Comparing Multi-Dimensional fNIRS Features Using Bayesian Optimization-Based Neural Networks for Mild Cognitive Impairment (MCI) Detection

Chutian Zhang, Hongjun Yang, Chen-Chen Fan, Sheng Chen, Chenyu Fan, Zeng-Guang Hou, *Fellow, IEEE*, Jingyao Chen, Liang Peng, Kexin Xiang, Yi Wu, and Hongyu Xie

***Abstract*— The diagnosis of mild cognitive impairment (MCI), a prodromal stage of Alzheimer's disease (AD), is essential for initiating timely treatment to delay the onset of AD. Previous studies have shown the potential of functional near-infrared spectroscopy (fNIRS) for diagnosing MCI. However, preprocessing fNIRS measurements requires extensive experience to identify poor-quality segments. Moreover, few studies have explored how proper multi-dimensional fNIRS features influence the classification results of the disease. Thus, this study outlined a streamlined fNIRS preprocessing method to process fNIRS measurements and compared multi-dimensional fNIRS features with neural networks in order to explore how temporal and spatial factors affect the classification of MCI and cognitive normality. More specifically, this study proposed using Bayesian optimization-based auto hyperparameter tuning neural networks to evaluate 1D channel-wise, 2D spatial, and 3D spatiotemporal features of fNIRS measurements for detecting MCI patients. The highest test accuracies of 70.83%, 76.92%, and 80.77% were achieved for 1D, 2D, and 3D features, respectively. Through extensive comparisons, the 3D time-point oxyhemoglobin feature was proven to be a more promising fNIRS feature for detecting MCI by using an fNIRS dataset of 127 participants. Furthermore, this study presented a potential approach for fNIRS data processing, and the designed models required no manual hyperparameter tuning, which promoted the general utilization of fNIRS modality with neural network-based classification to detect MCI.**

***Index Terms*— Convolutional neural networks (CNN), functional near-infrared spectroscopy (fNIRS), mild cognitive impairment (MCI), multi-dimensional feature evaluation, multilayer perceptron (MLP).**

Manuscript received 15 July 2022; revised 9 November 2022; accepted 22 December 2022. Date of publication 11 January 2023; date of current version 6 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC2001700; in part by the National Natural Science Foundation of China under Grant 62073319, Grant U1913601, and Grant 61720106012; in part by the Beijing Science and Technology Program under Grant Z211100007921021; in part by the Alliance of International Science Organizations (ANSO) Collaborative Research Project under Grant ANSO-CR-PP-2020-03; and in part by the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32040000. *(Corresponding author: Zeng-Guang Hou.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Institute of Automation, Chinese Academy of Sciences under Approval No. IA-201944.

Chutian Zhang and Jingyao Chen are with the Macau Institute of Systems Engineering, Macau University of Science and Technology, Macau, China, and also with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: 2109853pmi30005@student.must.edu.mo).

Hongjun Yang and Liang Peng are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

Chen-Chen Fan, Sheng Chen, and Kexin Xiang are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

Chenyu Fan, Yi Wu, and Hongyu Xie are with the Department of Rehabilitation Medicine, Huashan Hospital, Fudan University, Shanghai 200040, China.

Zeng-Guang Hou is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CASIA-MUST Joint Laboratory of Intelligence Science and Technology, Institute of Systems Engineering, Macau University of Science and Technology, Macau, China (e-mail: zengguang.hou@ia.ac.cn).

Digital Object Identifier 10.1109/TNSRE.2023.3236007

I. Introduction

WHILE the general awareness of mild cognitive impairment (MCI) is low, MCI's high lifetime cost of care for the patient and high conversion rate to Alzheimer's disease (AD) urges the need for screening and treating MCI worldwide. In September 2021, the World Health Organization (WHO) has reported that more than 55 million people live with dementia and nearly 10 million new cases globally every year. Among those dementia cases, 60-70% are AD [1]. According to a model created by the Lewin Group for Alzheimer's Association, for Americans age 65 and older with Alzheimer's or other dementias, the cost of care in 2022 is estimated at $321 billion. Furthermore, the caregivers of those individuals provided an estimated 16 billion hours of unpaid assistance in 2021, valued at $271.6 billion [2]. The above evidence shows the broad impact of AD on diagnosed patients, their families, caregivers, and society. Currently, WHO has concluded that dementia is underdiagnosed worldwide, and even if a diagnosis is made, the patient is typically at a

relatively late stage [3]. Several research groups have made estimations that the delayed onset of AD can significantly reduce the cost of care and increase the individual's lifespan [4], [5]. Thus, early AD diagnosis is crucial for patients to improve awareness and receive timely treatments. Both the Alzheimer's Association and the WHO have set the early diagnosis of AD as one of their principal goals [2], [3].

Individuals with MCI have cognitive decline beyond those expected based on an individual's age and education but may not affect their abilities to perform daily activities [6]. Petersen et al. identified MCI as an intermediate stage between normal aging and early AD [7]. A systematic review has conducted a random-effects meta-analysis of more than 30 studies of MCI and reported that 11.8-21.1% of people aged 60 and older have MCI [8]. Furthermore, studies suggest that approximately 10% to 15% of these individuals progress to AD annually [2], [8]. To diagnose MCI, physicians use patient questionnaires, cognitive assessments, neuroimaging methods, blood tests, and review the patient's medical history [2]. Compared to clinical tests and questionaries conducted by doctors for one single diagnosis, Model-based diagnosis through neuroimaging methods has the unmatched advantage of objectivity and time efficiency.

Neuroimaging methods, such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), and EEG-fNIRS hybrid, were already proven to be effective for detecting MCI by many works [7], [9], [10], [11], [12]. fNIRS signals are measured by transmitting near-infrared (NIR) light onto the scalp by NIR light sources, collecting the backscattered light by photo-detectors at a certain distance, and measuring changes in light attenuation. As the result of neural activities causing increases in oxygen metabolism and oversupplies of cerebral blood flow to compensate for the loss of oxygen in blood [13], there is an overall elevated oxyhemoglobin (HbO) concentration and decreased deoxyhemoglobin (HbR) concentration in the local brain area [14]. Studies have shown that the changes in light attenuations from fNIRS measurements are directly related to changes in hemoglobin concentrations, thus, directly related to neural activities [15], [16], suggesting that fNIRS measurements can provide promising biomarkers for measuring neural activity.

This work aimed to conduct community screenings of MCI by using the fNIRS signals. Pinti et al. have reviewed the pros and cons of using the fNIRS compared with other neuroimaging modalities. The fNIRS modality is lower in price than fMRI systems and non-invasive, perfectly safe, and more comfortable for subjects [17], making them suitable for our community screening set-up. However, fNIRS measurements have a lower temporal resolution, making them unsuitable for real-time systems as EEG signals are [18]. Moreover, fNIRS measurements have a shallower penetration depth compared to fMRI. Those two disadvantages were avoided in our offline classification of MCI because we needed to measure task-related cerebral neural activity rather than real-time classifications through neural activities of deeper brain tissues. Two more challenges for signal processing of the fNIRS measurements are: 1) poor signal quality due to variable signal-to-noise ratio between subjects [19] or caused by rapid head movements, and 2) lack of standardized data processing methods [17]. For researchers and medical staff new to the field, the two points mentioned above and their lack of experience will make it hard to judge the quality of fNIRS signals and lead to possible false diagnoses. Our work summarized and utilized a series of signal preprocessing methods from published works [20], [21], [22] to form an fNIRS data processing algorithm that required fewer manual interventions to remove or correct poor-quality signal segments. The streamlined signal processing algorithm aimed to improve signal quality and was applied to our dataset.

The results from signal processing were often used for identifying MCI by statistical analysis, machine learning, or deep learning methods. According to a review paper by Niu et al., MCI patients showed a delayed increase in averaged HbO concentration change compared with cognitively normal (CN) individuals at the group level [23]. While most works have concluded that MCI patients had lower local blood volume compared to CN individuals [13], [24], some works have observed a higher local blood volume from the MCI group or no difference between the two groups [23], [25]. For subject-level classifications of MCI vs. CN, Yang et al. have employed statistical analysis, machine learning, and deep learning methods to evaluate 15 biomarkers from task-related fNIRS signals. The comparison showed that the convolutional neural network (CNN) for classifying MCI vs. CN was superior to statistical analysis and traditional machine learning methods [26]. Our research found three main challenges we wanted to solve using fNIRS signals and deep learning methods to classify MCI vs. CN. Firstly, while many works used fNIRS signals with deep learning methods for other diagnoses (like detecting the pain intensity, epileptic seizure, and autism spectrum disorder), cortical analysis, and brain-computer interface and showed accurate classification results [27], [28], [29], only a few works utilized deep learning methods to detect MCI vs. CN [26], [30]. Secondly, deep learning methods require a large amount of data to prevent overfitting, yet the existing studies in our area mostly had small datasets of less than 50 participants [29]. Finally, few studies have investigated how multi-dimensional fNIRS features affect classification results, despite machine learning and deep learning having proven effective at diagnosing MCI.

This study compared 1D channel-wise, 2D spatial, and 3D spatiotemporal features of fNIRS measurements for detecting MCI patients. Various studies used 3D features extracted from human brain signals as inputs to their designed 3D CNN architectures [31], [32], [33], [34]. Generally, two types of 3D features were used in current studies regarding human brain signals. The first type was the most commonly used 3D image data based on the human head's x, y, and z coordinates. Yagis et al. utilized 3D MRI data as input for their VGG-16 architecture-inspired 17-hidden-layer deep 3D CNN model for diagnosing AD [31]. In their case, the input size was $176 \times 208 \times 176$. However, in our work, the 3D feature size was maximumly $10 \times 10 \times 7$, making a 3D CNN model at a such depth not feasible. Moreover, although the 3D

image features had abundant spatial information, they lacked temporal information that could help differentiate AD and cognitively normal individuals. The second type of the 3D feature was the 2D topology arrangement with sample points (or time points). Kumar et al. used topography-preserving EEG inputs for their 5-layer 3D-CNN to predict various components of hand movements [32]. Zhang et al. also used 2D spatial distributions of EEG electrodes and sample points to construct 3D EEG tensors as inputs for their cascade and parallel 3D CNN to classify attentive mental states [33]. Although fNIRS with 3D CNN was widely used to classify medical data, most used 3D image data as inputs [19]. To the best of our effort, we only found a few related works using 2D topology arrangement with sample points features. Kwak et al. constructed 3D fNIRS features from 1D fNIRS signals using fNIRS channel spatial locations and timesteps [34]. However, they only utilized the 3D fNIRS features for extracting spatially important regions. They applied that information to 3D EEG features for the classification of mental arithmetic and motor imagery tasks without fully exploring the 3D fNIRS features for classifications. To fully employ our spatial and temporal information of the measured fNIRS signal for classification, we designed our 3D feature as 2D spatial with time points or statistical temporal information for HbO and HbR. Then we utilized those different 3D spatiotemporal features with our (relatively shallower) 3D CNN to classify MCI vs. CN.

The primary task of this work is to use neural networks to classify MCI patients versus CN individuals by fNIRS measurements. The significant contributions of our work are as follows.

1) An fNIRS dataset of 127 MCI and CN participants from community screenings is preprocessed using streamlined processing steps, aiming to reduce the possibility of overfitting of our neural networks.
2) To our knowledge, this is the first work that constructs multilayer perceptron (MLP), 2D CNN, and 3D CNN networks with Bayesian optimization-based auto hyperparameter tuning mechanisms for MCI detection using fNIRS measurements.
3) Multi-dimensional (1D channel-wise, 2D spatial, and 3D spatiotemporal) features are extracted and evaluated through constructed multi-dimensional neural networks on our large dataset. The best performance of an 80.77% test accuracy with 76.92% sensitivity, 83.33% precision, and 80% F1 score is obtained, and the 3D time-point HbO feature with our auto hyperparameter tuning 3D CNN network is recommended to detect MCI patients.

In this article, Section II describes our dataset, including study participants, fNIRS equipment, and the experimental paradigm employed. Section III introduces our preprocessing, feature extracting, and model building methods. Section IV shows the experimental results and our analysis. Finally, Section V presents the conclusions.

II. Data Acquisition

A. Participant

In this study, 154 participants were enrolled for fNIRS data acquisition from Huashan Hospital. None of the enrolled participants had motor or other neurological diseases, and all of them had normal or corrected to normal visual acuity and normal color vision. Moreover, all participants were right-handed, with no acute hearing loss, and with no history of drug abuse, head injury, or use of psychoactive medications that could affect brain's blood flow. Before the experiment, informed consent was obtained according to the procedure approved by the Ethics Committee of Institute of Automation, Chinese Academy of Sciences (approval no. IA-201944). Senior doctors conducted each participant's clinical diagnosis of MCI or CN before fNIRS data acquisition according to the 2018 Guidelines for Diagnoses and Treatments of Dementias and Cognitive Impairments in China [35] and the Diagnostic and Statistical Manual of Mental Disorders—Fifth Edition (DSM-5) [36]. Fig. 1 shows fNIRS measurements were acquired from enrolled 154 participants. Seven subjects' fNIRS data were excluded due to incompletion or lack of tags. There were 20 more subjects' data excluded for bad data quality due to coefficient of variation (CV) rejection calculation. We included the resulting 127 subjects in our experiment for further data preprocessing, feature extraction, and model construction, and their demographics are shown in Table I.

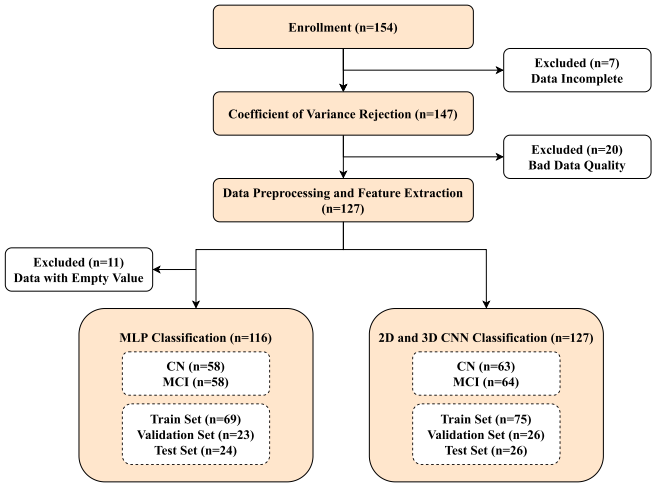

Fig. 1. Flowchart representing study participants, data inclusion criteria, and data partitions for models.

TABLE I
AGE AND GENDER DEMOGRAPHICS OF CN AND MCI GROUPS

	CN (N = 63)	MCI (N = 64)	p-value[a]
Age, years			
Mean	73.6	75.2	0.122
S.D.[b]	7.2	6.1	
Gender, # (%)			
Female	37 (58.7%)	37 (57.8%)	0.459
Male	26 (41.3%)	27 (43.2%)	

[a]Two sample t-test assuming unequal variances (significant level of 0.05)
[b]Standard Deviation

B. Equipment

We acquired fNIRS measurements using a continuous 70-channel fNIRS NirSmart system (Danyang Huichuang

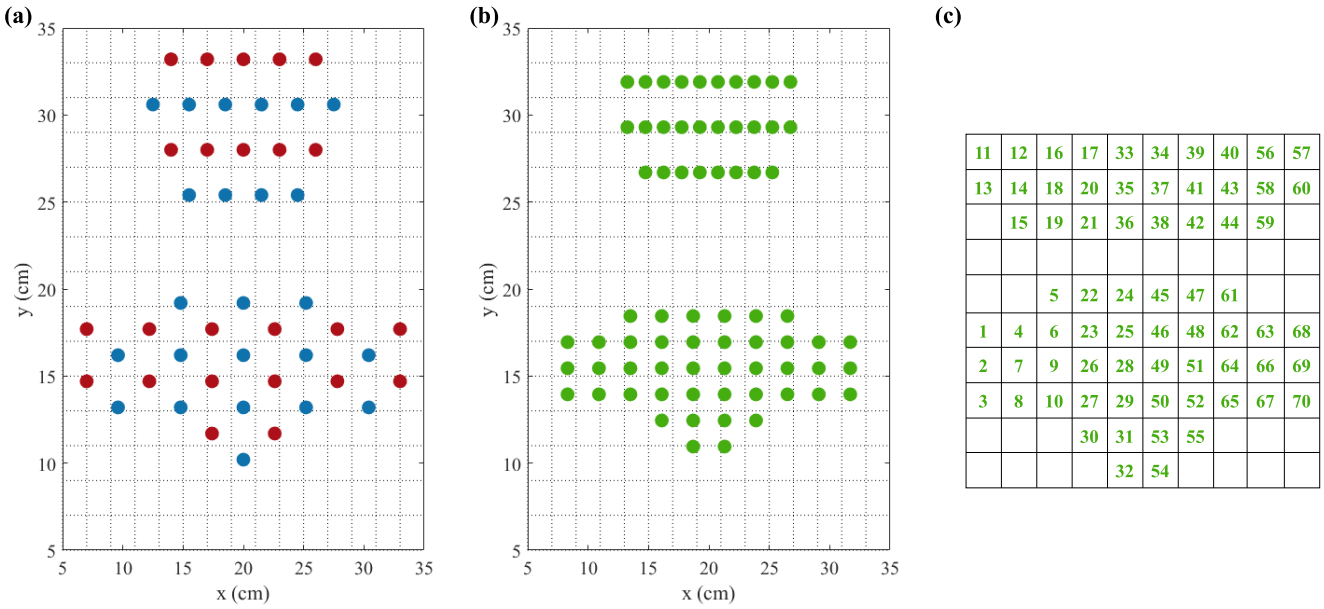

Fig. 2. (a) NIR light source(red) and detector(blue) locations in cm. The origin of the coordination is the right-back side of a head. (b) fNIRS channel locations in cm. Channel locations are approximated to be midpoints of each source-detector pair. (c) A 10 × 10 matrix of fNIRS 2D spatial feature with channel numbers. This arrangement preserved the channel location relationship within the prefrontal cortex and within the parietal cortex while the matrix's size remained small for a faster model processing speed.

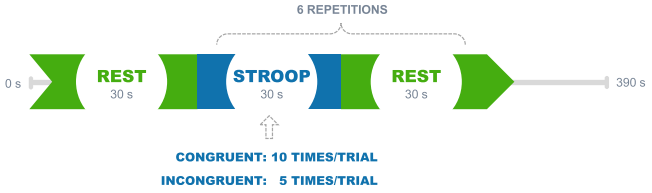

Fig. 3. Our Stroop paradigm lasts 6.5 minutes (390 seconds) for a session and we simultaneously record subject's hemoglobin changes through fNIRS. One session of the experiment contains thirteen 30-second parts starting from 30-second rest period, then 6 Stroop tasks with a rest period following each Stroop task.

Medical Equipment Corporation, China). The fNIRS system contains 24 NIR light sources and 24 photo-detectors and was placed according to the international 10-20 electrode placement system (see Fig. 2(a)). We placed the sensors on the scalp to measure hemoglobin changes of the prefrontal cortex and parietal cortex for cognitive functions. Seventy fNIRS channels (see Fig. 2(b)) were defined as the mid-point of source-detector pairs [14], and all sources and detectors were mounted on an elastic cap to ensure good contact with the subject's scalp. Two wavelengths (730 and 850 nm) at 11 Hz were used to measure the changes in HbO and HbR, respectively.

C. Experimental Paradigm

We asked subjects to sit in a comfortable chair and avoid sudden movements in our experiment. They needed to perform the color-word matching Stroop task while we recorded fNIRS signals simultaneously. Stroop tests were commonly used to test cognitive control and executive function inhibition, and the fNIRS measurements during a Stroop task were proven effective in detecting MCI patients [30]. For our Stroop task shown in Fig. 3, the E-Prime software displayed the stimulus presentation. Each subject was asked to perform one experiment session starting from a 30-second rest period, then six trials of 30-second Stroop tasks with 30-second rest periods following each Stroop task. We designed each Stroop task to have 10 congruent and 5 incongruent tests in a pseudo-random order. For a congruent Stroop test, the color of the displayed word (red, yellow, blue, or green) and the word's meaning match, whereas for an incongruent test, they mismatch. The entire session lasted 6.5 minutes (390 seconds).

III. Methods

A. fNIRS Signal Processing

We experimented and concluded a standardized fNIRS data processing method for our dataset to remove bad signals and artifacts to preserve the validity of the data. For our 154 participants' fNIRS data, we first checked for data completeness(as step-0 in Fig. 4(a)), ensuring they all had 6.5 minutes long of data and valid onset tags for each Stroop trial (6 in total). Seven participants had incomplete data and were excluded from further data processing. Since fNIRS can have poor signal quality due to variable signal-to-noise ratio, we then utilized the coefficient of variation (CV) to exclude channels and trials with poor signal quality [37].

1) CV Rejection: Coefficient of variation (CV) values were calculated to evaluate variable signal-to-noise ratios for unprocessed raw data(as step-1 in Fig. 4(a)). We made rejection strategies for trials, channels, and subjects based on the CV of trials and channels. There were 147 subjects' fNIRS data, each with 70 fNIRS channels of data, and each channel data contained six pieces of data from each Stroop trial.

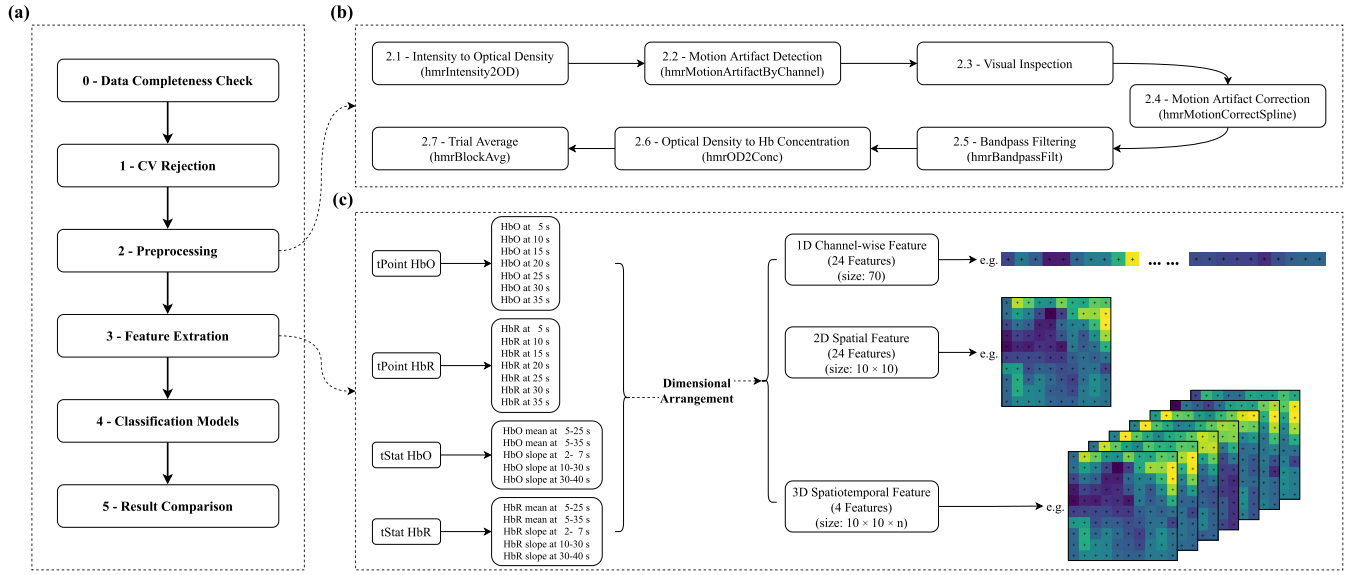

Fig. 4. (a) The overall processing steps. (b) The detailed preprocessing sub-steps (2.1 to 2.7). (c) Left: The total of 24 initially extracted features belong to 4 different feature classes. Right: Rearranging each extracted feature into 1D, 2D, and 3D shapes resulted in the finalized 1D channel-wise, 2D spatial, and 3D spatiotemporal features of fNIRS signal. There were 24 features each for 1D channel-wise and 2D spatial features that correspondingly converted from the initially extracted features and 4 3D spatiotemporal features converted from combining each class of the initially extracted features. The variable n for 3D feature was 7 for tPoint features and 5 for tStat features.

We calculated the CV of trials according to the following formula [37]:

$$CV_{trial}(\%) = \frac{\sigma_{trial}}{\mu_{trial}} \times 100\%, \tag{1}$$

where σ_{trial} and μ_{trial} are the standard deviation and mean of the same 30-second Stroop trial's fNIRS data. Trials with $CV_{trial} > 10\%$ [38] were rejected from further usage since a high CV represents a poor signal-to-noise ratio. Furthermore, since there were six Stroop trials in each channel's data, if more than 50% of trials (that is, more than three trials) were rejected, we rejected that channel for poor signal quality.

The next step was the CV rejection of channels. $CV_{channel}$ was defined as follows:

$$CV_{channel}(\%) = \frac{\sigma_{channel}}{\mu_{channel}} \times 100\%, \tag{2}$$

where $\sigma_{channel}$ and $\mu_{channel}$ are the standard deviation and mean of the same 6.5-minute session of channel data. Channels with $CV_{channel} > 15\%$ [37] were rejected, and if a subject had more than 10% channels (7 channels) rejected, we excluded that subject from further data processing. The calculations of CV were done in MATLAB. Throughout the entire CV rejection process, there were 20 subjects excluded because of poor signal quality, leaving us 127 subjects for data preprocessing.

2) Preprocessing: The data preprocessing of fNIRS was conducted using the HOMER2 toolbox [39]. Referring to Fig. 4(b), for step-2.1, the measured light intensity was converted to the change in optical density for each NIR wavelength by the *hmrIntensity2OD* function. The optical density change $\Delta OD(\lambda, t)$ (unitless) of each λ for time t (in seconds) was defined as [40] and citeScholkmann2014:

$$\Delta OD(\lambda, t) = \ln\left(\frac{I(\lambda, t_0)}{I(\lambda, t)}\right), \tag{3}$$

where λ denotes the NIR light wavelengths, which are 730 and 850 nm in this study. We assume the light intensity emitted by the NIR light source is constant. $I(\lambda, t_0)$ and $I(\lambda, t)$ (in units M) are detected light intensity at time t_0 and t (t_0 is the initial time point) for the corresponding wavelength λ.

For step-2.2, we performed channel-wise motion artifact detection using the function *hmrMotionArtifactByChannel*. In step-2.3, we performed visual inspections on the HOMER2 user interface to improve our signal quality, where the detected motion artifacts would be highlighted. Any highlighted artifact that lasted more than 5 seconds was selected and excluded from further processing. For step-2.4, detected motion artifacts were corrected by spline interpolation using the *hmrMotionCorrectSpline* function. Then, for step-2.5, we used the *hmrBandpassFilt* function to perform bandpass filtering of 0.005 to 0.1 Hz to remove low-frequency baseline drift and high-frequency physiological noise (i.e., Mayer signal-0.1 Hz; respiration-0.25 Hz; and heartbeat-1 Hz).

For step-2.6, we further converted the change of optical density to the change in hemoglobin concentrations (ΔHbO and ΔHbR) at time point t according to the Modified Beer-Lambert Law (MBLL) [41] using function *hmrOD2Conc*. The calculation was as follows:

$$\begin{bmatrix}\Delta HbO(t))\\ \Delta HbR(t))\end{bmatrix} = l^{-1}\begin{bmatrix}\varepsilon_{HbO}(\lambda_1) & \varepsilon_{HbR}(\lambda_1)\\ \varepsilon_{HbO}(\lambda_2) & \varepsilon_{HbR}(\lambda_2)\end{bmatrix}^{-1}\begin{bmatrix}\frac{\Delta OD(\lambda_1,t)}{d(\lambda_1)}\\ \frac{\Delta OD(\lambda_2,t)}{d(\lambda_2)}\end{bmatrix}, \tag{4}$$

where l is source-detector separation (in cm) and ε denotes the molar extinction coefficients in $\mu M^{-1}cm^{-1}$. $d(\lambda)$ (unitless) is the differential path length factor of each wavelength. λ_1 and λ_2 are 730 and 850 nm, respectively.

Finally, for step-2.7, we extracted the trial (block) averages from -5 to 40 seconds of each Stroop trial for each channel using the *hmrBlockAvg* function and exported the processed

data from the HOMER2 toolbox in MATLAB for further feature extractions.

B. Feature Extraction

In this work, there were two types of features in three arrangements. The two types of features were time point features (tPoint) and time-domain statistical features (tStat) for each channel of each subject. Since we were studying Stroop task-induced hemoglobin changes, we extracted tPoint features from t at 5/10/15/20/25/30/35 seconds (where t at 0 seconds is the onset time of the Stroop task) for each hemoglobin type (ΔHbO and ΔHbR) from trial averages resulting from the signal processing stage.

For tStat features, we used mean values of 5 to 25 seconds and 5 to 35 seconds and slope values of 2 to 7 seconds, 10 to 30 seconds, and 30 to 40 seconds for each hemoglobin type from the resulting block averages. The mean value of hemoglobin change of the chosen time range is computed as follows [26].

$$Mean(\Delta Hb(t_1 : t_2)) = \frac{Mean_{signal}(\Delta Hb(t_1 : t_2)) - Mean_{base}}{Mean_{base}}, \quad (5)$$

where ΔHb is ΔHbO or ΔHbR, t_1 and t_2 are the beginning and end times of the chosen time range, $Mean_{signal}$ is the signal mean of the hemoglobin change at the chosen time range, and $Mean_{base}$ is the baseline mean value at the time range of -5 to 0 seconds. The tStat feature of mean value of 5 to 25 seconds was chosen because the task-related initial peak of hemodynamic response usually occurs in the first 20 seconds since task onset [42]. The tStat feature of mean value of 5 to 35 seconds was chosen to represent the overall average of hemoglobin change due to the 30-second Stroop task. The slope values of hemoglobin changes were calculated using *NumPy polyfit* function in Python for each slope range. We used the 2 to 7 seconds slope feature since hemodynamic change usually was delayed by 1 to 2 seconds compared to neural activities and would reach its first peak at 4 to 6 seconds from a single neural response. Furthermore, we expected the hemodynamic change to be steady in the 10 to 30 second period and gradually drop to the baseline value in the 30 to 40 seconds interval, hence the use of 10 to 30 and 30 to 40 seconds slope features.

After we extracted all 24 initial fNIRS features from our processed Stroop-task fNIRS data (Fig. 4(c), left part), we reconstructed the initially extracted features into different dimensional arrangements, respectively. There were three ways we reconstructed the features, namely 1D channel-wise, 2D spatial, and 3D spatiotemporal features (examples shown in Fig. 4(c), right part). 1D channel-wise features were constructed as a series of tPoint or tStat values for each subject in the order of 70 channels from No. 1 to No. 70 (numbered by the equipment used). There was a total of 24 1D channel-wise feature datasets, whereas 7 each for HbO and HbR tPoint features and 5 each for HbO and HbR tStat features.

Channel locations were used to form 2D spatial features. The channel locations were calculated as the mid-point of each source-detector pair [14], and the resulting locations are shown in Fig. 2(b). As we reconstructed the features into 2D features, we wanted to preserve the relative location information while minimizing the size of the feature and the number of interpolated values for empty pixels. The latter was to reduce model computing time and maximize the percentage of the original data. The resulting 2D arrangement was a 10×10 image (shown in Fig. 2(c)), where the numbers represented the channel number. The empty pixels' values were filled by firstly using piecewise cubic interpolations (SciPy CloughTocher2DInterpolator function) and then using nearest-neighbor interpolation (*SciPy NearestNDInterpolator* function) for empty 'corners' (the lower left and lower right corners of our 2D image). Like the 1D channel-wise features, there was a total of 24 2D spatial feature datasets, whereas 14 for tPoint features and 10 for tStat features.

Finally, the 3D spatiotemporal features were constructed by layering each class's 2D spatial features into one 3D feature. There were 4 3D spatiotemporal features corresponded to each class, namely tPoint HbO, tPoint HbR, tStat HbO, and tStat HbR, where the 3D tPoint features had shapes of $10 \times 10 \times 7$ and 3D tStat features had shapes of $10 \times 10 \times 5$. The resulting 1D, 2D, and 3D features were later fed into our constructed models for the classification of MCI and CN.

C. Model Construction

An artificial neural network (ANN) like MLP or CNN contains an input layer, one or multiple hidden layers, and an output layer. For a classification task, an ANN aims to find underlying relationships between the input data and the output label based on layers of artificial neurons and the weights between neurons. That makes ANNs suitable for fNIRS signals because the measured signals are temporally and spatially related. Therefore, in this study, we classified 1D channel-wise features using MLP, 2D spatial features using 2D CNN, and 3D spatiotemporal features using 3D CNN. To the best of our knowledge, this was the first work to use 3D CNN to detect MCI and the first to compare 1D, 2D, and 3D ANNs on the same MCI dataset.

Furthermore, ANNs are suitable for diagnosing MCI using fNIRS signals because they can dynamically adjust parameters learned from errors and have good fault tolerance. Based on that advantage, we designed auto hyperparameter (HP) tuning ANNs to automatically evaluate HP choices and chose the best combinations for a higher classification accuracy with no manual HP tunings needed.

Our network structures are shown in Fig. 5. For our MLP network, the input layer was a dense layer with an input size of 70, which was our number of fNIRS channels. We set the number of neurons and activation function of the input layer as auto-tuned HPs. The activation functions were chosen from the *Keras* package in Python. The first layer of MLP's hidden layer was a normalization layer, where we set whether to normalize the batch dataset and the normalization rate as HPs to be auto-tuned. Next, we implemented a sequence of dense layers, where the number of neurons and activation function were the same as the auto-tuned HPs in the dense input layer.

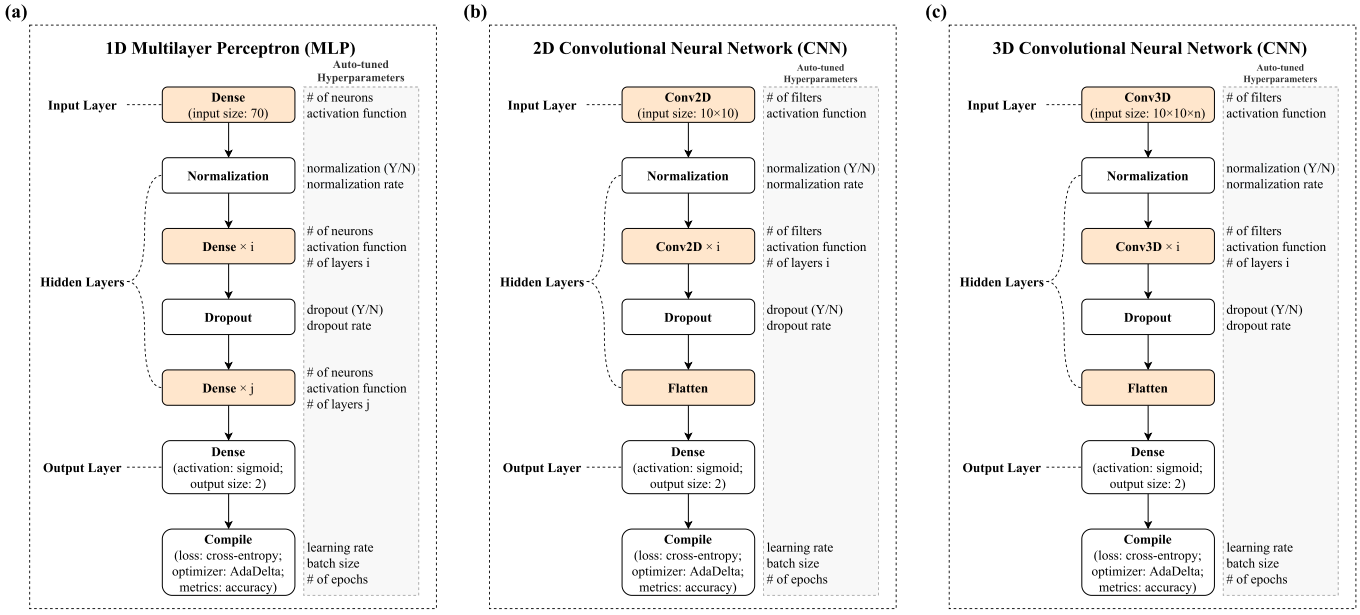

Fig. 5. (a), (b) and (c) are designed auto hyperparameter (HP) tuning MLP, 2D CNN, and 3D CNN network structures, respectively. For each structure, the middle part is the flowchart of the mainframe, and the right part marks the auto-tuned HP of the corresponding layer. The HPs and network structures will be further explained in part III. Methods. The variable n for 3D CNN's input is 7 for tPoint features and 5 for tStat features. Within one network, the values of HPs with the same name for different layers (i.e., number of neurons and activation functions) are identical. The colored (orange) layers of different networks mark the main distinctions between those structures.

However, in this sequence, we set the number of dense layers (as the number i in Fig. 5(a)) to be an auto-tune HP as well, where it was an integer from 1 to 3. Then we implemented a dropout layer and set whether to drop out and the dropout rate to be auto-tuned HPs. We added another sequence of dense layers next, and it was set up in the same manner as the first sequence of dense layers, but with the number of dense layers in this sequence as a separate HP (the number j in Fig. 5(a)). For the last layer of MLP, we added a dense output layer with activation function sigmoid and output size 2 for our binary classification task of MCI vs. CN.

Finally, to configure the model for training, we set the optimizer to be AdaDelta [43], the metrics to be the classification accuracy (the frequency with which the classification results matched the true labels of the subjects), and the loss function to be the cross-entropy loss. The cross-entropy loss $L(\mathbf{w})$ was defined as follows:

$$L(\mathbf{w}) = \frac{1}{N}\sum_{N}^{n=1}\left\{y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\right\}, \quad (6)$$

where $\mathbf{w}$ is the weights of the ANN, N is the dataset size, y_n is the true label, and $\hat{y}_n$ is the classification label. For this compile stage, we set the learning rate, batch size, and the number of epochs as HPs for auto-tuning. Unlike most works that chose typical HPs like the number of neurons, dropout rate, learning rates, and the number of epochs for auto-tuning [29], [44], [45], we also included activation functions, batch size, and most importantly, network structures as our auto-tuned HPs, which were as much we can auto-tune as possible.

The 2D and 3D CNN networks were constructed similarly to the MLP network. There were three differences. Firstly, for the 2D CNN, the input layer was a 2D convolution layer with input size 10×10 and filter size 2×2. Similarly, for the 3D CNN, the input layer was a 3D convolution layer with input size $10 \times 10 \times$ n (n is 7 for tPoint features and 5 for tStat features) and filter size $2 \times 2 \times 2$. The number of filters in 2D and 3D CNN was equivalent to the number of neurons in MLP. Secondly, the first sequence of dense layers in hidden layers of MLP was changed to a sequence of 2D convolution layers for 2D CNN with filter size 2×2 and a sequence of 3D convolution layers for 3D CNN with filter size $2 \times 2 \times 2$. Thirdly, the second sequence of dense layers in MLP was changed to a flatten layer for both 2D and 3D CNN.

We first utilized the Bayesian optimization method for our auto HP tuning networks for our data experiment to find the optimal combination of HPs. Bayesian optimization builds a posterior distribution of functions (Gaussian process) for HP values to the objective function and returns the best combination of HPs expected to be close to the optimum [46], [47]. We partitioned 60% of our dataset (train set in Fig. 1) and utilized 5-fold cross-validation for HP tuning.

The *BayesianOptimization* package in Python was used for our optimization problem. We included all of the hyperparameters (HPs) that needed to be auto-tuned (listed in the shadowed area of Fig. 5) with their respective value ranges as the domain space for optimizing each network structure. The choice range of activation functions included all commonly used functions (for 1D MLP: relu, sigmoid, softplus, softsign, tanh, selu, elu, and exponential, and for 2D CNN and 3D CNN: relu, sigmoid, and tanh). The number of neurons (for 1D MLP), filters (for 2D and 3D CNN), layers, batch size, and epochs were set to be integers at reasonable ranges that were large enough to consider all values to optimize the objective function, yet taken

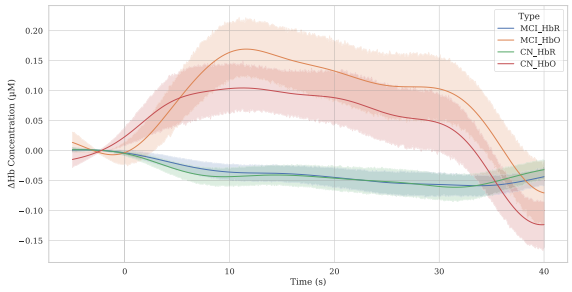

Fig. 6. MCI and CN subjects' average change in HbO and HbR of all channels. Solid lines represent the average hemoglobin changes for all subjects of the same type, and the corresponding shaded area is that type's 95% confidence interval. Time at 0 seconds is the onset of the 30-second Stroop task.

consideration of the size of our data. Moreover, the whether to use a dropout and normalization layer HPs had continuous ranges of 0 to 1, where a value below 0.5 was not to use such layer and vice versa. The dropout rate ranged from 0 to 0.99, the normalization rate ranged from 0 to 1, and the initial learning rate ranged from 0.0001 to 0.5 continuously.

The objective function of the Bayesian optimization was the 5-fold cross-validate average classification accuracy of the MLP, 2D CNN, or 3D CNN using one chosen set of HPs on the training set [48]. The algorithm randomly initialized 15 sets of HPs to evaluate the objective function and returned 15 corresponding classification accuracies. Then, the algorithm predicted how the objective function would vary with HPs and chose where to sample next in domain space with the highest probability to give a maximum classification accuracy [47], [48]. This process was iterated 15 times as well. In Bayesian optimization, the domain space was both explored and exploited, aiming to find a closer estimate of the global maximum of the average classification accuracy. Finally, we chose the set of HPs with the highest average cross-validate classification accuracy as our final hyperparameter set for each model.

After we acquired each set of optimized HPs, we fit the model using 80% of the dataset. At last, our final test results of each model using different features would be concluded by testing the finished model on the remaining 20% of the dataset. This process required fewer manual interventions to find each model's optimal combinations of HPs.

IV. Results

A. Data Observations and Model Training Tactics

Fig. 6 shows the averages and 95% confidence intervals of the change of hemodynamic responses due to the Stroop task of all channels after data processing. From Fig. 6, we can see that after the onset of our Stroop task, the subject's HbO started to increase and reached its peak value at around 8 to 20 seconds. HbO continuously increased during the task period and decreased about 5 seconds after the task (at the 35th second). By comparing HbO changes between CN and MCI, we observed that MCI individuals had a delayed initial increase and a delay in reaching the peak value. Furthermore, MCI patients' HbO change from the 6th second was generally greater than that of CN individuals. On the other hand, there was no significant group difference in the HbR change between MCI patients and CN individuals.

We noticed the overlaps of the confidence intervals not only for the already similar ΔHbR values but also for ΔHbR values of MCI and CN subjects. We went back and observed the ΔHbO and ΔHbR graphs for each subject and noticed vast differences between subjects within the same diagnosis group. The differences between individuals motivated us to conduct our training of the same model using the same feature with different random shuffling seeds to avoid bias caused by the level of similarity between training and testing datasets. We designed our model experiment to each run 5 times with different shuffling seed numbers for partitioning our training, validation, and test datasets. We showed the model experimental results in four aspects, 1) the 5-random-shuffle average performances of each feature (Table II and IV), 2) the best single-run performing features of each dimension of features (in-text and Table V), 3) the comparisons between features and class of features of the same dimensional structure (in-text and Table V), and 4) the comparisons between classes of features of different dimensions (Table III and IV).

To evaluate the performances of different dimensions of features with their corresponding models, we utilized 4 typical metrics: test accuracy, sensitivity, precision, and F1 score. The definition of those metrics was as follows.

$$Test\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{8}$$

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

$$F1\ Score = 2 \times \frac{Sensitivity \times Precision}{Sensitivity + Precision}, \tag{10}$$

where TP, TN, FP, and FN denote the count of true positives, true negatives, false positives, and false negatives, respectively.

B. Classification Results of 1D Channel-Wise and 2D Spatial Features

We performed data experiments on 24 1D channel-wise features for our 1D MLP HP auto-tune network and 24 2D spatial features for our 2D CNN HP auto-tune network. The detailed 5-random-shuffle average performances for each of the features are shown in Table II. We further calculated the overall average class performances of 1D and 2D features, in which we categorized the features into tPoint HbO, tPoint HbR, tStat HbO, and tStat HbR classes, and the result is shown in Table III.

For 1D channel-wise features, the best 5-random-shuffle average performance feature was the ΔHbO at 35 seconds, and the overall average test accuracy was 62.5%. The overall best-performed class was the tPoint HbO features (test accuracy was 59.17%). However, the class differences were close (within 1.34%). By comparing best single runs among all the 1D channel-wise features, the tPoint feature-ΔHbO at 35 seconds and the tStat feature-ΔHbR mean at 5-35 seconds

TABLE II

THE 5-RANDOM-SHUFFLE AVERAGED PERFORMANCE FROM ALL 1D CHANNEL-WISE FEATURES IN THE 1D MLP NETWORK (LEFT) AND ALL 2D SPATIAL FEATURES IN THE 2D CNN NETWORK (RIGHT)

1D Channel-wise Features (5 random shuffle average)		Test Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)
tPoint HbO	HbO at 5 s	59.17	51.67	64.00	55.39
	HbO at 10 s	56.67	63.33	57.31	58.06
	HbO at 15 s	60.83	51.67	65.67	54.91
	HbO at 20 s	61.67	51.67	64.22	52.00
	HbO at 25 s	58.33	63.33	67.08	55.32
	HbO at 30 s	55.00	65.00	56.65	57.65
	HbO at 35 s	62.50	56.67	65.33	59.72
tPoint HbR	HbR at 5 s	59.17	61.67	67.08	56.79
	HbR at 10 s	58.33	56.67	59.40	56.36
	HbR at 15 s	57.50	78.33	55.34	64.57
	HbR at 20 s	57.50	43.33	65.92	47.47
	HbR at 25 s	57.50	58.33	57.86	57.69
	HbR at 30 s	59.17	71.67	61.25	62.57
	HbR at 35 s	61.67	58.33	63.81	59.86
tStat HbO	HbO mean at 5-25 s	58.33	61.67	58.54	59.14
	HbO mean at 5-35 s	60.00	55.00	64.44	56.23
	HbO slope at 2-7 s	59.17	51.67	62.53	55.89
	HbO slope at 10-30 s	53.33	45.00	63.07	44.69
	HbO slope at 30-40 s	58.33	70.00	57.68	61.39
tStat HbR	HbR mean at 5-25 s	60.00	41.67	70.95	47.94
	HbR mean at 5-35 s	60.00	58.33	64.15	57.89
	HbR slope at 2-7 s	58.33	53.33	61.88	55.15
	HbR slope at 10-30 s	54.17	56.67	54.35	48.64
	HbR slope at 30-40 s	61.67	56.67	64.49	59.40

2D Spatial Features (5 random shuffle average)		Test Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)
tPoint HbO	HbO at 5 s	63.08	60.00	63.93	61.90
	HbO at 10 s	62.31	50.77	69.17	56.71
	HbO at 15 s	62.31	64.62	61.76	63.16
	HbO at 20 s	66.15	56.92	69.81	62.71
	HbO at 25 s	58.46	64.62	57.53	60.87
	HbO at 30 s	61.54	67.69	60.27	63.77
	HbO at 35 s	65.38	64.62	65.63	65.12
tPoint HbR	HbR at 5 s	60.77	70.77	58.97	64.34
	HbR at 10 s	60.77	52.31	62.96	57.14
	HbR at 15 s	56.15	72.31	54.65	62.25
	HbR at 20 s	54.62	67.69	53.66	59.86
	HbR at 25 s	53.85	76.92	52.63	62.50
	HbR at 30 s	57.69	60.00	57.35	58.65
	HbR at 35 s	56.92	60.00	56.52	58.21
tStat HbO	HbO mean at 5-25 s	63.08	52.31	66.67	58.62
	HbO mean at 5-35 s	60.00	58.46	60.32	59.38
	HbO slope at 2-7 s	60.77	64.62	60.00	62.22
	HbO slope at 10-30 s	57.69	56.92	57.81	57.36
	HbO slope at 30-40 s	59.23	70.77	57.50	63.45
tStat HbR	HbR mean at 5-25 s	58.46	66.15	57.33	61.43
	HbR mean at 5-35 s	54.62	67.69	53.66	59.86
	HbR slope at 2-7 s	56.15	55.38	56.25	55.81
	HbR slope at 10-30 s	52.31	83.08	51.43	63.53
	HbR slope at 30-40 s	63.08	75.38	60.49	67.12

TABLE III

THE OVERALL AVERAGED CLASS PERFORMANCE OF 4 CLASSES OF FEATURES AND THE TOTAL AVERAGE PERFORMANCE FROM 1D CHANNEL-WISE FEATURES IN THE 1D MLP NETWORK (LEFT) AND 2D SPATIAL FEATURES IN THE 2D CNN NETWORK (RIGHT)

1D Channel-wise Feature Classes (class average)	Test Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)
tPoint HbO	59.17	57.62	62.89	56.15
tPoint HbR	58.69	61.19	61.53	57.90
tStat HbO	57.83	56.67	61.25	55.47
tStat HbR	58.83	53.33	63.16	53.81
Total Average	**58.68**	**57.57**	**62.21**	**56.03**

2D Spatial Feature Classes (class average)	Test Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)
tPoint HbO	62.75	61.32	64.02	62.03
tPoint HbR	57.25	65.71	56.68	60.42
tStat HbO	60.15	60.62	60.46	60.21
tStat HbR	56.92	69.54	55.83	61.55
Total Average	**59.33**	**64.36**	**59.13**	**61.16**

performed the best, with 70.83% test accuracy and 74.07% F1 score.

Next, for 2D spatial features, the best 5-random-shuffle average performance feature was the HbO at 20 seconds, and its average test accuracy was 66.15%. The overall best-performed class was (again) the tPoint HbO features, with an average class test accuracy of 62.75%. The test accuracy of the 2D best-performed class was improved by 3.58% compared with 1D tPoint HbO features. Furthermore, the best-performed single run for 2D features was the ΔHbO at 35 seconds with 76.92% test accuracy and 78.57% F1 score, which was 6.09% and 4.5% improvements compared with 1D best-performed single run features.

Although the class differences between 1D channel-wise features were not distinct, the class differences between 2D spatial features showed some patterns. Table III shows that for 2D feature classes, 1) the HbO features outperformed the HbR features, and 2) the tPoint features had better average performances than the tStat features. The reason that HbO features outperformed the HbR features could be explained by our previous data observations (see Fig. 6), where the ΔHbO between MCI and CN did have distinct differences, yet the ΔHbR of MCI and CN were very similar. Moreover, from the observation that the tPoint features had better performances compared with tStat features, we could deduce that the tPoint features benefited more by adding spatial information compared with tStat features. It was possibly because tPoint features were the original fNIRS measurements with spatial correlations between channels, and tStat features were processed statistical features that may counteract the spatial correlations to a certain degree. Lastly, the difference was not as apparent as we expected by comparing the total average performances between 1D and 2D features. We deduced that the channel numbering from 1 to 70 for 1D channel-wise features still preserved some local spatial correlations (shown in Fig. 2(c)), which reduced the spatial information leverage that 2D spatial features had.

TABLE IV
THE 5-RANDOM-SHUFFLE AVERAGED PERFORMANCE FROM 3D SPATIOTEMPORAL FEATURES IN THE 3D CNN NETWORK

3D Spatiotemporal Features (5 random shuffle average)	Test Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)
tPoint HbO	73.85	69.23	77.06	72.56
tPoint HbR	61.54	64.62	60.87	62.69
tStat HbO	63.85	58.46	65.52	61.79
tStat HbR	60.77	66.15	59.72	62.77
Total Average	65.00	64.62	65.79	64.95

TABLE V
THE BEST SINGLE RUN PERFORMANCE FROM 3D SPATIOTEMPORAL FEATURES IN THE 3D CNN NETWORK

3D Spatiotemporal Features (best single run)	Test Accuracy (%)	Sensitivity (%)	Precision (%)	F1 score (%)
tPoint HbO	80.77	76.92	83.33	80.00
tPoint HbR	65.38	76.92	62.50	68.97
tStat HbO	69.23	61.54	72.73	66.67
tStat HbR	65.38	69.23	64.29	66.67
Total Average	70.19	71.15	70.71	70.57

C. Classification Results of 3D Spatiotemporal Features

Table IV shows the 4 3D spatiotemporal features' 5-random-shuffle classification results from our 3D CNN models. Again, the tPoint HbO feature showed the best average performance with a 73.85% test accuracy and a 72.56% F1 score. The 3D tPoint HbO feature test accuracy outperformed 1D and 2D tPoint HbO features by 14.68% and 11.1%, respectively. Table V shows the best-performed single run for the 3D features was the HbO tPoint feature with 80.77% test accuracy, 76.92% sensitivity, 83.33% precision, and 80% F1 score. The single-run best test accuracy was improved by 9.94% and 3.85% compared with 1D and 2D single runs, respectively.

Comparing the resulting models of 3D CNN with the models of 1D MLP and 2D CNN, they all utilized the same Bayesian optimization hyperparameter auto-tuning framework. The numbers of the second and the third sets of dense/convolution layers were set to be between 1 to 3. Other structural (if to use normalization and dropout layers) and traditional (number of neurons, dropout rate, learning rates, and the number of epochs) hyperparameters were also to be tuned with the same value ranges. Hence, the resulting 3D models benefited from neither network complexity nor superior model tuning techniques. Then, the main difference between the 3D CNN and 1D MLP or 2D CNN was the 3D input and the 3D convolution layers. The 1D MLPs took input data in series and thus failed to leverage the spatial information fully. The 2D CNNs took single matrices as input and failed to leverage context from adjacent matrices. Temporal information for tPoint features and generally more information for tStat features may be helpful for the classification of MCI. The 3D CNNs addressed this issue using 3D convolutional kernels to utilize volumetric patches of 3D inputs. Their ability to take advantage of information within a matrix and adjacent matrices was the main reason for 3D spatiotemporal features with 3D CNNs' leading classification performances.

We also observed that the HbO feature performed better than the HbR feature, and the tPoint feature outperformed the tStat feature for 3D spatiotemporal features. This confirmed our observations from 2D spatial features. The further explanation for our observations from 3D feature performances was that, firstly, the temporal information added to the 3D HbR features was not as effective because the temporal change of ΔHbR was significantly smaller than that of ΔHbO (see Fig. 6). Secondly, stacking 2D tStat features to form 3D tStat features was ineffective because different tStat features were not necessarily correlated. Thus, we could conclude that the fNIRS 3D tPoint feature is the best candidate for 3D CNN classification because of its apparent temporal change and strong spatial correlations.

V. CONCLUSION

In this study, we acquired and processed the Stroop task-induced fNIRS measurements of 127 MCI and CN participants. Then, we extracted 24 initial fNIRS features and constructed them into 1D channel-wise, 2D spatial, and 3D spatiotemporal features. Afterward, we designed MLP, 2D CNN, and 3D CNN networks with an auto hyperparameter tuning mechanism for fNIRS signal MCI detection. We fed the extracted 1D, 2D, and 3D features into the MLP, 2D CNN, and 3D CNN networks for classification and evaluated their performances. We concluded that the 3D tPoint HbO feature was the best fitted and best-performed feature among all our features. The highest performance was by the HbO tPoint feature with 80.77% test accuracy, 76.92% sensitivity, 83.33% precision, and 80% F1 score.

We provided the most promising fNIRS feature for clinical use based on our large dataset and comprehensive comparisons of different fNIRS features. Furthermore, with our streamlined data processing framework and Bayesian optimization-based auto hyperparameter tuning neural network structure requiring no manual intervention, we hope to encourage non-specialists to utilize the fNIRS methodology for MCI diagnosis. We will aim for more accurate classification of MCI and CN for our future work. We plan to further investigate the tPoint HbO features by refining extracted time points and evaluating their importance in detecting MCI patients.

REFERENCES

[1] World Health Organization. (2021). *Dementia Fact Sheet*, pp. 1–5. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] Alzheimer's Association, "2022 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 18, no. 4, pp. 700–789, 2022. [Online]. Available: https://alzjournals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.12638

[3] World Health Organization. *Global Action Plan Public Health Response to Dementia 2017–2025*, 2017. [Online]. Available: https://www.who.int/publications/i/item/9789241513487

[4] J. Zissimopoulos, E. Crimmins, and P. S. Clair, "The value of delaying Alzheimer's disease onset," *Forum Health Econ. Policy*, vol. 18, no. 1, pp. 25–39, 2014.

[5] Alzheimer's Association, "Changing the trajectory of Alzheimer's disease: How a treatment by 2025 saves lives and dollars," pp. 1–20, 2015. [Online]. Available: https://www.alz.org/media/Documents/changing-the-trajectory-r.pdf

[6] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild cognitive impairment: Clinical characterization and outcome," *Arch. Neurol.*, vol. 56, no. 3, pp. 303–308, 1999.

[7] R. C. Petersen and D. Bennett, "Mild cognitive impairment: Is it Alzheimer's disease or not?" *J. Alzheimer's Disease*, vol. 7, no. 3, pp. 241–245, 2005.

[8] R. C. Petersen et al., "Practice guideline update summary: Mild cognitive impairment," *Neurology*, vol. 90, no. 3, pp. 126–135, Jan. 2018.

[9] X. Cui et al., "Classification of Alzheimer's disease, mild cognitive impairment, and normal controls with subnetwork selection and graph kernel principal component analysis based on minimum spanning tree brain functional network," *Frontiers Comput. Neurosci.*, vol. 12, pp. 1–12, May 2018.

[10] T. K. K. Ho et al., "Improving the multi-class classification of Alzheimer's disease with machine learning-based techniques: An EEG-fNIRS hybridization study," *Alzheimer's Dementia*, vol. 17, no. S7, Dec. 2021, Art. no. e057565.

[11] B. Grässler et al., "Multimodal measurement approach to identify individuals with mild cognitive impairment: Study protocol for a cross-sectional trial," *BMJ Open*, vol. 11, no. 5, May 2021, Art. no. e046879.

[12] C.-C. Fan et al., "Group feature learning and domain adversarial neural network for aMCI diagnosis system based on EEG," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 9340–9346.

[13] K. Shaw et al., "Neurovascular coupling and oxygenation are decreased in hippocampus compared to neocortex because of microvascular differences," *Nature Commun.*, vol. 12, no. 1, pp. 1–16, May 2021.

[14] J. León-Carrión and U. León-Domínguez, "Functional near-infrared spectroscopy (fNIRS): Principles and neuroscientific applications," in *Neuroimaging-Methods*, P. Bright, Ed. Rijeka, Croatia: InTech, 2012, ch. 3. [Online]. Available: http://www.intechopen.com/books/neuroimaging-methods/functional-nearinfrared-spectroscopy-fnirs-brain-studies-and-others-clinical-uses

[15] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, pp. 150–157, Jul. 2001.

[16] O. J. Arthurs and S. J. Boniface, "What aspect of the fMRI BOLD signal best reflects the underlying electrophysiology in human somatosensory cortex?" *Clin. Neurophysiol.*, vol. 114, no. 7, pp. 1203–1209, Jul. 2003.

[17] P. Pinti et al., "The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience," *Ann. New York Acad. Sci.*, vol. 1464, no. 1, pp. 5–29, 2020.

[18] A. Kawala-Sterniuk et al., "Summary of over fifty years with brain-computer interfaces—A review," *Brain Sci.*, vol. 11, no. 43, pp. 1–41, 2021.

[19] M. Ferrari and V. Quaresima, "A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application," *NeuroImage*, vol. 63, no. 2, pp. 921–935, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2012.03.049

[20] Y.-C. Liu, Y.-R. Yang, Y.-A. Tsai, R.-Y. Wang, and C.-F. Lu, "Brain activation and gait alteration during cognitive and motor dual task walking in stroke—A functional near-infrared spectroscopy study," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 12, pp. 2416–2423, Dec. 2018.

[21] S. Jahani et al., "fNIRS can robustly measure brain activity during memory encoding and retrieval in healthy subjects," *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, Aug. 2017.

[22] S. Brigadoi et al., "Motion artifacts in functional near-infrared spectroscopy: A comparison of motion correction techniques applied to real cognitive data," *NeuroImage*, vol. 85, pp. 181–191, Jan. 2014.

[23] M. K. Yeung and A. S. Chan, "Functional near-infrared spectroscopy reveals decreased resting oxygenation levels and task-related oxygenation changes in mild cognitive impairment and dementia: A systematic review," *J. Psychiatric Res.*, vol. 124, pp. 58–76, May 2020.

[24] H. Niu, X. Li, Y. Chen, C. Ma, J. Zhang, and Z. Zhang, "Reduced frontal activation during a working memory task in mild cognitive impairment: A non-invasive near-infrared spectroscopy study," *CNS Neurosci. Therapeutics*, vol. 19, no. 2, pp. 125–131, Feb. 2013.

[25] K. H. Yap et al., "Visualizing hyperactivation in neurodegeneration based on prefrontal oxygenation: A comparative study of mild Alzheimer's disease, mild cognitive impairment, and healthy controls," *Frontiers Aging Neurosci.*, vol. 9, p. 287, Sep. 2017.

[26] D. Yang, K.-S. Hong, S.-H. Yoo, and C.-S. Kim, "Evaluation of neural degeneration biomarkers in the prefrontal cortex for early identification of patients with mild cognitive impairment: An fNIRS study," *Frontiers Human Neurosci.*, vol. 13, pp. 1–17, Sep. 2019.

[27] K. Khalil, U. Asgher, and Y. Ayaz, "Novel fNIRS study on homogeneous symmetric feature-based transfer learning for brain–computer interface," *Sci. Rep.*, vol. 12, no. 1, pp. 1–12, Feb. 2022.

[28] H. Hamid, N. Naseer, H. Nazeer, M. J. Khan, R. A. Khan, and U. S. Khan, "Analyzing classification performance of fNIRS-BCI for gait rehabilitation using deep neural networks," *Sensors*, vol. 22, no. 5, p. 1932, Mar. 2022.

[29] C. Eastmond, A. Subedi, S. De, and X. Intes, "Deep learning in fNIRS: A review," 2022, *arXiv:2201.13371*.

[30] M.-K. Kang and K.-S. Hong, "Application of deep learning techniques to diagnose mild cognitive impairment: Functional near-infrared spectroscopy study," in *Proc. 21st Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2021, pp. 2036–2042.

[31] E. Yagis, L. Citi, S. Diciotti, C. Marzi, S. W. Atnafu, and A. G. S. De Herrera, "3D convolutional neural networks for diagnosis of Alzheimer's disease via structural MRI," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 65–70.

[32] N. Kumar and K. P. Michmizos, "A neurophysiologically interpretable deep neural network predicts complex movement components from brain activity," *Sci. Rep.*, vol. 12, no. 1, pp. 1–12, Jan. 2022, doi: 10.1038/s41598-022-05079-0.

[33] Y. Zhang, H. Cai, L. Nie, P. Xu, S. Zhao, and C. Guan, "An end-to-end 3D convolutional neural network for decoding attentive mental state," *Neural Netw.*, vol. 144, pp. 129–137, Dec. 2021.

[34] Y. Kwak, W.-J. Song, and S.-E. Kim, "FGANet: fNIRS-guided attention network for hybrid EEG-fNIRS brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 329–339, 2022.

[35] Writing Group for Chinese Guidelines of Dementia and Cognitive Impairment Diagnoses and Treatments and Committee of Cognitive Disorders of Neurologist Branch of Chinese Medical Doctor Association, "2018 guidelines for diagnoses and treatments of dementias and cognitive impairments in China (5): The diagnosis and treatment of mild cognitive impairment," *Nat. Med. J. China*, vol. 98, no. 17, pp. 1294–1301, 2018. [Online]. Available: http://rs.yiigle.com/CN112137201817/1039079.htm

[36] American Psychiatric Association, "Neurocognitive disorders," in *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Arlington, VA, USA: American Psychiatric, 2013, ch. Sec. 2, pp. 591–643.

[37] S. K. Piper et al., "A wearable multi-channel fNIRS system for brain imaging in freely moving subjects," *NeuroImage*, vol. 85, pp. 64–71, Jan. 2014.

[38] C. F. Lu, Y. C. Liu, Y. R. Yang, Y. T. Wu, and R. Y. Wang, "Maintaining gait performance by cortical activation during dual-task interference: A functional near-infrared spectroscopy study," *PLoS ONE*, vol. 10, no. 6, pp. 1–22, 2015.

[39] T. J. Huppert, S. G. Diamond, M. A. Franceschini, and D. A. Boas, "HomER: A review of time-series analysis methods for near-infrared spectroscopy of the brain," *Appl. Opt.*, vol. 48, no. 10, pp. 1–33, 2009.

[40] W. B. Baker, A. B. Parthasarathy, D. R. Busch, R. C. Mesquita, J. H. Greenberg, and A. G. Yodh, "Modified Beer-Lambert law for blood flow," *Biomed. Opt. Exp.*, vol. 5, no. 11, pp. 4053–4075, 2014.

[41] F. Scholkmann et al., "A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology," *NeuroImage*, vol. 85, no. 1, pp. 6–27, Jan. 2014.

[42] D. Yang et al., "Detection of mild cognitive impairment using convolutional neural network: Temporal-feature maps of functional near-infrared spectroscopy," *Frontiers Aging Neurosci.*, vol. 12, no. 141, pp. 1–17, May 2020.

[43] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.

[44] S. D. Wickramaratne and M. S. Mahmud, "Conditional-GAN based data augmentation for deep learning task classifier improvement using fNIRS data," *Frontiers Big Data*, vol. 4, pp. 1–12, Jul. 2021.

[45] F. Eitel et al., "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation," *NeuroImage, Clin.*, vol. 24, Jan. 2019, Art. no. 102003.

[46] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, vol. 24, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2011, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf

[47] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, vol. 25, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf

[48] H. Cho, Y. Kim, E. Lee, D. Choi, Y. Lee, and W. Rhee, "Basic enhancement strategies when using Bayesian optimization for hyperparameter tuning of deep neural networks," *IEEE Access*, vol. 8, pp. 52588–52608, 2020.