An Evaluation of Hybrid Deep Learning Models for Classifying Multiple Lower Limb Actions

Zilu Wang, Ian Daly, and Junhua Li, *Senior Member, IEEE*

***Abstract*—Brain-computer Interfaces (BCIs) interpret electroencephalography (EEG) signals and translate them into control commands for operating external devices. The motor imagery (MI) paradigm is popular in this context. Recent research has demonstrated that deep learning models, such as convolutional neural network (CNN) and long short-term memory (LSTM), are successful in a wide range of classification applications. This is because CNN has the property of spatial invariance, and LSTM can capture temporal associations among features. A combination of CNN and LSTM could enhance the classification performance of EEG signals due to the complementation of their strengths. Such a combination has been applied to MI classification based on EEG. However, most studies focused on either the upper limbs or treated both lower limbs as a single class, with only limited research performed on separate lower limbs. We, therefore, explored hybrid models (different combinations of CNN and LSTM) and evaluated them in the case of individual lower limbs. In addition, we classified multiple actions: MI, real movements and movement observations using four typical hybrid models and aimed to identify which model was the most suitable. The comparison results demonstrated that no model was significantly better than the others in terms of classification accuracy, but all of them were better than the chance level. Our study informs the possibility of the use of multiple actions in BCI systems and provides useful information for further research into the classification of separate lower limb actions.**

I. Introduction

Brain-computer interfacing (BCI) is an emerging technology, enabling direct interactions between the human brain and external devices or the environment. One BCI application is to enable people who have lost their motor functions to communicate and control some devices, which would result in improvements in their quality of life. For instance, motor imagery (MI)-based BCI can convert the motion intentions of users into control commands [1]. Hence in this regard, MI-based BCI can also be used for rehabilitation. In other words, MI can be regarded as a new strategy of the motor system and a rehabilitation method for patients with movement impairment [2].

Whether the user's motion intention can be correctly identified is one of the important indexes to evaluate the performance of MI-based BCI systems. In the task of limb movement intention identification, electroencephalography (EEG) is widely used due to its advantages of high temporal resolution, cost-effectiveness, portability and noninvasive resolution [3]. Multiple studies have shown that during MI, the frequency band power of the EEG signal varies according to the content of the imaginary task [4]. In particular, the μ rhythm (8-13 Hz) and β rhythm (14-30 Hz) frequency bands are modulated by MI, showing frequency power elevation (ERS: Event-Related Synchronization) and attenuation (ERD: Event-Related Desynchronization) depending on the brain regions [5]. This phenomenon can be utilized to classify users' motion intentions based on EEG signals.

Deep learning (DL) models, such as convolutional neural network (CNN), autoencoder, or long short-term memory (LSTM), have been used in a number of domains, including speech recognition, audio processing, and computer vision [6],[7]. Li pointed out that DL models have great potential for neural signal analysis and classification [8]. These models have also been brought to EEG signal classification and have become popular in the BCI domain. For example, CNN is able to capture spatial patterns existing in EEG signals [9], while LSTM has the merit of tracing temporal relationships contained in EEG signals [10]. To improve the ability to extract spatial and temporal features simultaneously, hybrid models consisting of CNN and LSTM networks have been used to learn spatial and temporal features. In the application of hybrid CNN-LSTM models, the combination of models is diverse for different studies. For example, Zhang et al. [11] used a one-versus-rest filter bank common spatial mode to pre-extract the features of signals in a study with four classes of MI tasks. They segmented continuous EEG into time windows of 0.8 seconds in length. These time windows were fed into a hybrid DL network combining a CNN and LSTM. This model achieved a good classification accuracy. In other studies, researchers explored the model architecture. For example, cascaded and parallel structures were proposed. Currently, the most popular cascaded structure combining a CNN and LSTM is to use a CNN for extracting spatial features from the EEG signals, and then to use an LSTM for further feature extraction by capturing temporal information. Zhu et al. [12] used a cascaded hybrid model with a CNN and LSTM to classify the MI-related EEG signals of the left fist and right fist and achieved a classification accuracy of more than 80%. Additionally, this model was also used to convolve electrode channels to explore the effect of different electrode channel combinations on classification performance. The results showed that the more electrode channels in these combinations, the higher the classification accuracy. A parallel hybrid model is another way of combining CNN with LSTM. This method feeds EEG signals into the CNN and LSTM models respectively and then fuses the output features of the two models for classification. Li et al. [13] proposed a feature fusion algorithm based on the hybrid parallel CNN-LSTM model. They applied this hybrid CNN-LSTM model to a four-class MI dataset. The model shows good classification

Z. Wang and I. Daly are with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK (zw20774@essex.ac.uk, i.daly@essex.ac.uk).

J. Li is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK, and also with the Laboratory for Brain-Bionic Intelligence and Computational Neuroscience, Wuyi University, Jiangmen, 529020, China (Correspondence should be addressed to J. Li at junhua.li@essex.ac.uk).

performance, with an average accuracy across all participants of 87.68%.

There is evidence showing that MI-BCI is a promising method for improving the feasibility and effectiveness of the rehabilitation [14]. However, the success of lower limb MI-BCI research has not been comprehensively demonstrated, and there is relatively little research focused on the classification of lower limb actions using EEG signals. In our previous research [15], we designed an experiment to study the separability of multiple lower limb actions based on EEG signals. This experiment is of six different actions, namely, MI of left lower limb (Left-MI) and right lower limb (Right-MI), real movement (RM) of left lower limb (Left-RM) and right lower limb (Right-RM), movement observations (MO) of left lower limb (Left-MO) and right lower limb (Right-MO). Based on our previous exploration, we found that these six actions could be separable, but it was challenging to classify them due to subtle differences among them. The hybrid CNN-LSTM model has shown good classification performance in MI-EEG signal classification tasks. However, the types of the hybrid model are varied, and it is not known about their performance in the action classification. Therefore, in this paper, we evaluate four typical hybrid model frameworks and compare their classification performance to find out which is most suitable in the context of multiple lower-limb action classification.

II. Methods

A. Data Acquisition and Processing

The data used in this study were acquired from 28 participants when they performed six predefined actions. Each of them completed six experiment sessions. Each session consisted of 72 trials. A fixation cross appeared at the centre of a screen and lasted for a randomized period from 1.5-2.5 seconds, indicating the start of a trial. A condition cue (indicating one of the required actions: MI, RM or MO) followed and lasted for one second. After that, an arrow (indicating either the left or right lower limb) was added above the condition cue to instruct participants to start performing the required actions, which lasted for 6 seconds. EEG data related to actions were recorded by 62 electrodes arranged according to the layout of the international standard 10/20 system. The sampling rate for recording EEG signals was 250 Hz. A band-pass filter (0.5 Hz ~ 45 Hz) was applied to the EEG signals. Independent component analysis was used to remove artifacts and reconstruct the signals for the multichannel EEG. The EEG segments corresponding to the periods of action implementation (6 seconds) were retained and used as samples in this study. The experiment was reviewed and approved by the Institutional Review Board of the National University of Singapore, and the Humanities, Science and Health, or Social Science Ethics Sub-Committee at the University of Essex.

B. Models

We identified the existing combinations of CNN and LSTM based on the literature, and categorized them into four types of hybrid frameworks. According to a literature review [16], around 30% of the studies arranged EEG into a 2D format as input, and about 30% of studies use time-frequency maps as input. In this study, time-frequency maps obtained from a short-time Fourier transform (STFT) were used as the input for each type of hybrid CNN-LSTM framework.

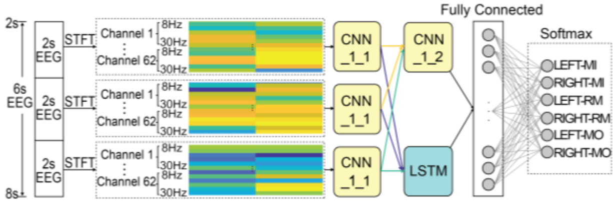

Figure 1. Model architecture of type_1.

We determined the STFT parameters according to the ERD/ERS time course associated with lower limb actions. Specifically, a one-second time window was used and slid over the segment with an overlapping of 0.5 seconds. Then, the portion of 8 to 30 Hz was extracted and used as features. This setting was the same for all types of models. Type_1 was shown in Fig. 1. EEG of each trial (6 seconds) was divided into 3 segments. Each segment was two seconds long, for which feature extraction was performed. The EEG signal from each channel was converted into a 2D time-frequency map with the help of STFT. The size of the time-frequency map was 23×3 $(frequency \times time)$. We then merged the maps from all channels into a larger 2D time-frequency map as the features of each segment. The size of the 2D time-frequency map obtained for each segment is 1426×3. The time-frequency maps were then fed into the first CNN block (marked as CNN_1_1 in Fig. 1), which was a monolayer including one convolutional layer and a batch normalization layer. The size of the convolution kernel was 3×3, and the number of filters was 32. The second CNN block (marked as CNN_1_2) had three hidden layers. Each hidden layer consisted of a convolution layer, a normalization layer and a max-pooling layer. A rectified linear unit (ReLU) was used as an activation function. The kernel size in each convolution layer was 3×3, and the number of filters was 4, 8, and 16, respectively. At each max-pooling layer, a kernel size of 2×2 was applied to reduce the size of the feature matrix. The features outputted from CNN_1_1 were not only fed into CNN_1_2 for decoding spatial features further but also fed into an LSTM layer with 100 neurons for decoding temporal features. The outputs of CNN_1_2 and the LSTM were fed into a fully connected layer with 512 neurons and a ReLU activation function. Finally, a softmax layer was put at the end of the network to achieve the classification results.

Another way for arranging the input EEG data is to retain the spatial layout of electrodes. That is, each electrode has physical neighbours around that electrode. Therefore, the EEG signals should be arranged according to the layout of the electrodes. In light of this fact, type_2 is with such input form. The entire type_2 model architecture is shown in Fig. 2. We kept channel locations and arranged them into a matrix with the padding of zeros for those locations without recording channels (e.g., around the corners). We applied STFT to each channel data to obtain a time-frequency map with the size of 23×11, and then reshaped this time-frequency map into a vector with the dimension of 253. After repeating STFT for each channel, we obtained a 3D tensor in the size of $9 \times 9 \times 253$. This 3D tensor was fed into a CNN with 3D kernel. It has two convolution layers and two max-pooling layers. The size of the kernels in the two convolution layers was $3 \times 3 \times 3$ .The numbers of filters were 40 and 80, respectively. The kernel size in the max-pooling layer was $2 \times 2 \times 2$. ReLU was used as an activation function, and the normalization layer was added after the CNN layer, then expressed as features through a fully connected layer. Its output was then fed into LSTM layer, which contained 100 neurons. Another fully connected layer with 100 neurons was followed, which was followed by a softmax layer for classification.

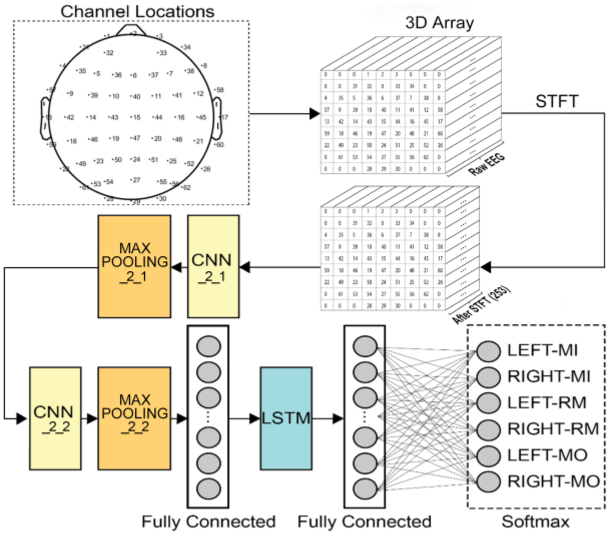

Figure 2. Model architecture of type_2 for extracting the spatio-temporal information of EEG.

In addition to the above two types of models, other researchers attempted to improve the model classification performance by changing the model architecture. These are the cascaded structure (see type_3 in Fig. 3) and the parallel structure (see type_4 in Fig. 4). Type_3 used 2D time-frequency map as input after feature extraction by STFT. A feature vector (253 in length) was extracted from each channel, which was reshaped from the time-frequency map 23×11. All vectors of 62 channels were then formed into a 2D matrix (62×253). This feature matrix was inputted to the cascaded structure, which consisted of two convolutional layers (CNN_3_1, CNN_3_2) and two max-pooling layers (MAX POOLING_3_1, MAX POOLING_3_2). The size of the CNN_3_1 kernel was set to a matrix of 62×4, which was mainly used to spatially convolve the input features. The CNN_3_2 kernel was set to 1×8, which was mainly to convolve the features along the temporal dimension. The numbers of filters were 40 and 80, respectively. The MAX POOLING_3_1 kernel was set to 1×4, and the MAX POOLING_3_2 kernel was set to 1×8. Then, the features were fed into LSTM layer with 100 neurons after a fully connected layer containing 1024 neurons, and then into another fully connected layer containing 1024 neurons. Lastly, the features were passed through softmax layer to have classification results.

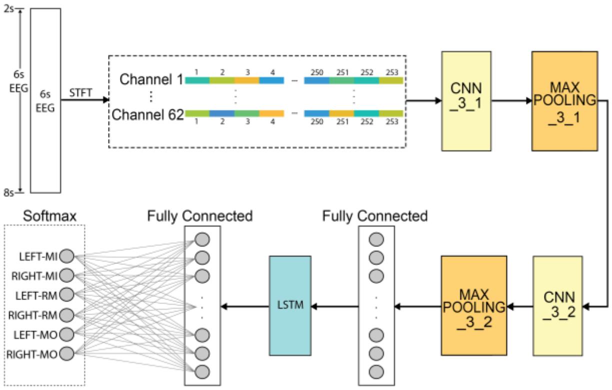

Figure 3. Model architecture of type_3 with the cascaded structure of CNN-LSTM.

Type_4 is of a parallel structure. That is, CNN and LSTM are used to process features in parallel. We obtained time-frequency maps (23×11) for each channel and then assembled them into a larger time-frequency map (1426×11), which was inputted into the model. CNN-LSTM parallel structure was used to extract spatial and temporal characteristics. The CNN consisted of two convolution layers (CNN_4_1 and CNN_4_2), two max-pooling layers and a fully connected layer. The size of the CNN_4_1 and CNN_4_2 kernels was 3×3, and the numbers of filters were 32 and 62, respectively. The max-pooling size was 3×3. The number of neurons in the fully connected layer was 512. The LSTM consisted of an LSTM layer with 100 neurons and a fully connected layer with 1024 neurons. After feature extraction through the CNN and LSTM, feature fusion was performed through a fully connected layer with 512 neurons. Finally, the combined features derived from the fully connected layer were input into a softmax layer for classification. In type_3 and type_4 models, ReLU activation and batch normalization were used after each CNN layer and fully connected layer.

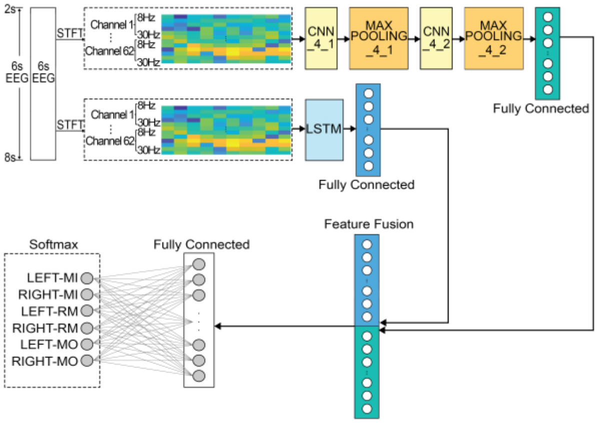

Figure 4. Model architecture of type_4 with the parallel structure of CNN-LSTM.

III. Results and Discussion

The classification performance of the four types of hybrid models was evaluated through five-fold cross-validation. The detailed classification results are shown in Table I.

TABLE I. Accuracies of Four Hybrid Models in the Context of Multiple Action Classification

participants	Accuracy (%)			
	Type_1	Type_2	Type_3	Type_4
1	50.87	55.11	50.18	51.11
2	37.27	32.87	37.74	33.11
3	34.95	36.57	34.69	34.74
4	50.97	51.61	50.21	50.72
5	27.99	30.08	29.43	30.07
6	28.00	34.05	31.71	31.27
7	38.42	36.36	34.46	35.73
8	27.82	22.79	21.57	20.89
9	38.87	17.36	22.94	29.37
10	37.96	44.17	39.11	38.13
11	39.12	39.16	37.94	36.57
12	34.27	40.95	37.30	34.57
13	29.66	30.81	36.76	31.72
14	44.92	46.77	42.11	45.63
15	35.19	40.29	38.17	43.31
16	41.18	37.74	37.51	39.36
17	31.71	31.72	33.56	31.95
18	27.99	30.57	35.86	31.24
19	23.61	27.29	28.24	27.09
20	34.47	34.23	35.64	32.43
21	29.42	27.27	28.50	30.54
22	38.62	39.06	39.72	38.76
23	29.62	33.17	34.07	33.80
24	31.26	38.40	37.28	35.90
25	50.24	53.23	54.20	48.13
26	35.64	36.09	34.25	34.46
27	39.56	17.61	19.06	20.20
28	34.02	35.40	32.98	33.09
Mean±STD	35.84±7.18	35.74±9.24	35.54±7.88	35.13±7.52

Statistical analysis was used to test whether there were significant differences between hybrid models in terms of classification accuracy. The statistical results showed that there was no significant difference between these four types of hybrid models ($F_{(3,108)} = 0.04, p = 0.988$). Fig. 5 shows the means and standard deviations of the four hybrid models. It may be observed that the classification performance of the models varies considerably across participants. This may be due to different participants exhibiting different sensitivities when performing the lower limb tasks. When checking the classification performance within each participant, it was very close for all four models. Overall, Type_1 had the highest average accuracy, which was only marginally higher than the other three types of models. All types of models performed significantly better than the chance level (16.67%), (paired t-test, all $p < 10^{-14}$). This result suggests that the four models are effective for multiple action classification.

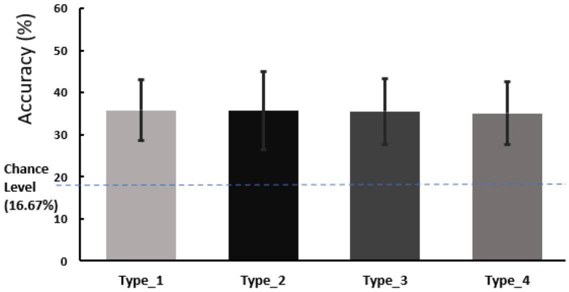

Figure 5. The means and standard deviations of the classification accuracies for four hybrid models.

IV. Conclusion

This paper evaluated the performance of hybrid models of CNN and LSTM for classifying multiple actions of separate lower limbs. Four typical models were identified and tested on the EEG dataset. The classification results showed that there was no significant difference in classification performance among the four types of models. However, all models performed significantly better than the chance level, implying that these four hybrid models were useful for classifying multiple actions.

References

[1] Z. Khademi, F. Ebrahimi, and H. M. Kordy, “A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals,” *Comput. Biol. Med.*, vol. 143, no. November 2021, p. 105288, 2022.

[2] S. H. Johnson, G. Sprehn, and A. J. Saykin, “Intact motor imagery in chronic upper limb hemiplegics: Evidence for activity-independent action representations,” *J. Cogn. Neurosci.*, vol. 14, no. 6, pp. 841–852, 2002.

[3] Teo, Wei-Peng, and Effie Chew. "Is motor-imagery brain-computer interface feasible in stroke rehabilitation?." PM&R: 723-728, 2014.

[4] G. Pfurtscheller and F. H. Lopes Da Silva, “Event-related EEG/MEG synchronization and desynchronization: Basic principles,” Clin. *Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, 1999.

[5] G. Pfurtscheller and C. Neuper, “Motor imagery and direct brain-computer communication,” *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[6] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, “Human Activity Recognition using Binary Motion Image and Deep Learning,” *Procedia Comput. Sci.*, vol. 58, pp. 178–185, 2015.

[7] S. Furui, L. Deng, M. Gales, H. Ney, and K. Tokuda, “Fundamental technologies in modern speech recognition,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 16–17, 2012.

[8] J. Li, “Thoughts on neurophysiological signal analysis and classification,” *Brain Sci. Adv.*, vol. 6, no. 3, pp. 210–223, 2020.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[10] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang, “LSTM-Based EEG Classification in Motor Imagery Tasks,” *IEEE Trans. NEURAL Syst. Rehabil. Eng.*, vol. 26, no. 11, 2086, 2018.

[11] R. Zhang, Q. Zong, L. Dou, and X. Zhao, “A novel hybrid deep learning scheme for four-class motor imagery classification,” *J. Neural Eng.*, vol. 16, no. 6, 2019.

[12] K. Zhu, S. Wang, D. Zheng, and M. Dai, “Study on the effect of different electrode channel combinations of motor imagery EEG signals on classification accuracy,” *J. Eng.*, vol. 2019, no. 23, pp. 8641–8645, 2019.

[13] H. Li, M. Ding, R. Zhang, and C. Xiu, “Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network,” *Biomed. Signal Process. Control*, vol. 72, no. PA, p. 103342, 2022.

[14] F. Malouin and C. L. Richards, “Mental practice for relearning locomotor skills,” *Phys. Ther.*, vol. 90, no. 2, pp. 240–251, 2010.

[15] Z. Wang, J. Li, I. Daly, and J. Li, “Machine Learning for Multi-Action Classification of Lower Limbs for BCI,” *Proc. - 2022 Int. Conf. Comput. Electron. Commun. Eng. iCCECE 2022*, no. Mi, pp. 84–89, 2022.

[16] H. Altaheri *et al.*, “Deep learning techniques for classification of electroencephalogram (EEG) motor imagery (MI) signals: a review,” Springer. Neural Computing and Applications, pp.1-42, 2021.