IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 24, 2020, accepted June 6, 2020, date of publication June 15, 2020, date of current version June 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002459

Motor Imagery Classification for Brain Computer Interface Using Deep Metric Learning

HAIDER ALWASITI[1], **(Member, IEEE)**, **MOHD ZUKI YUSOFF**[1], **(Member, IEEE)**, **AND KAMRAN RAZA**[2], **(Senior Member, IEEE)**

[1]Centre for Intelligent Signal and Imaging Research (CISIR), Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, Bandar Seri Iskandar 32610, Perak, Malaysia

[2]Faculty of Engineering, Sciences, and Technology, Iqra University, Karachi 75500, Pakistan

Corresponding author: Haider Alwasiti (hwasiti@ieee.org)

This work was supported in part by the Graduate Assistantship Scheme of Universiti Teknologi PETRONAS, in part by the Ministry of Education Malaysia through the Higher Institutional Centre of Excellence (HICoE) Scheme awarded to the Centre for Intelligent Signal and Imaging Research (CISIR), and in part by the YUTP Fund under Grant 0153AA-H39.

ABSTRACT Deep metric learning (DML) has achieved state-of-the-art results in several deep learning applications. However, this type of deep learning models has not been tested on the classification of electrical brain waves (EEG) for brain computer interface (BCI) applications. For the first time, we propose a triplet network to classify motor imagery (MI) EEG signals. Stockwell Transform has been used for converting the EEG signals in the time domain into the frequency domain, which resulted in improved DML classification accuracy in comparison to DML with Short Term Fourier Transform (0.647 vs. 0.431). DML model was trained with a topogram of concatenated 64 EEG channel spectrograms. The training batch was comprised of triplet pairs of the anchor, positive, and negative labeled epochs. The triplet network was able to train an embedding feature space that minimized the Euclidean distance between the embeddings of spectrograms of the same class and increased the distance between the embeddings of different labeled images. The proposed method has been tested on an EEG dataset of 109 untrained subjects. We showed that the DML classifier is able to converge with an extremely small number of training samples ($\sim$ 120 EEG trials) for only one subject per model, mitigating the well-known issue of the large inter-individual variability of human MI-BCI EEG which degrades the classification performance. The proposed preprocessing pipeline and the Triplet Network provide a promising method to classify MI-BCI EEG signals with much less training samples than the previous methods.

INDEX TERMS BCI, metric learning, EEG, Stockwell transform.

I. INTRODUCTION

Brain computer interfaces provide a direct control and communication path between brain and external devices. By analyzing electrical brain signals (EEG) during imagination to move hands, for instance, researchers have shown that motor imagery EEG waves are being modulated; the signals can be detected and used for assisting or replacing the normal muscular control, which is especially useful for patients with paralysis [76]. However, EEG signals are weak with a low signal-to-noise ratio and relatively low spatial resolution [8]. This is compounded by the fact that physiologically, the brain regions are not solely responsible for a single function, nor each function is performed by a single brain region [9], [76]. Large inter-individual variability and even intra-individual variability (EEG is not consistent from trial to trial) posed significant challenges for the classification of EEG signals for BCIs.

BCI EEG signals are known for high inter-individual variability [76]. Most previous BCI studies used relatively small to moderate training sample size ($\sim$ 8-30 subjects) and used all subjects' training data to train one model (training set usually divided into 80% of each user's EEG data and the remaining was held out for testing). Then, the accuracy of this trained model has been reported on each user's test data separately [28], [41], [45], [64], [66], [71], [72]. The motivation in the previous studies to train a single model on all users' training data, while the model was tested on each individual separately was likely, because the classification performance of most classifiers usually increase with more training data.

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

However, because of the inter-individual variability, there is an optimal balance between the effect of increasing the accuracy by enlarging the training set size (with the involvement of more subjects in training), and decreasing the classification accuracy due to the inter-individual variability, which its detrimental effect is directly related to the number of subjects involved. The obvious solution is to take a much larger training sample from each individual, and train a dedicated DL classifier for each subject alone. However, EEG data collection is an expensive procedure, and long data collection sessions rapidly lead to user fatigue, which deteriorate the EEG sample quality. Therefore, most previous studies have incorporated multiple subjects to train one model, which was the best strategy to get the highest classification accuracy with their approaches. There are only few studies reported to train a DL model on single subject EEG data, where they used a larger private dataset in comparison to the public MI-BCI datasets [15], [64]. The proposed method in this paper uses a publicly available EEG dataset with small EEG training set per each subject, and shows that it is possible to train a DL model with comparable performance with the previous methods using a very limited dataset of one subject only, which eliminates the need for expensive and long EEG data collection sessions. In the Results section, we have elaborated more on the relation of the number of subjects used for training the deep metric learning (DML) model versus the classification performance, and how the DML model performed better with only one subject with much less training data in comparison to the previous methods.

In light of this, the motivation in this study is to employ a DML model to classify MI-EEG signals due to its ability to converge with very limited training samples [31]. This particular advantage of this classifier made this approach possible to train a single model on each subject with a limited training size ($\sim$ 120 samples per model), effectively mitigating the issue of inter-individual variability of MI-BCI EEG.

Several studies [1], [7], [28], [32], [40], [42], [44], [45], [50], [51], [62], [66], [71], [72], [77] have investigated the classification of MI-EEG to develop a BCI system that can provide feedback during MI training and eventually use the BCI to enhance the life of patients with disabilities and paralysis.

Various methods for feature extraction have been tested in the literature. A common method of EEG feature extraction is to use Fast Fourier Transform to convert EEG signals into the frequency domain and using the Common Spatial Pattern (CSP) algorithm [28], [42], [44], [66], [77] to classify MI signals. The extracted features were used for training a classifier, where most researchers have applied classical machine learning classifiers such as Mahalanobis Linear Discrimination classifier [32], Support Vector Machine [28], [77], Bayesian classifier [50], Bayesian Linear Discriminant Analysis [42], Logistic Regression [40] or Linear Discriminant Analysis [51].

While the CSP feature extraction method focuses on a frequency selection to pick up the most significant features correlated with the motor imagery task, Deep Learning (DL) classifiers show promising new classification modalities that render such feature selection methods unnecessary. A feature selection in classical machine learning models was an essential requirement due to the curse of dimensionality [20]. In particular, Hughes Phenomenon showed that if the number of features increases after a certain threshold (based on the type and size of the parameter space of the model), the model's performance will start to decline. There are few theories that show how deep learning models are not affected by the curse of dimensionality [23]. The sparse coding theory and the manifold hypothesis are suggesting that the high dimensional feature space manifold actually sits on top of the lower dimensional feature space embedded in the higher dimensional manifold. Deep learning models are very good at exploiting the higher dimensional feature space and reducing it to the lower dimensional manifold. This was encouraging to include all possible EEG channels in this study, with large frequency bandwidth without feature selection, which is rarely adopted in previous BCI studies.

In literature, diverse classical machine learning classifiers have been explored for MI-EEG classification. For instance, Huang *et al.* achieved 0.57 classification accuracy using genetic algorithm based mahalanobis linear distance (MLD) classifier combined with a decision tree classifier. In addition, a model adaptation method was employed for decoding MI-EEG activity [32]. Handiru *et al.* proposed an iterative multiobjective optimization (IMOCS) for channel selection to reduce the high dimensionality of the EEG features [28]. Moreover, a reference candidate solution is initialized and subsequently finding a set of the most relevant channels in an iterative process is carried out. In addition, several other dimension reduction and channel selection algorithms are used in their study to achieve 0.61 classification accuracy with the aid of the heavy feature engineering performed before training the classifiers. They have reported 0.80 classification accuracy when the 35 best-performing subjects were selected from the total 109 subjects of the same public dataset used in this proposed method [22]. Morash *et al.* recruited 8 untrained subjects to develop their own EEG dataset [50]. Naïve Bayesian classifier was employed to classify BCI signals from 29 EEG channels with the frequency range of 1-40 Hz. They have selected for each subject the best EEG features in terms of the largest Bhattacharyya distances, and in terms of the best quantity (the number between 1 and 16 that produced the best training set results). The average classification accuracy achieved by the classifier was 0.56.

In a study by Weibo *et al.*, a total of 10 trained subjects have been employed to develop a support vector machine classifier that used event-related spectral perturbation (ERSP), power spectral entropy (PSE) and spatial distribution coefficient which achieved mean accuracy of 0.70 [77]. Another study reported the use of feature selection from the spectral power estimation computed in individualized frequency bands. The features are chosen by a criterion based on Mutual Information. A multinomial logistic regression classifier is employed and reported to achieve 0.75 classification accuracy on 8 trained subjects [40]. Lei *et al.* proposed an empirical Bayesian linear discriminant analysis (BLDA), in which the neurophysiological and experimental priors are considered simultaneously in order to reduce and simplify the feature selection of the EEG features. BLDA showed superior performance over the linear discriminant analysis (LDA) and SVM classifiers achieving 0.77 classification accuracy on 7 trained subjects [42].

Deep learning models have only recently been applied to BCI systems. There has been a developing interest in the utilization of deep learning techniques over the past few years to employ deep convolutional neural networks to MI-EEG classifications [1], [62], [66], [72] and a less number of studies used Recurrent Neural Networks (RNN) or Long Short Term Memory (LSTM) [44], [45].

DL models suffer less from the curse of dimensionality in comparison to the traditional shallow machine learning models. This encouraged several researchers to use DL and avoid the complex feature selection methods that are needed with the traditional machine learning models. One such deep learning approach is reported to use all the EEG channels covering the scalp of 12 subjects with 64 electrodes and a wide frequency range of 1-80 Hz [45]. The study proposed a recurrent neural network based classifier for encoding spatial and temporal sequential raw data with bidirectional Long Short Term Memory (bi-LSTM). The classifier resulted with 0.68 classification accuracy and showed superior performance in comparison to the standard LSTM method.

Similarly, another study used all the 64 EEG channels surrounding the scalp of 9 subjects [1]. They proposed a CNN model where CSP is utilized to discriminate interclass data and employing Fast Fourier Transform Energy Map (FFTEM) for feature selection and mapping of 1D data into 2D data (energy maps). Another interesting approach was reported by using the deep architecture of hierarchical semi-supervised extreme learning machine (HSS-ELM). The classifier used a semi-supervised ELM (SS-ELM) algorithm to classify the EEG signals, which could exploit the information from both labeled and unlabeled data [66]. The classifier achieved 0.67 average classification accuracy on 9 trained subjects. Sturm *et al.* proposed DL model with Layerwise Relevance Propagation (LRP) [71]. They reported classification accuracy of 0.75 which is comparable to those of CSP-LDA classifiers on trained subjects. They have used 58 EEG channels to collect and test EEG signals from 10 trained subjects.

Tabar *et al.* employed time-frequency maps from STFT as preprocessing, and by using CNN with stacked auto-encoders (SAE) they could achieve 0.75 average classification accuracy by training and testing on 9 trained subjects [72]. An approach with shallow CNN layers has been proposed by Schirrmeister *et al.* [64]. They compared the shallow CNN approach with only 2 layers versus deep CNN with 5 layers and ResNet with 31 layers. The shallow CNN outperformed all the other approaches by a few percent achieving 0.74 average accuracy with the public dataset of BCI Competition IV-2a, which employed 9 trained subjects (EEG frequency range 0-38 Hz). Later, after combining the public dataset with the author's private EEG dataset (20 trained subjects; freq.: 0-125 Hz; 1000 trials per subject), the average accuracy of the classifier increased to 0.84. The study concluded that high gamma waves (40-125 Hz) was encoding useful information for BCI, since the frequency range of 60 to 100 Hz are typically increased during movement execution and may contain useful movement-related information [17], [27], [61]. In light of these studies, including the gamma waves as part of the extracted features can potentially enhance the performance of any classifier. However in the literature, gamma waves were usually exploited with deep learning models more than the classical ML models. There is a tradeoff between the small possible performance enhancement of the gamma wave features and the detrimental effect of the increase of the number of features if gamma waves are included without increasing the training data set size. The classical machine learning models are much more sensitive to the issue of the curse of dimensionality in comparison to DL models. Increasing the feature dimensionality by incorporating gamma waves to the training set will have detrimental effect on the performance, unless accompanied by increasing the training data. Therefore only recently, with the advent of DL, the addition of gamma waves to the training set was more frequently exploited and attaining classification performance gain without preparing a large training set or increasing its size, which usually needs a substantial effort and cost.

Lawhern *et al.* proposed a one dimensional CNN arranged into 2 convolutional blocks to classify MI-EEG signals [41]. In the experiments, they have used 9 trained subjects from the public dataset of BCI Competition IV-2a with 22 EEG channels that covered the whole scalp of the subjects. They employed depthwise convolution to reduce the number of trainable parameters and reported 0.68 classification accuracy.

Finally, Luo *et al.* applied a deep RNN with a sliding window cropping strategy (SWCS) to classify MI-EEG signals [44]. Frequency features are extracted by the filter bank common spatial pattern (FB-CSP) algorithm and cropped by the SWCS into time slices. The feedback of the hidden layers has been processed by back-propagation through time (BPTT) algorithm [46]. However, the BPTT algorithm was extremely sensitive and the error flow tended to vanish, especially at the onset of the training phase. Hence, to overcome the vanishing gradient problem, a Long short-term memory (LSTM) unit [29] and a Gated recurrent unit (GRU) [16] were proposed. They found that CNN and SVM outperformed RNN in some subjects with high level (over 0.60) accuracies. However, in subjects with low-level (below 0.60) accuracies, RNN outperformed CNN and SVM.

To the best of our knowledge, deep metric learning with all its kind of varieties (Siamese Networks [38], Triplet Networks [30], few shot metric learning and prototypical networks [68], etc.) have not been applied to

MI-EEG classification yet. Despite that Deep Metric Learning models showed state-of-the-art results recently in a few deep learning applications like face identification and verification [36], [80].

In this work, we are presenting a deep metric learning approach to classify MI-EEG signals on untrained subjects. While most of previous BCI studies applied their proposed method on trained individuals who were exposed to BCI training before data collection, we have trained and tested our model on a large dataset of 109 untrained subjects. Trials of 64 channels EEG have been converted to time-frequency representation using Stockwell Transform [75]. It has an advantage over the other commonly used approaches in BCI research, like the Short Time Fourier Transform (STFT) [21]. Stockwell Transform is the only one preserving both the phase and magnitude information, which is considered as a lossless transform that is able to retrieve the original time domain signal without any phase distortion. This type of transform yielded better performance with our approach, although no BCI studies that we are aware of used Stockwell transform. While our current approach uses the signal amplitudes only, future work can be performed to extend the current study to involve the phase information as well. Loboda *et. al* investigated the use of phase values to classify MI-EEG signals for BCI applications with promising accuracy results (72%) [43]. Incorporating both EEG signal amplitude and phase is an exciting opportunity to combine two different signal information which most likely will boost accuracy more than if we consider each one alone. Beside this advantage of Stockwell Transform that allows to combine phase and amplitude information in one classifier (which has not been investigated in the literature yet), there is another property that makes it a better candidate than most other commonly used wavelet Transform methods. With Stockwell Transform, the window function is proportional to the frequency, which makes Stockwell Transform perform better in frequency domain analysis when the frequency of the input signal is low. And when the frequency input is high, Stockwell Transform has better clarity in the time domain when compared to Gabor Transform [60], Morlet Transform [14] and many other types of wavelet transforms. Stockwell Transform yielded better performance with our approach, although no BCI studies that we are aware of used Stockwell Transform, which is better known in the field of seismic signals analysis.

The time-frequency representation images are used for training and testing our triplet model in a similar approach like an image classification task. Triplet Network is trained using similarity based approach, where each training image is paired with another image from the same class (positive instance image), and another image belongs to different class (negative instance image). Triplet network encodes each of the 3 images using three CNN encoders with shared parameters to extract an embedding feature for each image. The training process aims to minimize the Euclidean distance of the embeddings of similar class images and increase the distance of different classes in the embedding feature space. Finally, in the prediction phase, the embedding features of the EEG spectrum images of the validation set are classified using the Nearest Neighbor classifier to estimate the nearest class in the embedding feature space.

In summary, the main contributions of this study are as follows:

- We show for the first time that DML classifier is able to converge and classify EEG signals with a small number of training samples (∼ 120 EEG trials) for only one subject per model, mitigating the well known issue of the large inter-individual variability of human MI-BCI EEG.
- Stockwell Transform has been utilized for the first time as a preprocessing step for classifying MI-BCI signals, which resulted with improved DML classification accuracy in comparison to the Short Term Fourier Transform.

Thus, the methods and findings described in this study are a first step to encourage the utilization of metric learning in BCI applications in particular or in any other EEG signal classification problem in general, especially when the training samples are extremely limited.

II. METHODS

A. DATASET

The EEG dataset used from [22] consists of 64 channel EEG according to 10-10 Electrode placement system [73], recorded from 109 subjects. Each subject performed imagination to move the right or left hand if a target on the right or left side of the screen appeared, respectively. There are trials where no target appeared and have been annotated as *rest*. The imagination to open and close the fist was performed until the target disappeared from the screen. Each trial lasted for 4 seconds, which was followed by a short duration of no activity. On average 150 EEG trials have been obtained from each user with roughly equal distribution of the left, right or rest labels.

B. APPROACH OVERVIEW

The aim of the model is to classify the EEG signals to detect the imagination of the user and label the trials into one of the three labels: *left, right* or *rest* classes. For each EEG channel, the trial segments have been converted into the frequency domain representing the EEG power for a certain frequency range plotted over time, yielding 64 images per EEG trial. The 64 spectrograms have been plotted on one larger blank image, each spectrogram image on its respective EEG placement according to the 10-10 EEG electrodes placement system to create a multi-spectral topographical plot for each trial, and used the plotted image for training and testing the deep metric learning (DML) model. The traditional method is to use the whole dataset to train and test a single model. However, we show that 10% higher accuracy can be achieved with our proposed method of training a single dedicated model for each user.

C. EEG SEGMENTATION

Epochs of 5 seconds have been segmented, which consist of 1 second before starting an event of imagery trial and lasting for 4 seconds. Fig. 1 shows 22 epochs of the concatenated C3 EEG channel series, which demonstrates voltage changes due to the imagination of moving the left hand of subject 22. The negative trend happens just before 1 second after the appearance of the target on the screen.

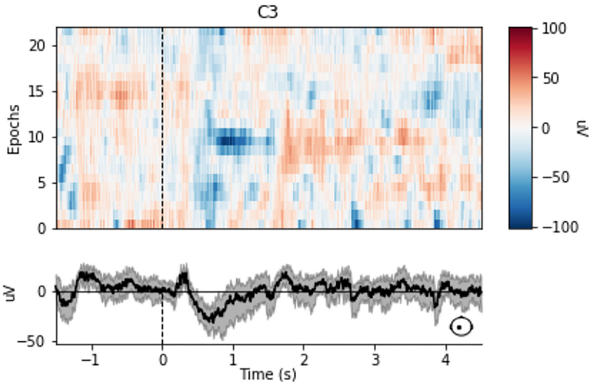

FIGURE 1. C3 EEG channel epochs of subject 22 showing voltage changes due to the imagination of moving the subject's left hand. The dotted line represents the time when the left target appeared on the screen.

EEG signals are known for low spatial resolution [69] and suffer from channel crosstalk [73]. Typically, EEG electrodes pick up signals from nearby areas within several centimeters [56]. Hence, we have referenced the raw EEG signals to the Common Average Reference (CAR), which is a common method for enhancing spatial resolution [76]. Concretely, the average of all equally spaced EEG channels covering the entire head, where the potentials of the EEG channels are generated by point sources inside the head, equals zero. However, the assumptions of point sources and complete coverage of the head are usually not met; but, CAR is still a good approximation for good spatial filtering that yields almost reference-free EEG signals [48]. Since it emphasizes components that are present in a large number of channels, CAR reduces such components and effectively acts as a high pass spatial filter.

D. TIME FREQUENCY REPRESENTATION

Each epoch with the 64 raw EEG data was transformed into the frequency domain, generating 64 spectrogram images. Commonly, the frequency of interest in BCI research is the *mu* (8-12 Hz) and the *beta rhythms* (18-25 Hz) [49]. However, we have seen better performance in using a range of frequencies from 2-78 Hz (y-log scaled) in the preprocessing step for the time frequency representation (TFR) of the EEG trials. One of the advantages of deep learning models in comparison to the classical machine learning techniques is the superior feature extraction ability, enabling them automatically learn complex features in an end-to-end learning system [12]. We found that without picking the frequency of interest, and let the DML model converge its trainable parameters to emphasize and diminish weights of features by its own training procedure, and incorporating larger frequency range improved the accuracy of the classification by 5%, which suggests that contrary to the common belief [49], there are useful features in the higher frequency range. Further study is needed to investigate the physiological basis for this observation.

A baseline correction was applied to all the EEG trials to increase the signal-to-noise ratio. The spectral subtraction method has been used, which is commonly used for background noise reduction in speech signals [13]. The stationary noise was estimated from the 1 second segment that preceded the screen target appearance and was subtracted from the whole epoch. Spectrum magnitudes were normalized by the logratio method in comparison to the 1 second segment baseline preceding the motor imagery, where the spectrum power was divided by the mean of the baseline power and the log of the result was taken. The upper and lower bound for the plot color range of the spectrograms are chosen to be +5 and −5 μV^2 to cover the whole range of the power spectrum of the dataset.

Stockwell Transform has been applied to transform the EEG signals from time domain to frequency domain. The discrete time Stockwell Transform is expressed by:

Let $\alpha = p\Delta_F$, $f = m\Delta_F$, $t = n\Delta_T$, where α is the width of the Gaussian window, Δ_F is the sampling frequency and Δ_T is the sampling interval; then:

$$S_x(n\Delta_T \, , m\Delta_F) = \sum_{p=0}^{N-1} X[(p+m)\,\Delta_F]\, e^{-\pi \frac{p^2}{m^2}}\, e^{\frac{j2pn}{N}} \qquad (1)$$

It has an important advantage over short-time Fourier transform (STFT), which is the implicit phase-normalized frequency bands, that makes the time information in the frequency domain distortion-free. Concretely, Stockwell Transform is known to be able to recover the input signal into time domain in a lossless way [70]. The width of the Gaussian window α tunes the tradeoff of the time and spectral resolution of the spectrogram. We found that 0.6 had the best tradeoff and gave the best performance. Fig. 2 shows an example of the Stockwell Transform of one trial with log-scaled frequency range of 2-78 Hz. Log-scaling resulted in a spectrogram that has more emphasis on the *mu* (8-12 Hz) and the *beta rhythms* (18-25 Hz) that are known for their correlation with the EEG changes correlated with motor imagery [49].

The 64 channels spectrograms have been plotted on one image for each trial. Since EEG channels suffer from crosstalk noise, and each electrode picks up signals from the nearby areas within few centimeters, we initially plotted the spectrograms according to the 10-10 EEG electrode placement system, which is the same placement that has been used in the data collection of the dataset. This produced a topogram image that preserved the location of channels (see Fig. 3). However, later we found that maximizing the spectrogram plots area and decreasing the background enhanced the performance of the classifier; therefore, we have adopted another

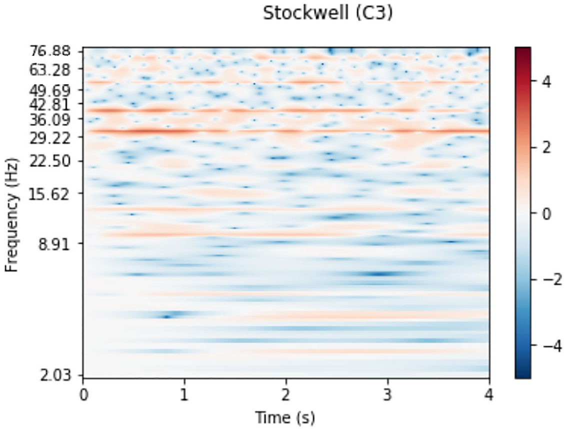

FIGURE 2. Stockwell power spectrogram of one trial (left hand movement imagination) log freq-scaled of subject 22 for EEG channel C3.

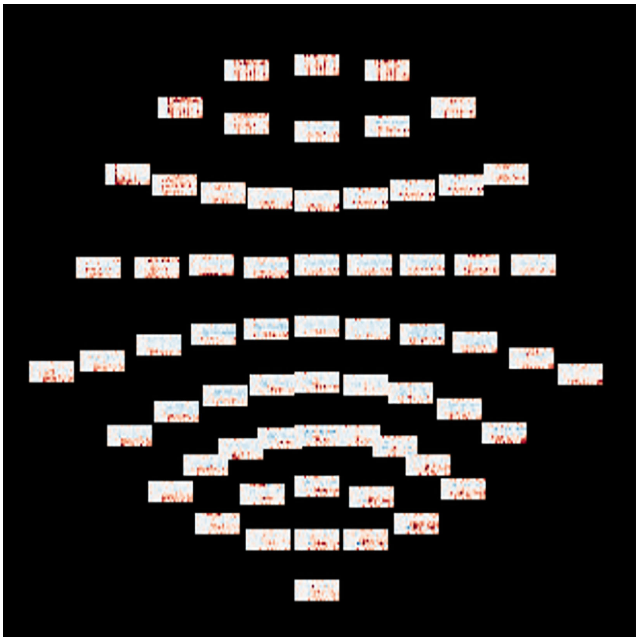

FIGURE 3. Stockwell power spectrogram of 64 channels placed according to the 10-10 EEG system placement for one trial of left hand movement imagination.

approach where we concatenated all the spectrograms on a 512 × 512 rectangular topogram area (see Fig. 4). The generated image dataset was normalized using the mean of pixel values and standard deviation.

E. DEEP METRIC MODEL ARCHITECTURE

The goal is to learn a metric feature space where two similar images correspond to two embedding feature vectors that are close together. We adopted the triplet network learning approach, which is an inspiration of Siamese network architecture [30].

A triplet neural network was trained using the triplets of inputs (x, x^+, x^-) which are the anchor image instance x, the positive instance image x^+ that is similar to x, and the negative instance image x^- that is different from x. The embedding function $f(.)$ that is learned by the network is directly optimizing the metric space such that the Euclidean distance (L2 norm) between the embedding function of the anchor image and the positive instance image is less than the difference between the Euclidean distance between the embedding function of the anchor image and the negative instance image:

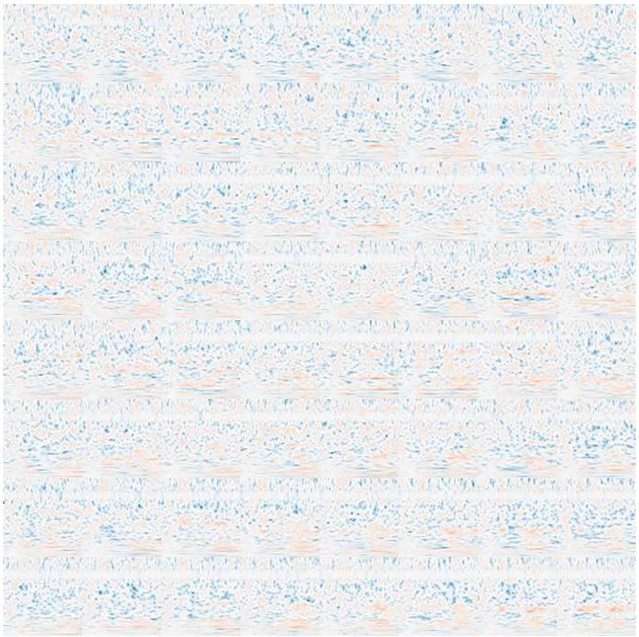

FIGURE 4. Example of a combined image generated from 64 channels of a single trial.

$$\|f(x) - f(x^+)\|_2 < \|f(x) - f(x^-)\|_2 \tag{2}$$

The Convolutional Neural Network (CNN) encoder after many epochs of training iterations will be converged to cluster similarly labeled images embedding as shown in Fig. 5. The Triplet Network has an advantage over the Siamese network approach, where it only optimizes the difference of embeddings between either the anchor and positive instances, or the anchor and negative instances embeddings. Hence, the Siamese networks are known for their sensitivity to calibration, where similarity vs. dissimilarity requires context. For instance, a person is regarded similar to another person in a dataset of random objects, while the two persons are deemed to be dissimilar in a dataset of individuals only. In Triplet Networks, such a calibration is not required [30].

Fig. 5 illustrates the general structure of the Triplet network where the triplet loss optimizes the ratio of the Euclidean distance between the anchor and the positive instance image embeddings $\Delta(a, p)$, and the distance between the anchor and the negative instance image embeddings $\Delta(a, p)$. The CNN encoders for the triplet backbones use shared parameters of the same architecture, which dramatically reduces the number of parameters that have to be learned. Parameter sharing speeds up the training process, as fewer gradients are needed to be computed at each iteration, and decreases the GPU memory requirements for training and inference. Moreover, the reduction of the model's parameter is a natural protection

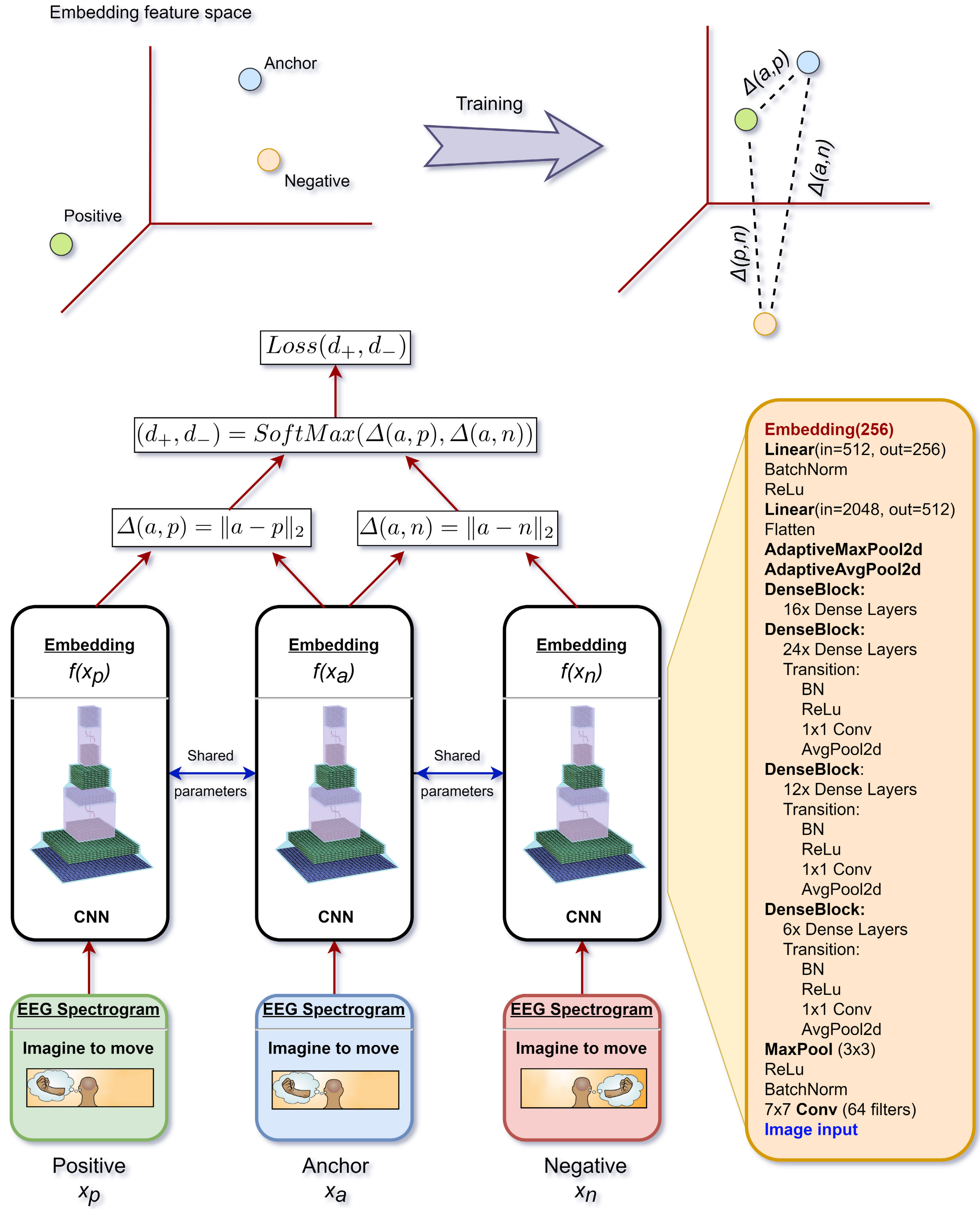

FIGURE 5. Triplet network training uses 3 kinds of images within each minibatch. The anchor image x_a, the similar positive instance image x_p, and the negative instance image x_n are encoded with 3 CNN encoders such that the training aims to reduce the Euclidean distance of the embeddings difference between anchor and positive instance image $\Delta(a, p)$, and increase the distance between the anchor and negative instance embeddings $\Delta(a, n)$.

against overfitting, without affecting the model's capacity for learning complex features [30]. Training optimizes for the following loss function:

$$Loss(d_+, d_-) = \Delta((d_+ - 0), (d_- - 1))^2 \tag{3}$$

which means the optimizing is minimizing the Mean Square Error (MSE) of the vector (d_+, d_-) compared to the vector $(0, 1)$. The aim is to train the model to make $\Delta(a, p)$ as close as possible to 0, while $\Delta(a, n)$ as large as possible. As such, in order to optimize this ratio, we applied SoftMax to both distances to get similarities that are bounded in the domain $[0, 1]$:

$$d_+ = \frac{e^{\Delta(a,p)}}{e^{\Delta(a,p)} + e^{\Delta(a,n)}} \tag{4}$$

$$d_- = \frac{e^{\Delta(a,n)}}{e^{\Delta(a,p)} + e^{\Delta(a,n)}} \tag{5}$$

However, it was observed that the network quickly learned an embedding feature space where d_- is close to 1, since most randomly chosen negative instance images are largely different from the anchor instance image. Thus, most of (a, n) pairs did not contribute to the gradients of the learning process, which led to underfitting where the network quickly stopped learning.

To solve this issue, SoftPN triplet loss function has been used which is inspired by the work of Balntas *et al.* [10]. The SoftPN loss replaces $\Delta(a, n)$ in Eq. 4 and Eq. 5 with $\min(\Delta(a, n), \Delta(p, n))$, so the optimization is trying to learn a metric space where both the anchor and the positive instance embeddings are as far as possible from the negative embedding. On the contrary, the original SoftMax ratio loss is only considering the anchor and the negative embedding distance.

The CNN encoder is comprised of 7 × 7 convolutional layer with 4 Dense blocks, followed by Adaptive Average Pool 2d layer, Adaptive Max Pool 2d layer and two fully connected linear layers. This is basically the DenseNet121 body [33] which has been employed for each CNN encoder of the DML model. The head of the DenseNet121 model has been modified to extract the embeddings and compute Softmax and the loss as shown in Fig. 5. The final embedding layer is a dense layer with 256 embedding features output. ReLu was chosen as the activation functions, which has the advantage of a less likelihood of gradient vanishing problem during training and faster training convergence [18].

In sum, the triplet network is optimizing the loss function in each iteration of the training. Therefore, in each iteration the training takes a small step to modify the weights of the CNN encoders in a way that decreases the loss. Moreover, the loss has been designed in such a way that it decreases if the Euclidean distance between the embeddings of the images with the same class decreases and the distance between the embeddings of different labeled images increases.

For the experiments, we used Intel Core i9-9900k 5.00 GHz with 8 Cores CPU and 64 GB RAM on Ubuntu 16.04. For deep learning, we used Nvidia GTX 1080-Ti GPU with 11 GB memory. DL model was implemented using Pytorch deep learning framework [57] on Python 3.6 with fastai library.

Triplets were generated on the fly while generating mini-batches of 7 triplets during training. In order to fit as many triplets in the GPU memory, we have switched the model into the half precision mode. The optimal learning rate of 1e-03 is chosen using a grid-search learning rate finder. Dynamically, changing the learning rate over iterations according to the one cycle policy [67] proved to give faster convergence rate. The learning rate is starting with $1/32^{th}$ maximum learning rate and reaches its maximum on around 30% of the total iterations and is going down again as shown in Fig. 6. Moreover, we have applied transfer learning to the model before training the CNN body from the DenseNet121 imageNet trained model, which proved to make the training even faster with less training epochs till the model converged.

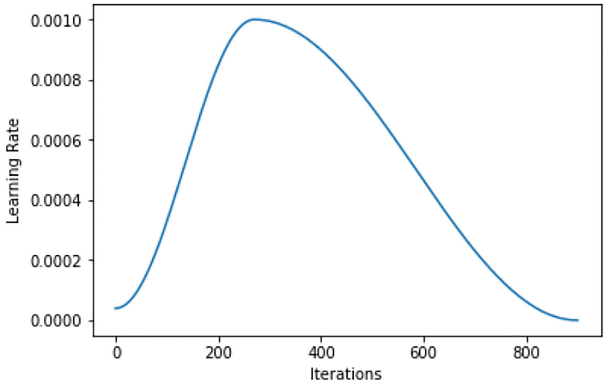

FIGURE 6. One cycle policy of learning rate.

The dataset has been divided into 80% training and 20% validation sets. Optimizing the model to find the best weight matrices and bias vectors is typically performed through an iterative gradient descent optimization. The computation of the gradient vector of the cost function (computing the deviation of the output from the desired output) with respect to each weight is performed using a backpropogation algorithm. The mathematical details and derivations are omitted here for brevity. For details we refer to [23] and [52]. Adam optimizer was used for this work, which is a variant of Stochastic Gradient Descent (SGD) [37].

The training has been carried out in 2 stages. In the first stage, the training was performed for 4 epochs while freezing all the convolutional and Dense Blocks, followed by the second stage, with a full model training for 48 epochs. Empirically, this did not yield better performance but made the convergence faster. Most likely, this was due to the linear dense layers being initialized with random weights, unlike the convolutional layers and Dense Blocks which had been initialized with pretrained imageNet weights [19].

Discriminative layer training with 3 learning rates has been performed in the second stage. The model was partitioned into 3 layer groups, where the CNN body partitioned into

2 layer groups, each with 2 Dense Blocks. The third layer group is comprised of the head of the model (linear dense layers with the Adaptive Average Pool 2d and Adaptive Max Pool 2d layers). The maximum learning rate max_{l_r} (that we have estimated by the learning rate finder method) on the third layer group equals to 1e-3; and $max_{l_r}/9$, $max_{l_r}/3$ was assigned to the first and second layer groups, respectively. Discriminative learning rates resulted in slightly better accuracy in comparison with one learning rate for all model's layers. This is likely because the imageNet pretrained weights of the first layer group have almost good weights for any dataset, since the lower convolutional layers are responsible for simple feature detection such as lines and edges which are the *most general* knowledge [78]. The second layer group of the model, which is responsible for extracting more complex features, is more different in our dataset from the imageNet, and higher learning rate is needed. Further, the third layer group, which has been initialized with random weights, needed the highest learning rate throughout the training. Gradient clipping of 0.1 has been applied in order to prevent exploding gradients.

Weight decay regularization has been applied and the best value has been estimated by grid search to be 0.1. Weight decay is an L2 type of regularization that is known to improve the generalization of deep learning models and decrease overfitting [39]. With this regularization term, the loss function is changed into:

$$\text{Loss}(w, x) = \text{DataLoss}(w, x) + \frac{1}{2}\ c\left\|w\right\|^2 \tag{6}$$

where w is the model weights, x is the mini-batch and c is the weight decay constant. The model weights update at each step during gradient descent would be:

$$w := w(1 - \alpha c) - \alpha \frac{d\,\text{DataLoss}(w, x)}{dw} \tag{7}$$

where α is the learning rate and $\frac{1}{2}\ c\left\|w\right\|^2$ is the L2 penalty term. Hence, the effect of the weight decay is to scale down the model's weight parameters and proportionally decay to zero. However, with the use of Batch Normalization (BN) [35], the effect of weight decay when used with BN is poorly understood [79]. BN is canceling out the scaling down effect of the weight decay, since Batch Normalization is basically *normalizing* the neural network outputs which makes them invariant to the scaling effect of the output of the previous layers. Yet, it is still a common method to regularize the network and prevent overfitting. Recent studies suggest that the effect of weight decay if used in combination with BN is not L2 regularization, but it is more likely preventing the decay of the effective learning rate over time. Therefore, with higher effective learning rate, weight decay results in better optimum generalization [79].

We found that training a dedicated model for each user performed better than one model for all users. This is due to the large inter-individual variations in terms of EEG correlation to motor imagery [24]. In a deployed BCI DL model, the user training can be performed once per user, and model parameters are saved as a user-specific profile and loaded before using the BCI for classification. After training, the prediction of the model was performed by generating the Stockwell spectrogram image for each epoch of the validation dataset, and feeding them in one CNN encoder to get the embedding features, which were then classified by Nearest Neighbor (NN) algorithm to find out to which class they belong. This is performed on the GPU using a custom parallelized solver in Pytorch, which made the NN search highly efficient. The aim of the NN algorithm is to search for the nearest embedding from the pool of the known labeled spectrogram embeddings. The embeddings of the known spectrogram images have been calculated and stored during the training process once the DML model training is completed. In the inference phase, the known embeddings are transferred to the GPU, and the Euclidean distance between these embeddings and the unknown item embedding are calculated efficiently by utilizing the parallel compute resources of the GPU. Finally, a sorting procedure is performed in ascending order from the nearest to the farthest embeddings, and the nearest known embedding spectrogram label is assigned as the prediction.

III. RESULTS

A. PERFORMANCE OF THE DML MODEL

The classification results of the validation data of 109 subjects are provided as a confusion matrix in Fig. 7, with classification accuracy of 0.647 (95% CI 0.624, 0.671).

The recall of the rest class is significantly better than the recall of the other two classes. Recall is the true positive rate, which is the ability of the model to correctly detect true positives, and hence it is also known as sensitivity:

$$\text{Recall} = \frac{tp}{tp + fn} \tag{8}$$

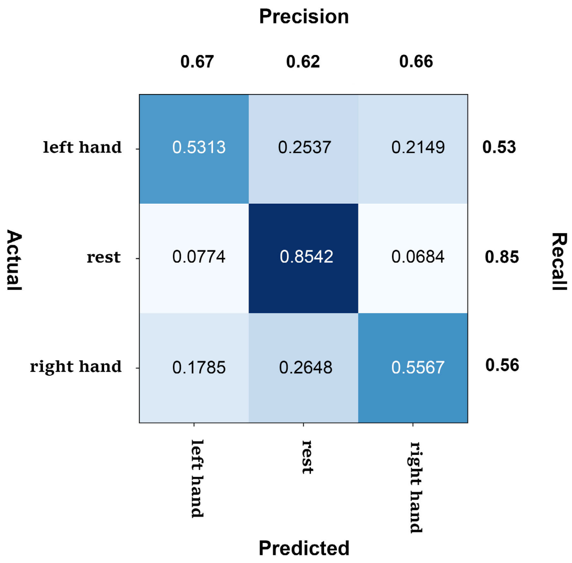

FIGURE 7. Normalized confusion matrix of the classification results of validation data.

While, recall represents the ability of the model to find all the relevant samples, precision represents the proportion of the samples that the model predicts as relevant was actually relevant. Precision (which is also known as positive predictive value) is defined as:

$$\text{Precision} = \frac{tp}{tp + fp} \tag{9}$$

where *tp* is the true positives which is the number of samples the model correctly labeled as positives and they were actually positives. False positive *fp* is the number of samples the model incorrectly labeled as positives, but actually they are negatives. False negative *fn* is the number of samples the model incorrectly labeled as negatives, but actually they are positives.

Intuitively, it makes sense that the *rest* class sensitivity is much better than the others, as this class has no motor imagery features when the subject is resting and not performing any task at all. So the rest spectrograms are more distinctive than the others. With real time simulation of muscles in online BCI systems, this is desirable. The model should not tend to trigger any output with action if the user is resting with no action.

Training the DML model took $\sim$ 30 minutes per each subject. In the inference, for each trial, preprocessing the EEG signals and plotting the 64 channel spectrogram images on 1 concatenated topogram took $\sim$ 360 ms on CPU. In addition, encoding from topogram image to embedding using the trained DML model on GPU took $\sim$ 8 ms. Finally, the NN algorithm on GPU took $\sim$ 10 μs.

The slowest stage in inference is the topogram image generation, which has been optimized to run on 16 logical cores of the Intel core i9-9900k CPU. The process can be accelerated further by utilizing a CPU with additional cores. Up to 64 CPU cores can be exploited to compute the spectrogram images of the 64 EEG channels in parallel. This can give $\sim$ 4x speedup over our reported runtime. Such improvement in inference time is important for a real time BCI system, where the total delay of Motor Imagery translation into control action preferably should be as small as possible for an optimal user experience.

B. PERFORMANCE OF TFR METHODS

The Stockwell Transform has been selected for the time frequency representation due to its phase-normalization property for the frequency bands, which makes the time information in the frequency domain distortion free [70]. Moreover, the Stockwell transform consistently performed better than Short Time Fourier Transform (STFT) in almost all subjects. The resulted average accuracy of the STFT method for all 109 subjects was 0.431 (95% CI 0.412, 0.451). Fig. 8 shows a comparison between the performance of STFT and Stockwell Transform.

The difference between the mean accuracy of the two groups is highly significant with over 0.2 difference ($p < 0.0001$). Table 1 shows the two-sample T-Test between the means.

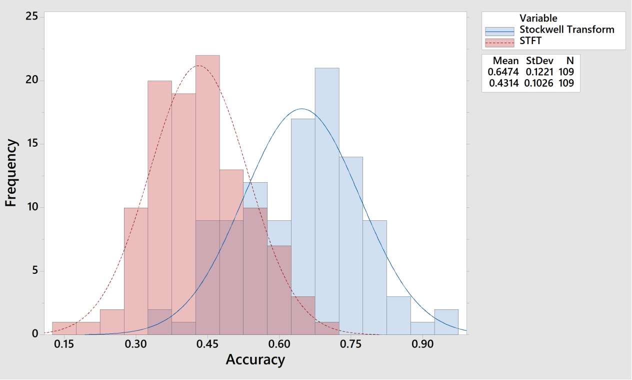

FIGURE 8. Performance comparison of the DML model with the Stockwell Transform and STFT.

TABLE 1. Two-sample T-Test assuming unequal variances for the performance of Stockwell Transform and STFT.

	N	Mean	StDev	SE Mean
Stockwell Transform	109	0.647	0.122	0.012
STFT	109	0.431	0.103	0.0098

Difference = μ (Stockwell Transform) - μ (STFT)
Estimate for difference: 0.2160
99.9% CI for difference: (0.1650, 0.2670)
T-Test of difference: T-Value = 14.14 P-Value < 0.0001 DF = 209

TABLE 2. Classification accuracies of stacking refinements one by one. If accuracy is not improved, the refinement is dropped and the next refinement is stacked with the best previous model. ΔAcc is the accuracy difference from the best previous model.

Refinements	ΔAcc	Acc
DenseNet169 + Adam optimizer		0.521
+ concatenated spectrograms	0.046	0.567
+ gradient clipping 0.1	0.035	0.602
+ rectified Adam optimizer	-0.115	0.487
+ grad clipping 1.0	-0.256	0.346
+ WD 0.1	0.018	0.620
+ DenseNet121	0.027	**0.647**
+ STFT	-0.216	0.431

C. ABLATION STUDY

Table 2 summarizes the collection of refinements that have been examined empirically throughout the study. Each refinement was added to the previous model settings. If the refinement enhanced the performance, it was kept in all next steps, unless the same refinement was changed. If the refinement effect was detrimental, we dropped the refinement and continued to apply the next refinement to the best previous model. Eventually, the best performance was performed by: DenseNet121, Adam optimizer, gradient clipping 0.1, WD 0.1 and concatenated spectrograms.

D. ACCURACY VERSUS SUBJECTS COUNT

Fig. 9 demonstrates the relationship of the accuracy performance of the DML model with the number of subjects

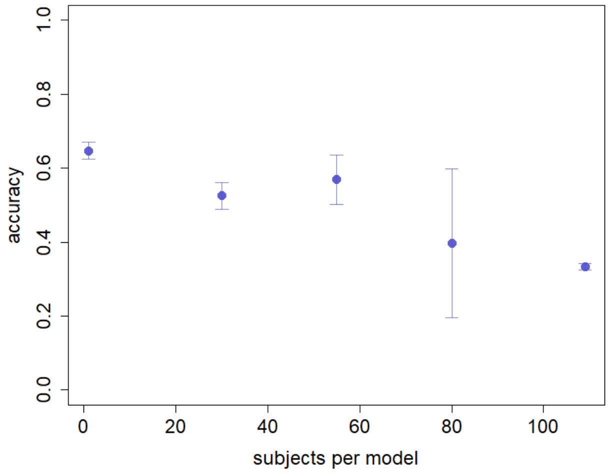

FIGURE 9. **Accuracy performance of DML model versus the number of subjects involved in training the model.**

involved in training the model. The experiments were carried out with 5 data points: 1, 30, 55, 80 and 109 subjects were involved in the training of a single DML model. In order to estimate the confidence interval, the experiments have been repeated 4 times. With each repetition, the subjects chosen were different or in the case of the 55 and 80 subjects, the chosen training sets were minimally overlapped. The 109 subjects point used the entire dataset. The figure shows an almost monotonic inverse relationship between the classification accuracy and the number of subjects involved in the training set. The best performing DML model was that with the single subject per model. This is clearly suggestive that the model could exploit the inherent ability of deep metric learning on converging with very small training data. In the 1 subject per model data point, the classifier had mitigated the issue of the inter-individual MI-EEG variability between users (which had negatively affected the performance of the classifier at more than 1 subject per model), despite the fact that the model used a limited training set of only $\sim$ 120 trials per model.

IV. DISCUSSION

Most previous BCI studies have been performed on trained subjects, who were exposed to training on modulating their EEG signals before the EEG data collection. It is known that motor imagery BCIs do not work well during the first session, and some training is necessary. While subjects differ in the performance and training time needed, most subjects perform better after 1-4 hours of training time [24].

The EEG dataset used for this study is screening data, in which the subjects had not been exposed to BCI training before collecting the data [22]. So the performance is expected to be lower than data collected from trained subjects. However, it is assumed that subjects who performed greater than 60% of accuracy could use BCI effectively for non-critical control maneuver tasks, like moving a mouse cursor on a screen [26].

To further inspect the efficiency of the proposed model, it was compared with various previous methods used for motor imagery BCI studies. Table 3 shows a comparative performance analysis of this approach with the previous studies. Studies that used EEG data from untrained subjects are listed in bold fonts. Numerous methods have been reported with comparable performance. However, studies that employed a large number of users are relatively rare. A large dataset of 109 subjects that is employed for this study and [28], ensures that the sample is more representative of the performance of the entire population, and will be more inclusive for outliers.

For comparison with the state-of-the-art methods of the previous literature that used the same dataset as the one utilized in this study, the results of two studies are reported herein; one that used classical machine learning approach and another with a deep learning approach. Handiru *et al.* used the Physionet public dataset [22] of 109 subjects to train Support Vector machine (SVM) classifier [28]. An iterative multi-objective optimization for channel selection (IMOCS) is employed to reduce the high dimensionality of EEG features. SVM classifier could achieve 0.61 classification accuracy when tested on all the 109 subjects. While the reported accuracy of the SVM classifier is impressive, the DML classifier outperformed the SVM by a small margin. Moreover, the researchers needed to tune the EEG channel selection algorithm for each subject, while our approach with the DML model is an end to end algorithm on all the 64 EEG channels without the need of feature selection. This approach was possible due to the advantage of deep learning over the classical machine learning methods, since DL suffers much less from the curse of dimensionality [23].

In another study, X. Ma *et al.* used a deep learning approach to classify MI-EEG signals of subjects from the Physionet public dataset [45]. Interestingly, the approach used a simpler preprocessing pipeline where the EEG signals are used in the time domain in the form of raw signal values without transforming the epochs to the frequency domain. Similar to our study, all the 64 EEG channels were used to train the classifier. The bidirectional Long Short Term Memory (bi-LSTM) classifier was employed which was able to achieve 0.68 accuracy, which is higher than our reported accuracy by $\sim$ 0.03. However, they have selected only 12 subjects from the dataset. In our observation, there are several subjects who performed very poorly in our experiments among the 109 subjects, and had significant detrimental effects on our reported accuracy. Many BCI studies reported that around 10-30% of subjects could not modulate their EEG for BCI control [3], [4], [6], [11], [54], [55], [58], [59], [63], [65], [74]. This has been called BCI illiteracy [55], [63]. BCI illiteracy is even worse in motor imagery BCI systems in comparison to other BCI systems based on P300 or steady state visually evoked potentials (SSVEP) [2], [3], [5], [26], [53]. To compare the model, X. Ma *et al.* additionally ran two baseline classical machine learning models experiments. They concluded that the proposed deep learning method outperformed

TABLE 3. Summary of related BCI studies. Methods including the proposed approach performed on untrained subjects including the proposed approach are listed in bold.

Study	DL/ML	Classifier	Subjects	Training	Electrodes	Freq (Hz)	Preprocessing	Accuracy
[32]	**ML**	**MLDC**	**3**	**untrained**	**27**	**0.1-100**	**WSD**	**0.57**
[28]	**ML**	**SVM**	**109**	**untrained**	**16**	**4-40**	**FB-CSP**	**0.61**
[50]	**ML**	**BC**	**8**	**untrained**	**29**	**1-40**	**DWT**	**0.56**
Proposed	**DL**	**TN**	**109**	**untrained**	**64**	**2-78**	**ST**	**0.65**
[45]	**DL**	**RNN-LSTM**	**12**	**untrained**	**64**	**1-80**	**SV**	**0.68**
[1]	DL	CNN (1 conv, 1 FC)	9	trained	64	8-30	FFM	0.61
[42]	ML	BLDA	3	trained	129	1-50	CSP	0.63
[66]	DL	MLP (3 hidden)	9	trained	22	8-30	CSP	0.67
[41]	DL	1d-CNN	9	trained	22	4-40	SV	0.68
[77]	ML	SVM	10	trained	3	1-35	MCSP	0.70
[62]	DL	CNN (4 conv)	9	trained	8	4-40	CW	0.71
[40]	ML	LR	8	trained	16	5-35	BPF	0.75
[71]	DL	MLP (2 hidden)	10	trained	58	9-13	SV	0.75
[72]	DL	CNN (1 conv, 6 FC)	9	trained	3	6-30	FFM	0.75
[51]	ML	LDA	23	trained	3	8-30	BPF	0.77
[44]	DL	RNN	7	trained	22	8-30	CSP	0.77
[64]	DL	CNN	29	trained	22	0-125	SV	0.84

Classifier type: MLDC (Mahalanobis Linear Discrimination classifier), GA (Genetic Algorithm), SVM (Support Vector Machine), BC (Bayesian classifier), RNN (Recurrent Neural Network), LSTM (Long Short Term Memory), CNN (Convolutional Neural Network), BLDA (Bayesian Linear Discriminant Analysis), LDA (Linear Discriminant Analysis), MLP (Multi-Layer Perceptron), LR (Logisic Regression), TN (Triplet Network).
DL/ML: Deep Learning or Classical Machine Learning method.
Trained/untrained: Subjects trained or untrained on BCI before experiments.
Preprocessing: WSD (Welch's Spectral Density estimation), CSP (Common Spatial Pattern), FB-CSP (Filter Bank CSP), MCSP (Multi Class CSP), DWT (Discrete Wavelet Transform), BPF (Band Pass Filter Power Features), SV (Signal Values in Time Domain), CW (Channel Wise approach), FFM (Fast Fourier Map), ST (Stockwell Transform).

the classical machine learning methods, namely CSP+LDA that achieved 0.59 classification accuracy and FBCSP+LDA with 0.60 accuracy.

In future research, it would be interesting to combine all the proposed deep learning models in the studies in Table 3 in one multi-modal ensemble network. The different methods combined will likely improve generalization and increase accuracy [81].

BCI systems aim to generate Central Nervous System (CNS) outputs that are fundamentally different from the normal CNS output [76]. The natural CNS output originates from the cooperation of several parts of the CNS, starting from the cerebral cortex to the spinal cord. No single place alone is responsible for the CNS output. However, BCI systems are controlled from the cerebral cortex only. This is due to the technical limit of EEG systems in picking up electrical signals from the deeper structures of the brain. For instance, walking is performed by the harmonic collaboration of the cerebral cortex, basal ganglia, thalamic nuclei, cerebellum, brainstem and spinal interneurons through the efferent motor neurons. While it is true that the cerebral cortex is where the initiation of the movement of the limbs starts, the high speed rhythmic sensorimotor neurons are essential for a stable and robust locomotion [25], [34], [47]. Furthermore, although that from trial to trial, the activity of one CNS region involved in the motor control may vary substantially, the collaboration and interaction among all the regions involved ensures that the muscular action by itself is very robust across trials [76]. However, BCIs, which are dependent on one CNS region, require a unique task from the user that has not been adapted throughout development.

Eventually, with the current EEG signal pickup technology, imperfect accuracy is inevitable in all types of BCI models. This is further complicated by the extremely weak and noisy EEG signals. This noise has the same statistical distribution as the signal of interest, which comes from the brain itself. While the baseline correction method that has been adopted in this project could resolve this issue to some extent, the stochastic nature of the EEG signals made the reduction of such noise a challenge. Moreover, the event related potential modulation is largely dependent on the attention and the mental fatigue of the user during the BCI trial. Some trials will be misclassified, no matter what, because the user did not pay enough attention, or because the event related potential is so small that it is buried in the background noise of the brain's natural activity.

The evidence to date shows that the adaptation of CNS for BCI control through direct output from the cerebral cortex is possible; however, it is still imperfect. We opine that, unless new techniques for picking up electro-physiological signals of the deep structures of the brain are developed, robust BCI systems that are on par with the natural coordination of the natural muscular movements of the human body is out of reach with the current BCI system approaches.

While robust BCI systems are desirable, it should not be perfectly accurate to be useful. There are BCI applications that do not require robust control with some room for inaccuracy, especially for cases where the normal muscular control of the body is not available. Completely paralyzed patients with locked-in syndrome could utilize such BCI for controlling a cursor on the screen.

V. CONCLUSION

In conclusion, we developed a triplet deep metric network for the classification of motor imagery BCI on a small training set of only ∼ 120 epochs per model. Our work highlights a novel model development technique that employs deep metric learning to compensate for a small dataset and may be utilized in future deep learning studies involving EEG signal classification. The complete process of the proposed method was presented in detail – including Stockwell Transform to convert the time domain of 64 channels EEG signals to the frequency domain which is, to the best of our knowledge, has not been adopted in MI-BCI systems before. Furthermore, baseline denoising correction, three convolutional encoders with shared parameters and a custom head have been employed. The aim is to train an embedding feature space, where spectrogram embedding features of similar classes are clustered near each other in the training phase. In the inference phase, one CNN encoder is used to compare the embedding features of the EEG spectrogram, and by using Nearest Neighbor classifier the class of the spectrogram is predicted. The BCI dataset used for this project is relatively large in comparison with the dataset used in most of the other BCI studies. Moreover, the 109 subjects involved are untrained and have not been exposed to any BCI training before. This makes the classification task quite challenging. DML classification with Stockwell Transform has achieved higher performance in comparison to DML with Short Term Fourier Transform (0.647% vs. 0.431%). We showed that the DML classifier is able to converge with an extremely small number of training samples (∼ 120 EEG trials) for only one subject per model. The proposed preprocessing pipeline and the Triplet Network provide a promising method to classify MI-BCI spectrogram image classes from noisy EEG signals with much less training samples than the previous methods. For future study, we aim to combine multiple DL models into one ensemble multi-modal network, which will combine the strengths of the various approaches used for DL models in BCI systems, and hopefully generate better accuracy than each one method alone.

REFERENCES

[1] W. Abbas and N. A. Khan, "DeepMI: Deep learning for multiclass motor imagery classification," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 219–222.

[2] B. Allison, T. Luth, D. Valbuena, A. Teymourian, I. Volosyak, and A. Graser, "BCI demographics: How many (and what kinds of) people can use an SSVEP BCI?" *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 2, pp. 107–116, Apr. 2010.

[3] B. Z. Allison, C. Brunner, V. Kaiser, G. R. Müller-Putz, C. Neuper, and G. Pfurtscheller, "Toward a hybrid brain–computer interface based on imagined movement and visual attention," *J. Neural Eng.*, vol. 7, no. 2, Apr. 2010, Art. no. 026007.

[4] B. Z. Allison, D. J. McFarland, G. Schalk, S. D. Zheng, M. M. Jackson, and J. R. Wolpaw, "Towards an independent brain–computer interface using steady state visual evoked potentials," *Clin. Neurophysiol.*, vol. 119, no. 2, pp. 399–408, 2008.

[5] B. Z. Allison and C. Neuper, "Could anyone use a BCI?" in *Brain-Computer Interface*. London, U.K.: Springer, 2010, pp. 35–54.

[6] B. Z. Allison, E. W. Wolpaw, J. R. Wolpaw, "Brain–computer interface systems: Progress and prospects," *Expert Rev. Med. Devices*, vol. 4, no. 4, pp. 463–474, 2007.

[7] H. Alwasiti and M. Z. Yusoff, "Shredded control of drones via motor imagery brain computer interface," *Compusoft*, vol. 9, no. 3, pp. 3606–3610, 2020.

[8] H. H. Alwasiti, I. Aris, and A. Jantan, "Brain computer interface design and applications: Challenges and future," *World Appl. Sci. J.*, vol. 11, no. 7, pp. 819–825, 2010.

[9] H. H. Alwasiti, I. Aris, and A. Jantan, "Eeg activity in Muslim prayer: A pilot study," *Maejo Int. J. Sci. Technol.*, vol. 4, no. 3, pp. 496–511, 2010.

[10] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, "PN-Net: Conjoined triple deep network for learning local image descriptors," 2016, *arXiv:1601.05030*. [Online]. Available: https://arxiv.org/abs/1601.05030

[11] N. Birbaumer and L. G. Cohen, "Brain-computer interfaces: Communication and restoration of movement in paralysis," *J. Physiol.*, vol. 579, no. 3, pp. 621–636, Mar. 2007.

[12] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," 2016, *rXiv:1604.07316*. [Online]. Available: https://arxiv.org/abs/1604.07316

[13] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[14] R. Büssow, "An algorithm for the continuous Morlet wavelet transform," *Mech. Syst. Signal Process.*, vol. 21, no. 8, pp. 2970–2979, Nov. 2007.

[15] A. M. Chiarelli, P. Croce, A. Merla, and F. Zappasodi, "Deep learning for hybrid EEG-fNIRS brain–computer interface: Application to motor imagery classification," *J. Neural Eng.*, vol. 15, no. 3, Jun. 2018, Art. no. 036028.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2067–2075.

[17] N. Crone, "Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band," *Brain*, vol. 121, no. 12, pp. 2301–2315, Dec. 1998.

[18] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8609–8613.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[20] D. G. Stork, P. E. Hart, and R. O. Duda, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[21] L. Durak and O. Arikan, "Short-time Fourier transform: Two fundamental properties and an optimal implementation," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1231–1242, May 2003.

[22] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[24] B. Graimann, B. Z. Allison, and G. Pfurtscheller, *Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction*. Springer, 2010.

[25] P. A. Guertin and I. Steuer, "Key central pattern generators of the spinal cord," *J. Neurosci. Res.* , vol. 87, no. 11, 2399–2405, 2009.

[26] C. Guger, S. Daban, E. Sellers, C. Holzner, G. Krausz, R. Carabalona, F. Gramatica, and G. Edlinger, "How many people are able to control a P300-based brain–computer interface (BCI)?" *Neurosci. Lett.*, vol. 462, no. 1, pp. 94–98, Sep. 2009.

[27] J. Hammer, T. Pistohl, J. Fischer, P. Kršek, M. Tomášek, P. Marusič, A. Schulze-Bonhage, A. Aertsen, and T. Ball, "Predominance of movement speed over direction in neuronal population signals of motor cortex: Intracranial EEG data and a simple explanatory model," *Cerebral Cortex*, vol. 26, no. 6, pp. 2863–2881, Jun. 2016.

[28] V. S. Handiru and V. A. Prasad, "Optimized bi-objective EEG channel selection and cross-subject generalization with brain–computer interfaces," *IEEE Trans. Human-Machine Syst.*, vol. 46, no. 6, pp. 777–786, Dec. 2016.

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

[30] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.* Cham, Switzerland: Springer, 2015, pp. 84–92.

[31] S. Horiguchi, D. Ikami, and K. Aizawa, "Significance of softmax-based features in comparison to distance metric learning-based features," 2017, *arXiv:1712.10151*. [Online]. Available: https://arxiv.org/abs/1712.10151

[32] D. Huang, K. Qian, S. Oxenham, D.-Y. Fei, and O. Bai, "Event-related desynchronization/synchronization-based brain-computer interface towards volitional cursor control in a 2D center-out paradigm," in *Proc. IEEE Symp. Comput. Intell., Cognit. Algorithms, Mind, Brain (CCMB)*, Apr. 2011, pp. 1–8.

[33] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "DenseNet: Implementing efficient convnet descriptor pyramids," 2014, *arXiv:1404.1869*. [Online]. Available: https://arxiv.org/abs/1404.1869

[34] A. J. Ijspeert, "Central pattern generators for locomotion control in animals and robots: A review," *Neural Netw.*, vol. 21, no. 4, pp. 642–653, May 2008.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[36] P. Jacob, D. Picard, A. Histace, and E. Klein, "Metric learning with HORDE: High-order regularizer for deep embeddings," 2019, *arXiv:1908.02735*. [Online]. Available: https://arxiv.org/abs/1908.02735

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

[38] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, Lille, France, vol. 2, 2015, pp. 1–8.

[39] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, 1992, pp. 950–957.

[40] R. Kus, D. Valbuena, J. Zygierewicz, T. Malechka, A. Graeser, and P. Durka, "Asynchronous BCI based on motor imagery with automated calibration and neurofeedback training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 6, pp. 823–835, Nov. 2012.

[41] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[42] X. Lei, P. Yang, and D. Yao, "An empirical Bayesian framework for brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 6, pp. 521–529, Dec. 2009.

[43] A. Loboda, A. Margineanu, G. Rotariu, and A. Mihaela, "Discrimination of EEG-based motor imagery tasks by means of a simple phase information method," *Int. J. Adv. Res. Artif. Intell.*, vol. 3, no. 10, pp. 1–66, 2014.

[44] T.-J. Luo, C.-L. Zhou, and F. Chao, "Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network," *BMC Bioinf.*, vol. 19, no. 1, p. 344, Dec. 2018.

[45] X. Ma, S. Qiu, C. Du, J. Xing, and H. He, "Improving EEG-based motor imagery classification via spatial and temporal recurrent neural networks," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1903–1906.

[46] J. Mazumdar and R. G. Harley, "Recurrent neural networks trained with backpropagation through time algorithm to estimate nonlinear load harmonic currents," *IEEE Trans. Ind. Electron.*, vol. 55, no. 9, pp. 3484–3491, Sep. 2008.

[47] D. A. McCrea and I. A. Rybak, "Organization of mammalian locomotor rhythm and pattern generation," *Brain Res. Rev.*, vol. 57, no. 1, pp. 134–146, Jan. 2008.

[48] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw, "Spatial filter selection for EEG-based communication," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 3, pp. 386–394, Sep. 1997.

[49] D. J. McFarland, L. A. Miner, T. M. Vaughan, and J. R. Wolpaw, "Mu and beta rhythm topographies during motor imagery and actual movements," *Brain Topography*, vol. 12, no. 3, pp. 177–186, 2000.

[50] V. Morash, O. Bai, S. Furlani, P. Lin, and M. Hallett, "Classifying EEG signals preceding right hand, left hand, tongue, and right foot movements and motor imageries," *Clin. Neurophysiol.*, vol. 119, no. 11, pp. 2570–2578, Nov. 2008.

[51] C. Neuper, R. Scherer, S. Wriessnegger, and G. Pfurtscheller, "Motor imagery and action observation: Modulation of sensorimotor brain rhythms during mental control of a brain–computer interface," *Clin. Neurophysiol.*, vol. 120, no. 2, pp. 239–247, Feb. 2009.

[52] M. A. Nielsen, *Neural Networks and Deep Learning*, vol. 25. San Francisco, CA, USA: Determination Press, 2015.

[53] F. Nijboer, N. Birbaumer, and A. Kübler, "The influence of psychological state and motivation on brain–computer interface performance in patients with amyotrophic lateral sclerosis—A longitudinal study," *Frontiers Neurosci.*, vol. 4, p. 55, Jul. 2010.

[54] F. Nijboer, A. Furdea, I. Gunst, J. Mellinger, D. J. McFarland, N. Birbaumer, and A. Kübler, "An auditory brain–computer interface (BCI)," *J. Neurosci. Methods*, vol. 167, no. 1, pp. 43–50, 2008.

[55] A. Nijholt and D. Tan, "Brain-computer interfacing for intelligent systems," *IEEE Intell. Syst.*, vol. 23, no. 3, pp. 72–79, May 2008.

[56] P. L. Nunez and B. A. Cutillo, *Neocortical Dynamics and Human EEG Rhythms*. New York, NY, USA: Oxford Univ. Press, 1995.

[57] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[58] G. Pfurtscheller, R. Leeb, D. Friedman, and M. Slater, "Centrally controlled heart rate changes during mental practice in immersive virtual environment: A case study with a tetraplegic," *Int. J. Psychophysiol.*, vol. 68, no. 1, pp. 1–5, Apr. 2008.

[59] G. Pfurtscheller, G. R. Muller-Putz, A. Schlogl, B. Graimann, R. Scherer, R. Leeb, C. Brunner, C. Keinrath, F. Lee, G. Townsend, C. Vidaurre, and C. Neuper, "15 years of BCI research at Graz university of technology: Current projects," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 205–210, Jun. 2006.

[60] S. Qian and D. Chen, "Discrete Gabor transform," *IEEE Trans. Signal Process.*, vol. 41, no. 7, pp. 2429–2438, Jul. 1993.

[61] F. Quandt, C. Reichert, H. Hinrichs, H. J. Heinze, R. T. Knight, and J. W. Rieger, "Single trial discrimination of individual finger movements on one hand: A combined MEG and EEG study," *NeuroImage*, vol. 59, no. 4, pp. 3316–3324, Feb. 2012.

[62] S. Sakhavi, C. Guan, and S. Yan, "Parallel convolutional-linear neural network for motor imagery classification," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 2736–2740.

[63] C. Sannelli, M. Braun, M. Tangermann, and K. R. Müller, "Estimating noise and dimensionality in BCI data sets: Towards BCI illiteracy comprehension," in *Proc. Int. Brain-Comput. Interface Workshop Training Course*, 2008, pp. 26–31.

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

[65] E. W. Sellers, A. Kubler, and E. Donchin, "Brain–computer interface research at the university of South Florida cognitive psychophysiology laboratory: The P300 speller," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 221–224, Jun. 2006.

[66] Q. She, B. Hu, Z. Luo, T. Nguyen, and Y. Zhang, "A hierarchical semi-supervised extreme learning machine method for EEG recognition," *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 147–157, Jan. 2019.

[67] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1—learning rate, batch size, momentum, and weight decay," 2018, *arXiv:1803.09820*. [Online]. Available: https://arxiv.org/abs/1803.09820

[68] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4077–4087.

[69] R. Srinivasan, "Methods to improve the spatial resolution of EEG," *Int. J. Bioelectromagnetism*, vol. 1, no. 1, pp. 102–111, 1999.

[70] R. G. Stockwell, L. Mansinha, and R. P. Lowe, "Localization of the complex spectrum: The S transform," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 998–1001, Apr. 1996.

[71] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial EEG classification," *J. Neurosci. Methods*, vol. 274, pp. 141–145, Dec. 2016.

[72] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, Feb. 2017, Art. no. 016003.

[73] M. Teplan, "Fundamentals of EEG measurement," *Meas. Sci. Rev.*, vol. 2, no. 2, pp. 1–11, 2002.

[74] T. M. Vaughan, D. J. McFarland, G. Schalk, W. A. Sarnacki, D. J. Krusienski, E. W. Sellers, and J. R. Wolpaw, "The wadsworth BCI research and development program: At home with BCI," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 229–233, Jun. 2006.

[75] Y. Wang and J. Orchard, "Fast discrete orthonormal stockwell transform," *SIAM J. Sci. Comput.*, vol. 31, no. 5, pp. 4000–4012, Jan. 2009.

[76] J. Wolpaw and E. W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice*. New York, NY, USA: Oxford Univ. Press, 2012.

[77] W. Yi, S. Qiu, H. Qi, L. Zhang, B. Wan, and D. Ming, "EEG feature comparison and classification of simple and compound limb motor imagery," *J. NeuroEng. Rehabil.*, vol. 10, no. 1, p. 106, 2013.

[78] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[79] G. Zhang, C. Wang, B. Xu, and R. Grosse, "Three mechanisms of weight decay regularization," 2018, *arXiv:1810.12281*. [Online]. Available: https://arxiv.org/abs/1810.12281

[80] X. Zhou, K. Jin, M. Xu, and G. Guo, "Learning deep compact similarity metric for kinship verification from face images," *Inf. Fusion*, vol. 48, pp. 84–94, Aug. 2019.

[81] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, nos. 1–2, pp. 239–263, May 2002.

HAIDER ALWASITI (Member, IEEE) graduated from the College of Medicine, Al-Mustansiriyah University, in 2001. He received the M.Sc. degree in biomedical engineering from Universiti Putra Malaysia. He is currently pursuing the Ph.D. degree in electrical and electronics engineering from Universiti Teknologi PETRONAS. His research interests include brain–computer interface systems, deep neural networks in medical applications, natural language processing, and self-driving cars. In his free time, he enjoys studying theoretical physics, running, traveling, and spending time with his family.

MOHD ZUKI YUSOFF (Member, IEEE) received the B.Sc. degree in electrical engineering from Syracuse University, in 1988, the M.Sc. degree in communications, networks, and software from the University of Surrey, in 2001, and the Ph.D. degree in electrical and electronic engineering from Universiti Teknologi PETRONAS (UTP), Malaysia, in 2010. He is currently an Associate Professor with UTP. He has international publications and holds patents. His research interests include transport safety and telecommunications. He is a member of the Tau Beta Pi and the Eta Kappa Nu.

KAMRAN RAZA (Senior Member, IEEE) received the B.E. degree in electrical engineering from the NED University of Engineering and Technology and the master's and Ph.D. degrees from Iqra University, Karachi, Pakistan. He has been associated with Iqra University as the Dean of the Faculty of Engineering, Sciences, and Technology. His research interests include intelligent control, signal processing, computer vision, and 3-D scene reconstruction. He is the author of over 45 peer-reviewed publications. He has more than 20 years of academic experience. He is a member of the Ph.D. Advisory Committee. He is also Pakistan's Higher Education Commission approved Ph.D. Supervisor. He is an approved Expert of Pakistan Engineering Council for evaluating Engineering Programs (relevant to Electrical, Electronic and Telecommunication Engineering disciplines) in Pakistan. He is also a Reviewer of various reputed international indexed journals. He has developed the Robotics Lab, the High Performance Computing Lab, the Industrial Automation Lab, the Telecommunication Laboratory, the Signal Processing Lab, the Wireless Communication Lab, and the Cisco Networking Lab in the Faculty of Engineering, Sciences, and Technology, Iqra University. He has supervised various funded projects and has conducted various professional training and tutorials.