whose MMSE score is equal to or less than 27 (indicative of a relatively high probability of mild cognitive impairment (MCI)).

INDEX TERMS Cognitive impairment, dementia, dual-task, machine learning, MCI, MMSE.

I. INTRODUCTION Dementia is characterized by a decline in memory, cognition, and behavioral ability, and is an increasingly prominent issue for older adults worldwide [1], [2]. According to statistics from the World Health Organization [1], there were around 50 million people with dementia worldwide in 2019, with 10 million new cases each year. This number is expected to reach 152 million in 2050. Dementia is caused by damage to brain cells or disruption of neural activity [1]. It is a progressive disease and the rate of progress varies for each individual. In the early stage of dementia, known as mild cognitive impairment (MCI), patients can still engage in complex activities, although they may need to exert relatively greater effort. In the late stage of dementia, impairments associated with cognitive deterioration could cause difficulties in daily life [1], [2]. Although dementia is considered incurable, early diagnosis and intervention could prevent the progression from the early to late stage [2]. However, early diagnosis can be overlooked because the symptoms of MCI are sometimes difficult to detect [2]–[5].

To address this issue, paper-based tests such as the Mini-Mental State Examination (MMSE) [6], [7], Montreal Cognitive Assessment [8], and Alzheimer’s Disease Assessment Scale-Cognitive Subscale [9] have been proposed to screen low cognitive status in clinical settings. These
tests contain various questions that are used to assess ability with respect to calculations, language, memory, and comprehension. For example, in the MMSE, which is a 30-point questionnaire, scores equal to or less than 27 and scores equal to or less than 23 indicate a relatively high probability of MCI and dementia, respectively [10]. The MMSE can be repeated every few months to check whether cognition decline has occurred. Because dementia can progress rapidly, i.e., within several months in some cases, more frequent assessments such as daily monitoring are beneficial. However, the MMSE cannot be used in daily monitoring because its questions are fixed, and subjects may memorize the answers. Moreover, it requires the presence of an examiner and is time-consuming [11]. To solve these problems, we propose a dual-task-based approach which can predict the MMSE score from daily dual-task performance data.

Dual-task, in which two different tasks are performed simultaneously, have been found effective in detecting low cognitive status [12]–[14]. The basic idea is that two different tasks performed simultaneously impose a heavier cognitive load compared with a single task, especially for people with low cognitive status [15]–[17]. Many studies have reported that gait is related to cognitive impairment [18]–[20]. Recently, Matsuura et al. proposed a novel dual-task that consists of a gait task and arithmetic questions [11]. They added various assessment features such as knee angles, and standard deviations of stepping speed and knee angles. However, as these features were extracted from single trial data, task performance was easily influenced by non-cognitive factors such as personal health condition or emotional state during the test. Mancioppi et al. investigated different types of motor activities such as gait task, toe-tapping, and forefinger-tapping in dual-task paradigms. They found that performance in all three motor tasks was related to cognition [21]. Lillian et al. proposed a novel dual-task containing three phases of walking accompanied with a cognitive task such as counting backwards in 3’s from 100 or reciting the alphabet. This paradigm was found to have 81.97% sensitivity and 67.74% specificity (the sum was 1.4971) [22]. Similarly, Digo et al. proposed a gait-based dual-task for screening cognitive decline. In their study, a type of wearable technology that termed “magnetic inertial measurement units” was adopted for precise gait analysis [23].

Recently, Mancioppi et al. proposed an improved paradigm for dual-task-based MCI detection. They proposed two types of dual-tasks: FTAP (Fore-finger tapping with cognitive task) and TTHP (Toe-tapping with cognitive task). MCI detection was implemented based on logistic regression models [24]. However, this system requires staff members to provide guidance to the subjects about how to wear the sensors and how to conduct the dual-tasks.

In the present study, we propose an automatic approach for detecting older adults with lower MMSE scores based on dual-task performance data collected using a dual-task system described in [11]. Compared with the MMSE, this dual-task system was faster and could be implemented without an examiner. Moreover, the calculation questions were created randomly (i.e., not fixed) to reduce the chance that the subjects would memorize the responses. When this system was installed in the community living facilities, older residents had the option to complete multiple trials on a daily basis. Leveraging this advantage, the proposed method can combine various dual-task-performance features across multiple trials in the same subject. To date, our system (shown in Fig. 1) has been installed in fourteen facilities for older adults to monitor the cognitive status of each resident.

Unlike the previous studies using single-trial features [11], [17], the present study investigated a way to effectively combine the extracted cognitive and locomotion features acquired during multiple continuous trials. After feature combination, we entered the combined cross-trial feature into a machine-learning algorithm to estimate MMSE scores. Then the proposed algorithm classifies older adults with low MMSE scores from those with high MMSE scores based on the estimated MMSE scores. The cross-trial feature selects the best performance among multiple trials, which is more statistically reliable, and hence reduces the influence of bad performances caused by non-cognitive factors. Therefore, the proposed approach achieves much better performance compared with previous methods that use single-trial features [11]. Another improvement compared with the previous research [11] is that the proposed approach used an effective machine learning method LightGBM [25] and a data augmentation technique SMOTE (Synthetic Minority Over-sampling Technique) [26] to solve the imbalance of the positive and negative samples. In addition, we analyzed the sensitivity of the number \( N \) of multiple trials and found the optimal \( N \) through experiments.

The contributions of this study are listed as follows:

1. We designed a cross-trial feature to detect older adults with lower MMSE scores, which was computed according to the dual-task features of multiple continuous trials. We found that the cross-trial feature was more statistically reliable than that obtained from a single trial, which improves the summation of sensitivity and specificity.

2. We introduced a validation technique to enable adaptive thresholding for classifications of older adults with lower MMSE scores. This technique yielded high performance in terms of the sensitivity and specificity, as shown in Tab. 5.

3. We analyzed the sensitivity of the number \( N \) of multiple trials and found the optimal \( N \) through experiments. The number of \( N \) could be manipulated to fit different applications by optimizing specific targets. For example, for applications in which higher specificity is desired, we could optimize the number \( N \) to maximize the specificity of detection (a detailed explanation is given in the Discussion section).

II. METHODS

A. DUAL-TASK EXPERIENCE SYSTEM

Here, we briefly recap the dual-task experience system we used, which was proposed in a previous study [11].
Figure 1 illustrates the architecture of this system, which consists of the Microsoft Kinect (a line of motion-sensing input devices produced by Microsoft and first released in 2010) V2 device, a personal computer (PC), a display screen, a QR code reader, hand rails, a force platform, and buttons for inputting responses. According to user instructions shown on the display screen, a subject can login with a prepared QR code, and then sequentially complete three stages of tasks that comprise the dual-task process. The first component is a single-task in which the subject answers arithmetic questions for 30 seconds. Next is a single-task in which the subject repeatedly steps in place for 20 seconds, and the third is a dual-task in which the subject steps repeatedly while answering arithmetic questions for 30 seconds. There are four difficulty levels for the arithmetic questions: easy (addition of two one-digit numbers), medium (subtraction of one one-digit number from one two-digit number smaller than 20), moderately hard (addition or subtraction of one one-digit number and one two-digit number smaller than 20), and hard (addition or subtraction of one one-digit number and one two-digit number). Subjects can choose any difficulty level. Responses to the questions are registered by pressing the buttons mounted on the handrails. While the subjects perform each task, their response-time and accuracy are recorded. Simultaneously, RGB color images, depth maps, and the coordinates of the subject’s body joints are captured using the Kinect. In addition, a force platform sensor on the floor registers and records contact with the subject’s feet.

**B. CROSS-TRIAL FEATURE EXTRACTION**

From the collected data, we found various features to represent performance in the cognitive task and locomotion task. These features were fused and optimized to generate the cross-trial feature. We then used the cross-trial feature to estimate MMSE scores for each subject. Finally, by comparing the estimated MMSE scores with a set threshold, we were able to classify subjects with lower MMSE scores.

As mentioned above, when the subjects engaged in the dual-task experience system, their stepping performance and arithmetic responses were recorded. From the various features of the cognitive and locomotion tasks, we sought to select those that would be most effective for estimating MMSE scores. We expected that accuracy for arithmetic questions would be an efficient feature because dementia can impair calculation abilities [28]. Gait features, such as walking speed and knee height, have also been found to be effective in dementia detection [29], [30]. To form a standard criterion, Matsuura *et al.* introduced a 12-dimensional feature (shown in Tab. 1) to describe the characteristics of performances during the single-task and dual-task processes [11]. Stepping speed was calculated using data from the force platform sensor, which recorded the moment of contact for each foot. The angle between the thigh and shank was calculated using the lateral view of 3D joint positions recorded by the Kinect [11]. Despite the advantages of this method, the feature identification was based on single trial data and thus could easily be influenced by the health or emotional state of the subject when performing the dual-task. In the present study, we investigated methods for aggregating a set of features obtained during multiple continuous trials into a cross-trial feature vector.

First, we investigated intra-subject variations in the estimated MMSE scores over the dual task trials. Specifically, we analyzed the relationship between the intra-subject standard deviation of the estimated MMSE scores and the ground-truth MMSE scores (termed “MMSE score” for simplicity), as shown in Fig. 2, where the MMSE scores were estimated by LightGBM [25]. LightGBM is described in the following section. From the figure, we can see that the standard deviation for the high MMSE scores was relatively small, which indicates that the estimated MMSE score did not vary substantially among multiple trials. Conversely, the standard deviation increased as the MMSE scores decreased. This means that the estimated MMSE scores among multiple trials
TABLE 1. Twelve features extracted from a dual-task process. Std. means standard deviation and N/A means not-applicable [11]: Step-Avg-S and Step-Avg-D denote the mean stepping speed during the single stepping task and dual-task, respectively. Step-Std-S and Step-Std-D denote the standard deviation of stepping speed during the single stepping task and dual-task, respectively. CalAns-Avg-S and CalAns-Avg-D denote the ratio of correct answers of the single calculation task and dual-task, respectively. CalSpd-Avg-S and CalSpd-Avg-D denote the calculation speed of the single calculation task and dual-task, respectively. Knee-Avg-S and Knee-Avg-D denote the mean knee angle of the single stepping task and dual-task, respectively. Knee-Std-S and Knee-Std-D denote the standard deviation of knee angles during the single stepping task and dual-task, respectively.

| Dual-task process | Single calculation task | Single stepping task | Dual-task (stepping + calculation) |
|-------------------|-------------------------|----------------------|-----------------------------------|
| Duration time     |                         |                      |                                   |
| Average stepping speed | N/A                    | Step-Avg-S           | Step-Avg-D                        |
| Std. of stepping speed | N/A                    | Step-Std-S           | Step-Std-D                        |
| Ratio of correct answers | CalAns-Avg-S           | N/A                  | CalAns-Avg-D                      |
| Calculation speed  | CalSpd-Avg-S            | N/A                  | CalSpd-Avg-D                      |
| Average knee angles | N/A                    | Knee-Avg-S           | Knee-Avg-D                        |
| Std. of knee angles | N/A                    | Knee-Std-S           | Knee-Std-D                        |

Analyzing the best performance over multiple trials is a practice used in the diagnosis of some diseases (e.g., the best walking speeds before and after a specific test are compared for the diagnosis of idiopathic normal-pressure hydrocephalus [18]). Hence, we considered the best performance in this study. However, the worst performance over multiple trials is also meaningful because subjects with lower MMSE scores tend to have unstable performance in the dual-task and may perform surprisingly poorly in one of the multiple trials, while subjects with higher MMSE scores are unlikely to perform poorly in any of the multiple trials. Therefore, we also considered the worst performance in this study.

In summary, we considered either the maximum or minimum values over multiple trials for each of the average and standard deviation feature and identified four patterns that contain all combinations of maximum or minimum values of average and standard deviation features, as shown in Tab. 2. We then determined the best combination from the four patterns, as well as the optimal number of trials through experiments, as explained in the Discussion section.

C. MMSE SCORE ESTIMATION MODEL

Given the cross-trial feature introduced in the above section, we applied a machine-learning algorithm to estimate MMSE scores. Previously [11], lower MMSE scores in older adults were detected based on single-trial features of performance data in a dual-task using several machine-learning algorithms such as the support vector machine, random forest, and neural networks. We found that LightGBM [25] achieved better results than previously compared algorithms and thus employed this framework in the present study. In our database, the MMSE scores were not evenly distributed among the subjects. Because most of the subjects were healthy people, there were fewer subjects with low MMSE scores compared with those with high MMSE scores. This imbalance could bias the trained model, leading the model to estimate higher MMSE scores for all samples. To address this, we augmented the samples so that the number of samples for each MMSE score was equal to the maximum number of samples among MMSE scores. We employed SMOTE [26] for this purpose, which chooses augmentation-target training samples randomly and finds the nearest neighbors of the chosen sample to generate new data. SMOTE is known to produce better results than those techniques using simple over-sampling by duplicating data points [26]. Notably, we could obtain different sets of augmented training samples because of the random property of this algorithm. Taking this property into account, we tried to estimate more reliable MMSE scores via aggregation.

Note that the augmented samples obtained via SMOTE were only used as training data. For testing, we used the original samples. In addition, to ensure that the test dataset
was independent from the training dataset, we used a leave-one-subject-out technique for training. Specifically, we used all real samples from one subject for testing, and samples (including the augmented ones) from the other subjects for training. Given a training set \( M \) of feature vectors and MMSE scores, we ran SMOTE \( N_{\text{SMOTE}} \) times to obtain \( N_{\text{SMOTE}} \) sets of augmented training samples. Next, we trained the LightGBM using each augmented set to obtain \( N_{\text{SMOTE}} \) sets of model parameters \( \Theta \) of LightGBM, as summarized in Algorithm 1. After obtaining the set of model parameter \( \Theta \), given a test sample, that is, a cross-trial feature vector \( x \in \mathbb{R}^D \), we estimated an MMSE score for each of model parameter in the set \( \Theta \) and then returned an averaged MMSE, as summarized in Algorithm 2.

**D. LOWER MMSE SCORES DETECTION MODEL**

In clinical settings, an MMSE score equal to or less than 23 indicates a relatively high probability of dementia, and an MMSE score equal to or less than 27 indicates a relatively high probability of MCI. In this study, we considered two types of binary classification problems: (1) subjects whose MMSE scores were less than or equal to 27 (we use MMSE \( \leq 27 \) in the following text for simplicity, and similar conventions are used for the other classes) vs. other subjects whose MMSE scores were larger than 27 (denoted as MMSE \( > 27 \)) and (2) subjects whose MMSE scores were less than or equal to 23 (denoted as MMSE \( \leq 23 \)) vs. other subjects whose MMSE scores were larger than 23 (denoted as MMSE \( > 23 \)). We assigned ground-truth positive/negative labels based on the ground-truth MMSE scores.

One straightforward way to evaluate classification (i.e., lower-MMSE-scores detection) performance is to employ the pre-defined thresholds also for the estimated MMSE scores. However, the estimated MMSE scores may be biased against the ground-truth MMSE scores because the pre-defined thresholds were not optimized. We therefore tuned the threshold to mitigate the bias. Matsuura et al. [11] tuned the threshold from 18 to 30 to find the best performance on test samples and observed an upper limit in terms of the classification performance of their method. However, this kind of tuning [11] was based on test samples and cannot be easily applied in a realistic setting. Unlike in previous research, the proposed method in the present study enabled us to tune thresholds using a validation set based on a leave-one-subject-out framework, shown in Algorithm 3. As when we used the leave-one-subject-out technique for training, we also used it to make the validation data independent from the training data. Through this technique, we obtained independence among the training, validation, and test data. Moreover, the proposed method included a two-stage tuning process to improve accuracy, as explained in detail in the following section.

We first prepared an original training set of \( D \) quadruplets. Each quadruplet \((x, s_{\text{GT}}, p_{\text{ID}}, y)\) was composed of a feature vector \(x \in \mathbb{R}^D\), the ground-truth MMSE score \(s_{\text{GT}}\), a subject ID \(p_{\text{ID}}\), and the positive/negative label \(y \in \{\text{positive}, \text{negative}\}\). We then selected a quadruplet of validation samples \((x, s_{\text{GT}}, p_{\text{ID}}, y)\) from \( D \) to obtain an exclusive training set \(M_{\text{ex}}\) containing pairs of a feature vector \(x\) and a ground-truth MMSE score, where the validation subject \(p_{\text{ID}}\) was excluded. We then trained a LightGBM model with the exclusive training set \(M_{\text{ex}}\) and estimated MMSE scores for the validation sample. We repeated this with new validation samples to obtain a set \(D_{\text{pair}}\) of pairs of estimated MMSE scores \(s\) and positive/negative labels \(y\). After obtaining set \(D_{\text{pair}}\), given a certain threshold \(t\) for the detection of lower MMSE scores,

---

**Algorithm 1 Training of an MMSE Score Estimator**

```plaintext
Algorithm 1: Training of an MMSE Score Estimator

```function` TRAINMMSEESTIMATOR(M, N_{\text{SMOTE}})
```
\[ \Theta = \phi \]
\[ \text{for } i = 1 \text{ to } N_{\text{SMOTE}} \text{ do} \]
\[ M_{\text{SMOTE}} \leftarrow \text{SMOTE}(M) \]
\[ \Theta \leftarrow \text{TrainLightGBM}(M_{\text{SMOTE}}) \]
\[ \Theta \leftarrow \Theta + \{\theta\} \]
\[ \text{end for} \]
\[ \text{return } \Theta \]
```

---

**Algorithm 2 MMSE Score Estimation**

```plaintext
Algorithm 2: MMSE Score Estimation

```function` ESTIMATEMMSE(x, \Theta)
```
\[ S = \phi \]
\[ \text{for each model parameter } \theta \in \Theta \text{ do} \]
\[ s \leftarrow \text{EstimateLightGBM}(x, \theta) \]
\[ S \leftarrow S + \{s\} \]
\[ \text{end for} \]
\[ \text{return } \left( \sum_{s \in S} s / |S| \right) \]
```

---

**Algorithm 3 Lower MMSE Scores Detection Model**

```plaintext
Algorithm 3: Lower MMSE Scores Detection Model

```function` DETECTMMSE(x, \Theta, D_{\text{valid}})
```
\[ \text{for each model parameter } \theta \in \Theta \text{ do} \]
\[ s_{\text{GT}} \leftarrow \text{EstimateGroundTruthMMSE(x)} \]
\[ s_{\text{pred}} \leftarrow \text{EstimateLightGBM}(x, \theta) \]
\[ \text{end for} \]
\[ \text{return } \left( \sum_{s \in S} s / |S| \right) \]
```

---

**Algorithm 4 One-Subject-Out Framework**

```plaintext
Algorithm 4: One-Subject-Out Framework

```function` ONSMOUTERMSE(x, \Theta, D_{\text{valid}})
```
\[ \text{for each model parameter } \theta \in \Theta \text{ do} \]
\[ s_{\text{GT}} \leftarrow \text{EstimateGroundTruthMMSE(x)} \]
\[ s_{\text{pred}} \leftarrow \text{EstimateLightGBM}(x, \theta) \]
\[ \text{end for} \]
\[ \text{return } \left( \sum_{s \in S} s / |S| \right) \]
```

---

**Algorithm 5 Leave-One-Subject-Out Framework**

```plaintext
Algorithm 5: Leave-One-Subject-Out Framework

```function` LONSMOUTERMSE(x, \Theta, D_{\text{valid}})
```
\[ \text{for each model parameter } \theta \in \Theta \text{ do} \]
\[ s_{\text{GT}} \leftarrow \text{EstimateGroundTruthMMSE(x)} \]
\[ s_{\text{pred}} \leftarrow \text{EstimateLightGBM}(x, \theta) \]
\[ \text{end for} \]
\[ \text{return } \left( \sum_{s \in S} s / |S| \right) \]
```

---

**Algorithm 6 Two-Stage Tuning Framework**

```plaintext
Algorithm 6: Two-Stage Tuning Framework

```function` TSTOUTERMSE(x, \Theta, D_{\text{valid}})
```
\[ \text{for each model parameter } \theta \in \Theta \text{ do} \]
\[ s_{\text{GT}} \leftarrow \text{EstimateGroundTruthMMSE(x)} \]
\[ s_{\text{pred}} \leftarrow \text{EstimateLightGBM}(x, \theta) \]
\[ \text{end for} \]
\[ \text{return } \left( \sum_{s \in S} s / |S| \right) \]
```

---

**Algorithm 7 One-Stage Tuning Framework**

```plaintext
Algorithm 7: One-Stage Tuning Framework

```function` OSTOUTERMSE(x, \Theta, D_{\text{valid}})
```
\[ \text{for each model parameter } \theta \in \Theta \text{ do} \]
\[ s_{\text{GT}} \leftarrow \text{EstimateGroundTruthMMSE(x)} \]
\[ s_{\text{pred}} \leftarrow \text{EstimateLightGBM}(x, \theta) \]
\[ \text{end for} \]
\[ \text{return } \left( \sum_{s \in S} s / |S| \right) \]
```

---

**Algorithm 8 Cross-Stage Tuning Framework**

```plaintext
Algorithm 8: Cross-Stage Tuning Framework

```function` CXTOUTERMSE(x, \Theta, D_{\text{valid}})
```
\[ \text{for each model parameter } \theta \in \Theta \text{ do} \]
\[ s_{\text{GT}} \leftarrow \text{EstimateGroundTruthMMSE(x)} \]
\[ s_{\text{pred}} \leftarrow \text{EstimateLightGBM}(x, \theta) \]
\[ \text{end for} \]
\[ \text{return } \left( \sum_{s \in S} s / |S| \right) \]
```
we could compute the numbers of true positive, false positive, true negative, and false negative samples as \( N_{TP}(D_{pair}; t) \), \( N_{FP}(D_{pair}; t) \), \( N_{TN}(D_{pair}; t) \), and \( N_{FN}(D_{pair}; t) \) via binary classification of set \( D_{pair} \) based on the threshold \( t \). We then computed an objective function \( f(D_{pair}; t) \), that is, the summation of a sensitivity, a.k.a. true positive rate, and specificity, a.k.a. true negative rate, as

\[
\begin{align*}
\text{Sen}(D_{pair}; t) &= \frac{N_{TP}(D_{pair}; t)}{N_{TP}(D_{pair}; t) + N_{FN}(D_{pair}; t)} \\
\text{Spec}(D_{pair}; t) &= \frac{N_{TN}(D_{pair}; t)}{N_{FP}(D_{pair}; t) + N_{TN}(D_{pair}; t)}
\end{align*}
\]

Finally, the optimal threshold was selected so as to maximize the objective function as

\[
t^*(D_{pair}) = \arg \max_t f(D_{pair}; t).
\]

Our optimization process can be described as a grid search, as summarized in Algorithm 3. For better efficiency, we further applied a two-stage coarse-to-fine grid search to optimize the parameter. Specifically, we first roughly tuned the threshold using a coarse search step \( \Delta t_{\text{coarse}} \) and then fine-tuned the threshold using a fine search step \( \Delta t_{\text{fine}} \), as shown in Algorithm 4. Once we obtained the optimal threshold \( t^* \), as well as the set of model parameters \( \Theta \), we could classify a test sample \( x \) as positive or negative by estimating an MMSE score and applying binary classification with the threshold \( t^* \), as summarized in Algorithm 5.

E. ETHICS DECLARATIONS

This study was approved by the Research Ethics Committee of SANKEN (The Institute of Scientific and Industrial Research), Osaka University (Osaka, Japan) under the authorization number H29-10. All experiments were performed in accordance with the appropriate guidelines and regulations. Informed consent was obtained from all subjects. For subjects whose MMSE score was under 23, the informed consent was also obtained from their next of kin/legally authorized representative.

III. EXPERIMENTS

In this section, we describe how we evaluated the proposed approach for detecting lower MMSE scores in older adults by conducting experiments based on data collected from our dual-task system.

A. EXPERIMENTAL SETUP

The experiments in this study were implemented based on data from three facilities for older adults, whose MMSE scores were provided by professional doctors. In total, data
from 883 trials were collected from 38 adults older than 65. Although we set several difficulty levels for our dual-task experience system, we only used data collected at the most difficult level to create the database. This was because performance in more difficult dual-task could more clearly distinguish subjects with low MMSE score from those with high MMSE score. This analysis is described more fully in the Discussion section.

The statistics describing our database are shown in Figure 3. Because each subject could use our dual-task system multiple times per day, we could obtain multiple trials of data from the same subject. From Figure 3, we can see that the low-MMSE-score-region has far fewer samples than the high-MMSE-score-region. This severe imbalance in the data made training difficult. To address this, we used the SMOTE data augmentation method and designed adaptive thresholds using validation data to improve the binary classification performance.

As part of the experimental protocol, we employed the leave-one-subject-out technique as mentioned above. As performance measures, we employed sensitivity, specificity, the summation of the sensitivity and specificity, and the F-measure. The sensitivity and specificity are defined as “Sen” and “Spec” in Eqs. (2) and (3). The precision (Pres) and F-measure were defined as follows:

$$\text{Pres}(D_{\text{pair}}, t) = \frac{N_{TP}(D_{\text{pair}}; t)}{N_{FP}(D_{\text{pair}}; t) + N_{TP}(D_{\text{pair}}; t)}$$

$$\text{F-measure} = \frac{2 \times \text{Sen} \times \text{Pres}}{\text{Sen} + \text{Pres}}.$$  \hspace{1cm} (6)

As for hyper-parameters, we set $N_{\text{SMOTE}} = 10$ of Algorithm 1 because the average of 10 implementations was sufficiently stable. Similarly, we set $t_{\text{max}} = 28.5$, $t_{\text{min}} = 23$, $\Delta t_{\text{course}} = 0.5$, and $\Delta t_{\text{fine}} = 0.1$ to tune the proper threshold for binary classification, as shown by the two-stage threshold tuning process in Algorithm 4.

**B. RESULTS**

In this section, we report the results of the ablation studies and sensitivity analyses in terms of five aspects: (1) four patterns of the cross-trial feature described in Tab. 2; (2) the number $N$ of continuous trials for the cross-trial feature; (3) presence and absence of threshold tuning; (4) comparison of different machine learning models; and (5) comparison of different methods for solving data imbalance. Note that bold numbers in the following tables denote the best results.

**C. COMPARISON OF FOUR PATTERNS**

Table 3 shows the result of the ablation study on different patterns of cross-trial features. In this experiment, we set the number of continuous trials at $N = 4$. Therefore, we can clearly see that pattern A outperforms the other patterns for all the criteria and for both the detection of subjects whose MMSE $\leq 23$ and the detection of subjects whose MMSE $\leq 27$. We therefore employed pattern A in the following experiments.

**D. COMPARISON OF DIFFERENT VALUES OF N**

Table 4 shows the results of the sensitivity analysis of the number $N$ of continuous trials by changing $N$ from 1 to 9. We compared the proposed method with Matsuura’s approach which used single-trial features and neural networks to predict MMSE scores [11]. The results in Tab. 4 showed that the proposed approach outperformed Matsuura’s method for every $N$ value. The summation of sensitivity and specificity reached 1.79 for the detection of older adults whose MMSE scored equal to or less than 23 when $N = 4$, and 1.75 for the detection of older adults whose MMSE scores were equal to or less than 27 when $N = 3$ (these were the best performances). Compared with the improved version of the existing method [11] ($N = 1$ in Tab. 4; applying LightGBM and SMOTE to the original method of [11]), the proposed approach achieved an improvement of 11.88% for detecting subjects whose MMSE $\leq 23$, and 8.70% for...
TABLE 3. Performances among four patterns of the cross-trial feature.

| Categories                              | Measures         | Pattern A | Pattern B | Pattern C | Pattern D |
|-----------------------------------------|------------------|-----------|-----------|-----------|-----------|
| Detection of subjects                   |                  |           |           |           |           |
| whose MMSE ≤ 23                         | Sensitivity      | 0.94      | 0.75      | 0.87      | 0.71      |
|                                         | Specificity      | 0.85      | 0.83      | 0.85      | 0.63      |
|                                         | Sensitivity+Specificity | 1.79 | 1.58      | 1.42      | 1.36      |
|                                         | F-measure        | 0.79      | 0.67      | 0.55      | 0.52      |
| Detection of subjects                   |                  |           |           |           |           |
| whose MMSE ≤ 27                         | Sensitivity      | 0.87      | 0.72      | 0.87      | 0.73      |
|                                         | Specificity      | 0.86      | 0.83      | 0.96      | 0.68      |
|                                         | Sensitivity+Specificity | 1.73 | 1.55      | 1.43      | 1.41      |
|                                         | F-measure        | 0.79      | 0.67      | 0.59      | 0.58      |

TABLE 4. Performance of the proposed method with different numbers of continuous trials.

| Categories                              | Methods          | Sensitivity | Specificity | Sensitivity+Specificity | F-measure |
|-----------------------------------------|------------------|-------------|-------------|------------------------|-----------|
| Detection of subjects                   | Matsuara’s Method [11] |             |             |                        |           |
| whose MMSE ≤ 23                         |                  | 0.98        | 0.47        | 1.45                   | 0.44      |
|                                         | Proposed         | N = 1       | 0.88        | 0.72                   | 1.60      | 0.65      |
|                                         |                  | N = 2       | 0.90        | 0.85                   | 1.75      | 0.77      |
|                                         |                  | N = 3       | 0.92        | 0.85                   | 1.77      | 0.78      |
|                                         |                  | N = 4       | 0.94        | 0.85                   | 1.79      | 0.79      |
|                                         |                  | N = 5       | 0.93        | 0.84                   | 1.77      | 0.77      |
|                                         |                  | N = 6       | 0.97        | 0.71                   | 1.68      | 0.69      |
|                                         |                  | N = 7       | 0.92        | 0.77                   | 1.69      | 0.73      |
|                                         |                  | N = 8       | 0.94        | 0.70                   | 1.64      | 0.68      |
|                                         |                  | N = 9       | 0.94        | 0.74                   | 1.68      | 0.71      |
| Detection of subjects                   | Matsuara’s Method [11] |             |             |                        |           |
| whose MMSE ≤ 27                         |                  | 0.87        | 0.56        | 1.43                   | 0.54      |
|                                         | Proposed         | N = 1       | 0.93        | 0.68                   | 1.61      | 0.70      |
|                                         |                  | N = 2       | 0.84        | 0.89                   | 1.73      | 0.80      |
|                                         |                  | N = 3       | 0.88        | 0.87                   | 1.75      | 0.80      |
|                                         |                  | N = 4       | 0.87        | 0.86                   | 1.73      | 0.79      |
|                                         |                  | N = 5       | 0.87        | 0.86                   | 1.73      | 0.78      |
|                                         |                  | N = 6       | 0.93        | 0.71                   | 1.64      | 0.71      |
|                                         |                  | N = 7       | 0.89        | 0.74                   | 1.66      | 0.73      |
|                                         |                  | N = 8       | 0.89        | 0.71                   | 1.60      | 0.68      |
|                                         |                  | N = 9       | 0.90        | 0.74                   | 1.64      | 0.71      |

FIGURE 4. Detection of subjects whose MMSE ≤ 23: ROC curves among different numbers of continuous trials. The black dotted line shows the chance level.

detecting subjects whose MMSE ≤ 27 when N = 4 and N = 3, with respect to the summation of sensitivity and specificity. This demonstrates that the cross-trial feature was effective in detecting low cognitive status. We computed the F-measure for each compared method, and found that N = 4 and N = 3 achieved the best F-measures for detecting MMSE ≤ 23 and MMSE ≤ 27, respectively. In addition, the receiver operating characteristics (ROC) curves, shown in Figs. 4 and 5, indicate that there was a trade-off between the false positive rate and the true positive rate when changing the threshold. Here, we show the ROC curves for only a subset of the continuous trials N = {3, 4, 5, 6} for better clarity.

E. COMPARISON OF PERFORMANCES WITH AND WITHOUT THRESHOLD TUNING

The results with and without threshold tuning are shown in Tab. 5. We used the pre-defined thresholds of 23 and 27 for the predicted MMSE scores in the case without threshold tuning. We found that the tuned threshold yielded better specificity, while the pre-defined threshold yielded better
sensitivity. Overall, the tuned threshold yielded better overall performance (i.e., sensitivity + specificity), and hence we can confirm that the tuned threshold achieved a better trade-off between sensitivity and specificity.

F. COMPARISON OF MACHINE LEARNING MODELS

We compared the performance observed using LightGBM with that using Long Short Term Memory (LSTM) [31] on our database, as shown in Tab. 6. Because LSTM training is time-consuming (the time cost of LSTM is about 80 times that of LightGBM), we set $N_{SMOTE} = 1$ in Algorithm 1 for both LightGBM and LSTM. Through comparison, we found that the training of LightGBM is much faster than that of LSTM. Moreover, LightGBM achieved better experimental results than LSTM. This may be because the size of the current database is too small to obtain complete learning with LSTM. Either way, the data indicate that LightGBM is more suitable for the current database.

G. COMPARISON OF METHODS FOR SOLVING THE DATA IMBALANCE PROBLEM

In this study, we first regressed the MMSE scores, and then implemented the binary classification by thresholding. The data imbalance problem occurred during the regression part of the process, i.e., the number of samples for each MMSE score was not identical (MMSE scores are integers from 0 to 30). Data re-sampling is frequently used to deal with data imbalances [26]. Among data re-sampling approaches, there are two categories: under-sampling and over-sampling methods. RUS (Random Under Sampling) is a typical under-sampling method while SMOTE, which is the approach we used, is a typical over-sampling method. Here, we compare the performances of SMOTE [26] and RUS [32]. SMOTE increases the size of the dataset to make the number of samples for each MMSE score equal to the maximum number of samples among MMSE scores [26]. In contrast, RUS deletes data randomly to make the number of samples for each MMSE score equal to the minimum number of samples among MMSE scores [32]. As a result, the number of samples obtained using SMOTE is generally about 200 times that obtained using RUS. Because the current database was small (only 38 subjects), the number of samples obtained using RUS was too insufficient to evaluate learning performance. Accordingly, the performance of RUS was far worse than that of SMOTE, as shown in Tab. 7. The performance of RUS is expected to be improved if the database is enlarged in future.

To investigate the degree to which the bias caused by the difference in the number of samples per subject or per MMSE score influenced the result, we trained a model by augmenting the training data so as to make the number of samples for each subject identical (termed “Augmentation based on subjects”). We then compared this model with our original augmentation where the number of samples for each MMSE was identical (termed “Augmentation based on MMSE”). Note that the number of samples for each MMSE could be different for “Augmentation based on subject” while the number of samples for each subject could be different for “Augmentation based on MMSE.” The comparison results are shown in Tab. 7, which indicates that SMOTE based on MMSE achieved better performance. The comparison of performance revealed that the influence of the imbalance caused by the number of samples for each subject was minor compared with that caused by the number of samples for each MMSE score. Therefore, the MMSE-imbalance problem was addressed in the first priority.

IV. DISCUSSION

In this section, we discuss (1) the explanation for why pattern A had the best performance, (2) feature importance and partial dependency of the used machine learning model LightGBM, (3) how to determine the value of $N$ according to different applications, (4) how the difficulty level of the dual-task influences locomotion and cognitive features, (5) the advantages and limitations of MMSE-based cognitive
impairment screening, and (6) peer comparison of the dual-task design.

A. SUPERIORITY OF PATTERN A

First, we discuss the results for the four patterns of cross-trial features, as shown in Tab. 3. As we have already confirmed in the previous section, pattern A, that is, the combination of maximum values of average features and maximum values of standard deviation features, achieved the best performance.

In terms of the average features, healthy people usually have faster response speeds and higher correct-answer-rates than those with lower MMSE scores. However, the values of these features could be slightly lower than usual if the subjects were influenced by some non-cognitive factors, such as a lack of familiarity with the system, emotional state, and so on. Therefore, the maximum values from several continuous trials can represent the true ability of healthy people. For people with lower MMSE scores, although their performance can vary sharply among trials, their best performance is still generally lower than that of healthy people. In other words, a subject is more likely to have cognitive impairment if he/she cannot achieve good dual-task performance when exerting the greatest amount of effort or when in the best-case situation. This explains why choosing the maximum average features can be efficient for detecting lower MMSE scores in older adults.

For standard deviation features, the gait of people with lower MMSE scores is expected to deteriorate when they focus on calculation questions, resulting in larger standard deviation values. This may occur even if they attempt the dual-task many times. Conversely, healthy people with high MMSE scores can generally step in a stable manner while engaged in computation, which leads to much smaller values of standard deviation features. By selecting the maximum values of standard deviation features (i.e., the worst ones) from multiple continuous trials, the tendencies that differ between healthy people and people with lower MMSE scores can be highlighted.

B. FEATURE IMPORTANCE AND PARTIAL DEPENDENCE

The factors that influence the performance of a specific machine learning model can be elusive. To address this issue, efforts to illustrate machine learning models have become more and more popular in recent years. Feature importance and Partial dependence plot (PDP) are two analytic criteria for illustrating the mechanisms of machine learning models [33], [34]. Feature importance shows how much a certain feature influences the performance of the machine learning model. PDP illustrates the relationship between the feature and the regression/classification target (e.g., linear or non-linear) [33], [34]. Here we use the two criteria to analyze the twelve basic features (shown in Tab. 1) used in the proposed approach.

Figure 6 shows the importance of all twelve basic features. From Fig. 6, it is apparent that the features of dual-task play a more important role in the regression of MMSE scores when compared with the features of single task. During the dual-task in the present study, the calculation speed, mean stepping speed, and standard deviation of the stepping speed were the three most important features. Next, we used PDP to illustrate how the three features influenced the predicted MMSE score. Figures 7, 8, and 9 show the PDP for the three features. The blue area in each figure shows the confidence interval. The number of grid points is 100 after interpolation. With respect to feature “CalSpd-Avg-D” shown in Fig. 7, we found that the predicted MMSE score increased when “CalSpd-Avg-D” increased from 0.63 to 1.35. This is expected because people with a lower MMSE score are likely to encounter greater difficulty during calculation. When the value of the feature exceeded 1.35, its influence on the predicted MMSE score approached 0. This demonstrated that the predicted MMSE score reached the maximum point when the value of feature “CalSpd-Avg-D” was larger than 1.35.

| Categories | Augmentation methods | Sensitivity | Specificity | Sensitivity×Specificity | F-measure |
|------------|----------------------|-------------|-------------|------------------------|-----------|
| Detection of subjects whose MMSE ≤ 23 | Augmented by RUS | 1.00 | 0.00 | 1.00 | 0.41 |
| | Augmented by SMOTIE (based on subjects) | 0.75 | 0.92 | 1.67 | 0.75 |
| | Augmented by SMOTIE (based on MMSE) | 0.94 | 0.85 | 1.79 | 0.79 |
| Detection of subjects whose MMSE ≤ 27 | Augmented by RUS | 0.97 | 0.05 | 1.02 | 0.45 |
| | Augmented by SMOTIE (based on subjects) | 0.70 | 0.93 | 1.63 | 0.74 |
| | Augmented by SMOTIE (based on MMSE) | 0.87 | 0.86 | 1.73 | 0.79 |
FIGURE 7. Partial dependence plot for feature “CalSpd-Avg-D” (Most important).

FIGURE 8. Partial dependence plot for feature “Step-Avg-D” (Second-most important).

FIGURE 9. Partial dependence plot for feature “Step-Std-D” (Third-most important).

With respect to feature “Step-Avg-D,” we found that its influence on the predicted MMSE score became zero when the value of the feature increased above 2.0. When “Step-Std-D” increased from 1.7 to 2.0, the predicted MMSE score increased in major areas and decreased in minor areas. Some of the people with cognitive impairment might have been relatively young, and thus able to engage in faster stepping. Nevertheless, for most subjects, faster stepping was related to higher MMSE scores. With respect to feature “Step-Std-D” shown in Fig. 9, the predicted MMSE score decreased as the standard deviation increased from 0.05 to 0.55. This is consistent with our claim that people with lower MMSE scores are more likely to exhibit unstable performance during multiple trials. When the value of the standard deviation exceeded 0.55, the predicted MMSE score reached the minimum value.

C. VALUE OF N FOR DIFFERENT APPLICATIONS

Here, we discuss the number of continuous trials based on the results in Tab. 4 as well as Figs. 4 and 5. These experimental results show that the value of $N$ could vary depending on different applications. For example, to maximize the sum of sensitivity and specificity, we could choose $N = 4$ and $N = 3$ to detect subjects whose MMSE score $\leq 23$ and subjects whose MMSE score $\leq 27$, respectively. However, a greater focus on sensitivity as opposed to specificity, e.g., in a system examining intelligence as a criteria for renewing driving licenses, a larger $N$ may be more suitable. We can see this point from Figs. 4 and 5, where the dark blue curves ($N = 5$) show the best specificity when the sensitivity is fixed at 0.9999 for detecting subjects whose MMSE $\leq 23$ and at 0.9625 for detecting subjects whose MMSE $\leq 27$. In some applications where the sensitivity must be higher than 0.99 for detecting subjects whose MMSE $\leq 23$, $N = 5$ is better than $N = 4$. Similarly, for applications where the sensitivity is required to be higher than 0.95 for detecting subjects whose MMSE $\leq 27$, $N = 5$ is better than $N = 3$. In summary, the optimal $N$ value varies according to the application. In other words, the $N$ parameter can be manipulated to make the proposed approach suitable for various systems.

However, larger $N$ value usually provides more information for detecting subjects with lower MMSE scores. The experimental results showed that the largest value of $N$ did not result in the best performance. This raises an interesting question: the degree to which the basic features are statistically reliable in terms of their dependency on the value of $N$. To illustrate this issue, we evaluated the statistical reliability of the features used in terms of their dependency on the number $N$ of sequential trials per input data. More specifically, we computed the intra-subject feature deviations over multiple sequential trials for the same subject and showed their transitions when $N$ was increased from 2 to 9.

Figure 10 (a) and (b) show the feature deviations for mean and std features, respectively. Note that the statistical reliability increases as the feature deviation decreases in these graphs. The results show that the statistical reliability increases as the $N$ increases in a near-monotonic fashion, which is consistent with our expectation. Moreover, it is apparent that the rate of decrease in the feature deviation is the largest when $N = 3$ or $N = 4$ in many cases, which means that the gain in statistical reliability is largest when $N = 3$ or $N = 4$. However, the number $N$ of sequential trials...
per unit of inputted data also affects the number of training samples. More specifically, if a specific subject completes $M$ trials in total, we can extract $M - N + 1$ training samples of sequential trials via sliding window-based sampling. This means that the number of training samples decreases as $N$ increases. Generally speaking, both the increase in the statistical reliability of the features by increasing $N$ and the increase in the number of training samples by decreasing $N$ contribute to the improved performance. Hence, the optimal $N$ is determined based on a trade-off between the above-mentioned two aspects. Consequently, the case of $N = 3$ or $N = 4$ yielded the best performance because these were associated with the largest gain in statistical reliability in our experiments. However, if the number $M$ of total trials per subject increased, the decreasing ratio of the number of training samples caused by the increase of $N$ would be little, making the increase in statistical reliability via increasing $N$ more favorable. We plan to continue this line of inquiry by further developing our database in future work.

**D. INFLUENCE OF DIFFICULTY LEVELS**

We considered how the difficulty level of the calculation questions in the dual-task influenced the measured locomotion and cognitive features. To demonstrate the different performance of the same feature under different difficulty levels, we show a comparison of the results from the easy to hard levels in Figure 11, taking the features Step-Std-D and CalSpd-Avg-D in Tab. 1 as examples. The figure illustrates the distributions of the two features under different ground-truths of the MMSE scores. From both graphs in Figure 11, we found that the average of the features tended to be consistently low in the region with a high MMSE score (larger than 28). Comparing the curves of the two difficulty levels, the difference in features between high MMSE score regions and low MMSE score regions was larger under the difficult level compared with the easy level. This means that the difficult level more easily highlights the differences between subjects with high MMSE scores and those with low MMSE scores. Therefore, we used the data from the most difficult level in this study.

**E. LIMITATION AND MERITS OF MMSE**

The MMSE is widely used to screen for cognitive impairment [35]. Nevertheless, clinical diagnoses of dementia or MCI should be based on multiple examinations, such as the CDRSB (Clinical Dementia Rating Sum of Boxes), MRI (Magnetic Resonance Imaging), MMSE, and so on. After a classification has been made based on MMSE scores, further clinical examinations are required prior to diagnosis. The sensitivity and specificity of the MMSE was reported to be 0.84 and 0.86 for detecting dementia, and 0.88 and 0.70 for detecting MCI [35]. Although the sensitivity and specificity of the MMSE are not high, it is still a useful tool in dementia prevention because of its convenience and low cost [36]–[38]. However, it cannot be applied to daily monitoring because it has fixed questions. The goal of the proposed dual-task-based approach was to address this limitation of the MMSE and achieve a system that enables convenient daily monitoring.

From the perspective of deep learning, which requires a large training dataset, data from clinical cases with definite diagnoses of dementia or MCI are not likely to be sufficient for training a learning model. In contrast, the collection of massive datasets of MMSE scores is relatively easy to implement. Indeed, MMSE scores should be supplemented with other measures when assessing low cognitive status. Nevertheless, the proposed system could achieve high accuracy for automatic monitoring of cognitive status because the collection of massive training datasets based on MMSE scores is relatively inexpensive.

The identification of individuals with cognitive impairment would be more accurate if MMSE scores were considered along with other examination results such as CDRSB and MRI data. Because this will require a new approach for...
data collection, we plan to address it in our future work. Specifically, we hope to refine the MMSE-based pre-trained model based on data regarding clinical diagnoses of dementia or MCI. The learning will be implemented in two steps: (1) coarse learning using massive datasets based on MMSE scores, and (2) refined learning using a small dataset based on clinical diagnoses.

F. DUAL-TASK DESIGN: A PEER COMPARISON

The design of dual-task paradigms is a popular research topic. A dual-task usually contains a motor task and a cognitive task. With respect to the cognitive task, those involving calculation, counting, and reciting are frequently used [11], [21]–[24]. With respect to the motor task, gait has been found to be highly related to cognitive impairment [18]–[20]. Besides the existing gait-based dual-task designs, Mancioppo et al. proposed two novel dual-tasks: (1) the FTAP (fore-finger tapping with cognitive task), and (2) the TTHP (toe tapping with cognitive task) [24]. They compared performance in these two new dual-tasks with MMSE scores and that in a gait-based dual-task, and found that the TTHP achieved the best performance (sensitivity + specificity = 1.84). The difference between the present paradigm and that of [24] is summarized in Tab. 8. Compared with the study by [24], the present paradigm has three major improvements:

1. The proposed system is totally automatic. Subjects can complete the examination without any staff present.
2. The cognitive questions are fixed in tasks used in many previous studies [24]. Thus, they cannot be reused frequently because the subjects may memorize the answers. In contrast, the questions in the cognitive task in the present study were created randomly. Therefore, the proposed dual-task paradigm can be implemented on a daily basis, i.e., it can be used as a daily tool to monitor cognitive status, as well as to screen for lower MMSE scores.
3. In a previous study [24], wearable sensors were used to conduct feature extraction. Although wearable sensors can facilitate high accuracy during feature analysis, they may cause discomfort for some people if their skin comes in contact with the sensors. Moreover, wearing and removing wearable sensors may be time-consuming for some subjects. Compared with [24], our proposed system is totally contactless. The subjects were required to step in place in the same location and answer some calculation questions. The total time cost of the whole process was 240s. In summary, our proposed system has fewer restrictions compared with wearable-sensor-based systems.

In addition, compared with [24] who only processed data for subjects with MMSE ≥24, the proposed system can be used to detect subjects whose MMSE score is ≤28 as well as subjects whose MMSE score is ≤24. For a similar gait-based dual-task, the proposed approach achieved a higher score (sensitivity + specificity = 1.75) than [24] when using three continuous trials. However, the experiments in [24] showed that the TTHP (not gait-based) achieved a higher score (1.84) compared with the gait-based dual-task. This may be related to the use of wearable sensors because they can be used to extract features more accurately than contact-less approaches. This evidence regarding the advantages of TTHP provides new directions for dual-task-based approaches. For instance, we may examine the use of a contact-less version of the TTHP with the proposed strategy in future research.

V. CONCLUSION

In this study, we proposed an approach for detecting subjects with lower MMSE scores. The data were collected using our dual-task experience system, which was installed in the facilities of older adults for long-term use. This system was designed to run without professional operators in attendance. Dual-task performance data could be automatically collected while the subjects performed the dual-task. In the proposed method, we designed one model to estimate MMSE scores from the collected dual-task performance data, and designed another model to detect subjects with low cognitive status based on the estimated MMSE scores. To guarantee high precision, we designed a cross-trial feature that fused all maximum features of several continuous trials. Our experimental results showed that the best performance (sum of sensitivity and specificity) reached 1.79 for detecting older adults whose MMSE scores were equal to or less than 27. To our knowledge, this study is the first to report that disturbance information of different trials is effective for the detection of lower MMSE scores in older adults, and our model achieved higher classification performance than those reported previously [11].

In our future work, we plan to extent the current research to include data from subjects who have received a diagnosis.

| Criteria                                      | Research [24] | The proposed paradigm |
|-----------------------------------------------|---------------|-----------------------|
| using wearable sensor                         | Yes           | No                    |
| Totally automatic (no need for a staff member present) | No            | Yes                   |
| Multiple types of dual-task                   | Yes           | No                    |
| Can be used for monitoring                    | No            | Yes                   |
| Best score (sensitivity+specificity) for screening MCI (MMSE≤ 28) | 1.84          | 1.75                  |
| Best score (sensitivity+specificity) for screening dementia (MMSE≤ 24) | Not conducted | 1.79                  |
| Dealing with data imbalance                   | No            | Yes                   |
| Using gait-based dual-task                    | Yes           | Yes                   |
| Best performance score (sensitivity+specificity) of gait-based dual-task for screening MCI | 1.73          | 1.75                  |
| Required walking space                        | 10 meters     | 1 meter               |
| Other feature as an input except dual-task data | Age           | None                  |
Also, we plan to compute the stepping speed using the Kinect instead of the force platform sensor because Kinect can collect more information and has fewer restrictions regarding setup. In addition, we plan to achieve miniaturization of the current system using a smartphone application.

ACKNOWLEDGMENT

The authors would like to thank the Misasagaki Social Welfare Foundation for giving them the opportunity to engage in data collection.

REFERENCES

[1] World Health Organization. (2019). Dementia. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia
[2] G. Livingston, A. Sommerlad, and V. Orgeta, “Dementia prevention, intervention, and care,” Lancet, vol. 390, no. 10113, pp. 2673–2734, 2017.
[3] L. Boise, R. Camicioli, D. L. Morgan, J. H. Rose, and L. Congleton, “Diagnosing dementia: Perspectives of primary care physicians,” Gerontologist, vol. 39, no. 4, pp. 457–464, Aug. 1999.
[4] I. Abrah, J. M. Rimland, F. M. Trotta, G. Dell’Aquila, A. Cruz-Jentoft, M. Petrovic, A. Gudmundsson, R. Soöza, D. O’Mahony, A. Guaita, and A. Cherburn, “Systematic review of systematic reviews of non-pharmacological interventions to treat behavioural disturbances in older patients with dementia. The SENATOR-OnTop series,” BMJ Open, vol. 7, no. 3, Mar. 2017, Art. no. e012759.
[5] M. Prince, A. Comas-Herrera, and M. Knapp, “World Alzheimer report 2016: Improving healthcare for people living with dementia: Coverage, quality and costs now and in the future,” Alzheimer’s disease International, vol. 39, no. 3, Mar. 2017, Art. no. e012759.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, 2002.
[7] F. Okura, I. Mitsugami, M. Niwa, K. Aoki, C. Zhou, and Y. Yagi, “[Paper] automatic collection of dementia subjects in single-task and dual-task with three different MIMU set-ups,” in Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA), Jun. 2021, pp. 1–6.
[8] E. Digo, E. Panero, V. Agostini, and L. Gastaldi, “Gait parameters of elderly subjects in single-task and dual-task with three different MIMU set-ups,” in Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA), Jul. 2020, pp. 3204–3207.
[9] Z. Nasreddine, N. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, J. K. Kueper, M. Speechley, and M. Montero-Odasso, “The Alzheimer’s disease assessment scale-cognitive subscale (ADAS-Cog): Modifications for diagnosing a dementia in highly educated population,” Clin. Neuropsychol., vol. 23, no. 2, pp. 189–197, 1987.
[10] 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., vol. 13, no. 4, pp. 209–213, Nov. 2000.
[11] 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, Dec. 2019, Art. no. 19927.
[12] V. C. Pangman, J. Sloan, and L. Guse, “An examination of psychometric properties of the mini-mental state examination,” J. Gerontol. A, vol. 60, no. 1, pp. 321–357, 2005.
[13] F. Okura, I. Mitsugami, M. Niwa, K. Aoki, C. Zhou, and Y. Yagi, “[Paper] automatic collection of dual-task human behavior for analysis of cognitive function,” TIE Trans. Media Technol. Appl., vol. 6, no. 2, pp. 138–150, 2018.
[14] A. Marterer, W. Danielczyk, M. Simaniy, and P. Fischer, “Calculation abilities in dementia of Alzheimer’s type and in vascular dementia,” Arch. Gerontol. Geriatrics, vol. 23, no. 2, pp. 189–197, Sep. 1996.
[15] T. Ijmker and C. J. C. Lamoth, “Gait and cognition: The relationship between gait stability and variability with executive function in persons with and without dementia,” Gait Posture, vol. 35, no. 1, pp. 126–130, Jan. 2012.
[16] F. Okura, M. C. Carlson, A. L. Fitzpatrick, L. H. Kuller, L. P. Fried, and P. P. Zandi, “Knee height and arm span: A reflection of early life environment and risk of dementia,” Neurology, vol. 70, no. 2, pp. 1818–1826, May 2008.
[17] H. Sepp and S. Jærgen, “Long short-term memory,” Neural Comput., vol. 9, no. 8, 1997, Art. no. 191504.
[18] Z. Mingzhu, X. Chao, and W. Yi-Fang, “IFME: Information filtering by multiple examples with under-sampling in a digital library environment,” in Proc. 13th ACM/ICDE/CEI Joint Conf. Digital Libraries, 2013, pp. 107–110.
[19] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Ann. Statist., vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[20] Q. Zhao and T. Hastie, “Causal interpretations of black-box models,” J. Bus. Econ. Statist., vol. 39, no. 1, pp. 272–281, Jan. 2021.
[21] I.-C. Tsai, C.-W. Chen, H. Chu, H.-L. Yang, M.-H. Chung, Y.-M. Liao, and K.-R. Chou, “Comparing the sensitivity, specificity, and predictive values of the Montreal cognitive assessment and mini-mental state examination when screening people for mild cognitive impairment and dementia in Chinese population,” Arch. Psychiatr. Nursing, vol. 39, no. 1, pp. 486–491, Aug. 2016.
[22] I. Arevalo-Rodriguez, N. Smailagic, M. Roqué i Figuls, A. Ciaponi, E. Sanchez-Pereze-A., G. Giannakou, O. L. Pedraza, X. B. Cos, and S. Cullum, “Mini-mental state examination (MMSE) for the detection of Alzheimer’s disease and other dementias in people with mild cognitive impairment (MCI),” BJPsych Adv., vol. 21, no. 6, p. 362, Nov. 2015.
[23] P. Pezzotti, “The accuracy of the MMSE in detecting cognitive impairment when administered by general practitioners: A prospective observational study,” BMC Family Pract., vol. 9, no. 1, pp. 1–11, Dec. 2008, doi: 10.1186/1471-2296-9-29.
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, majoring in computational intelligence and systems science, from the Tokyo Institute of Technology, Tokyo, Japan, in 2015. From 2015 to 2020, she was a Research Fellow with the Graduate School of Informatics, Kyoto University. Since 2020, she has been working as an Assistant Professor with SANKEN (The Institute of Scientific and Industrial Research), Osaka University. Her current research interests include biomedical signal processing, image processing, 3D reconstruction, pattern recognition, and machine learning.

TAKU MATSUURA received the M.S. degree from the Graduate School of Information Science and Technology, Osaka University, in 2020. His research topic is “cognitive impairment detection based on dual-task paradigm.” He is currently working on a company. His research interests include computer vision and machine learning.

FUMIO OKURA (Member, IEEE) received the M.S. and Ph.D. degrees in engineering from the Nara Institute of Science and Technology, in 2011 and 2014, respectively. He has been an Assistant Professor with SANKEN (The Institute of Scientific and Industrial Research), Osaka University, until 2020. He is currently an Associate Professor with the Graduate School of Information Science and Technology, Osaka University. His research interest includes the boundary domain between computer vision and computer graphics. He is a member of the IEICE, the IPSJ, and the VRSJ.

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 became a Research Associate, in 1990, a Lecturer, in 1993, an Associate Professor, in 1996, and a Professor, in 2003, with Osaka University, where he was the Director of SANKEN (The Institute of Scientific and Industrial Research), from 2012 to 2015. He was the Executive Vice President of Osaka University, from 2015 to 2019. His research interests include computer vision, human interfaces, and human behavior analysis. He is a member of the IEICE, the IPSJ, and the VRSJ.

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 became a Research Associate, in 1990, a Lecturer, in 1993, an Associate Professor, in 1996, and a Professor, in 2003, with Osaka University, where he was the Director of SANKEN (The Institute of Scientific and Industrial Research), from 2012 to 2015. He was the 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. He is a fellow of the IPSJ and a member of the IEICE and the RSJ. He is a member of the Editorial Board of the International Journal of Computer Vision. He is the Vice President of the Asian Federation of Computer Vision Societies. He was awarded 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 (2011 and 2013). International conferences for which he has served as the Chair include ROBIO2006 (PC), ACCV (2007PC and 2009GC), PSIVT2009 (PC), and ACPR (2011PC, 2013GC, 2021GC, and 2023GC). He has also served as an Editor for the IEEE ICRA Conference Editorial Board (2008 and 2011). He was the Editor-in-Chief of the IPSJ Transactions on Computer Vision and Applications.