IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 14, 2022, accepted May 4, 2022, date of publication May 9, 2022, date of current version May 16, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3173629

Machine Learning Approach for Classifying College Scholastic Ability Test Levels With Unsupervised Features From Prefrontal Functional Near-Infrared Spectroscopy Signals

JUNGGU CHOI[1], INHWAN KO[2], YOONJIN NAH[2], BORA KIM[3], YONGWAN PARK[4], JIHYUN CHA[5], JONGKWAN CHOI[5], AND SANGHOON HAN[1,2]

[1]Yonsei Graduate Program in Cognitive Science, Yonsei University, Seoul 03722, Republic of Korea
[2]Department of Psychology, Yonsei University, Seoul 03722, Republic of Korea
[3]Department of Counselling, Honam University, Gwangju 62399, Republic of Korea
[4]Department of Business Administration, Gyeongsang National University, Jinju 52828, Republic of Korea
[5]OBELAB Inc., Seoul 06211, Republic of Korea

Corresponding author: Sanghoon Han (sanghoon.han@yonsei.ac.kr)

This work was supported in part by the Yonsei Signature Research Cluster Program of 2021 under Grant 2021-22-0005, in part by the National Research Foundation of Korea (NRF) Grant through the Korea Government [Ministry of Science and ICT (MSIT)] under Grant 2019R1A2C1007399, in part by the Seoul Research and Development Program under Grant BT190004, in part by the Ministry of Education of the Republic of Korea, and in part by the NRF under Grant 2020S1A5A2A03042694.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board at Yonsei University under Application No. 7001988-202104-HR-659-06, and the guidelines of the Declaration of Helsinki.

ABSTRACT Learning ability evaluation has been critical in educational and medical fields to investigate learning achievement or cognitive impairment. Previous researchers utilized biosignal data such as functional near-infrared spectroscopy and an electroencephalogram to reflect neural variation in factors related to learning ability. Additionally, machine learning algorithms have been used to identify the inherent associations between learning ability and related factors. Herein, we propose a classification framework for college scholastic ability test levels using unsupervised features extracted from a functional near-infrared spectroscopy signal dataset based on machine learning models. To extract unsupervised features from functional near-infrared spectroscopy signals, we constructed a one-dimensional convolutional autoencoder with an electroencephalogram dataset as a transfer learning approach. Eight handcrafted features (signal mean, slope, minimum, peak, skewness, kurtosis, variance, and standard deviation) with various window length conditions were calculated to compare influences on classification performance. Five evaluation metrics (accuracy, precision, recall, F1-score, and area under the curve) were applied to evaluate the proposed framework's performance. Among the five classification algorithms (XGBoost classifier, support vector classifier, naive Bayes classifier, decision tree classifier, and logistic regression), the XGBoost classifier was the best at classifying college scholastic ability test levels. We found that unsupervised features extracted from deep learning algorithms are more usable for classification than handcrafted features. Furthermore, the applicability of transfer learning between two different neural modals was validated using the experimental results. The results of this study provide new insights into the relationships between hemodynamics in functional near-infrared spectroscopy signals and college scholastic ability test levels.

INDEX TERMS College scholastic ability test, functional near-infrared spectroscopy, learning ability, machine learning, transfer learning.

I. INTRODUCTION

Evaluation or validation of learning ability has been widely investigated by researchers in the educational and medical

The associate editor coordinating the review of this manuscript and approving it for publication was M. Shamim Kaiser.

domains. In particular, many researchers in the educational field have tried to identify education or study levels in student groups [1]–[3]. Kellaghan and Greaney [4] suggested an assessment of students' learning achievement to improve education quality. They focused on the models and standards of a national assessment. Pereira *et al.* [5] introduced learner-centered assessment methods for higher education through a review of previous studies. In addition, the influences of examinations and written tests in relation to learning evaluation were validated through an analysis.

In the medical domain, Prigoff *et al.* [6] assessed the effectiveness of medical education achievement in the increased virtual learning environment amid the coronavirus disease 2019 outbreak. Researchers have proposed the need for adjustments in curricula based on exam scores from student groups. Further, Barulli *et al.* [7] proposed test tools for memory capacity to detect cognitive impairment in a clinical setting. This memory test showed strength for neuropsychological evaluation.

Many methodologies (e.g., surveying, testing, and counseling) have been used in previous allied studies to measure the learning ability of study participants. Dermo [8] applied an online questionnaire to evaluate the use of e-assessment as a means to improve the quality of student learning. Le and Tam [9] validated eight assessment methods, including seminars and open-book tests, to compare the effectiveness of these methods with regard to students' understanding and attitudes. Additionally, Burnett [10] used counseling to assess the learning outcomes of participants. The author suggested that strategies and techniques need to maximize learning outcomes in counseling. Further, scholastic assessment tests have been widely used to evaluate overall academic achievement or level in specific subjects such as mathematics [11]–[14].

To evaluate diverse conditions in learning ability, previous studies have applied neural-related measurements (e.g., functional magnetic resonance imaging, electroencephalography, and functional near-infrared spectroscopy). Denervaud *et al.* [15] evaluated errors in learning in Montessori and traditionally schooled children. They used brain functional magnetic resonance imaging (fMRI) data to identify patterns associated with errors in children. Kim *et al.* [16] collected electroencephalogram (EEG) signals from college student groups to investigate influences of indoor thermal conditions on college students' learning performance. Relationships between the detailed ability of students, including working memory and executive ability, and thermal conditions, were examined. Firooz and Setarehdan [17] recorded functional near-infrared spectroscopy (fNIRS) and EEG signals from graduate students to estimate intelligence quotient (IQ) test scores. The researchers validated the usability of fNIRS and EEG as evaluation modalities in their research.

Various analysis methodologies have been utilized to analyze potential relations from the neural variations of participants. Howard *et al.* [18] examined associated brain regions with cognitive load for several tasks. Each region-related task was identified through a t-test and partial least squares analysis. Daly *et al.* [19] validated motivations for learning mathematics using EEG signals. Significant differences in prefrontal signals were verified using a t-test. Sugiura *et al.* [20] utilized fNIRS signals to evaluate performance in second-language-learning among young adolescents. Signals from the regions of interest were compared using a generalized linear modeling method.

Based on the aforementioned studies, recent studies have utilized machine learning and deep learning to identify latent patterns of neural data. Mao *et al.* [21] proposed a deep learning classification algorithm to classify the fMRI of attention deficit/hyperactivity disorder patients. The proposed framework of the authors showed a state-of-the-art performance compared with previously proposed algorithms. Evgin *et al.* [22] attempted to classify bipolar disorder using fNIRS signals on the basis of convolutional neural network models. They demonstrated the possibility of fNIRS analysis using feed-forward neural network algorithms.

Based on previous studies, we developed a classification framework using machine learning algorithms for learning ability levels based on fNIRS signals. To prescribe the operational definition of "learning ability," we collected and utilized college scholastic ability test (CSAT) scores from 73 participants. Additionally, the collected scores were set as a dependent variable of machine learning algorithms. Further, we hypothesized that unsupervised features extracted from deep learning algorithms can show better performance than the handcrafted features used in previous studies for fNIRS classification tasks. To validate this hypothesis, we included in our research design the construction of deep learning models as a feature extractor and comparisons between extracted features and calculated handcrafted features.

To construct deep learning models for feature extraction, the collected fNIRS dataset was insufficient to train and evaluate algorithms from scratch. We utilized EEG signals with characteristics similar to those of fNIRS signals for algorithm training in terms of transfer learning. Zhang *et al.* [23] adopted a transfer learning approach to evaluate deep convolutional neural networks using an EEG dataset. Moreover, EEG and fNIRS signals showed several common advantages, such as high temporal resolution, over other neural modals. Trambaiolli *et al.* [24] determined signal properties between neuro-electric (i.e., EEG) and neuro-hemodynamic (i.e., fNIRS) on the basis of their analysis results. They focused on not only the characteristics of time-series data, but also the capabilities of describing hemodynamic alterations in the occipital/visual cortex using EEG signals. As a result, we concluded that the application of deep learning algorithms trained by EEG signals was reasonable for feature extraction.

Furthermore, we attempted to examine the possibility of transfer learning without additional fine-tuning in deep learning algorithms between datasets collected from the same domain. Peng *et al.* [25] compared the model's performance based on transfer learning between five similar image datasets. They verified the potential of the transfer learning approach based on their experimental results. In addition, unsupervised algorithms trained with a single dataset were evaluated using four datasets, excluding the fine-tuning steps. Zhong *et al.* [26] used three classification algorithms with datasets collected from different domains in the transfer learning approach. Each trained algorithm was verified and compared without additional fine-tuning. Referring to these previous studies, deep learning algorithms trained with EEG signals were utilized without additional training with fNIRS signals as feature extractors.

In this study, we developed a five-step research scheme. First, suitable participants with regard to age and CSAT scores were recruited, and experiments using eight-session task materials were conducted to collect fNIRS signals. Second, the collected fNIRS signals were preprocessed and converted from raw signals to HbO (oxyhemoglobin) and HbR (deoxyhemoglobin) concentration signals. Third, one-dimensional convolutional autoencoder models were developed using the EEG dataset as a feature extractor for unsupervised feature extraction. Handcrafted and unsupervised features were extracted from the preprocessed HbO and HbR concentration signals. Fourth, five machine learning classifiers were trained with extracted features (handcrafted and unsupervised features) to classify CSAT levels. Finally, the classification performance of each classifier was compared to identify the optimized algorithms and conditions for our research topics. The overall research scheme is shown in Figure 1.

This work provides three main contributions to the field:

- We propose a novel classification framework based on machine learning algorithms for CSAT levels using fNIRS signals.
- The applicability of a one-dimensional convolutional autoencoder model trained with EEG signals as a feature extractor was validated in terms of transfer learning.
- We checked the usability of unsupervised features for classification through comparisons with handcrafted features.

The remainder of this paper is organized as follows. In Section II, we present the detailed procedures and methods for developing our proposed framework for CSAT-level classification. In Section III, we present the experimental results to evaluate our machine learning-based framework. In Section IV, we explain the significance and implications of our research. Finally, we conclude the paper in Section V.

II. METHODS

A. PARTICIPANTS FOR fNIRS DATASET COLLECTION

To collect fNIRS signals, we recruited participants from undergraduate freshman groups at three different universities (Yonsei University, Honam University, and Gyeongsang National University). Seventy-three healthy undergraduate students participated (mean age: 19.20; female: 41, male: 32).

We used the NIRSIT Lite device of OBELAB Inc. (Seoul) to collect information on the hemodynamic activities

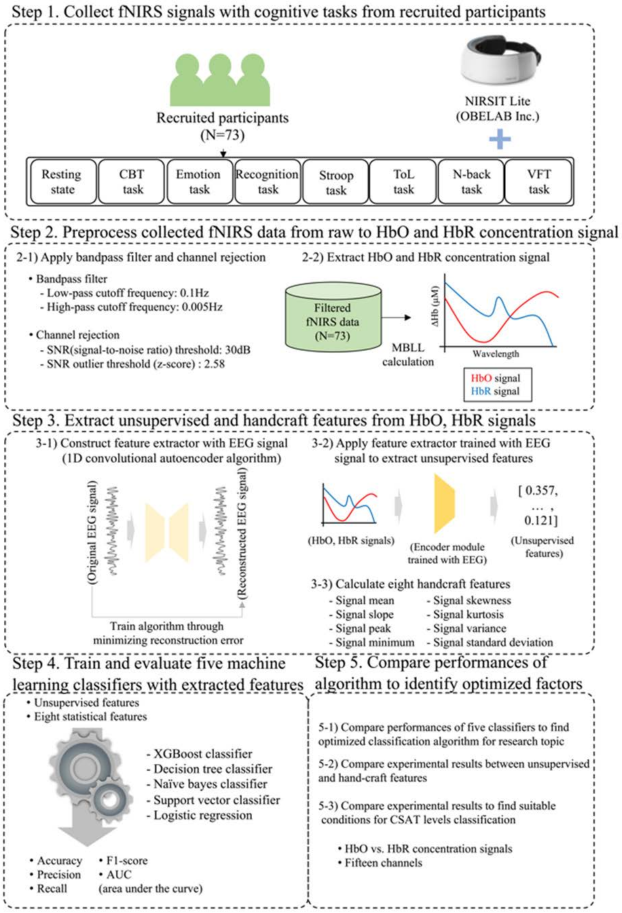
Step 1. Collect fNIRS signals with cognitive tasks from recruited participants
Recruited participants
(N=73)
NIRSIT Lite
(OBELAB Inc.)
Resting state
CBT task
Emotion task
Recognition task
Stroop task
ToL task
N-back task
VFT task
Step 2. Preprocess collected fNIRS data from raw to HbO and HbR concentration signal
2-1) Apply bandpass filter and channel rejection
• Bandpass filter
- Low-pass cutoff frequency: 0.1Hz
- High-pass cutoff frequency: 0.005Hz
• Channel rejection
- SNR(signal-to-noise ratio) threshold: 30dB
- SNR outlier threshold (z-score) : 2.58
2-2) Extract HbO and HbR concentration signal
Filtered fNIRS data (N=73)
MBLL calculation
ΔHb (μM)
Wavelength
HbO signal
HbR signal
Step 3. Extract unsupervised and handcraft features from HbO, HbR signals
3-1) Construct feature extractor with EEG signal (1D convolutional autoencoder algorithm)
(Original EEG signal)
(Reconstructed EEG signal)
Train algorithm through minimizing reconstruction error
3-2) Apply feature extractor trained with EEG signal to extract unsupervised features
(HbO, HbR signals)
(Encoder module trained with EEG)
[0.357, ..., 0.121]
(Unsupervised features)
3-3) Calculate eight handcraft features
- Signal mean
- Signal slope
- Signal peak
- Signal minimum
- Signal skewness
- Signal kurtosis
- Signal variance
- Signal standard deviation
Step 4. Train and evaluate five machine learning classifiers with extracted features
• Unsupervised features
• Eight statistical features
- XGBoost classifier
- Decision tree classifier
- Naïve bayes classifier
- Support vector classifier
- Logistic regression
• Accuracy
• Precision
• Recall
• F1-score
• AUC (area under the curve)
Step 5. Compare performances of algorithm to identify optimized factors
5-1) Compare performances of five classifiers to find optimized classification algorithm for research topic
5-2) Compare experimental results between unsupervised and hand-craft features
5-3) Compare experimental results to find suitable conditions for CSAT levels classification
• HbO vs. HbR concentration signals
• Fifteen channels

FIGURE 1. Overview of the research scheme in this study.

of participants. The sensor array of this device consisted of five dual-wavelength laser diodes (780 and 850 nm) and seven photodetectors separated by an 8 mm unit distance. The optical signal collected from each channel was sampled at 8.138 Hz. The laser and detector pairs were separated by a distance of 3 cm. The position schemes of optodes in the NIRSIT Lite device are depicted in Figure 2 [27].

Prior to the experiment, we explained the fNIRS signal collection procedure to the participants, and all experiments were conducted after obtaining their consent. The experiments were designed and conducted in accordance with the guidelines of the Declaration of Helsinki and institutional review board approval at Yonsei University (7001988-202104-HR-659-06).

B. EEG DATASET

In this study, we compared unsupervised features extracted using deep learning models and handcrafted features. To extract features from fNIRS signals, one-dimensional convolutional autoencoder models were trained and evaluated using EEG signals. An open-source brain-computer interface (BCI) IV competition EEG dataset was utilized for

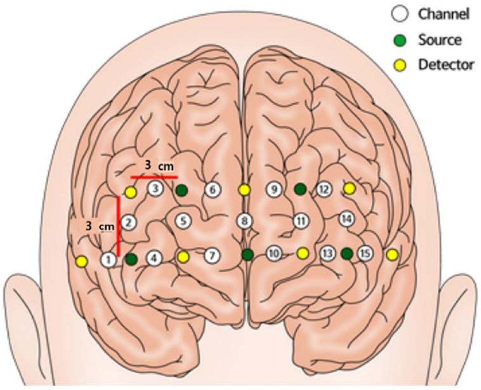

FIGURE 2. Optode positions of the NIRSIT Lite device. The yellow circles indicate detectors for reflected light. Green circles show the positions of the light source. The channels, lying between each source and detector pair, are shown in numbered white circles. The device mainly covers the frontopolar cortex, also roughly reflecting the signals from other adjacent regions in Prefrontal cortex including dorsolateral prefrontal cortex, ventrolateral prefrontal cortex, and orbitofrontal cortex.

our feature extractor [28]. This dataset was provided by the Berlin BCI group (Berlin Institute of Technology, Fraunhofer FIRST, and University Medicine Berlin). Researchers from the BCI group devised six experimental sessions to evaluate motor imagery in participants. Three options for the motor imagery task (the left hand, right hand, and foot) were sequentially assigned visual cues. EEG signals included in the BCI competition IV dataset were collected from seven healthy subjects (all male; aged 26 to 46 years) using 59 channel devices.

C. EXPERIMENTAL PROCEDURE

To examine hemodynamic variation in various tasks, we put together consecutive sessions with widely used cognitive tasks in related previous studies. Eight sessions (resting state and seven cognitive tasks) were finally selected and provided to participants. First (resting state step), brain activity in the resting state was measured to investigate hemodynamic changes before conducting the seven cognitive tasks [29]. Second (Corsi block-tapping task step), participants were instructed to select consecutive positions of colored boxes in the Corsi block-tapping task (CBT) [30]. The positions of the yellow box were changed sequentially to evaluate the memory of participants with regard to the stimulus sequence in the task. Third (emotion task step), human-face pictures with specific expressions were proposed to the participants for the selection of emotions. The participants were shown the pictures for a few seconds. Thereafter, the pictures disappeared and they were configured to offer a choice [31]. Fourth (recognition task step), to evaluate the recognition capacities of participants, facial pictures without expressions or emotions were shown on the monitor [32]. Recognition was assessed by having participants answer whether they saw the picture in the previous task by selecting appropriate buttons. Fifth (Stroop task step), participants were provided with colored words to select the color of words in the Stroop task [33]. In addition, participants were asked to verify the meaning of the words regardless of the word. Sixth (Tower of London task step), participants performed the Tower of London tasks and were provided with three beads and sticks to assess their planning and problem-solving capabilities [34]. Seventh (N-back task step), participants took part in the N-back task, which involved the sequential positions of stars in a grid [35]. Participants selected the previous positions of stars in three types of tasks (1-back, 2-back, and 3-back). Finally (verbal fluency task), participants' verbal fluency was evaluated by speaking related words in a limited period [36]. In addition, the time at which the word was spoken was recorded to identify the continuity of the answers. The procedures for the cognitive tasks and their time periods are listed in Table 1.

TABLE 1. Procedure and time periods of the eight sessions in this study.

No.	Cognitive task	Time periods
1	Resting state (no task)	5 min
2	Corsi block-tapping task	12 min
3	Emotion task	5 min
4	Recognition task	6 min
5	Stroop task	10 min
6	Tower of London task	8 min
7	N-back task	13 min
8	Verbal fluency task	8 min

D. fNIRS SIGNAL PREPROCESSING

After data were collected from participants via the cognitive tasks, fNIRS data were preprocessed to remove artifacts or noise in the signal. To exclude physiological and environmental noise, band-pass filtering with a range of 0.005 to 0.1 Hz was used. In addition, a 30 dB signal-to-noise ratio was applied to qualify the noise of detected channels. After applying two processes (band-pass filter and signal-to-noise ratio), we calculated relative changes in oxy-Hb (i.e., HbO concentration signal) and deoxy-Hb (i.e., HbR concentration signal) using the modified Beer–Lambert law (MBLL) [37], [38].

E. EEG SIGNAL PREPROCESSING

To construct a feature extractor (i.e., a one-dimensional convolutional autoencoder), we utilized the BCI IV EEG dataset. Unlike previous studies that analyzed in detail (e.g., power spectrum analysis or spectral analysis), we preprocessed only basic characteristics of EEG signals (e.g., sampling frequency and scale of signal values). The EEG signals included in the BCI IV dataset consisted of 1000 Hz signals. Based on previous studies [39], [40], we downsampled the data to 100 Hz and normalized the scale of values to range from -1 to $+1$ to achieve faster training of algorithms.

TABLE 2. Dimensions of preprocessed EEG datasets for developing one-dimensional convolutional autoencoder algorithms.

Length of values	Training dataset No. of rows, No. of columns)	Validation dataset (No. of rows, No. of columns)	Test dataset (No. of rows, No. of columns)
680	(227013, 680)	(37128, 680)	(37011, 680)
700	(213211, 700)	(36231, 700)	(36129, 700)
720	(206322, 720)	(35162, 720)	(35517, 720)
750	(198062, 750)	(33747, 750)	(34101, 750)
780	(190452, 780)	(32450, 780)	(32804, 780)

TABLE 3. Experimental conditions for hyperparameter setting of the feature extractor.

No.	Length of the input layer	Length of the latent vector	Batch size	Learning rate	Epochs	No. of layers
1	680	20	2048	0.0003	500	8
2	700	45	2048	0.0003	500	8
3	720	60	2048	0.0003	500	8
4	750	170	2048	0.0003	500	8
5	780	343	2048	0.0003	500	8

After the aforementioned steps for preprocessing, we composed three datasets for training and evaluation of deep learning algorithms. Signals collected from seven participants were divided into training (five participants), validation (single participant), and test (single participant) datasets. In the case of the training dataset, five participants were randomly assigned. Additionally, two other participants were also randomly assigned to the validation and test datasets.

To identify optimal hyperparameters such as the length of layers and size of latent vectors in deep learning models, we compared datasets and the different lengths of the EEG dataset. Referring to previous studies using similar algorithms, we selected five conditions for the length of the input layer (i.e., length of input signal data) and the length of latent vectors (i.e., length of extracted vectors) [41]. Accordingly, five datasets with different signal lengths were composed for the construction of feature extractors. The training, validation, and test datasets were included in each dataset. Further, we checked that the dimensions of the datasets were similar for each condition. The detailed dimensions of each dataset are listed in Table 2.

F. CONSTRUCTION OF FEATURE EXTRACTOR

To extract unsupervised features from fNIRS signals, we utilized one-dimensional convolutional autoencoder algorithms. Additionally, as mentioned in the previous paragraph, we compared five conditions to determine the optimal hyperparameters for the algorithms. The detailed conditions for the comparison are listed in Table 3.

The reconstruction performance of each condition was evaluated using three evaluation indices (root mean squared error, mean relative error, and mean absolute error) and a comparison of figures. Among the five conditions for hyperparameter setting, the third condition (720 length of input layer and 60 length of latent vector) showed lower index values than the other conditions (the detailed results are explained in Section III.). Based on these results, we trained and evaluated algorithms with a 720-length of input layers and 60 latent vector conditions. The detailed structures of the one-dimensional convolutional autoencoder are listed in Table 4. We utilized the encoder module (layers 1–4) as a feature extractor.

TABLE 4. Model structure of the one-dimensional convolutional autoencoder.

Layer (Module)	Layer	No. of input channels	No. of output channels	Kernel size	Stride	Other conditions
(input)	One-dimensional vector (720 length)					
1	1D Conv	1	8	30	2	Batch norm
2	1D Conv	8	16	20	2	Batch norm
3	1D Conv	16	32	30	2	Batch norm
4	1D Conv	32	64	20	2	Batch norm
5	1D transConv	64	32	20	2	Batch norm
6	1D transConv	32	16	30	2	Batch norm
7	1D transConv	16	8	20	2	Batch norm
8	1D transConv	8	1	30	2	Batch norm
(Output)	Reconstructed one-dimensional vector (720 length)					

1D conv: one-dimensional convolutional layer; 1D transConv: 1 dimensional transposed convolutional layer; Batch norm: batch normalization.

G. EXTRACTION OF UNSUPERVISED FEATURES

We obtained HbO and HbR signals following the previously mentioned preprocessing steps. To extract unsupervised features from the feature extractor, preprocessed HbO and HbR signals were divided into 720-length units. Single signals of 720-length were applied to the feature extractor. From the feature extractor (encoder module), one-dimensional 60-length vectors with 64 output channels were obtained. Extracted feature vectors were converted to single vectors by averaging channels without changes in length.

H. CALCULATION OF HANDCRAFTED FEATURES

To compare the influence of unsupervised features and handcrafted features on classification performance, we calculated eight handcrafted features used in previous studies.

1) SIGNAL MEAN

The signal means of HbO and HbR concentration signals were calculated as follows:

$$\mu_w = \frac{1}{N_w} \sum_{i=i_1}^{i_2} HbX(i) \tag{1}$$

In this formula, μ_w is the mean value for a given window. Subscript w indicates the window for the calculation. i_1 and i_2 denote the start and end points of the window, respectively. N_w is the number of signal values in the window, and HbX

refers to the HbO or HbR concentration signal data. In many previous studies, signal mean values were utilized for classification in BCI research [42], [43].

2) SIGNAL SLOPE

To extract the signal slope features, we referred to the calculation methods used in previous studies [44]. The highest and lowest signal values in the window were compared. The signal slope features were calculated as follows:

$$Slope_w = H_w - L_w \tag{2}$$

where subscript w indicates the window, and $Slope_w$ is the calculated signal slope feature value. H_w and L_w denote the highest and lowest values in the window, respectively.

3) SIGNAL PEAK

The signal peak feature is the peak value of the signal values in the window. Some previous studies have shown that peak value features worked best in fNIRS research [45], [46].

4) SIGNAL MINIMUM

The signal minimum feature is the minimum value of the signal in a given window. In associated studies on fNIRS-BCI, authors validated the usability of these features [47]–[49].

5) SIGNAL SKEWNESS AND KURTOSIS

The signal skewness feature was calculated as follows:

$$Skewness_w = \frac{E_x(HbX_w - \mu_w)^3}{\sigma^3} \tag{3}$$

where $Skewness_w$ indicates the skewness feature value calculated from the signal values in the window. σ^3 in the denominator represents the standard deviation of the HbO or HbR concentration signal value for the given window. In the numerator, μ_w denotes the mean value in the window, and E_x denotes the expectation of HbO or HbR signal. The signal kurtosis was computed as follows:

$$Kurtosis_w = \frac{E_x(HbX_w - \mu_w)^4}{\sigma^4} \tag{4}$$

where $Kurtosis_w$ indicates the calculated kurtosis feature value. These features (skewness and kurtosis) have been utilized in related fNIRS research [50], [51].

6) SIGNAL VARIANCE AND STANDARD DEVIATION

We calculated the variance and standard deviation values for a given window. These features have also been reported as being effective in fNIRS research [52], [53].

I. CLASSIFICATION ALGORITHMS

We applied five machine learning classifiers for our research topics (CSAT level classification using fNIRS features). The first classification algorithm was decision tree classifiers [54]. This classification algorithm is mainly composed of flow charts, such as tree structure flow charts (nodes and branches). The tree was built in two phases. First, in the build (growth) phase, the training dataset was split recursively based on local optimal criteria until the samples included in the dataset belonged to each of the partitions in the same class labels. Second, to prevent overfitting of the models, noise and outliers were removed in the pruning phase. Moreover, the second phase was conducted using fully grown trees. In terms of the model structure, three sub-structures (internal nodes, branches, and leaf nodes) consisted of these algorithms. We utilized decision tree classifiers with an iterative dichotomiser 3 (ID3) algorithm. ID3 algorithms use information gain to select the splitting attribute. Further, information gain represents the variation of entropy values. In summary, information gain was calculated using the difference in entropy before and after splitting.

The second classification algorithm was logistic regression [55]. A maximum likelihood estimation method was used to estimate the coefficients of the regression models. Subsequently, the regression model calculated a likelihood value L(x), where $0 \leq \text{L(x)} \leq 1$. The association between class label and input vectors was indicated by the likelihood values. If the likelihood values were higher than the threshold (0.5), the class was classified as having high CSAT levels in binary cases. In the three class condition Y, we considered Y as a specified value of either "low," "middle," or "high." As a result, the logistic regression model calculated the probability values to categorize each class under diverse class conditions.

The third classifier was a naive Bayes algorithm [56]. This probabilistic classifier utilizes the Bayes theorem. All attributes in the dataset are assumed to be independent.

Support vector classifiers (SVC) were used as the fourth classification algorithm [47]. In our study, this classifier was applied using a nonlinear kernel (radial basis kernel). The feature space of the dataset was classified using hyperplanes separated by class labels. In the research by Bhavsar & Panchal [58], the authors compared classification performance via SVC models under linear, polynomial, and radial basis kernel conditions. They showed advantages of radial basis kernels for high dimensional classification tasks. Therefore, to evaluate the classification performance of the different algorithms under various class conditions, we selected a radial basis kernel with non-linear characteristics. Additionally, the participants in the dataset were completely separated to prevent overfitting of the models.

Finally, the XGBoost classifier was utilized to compare the classification performance with the aforementioned algorithms [59]. This classifier was an ensemble of several decision-tree models. Furthermore, this model comprised gradient-boosting algorithms with regularized objectives. We minimized the regularized objective function to optimize the algorithms. The differences between the predicted y_i' and target yi were compared in differential convex loss function. Penalization term was added to adjust the complexity of the models. An additional regularization term smoothens the last learned weight to avoid overfitting. In our study, we assigned categories of CSAT levels (e.g., "low," "middle," and "high" in three class case) in y_i.

J. EVALUATION METRICS

To compare the classification performances between algorithms, we applied five evaluation metrics. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) were calculated from a confusion matrix to evaluate performance using other indices with accuracy. TP and TN values indicate the number of correctly classified samples. In contrast, FP and FN values represent incorrectly classified samples. Using the four basic values from the confusion matrix, we obtained four additional indicators (precision, recall, F1-score, and accuracy). Additionally, the true positive rate and false positive rate were checked to draw the receiver operating characteristic (ROC) curve. Further, the performance of the algorithms was evaluated using area under the curve values from the ROC curve.

K. TRAINING AND EVALUATION OF MACHINE LEARNING CLASSIFIERS

We utilized datasets consisting of features (both handcrafted and unsupervised) and class labels to train and evaluate five machine learning algorithms. To validate the classification performance of classifiers in various class conditions, we set three class conditions in our experiments (three, four, and five classes). Further, detailed experimental conditions based on the characteristics of fNIRS signals (HbO or HbR, channels of signal) and window length for handcrafted feature extraction were applied to compare the effects of the conditions on the classification performance. For example, in the case of the characteristics of fNIRS signals, the HbO and HbR signals were separately applied for feature extraction. Additionally, the fNIRS dataset used in our study consisted of distinguished signals collected from each of the 15 channels. Individual signals were separately applied for comparison. Further, nine different window length conditions (from 2 s length to 10 s length of window) were used for handcrafted features to compare and validate the influence of features on performance in our research settings. As a result, we conducted experiment with 32,400 conditions (8 features × 9window length × 15 channels × 5models × 3class labels × HbO and HbR = 32,400) for handcraft feature conditions and 450 conditions for unsupervised feature conditions (15 channels × 3class labels × 5 models × HbO and HbR = 450).

To train and evaluate the algorithms, we utilized 10-fold cross-validation to prevent overfitting. The number of rows in the dataset was the same (1,530 rows) for both handcrafted and unsupervised feature conditions. Additionally, the number of columns (i.e., features) differed according to the feature conditions. For example, in the case of unsupervised feature conditions, the dimension of the dataset was (1530, 61). Unlike unsupervised feature conditions, the number of columns in the handcrafted feature condition differed on the basis of length of the windows. The average number of columns in a handcrafted feature was 706. Furthermore, to handle imbalance of the number of instances in each class label, we applied weights to training of machine learning algorithms. The examples of distribution of the class label in the dataset were depicted in Figure 3.

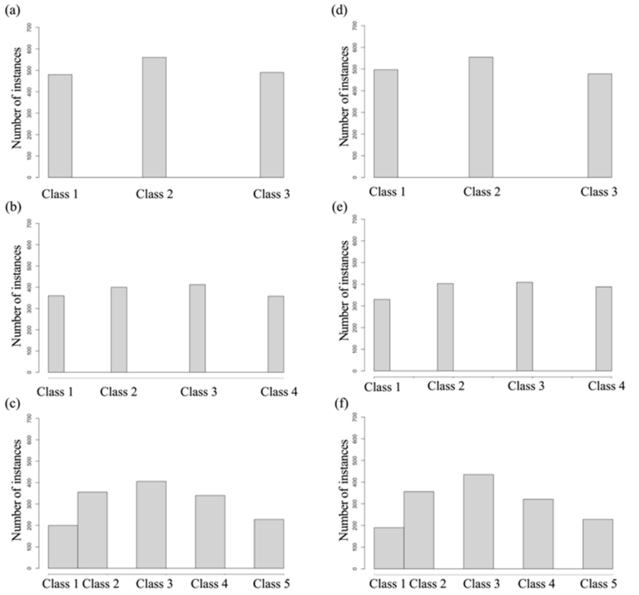

FIGURE 3. Distribution of class labels in the dataset for the training and evaluation of the machine learning models. (a), (b), (c): class label distribution of the dataset with the handcrafted features from the HbO and channel 1 condition signals. (d), (e), (f): class label distribution of the dataset with the handcrafted features from the HbR and channel 1 condition signals. (a), (d): three class labels. (b), (e): four class labels. (c), (f): five class labels.

L. TOOLS

All codes for deep learning models, ML classifiers, and data preprocessing were written using Pytorch (version 1.4.1), Python (version 3.7.1; scikit-learn, version 2.4.1), and R (version 4.0.3).

III. RESULTS

A. CONSTRUCTION OF THE FEATURE EXTRACTOR

To extract unsupervised features from fNIRS signals, we constructed a one-dimensional convolutional autoencoder as a feature extractor. Five hyperparameter conditions were compared to confirm the model structure and training parameters. Among the five conditions, the third condition (720 length of input layer and 60 length of latent vector) showed the lowest error values among the three error indices. The detailed error values are listed in Table 5. Additionally, we checked the similarity between the original and reconstructed signals in terms of visualization. The visualization of EEG signals is depicted in Figure 4.

B. CLASSIFICATION PERFORMANCE OF ML CLASSIFIERS

We evaluated the classification performances of five ML classifiers based on various experimental conditions for feature extraction and class labels. Among the five algorithms (decision tree classifier, logistic regression, naive Bayes classifier, support vector classifier, and XGBoost classifier) for classification, the XGBoost classifier showed the best classification performance. Detailed experimental results from XGBoost and other classifiers were presented in Appendix A and B. The performance of the XGBoost classifier in classifying CSAT levels under unsupervised feature conditions is shown in Tables 6-8.

In the case of handcrafted features, XGBoost classifiers showed that the averaged values of the evaluation metrics were approximately 79%. In contrast, the classification performance was relatively higher using the unsupervised features extracted from the deep learning algorithms under all experimental conditions (averaged evaluation metric value was 87%).

IV. DISCUSSION

In our study, we attempted to classify CSAT levels based on machine learning classifiers with several features extracted from fNIRS signals. Unsupervised features and handcrafted features were extracted from fNIRS signals in different processes to compare the effects for CSAT-level classification.

To propose reasonable evidence for our research topics (classifying CSAT levels through machine learning algorithms with fNIRS signals), we identified several associated previous studies on two aspects (learning ability evaluation with neuro-related dataset and analysis with machine or deep learning algorithms).

First, considering the relationship between neuro-related datasets (e.g., EEG or fNIRS) and learning ability, Kaewkamnerdpong [60] evaluated human learning ability using neuroimaging (EEG and fNIRS signals). He suggested that utilizing the real-time brain state for evaluation of the target learning ability was valuable from the experimental results. Artemenko *et al.* [61] collected fNIRS signals in event-related potential (ERP) measurements to investigate an individual's math ability. The authors compared the variations of fNIRS waves with arithmetic materials between high and low performers. Soltanlou *et al.* [62] examined cognitive development related to mathematics and language using fNIRS signals. Brain activation changes were measured during the language- and mathematics skills-related experiments in schoolchildren groups. Based on previous research, including the aforementioned studies, we concluded that the application of neuroimaging techniques (especially fNIRS) was suitable for the classification of CSAT levels as learning ability measurement.

Second, in terms of analysis through machine learning or deep learning algorithms, Benerradi *et al.* [63] applied machine learning and deep learning classifiers to classify mental workload status in a continuous human–computer interaction (HCI) research with an fNIRS dataset. They checked the promise of machine learning models for fNIRS analysis in their research. Hosseini *et al.* [64] discovered discriminative characteristics within fNIRS data collected from children with language disorders. A total of five machine learning classifiers were used to detect hemodynamic differences in healthy and disordered groups. Rojas *et al.* [65] suggested a classification framework for pain assessment using fNIRS signals collected from nonverbal patients. K-nearest neighbor algorithms were used for pain assessment. The authors focused on the advantages of machine learning models to investigate functional biomarkers for pain using fNIRS signals. Based on the previous mentioned studies, we verified that machine learning models have the potential to analyze fNIRS datasets for CSAT level classification. As a result, we confirmed that our research topic regarding CSAT level classification with fNIRS signals based on machine learning algorithms was well founded.

TABLE 5. Errors of five experimental conditions for the feature extractor.

Condition	RMSE	MRE	MAE
1	0.1968	0.0348	0.0128
2	0.2262	0.7871	0.0149
3	**0.0163**	**0.0622**	**0.0109**
4	0.1033	0.2654	0.0795
5	0.1432	0.1363	0.0942

RMSE: root mean squared error; MRE: mean relative error; MAE: mean absolute error.

To reflect variations in fNIRS signals for CSAT level classification, we utilized several features used in previous studies. Khan and Hong [66] extracted eight features (mean oxyhemoglobin, mean deoxyhemoglobin, skewness, kurtosis, signal slope, number of peaks, sum of peaks, and signal peak) from prefrontal fNIRS signals. The extracted features were applied to classify the neural states between alert and drowsy states. A total of 15 window conditions were used to extract the features. Yoo *et al.* [67] extracted mean, slope, kurtosis, and skewness features to decode multiple sound categories from fNIRS collected from the auditory cortex. Yang *et al.* [68] considered seven features (HbO mean, HbR mean, HbO slope, HbR slope, time to peak in hemodynamic response, skewness, and kurtosis) extracted from fNIRS signals as digital biomarkers to identify mild cognitive impairment (MCI). Among the diverse features utilized in previous studies, we found and extracted eight common features to compare the influences about classification performance with unsupervised features in our experimental settings. In addition, differences in window length conditions for feature extraction have been examined in previous studies [69]–[71]. In this regard, we extracted and applied handcrafted features with nine conditions (from 2 s length to 10 s length) for the length of windows.

In many previous studies that analyzed datasets using machine learning or deep learning algorithms, datasets collected from enough participants were utilized for research. For example, Jang *et al.* [72] used an electrocardiogram (ECG) signal dataset measured from 1,278 patients to train

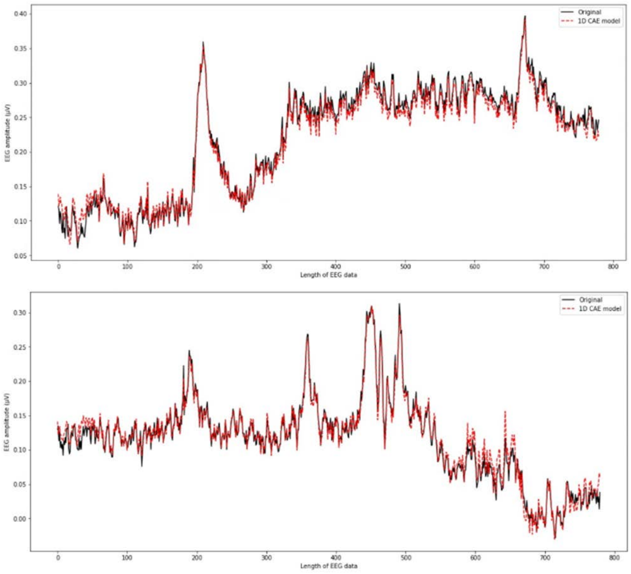

FIGURE 4. **Visualization results of the original EEG signal and reconstructed EEG signals from the one-dimensional convolutional autoencoder. Black graphs indicate the original EEG signal and red graphs indicate the reconstructed EEG signal.**

deep learning models. They showed that the amount of data was sufficient to train and evaluate algorithms. Additionally, Jang *et al.* [41] utilized an actigraphy dataset gathered from 14,482 healthy individuals for analysis. However, in our study, we obtained fNIRS signals from 73 participants. We considered that the number of participants could be relatively insufficient for analysis using deep learning algorithms directly. To overcome the shortage of datasets, we selected a transfer learning approach based on previous studies [73], [74]. To extract unsupervised features from deep learning algorithms, one-dimensional convolutional autoencoder models were trained using the BCI IV EEG dataset preferentially. Whether the characteristics of the EEG data were well reflected in model parameters was verified through a comparison of five hyperparameter conditions between the reconstructed and actual signals. As a result, an encoder module trained by EEG signals was used as a feature extractor to extract fNIRS unsupervised features. Moreover, in the case of the dataset collected from the same domain (i.e., EEG and fNIRS), we attempted to examine whether the pre-trained algorithm was suitable for feature extraction without additional fine-tuning.

From our experimental results, the overall classification performances were compared to find optimized algorithms for our research topics. Among the five classifiers, the XGBoost classifier showed the highest evaluation metric values under all experimental conditions. Similar results have been reported in previous studies. Zhu *et al.* [75] classified major depressive disorder groups using machine learning algorithms based on collected fNIRS signals. In this case, the performance of the XGBoost classifier was higher than that of the random forest classifier. Additionally, Khan *et al.* [76] verified the suitability of XGBoost classifiers in a finger movement classification task using an fNIRS dataset. As a result, we confirmed that the XGBoost algorithms were most suitable for fNIRS signal analysis for CSAT-level classification in our research scheme.

In experimental conditions with handcrafted features, averaged classification performances of XGBoost classifiers in eight features were compared to locate the common window length and signal conditions (i.e., HbO and HbR) for each feature extraction. First, in the case of signal slope features, a 4 s length window and HbO concentration signal condition were commonly found in three class conditions (three, four, and five classes). Noori *et al.* [77] utilized signal slope features with a relatively short window length for calculations. They verified that the experimental conditions using slope features showed the best classification performance.

Second, in conditions with signal peak features, we checked that HbR concentration signal and 3 s length of window was found in all conditions. Third, the HbO signal and 6 or 7 s length window (three class conditions: 6 s window length condition, four class conditions: 7 s window condition, and five class conditions: 6 s window condition) were found for signal standard deviation feature conditions.

Finally, under the HbR signal condition for signal standard deviation features, we found that 4 or 5 s length windows (three class conditions: 4 s window length condition, four class conditions: 5 s window condition, and five class conditions: 4 s window condition) commonly showed the best performance. Ghaffar *et al.* [78] used signal standard deviation features with a 5 s window for classification in their fNIRS-BCI research. The authors used the KNN and LDA algorithms with standard deviation features and verified higher accuracy than for other frameworks proposed in benchmark studies. By comparing the classification performances between our research and previous studies, we identified similar tendencies in our experimental results with regard to handcrafted features.

Further, to find suitable channel conditions for classification utilizing handcrafted features, we compared the frequency of channels based on the best classification performance in each condition. Fifteen channels were divided into three groups based on their position in the prefrontal regions. Channels 1, 4, 7, 10, 13, and 15 were included in the orbitofrontal cortex (OFC) group. Channels 2, 5, 8, 11, and 14 were sorted into the frontopolar prefrontal cortex group. The remaining channels (channels 3, 6, 9, and 12) were included in the dorsal prefrontal cortex group. Among the three region groups, we found that the frequency of channels belonging to the orbitofrontal cortex group was the largest. Based on these results, the signals collected from the orbitofrontal cortex groups (channels 1, 4, 7, 10, and 13) were found to be relatively more suitable for classifying CSAT levels than other channels in terms of handcrafted features. Spinella and Miley [79] examined the relationships between educational attainment and OFC regions. They found that reinforcing goal-directed behaviors and impulse control from the education process can influence the OFC regions.

TABLE 6. Classification performance of the XGBoost classifier with unsupervised features (three class labels).

Condition (signal)	Channel	Precision	Recall	F1-score	Accuracy	AUC	Condition (signal)	Channel	Precision	Recall	F1-score	Accuracy	AUC
HbO signal	Ch1	0.866	0.871	0.869	0.895	0.930	HbR signal	Ch1	0.891	0.821	0.875	0.896	0.930
	Ch2	0.864	0.870	0.865	0.932	0.930		Ch2	0.866	0.807	0.879	0.892	0.940
	Ch3	0.697	0.759	0.880	0.898	0.940		Ch3	0.881	0.870	0.866	0.896	0.930
	Ch4	0.695	0.763	0.778	0.891	0.930		Ch4	0.893	0.871	0.877	0.897	0.920
	Ch5	0.713	0.737	0.794	0.896	0.930		Ch5	0.887	0.876	0.876	0.886	0.940
	Ch6	0.867	0.870	0.871	0.927	0.930		Ch6	0.879	0.876	0.866	0.892	0.940
	Ch7	0.894	0.871	0.886	0.863	0.940		Ch7	0.889	0.876	0.898	0.897	0.940
	Ch8	0.876	0.877	0.892	0.897	0.930		Ch8	0.865	0.872	0.861	0.881	0.940
	Ch9	0.885	0.870	0.877	0.884	0.930		Ch9	0.845	0.870	0.886	0.897	0.930
	Ch10	0.853	0.879	0.886	0.892	0.930		Ch10	0.870	0.865	0.866	0.885	0.920
	Ch11	0.893	**0.893**	0.886	0.879	0.930		Ch11	0.891	0.870	0.886	0.889	0.940
	Ch12	0.854	0.870	**0.894**	0.885	0.930		Ch12	0.861	0.890	0.865	0.889	0.930
	Ch13	0.845	0.869	0.858	0.895	0.930		Ch13	**0.899**	0.876	0.872	0.890	0.920
	Ch14	0.870	0.886	0.890	0.886	0.940		Ch14	0.890	0.829	0.875	0.889	0.930
	Ch15	**0.898**	0.832	0.869	**0.899**	0.930		Ch15	0.871	**0.893**	**0.891**	**0.898**	0.940
Mean		0.838	0.847	0.866	0.895	0.932	Mean		0.879	0.864	0.876	0.892	0.933

AUC: Area under the curve

TABLE 7. Classification performance of the XGBoost classifier with unsupervised features (four class labels).

Condition (signal)	Channel	Precision	Recall	F1-score	Accuracy	AUC	Condition (signal)	Channel	Precision	Recall	F1-score	Accuracy	AUC
HbO signal	Ch1	**0.892**	0.863	0.851	0.864	0.940	HbR signal	Ch1	0.838	0.867	0.886	0.880	0.950
	Ch2	0.865	0.890	0.740	0.879	0.950		Ch2	0.855	0.867	0.855	0.868	0.940
	Ch3	0.861	0.899	0.859	0.869	0.950		Ch3	0.870	0.882	0.871	0.882	0.950
	Ch4	0.879	0.888	0.877	0.888	0.950		Ch4	0.850	0.889	0.861	0.863	0.940
	Ch5	0.860	0.884	0.867	0.884	0.950		Ch5	0.866	0.881	0.869	0.881	0.950
	Ch6	0.860	0.871	0.853	0.871	0.940		Ch6	0.851	0.866	0.855	0.866	0.940
	Ch7	0.880	0.898	0.884	0.890	0.950		Ch7	0.871	0.873	0.862	0.873	0.950
	Ch8	0.890	**0.900**	**0.890**	**0.900**	0.940		Ch8	**0.889**	0.864	0.851	0.864	0.950
	Ch9	0.870	0.889	0.877	0.889	0.940		Ch9	0.844	0.861	0.847	0.871	0.940
	Ch10	0.843	0.899	0.878	0.901	0.950		Ch10	0.779	0.815	0.872	0.890	0.940
	Ch11	0.851	0.880	0.854	0.864	0.940		Ch11	0.861	0.874	0.862	0.874	0.940
	Ch12	0.843	0.878	0.727	0.898	0.940		Ch12	0.860	0.880	0.868	0.880	0.940
	Ch13	0.880	0.890	0.887	0.891	0.940		Ch13	0.888	**0.899**	**0.889**	**0.901**	0.950
	Ch14	0.831	0.817	0.829	0.885	0.940		Ch14	0.860	0.882	0.867	0.882	0.940
	Ch15	0.867	0.881	0.869	0.881	0.950		Ch15	0.879	0.891	0.880	0.891	0.940
Mean		0.865	0.882	0.849	0.884	0.945	Mean		0.857	0.873	0.866	0.878	0.944

AUC: Area under the curve

TABLE 8. Classification performance of the XGBoost classifier with unsupervised features (five class labels).

Condition (signal)	Channel	Precision	Recall	F1-score	Accuracy	AUC	Condition (signal)	Channel	Precision	Recall	F1-score	Accuracy	AUC
HbO signal	Ch1	0.651	0.831	0.801	**0.877**	0.950	HbR signal	Ch1	0.877	0.846	0.808	0.846	0.950
	Ch2	0.880	0.831	0.803	0.851	0.950		Ch2	0.875	0.837	0.803	0.839	0.950
	Ch3	0.882	0.851	0.851	0.853	0.950		Ch3	0.873	0.829	0.896	0.829	0.950
	Ch4	0.871	0.848	0.810	0.848	0.950		Ch4	0.851	0.831	0.803	0.831	0.950
	Ch5	0.879	0.852	0.812	0.852	0.950		Ch5	0.861	0.839	**0.897**	**0.893**	0.940
	Ch6	0.814	0.830	0.806	0.839	0.950		Ch6	0.874	0.842	0.805	0.842	0.950
	Ch7	0.872	0.841	0.804	0.841	0.950		Ch7	0.890	0.843	0.813	0.842	0.950
	Ch8	0.804	0.845	0.820	0.845	0.950		Ch8	0.811	0.862	0.828	0.862	0.950
	Ch9	0.808	**0.865**	0.833	0.865	0.960		Ch9	0.861	0.827	0.893	0.827	0.950
	Ch10	0.876	0.847	**0.862**	0.847	0.960		Ch10	0.883	0.843	0.811	0.846	0.950
	Ch11	0.871	0.841	0.801	0.841	0.960		Ch11	0.872	0.846	0.804	0.846	0.950
	Ch12	0.817	0.835	0.813	0.838	0.950		Ch12	0.898	0.854	0.819	0.845	0.950
	Ch13	0.811	0.865	0.833	0.865	0.950		Ch13	0.884	**0.863**	0.821	0.862	0.960
	Ch14	0.871	0.837	0.802	0.837	0.950		Ch14	**0.893**	0.813	0.891	0.813	0.950
	Ch15	**0.891**	0.848	0.813	0.848	0.950		Ch15	0.883	0.846	0.812	0.846	0.950
Mean		0.840	0.844	0.818	0.850	0.952	Mean		0.872	0.841	0.834	0.845	0.950

AUC: Area under the curve

To compare the influence of each feature on classification performance, we applied unsupervised features from a feature extractor trained by EEG signals. With regard to the classification performances of the XGBoost classifier, the five evaluation metric values of the unsupervised feature condition were higher than those of the handcrafted features. We confirmed that unsupervised features extracted using deep learning algorithms were more appropriate for classification than the eight handcrafted features. Based on the aforementioned results, we verified that a deep learning algorithm can work well as a feature extractor without fine-tuning when the transfer learning approach is applied between similar domain datasets.

In addition, we sorted the experimental results for each channel in ascending order based on the classification performance to compare the appropriateness of signals collected from channels for classification. After sorting the results, we selected the channel with the highest number of cases for each metric. The frequency of channels in the OFC group was higher than that in the other groups. In the case of channel importance for classification, we found a similar trend with handcrafted features (i.e., signals of the OFC group were relatively appropriate for our research topic).

In summary, we compared the classification performances of five machine learning classifiers between handcrafted and unsupervised features to verify the usability of the features for CSAT classification. As a result, the XGBoost classifier was found to be most suitable for classification. We concluded that unsupervised features were more usable for classifying CSAT levels based on the experimental results. In addition, the applicability of the transfer learning approach without fine-tuning was verified in deep learning models between the same domain dataset (EEG and fNIRS). Furthermore, fNIRS signals measured from the OFC groups were more adequate than those measured from the other groups for our research topics.

V. CONCLUSION

In this work, we proposed a machine learning-based framework for classifying CSAT levels using unsupervised features extracted from deep learning algorithms. Based on previous studies on the relationship between learning ability and neural activities, hemodynamics in fNIRS signals using the NIRSIT Lite device were measured to extract handcrafted and unsupervised features. To evaluate our framework from various perspectives, we designed experiments using various class labels and feature extraction conditions. We found that the XGBoost classifier exhibited the best classification performance and that unsupervised features extracted by the feature extractor trained with EEG signals were suitable for classifying the CSAT levels.

The first strength of this study was the application of fNIRS signals, which are not widely used to classify CSAT levels. Second, we determined the ideal conditions for fNIRS signals for feature extraction. Third, in terms of transfer learning, we checked the usability of one-dimensional convolutional autoencoder algorithms as feature extractors with different neural modals (i.e., EEG and fNIRS). Fourth, an fNIRS dataset collected from undergraduate students in three different universities with eight cognitive task sessions was used to reflect variations in CSAT levels and neural activities.

Our study has some limitations. First, fNIRS signals can include detailed differences between diverse cognitive tasks. These differences can affect the learning ability evaluation results and CSAT levels. However, we considered overall characteristics instead of specific changes to classify the CSAT levels. Second, deep learning algorithms can be used to detect latent patterns in fNIRS signals for CSAT level classification. An additional fNIRS dataset needs to be collected for applying deep learning models in further studies. Finally, to generalize our framework, we need to consider external validation through fNIRS signals collected from other participant groups (e.g., other countries or societies) in further studies.

APPENDIX A

Experimental Results with handcraft features from three classification algorithms (XGBoost, logistic regression, and support vector classifier).

https://figshare.com/articles/dataset/Experiment_results_with_unsupervised_features/19100429. Choi, Junggu; Ko, Inhwan; Nah, Yoonjin; Kim, Bora; Park, Yongwan; Cha, Jihyun; Han, Sanghoon (2022): Experimental results with handcraft features. figshare. Dataset. https://doi.org/10.6084/m9.figshare.19100429. (XLSX)

APPENDIX B

Experimental Results with unsupervised features from three classification algorithms (XGBoost, logistic regression, and support vector classifier).

https://figshare.com/articles/dataset/Experimental_result_with_unsupervised_features/19100618. Choi, Junggu; Ko, Inhwan; Nah, Yoonjin; Kim, Bora; Park, Yongwan; Cha, Jihyun; Han, Sanghoon (2022): Experimental result with unsupervised features. figshare. Dataset. https://doi.org /10.6084/ m9.figshare.19100618. (XLSX)

REFERENCES

[1] M. Artigue, "Mathematics education research at university level: 3Achievements and challenges," in *Research and development in University Mathematics Education*. Routledge Park Square, Milton Park, Abingdon, Oxon Ox14 4RN: Viviane Durand-Guerrier, Reinhard Hochmuth, Elena Nardi, and Carl Winslow, 2021, pp. 2–21.

[2] R. S. Brown and D. T. Conley, "Comparing state high school assessments to standards for success in entry-level university courses," *Educ. Assessment*, vol. 12, no. 2, pp. 137–160, Apr. 2007.

[3] G. S. Aikenhead, "The measurement of high school students' knowledge about science and scientists," *Sci. Educ.*, vol. 57, no. 4, pp. 539–549, Oct. 1973.

[4] T. Kellaghan and V. Greaney, "Using assessment to improve the quality of education," in *Unesco, International Institute for Educational Planning*. United Nations Educational, Scientific and Cultural Organization 7 place de Fontenoy, Paris, France: Thomas Kellaghan and Vincent Greaney, 2001, p. 98.

[5] D. Pereira, M. A. Flores, and L. Niklasson, "Assessment revisited: A review of research inAssessment and evaluation in higher education," *Assessment Eval. Higher Educ.*, vol. 41, no. 7, pp. 1008–1032, Oct. 2016.

[6] J. Prigoff, M. Hunter, and R. Nowygrod, "Medical student assessment in the time of COVID-19," *J. Surgical Educ.*, vol. 78, no. 2, pp. 370–374, Mar. 2021.

[7] M. R. Barulli, M. Piccininni, A. Brugnolo, C. Musarò, C. D. Dio, R. Capozzo, R. Tortelli, U. Lucca, and G. Logroscino, "The Italian version of the test your memory (TYM-I): A tool to detect mild cognitive impairment in the clinical setting," *Frontiers Psychol.*, vol. 11, pp. 1–8, Jan. 2021.

[8] J. Dermo, "e-Assessment and the student learning experience: A survey of student perceptions of e-assessment," *Brit. J. Educ. Technol.*, vol. 40, no. 2, pp. 203–214, 2009.

[9] K. N. Le and V. W. Y. Tam, "A survey on effective assessment methods to enhance student learning," *Australas. J. Eng. Educ.*, vol. 13, no. 2, pp. 13–20, Jan. 2007.

[10] P. C. Burnett, "Assessing the structure of learning outcomes from counselling using the SOLO taxonomy: An exploratory study," *Brit. J. Guid. Counselling*, vol. 27, no. 4, pp. 567–580, Nov. 1999.

[11] N. J. Jenkins, "The scholastic aptitude test as a predictor of academic success: A literature review," Educational Resources Information Center (ERIC), Institute of Education Sciences, 550 12th Street, SW, Washington, DC, USA, Tech. Rep. ED354243, 1992.

[12] A. M. Gallagher and R. De Lisi, "Gender differences in scholastic aptitude test: Mathematics problem solving among high-ability students," *J. Educ. Psychol.*, vol. 86, no. 2, p. 204, 1994.

[13] L. J. Stricker, D. A. Rock, and N. W. Burton, "Sex differences in predictions of college grades from scholastic aptitude test scores," *J. Educ. Psychol.*, vol. 85, no. 4, p. 710, 1993.

[14] C. Cliffordson, "Effects of practice and intellectual growth on performance on the Swedish scholastic aptitude test (SweSAT)," *Eur. J. Psychol. Assessment*, vol. 20, no. 3, pp. 192–204, Jan. 2004.

[15] S. Denervaud, E. Fornari, X.-F. Yang, P. Hagmann, M. H. Immordino-Yang, and D. Sander, "An fMRI study of error monitoring in montessori and traditionally-schooled children," *NPJ Sci. Learn.*, vol. 5, no. 1, pp. 1–10, Dec. 2020.

[16] H. Kim, T. Hong, J. Kim, and S. Yeom, "A psychophysiological effect of indoor thermal condition on college students' learning performance through EEG measurement," *Building Environ.*, vol. 184, Oct. 2020, Art. no. 107223.

[17] S. Firooz and S. K. Setarehdan, "IQ estimation by means of EEG-fNIRS recordings during a logical-mathematical intelligence test," *Comput. Biol. Med.*, vol. 110, pp. 218–226, Jul. 2019.

[18] S. J. Howard, H. Burianová, J. Ehrich, L. Kervin, A. Calleia, E. Barkus, J. Carmody, and S. Humphry, "Behavioral and fMRI evidence of the differing cognitive load of domain-specific assessments," *Neuroscience*, vol. 297, pp. 38–46, Jun. 2015.

[19] I. Daly, J. Bourgaize, and A. Vernitski, "Mathematical mindsets increase student motivation: Evidence from the EEG," *Trends Neurosci. Educ.*, vol. 15, pp. 18–28, Jun. 2019.

[20] L. Sugiura, M. Hata, H. Matsuba-Kurita, M. Uga, D. Tsuzuki, I. Dan, H. Hagiwara, and F. Homae, "Explicit performance in girls and implicit processing in boys: A simultaneous fNIRS–ERP study on second language syntactic learning in young adolescents," *Frontiers Hum. Neurosci.*, vol. 12, pp. 1–19, Mar. 2018.

[21] Z. Mao, Y. Su, G. Xu, X. Wang, Y. Huang, W. Yue, L. Sun, and N. Xiong, "Spatio-temporal deep learning method for ADHD fMRI classification," *Inf. Sci.*, vol. 499, pp. 1–11, Oct. 2019.

[22] H. B. Evgin, O. Babacan, I. Ulusoy, Y. Hosgoren, A. Kusman, D. Sayar, B. Baskak, and H. D. Ozguven, "Classification of fNIRS data using deep learning for bipolar disorder detection," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.

[23] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network," *Neural Netw.*, vol. 136, pp. 1–10, Apr. 2021.

[24] L. R. Trambaiolli, R. Cassani, and T. H. Falk, "EEG spectro-temporal amplitude modulation as a measurement of cortical hemodynamics: An EEG-fNIRS study," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 3481–3484.

[25] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1306–1315.

[26] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and O. Ren, "Cross validation framework to choose amongst models and datasets for transfer learning," in *Proc. Joint Eur. Conf. Machine Learn. Knowl. Discovery Databases*, Berlin, Germany: Springer, Sep. 2010, pp. 547–562.

[27] S. Bak, J. Jeong, and J. Shin, "Mutual interaction between genders with stress or non-stress by positive stimulus characteristics using fNIRS," in *Proc. 9th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2021, pp. 1–5.

[28] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539–550, 2007.

[29] C.-M. Lu, Y.-J. Zhang, B. B. Biswal, Y.-F. Zang, D.-L. Peng, and C.-Z. Zhu, "Use of fNIRS to assess resting state functional connectivity," *J. Neurosci. Methods*, vol. 186, no. 2, pp. 242–249, Feb. 2010.

[30] S. Lancia, V. Cofini, M. Carrieri, M. Ferrari, and V. Quaresima, "Are ventrolateral and dorsolateral prefrontal cortices involved in the computerized corsi block-tapping test execution? An fNIRS study," *Neurophotonics*, vol. 5, no. 1, p. 1, Jan. 2018.

[31] A. Manelis, T. J. Huppert, E. Rodgers, H. A. Swartz, and M. L. Phillips, "The role of the right prefrontal cortex in recognition of facial emotional expressions in depressed individuals: FNIRS study," *J. Affect. Disorders*, vol. 258, pp. 151–158, Nov. 2019.

[32] J. D. Schaeffer, A. S. Yennu, K. C. Gandy, F. Tian, H. Liu, and H. Park, "An fNIRS investigation of associative recognition in the prefrontal cortex with a rapid event-related design," *J. Neurosci. Methods*, vol. 235, pp. 308–315, Sep. 2014.

[33] M. Laguë-Beauvais, J. Brunet, L. Gagnon, F. Lesage, and L. Bherer, "A fNIRS investigation of switching and inhibition during the modified stroop task in younger and older adults," *NeuroImage*, vol. 64, pp. 485–495, Jan. 2013.

[34] A. C. Ruocco, A. H. Rodrigo, J. Lam, S. I. Di Domenico, B. Graves, and H. Ayaz, "A problem-solving task specialized for functional neuroimaging: Validation of the scarborough adaptation of the tower of London (S-TOL) using near-infrared spectroscopy," *Frontiers Hum. Neurosci.*, vol. 8, p. 185, Mar. 2014.

[35] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task—Quantified in the prefrontal cortex using fNIRS," *Frontiers Human Neurosci.*, vol. 7, p. 935, Jan. 2014.

[36] M. K. Yeung and J. Lin, "Probing depression, schizophrenia, and other psychiatric disorders using fNIRS and the verbal fluency test: A systematic review and meta-analysis," *J. Psychiatric Res.*, vol. 140, pp. 416–435, Aug. 2021.

[37] D. T. Delpy, M. Cope, P. V. D. Zee, S. Arridge, S. Wray, and J. Wyatt, "Estimation of optical pathlength through tissue from direct time of flight measurement," *Phys. Med. Biol.*, vol. 33, no. 12, pp. 1433–1442, Dec. 1988.

[38] M. Cope, *The Development of a Near Infrared Spectroscopy System and its Application for Non Invasive Monitoring of Cerebral Blood and Tissue Oxygenation in the Newborn Infants*. London, U.K.: Univ. London, 1991.

[39] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "Hierarchical internal representation of spectral features in deep convolutional networks trained for EEG decoding," in *Proc. 6th Int. Conf. Brain-Comput. Interface (BCI)*, Jan. 2018, pp. 1–6.

[40] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *J. Neural Eng.*, vol. 16, no. 5, pp. 1–37, 2019.

[41] J.-H. Jang, J. Choi, H. W. Roh, S. J. Son, C. H. Hong, E. Y. Kim, T. Y. Kim, and D. Yoon, "Deep learning approach for imputation of missing values in actigraphy data: Algorithm development study," *JMIR mHealth uHealth*, vol. 8, no. 7, Jul. 2020, Art. no. e16113.

[42] K.-S. Hong, M. J. Khan, and M. J. Hong, "Feature extraction and classification methods for hybrid fNIRS-EEG brain-computer interfaces," *Frontiers Hum. Neurosci.*, vol. 12, p. 246, Jun. 2018.

[43] N. Naseer and K.-S. Hong, "FNIRS-based brain-computer interfaces: A review," *Frontiers Human Neurosci.*, vol. 9, p. 3, Jan. 2015.

[44] A. P. Buccino, H. O. Keles, and A. Omurtag, "Hybrid EEG-fNIRS asynchronous brain-computer interface for multiple motor tasks," *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0146610.

[45] M. Stangl, G. Bauernfeind, J. Kurzmann, R. Scherer, and C. Neuper, "A haemodynamic brain–computer interface based on real-time classification of near infrared spectroscopy signals during motor imagery and mental arithmetic," *J. Near Infr. Spectrosc.*, vol. 21, no. 3, pp. 157–171, Jun. 2013.

[46] J. Shin and J. Jeong, "Multiclass classification of hemodynamic responses for performance improvement of functional near-infrared spectroscopy-based brain–computer interface," *J. Biomed. Opt.*, vol. 19, no. 6, Jun. 2014, Art. no. 067009.

[47] M. J. Khan and K.-S. Hong, "Hybrid EEG–fNIRS-based eight-command decoding for BCI: Application to quadcopter control," *Frontiers Neurorobot.*, vol. 11, p. 6, Feb. 2017.

[48] R. Li, T. Potter, W. Huang, and Y. Zhang, "Enhancing performance of a hybrid EEG-fNIRS system using channel selection and early temporal features," *Frontiers Human Neurosci.*, vol. 11, p. 462, Sep. 2017.

[49] A. Zafar and K. S. Hong, "Detection and classification of three class initial dips from prefrontal cortex," *Biomed. Opt. Exp.*, vol. 8, pp. 367–383, Jan. 2017.

[50] K.-S. Hong and H. Santosa, "Decoding four different sound-categories in the auditory cortex using functional near-infrared spectroscopy," *Hearing Res.*, vol. 333, pp. 157–166, Mar. 2016.

[51] H.-J. Hwang, H. Choi, J.-Y. Kim, W.-D. Chang, D.-W. Kim, K. Kim, S. Jo, and C.-H. Im, "Toward more intuitive brain–computer interfacing: Classification of binary covert intentions using functional near-infrared spectroscopy," *J. Biomed. Opt.*, vol. 21, no. 9, Apr. 2016, Art. no. 091303.

[52] L. Holper and M. Wolf, "Single-trial classification of motor imagery differing in task complexity: A functional near-infrared spectroscopy study," *J. NeuroEng. Rehabil.*, vol. 8, no. 1, p. 34, 2011.

[53] B. Abibullaev, J. An, and J.-I. Moon, "Neural network classification of brain hemodynamic responses from four mental tasks," *Int. J. Optomechtron.*, vol. 5, no. 4, pp. 340–359, 2011.

[54] D. Lavanya and K. U. Rani, "Performance evaluation of decision tree classifiers on medical datasets," *Int. J. Comput. Appl.*, vol. 26, no. 4, pp. 1–4, 2011.

[55] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[56] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, 2001, pp. 41–46.

[57] Q. Wu and D. X. Zhou, "Analysis of support vector machine classification," *J. Comput. Anal. Appl.*, vol. 8, no. 2, pp. 1–21, 2006.

[58] H. Bhavsar and M. H. Panchal, "A review on support vector machine for data classification," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 10, pp. 185–189, 2012.

[59] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[60] B. Kaewkamnerdpong, "A framework for human learning ability study using simultaneous EEG/fNIRS and portable EEG for learning and teaching development," in *Smart Education and e-Learning*. Cham, Switzerland: Springer, 2016, pp. 155–165.

[61] C. Artemenko, M. Soltanlou, T. Dresler, A.-C. Ehlis, and H.-C. Nuerk, "The neural correlates of arithmetic difficulty depend on mathematical ability: Evidence from combined fNIRS and ERP," *Brain Struct. Function*, vol. 223, no. 6, pp. 2561–2574, Jul. 2018.

[62] M. Soltanlou, M. A. Sitnikova, H.-C. Nuerk, and T. Dresler, "Applications of functional near-infrared spectroscopy (fNIRS) in studying cognitive development: The case of mathematics and language," *Frontiers Psychol.*, vol. 9, p. 277, Apr. 2018.

[63] J. Benerradi, H. A. Maior, A. Marinescu, J. Clos, and M. L. Wilson, "Exploring machine learning approaches for classifying mental workload using fNIRS data from HCI tasks," in *Proc. Halfway Future Symp.*, 2019, pp. 1–11, 2019.

[64] R. Hosseini, B. Walsh, F. Tian, and S. Wang, "An fNIRS-based feature learning and classification framework to distinguish hemodynamic patterns in children who stutter," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 6, pp. 1254–1263, Jun. 2018.

[65] R. F. Rojas, X. Huang, J. Romero, and K. L. Ou, "FNIRS approach to pain assessment for non-verbal patients," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 778–787.

[66] M. J. Khan and K.-S. Hong, "Passive BCI based on drowsiness detection: An fNIRS study," *Biomed. Opt. Exp.*, vol. 6, no. 10, pp. 4063–4078, Oct. 2015.

[67] S.-H. Yoo, H. Santosa, C.-S. Kim, and K.-S. Hong, "Decoding multiple sound-categories in the auditory cortex by neural networks: An fNIRS study," *Frontiers Hum. Neurosci.*, vol. 15, p. 211, Apr. 2021.

[68] 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 Hum. Neurosci.*, vol. 13, p. 317, Sep. 2019.

[69] H. Aghajani, M. Garbey, and A. Omurtag, "Measuring mental workload with EEG+fNIRS," *Frontiers Hum. Neurosci.*, vol. 11, p. 359, Jul. 2017.

[70] F. Putze, S. Hesslinger, C.-Y. Tse, Y. Huang, C. Herff, C. Guan, and T. Schultz, "Hybrid fNIRS-EEG based classification of auditory and visual perception processes," *Frontiers Neurosci.*, vol. 8, p. 373, Nov. 2014.

[71] D. Heger, C. Herff, and T. Schultz, "Combining feature extraction and classification for fNIRS BCIs by regularized least squares optimization," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 2012–2015.

[72] J.-H. Jang, T. Y. Kim, H.-S. Lim, and D. Yoon, "Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder," *PLoS ONE*, vol. 16, no. 12, Dec. 2021, Art. no. e0260612.

[73] J.-H. Jang, T. Y. Kim, and D. Yoon, "Effectiveness of transfer learning for deep learning-based electrocardiogram analysis," *Healthcare Informat. Res.*, vol. 27, no. 1, pp. 19–28, Jan. 2021.

[74] S. Raghu, N. Sriraam, Y. Temel, S. V. Rao, and P. L. Kubben, "EEG based multi-class seizure type classification using convolutional neural network and transfer learning," *Neural Netw.*, vol. 124, pp. 202–212, Apr. 2020.

[75] Y. Zhu, J. K. Jayagopal, R. K. Mehta, M. Erraguntla, J. Nuamah, A. D. McDonald, H. Taylor, and S.-H. Chang, "Classifying major depressive disorder using fNIRS during motor rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 961–969, Apr. 2020.

[76] H. Khan, F. M. Noori, A. Yazidi, M. Z. Uddin, M. N. A. Khan, and P. Mirtaheri, "Classification of individual finger movements from right hand using fNIRS signals," *Sensors*, vol. 21, no. 23, p. 7943, Nov. 2021.

[77] F. M. Noori, N. Naseer, N. K. Qureshi, H. Nazeer, and R. A. Khan, "Optimal feature selection from fNIRS signals using genetic algorithms for BCI," *Neurosci. Lett.*, vol. 647, pp. 61–66, Apr. 2017.

[78] M. S. B. A. Ghaffar, U. S. Khan, J. Iqbal, N. Rashid, A. Hamza, W. S. Qureshi, M. I. Tiwana, and U. Izhar, "Improving classification performance of four class FNIRS-BCI using mel frequency cepstral coefficients (MFCC)," *Infr. Phys. Technol.*, vol. 112, Jan. 2021, Art. no. 103589.

[79] M. Spinella and W. M. Miley, "Orbitofrontal function and educational attainment," *College Student J.*, vol. 38, no. 3, pp. 333–339, 2004.

JUNGGU CHOI was born in 1993. He is currently pursuing the Ph.D. degree in cognitive science with the Applied Brain Cognitive Laboratory, Yonsei University Graduate Program. His research interests include brain science and data mining.

INHWAN KO received the M.S. degree in psychology from Yonsei University. He has diverse working experience with the Human Resource Management Practitioner in multinational companies to a certified public labor attorney. His research interests include converging human resource management and cognitive neuroscience.

YOONJIN NAH is currently a Post-Master's Researcher with the Department of Psychology, Yonsei University. He has various experiences in decoding of differential cognitive states or clinical groups with machine learning algorithms using fMRI connectivity data.

BORA KIM received the Ph.D. degree in psychology from Washington State University, in 2016. She is currently an Assistant Professor at Honam University. Her research interests include human judgment and decision-making behavior in technological environments and recently expands her research to the social neurocognition field.

YONGWAN PARK received the Ph.D. degree in marketing from Virginia Tech. He is currently an Assistant Professor of marketing with the College of Business, Gyeongsang National University. His research interests include consumer judgment and decision based on behavioral decision theory, and consumer perception about IT products.

JIHYUN CHA received the Ph.D. degree in cognitive and brain sciences from Washington University, St.Louis. She is a Researcher at OBELAB, Inc. Her research interests include investigating cognitive and neural biomarkers of individual differences and clinical symptoms through academic, clinical, and commercial applications of fNIRS.

JONGKWAN CHOI received the Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST). His research interests include portable functional near-infrared spectroscopy (fNIRS), biomedical integrated circuits, and optical communication system. He received the IEEE International Symposium on Circuits and Systems Second Best Paper Award of the Biomedical and Life Science Circuits, in 2012. Since 2016, he has been with OBELAB, Inc., Seoul, South Korea, a bio start-up that manufactures portable functional brain imaging systems and working on developing a system architecture.

SANGHOON HAN was born in 1977. He is currently a Professor with the Department of Psychology, Yonsei University. His research interests include decision making and cognitive science.