Expert Systems With Applications 213 (2023) 119032

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

LSTM-CNN model of drowsiness detection from multiple consciousness states acquired by EEG

Chungho Lee [a], Jinung An [a,b,*]

[a] Division of Intelligent Robot, DGIST, Daegu, Republic of Korea
[b] Interdisciplinary Studies, Graduate School, DGIST, Daegu, Republic of Korea

ARTICLE INFO

Keywords:
EEG (electroencephalogram)
Drowsiness detection
Multiclass classification
Input vector length optimization
LSTM (long-short term memory)
CNN (convolutional neural network)

ABSTRACT

This study aimed to design a deep neural network for electroencephalography (EEG)-based drowsiness detection in multiple consciousness states, i.e., "awake," "sleep," and "drowsiness." Few studies have seriously considered the optimal input vector size or labeling method in classifying multiple consciousness states, which may affect classification performance. To determine the optimal input vector length, i.e., window length, three neural network models (long short-term memory [LSTM], convolutional neural network [CNN], and combined LSTM and CNN) and four feature-based models were tested with six different levels of window length. The EEG dataset was acquired from 19 participants with randomly assigned auditory stimuli and button responses. The EEG data were labeled into three classes (awake, sleep, and drowsiness) based on the defined button response pattern corresponding to the stimuli. The results demonstrated that when the input vector size exceeded 8 sec, the performance of the neural network models dropped rapidly; however, when the window size was less than 8 sec, the performance change according to the window size was small. In contrast, the performance of feature-based models increased continuously as the window size increased. The LSTM model yielded the best accuracy (86%) for a 1 sec window length, and the LSTM-CNN model yielded the best kappa index (0.77) for a 4 sec window length. In addition, the proposed model was applied to the binary classification of normal consciousness (awake) and low consciousness (drowsiness and sleep) states to determine whether this model works appropriately in actual applications such as drowsiness detection in a driving environment. For binary classification, the LSTM-CNN model resulted in 0.95 F_1 scores in 4000-ms. When a short input data (500 msec) is used, the LSTM-CNN model resulted in an average accuracy of 85.6% and a kappa index of 0.77 for the three-class classification problem and 0.94 F_1 scores for the binary classification problem. In conclusion, we demonstrated that the proposed model could effectively detect drowsiness. Furthermore, a significant correlation was found between reaction time and drowsiness. However, using the reaction time as an index for labeling drowsiness was challenging because of the high false-negative ratio.

1. Introduction

Drowsiness can be defined as a progressive loss of cortical processing efficiency (Slater, 2008). This inefficiency leads to a gradual loss of cognitive function. According to a previous study, 56 % of night-shift nurses reported sleep deprivation (Johnson, Brown, & Weaver, 2010), resulting in a significantly higher probability of patient care errors (Johnson, Jung, Brown, Weaver, & Richards, 2014). Another study that analyzed industrial injury incidents reported 1.23 times higher injury risk on the night-shift than on the morning-shift under a three-shift system (Smith, Folkard, & Poole, 1994).

Electroencephalography (EEG) is a method used to record brain activity by measuring voltage fluctuations resulting from ionic currents within the brain's neuron (Henry, 2006). Owing to the EEG patterns that appear according to sleep stages (Rodenbeck, Binder, Geisler, Danker-Hopfe, Lund, Raschke, Weeß, & Schulz, 2006), EEG has been adopted in drowsiness detection systems in several studies.

In a previous study by Yeo, Li, Shen, and Wilder-Smith (2009), EEG signals were recorded for one hour from 20 subjects during a driving simulation (Yeo et al., 2009). The recorded data were manually labeled as "alert" or "drowsy" with specific rules based on eye blink frequency and EEG activity, segmented into 10-s epochs. For each epoch, delta,

* Corresponding author at: R4-708, 333, Techno jungang-daero, Hyeonpung-eup, Dalseong-gun, Daegu, Republic of Korea.
E-mail addresses: filomaq@dgist.ac.kr (C. Lee), robot@dgist.ac.kr (J. An).

https://doi.org/10.1016/j.eswa.2022.119032
Received 8 June 2021; Received in revised form 7 August 2022; Accepted 10 October 2022
Available online 15 October 2022

theta, alpha, beta, and gamma frequency band powers (BPs) were calculated for the feature space and classified at 99.30 % accuracy with a support vector machine (SVM).

Zhao, Zheng, Zhao, Tu, and Liu (2011) calculated a multivariate autoregressive (MVAR) model to extract EEG features (Zhao et al., 2011). EEG signals were recorded for 150 min from 13 subjects during a driving simulation. The coefficients of the third-order MVAR model were used as EEG features. These features were tested using various classification strategies. Linear principal component analysis and kernel principal component analysis (KPCA) were adopted as feature reduction algorithms, and the radial basis function (RBF) network and SVM were selected for classification. The combination of KPCA and SVM performed best (81.64 %), with 25 features.

Budak, Bajaj, Akbulut, Atila, and Sengur (2019) combined three long short-term memory (LSTM) networks to detect drowsiness (Budak et al., 2019). An expert observer labeled data from the Massachusetts Institute of Technology-Beth Israel Polysomnographic EEG database into awake and multiple sleep stages, and EEG signal and data with "alert" and "sleep stage 1" were used. We assumed that "sleep stage 1" was drowsy. The following input data were fed for each LSTM network: the histogram of zero-crossing rate, energy, spectral entropy, an instantaneous frequency for the first LSTM network, output data of AlexNet and VGG16 with spectrogram image input for the second LSTM network, and coefficients of the tunable-Q wavelet transformed for the third LSTM network. Each LSTM network contained a fully-connected node for binary classification, and the final classification conducted by majority voting was 94.31 %.

Before this study, we implemented a drowsiness detection system using LSTM (Lee, Choi, & An, 2021). EEG signals were acquired using a button response to a randomly given auditory stimulus. The acquired EEG signal was segmented with a 4-second window and a second shift and labeled with one of three cognitive states, including "sleep, awakeness, and drowsiness". The cognitive states of "awakeness" and "sleep" were labeled according to the subject's response. As a transition of cognitive states, "drowsiness" was labeled when "awakeness" and "sleep" occurred sequentially. The labeled segments were classified using the LSTM neural network model, resulting in 81.1 % with every channel and 79.8 % with three channels in the right hemisphere of the premotor cortex.

However, the previous study had two limitations. First, the window length of 4 sec was not optimal. The segmented data were labeled according to the last timestamp of the data. Thus, the window length represents the length of past data for classification. If the window length is too short, insufficient data are fed into a classifier. If the window is too long, a large proportion of the data will not contribute to classification and will interfere in the worst case. Therefore, the optimal window length must be determined. Second, additional classification methods should be tested to verify the performance of the proposed neural network. The performance of drowsiness detection systems depends on the experimental process and labeling method. Balandong, Ahmad, Mohamad Saad, and Malik (2018) compared several studies that implement drowsiness detection systems, and the accuracies of drowsiness detection systems using SVM ranged from 81.64 % to 98.00 % (Balandong et al., 2018). Despite the differences in kernel type or other hyperparameters of the classifier, the influence of the difference between the experimental process and the labeling method should be considered. Therefore, examining the proposed neural network model's performance is necessary. An identical dataset was tested for convolutional neural network (CNN), LSTM, feature-based SVM, and linear discriminant analysis (LDA). This study uses classification results from LDA, SVM, CNN, LSTM, and LSTM-CNN models, and six levels of window length (500, 1000, 2000, 4000, 8000, and 16000 ms) were compared.

The remainder of this paper is organized as follows. Section 2 describes the methods used to acquire the data. Section 3 explains the signal processing method and structure that classifies the collected data. The experimental results are discussed in Section 4, and the conclusions and direction for future work are provided in Section 5.

2. Materials

2.1. Data acquisition

2.1.1. Experimental procedure

This study aimed to detect and classify the moments of transition from an awake to a sleep state. Therefore, continuous measurement of the subject's consciousness states (awake, sleep, drowsiness) was required. This study assumes that the physiological response to physical stimuli such as sound and vibration will change according to a consciousness state. The experimental procedure was designed to mark the three conscious states by recording the corresponding physical response (button press) when auditory stimuli were repeatedly presented.

The participants rested in an armchair in a dark room with their eyes covered with a black blindfold to induce sleep naturally. The participants wore stereo earphones. Two buttons were closed to the index fingers of their left and right hands (Fig. 1). The earphones and buttons were connected to a computer through a sound card and a digital input/output interface (USB-6211, NATIONAL INSTRUMENTS Corp.). Eighteen EEG electrodes were used for EEG signal measurement. The EEG electrodes are located in the frontal lobe (Fig. 1). This study focused on the frontal lobe to measure consciousness-related brain activity. The frontal lobe plays an essential role in the control of consciousness (Miller & Cohen, 2003). It has been reported that frontal theta activity reflects mental fatigue (Wascher et al., 2014). Channels on the occipital lobe were excluded because the head was reclined on a chair to investigate the natural transition from awake to sleep. The accurate coordinates of the EEG electrodes measured by the 3D digitizer were converted to the MNI coordinate system using the standard MNI coordinate values provided by the NIRS-SPM MATLAB software package (Ye, Tak, Jang, Jung, & Jang, 2009). The MNI coordinates of the EEG electrodes were projected onto the cortical surface marked with the Brodmann area and displayed in Fig. 1. At the start of the experiment, auditory stimuli were provided through the left or right side of the earphone, and subjects were instructed to respond by pushing the button on the corresponding side. Auditory stimuli were given at random intervals of 12 to 22 s; the side of the auditory stimuli was randomly selected. The experiment ended when the subject fell into a deep sleep and later awakened. The average experimental time was 74.4 min. The shortest experiment took 43.9 minutes, and the most extended experiment took 102.5 minutes.

2.1.2. Ethic statement

This study was approved by the Institutional Review Board (IRB) of DGIST (DGIST-171011-HR-035–01). All subjects understood the purpose of the study, and informed consent was accepted from all individual participants. The methods used in this study were performed under the guidelines approved by the mentioned IRB.

2.1.3. Participants

Twenty subjects participated in this experiment; however, the dataset from one participant was discarded because of a technical problem. Therefore, a dataset acquired from 19 participants was used in this study. Among the 19 subjects, 13 were men, six were women, and the average age was 27.6. The participants were asked about their medication status and sleep conditions; two participants reported taking painkillers or cold medicines in the previous week, and none had sleep disorders or drank alcohol or caffeinated drinks.

2.1.4. Apparatus

The Biosemi ActiveTwo system was used to acquire data. It has 18 EEG electrodes, and 8-bit event channels were digitized at a 512 Hz sampling rate. The location of the EEG channels was mapped using a Fastrak 3D digitizer. At the start and end of the experiment, a negative

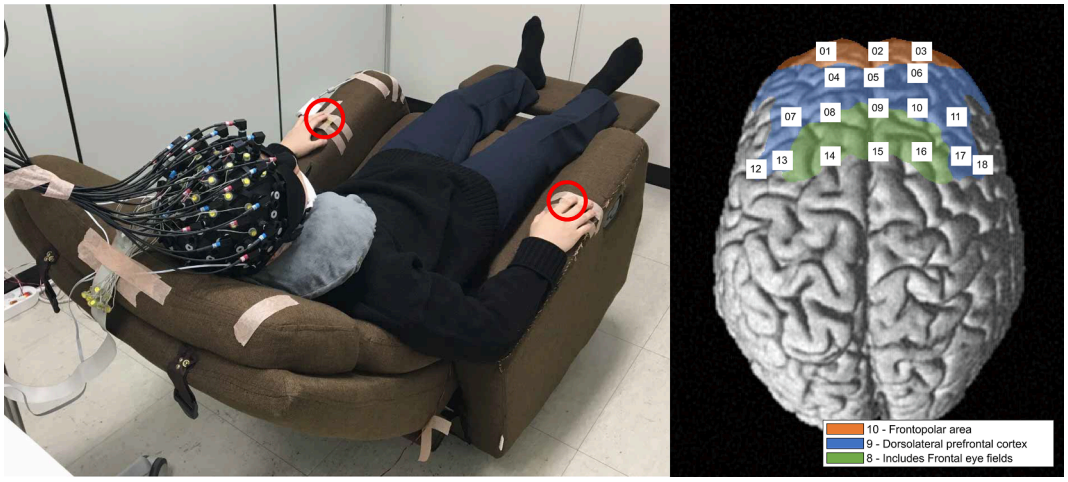

Fig. 1. Experimental setup. *left.* A subject is seated in an armchair wearing a blindfold and earphones. Buttons are installed on the left and right armrests of the armchair (red circles). *right.* Brodmann areas and the location of the EEG channel are shown. Brodmann areas in which the EEG channels are distributed are colored according to their numbers. (The reader can be referred to this article's web version to interpret the color references in this figure.)

edge occurred at the event channel at the start of the left/right audio stimuli and left/right button responses. The timestamps of the audio stimulus and the button response time were collected based on the event channel data.

2.2. Preprocessing

The EEG signals were filtered using a finite impulse response band-pass filter with a frequency band from 0.1 to 50 Hz. The EEG signal was segmented with six different levels of window length. The shift time of the window is fixed to 1 s. The window length was determined as two seconds (1/2, 1, 2, 4, 8, and 16 s). The segmented data were labeled according to auditory stimuli and button responses.

2.3. Ground truth definition

For each auditory stimulus, a single-button or no-button response was assumed. Only the first button response was recorded if multiple button responses were recorded. Thus, the first button response was detected within 10 s from the stimulus, represented by "R." If no button response was detected within 10 s from the stimulus, that stimulus was represented by "N." Then, a series of auditory stimuli and button responses were represented with a character string of "R" and "N." From that character string, the patterns that contained three or more "R" were defined as awake, and three or more "N" were defined as sleep. When the "RRRNNN" pattern occurred, the transition range between "RN" was defined as drowsiness. The number of repetitions of "R" and "N" was determined considering mistakes by the subjects and the number of labeled data. If the number of repetitions is too large, the number of labeled data is too small for training. Conversely, if the number of repetitions is too tiny, mistakenly missed auditory stimulus will be counted as sleep. Table 1 presents the maximum number of repetitions selected in the trainable number range of labeled data.

After defining the consciousness state of each time range, the segmented data were labeled according to the time range to which the last timestamp of the segmented data belongs. As only the last timestamp of the segmented data was used for the labeling process, the segmented data used for classification could be regarded as past data.

The data would be labeled as awake or sleep when a subject repeatedly responded or did not respond to auditory stimulus n times. The data would be labeled as drowsiness when the subject responded n time and not responded n times in succession. This table shows the total time of awake, sleep, and drowsiness data labeled according to repetition number n.

This table shows the time distribution of labeled data for each subject when the repetition number n is 3. When the subject fails to press the button more than three times, he/she is considered asleep. If the subject has never been asleep, the sleep and drowsiness times are zero.

3. Methods

3.1. Data manipulation

The raw data collected from the EEG device included motion artifacts when subjects tossed and turned. Motion artifacts were manually marked based on the accelerometer and gyroscope signals. In the subsequent analysis, windows containing motion artifacts were excluded.

The labeled data were allocated to five groups to measure classifier performance using the cross-validation method. In the process of grouping the data, each label was considered separately to prevent deterioration of the data imbalance by the probabilistic factor. A grouped dataset was created once and applied equally to multiple classifiers under the same conditions.

Table 2 shows the time distribution of labeled data for each subject. It shows a considerable difference in the data between the three classes for each subject. In some subjects, there is no sleep and drowsiness data. Therefore, cross-validation was conducted in a subject-independent manner.

Finally, since the sampling rate is 512 Hz and the number of channels is 18, the shape of the input vector is (512w, 18), where w is the window

Table 1
Total time of labeled data for the number of repetitions.

	Time of labeled data for the number of repetitions (min)								
Repetition (n)	2	3	4	5	6	7	8	9	10
Awake	1159.70	1114.93	1081.36	1054.57	962.81	882.64	842.63	829.35	808.69
Sleep	342.26	324.12	287.45	266.04	254.54	247.31	239.22	234.11	228.81
Drowsiness	77.86	50.00	24.74	13.18	13.18	8.28	5.16	3.53	3.53

Table 2
Time distribution of labeled data for each subject.

	Time of labeled data (min)		
Subject ID	Awake	Sleep	Drowsiness
1	52.15	29.12	4.83
2	26.39	26.17	4.98
3	84.29	16.15	6.87
4	89.58	19.10	5.22
5	117.15	6.49	1.45
6	36.95	50.22	8.46
7	40.85	47.96	1.49
8	76.89	10.96	1.61
9	109.93	0	0
10	21.12	33.24	0
11	21.65	19.58	3.23
12	60.73	21.60	3.16
13	23.72	18.06	1.69
14	53.23	7.51	1.99
15	58.33	0	0
16	47.87	0	0
17	56.61	12.09	5.03
18	63.23	5.88	0
19	74.27	0	0

length in seconds.

3.2. Feature-based machine learning models

3.2.1. Feature extraction

The use of EEG for drowsiness detection requires proper preprocessing and a feature extraction process to reveal the intended information from signals containing noise caused by motion artifacts (Symeonidou, Nordin, Hairston, & Ferris, 2018), electromagnetic field interferences (Anderer et al., 1999), or other mechanical defects(Urigüen & Garcia-Zapirain, 2015). This study performed feature extraction using sub-band wave energy and an MVAR.

Generally, EEG signals are divided into several sub-bands: delta (0.5–4 Hz), theta (4–7 Hz), alpha (7–15 Hz), beta (15–30 Hz), and gamma (30–50 Hz). It has been reported that EEG sub-bands contain information about consciousness states. The EEG alpha, theta, and gamma BP change according to sleep phase changes, and this phenomenon is used to determine the sleep stage (Boostani, Karimzadeh, & Nami, 2017). When an anesthetic artificially induced the change in consciousness, the EEG BP showed a significant difference (Maksimow et al., 2006). Assuming that these sub-bands include information related to the drowsiness level, the wave energy of the sub-bands was used as the EEG feature. For data of a specified length, the signal of each channel was separated into the delta, theta, alpha, beta, and gamma bands by an infinite impulse response Butterworth band-pass filter. The wave energy was defined as the sum of the squared magnitudes of each signal component. The BP was calculated for 18 channels and five EEG sub-bands, and 90 features were used for machine learning.

The MVAR model is an extended form of the univariate autoregressive model (Anderson, Stolz, & Shamsunder, 1998). The use of MVAR is intended for research on the synchronization of brain structures, degree of coupling between channels, estimation of phase delays, and eventually, the direction of spreading brain activity (Franaszczuk, Blinowska, & Kowalczyk, 1985; Neumaier & Schneider, 2001). The EEG features were extracted using Zhao et al.'s approach (Zhao et al., 2011). For the given EEG signal v, the coefficient matrix A and constant matrix C were determined by minimizing the sum of squares of ERR as follows:

$$ERR_n = v_n - \left(\sum_{i=1}^{n} A_i v_{n-i} + C \right),$$

where the order of MVAR is p, the dimensionality of the signal is m, the shape of the coefficient matrix A and constant matrix C is m by m, and the number of estimated coefficient matrices is p. In this study, concatenation of the flattened coefficient matrix A was used as the MVAR feature, and the number of features was pm^2. The EEG channels were merged into five groups according to the regions of the channels to create a noise-tolerant feature set. Each group consists of several channels, such as Group 1: CH1, CH2, and CH3; Group 2: CH4, CH7, and CH8; Group 3: CH6, CH10, and CH11; Group 4: CH12, CH13, and CH14, and Group 5: CH16, CH17, and CH18. The numbers corresponding to the EEG channel numbers of the brain montage are shown on the right of Fig. 1. The order of the MVAR was set to three; thus, the number of MVAR features was 75.

3.2.2. Model implementation

The two feature sets extracted above were adopted for LDA and SVM (Fig. 2). Therefore, four combinations of feature sets and classifiers were tested. The implementation of LDA is LIBLINEAR, and that of SVM is LIBSVM. A radial basis kernel was used for SVM classification, the gamma value was 1/3, and the cost was 1. The weight of each label was set to the inverse of the data imbalance ratio for compensating the data imbalance.

3.3. Deep neural network models

3.3.1. Network architecture

In this study, a CNN and LSTM were adopted for classification. In the proposed model, three output layers are present. In Fig. 3, output layers 1, 2, and 3 classify the labels based on the CNN, LSTM, and LSTM-CNN models. The structure of the CNN-based model refers to the residual term proposed by ResNet (He, Zhang, Ren, & Sun, 2016). In the CNN-based model, two one-dimensional convolution layers, a batch normalization layer, and a rectified linear unit activation layer comprised one residual block. Six blocks were stacked in the proposed network. For each block, the number of filters gradually increased, and the signal length gradually decreased. The kernel size was 9 for the first two residual blocks and seven for the others. The stride of all blocks was 2, except the last block's stride was set so that the signal length of the output vector was 1. For the last block, the number of filters was set to 2048, and the signal length was 1.

The LSTM-based model is expected to detect periodic brain wave patterns. The structure of the LSTM-based model contains four serially stacked LSTM layers. After each LSTM layer, a dropout layer was adopted, and the batch was normalized after four LSTM layers to alleviate the overfitting problem and improve performance (Ioffe & Szegedy, 2015; Srivastava, Hinton, Krizhevsky, & Salakhutdinov, 2014). The number of nodes in each LSTM layer is 32, and the dropout probability is 0.3. These parameters were determined experimentally by the size and complexity of the EEG data.

3.3.2. Model implementation and training

For classification, three neural network models were used in this study: a CNN-based model (CNN model), an LSTM-based model (LSTM model), and a model with the concatenation of each output of the CNN and LSTM models (LSTM-CNN model). In the CNN- and LSTM-based models, the fully connected layer with a SoftMax activation function and three units were attached directly to the flattened layer for the three-class classification. For the LSTM-CNN model, flattened layers of two branches were concatenated and then connected to a fully connected layer with a Softmax activation function and three units. By setting the shapes of the flattened vectors of the two models to be identical, the LSTM-CNN model was intended to balance the possibility of the contribution of each model.

The neural network models were optimized to minimize the categorical cross-entropy loss function using the Adam optimizer, with a learning rate of 0.001. The maximum number of epochs was set to 200 but stopped before reaching the maximum number of epochs if the loss was saturated. When the validation loss was not improved even for ten epochs, the training was stopped, and the best model was selected. The

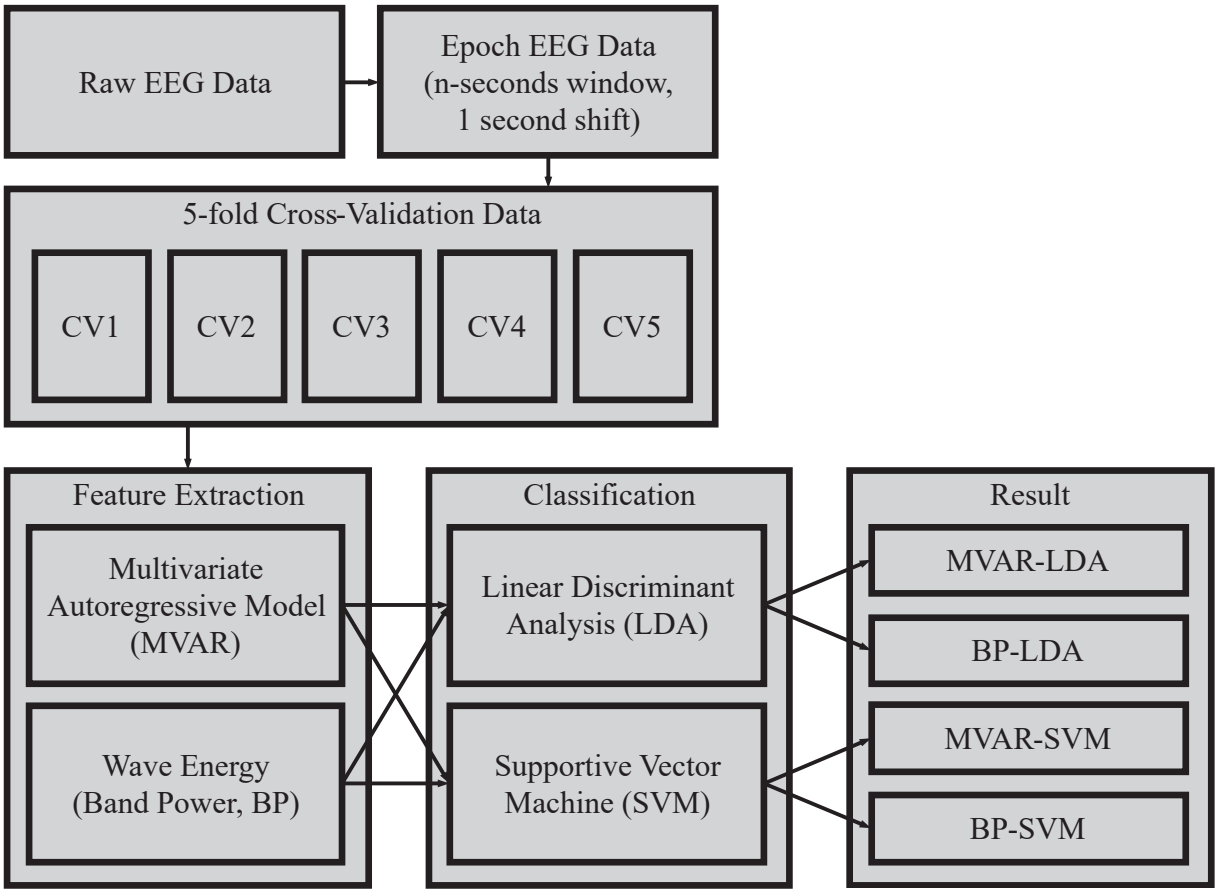

Fig. 2. Flowchart of feature-based classification. A brief illustration of how raw EEG data processed to be fed into classifiers.

weights of the CNN and LSTM models were initialized using the method presented by Glorot (Glorot & Bengio, 2010). The LSTM-CNN model was set to the initial weights from those trained results of CNN and LSTM models. 10 % of the training data were assigned as the validation data to determine the early stop point. Therefore, because the training and test data were determined by 5-fold cross-validation, the ratio between the training, validation, and test data was 3.9:0.1:1.

The confusion matrices for the 3-classes classification results with 500 msec window length are shown in percent. Each row represents the ground truth, and each column represents a predicted label. For example, the second column of the first row is the ratio of samples labeled as "awake" and predicted as "sleep" to the number of samples with the ground truth of awake. Refer to the Supplementary material for more detailed information.

4. Results and discussions

4.1. Response time

To label drowsiness data, Lin et al. (2014) proposed an approach that measures the response time for a randomly given event (Lin et al., 2014). However, since the relationship between response time and drowsiness level was not revealed, we investigated the correlation between response time and the period before falling asleep. Our basic assumption is that the level of drowsiness increases as the point of falling asleep gets closer. Based on this assumption, the correlation between response time and the period before falling asleep was examined to reveal the relationship between response time and drowsiness level. Fig. 4 shows the distribution of responses where the x-axis is the period before falling asleep in minutes and the y-axis is the response time in seconds. In this case, "falling asleep" is defined as the first timestamp of three consecutive auditory stimuli without any response. As expected, the response time in seconds and the period before falling asleep in minutes had a negative correlation coefficient (−0.27), and the p-value was less than 0.01. However, in the case of responses in the range of 0 to 30 s before falling asleep, the ratio of fast responses (response time < 2 s) was still high (62.2 %). Therefore, despite the significant negative correlation between response time and the period before falling asleep, using response time as the ground truth of the drowsiness state is not recommended. Therefore, our proposed approach defines drowsiness as a transitional state from awake to sleep. It is expected that this approach will be able to cope with the problems mentioned above.

4.2. Three classes classification

The average accuracy and kappa index were investigated for each combination of neural network models and window length with classification problems that classified the three classes: awake, drowsiness, and sleep. The accuracy averaged the accuracies of each of the three classes (Fig. 5) because the number of classes was three, and the data for each class were imbalanced. The feature set of the wave energy of the sub-bands is expressed as BP. The average accuracy was the lowest for each neural network model when a 16-second window length was used, whereas the window length that showed the highest performance was not unified. The CNN-based model showed the highest performance (82.5 %) when the window length was 500 msec, 1000 msec for the LSTM-based model (86.2 %), and 500 msec for the LSTM-CNN model (85.6 %). However, for feature-based classifiers, the overall accuracies were improved with increased window length. The performance of the feature-based classifier was maximized when the window length was 16 seconds. In contrast, neural network models performed worst under the same conditions. However, the neural network models showed better

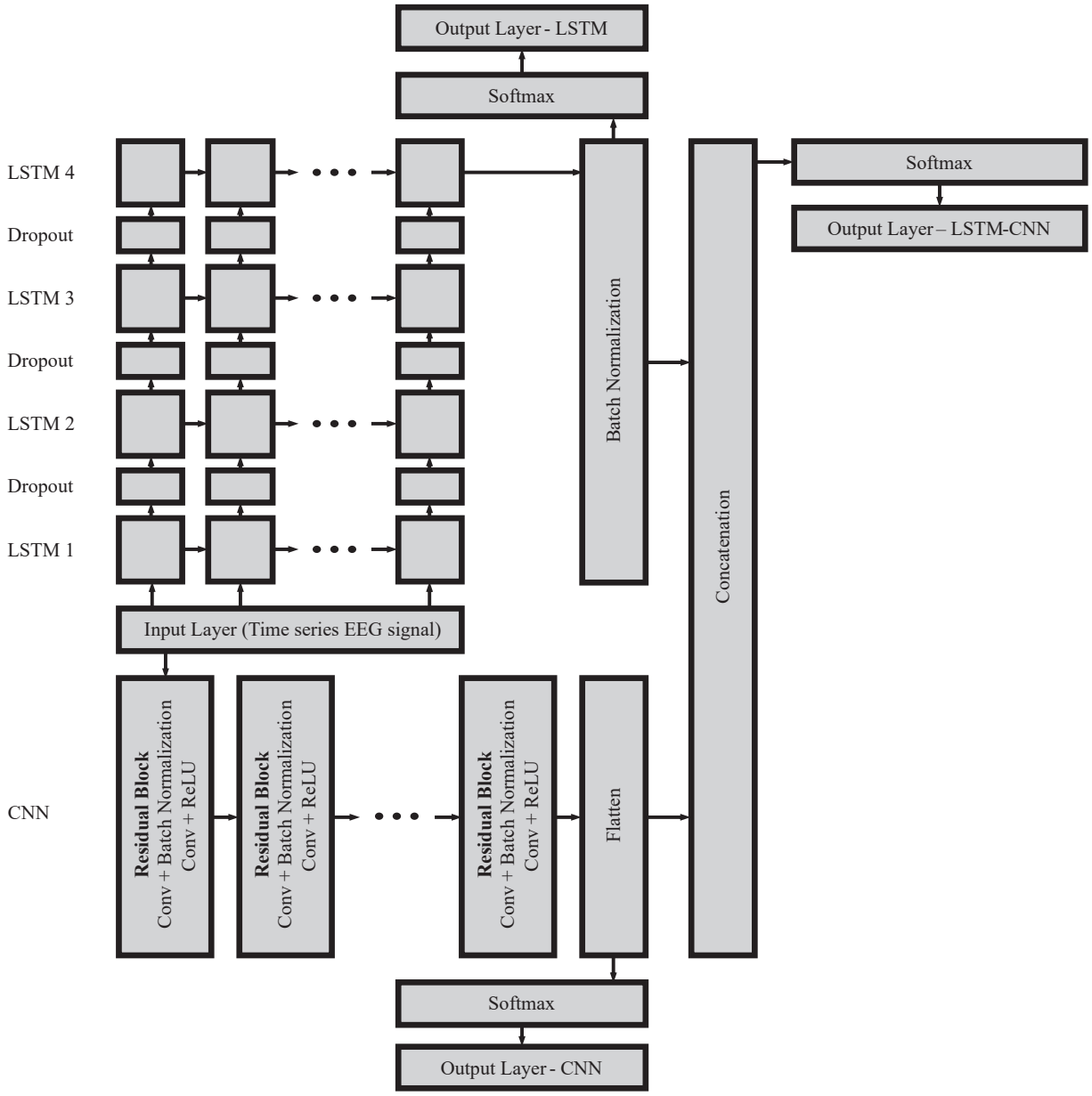

Fig. 3. Structure of neural network models. The CNN path (below the input layer) and LSTM path (above the input layer) are connected from the input layer. The flattened layers of each path are connected to corresponding output layers. The concatenated layer is connected to the output layer of the LSTM-CNN model.

accuracy for every window length than any feature-based classifier.

Confusion matrices were created when the window length was 500-ms (Table 3) to compare the difference between the models. When compared under the shortest window condition, the LSTM-CNN model had the highest performance (85.60 %). To compare proposed models with well-known models, we tested the EEGNet-4,2. The tested model implements EEGNet-4,2 used by Lawhern et al. (2018). In the study of Lawhern et al., EEGNet-4,2 had fewer trainable parameters, so the possibility of overfitting was low, and the performance was similar to other models. The EEGNet-4,2 model accurately classified awake data as awake (94.24 %). However, 41.56 % of sleep data and 73.53 % of drowsiness data were classified as awake, so the overall accuracy was 50.86 %. It seems that the EEGNet-4,2 model was affected by the amount of data. The data in use is extensive in the order of awake, sleep, and drowsiness and is similar to the classification accuracy of EEGNet-4,2. The proposed model parameters were adjusted considering data imbalance, but EEGNet-4,2 was not. As a result, the parameters of the EEGNet-4,2 model did not correctly reflect the sleep and drowsiness data, where the amount of data is relatively small. Confusion matrices in other window conditions are attached in the "Supplementary Material."

As mentioned above, the time position of the label was the most current time position of the input vector. Therefore, an increase in the window size represents an increase in the difference between the average time of the input vector and labeling time. In neural network models, this difference apparently decreases accuracy as the window length increases. In contrast, in the case of feature-based models, the increase in accuracy appears to be affected by the increased amount of information as the number of data increases. A trend similar to neural network models is expected if a sufficient number of features are extracted, and appropriate feature selection techniques are applied.

For a detailed comparison, Cohen's kappa index was calculated (Cohen, 1960). Cohen's kappa index measures the classifier performance in a multiclass classification problem (Carletta, 1996). According to Carletta, the kappa index measures the agreement among a set of coders making category judgments. For the range of the kappa index of reliable classifier, Carletta reported that a kappa index over 0.8

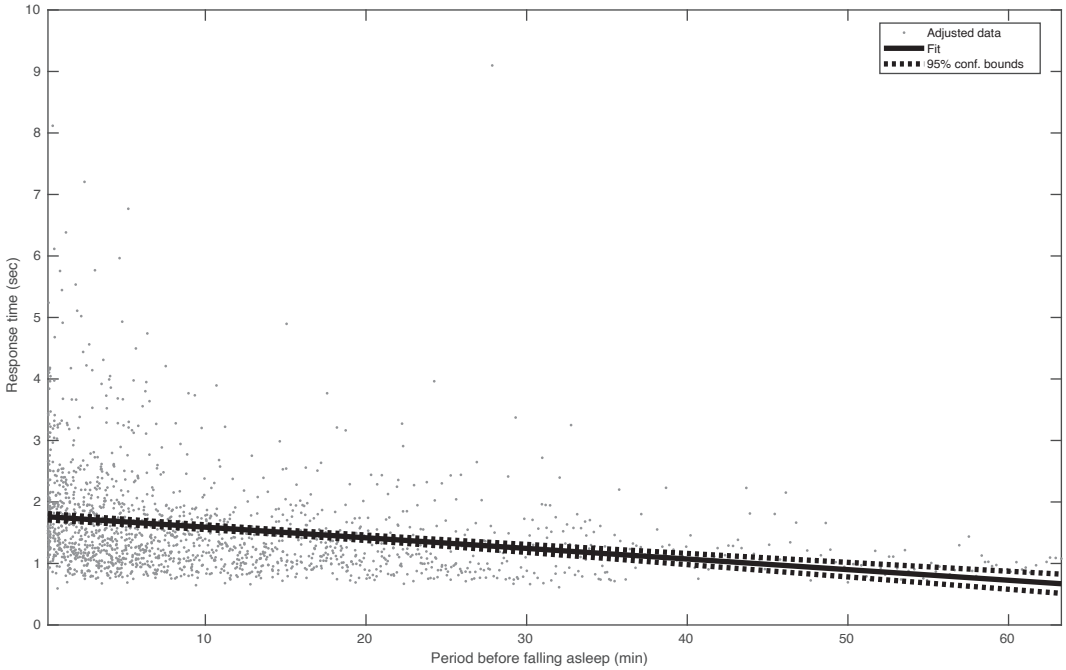

Fig. 4. Response time corresponding to period before falling asleep. Each button response was plotted with response time(y-axis) in seconds and the period before falling asleep(x-axis) in minutes. In the shorter period before falling asleep, subjects tend to react slowly.

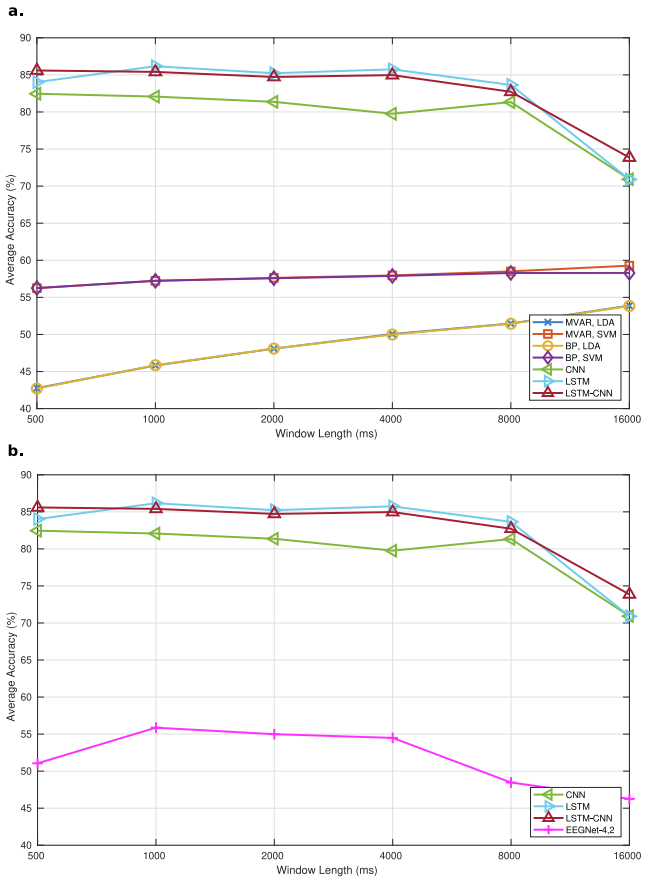

Fig. 5. Average accuracy of three-classes classification. Average accuracy is the mean of classification accuracies of three classes. The accuracies of the proposed models were compared with SVM and LDA (a) and EEGNet-4,2 (b). The x-axis is the window length in milliseconds and is expressed as a log scale.

Table 3
Confusion matrix of 3-classes classification problem.

True Labels \ Predicted Labels		Awake	Sleep	Drowsiness	Accuracy
MVAR, LDA	Awake	99.71 %	0.17 %	0.12 %	42.77 %
	Sleep	71.77 %	28.21 %	0.02 %	
	Drowsiness	98.14 %	1.49 %	0.37 %	
MVAR, SVM	Awake	94.89 %	5.09 %	0.01 %	56.24 %
	Sleep	26.13 %	73.84 %	0.03 %	
	Drowsiness	82.90 %	17.10 %	0.00 %	
BP, LDA	Awake	99.71 %	0.17 %	0.12 %	42.70 %
	Sleep	71.96 %	28.02 %	0.02 %	
	Drowsiness	98.51 %	1.12 %	0.37 %	
BP, SVM	Awake	93.74 %	6.26 %	0.00 %	56.25 %
	Sleep	24.98 %	75.02 %	0.00 %	
	Drowsiness	83.64 %	16.36 %	0.00 %	
CNN	Awake	88.12 %	3.95 %	7.94 %	82.46 %
	Sleep	3.74 %	80.33 %	15.93 %	
	Drowsiness	7.73 %	13.34 %	78.92 %	
LSTM	Awake	86.64 %	5.06 %	8.30 %	84.02 %
	Sleep	2.52 %	85.31 %	12.18 %	
	Drowsiness	5.79 %	14.08 %	80.13 %	
LSTM-CNN	Awake	89.47 %	4.92 %	5.61 %	85.60 %
	Sleep	2.24 %	89.43 %	8.32 %	
	Drowsiness	4.55 %	17.57 %	77.88 %	
EEGNet-4,2	Awake	94.24 %	4.31 %	1.46 %	50.86 %
	Sleep	41.56 %	54.03 %	4.42 %	
	Drowsiness	73.53 %	22.17 %	4.30 %	

represents “good reliability,” and the range between 0.67 and 0.8 represents “tentative conclusions to be drawn.” Similarly, (Landis & Koch, 1977) reported that a kappa statistic between 0.61 and 0.8 represents “substantial” strength of agreement and almost perfect agreement for a kappa statistic over 0.81. In this study, the kappa index of each neural network model was 0.69, 0.72, and 0.76 for CNN, LSTM, and LSTM-CNN, respectively, when averaging the kappa index across the window length from 500 ms to 8000 ms (Fig. 6). With the combination of the LSTM-CNN model and 4000 ms window length, the highest kappa index (0.77) was observed. In the case of feature-based classifiers, the kappa index increased with the window length. When the window length was 16 s, the kappa index of the LDA classifiers was 0.67, while

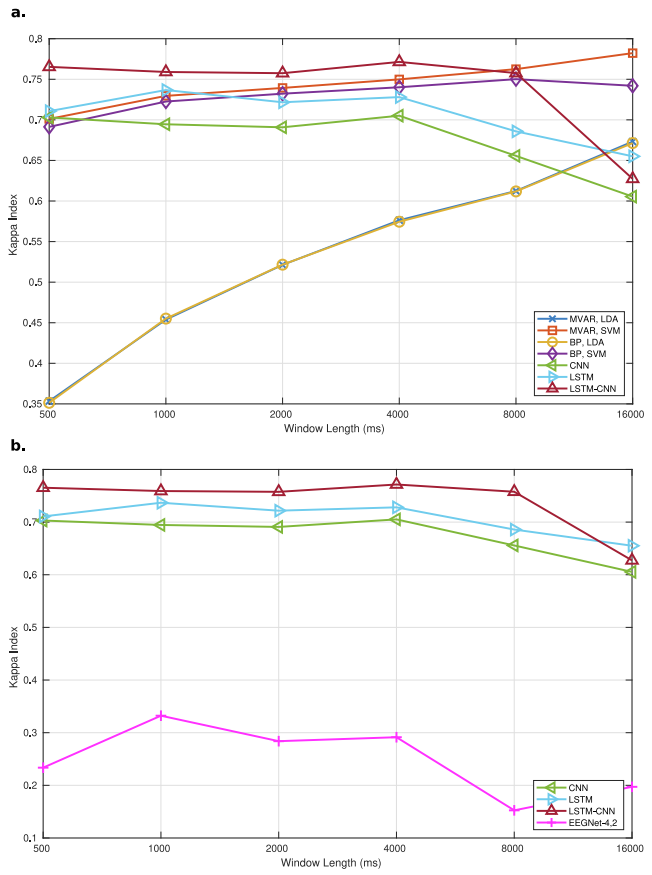

Fig. 6. Kappa index of three-classes classification. The Cohen's kappa index of each classification result. The kappa indexes of the proposed neural network models were compared with the SVM and the LDA (a) and were compared with those of EEGNet-4,2 (b). The x-axis is the window length in milliseconds and is expressed as a log scale.

SVM with MVAR reported 0.78 and SVM with BP reported 0.74. According to Landis, three neural network models resulted from "substantial" strength of agreement, and the LSTM-CNN model showed the best performance among the three models, while the LDA model showed less than "good reliability." The kappa index of the SVM classifier showed "good reliability" for every window length level. In the case of the EEGNet-4,2 model, the kappa index was low due to the poor performance of the drowsiness data.

As expected, the values and patterns of the kappa index were similar to those of average accuracy. However, despite a low average accuracy compared to that of the neural network models, the kappa index of the SVM classifier was similar to that of the neural network models. The imbalanced data and performance caused a mismatch between the averaged accuracy and the kappa index of the SVM classifier. In the case of the SVM classifier, the true positive ratio for drowsiness was 0.03 %, whereas that for awake was 97.36 %.

The performances of BP-LDA and MVAR-LDA were nearly the same, and those of BP-SVM and MVAR-SVM were nearly the same. The difference in accuracy between the same classifiers was less than 1 %, and the difference in the kappa index was below 0.05. It implies that the BP and MVAR features share much information.

Overall, there was little difference in performance when the window length was less than 8 s in the deep-learning models. It is advantageous to have a small window length when the detection accuracy is the same, considering the reaction speed of the drowsiness detection system. Therefore, among the tested methods, the LSTM-CNN model with a window of 500 ms is most suitable for the drowsiness detection system.

4.3. Binary classification

One of the goals of drowsiness detection systems is to prevent problems caused by drowsiness in real life. However, the proposed three-class classification problem does not fit real-world applications. For example, classifying drowsiness and sleep does not require drowsiness detection in real applications. Therefore, we re-labeled the awake state as the normal state and the drowsiness state and the sleep state as the abnormal states and evaluated the binary classification performance. This labeling assumes an environment where it is crucial to detect a decline in cognitive ability. For example, in long-distance driving situations, all abnormal cognitive states, such as drowsiness and sleep, can be fatal to the driver. However, because drowsiness and sleep are different states, it is necessary to test whether they can be combined into one label. When we evaluated the classification performance between each state, the proposed neural network model tended to distinguish awake-drowsiness better than awake-sleep difference in F_1 scores was approximately 0.013 (Fig. 7). To evaluate the performance of the normal-abnormal binary classifier, canonical recall, precision, F_1 score, receiver operating characteristic (ROC) curves were used.

In binary classification, recall and precision can be represented by the number of true positives, false positives, and false negatives. In this study, a true positive represents the correct classification of awake, a false positive represents labeled as not awake but classified as awake, and a false negative represents labeled as awake but classified as not awake. Therefore, recall represents the ratio of the number of correctly classified awakes to the number of real awakes. Furthermore, precision represents the ratio between the number of correctly classified awakes and the number of data classified as awake.

The LSTM and LSTM-CNN models showed the highest recall for every window length level among the three neural network models. The recall value of LSTM was the highest (0.99) with an 8-second window length,

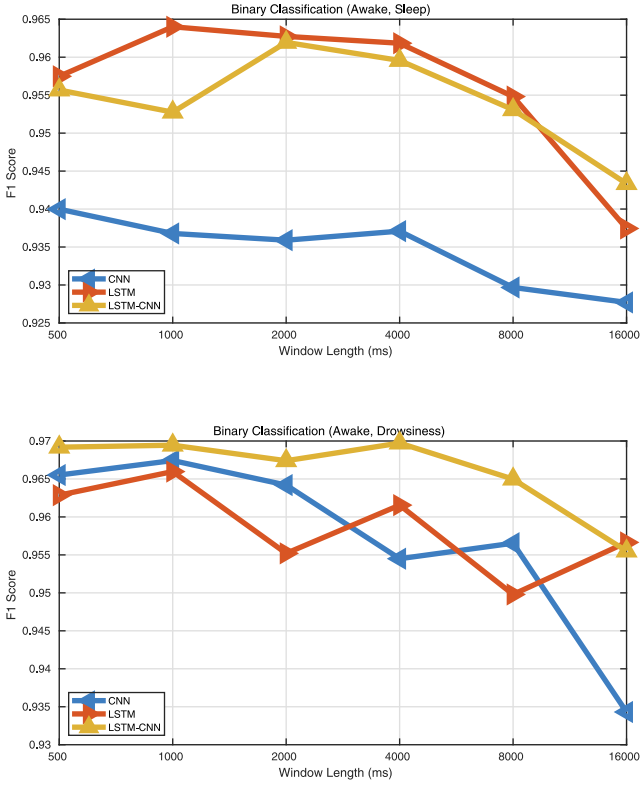

Fig. 7. F_1 scores of awake-drowsiness and awake-sleep binary classification. The F_1 score for awake-drowsiness and awake-sleep binary classification. The x-axis represents the window length. The unit is milliseconds and is expressed on a log scale.

and LSTM-CNN (0.98) showed a similar recall (Fig. 8a). However, in terms of precision, the LSTM-CNN model showed the best performance across every level of window length among the three neural network models (Fig. 8b). Unlike neural network models, feature-based models have low recall and high precision values. Because feature-based models hardly classify data as awake, the ratio of data correctly classified as awake was relatively high. Therefore, the F_1 score and harmonic mean of the precision and recall were measured for comprehensive scoring of binary classification performance. The overall performance of LSTM-CNN was highest for every level of window length except 16 s and showed the best performance (F_1 score = 0.95) with a 4000-ms window length (Fig. 8c). Similar to the three-class classification result, the F_1 scores of the feature-based classifiers increased with the window length. With a 16-second window length and the MVAR feature set, the SVM classifier resulted in the best F_1 score in the binary classification test. In the case of the EEGNet-4,2 model, the recall rate was relatively high (Fig. 8d), but the precision rate (Fig. 8e) and F1 score (Fig. 8f) were low because there were many cases of misclassifying abnormal data as normal data.

The ROC curves and the area under the curve (AUC) were investigated to compare the classifiers in detail. The LSTM-CNN model with a 0.5-second window length saturated rapidly as the false-positive rate increased (Fig. 9) compared to the feature-based classifiers with a 16 sec window length. In other words, the LSTM-CNN model achieved a higher true-positive rate with a low false-positive rate. These properties were quantified using the AUC (Fig. 10). Except for the 16000-ms window length, the LSTM and LSTM-CNN model ranked first and second in AUC values. Though the LSTM and LSTM-CNN classifiers showed no notable change in the window length between 0.5 and 8 s, the LSTM classifier showed the best AUC at 1 s and the LSTM-CNN classifier at 4 s.

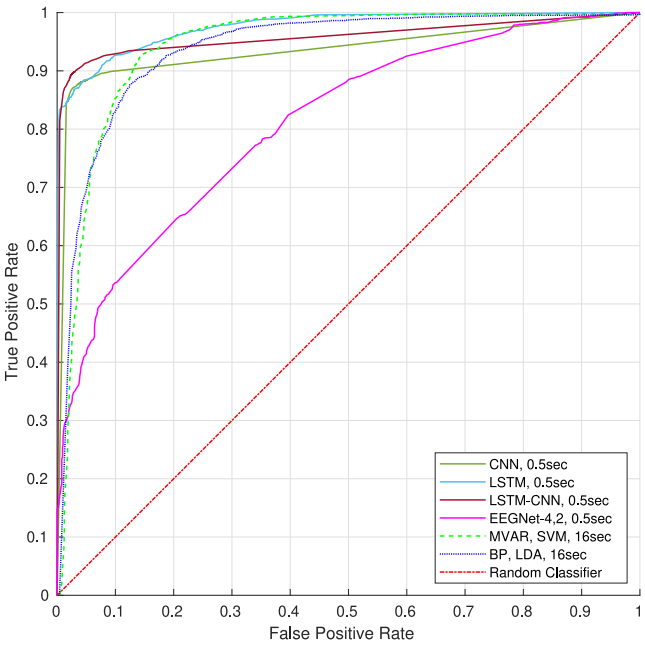

Fig. 9. Receiver operating characteristic curve of binary classification. The red dash-dot line represents the random classifier. The LSTM-CNN, LSTM, and CNN models show the most significant area under the curve, followed by MVAR-SVM, BP-LDA, and EEGNet-4,2. (The reader can be referred to this article's web version to interpret the color references in this figure.)

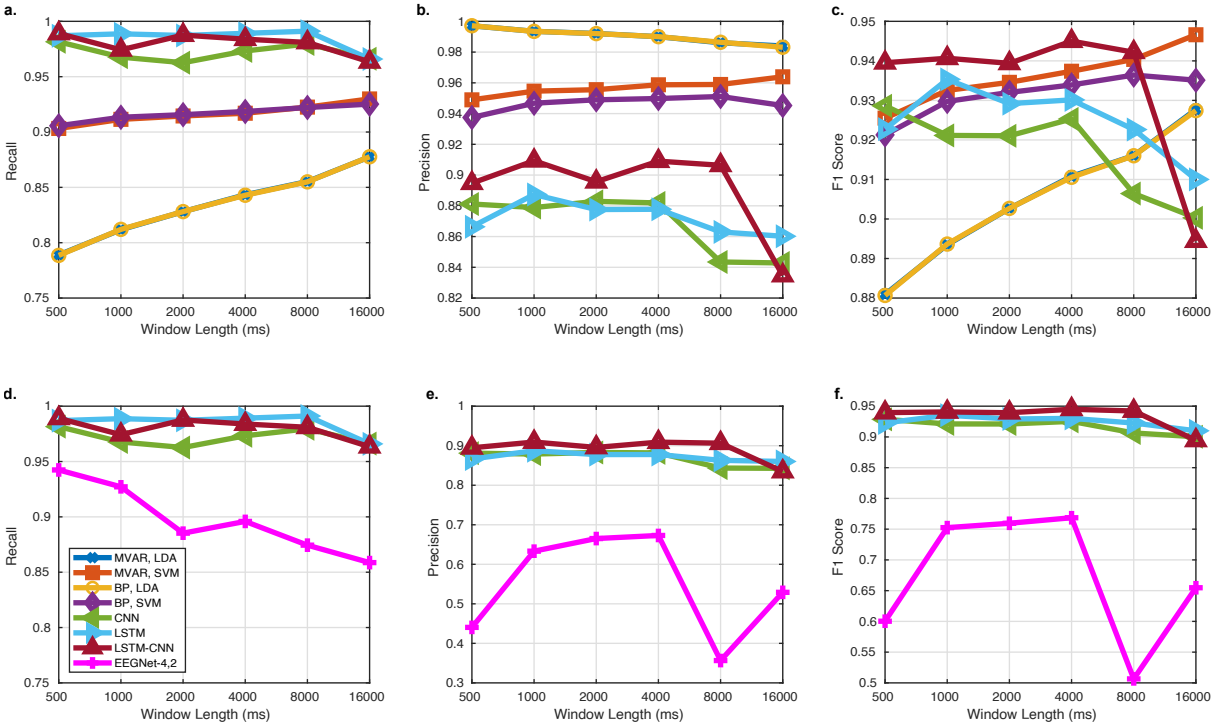

Fig. 8. Recall rate, precision rate, and F_1 score of binary classification. The recall rate was calculated as TP/(TP + FN) and the precision rate was calculated as TP/(TP + FP) where TP is the number of data labeled as awake and classified as awake, FN is the number of data labeled as awake but classified as not awake, and FP is the number of data labeled as not awake but classified as awake. The F_1 score is the harmonic mean of precision and recall. The binary classification performance rates of the proposed neural network models were compared with those of the support vector machine, the linear discriminant analysis model (a, b, c), and were compared with those of EEGNet-4,2 (d, e, f). The x-axis is window length in milliseconds and is expressed as a log scale.

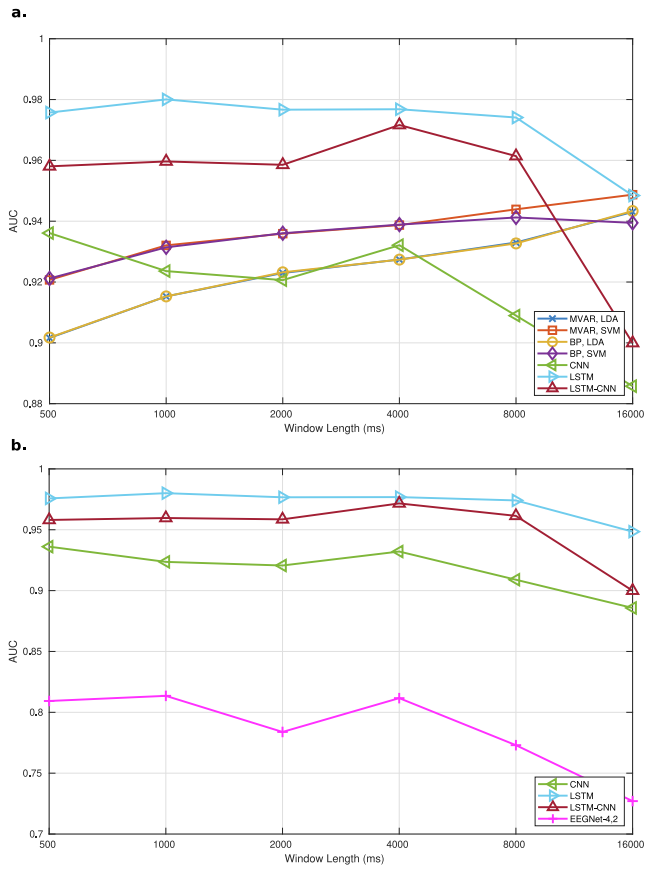

Fig. 10. The area under the curve of binary classification. The area under the curve of every tested model. The area under the curve of the proposed neural network models was compared with that of the support vector machine, the linear discriminant analysis model (a), and was compared with that of EEGNet-4,2 (b). The x-axis is the window length in milliseconds and is expressed as a log scale.

5. Conclusions and future works

This study aimed to implement a drowsiness detection system with a proper neural network model. In this study, the subjects' states of consciousness were automatically collected by timestamping their responses to auditory stimuli. The proposed approach predicts the labeled consciousness states, and the performance according to the input vector size was investigated. The LSTM-CNN hybrid model resulted in an average accuracy of 85.6 % and a kappa index of 0.77 for the three-class classification problem and 0.94 F_1 scores for the binary classification problem with a short input data (500-ms).

A few obstacles should be considered to apply the EEG-based drowsiness detection system in the real world. For instance, to measure the EEG signal while driving, the number of attached electrodes should be reduced to minimize the effect of irritation caused by the measurement. It is crucial to find effective channels for drowsiness detection to minimize EEG channels. An approach to reduce the number of EEG channels according to the statistics or performance should be considered in future studies. In addition, the noise caused by motion artifacts and other ambient factors, such as artificially generated electromagnetic fields, may be obstacles to real-world application. Because the signal-to-noise ratio of the EEG signal is relatively low, its quality is vulnerable to ambient factors. Other signals based on different dynamics should be considered to alleviate the above issue. Functional near-infrared spectroscopy (fNIRS) estimates cortical activation from concentration changes in the oxygenated and deoxygenated hemoglobin of the cerebral cortices. Generally, fNIRS is known to be more robust to motion artifacts than to EEG signals (Lee, Lee, Jin, & An, 2018). Therefore, since the cortical metabolic features extracted from fNIRS signals can expand the feature space based on brain wave patterns obtained from EEG signals, the simultaneous use of EEG and fNIRS deserves careful study.

Because our proposed approach defines the drowsiness as a transitional state from awake to sleep, the amount of drowsiness data is inevitably low. In this study, the percentage of drowsiness data was approximately 3 %. One of the reasons why the feature-based classifier's drowsiness classfication performance was low is related to this extreme data imbalance. Therefore, to alleviate this problem, a data augmentation method should be considered. Generative adversarial networks (GANs) have been widely used to solve data imbalance problems (dos Tanaka, 2019). However, there are several obstacles to generating EEG signals using GANs. First, it isn't easy to check whether the EEG signal generated through the GAN contains information in a human-generated EEG signal. In addition, ensuring that the GAN model of each class consistently generates unique features is not simple. If the EEG signal can be generated through a GAN by solving these obstacles, we expect that the classification performance can be improved.

CRediT authorship contribution statement

Chungho Lee: Software, Formal analysis, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jinung An:** Conceptualization, Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451; Development of BCI based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning), the DGIST R&D Program of the Ministry of Science and ICT of Korea (21-IT-03), and the '2021 Joint Research Project of Institutes of Science and Technology'.

Special thanks to Gihyoun Lee and Sang Hyeon Jin for the experimental set up and data acquisition.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2022.119032.

References

Anderer, P., Roberts, S., Schlögl, A., Gruber, G., Klösch, G., Herrmann, W., … Saletu, B. (1999). Artifact processing in computerized analysis of sleep EEG - A review. *Neuropsychobiology, 40*(3), 150–157. https://doi.org/10.1159/000026613

Anderson, C. W., Stolz, E. A., & Shamsunder, S. (1998). Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. *IEEE Transactions on Biomedical Engineering, 45*(3), 277–286. https://doi.org/10.1109/10.661153

Balandong, R. P., Ahmad, R. F., Mohamad Saad, M. N., & Malik, A. S. (2018). A Review on EEG-Based Automatic Sleepiness Detection Systems for Driver. In *IEEE Access* (pp.

22908–22919). Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/ACCESS.2018.2811723

Boostani, R., Karimzadeh, F., & Nami, M. (2017). A comparative review on sleep stage classification methods in patients and healthy individuals. In *Computer Methods and Programs in Biomedicine* (pp. 77–91). Elsevier Ireland Ltd.. https://doi.org/10.1016/j.cmpb.2016.12.004

Budak, U., Bajaj, V., Akbulut, Y., Atila, O., & Sengur, A. (2019). An effective hybrid model for EEG-based drowsiness detection. *IEEE Sensors Journal, 19*(17), 7624–7631. https://doi.org/10.1109/JSEN.2019.2917850

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *ArXiv Preprint Cmp-Lg/9602004*. doi:10.48550/arXiv.cmp-lg/9602004.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Tanaka, F. H. K. dos S., & Aranha, C. (2019). Data Augmentation Using GANs. *ArXiv Preprint*. doi:10.48550/arXiv.1904.09135.

Franaszczuk, P. J., Blinowska, K. J., & Kowalczyk, M. (1985). The application of parametric multichannel spectral estimates in the study of electrical brain activity. *Biological Cybernetics, 51*(4), 239–247. https://doi.org/10.1007/BF00337149

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 249–256).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Henry, J. C. (2006). Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Fifth Edition. *Neurology, 67*(11), 2092. https://doi.org/10.1212/01.wnl.0000243257.85592.9a

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 448–456).

Johnson, A. L., Brown, K., & Weaver, M. T. (2010). Sleep deprivation and psychomotor performance among night-shift nurses. *AAOHN Journal: Official Journal of the American Association of Occupational Health Nurses, 58*(4), 147–156. https://doi.org/10.3928/08910162-20100316-02

Johnson, A. L., Jung, L., Brown, K. C., Weaver, M. T., & Richards, K. C. (2014). Sleep deprivation and error in nurses who work the night shift. *Journal of Nursing Administration, 44*(1), 17–22. https://doi.org/10.1097/NNA.0000000000000016

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159. https://doi.org/10.2307/2529310

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering, 15*(5), Article 056013. https://doi.org/10.1088/1741-2552/AACE8C

Lee, C., Choi, R. H., & An, J. (2021, February 22). Deep Neural Network for Drowsiness Detection from EEG. In *9th IEEE International Winter Conference on Brain-Computer Interface, BCI 2021*. doi:10.1109/BCI51272.2021.9385368.

Lee, G., Lee, S. H., Jin, S. H., & An, J. (2018). Robust functional near infrared spectroscopy denoising using multiple wavelet shrinkage based on a hemodynamic response model. *Journal of Near Infrared Spectroscopy, 26*(2), 79–86. https://doi.org/10.1177/0967033518757231

Lin, C. T., Chuang, C. H., Huang, C. S., Tsai, S. F., Lu, S. W., Chen, Y. H., & Ko, L. W. (2014). Wireless and wearable EEG system for evaluating driver vigilance. *IEEE Transactions on Biomedical Circuits and Systems, 8*(2), 165–176. https://doi.org/10.1109/TBCAS.2014.2316224

Maksimow, A., Särkelä, M., Långsjö, J. W., Salmi, E., Kaisti, K. K., Yli-Hankala, A., … Jääskeläinen, S. K. (2006). Increase in high frequency EEG activity explains the poor performance of EEG spectral entropy monitor during S-ketamine anesthesia. *Clinical Neurophysiology, 117*(8), 1660–1668. https://doi.org/10.1016/j.clinph.2006.05.011

Miller, E. K., & Cohen, J. D. (2003). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167–202. https://doi.org/10.1146/ANNUREV.NEURO.24.1.167

Neumaier, A., & Schneider, T. (2001). Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software, 27*(1), 27–57. https://doi.org/10.1145/382043.382304

Rodenbeck, A., Binder, R., Geisler, P., Danker-Hopfe, H., Lund, R., Raschke, F., Weeß, H. G., & Schulz, H. (2006). A review of sleep EEG patterns. Part I: A compilation of amended rules for their visual recognition according to Rechtschaffen and Kales. In *Somnologie* (Vol. 10, Issue 4, pp. 159–175). John Wiley & Sons, Ltd. doi:10.1111/j.1439-054X.2006.00101.x.

Slater, J. D. (2008). A definition of drowsiness: One purpose for sleep? *Medical Hypotheses, 71*(5), 641–644. https://doi.org/10.1016/j.mehy.2008.05.035

Smith, L., Folkard, S., & Poole, C. J. M. (1994). Increased injuries on night shift. *The Lancet, 344*(8930), 1137–1139. https://doi.org/10.1016/S0140-6736(94)90636-X

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*.

Symeonidou, E. R., Nordin, A. D., Hairston, W. D., & Ferris, D. P. (2018). Effects of cable sway, electrode surface area, and electrode mass on electroencephalography signal quality during motion. *Sensors (Switzerland), 18*(4), 1073. https://doi.org/10.3390/s18041073

Urigüen, J. A., & Garcia-Zapirain, B. (2015). EEG artifact removal - State-of-the-art and guidelines. *Journal of Neural Engineering, 12*(3), Article 031001. https://doi.org/10.1088/1741-2560/12/3/031001

Wascher, E., Rasch, B., Sänger, J., Hoffmann, S., Schneider, D., Rinkenauer, G., … Gutberlet, I. (2014). Frontal theta activity reflects distinct aspects of mental fatigue. *Biological Psychology, 96*(1), 57–65. https://doi.org/10.1016/J.BIOPSYCHO.2013.11.010

Ye, J. C., Tak, S., Jang, K. E., Jung, J., & Jang, J. (2009). NIRS-SPM: Statistical parametric mapping for near-infrared spectroscopy. *NeuroImage, 44*(2), 428–447. https://doi.org/10.1016/J.NEUROIMAGE.2008.08.036

Yeo, M. V. M., Li, X., Shen, K., & Wilder-Smith, E. P. V. (2009). Can SVM be used for automatic EEG detection of drowsiness during car driving? *Safety Science, 47*(1), 115–124. https://doi.org/10.1016/j.ssci.2008.01.007

Zhao, C., Zheng, C., Zhao, M., Tu, Y., & Liu, J. (2011). Multivariate autoregressive models and kernel learning algorithms for classifying driving mental fatigue based on electroencephalographic. *Expert Systems with Applications, 38*(3), 1859–1865. https://doi.org/10.1016/j.eswa.2010.07.115