IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 29, 2019, accepted January 12, 2020, date of publication January 15, 2020, date of current version January 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2966834

Multilevel Feature Fusion With 3D Convolutional Neural Network for EEG-Based Workload Estimation

YOUNGCHUL KWAK[1], KYEONGBO KONG[1], WOO-JIN SONG[1], (Member, IEEE), BYOUNG-KYONG MIN[2], (Member, IEEE), AND SEONG-EUN KIM[3], (Member, IEEE)

[1]Department of Electronics Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, South Korea
[2]Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea
[3]Department of Electronics and Control Engineering, Hanbat National University, Daejeon 34158, South Korea

Corresponding author: Seong-Eun Kim (sekim@hanbat.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by Korea Government (MSIT), under Grant 2017R1C1B5017254, Grant 2018M3C1B8017549, and Grant 2018R1A2B6004084, and in part by the Institute for Information & Communications Technology Promotion (IITP) through the Korea Government under Grant 2017-0-00451, Grant 2017-0-00432, and Grant 2019-2016-0-00464.

ABSTRACT Mental workload is defined as the proportion of the information processing capability used to perform a task. High cognitive load requires additional resources to process information; this demand for additional resources may reduce the processing efficiency and performance. Therefore, the technique of workload estimation can ensure a proper working environment to promote the working efficiency of each person. In this paper, we propose a three-dimensional convolutional neural network (3D CNN) employing a multilevel feature fusion algorithm for mental workload estimation using electroencephalogram (EEG) signals. The 1D EEG signals are converted to 3D EEG images to enable the 3D CNN to learn the spectral and spatial information over the scalp. The multilevel feature fusion framework integrates local and global neuronal activities by workload tasks in the 3D CNN algorithm. Multilevel features are extracted in each layer of the 3D convolution operation and each multilevel feature is multiplied by a weighting factor, which determines the importance of the feature. The weighting factor is adaptively estimated for each EEG image by a backpropagation process. Furthermore, we generate subframes from each EEG image and propose a temporal attention technique based on the long short-term memory model (LSTM) to extract a significant subframe at each multilevel feature that is strongly correlated with task difficulty. To verify the performance of our network, we performed the Sternberg task to measure the mental workload of the participant, which was classified according to its difficulty as low or high workload condition. We showed that the difficulty of the workload was well designed, which was reflected in the behavior of the participant. Our network is trained on this dataset and the accuracy of our network is 90.8 %, which is better than that of conventional algorithms. We also evaluated our method using the public EEG dataset and achieved 93.9 % accuracy.

INDEX TERMS Convolutional neural network, electroencephalogram (EEG), feature fusion, mental workload, working memory.

I. INTRODUCTION

Mental workload is defined as the proportion of information processing capability that is utilized when a person is performing a task [1]. Studies on mental workload have aimed to develop methods that efficiently use the limited information processing capacity of humans [2], [3]. For instance, when a person acquires new skills, low workload can cause wastage of mental resources. Therefore, providing higher difficulty of learning can enhance the speed of learning. In contrast, high work intensity generates mental overload and reduces the efficiency and performance of the task. In many cases, high workload can cause serious accidents owing to failure of the task or poor decision-making [3]. Monitoring the workload is essential to assist people in enhancing their work performance or learning speed by properly adjusting the difficulty of the workload. It can also improve workers' well-being and safety at work by measuring excessive workload in real-time.

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyu Zhou.

Therefore, workload estimation has been widely studied in various areas such as air vehicle task [4], text reading [5], and multitasking environments [6].

Because cognitive tasks are processed by neuronal activities, brain signals can be effectively used to estimate the workload. Brain signals can be measured by several methods including electroencephalography (EEG), functional near-infrared spectroscopy (fNIR), and functional magnetic resonance imaging (fMRI). Among the methods, EEG has been mostly used in studies on working memory because the EEG signals have a sensitive indicator to distinguish the working memory process. EEG has high temporal resolution, which can be applied to real-time workload estimation. In addition, portable EEG devices allow EEG signals to be easily acquired during a task and can be easily adapted to real-world applications.

Several neuroscience researchers have found that the spatiotemporal dynamics of the EEG power spectrum over the scalp are strongly related to the mental workload [7]–[10]. Some studies have reported that the power of the theta (4-8 Hz) oscillations increases in the frontal lobe as the workload increases [7], [9]; the alpha (8-12 Hz) power is linked to idling [11] and cortical inhibition [12]; and the beta (12-30 Hz) power is increased in proportion to the workload [7]. The findings imply that the structure of the power spectrum over the frequency bands (theta, alpha, and beta) is closely related to different levels of the workload. Based on these findings, early studies extracted the spectral features related to workload estimation and attempted to classify the workload as ‘‘low workload’’ and ‘‘high workload’’ through classical machine learning algorithms such as support vector machine (SVM) or linear discriminant analysis (LDA) [13]–[15]. However, SVM and LDA require handcrafted features representing the spatial and spectral information that greatly influence the classification performance.

In recent years, deep learning has emerged as one of the most powerful techniques for EEG decoding of brain signals in applications such as brain-computer interface, seizure detection, and workload estimation [16]–[23]. Deep learning methods can extract high-level representations from a large dataset and their performance has been verified in various areas including speech recognition, image classification, and video recognition [24]–[27]. In particular, convolutional neural network (CNN) is the most widely used deep learning structure that is analogous to the organization of neurons in the visual cortex. The CNN is capable of capturing the spatial and temporal dependencies in an image through the relevant filters and efficiently extracting high-level features with a small number of parameters. In addition, CNN is easily applicable to any dimension of data (one-dimensional (1D): speech data, two-dimensional (2D): image data, and three-dimensional (3D): video data) by adjusting the dimension of the convolution operation.

Therefore, CNN has been widely utilized in EEG decoding applications because EEG signals are characterized by spectral, spatial, and temporal information. For example, 1D CNN is used to extract the temporal information from EEG signals [20]. On the other hand, some studies have converted 1D EEG signals to 2D EEG images (EEG topographic map) to extract the spatial information from multi-channel EEG data over the scalp [16], [23]. In addition, 3D CNN has been utilized to simultaneously extract spectral and spatial information from spectral topographic maps across all frequencies [18], [21]. Recent studies on EEG decoding applications have shown that 3D CNN structures have superior performance in EEG decoding applications [18], [21], [28].

Most of the CNN algorithms perform classification using only the deep feature extracted from the last convolutional layer. The deep feature certainly contains essential high-level structure information for classification, but a recent study suggests that low-level features that are extracted from the intermediate convolutional layers contain abundant local structure information [29]. Because these local features are useful in improving the deep learning performance, multilevel feature fusion methods have been proposed for various applications [22], [29]–[32]. For example, in [31], multilevel features were aggregated for a music auto-tagging problem to utilize different time-scale features. In remote sensing scene classification, a feature fusion framework is used to aggregate multilayer features to build a more discriminative feature [29]. Further, in object detection, a multilevel feature containing semantics and fine details is utilized to detect a salient object [32]. However, few studies have developed CNN algorithms exploiting multilevel features for EEG decoding applications, despite the well-known fact that the human brain functions through both local and global neuronal connections [33].

In neuroscience, there is a general consensus that the human brain maintains two properties—functional segregation and integration [33], [34]. Functional segregation refers to locally specialized information processing of the brain and integration refers to the global integration of information across the entire brain. To extract spatial information of the spectral power represented by integration and segregation, utilizing both low-level and high-level features is indispensable in the CNN structure: low-level features are useful for extracting the local features of the brain’s segregation, and high-level features are analogous to the integration of the brain’s global activities. Recently, the 3D CNN structure with a multilevel fusion method was proposed for EEG decoding applications [22], but the determination of the importance of each low-level feature for effective integration was not fully investigated.

Moreover, according to the previous study [35], processes of information manipulation and holding information can be separated during the working memory task. Because these two processes have different spectral features, the spectral power changes over time during the transition from information manipulation to holding information. In particular, the spectral feature that is strongly correlated with task difficulty appears in an individually different time when testing ten participants [36]. Therefore, it is important to capture the time interval that has the significant spectral feature related with the task difficulty.

This study proposes a new multilevel feature fusion method for EEG based workload estimation. We recorded the EEG signals from 62 participants during a Sternberg working memory task consisting of an easy and a hard task. Multi-channel EEG data were transformed into a spectral topographical map containing spectral and spatial information. Then, we constructed a 3D CNN based feature fusion network to learn the spatial and spectral representations using low-level and high-level features. Low-level features, which are extracted from low-level convolutional layers, represent the local activities of the brain, and high-level features can extract the global activities of the brain. Therefore, we aggregated both low-level and high-level features and proposed a fusion stream with a weighting factor, which determined the importance of each feature. The weighting factor was optimally adjusted according to the EEG image by the proposed network. In addition, to extract the time interval that has the spectral feature representing the task difficulty, we propose a temporal attention method based on the long short-term memory model (LSTM) [37]. The experimental results show that the proposed multilevel feature fusion model with temporal attention improves the performance of conventional CNN models and outperforms the classical classifiers based on handcrafted features. In summary, the main contributions of our paper are following as:

- We propose a 3D CNN based multilevel feature fusion architecture to extract the low-level and high-level features which represent the brain's segregation and integration, respectively.
- We propose a learning structure for the weighting factor that indicates the importance of each multilevel feature in the proposed multilevel feature fusion architecture. This weighting factor is adaptively adjusted to each 3D EEG image and leads to superior performance compared to the fixed weighting factor.
- We propose a LSTM based temporal attention technique to pick out a highly correlated time segment with the task difficulty.

In this paper, we represent a vector by a lowercase boldface letter and a matrix by an uppercase boldface letter. The remainder of this paper is organized as follows. Section II describes our experimental setup and data collection to study the mental workload in working memory tasks. Section III proposes the multilevel feature fusion method based on the 3D CNN structure. The performance of the proposed method and performance comparisons with conventional algorithms are presented in Section IV. Finally, Section V provides the conclusion.

II. MATERIALS

A. EXPERIMENTAL SETUP

Participants were instructed to perform the Sternberg task, which has been widely used for studying mental workload, especially in working memory [38], [39]. Participants were required to memorize letter-number combined stimuli during an encoding phase, where a stimuli consisted of a combination of an alphabet (A to Z) and a number (0 to 9), such as T3, A1, and G7. The difficulty of the workload was determined by the memory set size. In present study, the low and high workload conditions were composed of three and seven stimuli, respectively.

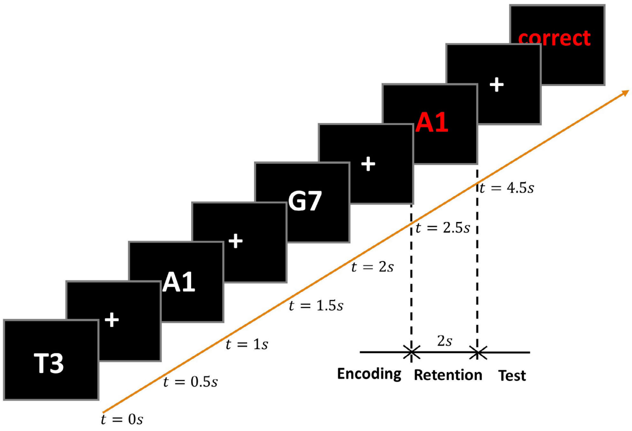

FIGURE 1. Framework of the Sternberg task.

Fig. 1 shows the experimental protocol for the low workload, where three stimuli are displayed sequentially. Each stimulus was presented for 0.5 s followed by an empty black screen with a fixation cross appeared for 0.5 s. Then, the next stimulus was presented and the process was repeated till the last stimulus (for this case, the 3rd stimulus) appeared on the screen. After the last stimulus disappeared, the participants were required to retain the three presented stimuli for 2 s (i.e., the retention period) when participants should hold all the information of presented stimuli for a short term time window. After the retention period, only one test stimulus was presented, and the participants were asked to press the "yes" button if the test stimulus was one of the three previously presented stimulus; otherwise, they press the "no" button. In the last phase of each trial, the participants received feedback on whether their responses were correct.

All participants underwent both low and high workload conditions, and each workload condition was repeated independently for 60 trials. The order of the workload was randomized to avoid the influence of mental fatigue.

B. DATA ACQUISITION AND PARTICIPANTS

Starstim R32 (Neuroelectrcis, Spain) was utilized to record the EEG signals. The EEG signals were recorded at a sampling rate of 500 Hz using 32 electrodes in accordance with the standard 10-10 system. As shown in Fig. 2, thirty electrodes except the Fp1 and Fp2 electrodes were included in the further analysis. All the electrodes were referenced to the electrode located at a right earlobe.

The experiment was approved by the Institutional Review Board (IRB) of Korea University. Eighty-four healthy people voluntarily participated in the Stenberg working memory

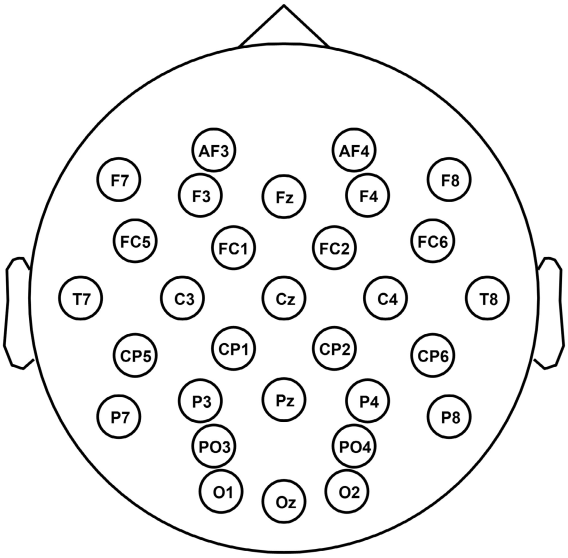

FIGURE 2. Locations of the 30 electrodes.

task and none of them had a history of neurological and psychiatric disorders. All participants were instructed on how to perform the task before the experiment was commenced. In addition, some participants were unable to adapt to the difficulty of the high workload condition, resulting in poor performance. These participants were thus excluded from the further analysis. Taken together, three participants were excluded. Finally, the data on the remaining sixty-two participants (mean age, 23.2 ± 3.0 s.d., 27 females) were used for this study.

We considered only those trials in which the participants pressed the correct button because incorrect trials would not guarantee that they held the stimulus information correctly during the retention period. In addition, trials that showed amplitudes exceeding $-70 - 70\ \mu$ V were excluded from the analysis. As a result, 5,301 samples collected from 62 participants were utilized to train our network.

III. METHOD

A. DATA PREPROCESSING

Preprocessing was performed with MATLAB (MathWorks Inc., Natick, MA). In the working memory task, memory operations are primarily related to oscillations with frequency between 4 Hz and 30 Hz [7]–[10], [40]. Based on prior knowledge, the EEG signals are first band-pass filtered in a frequency band from 0.5 Hz to 40 Hz, which removes the direct current component and high-frequency noises. Then, the filtered signals are down-sampled to 100 Hz.

The EEG signal is the best window to the cortical brain activity. The cerebral cortex is broadly divided into four areas, namely, the frontal lobe, temporal lobe, parietal lobe, and occipital lobe. The lobes of the brain do not function alone. The relationships between the lobes of the brain and between the right and left hemispheres are very complex. Because the memory operations are also closely related with the four lobes in some way, it is very important to analyze the spatial dynamics of the EEG data for understanding the functioning clearly. Therefore, as shown in Fig. 3, we create a topographical map to utilize the spatial characteristics of the EEG spectral power using all 30 electrodes over the scalp. The preprocessing procedures are divided into four steps: a) Laplacian filtering, b) power spectrum estimation, c) 3D to 2D mapping of the EEG electrodes, and d) interpolation.

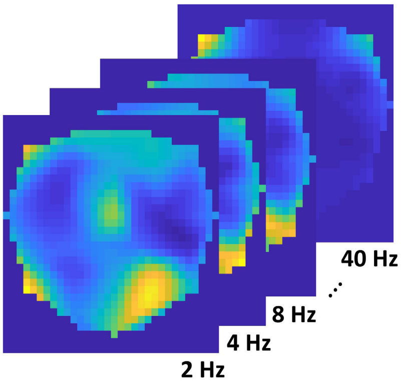

FIGURE 3. Topological maps of EEG spectral powers.

The first step, Laplacian filtering, is a spatial filtering method to enhance localized activity while suppressing the diffused activity [41]. There are multiple sources of signals in the brain, and the sum of the signals generated from multiple sources is measured at a single EEG electrode. Recovering the target source from the superposition of the sources is a typical source localization problem. Laplacian filtering is a simple and effective method to address the problem in applying EEG data [42], [43]. By Laplacian filtering, we subtract the weighted sum of the nearest neighbors of each electrode, as follows [44]:

$$\hat{\boldsymbol{v}}_i(t) = \boldsymbol{v}_i(t) - \sum_{j \in N_i} w_{i,j}\boldsymbol{v}_j(t), \tag{1}$$

where $V_i(t)$ denotes the signals of the i-th EEG electrode, N_i denotes the nearest neighbors of the i-th EEG electrode at time t, and $w_{i,j}$ is a weight determined by the reverse of the distance between i-th and j-th EEG electrodes and is normalized to satisfy $\sum_j w_{i,j} = 1$ for a given i. We manually define the set of the nearest neighbors N_i for each electrode.

The next step is to transform the EEG time series into frequency domain data. The structure of the neural oscillations is strongly linked to cognitive processes, and investigation of frequency-specific changes in the EEG data leads to valuable insights into the working of the brain. We applied Welch's method to 2-s EEG data recorded during the retention period and obtained its power spectral density over the frequency range of 2 Hz to 40 Hz spaced 2 Hz apart. In the third step, the 3D electrode locations over the scalp are projected to a 2D image. Supposing a human head can be approximately modeled as a sphere, we exploit the azimuthal equidistant projection by defining that all points on the electrode are at the proportionally correct distance from the center point

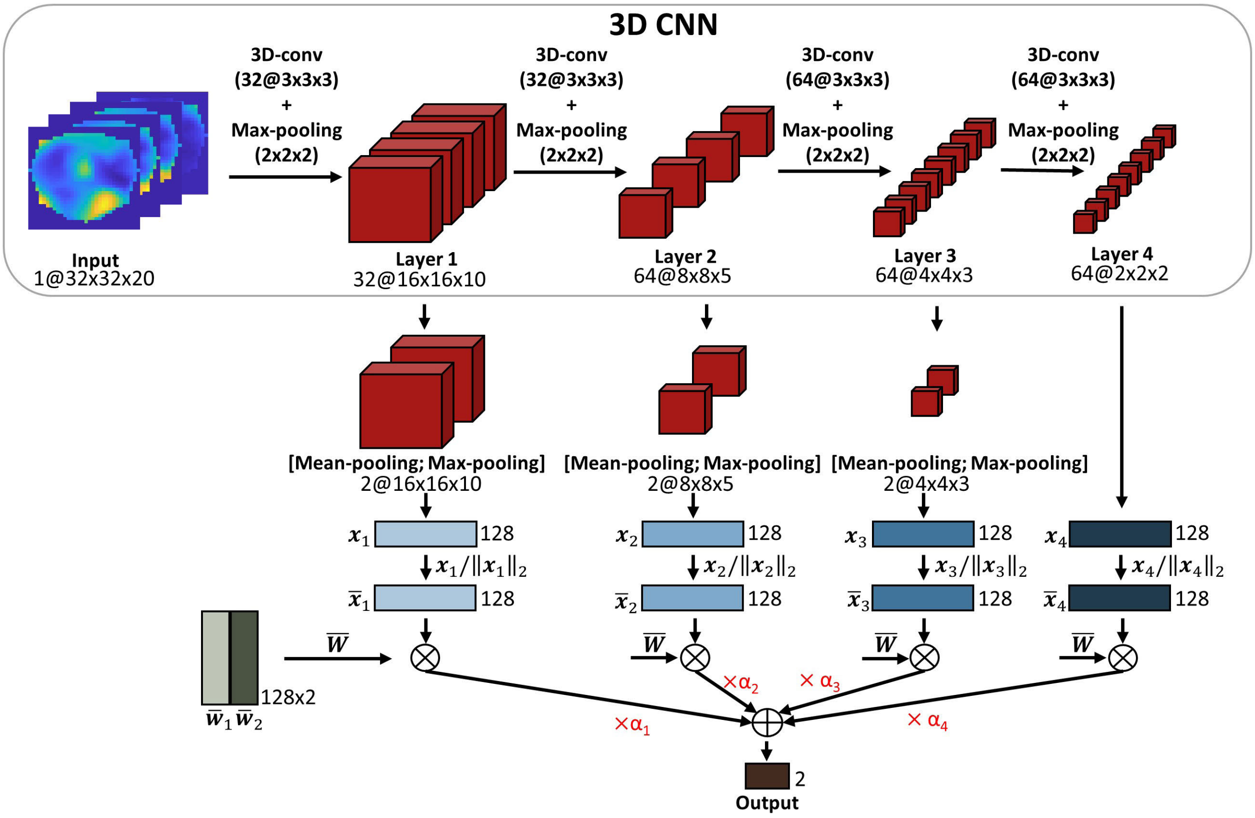

FIGURE 4. Architecture of the proposed algorithm.

(center electrode, Cz). Then, the projected point is mapped to a 32 × 32 mesh and each point is assigned the power spectral density of the corresponding electrode at each frequency [21]. The last step interpolates the empty values between the electrodes using cubic spline interpolation [21]. Finally, the preprocessing transforms the 2-s retention segments of the EEG data from 30-electrodes into a 32 × 32 × 20 EEG image that contains spatial and spectral information.

B. OVERALL ARCHITECTURE OF MULTILEVEL FEATURE FUSION

3D convolution operations are commonly used to extract both spatial and temporal features from video data [25], [45]. Because our EEG data were transformed into 3D image data that contained spatial and spectral features, we propose a 3D CNN based multilevel feature fusion scheme, as shown in Fig. 4. The proposed 3D CNN architecture consists of four convolutional layers and four max-pooling layers based on the 3D kernels to learn spatial and spectral characteristics. The kernel size of all convolutional layers is 3 × 3 × 3 with unit stride and no padding. The max-pooling layers have a 2 × 2 × 2 kernel size with a stride of two. We use the exponential linear unit (ELU) to activate the outputs of all convolutional layers. Like other activation functions such as rectified linear unit (ReLU) and leaky ReLU, the ELU can alleviate the vanishing gradient problem via the identity function for positive values [46]. However, the ELU has negative values to force the mean activated value closer to zero as in batch normalization but with lower computational complexity, although it saturates to a negative value with smaller arguments [46]. The saturation decreases the forward propagation of variations of the deactivated units. Thus, the ELU is more robust against noise; moreover, it facilitates faster learning and has higher classification accuracy than ReLU in EEG decoding applications [19]. The ELU can be expressed as

$$f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \beta(e^x - 1), & \text{otherwise.} \end{cases} \tag{2}$$

where β is a positive hyperparameter to control a saturated value of the negative input x. We set the fixed hyperparameter $\beta = 1$ [46].

In most deep learning studies on EEG data [16]–[19], [47], the high-level feature is adopted only in fully connected layers and used for classification. However, we construct a 3D CNN based multilevel feature fusion method to exploit the intermediate features that contain important information from local and global properties of 3D EEG data [29]. Low-level convolutional features are suitable for expressing the local brain activities in each lobe, and high-level features correspond to global core activities over the entire brain. The new structure combining multilevel features, shown in Fig. 4, integrates low-level and high-level features for classification.

In the proposed structure, the low-level convolutional features (Layers 1, 2, and 3 in Fig. 4) have larger dimensions;

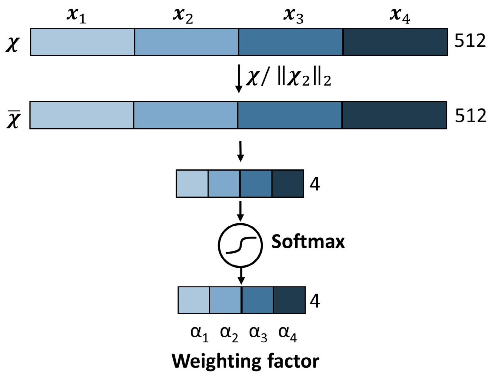

FIGURE 5. Process of extracting the weighting factors.

consequently, the direct connection between low-level convolutional features and the fully connected layer greatly increases the parameters to be estimated in the learning process. To decrease the dimension of the low-level convolutional features, we apply mean-pooling and max-pooling to the low-level convolutional features by channels, and both mean-pooled and max-pooled features are concatenated and fed to the fully connected layer. The dropout [48] is applied before all fully connected layers to prevent overfitting problems. After the fully connected layer, all multilevel features are transformed to a 1D vector ($\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4 \in \mathbb{R}^{128}$).

In the next stage of the multilevel feature fusion, we calculate the logit of each multilevel feature, where the logit is the unnormalized final score of the model before the softmax function. To obtain the logit, we perform L_2-normalization of the multilevel features and each column of the classifier ($\boldsymbol{w}_j \in \mathbb{R}^{128}$), and multiply the normalized terms together. L_2-normalization makes the logit of each multilevel feature fall in the range between -1 and 1, i.e., $-1 < \boldsymbol{w}_j^T \boldsymbol{x}_i < 1$ for $^\forall i, j$. Then, to weigh the logit of multilevel features that are equally scaled, we propose a model to compute the weighting factor $\boldsymbol{\alpha}$ corresponding to the logit. To obtain $\boldsymbol{\alpha}$, we aggregate the multilevel features ($\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$) and then the combined features are L_2-normalized. The L_2-normalization is performed after aggregating the features to utilize the relative magnitude of each multilevel feature. Then, the normalized feature is forwarded to a multilayer perceptron with one hidden layer and a softmax function, as shown in Fig. 5. It is expressed as

$$\boldsymbol{\alpha} = \sigma(\boldsymbol{M}^T \bar{\boldsymbol{\chi}} + \boldsymbol{b}), \tag{3}$$

where $\boldsymbol{\chi} \in \mathbb{R}^{512}$ is a concatenation of the multilevel features $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$, and $\boldsymbol{x}_4$; $\bar{\boldsymbol{\chi}}$ is the L_2-normalized $\boldsymbol{\chi}$; $\boldsymbol{M} \in \mathbb{R}^{512\times 4}$ is a hidden layer; $\boldsymbol{b} \in \mathbb{R}^4$ is a bias parameter; $\boldsymbol{\alpha} \in \mathbb{R}^4$ are the weighting factors; $\sigma()$ is a softmax function. The sum of each component $\boldsymbol{\alpha}$ is equal to one and it indicates the importance of the multilevel feature. The $\boldsymbol{\alpha}$ is determined by adding $\boldsymbol{b}$ to the product of the multilevel features and $\boldsymbol{M}$, where $\boldsymbol{M}$ and $\boldsymbol{b}$ are learnable parameters. Therefore, the appropriate weighting factor is adaptively determined depending on $\bar{\boldsymbol{\chi}}$ extracted from the 3D EEG images. The final prediction of class j is made by summing each output after weighting it by the weighting factor, as follows:

$$\text{Pred}(j) = s \cdot \sigma\left(\sum_{i=1}^{4} \alpha_i \bar{\boldsymbol{w}}_j^T \bar{\boldsymbol{x}}_i\right), \tag{4}$$

where $\boldsymbol{w}_i$ denotes the i-th column of the classifier $\boldsymbol{W} \in \mathbb{R}^{512\times 2}$, α_i denotes the i-th component of $\boldsymbol{\alpha}$, and $s > 0$ is a rescaling parameter with a positive real number. The rescaling parameter $s > 0$ is introduced to help the network converge. This is explained in detail in Section IIIC.

C. TRAINING METHOD IN MULTILEVEL FEATURE FUSION

In an earlier study on multilevel feature fusion [22], the length of a multilevel feature, which is selected from the pre-defined value, is considered as the importance of the feature. The study concatenated multilevel features and integrated them into one feature by using a fully connected layer to be used for classification. However, it is difficult to explain the effects of the multilevel features in the classification process. Therefore, to observe the effect of each multilevel feature separately during the classification process, we utilize the method of classifying each multilevel feature and then weighting each classification result. To do this, we decompose the logit into magnitude and direction of the multilevel feature and classifier, i.e., $\bar{\boldsymbol{w}}_j^T \bar{\boldsymbol{x}}_i = \|\bar{\boldsymbol{w}}_j^T\|_2 \|\bar{\boldsymbol{x}}_i\|_2 \cos\theta$, where $\|\bar{\boldsymbol{x}}\|_2$ is the L_2-norm of $\bar{\boldsymbol{x}}$. Because the norm of an L_2-normalized vector is one, only the angular information is used for the classification. Then, the classification results of each multilevel feature are weighted by multiplying the weighting factors. The detailed expression for the classification loss is described in 1) below and the gradient compensation method to alleviate the problem caused by the weighting factor is described in 2).

1) WEIGHTED MULTILEVEL ANGULAR LOSS

The most widely used classification loss function, cross-entropy loss with softmax function, is expressed as follows:

$$L = -\sum_{i=1}^{n} y_i \log \frac{e^{\boldsymbol{w}_i^T \boldsymbol{x} + b_i}}{\sum_{j=1}^{n} e^{\boldsymbol{w}_j^T \boldsymbol{x} + b_j}}, \tag{5}$$

where $\boldsymbol{x}$ denotes the deep feature, y_i is one hot-encoded class label of $\boldsymbol{x}$, b_j is a biased term of the j-th column, and n is the number of the class. For simplicity, we express the cross-entropy loss for a single instance, i.e., the batch size as one. In general, $\boldsymbol{x}$ is the feature extracted from the last convolutional layer and the logit, which is the final output of the network without the softmax function, can be expressed as $\boldsymbol{w}_i^T \boldsymbol{x} + b_i$. However, because our network utilizes all multilevel features, the logit of our network can be expressed as $\sum_{k=1}^{4}(\bar{\boldsymbol{w}}_i^T \bar{\boldsymbol{x}}_k + b_i)$. Note that $\bar{\boldsymbol{w}}_i^T$ and $\bar{\boldsymbol{x}}_i$ are L_2-normalized. We set $b_i = 0$ for $^\forall i$ as in [49]–[52]. Then, the logit can be expressed as

$$\begin{aligned}\sum_{i=1}^{4}(\bar{\boldsymbol{w}}_j^T \bar{\boldsymbol{x}}_i + b_j) &= \sum_{i=1}^{4}(\|\bar{\boldsymbol{w}}_j^T\|_2 \|\bar{\boldsymbol{x}}_i\|_2 \cos\theta_{j,i}) \\ &= \sum_{i=1}^{4} \cos\theta_{j,i},\end{aligned} \tag{6}$$

where $\theta_{j,i}$ is the angle between two vectors $\bar{\boldsymbol{w}}_j^T$ and $\bar{\boldsymbol{x}}_i$. In this logit, only the angular information $\cos\theta_{j,i}$ is used. It means L_2-normalization forces the prediction to depend only on the angle $\theta_{j,i}$ between the classifier $\bar{\boldsymbol{w}}_j^T$ and features $\bar{\boldsymbol{x}}_i$. Using the above Eq. (6), the multilevel angular loss can be written as

$$L = -\sum_{i=1}^{n} y_i \log \frac{\exp\left(\sum_{k=1}^{4} \cos\theta_{i,k}\right)}{\sum_{j=1}^{n} \exp\left(\sum_{k=1}^{4} \cos\theta_{j,k}\right)}. \tag{7}$$

Next, to allocate the importance of each multilevel feature, the component of the weighting factor α_i is multiplied with the corresponding cosine values $\cos\theta_{j,i}$, as follows:

$$L = -\sum_{i=1}^{n} y_i \log \frac{\exp\left(\sum_{k=1}^{4} \alpha_k \cos\theta_{i,k}\right)}{\sum_{j=1}^{n} \exp\left(\sum_{k=1}^{4} \alpha_k \cos\theta_{j,k}\right)}. \tag{8}$$

Because $\cos\theta_{j,i}$ represents the logit of each multilevel feature $\bar{x}_i$, we can determine the importance of the multilevel feature $\bar{x}_i$ by multiplying the corresponding component of the weighting factor α_i.

In [49], it was proved that if the feature and each column of the classifier are L_2-normalized, the cross entropy loss with the softmax function will have a lower bound, $\log(1 + (n-1)e^{-n/(n-1)})$, where n is the number of classes. When n is large, this loss converges to a very large value after few thousands of iterations and the network fails to reach its optimal performance. Therefore, to generalize the model for a large number of classes, we compensate the logit by multiplying the rescaling parameter s that is trained by the backpropagation algorithm as in [49]. Therefore, our final loss equation can be obtained by rescaling the logit as follows:

$$L = -\sum_{i=1}^{n} y_i \log \frac{\exp\left(s\sum_{k=1}^{4} \alpha_k \cos\theta_{i,k}\right)}{\sum_{j=1}^{n} \exp\left(s\sum_{k=1}^{4} \alpha_k \cos\theta_{i,k}\right)}, \tag{9}$$

where the rescaling parameter s is a learnable parameter.

In summary, the classification of each feature is conducted with only angular information and the importance of each feature is determined by multiplying the weighting factors.

2) GRADIENT COMPENSATION FOR TRAINING MULTILEVEL FEATURES

We multiply the component of the weighting factors $\boldsymbol{\alpha}_i$ by each logit of the multilevel feature $\bar{\boldsymbol{x}}_i$. During the backpropagation process, the training speed of the multilevel features depends on $\boldsymbol{\alpha}_i$. Higher $\boldsymbol{\alpha}_i$ stimulates faster training speed whereas smaller $\boldsymbol{\alpha}_i$ prevents the convergence of the learning of the multilevel feature. Therefore, in this subsection, we clearly define the problem and describe a gradient compensation algorithm to address the problem. We can rewrite Eq. (9) as

$$L = -\sum_{i=1}^{n} y_i \log \frac{\exp\left(s\sum_{k=1}^{4} \alpha_k \bar{\boldsymbol{w}}_i^T \bar{\boldsymbol{x}}_k\right)}{\sum_{j=1}^{n} \exp\left(s\sum_{k=1}^{4} \alpha_k \bar{\boldsymbol{w}}_j^T \bar{\boldsymbol{x}}_k\right)}. \tag{10}$$

The gradient flow to component m of multilevel feature $\boldsymbol{x_l}$ can be computed as

$$\frac{\partial L}{\partial x_{l,m}} = \left(\bar{x}_{l,m}\frac{\partial \alpha_l}{\partial x_{l,m}} + \alpha_l \frac{\partial \bar{x}_{l,m}}{\partial x_{l,m}}\right)\psi, \tag{11}$$

where,

$$\psi = -\sum_{i=1}^{n} s y_i \frac{\sum_{j=1}^{n}(w_{i,m}^T - w_{j,m}^T)\exp\left(s\sum_{k=1}^{4}\alpha_k \bar{\boldsymbol{w}}_i^T \bar{\boldsymbol{x}}_k\right)}{\sum_{j=1}^{n}\exp\left(s\sum_{k=1}^{4}\alpha_k \bar{\boldsymbol{w}}_j^T \bar{\boldsymbol{x}}_k\right)}, \tag{12}$$

$x_{l,m}$ denotes component m of the vector $\boldsymbol{x}_l$, $l \in \{1, 2, 3, 4\}$, and $m \in \{1, 2, \cdots, 128\}$. Gradient flow of $x_{l,m}$ is divided into the gradient flows of α_l and $\bar{x}_{l,m}$. Because α_l is multiplied to $\bar{x}_{l,m}$, the component of the weighting factor α_l serves to regulate the learning speed of the multilevel feature $\bar{\boldsymbol{x}}_l$ similar to the learning rate. Therefore, CNN cannot be trained properly to extract the multilevel feature $\bar{\boldsymbol{x}}_l$ if the component of the weighting factor α_l is much smaller than the other components of the weighting factor. In the backpropagation process, we multiply the inverse of α_l to compensate for the different learning speeds of the multilevel features $\bar{\boldsymbol{x}}_l$:

$$\frac{\partial L}{\partial x_{l,m}} \leftarrow \left(\bar{x}_{l,m}\frac{\partial \alpha_l}{\partial x_{l,m}} + \frac{1}{\alpha_l}(\alpha_l \cdot \frac{\partial \bar{x}_{l,m}}{\partial x_{l,m}})\right)\psi. \tag{13}$$

Only the weights of the convolutional and fully connected layers used to extract $\bar{\boldsymbol{x}}_i$ are affected by the gradient compensation, and $\boldsymbol{M}$ and $\boldsymbol{b}$, which are used to extract $\boldsymbol{\alpha}$ and the classifier $\boldsymbol{W}$, are not affected by the gradient compensation.

D. MULTILEVEL FEATURE FUSION WITH TEMPORAL ATTENTION

It is crucial to consider temporal information of EEG data in workload estimation because the brain functioning for working memory task consists of two processes: manipulation and retention. After stimulation showing all words (Fig. 1), the process of information manipulation, which has different spectral patterns compared to the retention process, begins immediately [44]. In addition, according to [35], the difference in spectral power due to task difficulty appears in a certain time during the retention process, and the timing varies from person to person. Therefore, it is necessary to find the individual time interval that has the spectral feature related with the difficulty of the task. To extract the time interval, we propose an LSTM based temporal attention method.

The 3D EEG image is modified by the following procedure to contain temporal information. We divided the raw 2 s EEG data into seven segments through an overlapping sliding window with a window size of 0.5 s and a stride of 0.25 s. Then, each segment is converted into a 3D EEG image as done in Section III-A. Therefore, we can get seven frames of 32 × 32 × 3 EEG images. Each frame is fed to the 3D CNN structure which shares the same weights with all frames. The multilevel features and weighting factor of each frame are computed as shown in Fig. 4.

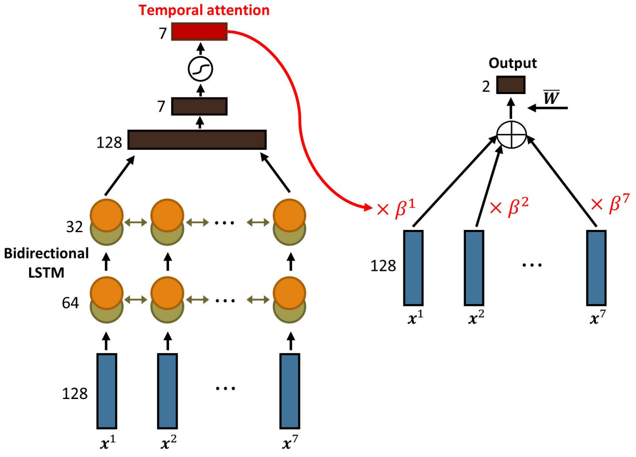

FIGURE 6. Framework of the temporal attention.

The final feature of each frame is the weighted sum of multilevel features as follows:

$$\boldsymbol{x}^t = \sum_{i=1}^{4} \alpha_i^t \bar{\boldsymbol{x}}_i^t, \tag{14}$$

where $\boldsymbol{x}^t$ is the final feature at frame t, and α_i^t and $\bar{\boldsymbol{x}}_i^t$ are t-th frame of α_i and $\bar{\boldsymbol{x}}_i$, respectively. $\boldsymbol{x}^t$ is fed into two stacked bidirectional LSTM to calculate the importance of each frame. The temporal attention β^t is estimated by putting the output of the LSTM layer into the fully connected layer and softmax function. This temporal attention β^t is multiplied by the corresponding feature $\boldsymbol{x}^t$, and the result is multiplied by the L_2-normalized classifier as shown in Fig. 6. Then, the final loss function of the proposed multilevel feature fusion with temporal attention can be written as:

$$L = -\sum_{i=1}^{n} y_i \log \frac{\exp\left(s \sum_{t=1}^{7} \sum_{k=1}^{4} \alpha_k^t \beta^t \bar{\boldsymbol{w}}_i^T \bar{\boldsymbol{x}}_k^t\right)}{\sum_{j=1}^{n} \exp\left(s \sum_{t=1}^{7} \sum_{k=1}^{4} \alpha_k^t \beta^t \bar{\boldsymbol{w}}_j^T \bar{\boldsymbol{x}}_k^t\right)}. \tag{15}$$

As shown in the above equation, each feature $\bar{\boldsymbol{x}}_i^t$ is multiplied by two factors α_i^t and β^t, which means that features are weighted based on the information included in their time segment and level of the layer.

IV. RESULTS

A. BEHAVIOR RESULTS

In this study, we performed the Sternberg task to measure the mental workload. We analyzed the subject's behavior by accuracy and reaction time, which is summarized in Fig. 7. The accuracy indicates the percentage of correct answers in the Sternberg task and the reaction time represents the duration between the display of the test stimulus and pressing of the "yes" or "no" button. The basic hypothesis is that as the difficulty of the workload increases, the accuracy tends to decrease and the reaction time tends to increase. As shown in Fig. 7, the behavior results are consistent with the hypothesis: the accuracy decreases from 94.5% to 82.5% and the reaction time increases from 758.7 ms to 882.9 ms. It shows that both accuracy and reaction time are significantly different (paired t-test, $p < 0.001$) between the two tasks (low and high workload conditions). Therefore, the results confirm that 3-stimuli and 7-stimuli tasks have different difficulty levels, and the Sternberg task is well designed to distinguish between low and high workload conditions.

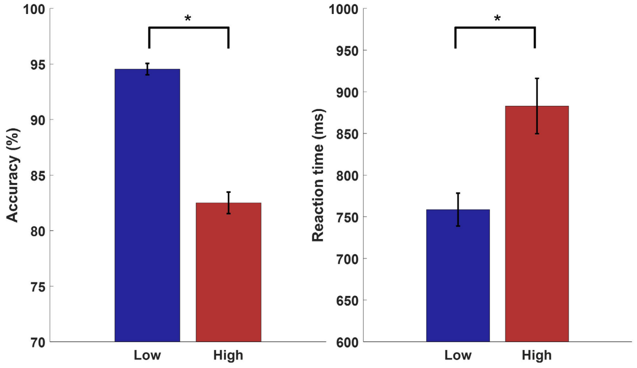

FIGURE 7. Average of accuracy and reaction time for low and high workload conditions. Error bar represents the standard error and asterisk indicates a significant difference (paired t-test, $p < 0.001$) between low and high workload conditions.

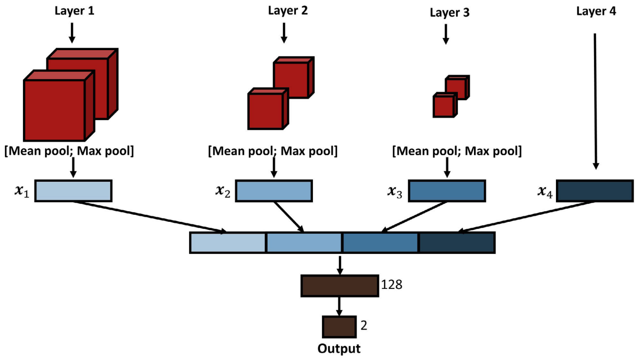

FIGURE 8. Conventional multilevel fusion method [22].

B. CLASSIFICATION RESULTS

We adopted the 10-fold cross-validation to verify the performance of the proposed method. We created a dataset by collecting all data of 62 participants and randomly divided the dataset into 10 non-overlapping sets. One set was selected for the test, and the others were used for training. We repeated the training 10 times, and then averaged the classification results of the ten sets. We compared the classification results with those of other algorithms, which were implemented as follows:

SVM: To extract the spectral feature, EEG signals were re-montaged to a nearest neighbor Laplacian filter. We calculated the power spectral density from 2 Hz to 40 Hz in steps of 2 Hz using the Welch's method. The power spectral densities of 30 EEG electrodes were combined and transformed into a 1D vector of 600 (20 × 30)-dimension. Then, the standard normalization features of these vectors were used for SVM training. We used a radial basis function with kernel size of $\{0.1, 0.5, 1, 5, 10\}$. The best result among the parameters was used for the analysis.

k-nearest neighbor (K-NN): We tested the KNN algorithm using 600-dimensional features exploited for the SVM algorithm. In addition, we set the parameter $k \in \{3, 5, 7, 10\}$, which determines the number of nearest neighbors, and the best result was used for comparison.

Baseline: For the baseline algorithm, we exploited the deep feature $\boldsymbol{x_4}$ extracted from the last convolutional layer in the proposed 3D CNN structure and used it for classification. The CNN parameters were randomly initialized with Xavier initialization [53]: the parameters were standard normalized with Gaussian distributions with zero mean and standard deviation $\sqrt{2/(n_{in}+n_{out})}$, where n_{in} is the dimension of the neuron input and n_{out} is the dimension of the neuron output. We trained the proposed network for 400 epochs with an Adam optimizer [54] with an initial learning rate of 0.001 and the learning rate was decreased to 0.0001 after 200 epochs. The batch size was set to 32.

2DCNN + LSTM/1D-Conv [16]: We produced seven frames from one EEG image by using a 0.5 s window with an overlap of 0.25 s. Each frame is generated by stacking topology maps on three frequency bands of theta (4-7 Hz), alpha (8-13 Hz) and beta (13-30 Hz), thus the size of a single frame is $32 \times 32 \times 3$ as done in [16]. The 2D convolutional layer extracts spatial information, and its output is fed to the LSTM and 1D convolutional layer to extract temporal information. The output of the LSTM and 1D convolutional layer is concatenated and fed to a fully connected layer for workload classification.

3DCNN + LSTM [21]: This method extracts spatial and spectral information using a 3D convolutional layer and then utilizes a bidirectional LSTM to extract temporal information. For a fair comparison, we replace the 3D convolutional layer used in [21] with our baseline architecture, and its output is fed into two stacked bidirectional LSTM layers as done in [21].

Conventional multilevel fusion [22]: The EEG data used in [22] contains 5,184 samples collected from 9 participants using 25 electrodes, which has a similar data size with our dataset. For a fair comparison, both methods have the same learning structure except for the feature fusion part. The feature extraction process, which extracts multilevel features, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$, from the 3D CNN structure, was the same as in the proposed method. However, the length of each feature was used to determine the weight of each feature [22]. As in [22], we predefined a set of feature lengths as $l = \{32, 64, 128\}$, and assigned a value from this set as the length of each feature ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_4$) using the greedy algorithm. Then, all features were concatenated and followed by the fully connected layer and softmax function, as shown in Fig. 8. The CNN parameters were initialized as in the baseline method.

Proposed method: The rescaling factor s was initialized to one and the other parameters were initialized in the same way as in the baseline initialization method.

TABLE 1. Performance comparison of different algorithms.

Algorithm	Accuracy
SVM	78.8%
KNN	80.6%
2DCNN+LSTM/1D-Conv [16]	80.8%
Conventional multilevel fusion [22]	88.3%
3DCNN+LSTM [21]	89.0%
Baseline (3DCNN)	85.6%
3DCNN+multilevel fusion	90.3%
3DCNN+multilevel fusion + temporal attention	**90.8%**

TABLE 2. Confusion matrix for the proposed algorithm (3DCNN + multilevel fusion + temporal attention) with average of 10-fold cross-validation.

	Low workload	**High workload**	**Sensitivity**
Low workload	261.8	19.8	92.3%
High workload	29.2	219.3	88.3%
Precision	90.0%	91.7%	

The classification accuracies of different algorithms are summarized in Table 1. The result shows that the CNN based algorithms outperform the traditional classifiers, SVM and KNN. It implies that the handcrafted features are not optimal for the classifiers, but CNN can extract the good features using network training. In addition, the conventional multilevel fusion [22] and the proposed algorithm performs at least 3.0 % better than the baseline. This result shows that the multilevel features are useful in extracting robust EEG features across two mental tasks and achieve higher accuracy of workload estimation. Also, we verify that our multilevel feature fusion method is superior to the conventional multilevel feature fusion algorithm. It can be inferred that the optimal weight of each multilevel feature obtained by the proposed method can improve the accuracy by 2.0 %. The proposed temporal attention put higher weighting on strongly correlated frames with workload estimation and improves the accuracy by 0.5 % compared to the proposed multilevel fusion without the temporal attention. Finally, our multilevel feature fusion with temporal attention achieves the highest classification accuracy among conventional deep learning algorithms. To provide more detailed classification performance, we build a confusion matrix of the proposed algorithm as shown in Table 2 and plot the training loss and test accuracy along with epoch as shown in Fig. 9.

C. ANALYSIS OF WEIGHTING FACTOR

Here, we analyze the effect of the weighting factor $\boldsymbol{\alpha}$ on the classification performance. To show the difference in performance between a predefined weighting factor and a learned weighting factor, we train our network with a fixed weighting factor with equal components, $\boldsymbol{\alpha} = [0.25, 0.25, 0.25, 0.25]$, i.e., the weighting factor is not optimally adjusted by the proposed learning process expressed in Eq. (3) (see Fig. 5). As shown in Table 3, the classification accuracy is drastically reduced from 90.3 % to 88.5 % when the weighting factor, α, is fixed at a suboptimal value. It reveals that the importance

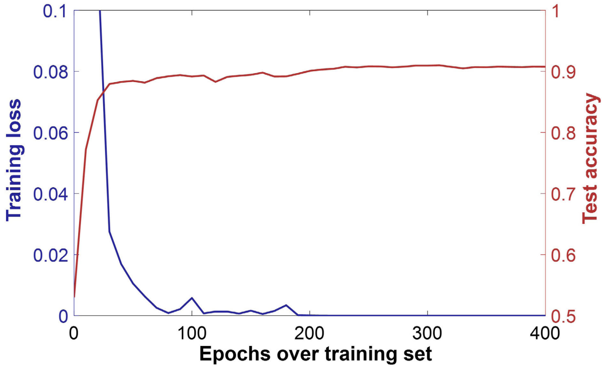

FIGURE 9. Training loss and test accuracy with training epochs of the proposed algorithm (3DCNN + multilevel fusion + temporal attention). Blue line indicates training loss and Red line indicates test accuracy. Each line is calculated of averaging all cross-validation folds.

TABLE 3. Classification accuracy according to the method for determining the weighting factor.

Algorithm	Accuracy
3DCNN+multilevel fusion with fixed $\boldsymbol{\alpha}$	88.5%
3DCNN+multilevel fusion	**90.3%**

of the intermediate features should be adaptively determined by the learning process that optimizes the cross-entropy loss function.

Fig. 10 shows the cumulative distribution function (CDF) of each component ($\alpha_1, \alpha_2, \alpha_3$, and α_4) of $\boldsymbol{\alpha}$ for all training data. To compute the CDF, we analyzed $\boldsymbol{\alpha}$ of the network for determining the best performance among ten networks trained in 10-fold cross-validation. The weighting factor $\boldsymbol{\alpha}$ was not fixed to one specific value during the training; it changed with the test data, as shown in Fig. 10. Because the weighting factor is obtained by multiplying the parameter $\boldsymbol{M}$ and the multilevel feature $\bar{\boldsymbol{x}}_k$ determined by the EEG image, different weighting factors are extracted according to the EEG images. Interestingly, we can show that the components of the weighting factors do not have equal values. The weighting factors, α_2 and α_3 for the intermediate features $\bar{\boldsymbol{x}}_2$ and $\bar{\boldsymbol{x}}_3$ are mainly distributed between 0 and 0.05, which are very small. However, the weighting factors, α_1 and α_4 for the first and last features $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_4$ are mainly distributed between 0.2 and 0.8, which are relatively high. It can be deduced from this result that the first convolutional layer has richer local structure information and the last convolutional layer has the most abundant global structure information. In summary, performance improvement can be achieved by determining appropriate $\boldsymbol{\alpha}$ for each image through the learning process.

D. APPLICATION TO THE PUBLIC EEG DATASET

To verify the generalization of the proposed multilevel feature fusion method, we evaluate the performance using the public EEG dataset for workload estimation published in [16]. The dataset contains EEG signals recorded at a sampling rate of 500 Hz using 64 electrodes under the standard 10-10 system from fifteen participants. They conducted the Sternberg task to measure workload of working memory and its experimental protocol is briefly described below.

The experimental protocol is divided into three phases: encoding phase, retention phase, and test phase. In the encoding phase, participants were required to memorize English characters, where the number of characters was randomly chosen to be 2, 4, 6 or 8 for each trial. Then, the characters were disappeared, and the participants were required to retain the presented characters for 3.5 s (i.e., retention period). In the last phase, the test character was shown in the monitor and participants were asked to answer whether the test character belonged to the previously presented characters. The number of trials for each participant is 240. Two participants are excluded from the dataset because of excess noise and artifacts in EEG signals [16], [17]. In addition, only those trials in which the participants correctly answered were used for analysis. As a result, 2,670 samples collected from 13 participants were utilized for a 4-class classification: workloads of 2, 4, 6, and 8. We train the networks using EEG data recorded during the retention period and use the 13-fold leave-one-subject-out cross validation strategy presented in [16] and [17].

For a fair comparison, the proposed method is based on the 2D CNN structure proposed in [16]. We extract multilevel features from all pooling layers and apply our multilevel feature fusion method to those features. Our temporal attention is also applied to sequential features extracted from each frame of EEG images. The performance results of the traditional classifiers implemented in [16], the deep learning algorithms [16], [17] and the proposed method are summarized in Table 4.

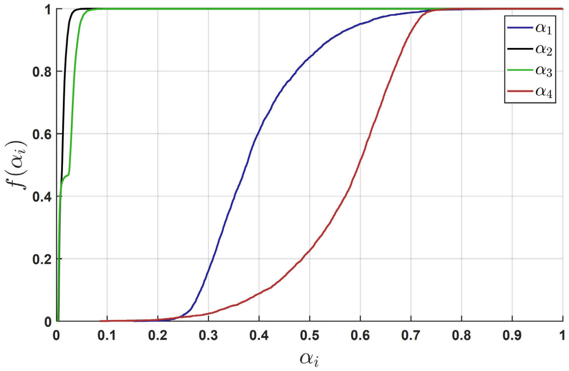

FIGURE 10. Cumulative distribution function of α.

TABLE 4. Performance comparisons on the public EEG dataset.

Algorithm	Accuracy
SVM	85.3%
Logistic Regression	85.5%
Random Forest	85.6%
2DCNN+LSTM/1D-Conv [16]	91.1%
Fused CNNs [56]	92.4%
Proposed algorithm	**93.9%**

The results show that the proposed method has 2.8 % higher accuracy than the original 2D CNN + LSTM/1D-Conv structure [16], although the proposed method is based on the same CNN structure of the 2DCNN + LSTM/1D-Conv. Moreover, the accuracy of the proposed method is 1.5% higher than the Fused CNNs [17] that is the state-of-the-art method using the same EEG dataset. The significant improvement verifies that the proposed multilevel feature fusion with temporal attention could enhance the performance of conventional CNN structures.

V. CONCLUSION

To simultaneously extract EEG features contain both local and global structure information, we proposed a 3D CNN based multilevel feature fusion algorithm for mental workload estimation. The multi-channel EEG data was transformed to 3D EEG images that contained spectral and spatial information. Then, we extracted the multilevel features from the 3D convolutional operation and each multilevel feature was multiplied by the weighting factor, which determines the importance of each multilevel feature. The weighting factors were adaptively optimized by the proposed learning process according to the EEG image, which is essential to enhance the performance of the proposed structure when compared with the case where a fixed weighting factor is used. In addition, the proposed temporal attention extracts the significant time interval with spectral features that are strongly correlated with task difficulty. The results proved that the multilevel feature fusion method could improve the performance of the 2D/3D CNN structure for mental workload estimation. Moreover, the proposed model achieved an accuracy of 90.8 % on our dataset and state-of-the-art accuracy of 93.9 % on the public dataset. Thus, it outperformed the traditional classifiers that use handcrafted features, the conventional 2D/3D CNN algorithms, and the conventional multilevel fusion algorithm. These findings coupled with our successful application of the proposed method to predict mental workload using EEG data suggest that our method for optimizing the weighting factor will be a useful tool for image classification in other fields.

REFERENCES

[1] N. Moray, *Mental Workload: Its Theory and Measurement*. New York, NY, USA: Plenum, 1979.

[2] F. G. W. C. Paas and J. J. G. Van Merriënboer, "Instructional control of cognitive load in the training of complex cognitive tasks," *Educ. Psychol. Rev.*, vol. 6, no. 4, pp. 351–371, Dec. 1994.

[3] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educ. Psychologist*, vol. 38, no. 1, pp. 63–71, Mar. 2003.

[4] R. Parasuraman, K. A. Cosenzo, and E. De Visser, "Adaptive automation for human supervision of multiple uninhabited vehicles: Effects on change detection, situation awareness, and mental workload," *Mil. Psychol.*, vol. 21, no. 2, pp. 270–297, Apr. 2009.

[5] A. Saha, V. Minz, S. Bonela, S. R. Sreeja, R. Chowdhury, and D. Samanta, "Classification of EEG signals for cognitive load estimation using deep learning architectures," in *Proc. Int. Conf. Intell. Hum. Comput. Interact.* Cham, Switzerland: Springer, 2018, pp. 59–68.

[6] E. Solovey, P. Schermerhorn, M. Scheutz, A. Sassaroli, S. Fantini, and R. Jacob, "Brainput: Enhancing interactive systems with streaming fnirs brain input," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 2193–2202.

[7] J. B. Brookings, G. F. Wilson, and C. R. Swain, "Psychophysiological responses to changes in workload during simulated air traffic control," *Biol. Psychol.*, vol. 42, no. 3, pp. 361–377, Feb. 1996.

[8] L. R. Fournier, G. F. Wilson, and C. R. Swain, "Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training," *Int. J. Psychophysiol.*, vol. 31, no. 2, pp. 129–145, Jan. 1999.

[9] A. Gundel and G. F. Wilson, "Topographical changes in the ongoing EEG related to the difficulty of mental tasks," *Brain Topogr.*, vol. 5, no. 1, pp. 17–25, 1992.

[10] O. Jensen and C. D. Tesche, "Frontal theta activity in humans increases with memory load in a working memory task," *Eur. J. Neurosci.*, vol. 15, no. 8, pp. 1395–1399, Apr. 2002.

[11] G. Pfurtscheller, A. Stancák, Jr., and C. Neuper, "Event-related synchronization (ERS) in the alpha band—An electrophysiological correlate of cortical idling: A review," *Int. J. psychophysiol.*, vol. 24, nos. 1–2, pp. 39–46, 1996.

[12] H. Van Dijk, J.-M. Schoffelen, R. Oostenveld, and O. Jensen, "Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability," *J. Neurosci.*, vol. 28, no. 8, pp. 1816–1823, Feb. 2008.

[13] C. Mühl, C. Jeunet, and F. Lotte, "EEG-based workload estimation across affective contexts," *Frontiers Neurosci.*, vol. 8, p. 114, Jun. 2014.

[14] A.-M. Brouwer, M. A. Hogervorst, J. B. F. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using EEG spectral power and ERPs in the n-back task," *J. Neural Eng.*, vol. 9, no. 4, Aug. 2012, Art. no. 045008.

[15] J. C. Christensen, J. R. Estepp, G. F. Wilson, and C. A. Russell, "The effects of day-to-day variability of physiological data on operator functional state classification," *NeuroImage*, vol. 59, no. 1, pp. 57–63, Jan. 2012.

[16] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*. [Online]. Available: https://arxiv.org/abs/1511.06448

[17] Z. Jiao, X. Gao, Y. Wang, J. Li, and H. Xu, "Deep convolutional neural networks for mental load classification based on EEG data," *Pattern Recognit.*, vol. 76, pp. 582–595, Apr. 2018.

[18] Y.-C. Hung, Y.-K. Wang, M. Prasad, and C.-T. Lin, "Brain dynamic states analysis based on 3D convolutional neural network," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 222–227.

[19] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[20] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, Nov. 2018.

[21] P. Zhang, X. Wang, W. Zhang, and J. Chen, "Learning spatial–spectral–temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 1, pp. 31–42, Jan. 2019.

[22] S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Bencherif, and M. S. Hossain, "Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification," *IEEE Access*, vol. 7, pp. 18940–18950, 2019.

[23] Y. Yang, Q. Wu, Y. Fu, and X. Chen, "Continuous convolutional neural network with 3D input for EEG-based emotion recognition," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2018 pp. 433–443.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[26] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[27] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[28] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3D convolutional neural network for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2164–2177, Oct. 2019.

[29] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[30] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3469–3476.

[31] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1208–1212, Aug. 2017.

[32] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[33] G. Tononi, O. Sporns, and G. M. Edelman, "A measure for brain complexity: Relating functional segregation and integration in the nervous system," *Proc. Nat. Acad. Sci. USA*, vol. 91, no. 11, pp. 5033–5037, May 1994.

[34] K. J. Friston, "Modalities, modes, and models in functional neuroimaging," *Science*, vol. 326, no. 5951, pp. 399–403, Oct. 2009.

[35] P. Sauseng, W. Klimesch, M. Doppelmayr, T. Pecherstorfer, R. Freunberger, and S. Hanslmayr, "EEG alpha synchronization and functional coupling during top-down processing in a working memory task," *Hum. Brain Mapping*, vol. 26, no. 2, pp. 148–155, Oct. 2005.

[36] O. Jensen, "Oscillations in the alpha band (9-12 Hz) increase with memory load during retention in a short-term memory task," *Cerebral Cortex*, vol. 12, no. 8, pp. 877–882, Aug. 2002.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] S. Sternberg, "High-speed scanning in human memory," *Science*, vol. 153, no. 3736, pp. 652–654, Aug. 1966.

[39] O. Jensen and J. E. Lisman, "An oscillatory short-term memory buffer model can account for data on the sternberg task," *J. Neurosci.*, vol. 18, no. 24, pp. 10688–10699, 1998.

[40] P. Bashivan, G. M. Bidelman, and M. Yeasin, "Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity," *Eur. J. Neurosci.*, vol. 40, no. 12, pp. 3774–3784, Dec. 2014.

[41] P. L. Nunez and K. L. Pilgreen, "The spline-laplacian in clinical neurophysiology: A method to improve EEG spatial resolution," *J. Clin. Neurophysiol., Off. Publication Amer. Electroencephalogr. Soc.*, vol. 8, no. 4, pp. 397–413, 1991.

[42] A. Gevins, M. E. Smith, H. Leong, L. Mcevoy, S. Whitfield, R. Du, and G. Rush, "Monitoring working memory load during computer-based tasks with EEG pattern recognition methods," *Hum. Factors*, vol. 40, no. 1, pp. 79–91, Mar. 1998.

[43] J. Le, V. Menon, and A. Gevins, "Local estimate of surface Laplacian derivation on a realistically shaped scalp surface and its performance on noisy data," *Electroencephalogr. Clin. Neurophysiol./Evoked Potentials Sect.*, vol. 92, no. 5, pp. 433–441, Sep. 1994.

[44] L. Cornelissen, S. Kim, J. Lee, E. Brown, P. Purdon, and C. Berde, "Electroencephalographic markers of brain development during sevoflurane anaesthesia in children up to 3 years old," *Brit. J. Anaesthesia*, vol. 120, no. 6, pp. 1274–1286, Jun. 2018.

[45] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[46] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: https://arxiv.org/abs/1511.07289

[47] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Comput. Biol. Med.*, vol. 100, pp. 270–278, Sep. 2018.

[48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[49] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L_2 hypersphere embedding for face verification," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1041–1049.

[50] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[51] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," 2017, *arXiv:1703.09507*. [Online]. Available: https://arxiv.org/abs/1703.09507

[52] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.

[53] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

YOUNGCHUL KWAK received the B.S. and M.S. degrees in electrical engineering from the Pohang University of Science Technology (POSTECH), Pohang, South Korea, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree. He has been a Research Assistant with the Department of Electrical Engineering, POSTECH, since 2016. His current research interests include biomedical signal processing, machine learning, and deep learning.

KYEONGBO KONG received the B.S. degree in electronics engineering from Sogang University, Seoul, South Korea, in 2015, and the M.S. degree in electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2017, where he is currently pursuing the Ph.D. degree. He has been a Research Assistant with the Department of Electrical Engineering, POSTECH, since 2015. His current research interests include image processing, computer vision, machine learning, and deep learning.

WOO-JIN SONG (Member, IEEE) was born in Seoul, South Korea, in 1956. He received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, in 1979 and 1981, respectively, and the Ph.D. degree in electrical engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1986.

From 1981 to 1982, he was with the Electronics and Telecommunication Research Institute (ETRI), Daejeon, South Korea. In 1986, he was employed by Polaroid Corporation, a Senior Engineer, working on digital image processing, where he was promoted to a Principal Engineer, in 1989. In 1989, he joined the Faculty of the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, where he is currently a Professor of electronic and electrical engineering. His current research interests include digital signal processing, in particular, radar signal processing, signal processing for digital television and multimedia products, and adaptive signal processing.

BYOUNG-KYONG MIN (Member, IEEE) received the M.S. degree in neurobiology and physiology from Northwestern University, Evanston, IL, USA, in 1998, and the Ph.D. degree in biological psychology from Magdeburg University, Germany, in 2007.

From 2007 to 2009, he was a Postdoctoral Fellow with the Yonsei University College of Medicine (Severance Hospital), Seoul, South Korea. From 2009 to 2011, he was a Research Fellow with the Harvard Medical School (Brigham and Women's Hospital), Boston, MA, USA. Since 2017, he has been an Associate Professor with the Department of Brain and Cognitive Engineering, Korea University, Seoul. From 2018 to 2019, he was a Visiting Scholar with the McGovern Institute for Brain Research, Massachusetts Institute of Technology, Boston, USA. His research interests include spectral analysis of brain electrical activity (EEG) and EEG-based brain–machine interfaces (BMIs). He has combined ultrasound sonication with an EEG-based BMI to accomplish a non-invasive human brain-to-brain interface. He also serves as an Editor of the journal *Medicine* and as the publication chair of two IEEE international conferences and workshops.

SEONG-EUN KIM (Member, IEEE) received the B.S. and Ph.D. degrees in electronics and electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2004 and 2010, respectively.

From 2011 to 2014, he was a Research Staff Member with the Samsung Advanced Institute of Technology, Yongin, South Korea. From 2014 to 2017, he worked as a Postdoctoral Associate at the MIT/Harvard Neuroscience Statistics Research Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA, where he was involved in developing neural signal processing algorithms. In 2017, he moved to Hanbat National University, Daejeon, South Korea, where he is currently an Assistant Professor with the Department of Electronics and Control Engineering. His research interests include statistical and adaptive signal processing, biomedical signal processing and engineering, and systems and computational neuroscience.