A Combinatorial Deep Learning Structure for Precise Depth of Anesthesia Estimation from EEG Signals

Sara Afshar, Reza Boostani, and Saeid Sanei

***Abstract*—Electroencephalography (EEG) is commonly used to measure the depth of anesthesia (DOA) because EEG reflects surgical pain and state of the brain. However, precise and real-time estimation of DOA index for painful surgical operations is challenging due to problems such as postoperative complications and accidental awareness. To tackle these problems, we propose a new combinatorial deep learning structure involving convolutional neural networks (inspired by the inception module), bidirectional long short-term memory, and an attention layer. The proposed model uses the EEG signal to continuously predicts the bispectral index (BIS). It is trained over a large dataset, mostly from those under general anesthesia with few cases receiving sedation/analgesia and spinal anesthesia. Despite the imbalance distribution of BIS values in different levels of anesthesia, our proposed structure achieves convincing root mean square error of 5.59 ± 1.04 and mean absolute error of 4.3 ± 0.87, as well as improvement in area under the curve of 15% on average, which surpasses state-of-the-art DOA estimation methods. The DOA values are also discretized into four levels of anesthesia and the results demonstrate strong inter-subject classification accuracy of 88.7% that outperforms the conventional methods.**

***Index Terms*—Attention mechanism, convolutional neural network, deep learning, depth of anesthesia, electroencephalogram, long short-term memory.**

I. Introduction

BRAIN is the main human organ affected by anesthetic drugs during surgery [1]. Specialists usually measure the depth of anesthesia (DOA) during surgery by monitoring physiological measures such as blood pressure, heart rate, and blood oxygen level. These clinical parameters, however, vary in patients and surgeries and depend on age, body weight, gender, and medical history. Therefore, assessing the consciousness level by observing the vital signs of patients, as primary inputs, is quite challenging. As a supplement monitoring, the effect of anesthetic agents by processing the online electroencephalogram (EEG) during total intravenous anesthesia is recommended to reduce the incidence of awareness [2], [3]. This is done by bispectral index (BIS).

BIS is the commercial EEG-based monitor. The BIS algorithm, however, is not publicly available since it is still subject to patent access limitation. BIS monitor's electrodes are molded on the forehead to capture raw EEG signals and generate DOA index in the range of 0-100. BIS is recognized as the most commonly used EEG-based DOA estimator.

Machine learning techniques, including deep learning methods have shown impressive results in processing complex data [4]–[6]. This study aims at mimicking the BIS index in an online manner by developing a new hybrid deep learning structure. The network receives raw EEG and generates the DOA index without eliciting any hand-crafted EEG feature. Using large patient datasets we show the superior performance of our deep neural network (DNN) to other DNN structures as well as feature-based classification systems. Utilizing DNN in the field of anesthesia is relatively new [7], especially for forecasting continuous BIS score in a real-time manner. The main goal of our work is to estimate the BIS index by a proper regression model, based on the combination of deep learning methods.

A wide range of features in diverse domains have been proposed for DOA assessment over past years. For example, wavelet coefficient energy entropy and wavelet weighted median frequency are introduced in order to obtain high correlated indexes with BIS [8], [9]. Burst suppression is a key feature to detect deep level of anesthesia. Sarkela *et al.* [10] developed a method for detection and segmentation of burst suppression automatically based on Nonlinear Energy Operator. Sample and permutation entropy features are commonly applied in several studies [11]–[13]. Instantaneous frequency (IF) is introduced as a highly positive correlated feature to the BIS score [14]. Short-time Fourier transform is also applied to EEG with the aim of IF estimation. In addition, Kalman filter is executed to predict cutoff frequencies of the band-pass filter through consecutive windows in order to better IF estimation [14]. For decision making, different types of classifiers and regressors such as, artificial neural network [11], adaptive neuro-fuzzy inference system with linguistic hedges [13] and random forest [12] have been utilized. The datasets in anesthesia research field, however, are mostly private. DOA labels in datasets are assessed by anesthesiologists [15] or extracted from automatic EEG monitoring systems [7], [12].

Lee *et al.* [7] propose a deep learning model, which is trained over 231 subjects that receive total intravenous anesthesia during surgeries. The inputs of the network are infusion histories of propofol and remifentanil besides subject's characteristics. It predicts the BIS score in terms of continuous values. Their developed method outperforms the pharmacokinetic–pharmacodynamic model's results. Liu *et al.* [15] determine the spectrogram of EEGs for 50 subjects as the

S. Afshar (corresponding author) and R. Boostani are with the School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran (e-mail: S. Afshar: (sara.afshar191@gmail.com); R. Boostani: (boostani@shirazu.ac.ir)).

S. Sanei is with the School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham, U.K. (e-mail: saeid.sanei@ntu.ac.uk).

inputs of convolutional neural networks (CNNs) like CifarNet, AlexNet, and VGGNet. They rely on converting EEG signals to 2D images, which is computationally intensive for big datasets. They reach 93.5% classification accuracy over just three levels of anesthesia, while it is more common to consider four anesthetized states before any classification performance can be reported [11]–[13]. Lee *et al.* [16] build a decision tree with four subparameters driven from the BIS monitor with the aim of classifying BIS ranges. Thereafter, multiple regression models are used for BIS values calculation. The model is being trained on a big dataset with size of 5,427 subjects. This method is less generalized and more sensitive to noise in comparison with our end-to-end deep learning model.

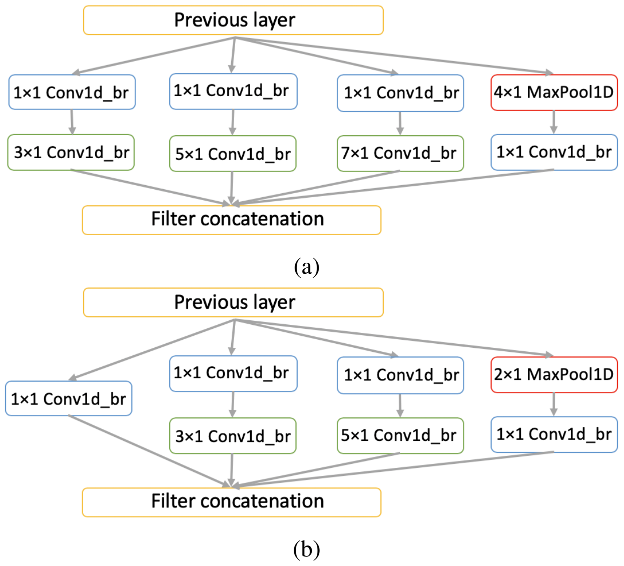
The rest of the paper is organized as follows: Section II describes the components of proposed architecture. The characteristics of the deployed dataset and the description of preprocessing phase are expressed in Section III. Experimental results and discussion are presented in Section IV and Section V, respectively. Finally, Section VI concludes the paper.

II. Methodology

A. *Proposed method*

Our proposed method is a combination of three DNNs. We select CNN due to its outstanding performance in various fields. In addition, CNN is capable of hierarchical feature learning from unstructured data and its flexibility allows it to be customized for specific purposes. From another perspective, recurrent neural network (RNN) is particularly developed to capture temporal features and memorize information through the time in sequential data. Since EEG is a time series with complex underlying information, combination of these two DNNs can fulfill our purpose. The proposed model is formed by concatenating CNN and bidirectional long short-term memory (Bi-LSTM) in conjunction with an attention layer. Next, we provide more details about each module.

*1) **CNN/Inception**:* The main body of our model is formed by CNN and it is inspired by the inception module in Inception V1 [17]. The inception module is the result of convolutional layers concatenated with different kernel sizes and a max pooling layer. Merging the pooling layer to the convolutional layers increases the number of stage's outputs successively. Consequently, the computational complexity is drastically grown. Therefore, a 1×1 convolution filter is added before 3, 5 convolution layers, as described in Fig. 1b. It helps to reduce the computational complexity and dimension of filters [17]. One of the advantages of inception network is that a DNN becomes a bit wider than deeper. Using a very deep model leads to encounter with vanishing gradient problem. CNN part of the proposed method is formed by stacking of inception modules as illustrated in Fig. 1. Employing the inception module instead of vanilla plain CNN [17] in the network stem leads to better results. Also, we substitute each simple convolution layer with a convolution layer consisting of regularization and batch normalization [18]. A newer version of inception has a large number of parameters which may over-fit the classifier for our dataset. Therefore, we prefer to use the aforementioned blocks.



(a)

(b)

Fig. 1: Inception blocks in the proposed method. (a) inception module with dimension reduction. 1D convolution with batch normalization and regularization (br), (Incp1). (b) Another applied module, (Incp2).

*2) **Bi-LSTM**:* In practice, an early version of RNN neither handles the long-term dependency nor addresses the issue of vanishing gradient [19]. To overcome these drawbacks, the application of long short-term memory (LSTM) [20] is proposed, which is made up of memory cells, an input gate, an output gate and a forgetting gate which keeps tracking of dependencies between the elements of input sequence and remembering previous state. LSTM takes the previous states and inputs into account. Bidirectional [21] wrapper help the model be trained over both forward and backward directions. It utilizes the future information, additionally, and merges the results of the two sequences of past-to-future and future-to-past. In Fig. 2 $\{\ldots, \overrightarrow{h}_{t-1}, \overrightarrow{h}_{t}, \overrightarrow{h}_{t+1}, \ldots\}$ and $\{\ldots, \overleftarrow{s}_{t+1}, \overleftarrow{s}_{t}, \overleftarrow{s}_{t-1}, \ldots\}$ are, respectively, the hidden forward and backward LSTM layer state sequences.

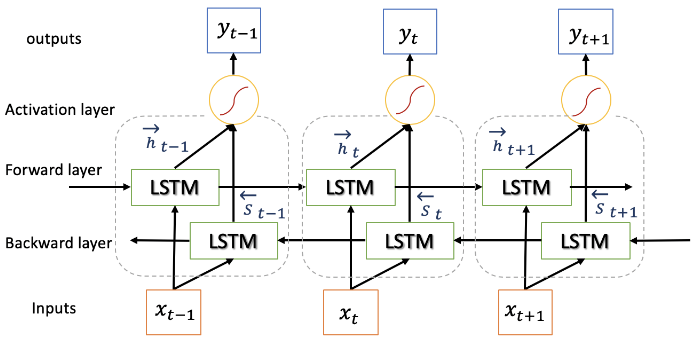


Fig. 2: Unfolded Bi-LSTM in time for three time steps, dotted boxes encompass hidden layers of forward and backward layers.

*3) **Attention layer**:* The attention mechanism is proposed to help memorize long source sequences. Therefore, we do not have any concerns about the forgetting problem. We have implemented a simple form of additive attention [22] which

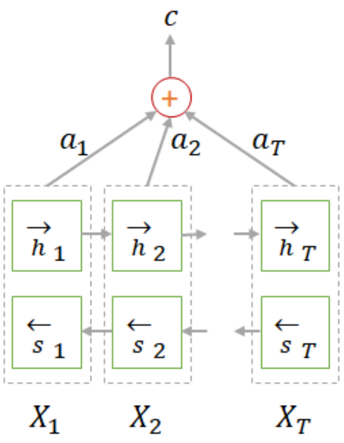


Fig. 3: Illustration of the attention layer. $\{X_1, X_2, ..., X_T\}$ is an input sequence. $\overrightarrow{h_j}$ and $\overleftarrow{s_j}$ are Bi-LSTM hidden states. a_j and c are alignment model and contex vector, respectively. c is the weighted sum of the encoder hidden states.

is more similar to feed-forward attention model [23]. In fact, our implantation does not have any decoder (Fig. 3). Here we show the differences; the attention layer consists of:

- **Alignment model/score e_j:**

$$e_{ij} = a(s_{i-1}, h_j) \;\; => \;\; e_j = a(h_j) \qquad (1)$$

  s_{i-1} is the decoder's previous hidden state, h_j is the j-th hidden state of the input sequence. a is parametrized by a single hidden layer feed-forward neural network which is jointly trained with other parts of the model. The attention model does not use RNN as a decoder; consequently, it does not have s_{i-1} as its hidden state.
- **Attention weights α_j:**

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \;\; => \;\; \alpha_j = \frac{exp(e_j)}{\sum_{k=1}^{T_x} exp(e_k)} \qquad (2)$$

  is a probability to represent the weight of each input and obtained by applying softmax to the alignment score. T_x represents the number of inputs into the encoder. Therefore, a higher weight has a stronger influence on the prediction results.
- **Context vector c:**

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \;\; => \;\; c = \sum_{j=1}^{T_x} \alpha_j h_j \qquad (3)$$

  Context vector solves the forgetting problem of long sequences in RNN/LSTM. It has access to the entire input's hidden states instead of only the last one and adds importance to each of the hidden states by using the attention weights. In fact, the attention makes a shortcut between the source hidden states and the contex vector.

The mechanism of additive attention is then applied to the top of Bi-LSTM. To sum up, attention layer not only considers all of the source hidden states, but also adds importance to them. Fig. 4 illustrates a flow chart for our proposed model, encapsulating the attributes presented in figures 1, 2, and 3.

B. *Experimental design*

We utilized Matlab 2018a for the preprocessing phase. The network was implemented in Python 3.6 using Keras API 2.2.4 with TensorFlow (version 1.13) backend. The training was conducted on a system with 28 GB RAM and NVIDIA GeForce GTX 1060.

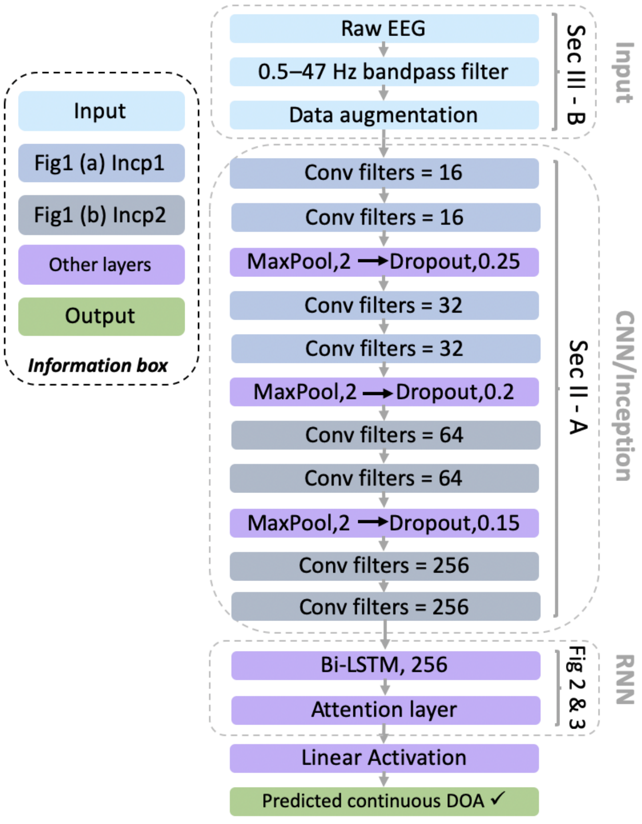


Fig. 4: High level view of the proposed method.

1) ***Regularization:*** This is to control the complexity of neural network to obtain a better generalization result [24]. To achieve this, the proposed model has: 1) Dropouts, 2) Weight decay or kernel regularization type L2 (λ is equal to 0.001), 3) Early stopping which is a termination strategy triggered when the loss function over validation set does not improve after some epochs (4 epochs in our case).

2) ***Hyper-parameter tuning:*** In our model, the weights and biases are calculated and updated to minimize the error using Adam optimizer—a gradient-based optimization of stochastic objective functions [25]. Our model shows the best performance for initial Adam's learning rate 0.0005 and batch size 64. Batch size obtained from the grid search is small causing the learning rate to be relatively small. The learning rate decays over epochs. Each epoch takes approximately 10200 sec and the model is trained for 28 epochs with early stopping criteria. The activation function is nonlinear. We exploit exponential linear unit for the non-linearity. Among mean squared error, mean absolute error, and hubber loss functions, mean square error loss shows the best result.

C. *Compared methods*

It is hard to find two anesthesia studies with an identical open-access database. Therefore, for meaningful comparison, we have to implement various methods and execute them on our database. Accordingly, We compare our proposed method to two state-of-the-art methods [11], [12]. Liu *et al.* [12] select sample entropy as a feature and random forest as a regression model to predict the continuous value of DOA. They use Multivariate empirical mode decomposition for noise reduction. Shalbaf *et al.* [11] extract sample and permutation entropy and classify them with an artificial neural network. In addition to our developed method that embodies two sub-architectures (CNN/Inception and CNN+Bi-LSTM in Fig. 4), we evaluated residual network (ResNet) 18 and 34 layers [26] with 1D convolutions.

III. Experimental setup

A. *Dataset*

The experimental dataset in our study was collected and registered by the Department of Anesthesiology and Pain Medicine, Seoul National University Hospital, College of Medicine, Seoul, Republic of Korea [27]. We chose subjects who received vascular surgery with various types of diagnosis and operations. Among 176 subjects, 74 are female and the rest are male. The physical status of the patients were ASA (American Society of Anesthesiologists) scores I, II, III and IV. Multiple techniques were conducted during anesthesia induction and maintenance. They received target-controlled infusion of propofol and/or remifentanil, administration of inhaled volatile anesthetics (sevoflurane and/or desflurane) or combination of both techniques in conjunction with another essential drug. Main characteristics of patients like age, height, weight and anesthesia duration are shown in Table I.

The data was collected using the BIS monitoring system (BIS Vista, Medtronic, Minneapolis, MN, USA) and BIS Quatro sensor. BIS represents DOA in terms of continuous values within the range of 0 (flat line EEG) to 100 (awake). The BIS sensor has 4 EEG electrodes (2 channels), which are placed directly over the forehead. The EEG sampling rate is 128 samples/sec for each channel and BIS is updated every 1 second.

B. *Data prepossessing*

EEG signals are often contaminated by another biological signals (e.g., eye blink, muscle activity, heart beat) and power line noise [30]. To deal with contamination, although the frequency range of EEG signal is from 0.5 to around 70 Hz, most of anesthesia relevant content of EEG signal occurs between 0.5 to 47 Hz [31]; consequently, EEGs are filtered by a minimum order band-pass FIR filter (bandpass function in Matlab) with cut-off frequencies of 0.5 and 47 Hz. The following cases, which contain very few samples, are excluded from dataset in order to improve the model performance.

- Segments with missing values and its corresponding BIS
- BIS with zero value and its corresponding EEG segment (Although BIS value of zero represents an isoelectric EEG signal, most zero BIS values are before the beginning of anesthesia; therefore, they do not contain any useful information.)
- EEG segments with amplitudes larger than 200 μV and smaller than -200 μV [11]
  (These outliers have very large amplitudes around 1000 μV. Few exclusions have been considered in order to improve the accuracy for the regression fit which is sensitive to outliers.)

An EEG segment refers to 1s of two EEG channels (128×2 samples).

Input representation: Before feeding the data to the model, we process the data from successive windows without overlap to the successive windows with overlap. Accordingly, we assess the spectrum over different window lengths of 1, 5, 10, 30, and 60 seconds and select 10 sec window based on performance, computational time and memory criteria. Then, the 10 sec window is moved 1 sec at a time so that the consecutive windows have 90% overlap. The detailed results are discussed in Section IV. Note that, the DNN input configuration changes from 128 $\times$ 2 to 1280 $\times$ 2, as demonstrated in Fig. 5

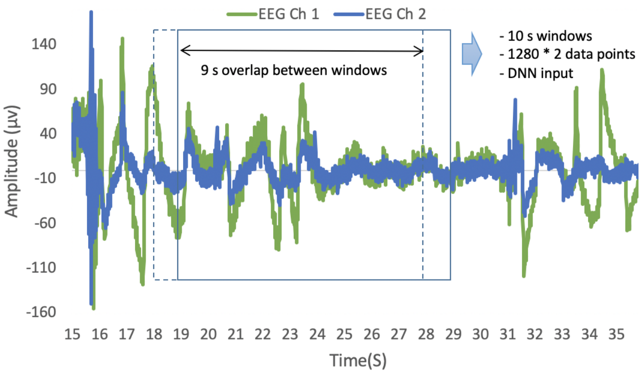


Fig. 5: Sliding of overlapping windows over the input sequence.

C. *Evaluation metrics*

After obtaining the regression outputs, the resulting predicted values are categorized into 4 levels of consciousness following the BIS manual for performance analysis using classification metrics. Accordingly, the four stages are deep anesthesia (DA, BIS: 0-40), general anesthesia (GA, BIS: 40-60), light sedation (S, BIS: 60-80), and awake (W, BIS: 80-100). The model selection is carried out through cross validation. The results are expressed in terms of regression error and classification accuracy. To evaluate the regression performance of the models, the predicted values are measured by the following most common evaluation metrics [7], [12]: mean square error (MSE), root MSE (RMSE), mean absolute error (MAE), R-squared (R^2) and correlation coefficient (CC). We use several classification metrics like accuracy (ACC), sensitivity (SE) and area under the curve (AUC). In addition, F1 score (F1), specificity (SP), and precision (PR) are used to measure and compare the performances.

IV. Experimental Results

In the validation procedure, the changes in pattern of hyperparameters are important for measuring the generalization of our deep learning model. To have an unbiased evaluation, we need to construct a separated set (from test set) called validation set. Accordingly, the dataset (described in Section III), is divided into three separated subsets for single fold cross

Table I: Patient characteristics, mean ± standard deviation (min-max)

	Training Set	Validation Set	Test Set
Age $(year)$	58.72 ± 19.64 (15-92)	57.65 ± 14.52 (17-81)	58.77 ± 14.09 (17-88)
Height (cm)	160.99 ± 15.94 (140-184)	162.98 ± 9.51 (144.8-178.2)	164.99 ± 8.73 (151-182.9)
Weight (kg)	61.26 ± 14.82 (30.7-103)	61.88 ± 11.4 (38.283.4)	64.23 ± 13.71 (43.3-101.3)
Anesthesia Duration(h)	3.34 ± 1.85 (0.67-8.83)	3.06 ± 2.15 (0.92-9.83)	2.93 ± 1.34 (0.75-8)

validation [1]. The total number of data points is 1,938,940: 20% for test subset (35 patients: 357,619 samples), 15% for validation subset are (23 patients: 244,403 samples) randomly selected and the remaining data is considered for the training subset (118 patients: 1,336,918 samples). Fig. 6 illustrates the distribution of the whole dataset (consisting of 176 patients), which is divided into 4 levels of anesthesia. Table II represents the distributions of labels in each class and set.

Then, each model predicts the BIS values of a test patient in the evaluation phase. Therefore, we report regression results in terms of mean and standard deviation for all patients in the test set. We have two approaches for comparing the results: (a) the performance of our model in tracing the BIS score is compared with the outcome of other deep learning models, (b) the selected deep learning model outcome with a feature extraction-regression method. All models are trained and tested on the same dataset as used in this study. We will analysis the results in next section.

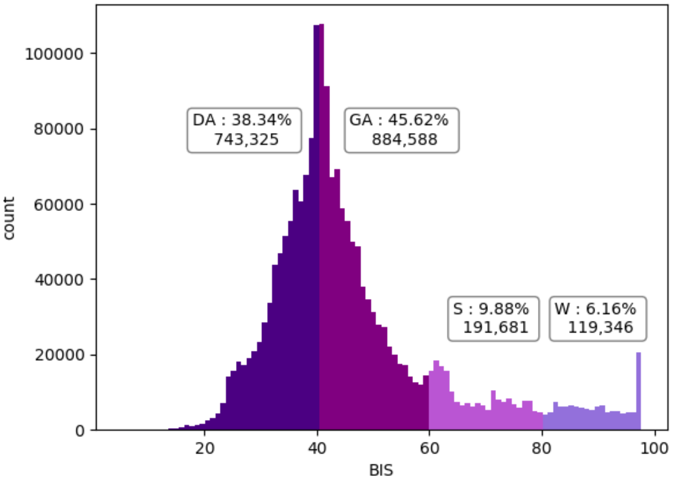


Fig. 6: Histogram of label distribution, the number of data points and percentage of each class are annotated.

Table II: BIS values in training, validation, and test sets.

DOA	Training Set	Validation Set	Test Set
DA	468065 (35.01%)	91573 (37.47%)	183687 (51.36%)
GA	658949 (49.29%)	91525 (37.45%)	134114 (37.5%)
S	130197 (9.74%)	38219 (15.64%)	23265 (6.5%)
W	79707 (5.96%)	23086 (9.44%)	16553 (4.6%)

Fig. 5 presents a graphical illustration of epoch's size modification as an input of the proposed model. Note that each EEG channel is a time series and BIS monitoring utilizes previous information to predict DOA index, a point also elaborated by other authors: e.g. Rampil [28] states "BIS value represents an average value derived from the previous 60 sec of useable data". Accordingly, the overlapping windows method significantly improves the results. There is a mean comparison of datasets with 1, 5, and 10 sec epochs in Table III. All datasets are trained on the CNN part of Fig. 4 with equivalent test set. Subsequently, CNN can better learn for larger windows using a bigger slice of samples as input. We also test 30 and 60 sec epochs. Their results are relatively close to 10 sec but their computational time and memory costs are too high. Consequently, 10 sec window provides the best results in terms of time, memory and efficiency, so we present these results.

Table III: Means ± standard deviations of window's lengths

	1 sec	5 sec	10 sec
MSE	66.26 ± 27.35	60.86 ± 24.89	36.35 ± 12.22
MAE	6.13 ± 1.35	6.03 ± 2.09	4.65 ± 0.86
RMSE	7.98 ± 1.58	7.86 ± 2.65	5.95 ± 0.99
R^2	61.89 ± 16.98	65.28 ± 15.65	79.25 ± 9.96
CC	82.23 ± 7.45	80.3 ± 12.65	91.23 ± 4.04
AUC	70.9 ± 4.53	71.38 ± 7.54	79.28 ± 4.57

Moreover, Fig. 7a shows one of the best prediction results in terms of error loss. It illustrates that the proposed model tracks out of range peaks in BIS index (level of hypnotic), which may stem from the skin incision during surgery.

V. Discussion

CNN as a deep learning model has been a choice of the most recent research papers. CNN shows great performance on sequential data like biomedical signals, especially in EEG applications [29]–[31]. There is some resemblance between hand-crafted frequency-based features and learned temporal filters by a CNN [32], [33]. The former results from convolutional operation by Fourier filters and the later are temporal filters learned by CNNs [34]. As shown in Fig. 1, the advantage of inception modules is increasing convolutional filters' learning ability by different dimensions, resulting in more complex filters. Batch normalization in convolutional layers of the proposed method helps decrease the training time. Moreover, dropout in conjunction with batch normalization reduces overfitting. Without dropout the model is biased toward learning noise patterns.

[1] We divided the dataset randomly between the three subsets with one constraint which forces the testing subset to include patients with all 4 levels of consciousness. This is done so that the absence of a level would not artificially distort the classification metrics.

Table IV: Comparing the proposed method with the benchmark approaches. All models used an identical dataset in Sec IV.

(a) Classification metrics, Values are in percentage %

	ResNet18		ResNet34		CNN/Inception		CNN+Bi-LSTM		Liu *et al.* [12]		Shalbaf *et al.* [11]		Proposed method	
stage	SE	ACC	SE	ACC	SE	ACC	SE	ACC	SE	ACC	SE	ACC	SE	ACC
DA	84.92	79.17	65.72	76.32	76.37	79.45	83.54	81.15	37.73	62.79	56.01	70.1	78.46	80.95
GA	64.09	76.3	82.8	73.5	75.19	76.78	70.07	78.54	83.73	57.5	72.49	66.86	76.57	78.43
S	50.63	95.96	54.36	96.04	55.52	95.91	57.85	96.19	41.78	92.96	61.4	92.13	64.7	96.47
W	90.62	98.86	86.39	98.88	93.42	98.58	94.78	98.8	23.25	95.89	68.76	97.16	90.71	98.98
Overall	72.57	87.57	72.32	86.17	75.13	87.68	76.56	88.67	46.62	77.28	64.67	81.57	77.62	88.71

(b) Regression metrics and AUC classifcation metric (below divider)

Metric	ResNet-18	ResNet-34	CNN/Inception	CNN+Bi-LSTM	Liu *et al.* [12]	Shalbaf *et al.*[11]	Proposed method
MSE	36.58 ± 13.7	38.67 ± 13.64	36.35 ± 12.22	32.85 ± 12.62	142.31 ± 85.52	98.83 ±69.29	32.44 ± 12.28
MAE	4.58 ± 0.94	4.83 ± 0.93	4.65 ± 0.86	4.36 ± 0.87	8.44 ± 2.37	7.38 ± 2.32	4.3 ± 0.87
RMSE	5.95 ± 1.1	6.13 ± 1.04	5.95 ± 0.99	5.61 ± 1.07	11.44 ± 3.38	9.5 ± 2.91	5.59 ± 1.04
R^2	79.25 ± 10.77	78.02 ± 10.24	79.25 ± 9.96	81.27 ± 8.9	32.09 ± 25.7	50.57 ± 23.02	81.55 ± 8.9
CC	90.43 ± 4.59	90.9 ± 4.24	91.23 ± 4.04	91.79 ± 3.93	66.33 ± 14.47	81.84 ±8.4	91.87 ± 4.03
AUC	78.33 ± 4.71	78.27 ± 5.83	79.28 ± 4.57	80.89 ± 4.34	62.29 ± 5.31	72.64± 5.71	81.11 ± 5.27

The proposed method is compared with multiple deep learning models. Results reported in Table IV suggest that RestNet18 and 34 lead to over-fitting during the learning phase because ResNet18 has a high SE in the DA stage (84.92%) and ResNet34 has a high SE in the GA stage (82.8%). However, the sensitivities are much weaker in other stages. In addition, in the training phase of these two residual networks, training loss drops after a few epochs with no improvement in validation loss. In terms of SE and ACC metrices, after multiple model training processes for each network, CNN/inception with SE 75.13% surpasses both ResNet18 and 34 with SE 72.57% and 72.32% [2]. Consequently, we choose our developed CNN/inception for the convolutional part of the proposed model.

Primarily, attention layer on top of the RNN captures and selects the most relevant extracted features by evaluation of all last hidden states. As reported in Table IV, despite a small size of S level, the proposed model shows noticeably improved SE after adding an attention layer to CNN+Bi-LSTM, from 57.85% to 64.7%.

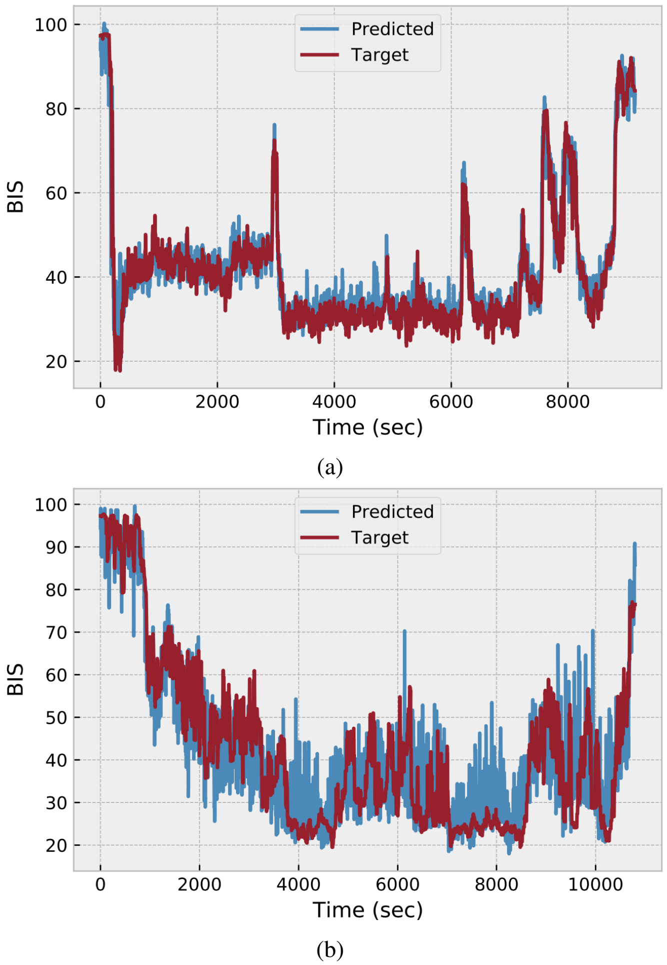
Based on the regression and classification criteria, the best model is selected through the cross validation. The conventional method needs to extract features after artifact removal. The method by Liu *et al.* [12] suffers from a complex preprocessing stage. Accordingly, this method cannot be applied to real-time situations due to its time-consuming noise reduction phase. In contrast, DNN has end-to-end learning. Exploiting a DNN like CNN does not have that tedious part, since the backpropagation weight updating decreases the effect of noise in the raw data [35]. In this application, for assessing DOA, it is crucial to have an online prediction. Our method can be applied in an online manner.

Moreover, according to Table IV both traditional methods are biased toward the training set's largest class (GA, BIS: 40-60). For example, the high SE of GA level in [12] is 83.73%, and in [11] is 72.49% with low sensitivities in other levels. Furthermore, as shown in Fig. 8, F1, PR and SP in our proposed method are significantly higher than those achieved by the other two methods: e.g. 78.31%, 80.45% and 90.3% versus 61.87%, 61.6% and 85.32% for [11], and 47.16%, 59.15% and 81.74% for [12], respectively. The AUC score for our our methods, Liu *et al.*, and Shalbaf *et al.* are 81.11 ± 5.27, 62.29 ± 5.31, and 72.64 ± 5.71, respectively. Therefore, our proposed method separates the consciousness levels more accurately than the other methods. These results suggest that there are limitations to conventional methods, which are feature-based, to much larger datasets. Unlike these approaches, DNNs elicit a set of customized features for each dataset. As a result, the proposed model does not need any experts' knowledge of feature extraction. Besides, Fig. 9 illustrates that training and validation losses of our proposed model during the learning phase provides a good generalization behavior.

Our regression scores in Table IVb are significantly better ($p < 0.001$) than the conventional methods as discuss below. For instance, we note that MSE (32.44 ± 12.28), MAE (4.3 ± 0.87), R^2 (81.55 ± 8.9), and CC (91.87 ± 4.03) of our proposed method far surpass MSE (142.31 ± 85.52), MAE (8.44 ± 2.37) , R^2 (32.09 ± 25.7), and CC (66.33 ± 14.47) of [12], and MSE (98.83 ± 69.29), MAE (7.38 ± 2.32), R^2 (50.57 ± 23.02), and CC (81.84 ± 8.4) of [11]. The results show that the proposed method predicts DOA index more precisely.

We now examine the performance of our proposed model by its confusion matrix. More specifically, all of the test results are combined and expressed in terms of confusion matrix for a better visualization in Fig. 10. Our model does not directly perform the classification; therefore, many of misclassifications have close distances with their true BIS indexes. It is noted that demarcation of the four states of consciousness is an artifact of classification and it can lead to misclassification with small changes in BIS scores. For example, we note that in Fig. 11 the true BIS is 38.2 (DA stage) and the predicted index is 41.12 (GA stage). Although there is only a small difference between the indices, this index is counted as a misclassification in the confusion matrix. Subsequently, the SE scores may not

[2] Although SE of proposed method is better than those of other deep learning methods both numerically and statistically, ACC is only better numerically but not statistically.



(a)

(b)

Fig. 7: Comparison of true and predicted BIS by the proposed method. (a) One of the best prediction with MAE 2.88, age: 48 year-old. (b) Worst prediction with MAE 6.4, age: 63 year-old.

express the exact class.

VI. Conclusion

Real-time monitoring of DOA by automatic analysis of EEG is very useful in assisting anesthesiologists with there decision and judgment. EEG monitoring and online DOA assessment during surgery have great potential for improving neurologic outcomes. We propose a deep learning architecture based on inception building blocks, Bi-LSTM, and an attention layer, to accurately measure the DOA automatically. Our dataset contains a wide range of patients with various characteristics and different types of anesthesia. Statistical analysis shows that in terms of both regression and classification errors our model outperforms the competitive methods particularly, for large EEG datsets.

In future studies, informative features like those from EMG can be added to our proposed model to further improve the performance in terms of both error types. Moreover, it would be useful to evaluate the effect of different attention networks with encoder-decoder mechanism.

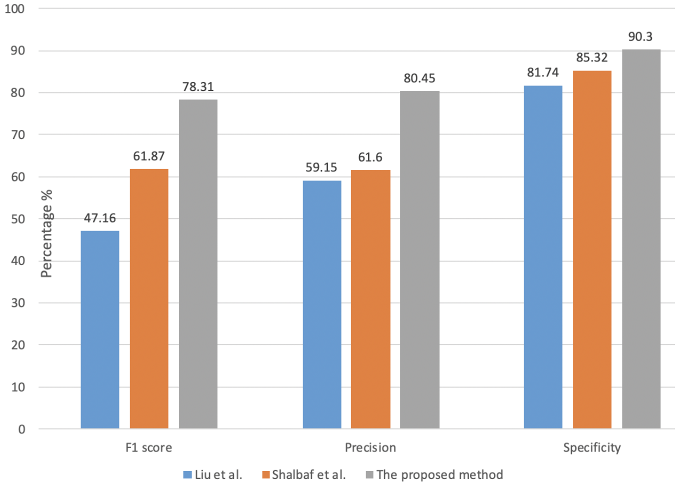


Fig. 8: F1, PR and SP of the mentioned models which are trained and tested on this study dataset.

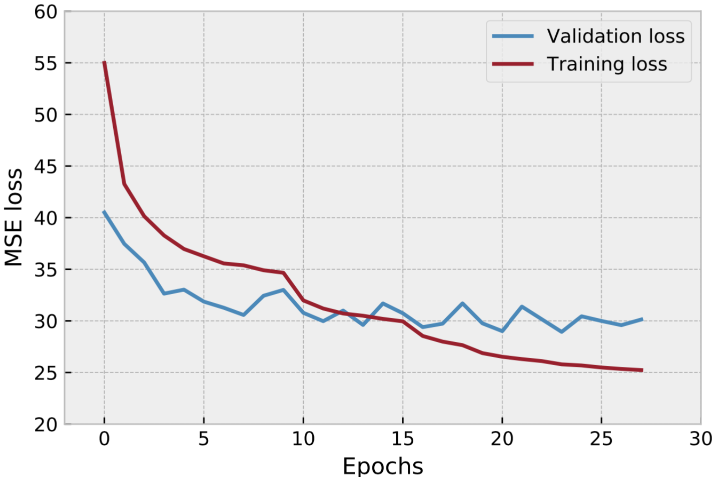


Fig. 9: Training and validation losses during the learning process.

References

[1] A. M. Brambrink and J. R. Kirsch, *Essentials of neurosurgical anesthesia & critical care: strategies for prevention, early detection, and successful management of perioperative complications*. Springer Nature, 2019.

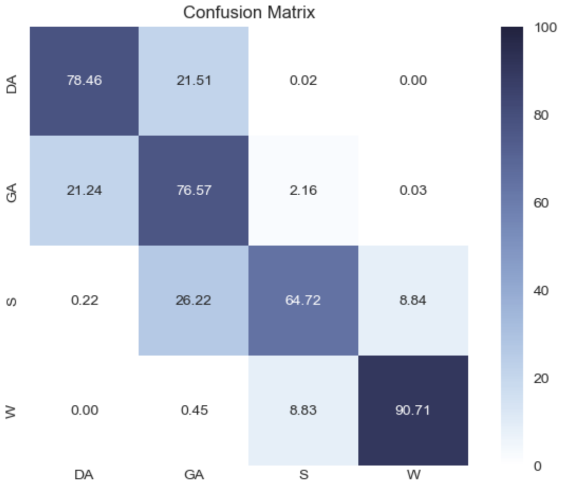
[2] A. Nimmo, A. Absalom, O. Bagshaw, A. Biswas, T. Cook, A. Costello, S. Grimes, D. Mulvey, S. Shinde, T. Whitehouse *et al.*, "Guidelines for the safe practice of total intravenous anaesthesia (TIVA) joint guidelines from the association of anaesthetists and the society for intravenous anaesthesia," *Anaesthesia*, vol. 74, no. 2, pp. 211–224, 2019.

[3] Y. Punjasawadwong, A. Phongchiewboon, and N. Bunchungmongkol, "Bispectral index for improving anaesthetic delivery and postoperative recovery," *Cochrane Database of Systematic Reviews*, no. 6, 2014.

[4] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2016.

[5] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," *Computers in Biology and Medicine*, p. 103801, 2020.

[6] H. Korkalainen, J. Aakko, S. Nikkonen, S. Kainulainen, A. Leino, B. Duce, I. O. Afara, S. Myllymaa, J. Töyräs, and T. Leppänen, "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2073–2081, 2019.



	DA	GA	S	W
DA	78.46	21.51	0.02	0.00
GA	21.24	76.57	2.16	0.03
S	0.22	26.22	64.72	8.84
W	0.00	0.45	8.83	90.71

Fig. 10: Confusion matrix of test set. True positive and false negative percentages are shown. The columns show the predicted levels. The accuracy is 88.71%.

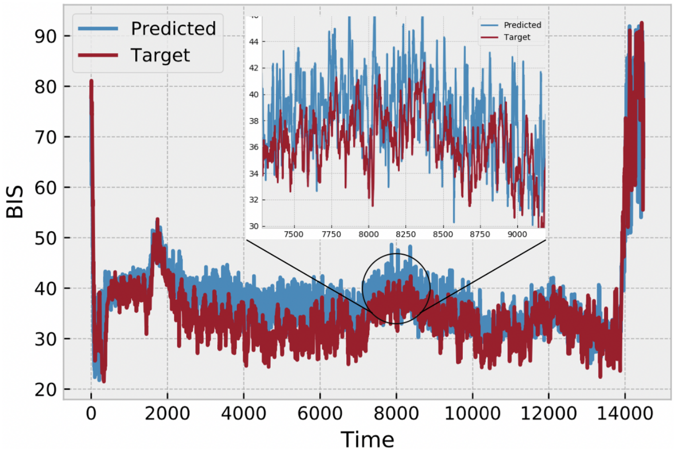


Fig. 11: Closer look at the difference between true and predicted labels of randomly selected patient.

[7] H.-C. Lee, H.-G. Ryu, E.-J. Chung, and C.-W. Jung, "Prediction of bispectral index during target-controlled infusion of propofol and remifentanil," *Anesthesiology*, vol. 128, no. 3, pp. 492–501, 2018.

[8] T. Zoughi and R. Boostani, "Presenting a combinatorial feature to estimate depth of anesthesia," *Int. Journal of Signal Processing*, vol. 6, no. 2, pp. 10–14, 2010.

[9] S. Afrasiabi, R. Boostani, S. Koochaki, and F. Zand, "Presenting an effective EEG-based index to monitor the depth of anesthesia," in *The 16th CSI Int. Symposium on Artificial Intelligence and Signal Processing (AISP 2012)*. IEEE, 2012, pp. 557–562.

[10] M. Särkelä, S. Mustola, T. Seppänen, M. Koskinen, P. Lepola, K. Suominen, T. Juvonen, H. Tolvanen-Laakso, and V. Jäntti, "Automatic analysis and monitoring of burst suppression in anesthesia," *Journal of Clinical Monitoring and Computing*, vol. 17, no. 2, pp. 125–134, 2002.

[11] R. Shalbaf, H. Behnam, J. W. Sleigh, A. Steyn-Ross, and L. J. Voss, "Monitoring the depth of anesthesia using entropy features and an artificial neural network," *Journal of Neuroscience Methods*, vol. 218, no. 1, pp. 17–24, 2013.

[12] Q. Liu, L. Ma, S.-Z. Fan, M. F. Abbod, and J.-S. Shieh, "Sample entropy analysis for the estimating depth of anaesthesia through human EEG signal at different levels of unconsciousness during surgeries," *PeerJ*, vol. 6, p. e4817, 2018.

[13] A. Shalbaf, M. Saffar, J. W. Sleigh, and R. Shalbaf, "Monitoring the depth of anesthesia using a new adaptive neurofuzzy system," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 671–677, 2017.

[14] A. Lashkari and R. Boostani, "A kalman-based instantaneous frequency estimation for anesthetic depth measurement," in *2017 22nd Int. Conf. on Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–4.

[15] Q. Liu, J. Cai, S.-Z. Fan, M. F. Abbod, J.-S. Shieh, Y. Kung, and L. Lin, "Spectrum analysis of EEG signals using cnn to model patient's consciousness level based on anesthesiologists' experience," *IEEE Access*, vol. 7, pp. 53 731–53 742, 2019.

[16] H.-C. Lee, H.-G. Ryu, Y. Park, S. B. Yoon, S. M. Yang, H.-W. Oh, and C.-W. Jung, "Data driven investigation of bispectral index algorithm," *Scientific Reports*, vol. 9, no. 1, pp. 1–8, 2019.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of the 32nd Int. Conf. on Int. Conf. on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, p. 448–456.

[19] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

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

[21] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd Int. Conf. on Learning Representations, ICLR 2015, San Diego, CA, USA, Conf. Track Proc.*, 2015.

[23] C. Raffel and D. P. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *arXiv preprint arXiv:1512.08756*, 2015.

[24] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford University Press, 1995.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd Int. Conf. on Learning Representations, ICLR 2015, San Diego, CA, USA, May, Conf. Track Proc.*, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] H.-C. Lee and C.-W. Jung, "Vital recorder—a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices," *Scientific Reports*, vol. 8, no. 1, pp. 1–8, 2018.

[28] I. J. Rampil *et al.*, "A primer for EEG signal processing in anesthesia," *Anesthesiology: The Journal of the American Society of Anesthesiologists*, vol. 89, pp. 980–1002, 1998.

[29] A. Antoniades, L. Spyrou, D. Martin-Lopez, A. Valentin, G. Alarcon, S. Sanei, and C. C. Took, "Deep neural architectures for mapping scalp to intracranial EEG," *Int. Journal of Neural Systems*, vol. 28, no. 08, p. 1850009, 2018.

[30] 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," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[31] D. Jarchi, R. Boostani, M. Taheri, and S. Sanei, "Seizure source localization using a hybrid second order blind identification and extended rival penalized competitive learning algorithm," *Biomedical Signal Processing and Control*, vol. 4, no. 2, pp. 108–117, 2009.

[32] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns." in *INTERSPEECH*. ISCA, 2015, pp. 1–5.

[33] A. I. Humayun, S. Ghaffarzadegan, M. I. Ansari, Z. Feng, and T. Hasan, "Towards domain invariant heart sound abnormality detection using learnable filterbanks," *IEEE journal of biomedical and health informatics*, vol. 24, no. 8, pp. 2189–2198, 2020.

[34] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of Neural Engineering*, vol. 16, no. 5, p. 051001, 2019.

[35] G. Li, C. H. Lee, J. J. Jung, Y. C. Youn, and D. Camacho, "Deep learning for eeg data analytics: A survey," *Concurrency and Computation: Practice and Experience*, p. e5199, 2019.