Deep Learning-Based Multilevel Classification of Alzheimer's Disease Using Non-invasive Functional Near-Infrared Spectroscopy

Thi Kieu Khanh Ho[1], Minhee Kim[2], Younghun Jeon[3], Byeong C. Kim[4], Jae Gwan Kim[2], Kun Ho Lee[5,6,7], Jong-In Song[3] and Jeonghwan Gwak[1,8,9,10]*

[1] Department of Software, Korea National University of Transportation, Chungju, South Korea, [2] Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology, Gwangju, South Korea, [3] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, South Korea, [4] Department of Neurology, Chonnam National University Medical School, Gwangju, South Korea, [5] Gwangju Alzheimer's Disease and Related Dementias Cohort Research Center, Chosun University, Gwangju, South Korea, [6] Department of Biomedical Science, Chosun University, Gwangju, South Korea, [7] Korea Brain Research Institute, Daegu, South Korea, [8] Department of Biomedical Engineering, Korea National University of Transportation, Chungju, South Korea, [9] Department of AI Robotics Engineering, Korea National University of Transportation, Chungju, South Korea, [10] Department of IT and Energy Convergence (BK21 FOUR), Korea National University of Transportation, Chungju, South Korea

The timely diagnosis of Alzheimer's disease (AD) and its prodromal stages is critically important for the patients, who manifest different neurodegenerative severity and progression risks, to take intervention and early symptomatic treatments before the brain damage is shaped. As one of the promising techniques, functional near-infrared spectroscopy (fNIRS) has been widely employed to support early-stage AD diagnosis. This study aims to validate the capability of fNIRS coupled with Deep Learning (DL) models for AD multi-class classification. First, a comprehensive experimental design, including the resting, cognitive, memory, and verbal tasks was conducted. Second, to precisely evaluate the AD progression, we thoroughly examined the change of hemodynamic responses measured in the prefrontal cortex among four subject groups and among genders. Then, we adopted a set of DL architectures on an extremely imbalanced fNIRS dataset. The results indicated that the statistical difference between subject groups did exist during memory and verbal tasks. This presented the correlation of the level of hemoglobin activation and the degree of AD severity. There was also a gender effect on the hemoglobin changes due to the functional stimulation in our study. Moreover, we demonstrated the potential of distinguished DL models, which boosted the multi-class classification performance. The highest accuracy was achieved by Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) using the original dataset of three hemoglobin types (0.909 ± 0.012 on average). Compared to conventional machine learning algorithms, DL models produced a better classification performance. These findings demonstrated the capability of DL frameworks on the imbalanced class distribution analysis and validated the great potential of fNIRS-based approaches to be further contributed to the development of AD diagnosis systems.

Keywords: Alzheimer's disease, fNIRS, multi-class classification, deep learning – artificial neural network (DL-ANN), CNN-LSTM

INTRODUCTION

The incidence of age-related dementia has increased dramatically as the world population is aging. Dementia was estimated to affect 50 million people worldwide in 2018 and is expected to exceed 150 million people within 30 years (Alzheimer's Disease International, 2018). Alzheimer's disease (AD) is one of the most well-known causes of dementia, accounting for nearly two-thirds of all dementia patients (Alzheimer's Association, 2019). AD is characterized by the progressive impairment of both cognitive and memory abilities due to a consequence of the significant loss of neurons in the nervous system (Veitch et al., 2019) or disruption of nerve cell communications by the presence of extracellular Amyloid-Beta (Aβ) peptide plaques and neurofibrillary tangles (NFT), which typically develop decades before the symptoms manifest (Lowe et al., 2019). Since AD first destroys brain cells related to language and memory brain regions, patients suffer from confusion, memory loss, speech impediment, poor problem-solving skills, and difficulty in daily communication. Then, other brain regions for controlling breathing and heart functionality would deteriorate, eventually leading to death. Patients with AD are generally diagnosed in late stages, by which point existing treatments can only decelerate the speed of cognitive declines. As such, a prompt diagnosis to facilitate proper treatments of preclinical and its prodromal stages is significantly crucial. Besides fundamental exams, such as patient interview, Mini-Mental State Examination (MMSE), and physical and neurobiological exams, reliable and powerful neuroimaging techniques, such as from functional magnetic resonance imaging (fMRI) (Cheng et al., 2019), structural magnetic resonance imaging (sMRI) (Wang et al., 2016), positron-emission tomography (PET) (Alster et al., 2019), electroencephalography (EEG) (Mazaheri et al., 2018), or diffusion tensor imaging (DTI) (Hwang et al., 2019) techniques, are desperately needed to furnish more informative diagnosis approach. However, these techniques usually appeal to the clinical interaction between doctors and patients, whilst a more flexible paradigm of doctor-patient interaction is preferable.

Functional near-infrared spectroscopy (Naseer and Hong, 2015) is a scalp-located and non-invasive technique which measures the neural activity using the concentration of oxygenated (HbO) and deoxygenated (Hb) hemoglobin in the brain by calculating by the blood-oxygen-level-dependent effect (Ferrari and Quaresima, 2012). fNIRS offers a variety of advantages over the aforementioned neuroimaging methods, such as a relatively high temporal resolution, absence of ionizing radiation, low cost, high portability, lightweight implementation, and lower susceptibility to motion artifacts during experiments (Lancia et al., 2018; Ho et al., 2019). Unlike commonly sizeable systems (e.g., fMRI, PET, EEG), which prevent the daily activity acquisition in the brain, fNIRS enables the brain measurement with much less intrusive form factors when performing tasks to not interrupt the actual behavior of participants. In AD studies, specifically, fNIRS has verified its own potential (Zeller et al., 2010; Perpetuini et al., 2017). Araki et al. (2014) presented a significant difference in the cerebral blood flows of patients with AD under the effects of memantine. Hemodynamic responses between healthy controls (HC) and patients with AD were also distinguished (Metzger et al., 2016; Li et al., 2019), whilst mild cognitive impairment (MCI) – an early stage of AD – was differentiated from those in the HC group (Ghafoor et al., 2019). These studies have demonstrated that patients with AD typically exhibit lower levels of activation in specific brain regions compared to the HC group during cognitive tasks. In addition, fNIRS is well operable with EEG in the multimodal integration to reinforce the diagnosis accuracy (Fazli et al., 2012). Nevertheless, these previous studies have solely distinguished AD from HC using cognitive tests, whereas different AD stages remain unknown and the pathological mechanism of AD progression has not yet been thoroughly documented. Therefore, in this study, multiple subject groups were recruited and participated in different evaluation tasks (cognitive, memory, and language) to comprehensively compare and evaluate the capability of fNIRS in AD diagnosis.

Machine learning (ML) techniques have been a fundamental element involved in any computer-aided diagnosis systems for the automated prediction of neurological disorders. A Bayes classifier was applied to determine the utility of the eye-tracker, heat flux and median absolute deviation, electrocardiogram armband, EEG headset, and heart rate monitor (Haapalainen et al., 2010). A Gaussian mixture model was used to measure the mental workload of a mental arithmetic task through eye-tracking data (Chen and Epps, 2013). The capability of multiple learning algorithms, including logistic regression, decision trees, naïve Bayes, 1-nearest neighbor, and multilayer perceptron, was evaluated on heart rate data. Subsequently, ML has gained much attention in fNIRS studies. Support vector machines (SVMs) and adaptive boosting as traditional ML algorithms were utilized to distinguish stress levels from the resting states of subjects during Stroop task experiments (Solovey et al., 2014). Logistic regression and SVMs in mental workload classification were also verified (Benerradi et al., 2019).

Despite being successfully applied on signal domains, extracting and selecting features in appropriate ways are still considered as main drawbacks of conventional ML. The cost of these supervised learning methods increases with respect to parameters. As such, these algorithms require a huge amount of labeled training data. Meanwhile, theoretical and biological arguments have strongly recommended a model which is composed of multilayers of nonlinear processing with few labeled inputs in correspondence with human brain activity. Therefore, deep architectures such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) as two common forms of Artificial Neural Network (ANN) have unsurprisingly come to our top priorities in terms of solving optimization problems, automated feature extractions, and multi-class classification tasks. Deep learning (DL) models have been adopted a diversity of applications in practice. CNNs have drawn extensive attention in image classification using various frameworks, such as VGG-16 (Simonyan and Zisserman, 2014) and Residua-Net (He et al., 2016), while RNNs depict connections between computing units that formulate a directed graph along a sequence, thus is popular in time-series data analyses, such as machine translation,

time-series prediction, or speech recognition (Lokesh et al., 2019; Wright et al., 2019).

Although the DL-based approach has earned great success in various computer vision tasks, its potential in AD classification using fNIRS are not substantially achievable. First, various preprocessing steps are required since the amplitude and the length of signals are different from patient to patient. Second, a sufficiently large number of datasets is mandatory to efficiently train a DL model. Thus, either an adequate number of original data or artificial data from augmentation or resampling techniques is required. Third, it is arduous to obtain an efficient conversion approach to express medical signals in the form of carrying pertinent information in CNN. Moreover, feeding inputs to a RNN network and training it with gradient descent algorithms (Ruder, 2016), such as long-term independence, remain challenges. To address these challenges, researchers have recently paid exceptional efforts to develop the 1 Dimensional-Convolutional Neural Network (1D-CNN) which can capture spatial-temporal information of signals and long short-term memory (LSTM), as one type of RNNs, which can train and avoid the so-called vanishing gradient issue. Therefore, it is of great interest and importance to address the aforementioned challenges in our study.

The main contributions of this study are summarized as follows: (1) we presented the comprehensive experimental protocol, including cognitive, memory, and verbal tests to investigate the diverse patterns of AD progression, thereby facilitating the investigation of hemodynamic response differences between the HC group and three stages of patients with AD; (2) we inspected the significant differences of oxygen hemodynamic concentrations between subject groups and between gender; and (3) we demonstrated the potential of fNIRS in AD multi-class classification using four DL architectures.

The remainder of this paper is organized as follows: Section "Related Works" briefly presents the related works concerning AD studies using fNIRS and DL techniques. In section "Materials and Methods," we describe materials, including the experimental protocols, fNIRS datasets, pre-processing steps, and DL models for multi-class classification tasks in detail. Section "Experimental Results" summarizes the experimental classification results. In section "Conclusions and Future Works," we conclude the paper with our main contributions and some suggestions for future works.

RELATED WORKS

Functional near-infrared spectroscopy is a non-invasive way of measuring cerebral hemodynamic change using near-infrared rays (Yücel et al., 2017). Irani et al. (2007) proved the relatively superior spatial-resolution of fNIRS compared to EEG/Magnetoencephalography (MEG). fNIRS was used to successfully distinguish signals from nearby measured brain regions, thus avoiding fake correlations as induced by EEG/MEG (Sonkaya, 2018). Moreover, from a practical point of view, fNIRS is portable, safe, quiet, relatively inexpensive, easy to handle, has fewer restrictions on subjects, compatible with iron metals, and

feasible for long-term continuous and repeated measurements at short intervals. These advantages outweigh the merits of other neuroimaging techniques in the study of neurological disorders and psychiatric disorders such as dementia and brain-related disorders. Many studies have examined the validity of using fNIRS to compare the hemodynamic response in HC and patients with AD, and demonstrated that patients with AD showed lower activation in specific brain regions during various cognitive tasks than normal people's (Beishon et al., 2017; Vermeij et al., 2017; Liao et al., 2019). Hock et al. (1997) reported that there was a decrease in HbO and Total Hemoglobin (HbT) given a verbal fluency task (VFT) in the parietal cortex of patients with AD. Arai et al. (2006) revealed that HbO concentrations were significantly reduced in the frontal and bilateral parietal areas of the AD group, whereas the hemoglobin activation was only lower in the right parietal area of the MCI group. These findings suggest that fNIRS holds the promising potential to detect AD, even at early stages.

To monitor the declined cognitive function of AD, it is essential to further define a classification scheme. Recent studies have adopted ML on the classification of AD stages. With the development of computational resources, the application of ML algorithms has enabled the work of AD stage labeling based on a classification model, which is beyond the traditional analysis. Cicalese et al. (2020) presented a linear discriminant analysis (LDA) algorithm applied on a hybrid EEG-fNIRS dataset to classify four subject groups (8 HC, 8 MCI, 6 MAD, and 7 MSAD) given a random digit encoding-retrieval task. Their results indicated that the right prefrontal and left parietal regions were relevant to AD progression. The integrated EEG-fNIRS feature set achieved a higher accuracy (79.31%) compared to using EEG (65.52%) or fNIRS (58.62%) alone. Yang et al. (2018) aimed to investigate MCI assessment using statistical analysis and an LDA algorithm to differentiate MCI from HC. They yielded an effective ML tool for early prediction of AD using digital biomarkers and fNIRS (Yang and Hong, 2019). Despite its popularity, conventional ML algorithms have been criticized for their poor performance on raw data, not being limited to on bio-signal domains, and for the prerequisite step of manually extracting informative features (Plis et al., 2014; LeCun et al., 2015).

The ability to achieve features with higher orders of abstraction and complexity compared to conventional ML, such as SVM, makes DL better suited for detecting scattered, complex, and subtle patterns of the data. In addition, an integral facet of DL to differentiate it from other ML methods is that features are not manually engineered. Instead, DL learns the data in an end-to-end manner, resulting in a more objective and less bias-prone process (Plis et al., 2014). Proven for its superior image classification and recognition capabilities, CNN has motivated the development of a CNN-based taxonomy for early-stage AD detection (Vieira et al., 2017). This review paper identified a total of 25 articles relevant to DL studies of psychiatric and neurological disorders, including AD-related concerns. First, the diagnostic studies aimed to classify patients from HC. Second, studies on conversion to illness used the baseline scan from individuals who were predicted as being at high risk of developing a psychiatric or neurologic abnormality. Finally, studies on

predicting the treatment response employed the baseline scan from subjects with a neurological or psychiatric diagnosis to predict the subsequent treatment response. The accuracies of DL architectures from those articles were also reported to testify DL feasibility in discriminating between more than two classes (HC, MCI, and AD).

Oh et al. (2019) presented volumetric CNN-based approaches [convolutional autoencoder (CAE)-based unsupervised learning and supervised transfer learning] for four binary classification tasks (AD vs. HC, progressive MCI vs. HC, stable MCI vs. HC, and progressive MCI vs. stable MCI) and a gradient-based visualization task of the spatial attention. They discovered the important AD-related biomarker using an MRI dataset from Alzheimer's Disease Neuroimaging Initiative (ADNI) database without human intervention (Oh et al., 2019). The experimental results showed that their proposed models achieved favorable performance and comparable efficiency to current state-of-the-art models. Meanwhile, the multimodal DL approach has exerted to incorporate different types of inputs and DL models to boost the AD classification accuracy. Lee et al. (2019) extracted multiple features from MRI, cohort data, and cerebrospinal fluid (CSF) data and adopted an RNN to predict AD. Suk and Shen (2013) employed MMSE, MRI, PET, and CSF to discriminate AD from MCI. Feng et al. (2019) fed MRI and PET data to the incorporated framework of 3D-CNN and LSTM.

Previous AD studies have validated not only the potential of fNIRS in assisting clinicians/practitioners to distinguish different AD groups from a HC group, but also the potential of DL on AD analysis. However, to the best of our knowledge, there are several issues that remain unsolved and needed to be addressed. The comprehensive experiment design with different tasks to investigate AD patterns and the DL classification frameworks for AD studies using imbalanced fNIRS datasets have not been thoroughly documented. We, therefore, conducted entire experiments (cognitive, memory, and verbal tests) and verified a collection of four DL models to differentiate the HC group from three levels of AD progression using the fNIRS dataset.

MATERIALS AND METHODS

Subjects

Early-phase AD is hardly characterized since it precedes for years as an asymptomatic AD that is not comparable to the normal aging process. However, using neuropathological biomarkers from the cerebrospinal fluid and several brain imaging techniques, the sequence of AD development is distinguishable from that of aging in a cognitively unimpaired healthy state that does not have AD biomarkers. Hence, the National Research Center for Dementia and Chonnam National University Hospital (Gwangju, South Korea) recruited senior citizens living in Gwangju and adjacent cities to approach this objective. By conducting a series of medical examinations (MMSE, PET, MRI, and patient interview), subjects were diagnosed with disease stages according to the guidelines from the National Institute of Neurological and Communicative Disorders and Stroke (NINCDS)-Alzheimer's Disease and Related Disorders Association (ADRDA) Work

Group (McKhann et al., 1984), the National Institute on Aging and Alzheimer's Association (NIA-AA) (Jack et al., 2011), and International Working Group (IWG) (Cummings et al., 2013). Subjects with a mental and/or behavioral disorder were excluded from this cohort.

The final subject groups included HCs (cognitively normal individuals), asymptomatic AD (abbreviated as aAD: cognitively normal individuals, the amyloid positive in PET was found), prodromal AD (abbreviated as pAD: mild brain dysfunction symptoms such as short-term memory deficits, impaired insight, irritability, dysphoric mood, and anxiety), and AD dementia (abbreviated as ADD: deterioration of memory, language, social abilities that severely effect on daily life) groups. A total of 140 subjects were involved in this study: HC class ($n = 53$, age = 72.7 ± 5.3 years), aAD class ($n = 28$, age = 74.5 ± 4.3 years), pAD class ($n = 50$, age = 75.8 ± 3.9 years), and ADD class ($n = 9$, age = 75.4 ± 6.8 years).

All subjects had no previous experience with our experimental protocol. The purpose of the research and the written consent form were furnished and agreed upon by each subject prior to conducting the experiment. The experimental procedure was approved by the International Review Board at the Gwangju Institute of Science and Technology. The demographic information of participants is summarized in **Table 1**.

Functional Near-Infrared Spectroscopy Device and Data Preprocessing

Figure 1A represents the fNIRS device setup. fNIRS signals from six channels located in the prefrontal cortex region were recorded. The device was built in the laboratory, and the probe consisted of a pair of source and detector, light-emitting diode (LED; OE-MV7385-P, Opto ENG, South Korea) and photodiodes (Opt101, Texas Instruments). The LED had dual wavelengths (730 and 850 nm). Channels 1, 2, and 3 measured signals in the right prefrontal region while channels 4, 5, and 6 measured signals in the left prefrontal region. While channels 2, 3, 5, and 6 were far-channels (30 mm) and the signals obtained from these channels presented the cerebral hemodynamic response, channels 1 and 4 were known as near-channels since the distance between the source and the detector was very close (8 mm) in order to measure the superficial layer from the head. The photodiode located in between channel 3 and channel 5 was 15 mm distant from the Fpz point in the 10–20 EEG system.

The signal sampling rate was set to 8 Hz. Each channel was visually inspected during the experiment, and channels exhibiting large spikes were marked as noise and excluded from the study. The changes in the concentration of hemoglobin [HbO, Hb, and

TABLE 1 | Subject demographics.

	HC	aAD	pAD	ADD
Number	53	28	50	9
Age ± SD (years)	72.7 ± 5.3	74.5 ± 4.3	75.8 ± 3.9	75.4 ± 6.8
Gender (M/F)	21/32	15/13	31/17	4/5
MMSE ± SD	27.0 ± 4.2	26.9 ± 2.5	26.0 ± 3.2	20.2 ± 4.8
Education ± SD	9.8 ± 4.7	10.2 ± 5.2	10.6 ± 5.2	8.5 ± 5.3
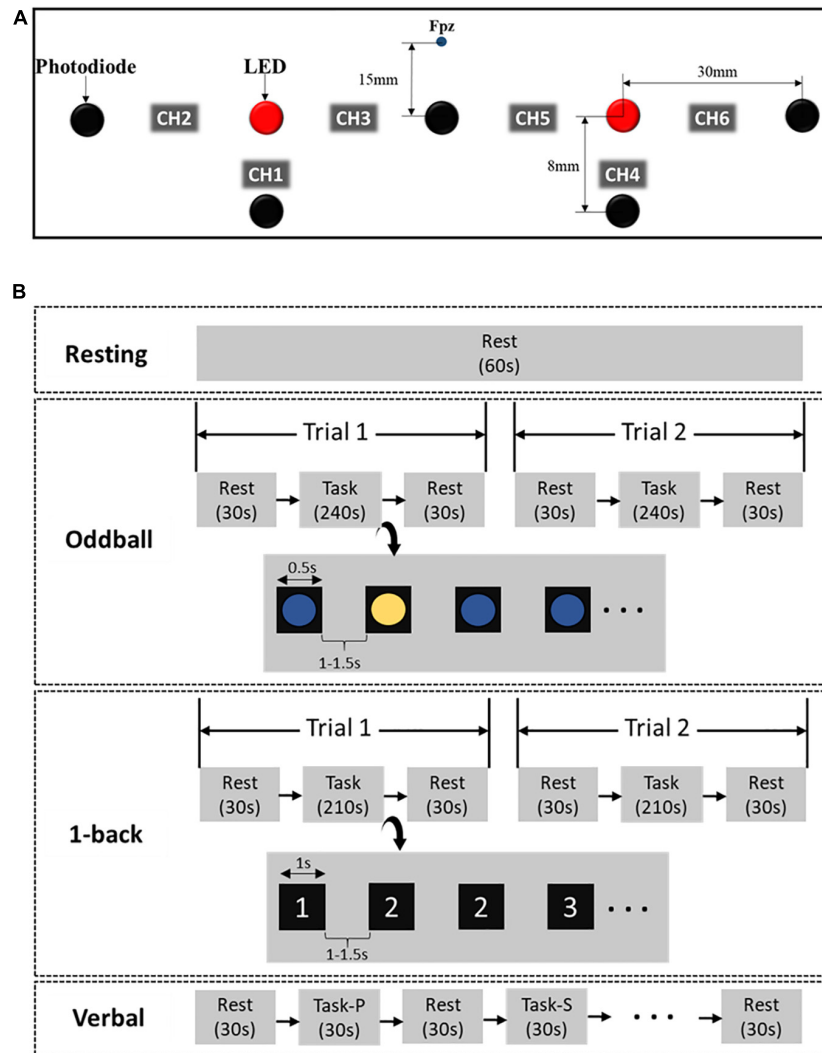
FIGURE 1 | fNIRS device setup **(A)** and experimental protocol **(B)** (s, second; P, phonemic; S, semantic).

total hemoglobin (THb)] were computed by the modified Beer–Lambert law (Swinehart, 1962). To remove artifacts, a low-pass filter with a cutoff frequency of 0.5 Hz was applied. Although we utilized four channels in our previous studies (Nguyen et al., 2019) where the near-channels were excluded by employing the extraction algorithm since their function was to get rid of the skin and skull noise from the far-channels, it is crucial to consider inputs from all channels to be fed to proposed DL frameworks in this present study.

Experimental Protocol

Subjects were requested to sit in a comfortable chair situated in a confined room to reduce environmental disturbances and minimize their body movements during experiments. In our experiment, subjects underwent a resting stage and three brain functional tasks, including Oddball (cognitive ability test), 1-back (memory ability test), and Verbal fluency (language ability test). First, subjects calmly watched a white cross appeared on the

monitor screen for 60 s during the resting period. Subjects also had 30 s for resting in a similar way before and after each experimental task.

Second, the blue and yellow circles appeared alternately, and subjects were asked to report the color of each ball in the Oddball task. The yellow circle rarely appeared compared to the blue one. Subjects were asked to pay attention to the occurrence of the circle and press "0" if the yellow circle presented. Each colored circle appeared for 0.5 s, and a blank screen appeared for 1–1.5 s between the circles. The task took 5 min and was recorded twice.

Third, a consecutive series of numbers, from 1 to 3, randomly appeared on the screen in the 1-back task. The appearance of each number lasted for only a second. Subjects were requested to remember a number which was selected and previously presented in one trial. Subjects then pressed the "0" button likewise. The task took 5 min and was also recorded twice.

Lastly, a total of six Korean phonemic and semantic words alternately appeared on the screen in the Verbal fluency task.

Initially, a letter appeared alone (i.e., "a"), and subjects would continuously recite complete words that start with the letter (i.e., "anniversary"). If a white cross appeared on the screen, they would stop speaking and carefully glare at the screen until the next semantic word appeared. When the semantic word (i.e., "animal") appeared, the subjects would continuously recite certain animals (i.e., "Lion"). Each trial block took for 30 s, and there was a 30-s resting state between trials. This process was repeatedly conducted, and the Verbal task took approximately 7 min.

In total, every subject performed six sub-tasks in the order of Resting, Oddball (trial 1), Oddball (trial 2), 1-back (trial 1), 1-back (trial 2), and Verbal fluency. All procedures took around 30 min, including the break time between tasks (see **Figure 1B**). Any subject who was not able to fulfill any of these sub-tasks was excluded from our analysis.

Proposed Deep Learning Models

As a subcategory of ML, DL is inspired by the structure and function of the brain, the so-called biological neural networks (NNs). A DL model is made of the basic building block of NNs which are trained on the large-scale labeled dataset(s) to learn functions from the original data without manually extracting features, as most of the conventional ML algorithms do. CNN, denoted as one of the most well-known forms of DL, has been demonstrating its robust performances in the computer vision community where the inputs usually include high dimensional features (i.e., each pixel of an image is a dimension). Hence, it is pertinent to perform on a long sequence of data and is easy to implement. Meanwhile, employing RNN, which can handle the temporal data through modeling the underlying dynamic behaviors of time series sequences, is natural. However, since the sequence data is generated from electrodes that are closely located, there is a strong correlation in time within the sequence. Directly learning sequences with noises from RNNs is very challenging. LSTM or the recently invented Generated Recurrent Units (GRUs) can settle the vanishing or explosion of gradient loss. Due to the early-stopping strategy used during the training process, it is certainly expected that no severe overfitting can be observed. Lastly, to take synergetic advantage of CNN and LSTM, a mixed model including CNN followed by LSTM, where CNN acts as an effective decoder for LSTM layers, is seemingly interesting to build.

1D-Convolutional Neural Network

Since the input size of fNIRS data is different from acquainted image datasets with more than one dimension, a 1D-CNN is necessarily done in the time domain. In this work, with the idea of using 2 dimensional stacks for time-specific activity, the CNN model was employed to decompose and feed-forward a 1D fNIRS signal that was known as a precomputed spectrogram or time-frequency. CNN consisted of one 1D convolutional layer, one 1D max-pooling layer, and two linear layers. Batch normalization was used before the activation function, and dropout was used for regularization. A grid search of convolutional filter number, filter size, pooling size, batch size, number of units per layer, stride, and

learning rate was done to optimize hyper-parameters, as shown in **Supplementary Table 1**.

Long Short-Term Memory

To discriminate the subtle spatial-temporal change in the fNIRS feature space associated with different progressive stages of AD, a solution capable of remembering and ultimately aggregating transitions across the dataset is required. LSTM, with the attention mechanism, has been known as a wide-ranging method utilized for distinctly learning and classifying bio-signal time-series datasets (Tsiouris et al., 2018; Zhang et al., 2018, 2019) on fNIRS (Yoo et al., 2018; Sirpal et al., 2019). Thus, it triggered us to apply LSTM to improve the classification performance using fNIRS by focusing on the crucial task-relevant features from different time-steps. Compared to classic approaches in such aforementioned classifiers, RNN or LSTM particularly requires no or almost no feature engineering that an fNIRS dataset is able to feed directly into the network. We utilized the "many to one" architecture, which means using time series of feature vectors (one vector per time step) to convert them to a probability matrix at the output layer.

An LSTM network consists of cells whose outputs evolve through the network based on the content of past memory. The cell has a common cell state, keeping long-term dependences along the entire LSTM chain. The flow information is then monitored by the input gate i_t and the forget gate f_t, hence allowing the network to decide whether to forget the previous state C_{t-1} or to update the current state C_t with new information. The output of each cell, namely, hidden state, is controlled by an output gate o_t that allows the cell to compute its output given the updated cell state (see **Figure 2A**). These states of an LSTM model are formulated by the following formula:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{1}$$
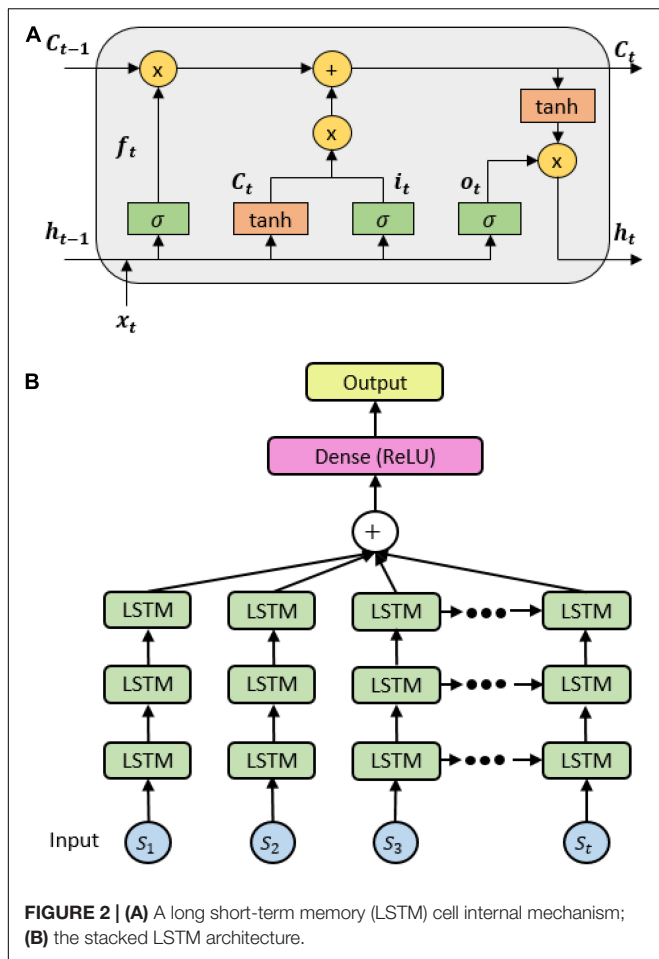
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{2}$$

$$C_t = f_t * C_{t-1} + i_t * tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \tag{3}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{4}$$

$$h_t = o_t * tanh(C_t), \tag{5}$$

where $\sigma(x) = \frac{1}{1 + e^{-x}}$, x_t is the input features fed into the cell, h_t is the hidden state at time step t, C_{t-1} is the cell state at time step t, W_i, W_f, W_c, W_o are the weights, and b_i, b_f, b_c, b_o are the biases obtained during the backpropagation process. We thereby proposed an LSTM architecture with three stacked layers and a fully connected layer with rectified linear units (ReLU) function to predict the probability of each class (see **Figure 2B**).

Several hyper-parameters for the LSTM network were thoroughly explored and tuned to earn the best classification results. Concretely, these hyper-parameters encompassed the recurrent depth, LSTM hidden layer size, batch size, and training epochs. Three stacked LSTM layers (D_0, D_1, D_2) and the weight matrix L_2 regularization coefficient was applied for each LSTM layer. In addition, other variables such as learning rate (l_r) and the amount of lambda loss were used for stochastic Adam

FIGURE 2 | (A) A long short-term memory (LSTM) cell internal mechanism; **(B)** the stacked LSTM architecture.

FIGURE 3 | (A) A gated recurrent unit (GRU) cell internal mechanism; **(B)** the stacked GRU architecture.

optimizer. **Supplementary Table 2** presents the optimal values for these parameters.
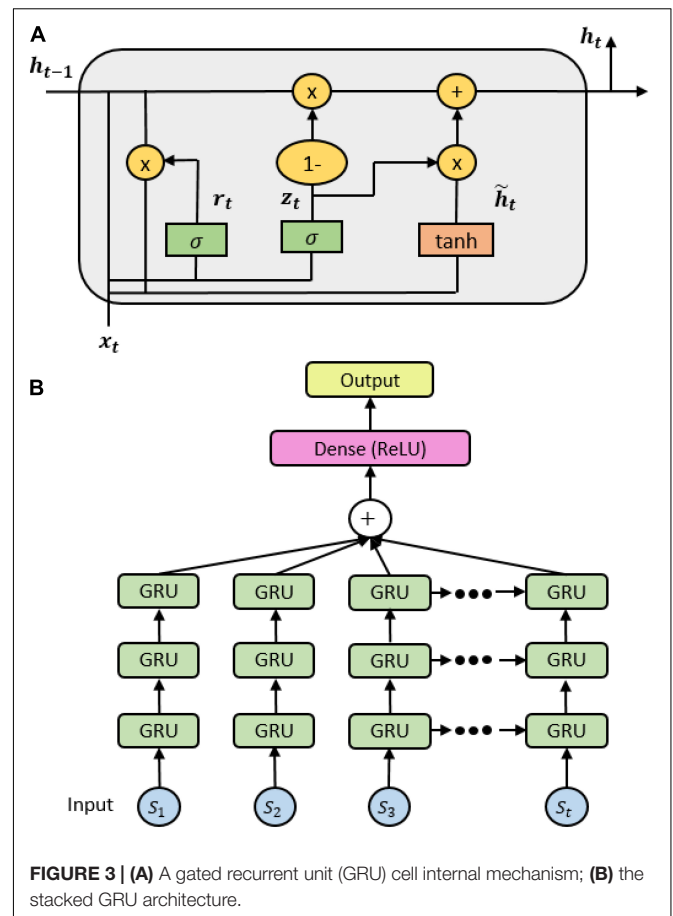
Gated Recurrent Units

As similar to the precedent LSTM, gated recurrent units (GRU) is newly created for the solution of short-term memory. An internal mechanism called gates is also needed to regulate the flow of information. However, GRU gets rid of the cell state and uses the hidden state to transfer information. It includes only two gates: a reset gate which is another gate used to decide how much past information can be forgotten, and an update gate which acts similar to the forget and input gate of LSTM which decides what information can be discarded and what new information can be complemented (see **Figure 3A**). These states of a GRU model are formulated by following formulas:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \qquad (6)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \qquad (7)$$

$$\tilde{h}_t = tanh(W \cdot [r_t * h_6 t - 1, x_t]), \qquad (8)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t) \qquad (9)$$

Due to its fewer tensor operations, GRU trains data faster than LSTM. It also has a faster convergence, but it is unlikely to guarantee which one outperforms others. In practice, we built a GRU model with the same architecture of LSTM, excluding the module used in the current layers. Three GRU layers were stacked, followed by a fully connected layer and an output layer (see **Figure 3B**). Batch-normalization and the ReLU activation function were applied in the fully connected layer. The output layer was computed by the cross-entropy loss with a soft-max layer. More details of chosen parameters are depicted in **Supplementary Table 3**.

A Mixed Model: Long Short-Term Memory Model Coupled With Convolutional Neural Network Decoder

Inspired by the phenomenal performance of the hybrid CNN-RNN in both the spatial domain (Dutta et al., 2018) and the spatial-temporal domain (Yang et al., 2018; Shi et al., 2019), we employed a CNN layer as a decoder for original input data, followed by multiple LSTM layers to cope with the decoded temporal data, and a fully connected layer where the cross-entropy loss with a soft-max function was computed for classification (see **Supplementary Table 4**). We let a CNN layer extract multiple high-level features from the raw data by using a substantially large filter size and stride. The sequence was thus
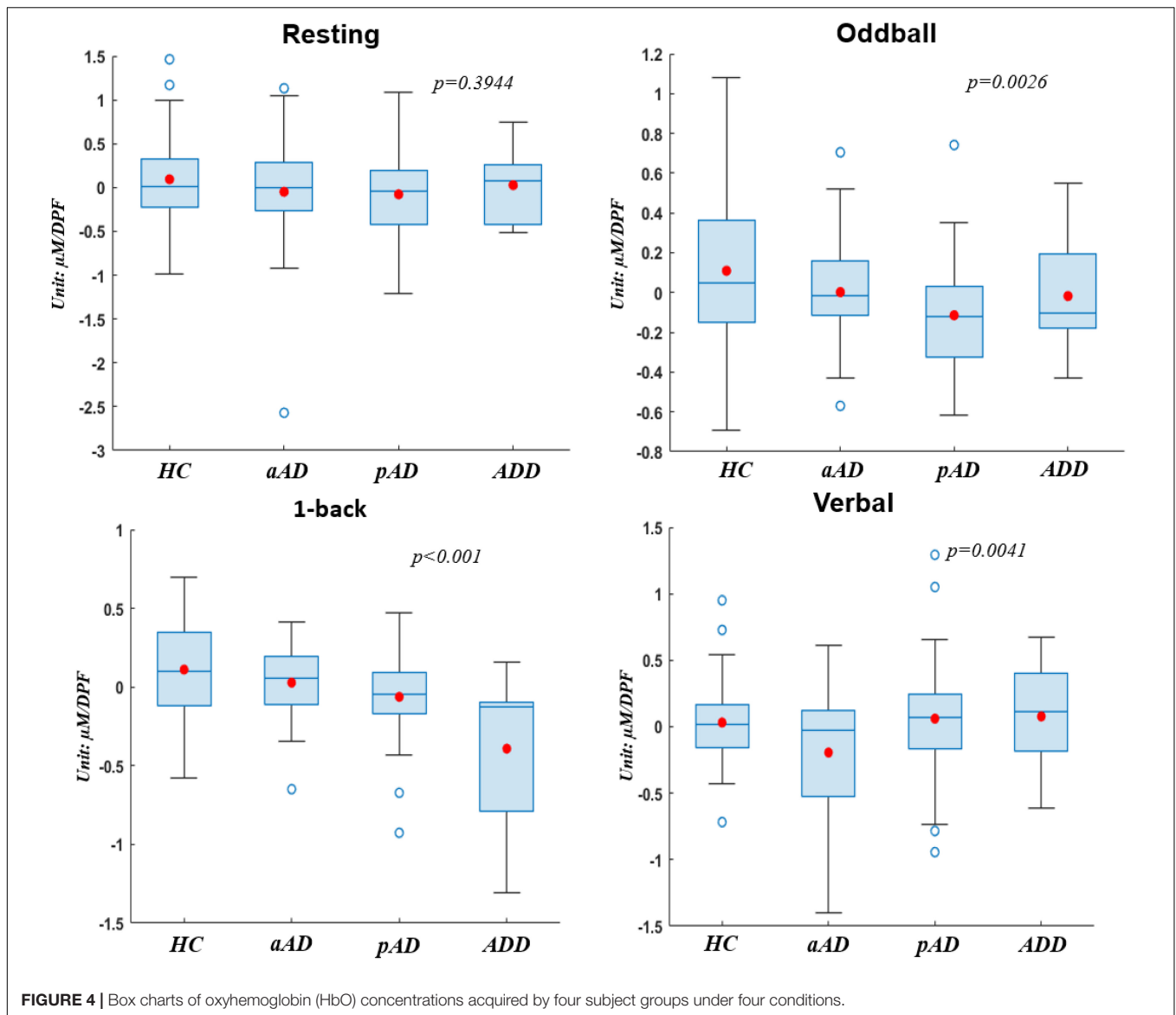
FIGURE 4 | Box charts of oxyhemoglobin (HbO) concentrations acquired by four subject groups under four conditions.

decoded into multiple shorter sequences symbolizing different high-level temporal features.

EXPERIMENTAL RESULTS

Statistical Analysis: The Change of Oxygen Hemoglobin Concentrations

To clearly evaluate the change of hemoglobin concentrations among four subject groups who underwent four experimental stages, we utilized the utmost distinguishable observation from HbO over other hemoglobin types of reduced hemoglobin (HbR) and THb in this part. **Figure 4** presents the box chart distributions of HbO obtained from each experimental protocol to compare HC with each level of AD severity. Note that the typical hemodynamic response of brain activation is accompanied by an increase of HbO concentrations

(Mangrum et al., 2018). On each box, the central mark is median, the red filled circle is the mean, the edges of each box are the 25th and 75th percentiles, the whiskers denote the standard error of the mean, and the empty blue circle show outliners. All statistical analyses were performed with a significant level at 0.05 as the cutoff of significance, and each box chart was given a p-value.

Regarding the difference among subject groups, all groups did not show significant differences ($p > 0.05$) during the resting period, whereas there existed significant differences during the other three experimental tasks ($p < 0.05$). The HC group (values were greater than zero) consistently presented much higher hemodynamic activations than other groups. This indicates that normal people had the highest response and had no difficulty when they performed all the brain ability tasks. Meanwhile, on the second place of AD severity level, pAD patients showed lower values than HC and aAD during resting,
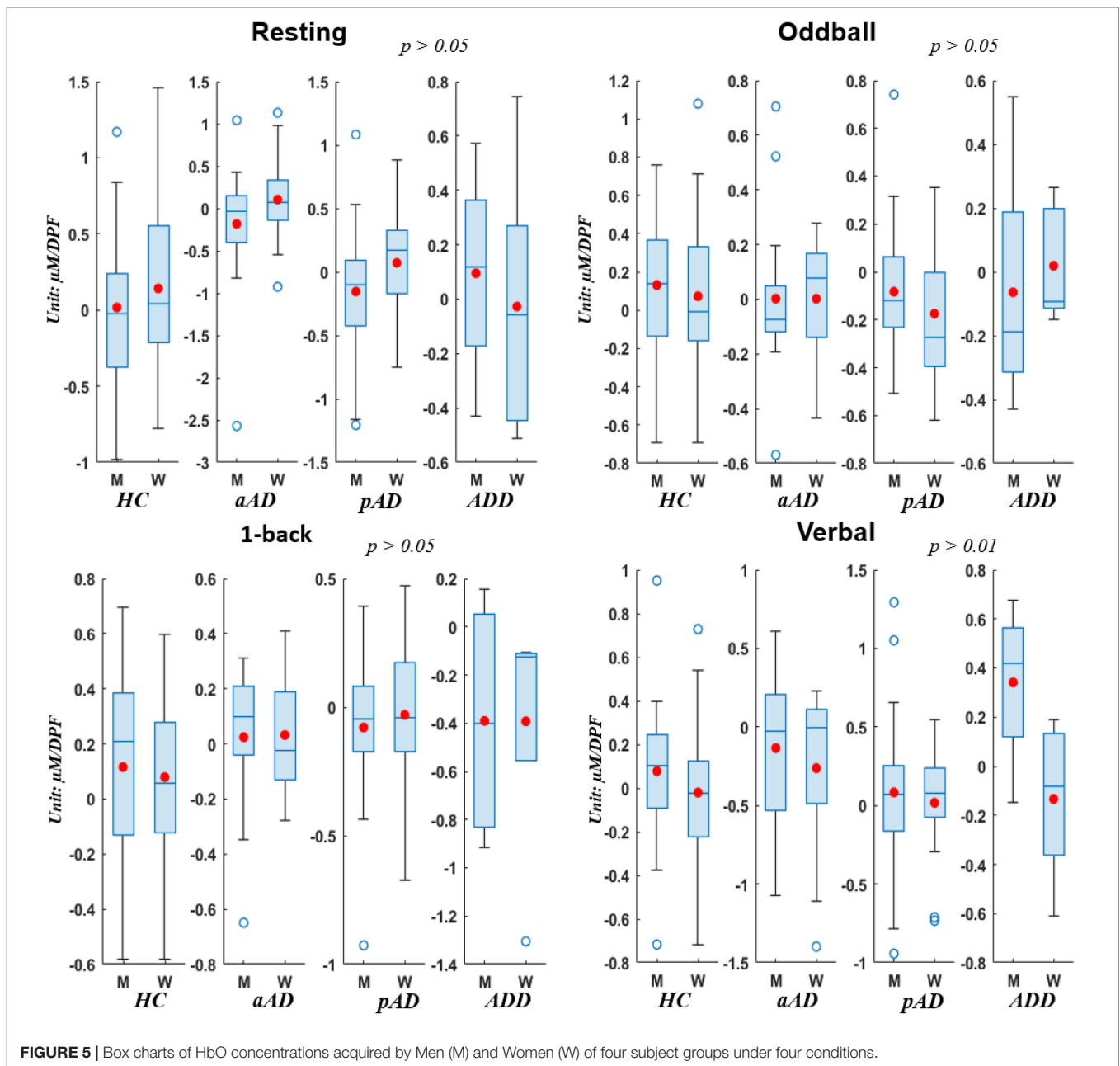
FIGURE 5 | Box charts of HbO concentrations acquired by Men (M) and Women (W) of four subject groups under four conditions.

oddball, and 1-back tasks. This verifies that pAD patients noted lower activations of hemodynamic responses compared with HC and patients with aAD in the resting period, cognitive ability, and memory ability tests. More interestingly, the significantly statistical difference between subject groups did exist during the 1-back task ($p < 0.001$). Each group behaved and responded differently in this memory task, and their response was perfectly correlated with the degree of AD severity. ADD patients, in particular, denoted the lowest hemodynamic activation, followed by pAD, aAD, and HC.

In addition, we testified the change of HbO concentrations via box chart distributions for all conditions when subjects were sub-grouped by gender (see **Figure 5**). In general, there was no

significant difference between men and women who underwent all three conditions ($p > 0.05$ for resting, oddball, and 1-back stages) whilst there existed relatively important disparity among gender in the verbal stage. Further, men from all four subject groups presented much higher levels of hemodynamic activations than women in most conditions except the resting period in which men from HC, aAD, and pAD groups showed lower hemodynamic activations. This finding suggested that the frontal changes in HbO concentrations during cognitive, memory, and verbal tasks were stronger in men compared to women. Hence, there was a gender effect on relative hemoglobin changes due to the functional stimulation in our study that achieved similar observations from previous studies on evaluating hemodynamic

TABLE 2 | Comparison of classification metric results as shown by four proposed deep learning (DL) models using original data.

Class	Accuracy			Precision			Recall			F1-Score		
	HbO	HbR	THb	HbO	HbR	THb	HbO	HbR	THb	HbO	HbR	THb
1D-Convolutional Neural Network (1D-CNN)												
HC	0.863	0.853	0.866	0.884	0.862	0.870	0.863	0.853	0.866	0.848	0.852	0.862
aAD	0.826	0.817	0.779	0.847	0.879	0.819	0.826	0.817	0.779	0.806	0.810	0.774
pAD	0.854	0.861	0.876	0.878	0.842	0.912	0.854	0.861	0.876	0.849	0.860	0.875
ADD	0.774	0.731	0.786	0.778	0.864	0.772	0.754	0.711	0.736	0.749	0.710	0.735
Mean	0.829	0.815	0.827	0.847	0.862	0.843	0.824	0.810	0.814	0.813	0.808	0.811
Long Short-term memory (LSTM)												
HC	0.897	0.855	0.875	0.913	0.835	0.903	0.897	0.855	0.875	0.893	0.850	0.874
aAD	0.830	0.836	0.848	0.904	0.907	0.927	0.830	0.836	0.848	0.837	0.835	0.836
pAD	0.867	0.863	0.879	0.887	0.900	0.925	0.867	0.863	0.879	0.858	0.874	0.872
ADD	0.781	0.761	0.795	0.764	0.762	0.747	0.751	0.731	0.735	0.751	0.721	0.732
Mean	0.844	0.829	0.849	0.867	0.851	0.875	0.836	0.821	0.834	0.835	0.820	0.828
Gated Recurrent Units (GRUs)												
HC	0.816	0.765	0.771	0.839	0.850	0.810	0.816	0.765	0.771	0.812	0.760	0.774
aAD	0.792	0.727	0.741	0.811	0.808	0.785	0.792	0.727	0.741	0.770	0.719	0.739
pAD	0.811	0.779	0.805	0.837	0.868	0.849	0.811	0.779	0.805	0.808	0.778	0.805
ADD	0.705	0.695	0.705	0.721	0.764	0.736	0.705	0.695	0.705	0.706	0.697	0.708
Mean	0.781	0.742	0.756	0.802	0.823	0.795	0.781	0.742	0.756	0.774	0.738	0.756
CNN-LSTM												
HC	0.918	0.907	0.917	0.938	0.905	0.918	0.918	0.907	0.917	0.898	0.902	0.913
aAD	0.879	0.855	0.889	0.865	0.872	0.917	0.839	0.855	0.859	0.835	0.853	0.864
pAD	0.880	0.893	0.907	0.902	0.890	0.910	0.880	0.893	0.907	0.876	0.894	0.907
ADD	0.797	0.799	0.793	0.781	0.839	0.771	0.767	0.799	0.753	0.767	0.806	0.746
Mean	0.868	0.864	0.877	0.871	0.877	0.879	0.851	0.864	0.859	0.844	0.864	0.858

activations based on gender inclusion criteria (Shoemaker et al., 2001; Auger et al., 2016).

Classification Performance of Deep Learning Models

To precisely validate the potential of deep learning approaches, the original dataset (three hemoglobin types) was used. Four models were built, including 1D-CNN, stacked LSTM, stacked GRU, and a mixed CNN-LSTM. To enhance the classification performance, we segmented the training signals into nonoverlapping fragments with a specified window length. Each split window was assigned a label associated with its ground truth. Then, we ran eight classifier algorithms to differentiate the HC group from the three levels of AD severity. DL models were then adopted to differentiate the HC group from the three levels of AD severity. Five-fold cross-validation was used to evaluate the model's performance using accuracy-based measurements (Accuracy, Precision, Recall, and F1-score).

In regard to the efficacy of the DL models, all models achieved average accuracies above 80% using all three types of hemoglobin (i.e., accuracies of 82.9, 84.4, 78.1, and 86.8% with HbO type

are associated with the performance of 1D-CNN, LSTM, GRU, and CNN-LSTM, respectively). Based on our observations, although LSTM and GRU models had similar concepts of preventing the vanishing gradient problem, their performance in AD multi-classification using fNIRS were exposed differently. GRU was simpler with only two gates, was trained faster, and more computationally efficient, but GRU obtained much lower accuracies than LSTM on all three hemoglobin classes. This indicates that LSTM could outperform GRU since LSTM resembled the theory of remembering longer sequences with a memory unit. This made LSTM more sophisticated and able to produce the network stability in dealing with the fNIRS task, which requires modeling the long-distance relation. In addition, LSTM achieved a higher classification accuracy than 1D-CNN. That is, LSTM, which included less feature compatibility, handled arbitrary input/output lengths, and used its internal memory and time-series information to process the arbitrary sequences of the input, was well-suited for the fNIRS temporal data compared to 1D-CNN.

More importantly, CNN-LSTM outperformed others (86.8, 86.4, and 87.7% corresponding to HbO, HbR, and THb, respectively). CNN and LSTM were not mutually exclusive since

either of them could perform image and time-series classification tasks. The model integration could yield an opportunity to bond the two network types to facilitate effectiveness. A CNN alone would be unable to process fNIRS as the input to be classified as it is visually complicated with the temporal characteristic. Meanwhile, the high-level features that were extracted from long sequences of fNIRS signals by CNN were considerably easy for LSTM to deal with due to its shorter lengths. In detail, the highest accuracy averaged among three hemoglobin types was 87.7% and could occasionally peak to values up to 88.9% at some moments of luck during the training process. This relied on how LSTM weights got initialized at the beginning of training. This means that LSTM was able to correctly identify the movement type using decoded CNN features. Thus, CNN-LSTM was more robust in classifying HC with three different AD stages compared to 1D-CNN and LSTM alone. **Table 2** generally denotes the robust performance of the DL model of CNN-LSTM on all three hemoglobin types with higher means in terms of different classification metrics than those obtained by other DL models. Regarding the effects of subject group types, since we had an extremely imbalanced dataset (for all three types of hemoglobin concentrations), it is predictable that DL models could easily classify the majority classes (i.e., HC and pAD). As a result, we obtained accuracies with higher means and lower SD on HC and pAD, while groups of patients with aAD and patients with ADD achieved lower classification accuracies.

The training's session progress of CNN-LSTM, including Training and Validation accuracies (**Figure 6A** and losses **Figure 6B**), is presented. Initially, the validation loss was slightly similar or lower than the training loss. As long as the validation loss was higher, we would stop the training. In addition, both training and validation accuracies should approach to 1. This demonstrates that the training of CNN-LSTM was exposed to have a stable and accurate performance. As shown in **Figure 6C**, the overall confusion matrix which was obtained visualized the prediction accuracy of CNN-LSTM by comparing the actual and predicted classes. The overall confusion matrix was produced by the full set of actual and predicted classes to visualize the performance and effectiveness of the algorithms. In particular, for each of five splits, we fitted the training data to the model, and the fitted model was used to predict the classes in the current fold. All actual classes and predicted classes were then appended after fivefolds. The model nearly predicted HC correctly (six misclassified samples), aAD (five misclassified samples), pAD class (five misclassified samples), and ADD (three misclassified samples) in the total of 140 patients.

Comparison of the Performance of Deep Learning-Based Approaches With Traditional Machine Learning Classifiers
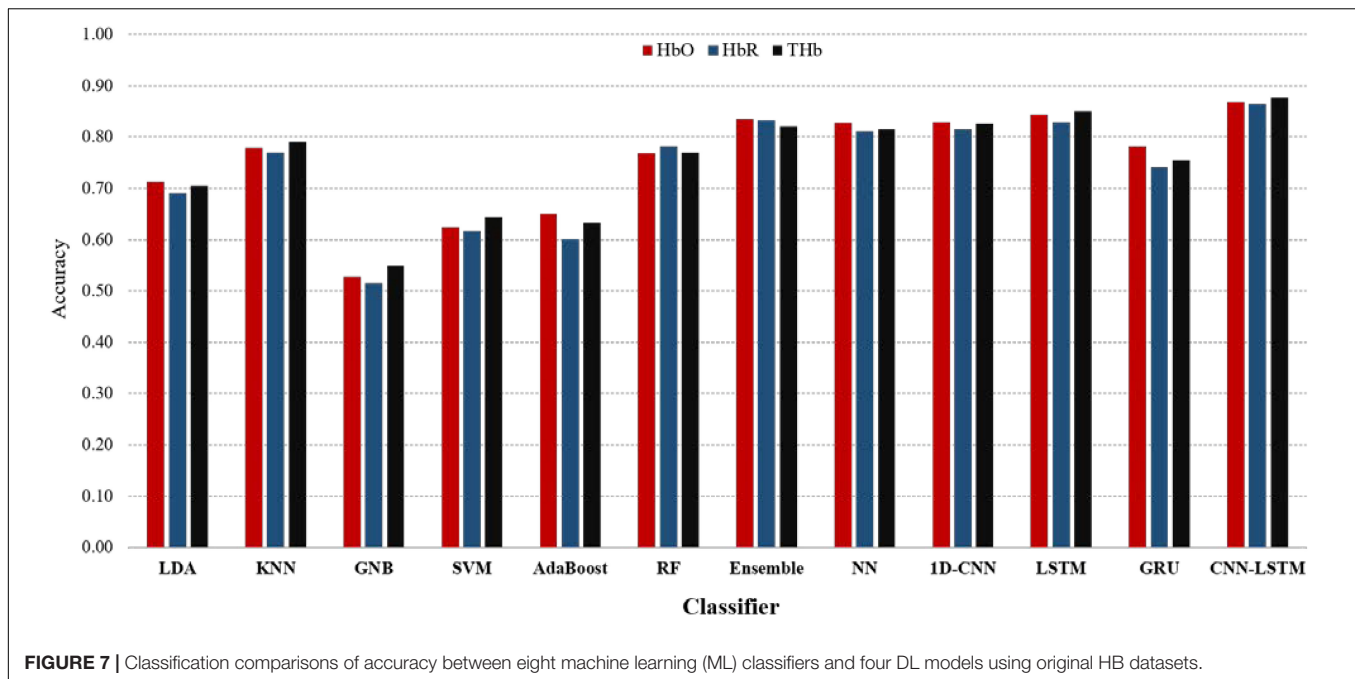
To empirically verify the potential of DL techniques in AD classification using fNIRS, we compared them with the performances of seven conventional ML algorithms, including:

- Linear Discriminant Analysis (LDA) (Li et al., 2006): A set of n samples $\{x_1, x_2, \ldots, x_n\}$ was assigned to four AD classes. We calculated the intra-class M_{intra}
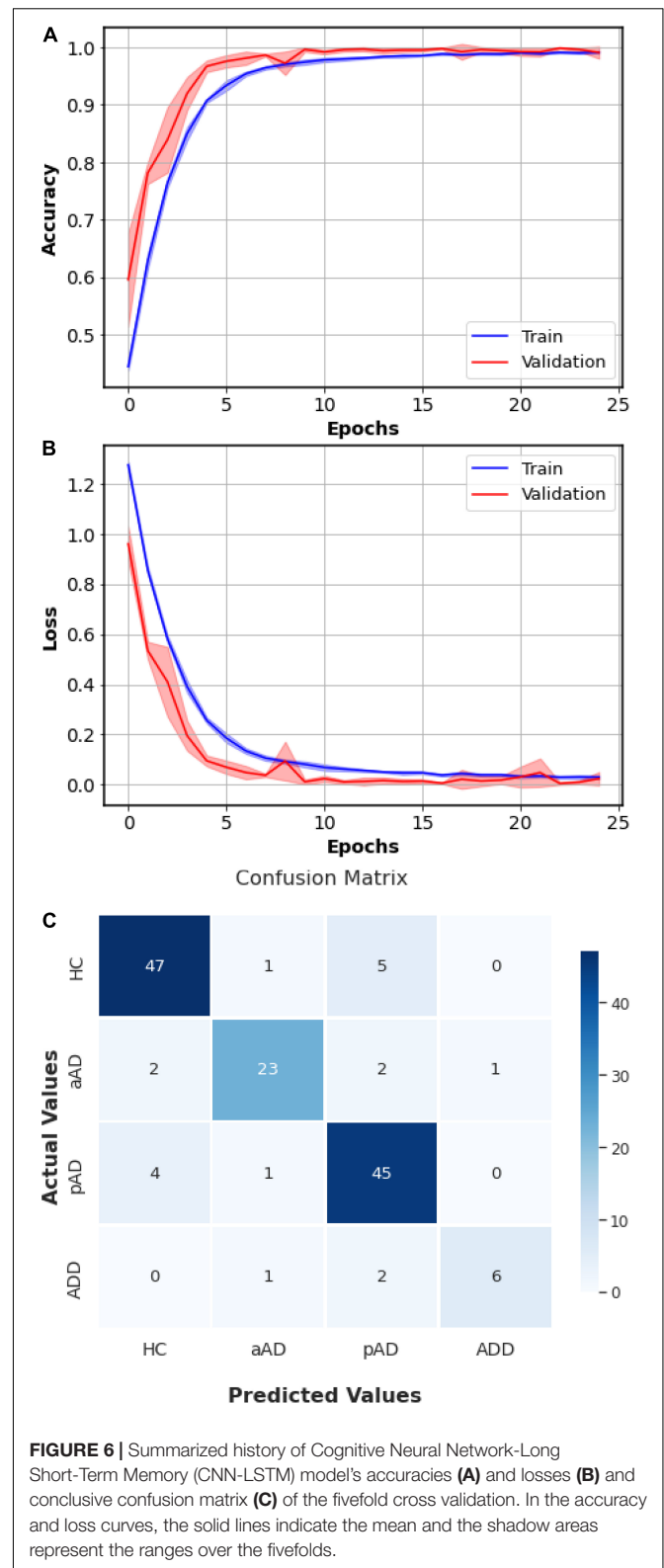
FIGURE 6 | Summarized history of Cognitive Neural Network-Long Short-Term Memory (CNN-LSTM) model's accuracies **(A)** and losses **(B)** and conclusive confusion matrix **(C)** of the fivefold cross validation. In the accuracy and loss curves, the solid lines indicate the mean and the shadow areas represent the ranges over the fivefolds.

and inter-class M_{inter} matrices samples (discriminant features) and performed the classification task on the transformed space using the Euclidean distance with the

FIGURE 7 | Classification comparisons of accuracy between eight machine learning (ML) classifiers and four DL models using original HB datasets.

200 epochs training. LDA offered an elegant way to reduce dimension and to possibly classify four classes using those discriminant features.

- K-Nearest Neighbors (KNN) (Zhang and Zhou, 2007): $K = 4$ (number of clusters) was assigned as the most optimal value. The KNN algorithm computed the distance between each data point and its cluster centroid and minimized the error. The point located at a minimum distance from the testing point was supposed to belong to the same class. After 200 epochs, KNN could capture the distant proximity for four class-classification. However, its performance gradually slowed down as the number of data points increased.
- Gaussian Naive Bayes (GNB) (Griffis et al., 2016): We computed the prior $P(c)$ and posterior probability $P(x|c)$. Then, the conditional probability for each class (given a test sample) was calculated $P(c_i|x)$ based on the Naive Bayes (NB) theory. The class presenting the highest probability among the four classes was predicted as the target class. GNB was the simplest and easiest algorithm among all conventional ML methods.
- Support Vector Machine (SVM) (Lee and Lee, 2007): To non-linearly separate four classes, SVM was applied to construct the decision hyperplanes whereby the margin of the classifier was maximized. The sigmoid kernel trick was employed to do the data transformation and seek optimal boundaries between classes, although this mimicked the idea of the two-layer perception and generated less errors when dealing with non-linear fNIRS data. The number of epochs was not assigned since SVM would stop training when the margin error was very trivial. In addition, no further optimization was required. The

time-consuming and computationally intensive resources were more demanding compared to previous classifiers.
- Adaptive Boosting (AdaBoost) (Hastie et al., 2009): A weak classifier was initially built to generate class labels using unweighted training samples. If any misclassified data point was found during training, the weight of that training point was boosted and updated to the next classifier. The procedure was repeated while each classifier had its own score. The final classifier was defined by the linear combination of 1,000 weak classifiers to boost accuracy. The number of estimators was 200 to control the boosting process, and the higher amount of training time was inevitable.
- Random Forest (RF) (Gray et al., 2013): Starting with the selection of arbitrary samples from the given dataset, a decision tree was built for each sample and produced predicted classes. RF consisted of multiple decision trees that would vote for each class. The class was then selected based on the most votes to be the final prediction. The max depth of the tree was set as 8 and the number of estimators was 200. RF was trained several times, and the random state was fixed to guarantee the same sequence of samples and to achieve a deterministic behavior.
- Ensemble Learning (Kang et al., 2015): Six base classifiers (aforementioned algorithms) were integrated to yield a strong and diverse set of base learners to find the optimal combination. By selecting features from these weak learners via a meta-classifier, it was not easily vulnerable to overfitting. We predicted the final class label using the hard (majority) voting method. The result was explicitly equal or better than the best of base classifiers although, it costed longer training time and features. demanding memory.

- Neural Network (NN) (Amato et al., 2013): Compared to previous ML algorithms, NN could learn non-linear and complex relationships of inputs and outputs to constitute better generalization and classification accuracies. After training NN with different settings of hyper-parameters, we came up with the final set of neurons: an input layer, the set of hidden neurons (100-80-50) organized in the form of three hidden layers, and an output layer (four neurons). We manually selected a ReLU activation function, Adam optimizer, 1e-5 learning rate, and 500-iteration duration. NN was relatively hard to train and required a large number of parameters.

As presented in **Figure 7**, among traditional ML algorithms, the highest accuracy was obtained by Ensemble (82.9% in average), followed by NN, KNN, RF, LDA, SVM, AdaBoost, and GNB (81.8, 77.9, 77.2, 70.3, 63.9, 62.7, and 53.0%, respectively). This indicates that NN was substantially flexible to adopt our adequate data size without considering any feature-engineering steps or structured data and was able to learn high-level features in an end-to-end manner. Meanwhile, most of the features to be identified (to reduce the complexity), split into different parts (to make patterns more tangible), and recombined in the final stage were all preliminary steps for other ML classifiers. However, the obtained result was somewhat poor. In contrast, DL or CNN-LSTM, in particular, validated its robust performance to scale the accuracy up to 87.7%. The results revealed that CNN-LSTM obtained the highest accuracies, tallying with three hemoglobin types, followed by LSTM, NN, 1D-CNN, and Ensemble. In general, the DL algorithms outperform the ML algorithms. Among them, CNN-LSTM elevates accuracy to 3% compared to NN (as the best performance of the ML algorithms).

CONCLUSION AND FUTURE WORKS

Developing an approach to identify individuals before or during the earliest stages of AD, with the hope that the urgent intervention could significantly prevent the onset of clinical symptoms, is desperately needed. Due to somewhat similar symptomologies shared by moderate AD with other neurological disorders, the difficulty of monitoring AD progression that results in diagnostic uncertainties and the subsequent suspension of prompt treatments still remains. Therefore, this present study, on one hand, aimed to present the potential capability of fNIRS in AD analysis. A cohort of 140 participants ranging from HC to three severity levels of patients with AD was included. We verified that there were significant oscillations of hemoglobin concentrations recorded from the prefrontal cortex regions among four subject groups and among gender using an intensive experimental design (cognitive, memory, and verbal fluency tests). On the other hand, a collection of DL architectures demonstrated its potential application to fundamentally advance the search for AD multi-classification using fNIRS signals. Three groups of pathological patients with AD were remarkably distinguished from HC with the top highest accuracy of 90.91 and 90.04% achieved by CNN-LSTM and

LSTM, respectively. More importantly, DL could better learn complex and abstract representations of an imbalanced fNIRS dataset and presented its outperformance compared to seven traditional ML classifiers in AD analysis. These findings, thereby, furnish preliminary evidence supporting the potential roles of fNIRS coupled with DL to accurately diagnose and assess the AD severity in future development.

Although this study has yielded the feasibility of utilizing different DL models on the fNIRS dataset to discriminate patients at diverse AD stages, several improvements would be required for the future works. First, it is critically crucial to represent the property of hemodynamic responses of three hemoglobin types of fNIRS signals, such as variance, kurtosis, skewness, or initial dips, to precisely quantify the difference among four groups. Second, since the DL performance counted on the relatively small sample size, which could easily lead to the misclassification problem, either a larger cohort or advanced augmentation techniques used for bio-signal domains should be carried out to extend our present findings. Lastly, although CNN-LSTM was very powerful to result in competitive classification performance, other advanced DL architectures should be developed.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the research was conducted with the research support of the National Research Foundation of Korea and Gwangju Metropolitan City, and approval from the relevant institutions is required. Requests to access the datasets should be directed to JG, james.han.gwak@gmail.com; jgwak@ut.ac.kr.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by IRB Committee of Gwangju Institute of Science and Technology (from IRB No. 20180629-HR-36-07-01 to 20180629-HR-36-07-05). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TH and JG designed the study and wrote the original draft. MK, JK, and JG designed the experimental protocol. MK and JK designed the NIRS system. BK, KL, and J-IS contributed to the collection of patient samples and related data. MK conducted fNIRS data preprocessing steps. TH conducted the experiments and analyzed the obtained results. JG edited the final manuscript. All authors reviewed the manuscript and gave final approval for the manuscript submission.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2022.810125/full#supplementary-material

REFERENCES

Alster, P., Madetko, N. K., Koziorowski, D. M., Królicki, L., Budrewicz, S., and Friedman, A. (2019). Accumulation of tau protein, metabolism and perfusion—application and efficacy of positron emission tomography (PET) and single photon emission computed tomography (SPECT) imaging in the examination of progressive supranuclear palsy (PSP) and corticobasal syndrome (CBS). *Front. Neurol.* 10:101. doi: 10.3389/fneur.2019.00101

Alzheimer's Association (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers Dement.* 15, 321–387. doi: 10.1016/j.jalz.2019.01.010

Alzheimer's Disease International (2018). *World Alzheimer Report 2018. The State of the Art of Dementia Research: New Frontiers.* London: Alzheimer's Disease International, 32–36.

Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., and Havel, J. (2013). Artificial neural networks in medical diagnosis. *J. Appl. Biomed.* 11, 47–58.

Arai, H., Takano, M., Miyakawa, K., Ota, T., Takahashi, T., Asaka, H., et al. (2006). A quantitative near-infrared spectroscopy study: a decrease in cerebral hemoglobin oxygenation in Alzheimer's disease and mild cognitive impairment. *Brain Cogn.* 61, 189–194. doi: 10.1016/j.bandc.2005.12.012

Araki, T., Wake, R., Miyaoka, T., Kawakami, K., Nagahama, M., Furuya, M., et al. (2014). The effects of combine treatment of memantine and donepezil on Alzheimer's disease patients and its relationship with cerebral blood flow in the prefrontal area. *Int. J. Geriatr. Psychiatry* 29, 881–889. doi: 10.1002/gps.4074

Auger, H., Bherer, L., Boucher, É, Hoge, R., Lesage, F., and Dehaes, M. (2016). Quantification of extra-cerebral and cerebral hemoglobin concentrations during physical exercise using time-domain near infrared spectroscopy. *Biomed. Opt. Express* 7, 3826–3842. doi: 10.1364/BOE.7.003826

Beishon, L., Haunton, V. J., Panerai, R. B., and Robinson, T. G. (2017). Cerebral hemodynamics in mild cognitive impairment: a systematic review. *J. Alzheimers Dis.* 59, 369–385. doi: 10.3233/JAD-170181

Benerradi, J., Maior, H. A., Marinescu, A., Clos, J., and Wilson, M. L. (2019). "Exploring machine learning approaches for classifying mental workload using fNIRS data from HCI tasks," in *Proceedings of the Halfway to the Future Symposium*, Vol. 2019, Nottingham, 1–11.

Chen, S., and Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Comput. Methods Programs Biomed.* 110, 111–124. doi: 10.1016/j.cmpb.2012.10.021

Cheng, J., Yang, H., and Zhang, J. (2019). Donepezil's effects on brain functions of patients with Alzheimer disease: a regional homogeneity study based on resting-state functional magnetic resonance imaging. *Clin. Neuropharmacol.* 42:42. doi: 10.1097/WNF.0000000000000324

Cicalese, P. A., Li, R., Ahmadi, M. B., Wang, C., Francis, J. T., Selvaraj, S., et al. (2020). An EEG-fNIRS hybridization technique in the four-class classification of Alzheimer's disease. *J. Neurosci. Methods* 336:108618. doi: 10.1016/j.jneumeth.2020.108618

Cummings, J. L., Dubois, B., Molinuevo, J. L., and Scheltens, P. (2013). International work group criteria for the diagnosis of Alzheimer disease. *Med. Clin.* 97, 363–368. doi: 10.1016/j.mcna.2013.01.001

Dutta, K., Krishnan, P., Mathew, M., and Jawahar, C. (2018). "Improving cnn-rnn hybrid networks for handwriting recognition," in *Proceedings of the 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (Niagara Falls, NY: IEEE), 80–85.

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K. R., et al. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *Neuroimage* 59, 519–529. doi: 10.1016/j.neuroimage.2011.07.084

Feng, C., Elazab, A., Yang, P., Wang, T., Zhou, F., Hu, H., et al. (2019). Deep learning framework for Alzheimer's disease diagnosis via 3D-CNN and FSBi-LSTM. *IEEE Access* 7, 63605–63618. doi: 10.1109/access.2019.2913847

Ferrari, M., and Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage* 63, 921–935. doi: 10.1016/j.neuroimage.2012.03.049

Ghafoor, U., Yaqub, M. A., and Hong, K.-S. (2019). "Performance evaluation of temporal features for detection of mild cognitive impairment: an fNIRS study," in *Proceedings of the 2019 International Automatic Control Conference (CACS)* (Keelung: IEEE), 1–6. doi: 10.3389/fnagi.2020.00141

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., Rueckert, D., and Alzheimer's Disease Neuroimaging Initiative (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *Neuroimage* 65, 167–175. doi: 10.1016/j.neuroimage.2012.09.065

Griffis, J. C., Allendorfer, J. B., and Szaflarski, J. P. (2016). Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. *J. Neurosci. Methods* 257, 97–108.

Haapalainen, E., Kim, S., Forlizzi, J. F., and Dey, A. K. (2010). ", Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Copenhagen, 301–310.

Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class adaboost. *Stat. Interf.* 2, 349–360. doi: 10.4310/sii.2009.v2.n3.a8

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 770–778.

Ho, T. K. K., Gwak, J., Park, C. M., and Song, J.-I. (2019). Discrimination of mental workload levels from multi-channel fNIRS using deep leaning-based approaches. *IEEE Access* 7, 24392–24403. doi: 10.1109/access.2019.2900127

Hock, C., Villringer, K., Müller-Spahn, F., Wenzel, R., Heekeren, H., Schuh-Hofer, S., et al. (1997). Decrease in parietal cerebral hemoglobin oxygenation during performance of a verbal fluency task in patients with Alzheimer's disease monitored by means of near-infrared spectroscopy (NIRS)—correlation with simultaneous rCBF-PET measurements. *Brain Res.* 755, 293–303. doi: 10.1016/s0006-8993(97)00122-4

Hwang, S. J., Adluru, N., Kim, W. H., Johnson, S. C., Bendlin, B. B., and Singh, V. (2019). Associations between positron emission tomography amyloid pathology and diffusion tensor imaging brain connectivity in pre-clinical Alzheimer's disease. *Brain Connect.* 9, 162–173. doi: 10.1089/brain.2018.0590

Irani, F., Platek, S. M., Bunce, S., Ruocco, A. C., and Chute, D. (2007). Functional near infrared spectroscopy (fNIRS): an emerging neuroimaging technology with important applications for the study of brain disorders. *Clin. Neuropsychol.* 21, 9–37. doi: 10.1080/13854040600910018

Jack, C. R. Jr., Albert, M. S., Knopman, D. S., McKhann, G. M., Sperling, R. A., Carrillo, M. C., et al. (2011). Introduction to the recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 257–262.

Kang, S., Cho, S., and Kang, P. (2015). Multi-class classification via heterogeneous ensemble of one-class classifiers. *Eng. Appl. Artif. Intell.* 43, 35–43. doi: 10.1016/j.engappai.2015.04.003

Lancia, S., Cofini, V., Carrieri, M., Ferrari, M., and Quaresima, V. (2018). Are ventrolateral and dorsolateral prefrontal cortices involved in the computerized Corsi block-tapping test execution? An fNIRS study. *Neurophotonics* 5:011019. doi: 10.1117/1.NPh.5.1.011019

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

Lee, D., and Lee, J. (2007). Domain described support vector classifier for multi-classification problems. *Pattern Recogn.* 40, 41–51.

Lee, G., Nho, K., Kang, B., Sohn, K.-A., and Kim, D. (2019). Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci. Rep.* 9:1952. doi: 10.1038/s41598-018-37769-z

Li, R., Nguyen, T., Potter, T., and Zhang, Y. (2019). Dynamic cortical connectivity alterations associated with Alzheimer's disease: an EEG and fNIRS integration study. *Neuroimage Clin.* 21:101622. doi: 10.1016/j.nicl.2018.101622

Li, T., Zhu, S., and Ogihara, M. (2006). Using discriminant analysis for multi-class classification: an experimental investigation. *Knowl. Inf. Syst.* 10, 453–472. doi: 10.1007/s10115-006-0013-y

Liao, J., Li, T., Dong, W., Wang, J., Tian, J., Liu, J., et al. (2019). Reduced prefrontal-temporal cortical activation during verbal fluency task in obsessive-compulsive disorder: a multi-channel near-infrared spectroscopy study. *J. Psychiatr. Res.* 109, 33–40. doi: 10.1016/j.jpsychires.2018.11.006

Lokesh, S., Kumar, P. M., Devi, M. R., Parthasarathy, P., and Gokulnath, C. (2019). An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. *Neural Comput. Appl.* 31, 1521–1531. doi: 10.1007/s00521-018-3466-5

Lowe, V. J., Lundt, E. S., Albertson, S. M., Przybelski, S. A., Senjem, M. L., Parisi, J. E., et al. (2019). Neuroimaging correlates with neuropathologic schemes in neurodegenerative disease. *Alzheimers Dement.* 15, 927–939. doi: 10.1016/j.jalz.2019.03.016

Mangrum, W., Hoang, Q. B. P. B., Amrhein, T. J., Maxfield, C. M., Song, A. W., and Merkle, E. (2018). *Duke Review of MRI Principles: Case Review Series E-Book.* Philadelphia, PA: Elsevier Health Sciences.

Mazaheri, A., Segaert, K., Olichney, J., Yang, J. C., Niu, Y. Q., Shapiro, K., et al. (2018). EEG oscillations during word processing predict MCI conversion to Alzheimer's disease. *Neuroimage Clin.* 17, 188–197.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group* under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology* 34, 939–939. doi: 10.1212/wnl.34.7.939

Metzger, F. G., Schopp, B., Haeussinger, F. B., Dehnen, K., Synofzik, M., Fallgatter, A. J., et al. (2016). Brain activation in frontotemporal and Alzheimer's dementia: a functional near-infrared spectroscopy study. *Alzheimers Res. Ther.* 8, 1–12. doi: 10.1186/s13195-016-0224-8

Naseer, N., and Hong, K.-S. (2015). fNIRS-based brain-computer interfaces: a review. *Front. Hum. Neurosci.* 9:3.

Nguyen, T., Kim, M., Gwak, J., Lee, J. J., Choi, K. Y., Lee, K. H., et al. (2019). Investigation of brain functional connectivity in patients with mild cognitive impairment: a functional near-infrared spectroscopy (fNIRS) study. *J. Biophotonics* 12:e201800298. doi: 10.1002/jbio.201800298

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., and Oh, I.-S. (2019). Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Sci. Rep.* 9:18150.

Perpetuini, D., Bucco, R., Zito, M., and Merla, A. (2017). Study of memory deficit in Alzheimer's disease by means of complexity analysis of fNIRS signal. *Neurophotonics* 5:011010. doi: 10.1117/1.NPh.5.1.011010

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv* [preprint] arXiv:1609.04747,

Shi, X., Wang, T., Wang, L., Liu, H., and Yan, N. (2019). "Hybrid convolutional recurrent neural networks outperform CNN and RNN in task-state EEG detection for Parkinson's disease," in *Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (Lanzhou: IEEE), 939–944.

Shoemaker, J. K., Hogeman, C. S., Khan, M., Kimmerly, D. S., and Sinoway, L. I. (2001). Gender affects sympathetic and hemodynamic response to postural stress. *Am. J. Physiol. Heart Circ. Physiol.* 281, H2028–H2035. doi: 10.1152/ajpheart.2001.281.5.H2028

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv* [preprint] arXiv:1409.1556,

Sirpal, P., Kassab, A., Pouliot, P., Nguyen, D. K., and Lesage, F. (2019). fNIRS improves seizure detection in multimodal EEG-fNIRS recordings. *J. Biomed. Opt.* 24:051408. doi: 10.1117/1.JBO.24.5.051408

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., and Mehler, B. (2014). "Classifying driver workload using physiological and driving performance data: two field studies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Toronto, ON, 4057–4066.

Sonkaya, A. R. (2018). The use of functional near infrared spectroscopy technique in neurology. *NeuroQuantology* 16:87. doi: 10.1255/jnirs.266

Suk, H.-I., and Shen, D. (2013). "Deep learning-based feature representation for AD/MCI classification," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (Berlin: Springer), 583–590. doi: 10.1007/978-3-642-40763-5_72

Swinehart, D. F. (1962). The beer-lambert law. *J. Chem. Educ.* 39:333.

Tsiouris, KM, Pezoulas, V. C., Zervakis, M., Konitsiotis, S., Koutsouris, D. D., and Fotiadis, D. I. (2018). A long short-term memory deep learning network for the prediction of epileptic seizures using EEG signals. *Comput. Biol. Med.* 99, 24–37. doi: 10.1016/j.compbiomed.2018.05.019

Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2019). Understanding disease progression and improving Alzheimer's disease clinical trials: recent highlights from the Alzheimer's disease neuroimaging initiative. *Alzheimers Dement.* 15, 106–152. doi: 10.1016/j.jalz.2018.08.005

Vermeij, A., Kessels, R. P., Heskamp, L., Simons, E. M., Dautzenberg, P. L., and Claassen, J. A. (2017). Prefrontal activation may predict working-memory training gain in normal aging and mild cognitive impairment. *Brain Imaging Behav.* 11, 141–154. doi: 10.1007/s11682-016-9508-7

Vieira, S., Pinaya, W. H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* 74, 58–75. doi: 10.1016/j.neubiorev.2017.01.002

Wang, S., Zhang, Y., Liu, G., Phillips, P., and Yuan, T.-F. (2016). Detection of Alzheimer's disease by three-dimensional displacement field estimation in structural magnetic resonance imaging. *J. Alzheimers Dis.* 50, 233–248. doi: 10.3233/JAD-150848

Wright, A., Damskägg, E.-P., and Välimäki, V. (2019). "Real-time black-box modelling with recurrent neural networks," in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, Birmingham.

Yang, D., and Hong, K.-S. (2019). "Comparison of neural biomarker assessment methods for early detection of Alzheimer's disease," in *Proceedings of the 2019 International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE)* (Bali: IEEE), 34–39.

Yang, Y., Wu, Q., Qiu, M., Wang, Y., and Chen, X. (2018). "Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network," in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)* (Lanzhou: IEEE), 1–7.

Yoo, S.-H., Woo, S.-W., and Amad, Z. (2018). "Classification of three categories from prefrontal cortex using LSTM networks: fNIRS study," in *Proceedings of the 2018 18th International Conference on Control, Automation and Systems (ICCAS)* (PyeongChang: IEEE), 1141–1146.

Yücel, M. A., Selb, J. J., Huppert, T. J., Franceschini, M. A., and Boas, D. A. (2017). Functional near infrared spectroscopy: enabling routine functional brain imaging. *Curr. Opin. Biomed. Eng.* 4, 78–86. doi: 10.1016/j.cobme.2017.09.011

Zeller, J. B., Herrmann, M. J., Ehlis, A.-C., Polak, T., and Fallgatter, A. J. (2010). Altered parietal brain oxygenation in Alzheimer's disease as assessed with near-infrared spectroscopy. *Am. J. Geriatr. Psychiatry* 18, 433–441. doi: 10.1097/JGP.0b013e3181c65821

Zhang, D., Yao, L., Zhang, X., Wang, S., Chen, W., and Boots, R. (2018). "Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface," in *Proceedings of the The 32nd AAAI Conference on Artifical Intelligence (AAAI-18)*, New Orleans, LA, 1703–1710. doi: 10.1109/TCYB.2019.2905157

Zhang, G., Davoodnia, V., Sepas-Moghaddam, A., Zhang, Y., and Etemad, A. (2019). Classification of hand movements from EEG using a deep attention-based LSTM network. *IEEE Sensors J.* 20, 3113–3122. doi: 10.1109/jsen.2019.2956998

Zhang, M.-L., and Zhou, Z.-H. (2007). ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recogn.* 40, 2038–2048. doi: 10.1016/j.patcog.2006.12.019

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.