IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 June 2022, accepted 27 July 2022, date of publication 1 August 2022, date of current version 8 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195513

METHODS

Subject-Independent Classification of Motor Imagery Tasks in EEG Using Multisubject Ensemble CNN

IRINA DOLZHIKOVA[1], BERDAKH ABIBULLAEV[2], (Senior Member, IEEE), REZA SAMENI[3], (Senior Member, IEEE), AND AMIN ZOLLANVARI[1], (Senior Member, IEEE)

[1]Electrical and Computer Engineering Department, School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan 010000, Kazakhstan
[2]Department of Robotics and Mechatronics, School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan 010000, Kazakhstan
[3]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA 30322, USA

Corresponding author: Amin Zollanvari (amin.zollanvari@nu.edu.kz)

This work was supported by the Nazarbayev University Faculty Development Competitive Research Grant under Award 021220FD1151.

ABSTRACT Subject-independent (SI) classification is a major area of investigation in Brain-Computer Interface (BCI) that aims to construct classifiers of users' mental states based on collected electroencephalogram (EEG) of independent subjects. Significant inter-subject variabilities in the EEG are among the most challenging issues in designing SI BCI systems. In this work, we propose and examine the utility of Multi-Subject Ensemble Convolutional Neural Network (MS-En-CNN) for SI classification of motor imagery (MI) tasks. The base classifiers used in MS-En-CNN have a fixed CNN architecture (referred to as DeepConvNet) that are trained using data collected from multiple subjects during the training process. In this regard, training subjects are divided into K-folds using which K base DeepConvNets are trained based on data from $K-1$ folds, whereas the hyperparameter optimization is performed using the held-out fold. We evaluate the performance of the MS-En-CNN on the large open-access MI dataset from the literature, which includes 54 participants and a total number of 21,600 trials. The result shows that the MS-En-CNN achieves the highest single-trial SI classification performance reported on this dataset. In particular, we obtained SI classification performances with average and median accuracies of 85.42% and 86.50% ($\pm$ 10.16%), respectively. This result exhibits a statistically significant improvement ($p < 0.001$) over the best previously reported result with an average and a median accuracy of 84.19% and 84.50% ($\pm$10.08%), respectively.

INDEX TERMS Brain-computer interface, deep learning, convolutional neural network, multi-subject ensemble.

I. INTRODUCTION

Electroencephalogram (EEG)-based Brain-Computer Interface (BCI) is a non-invasive technology that can be used to decode electrical activities of the brain and translate them into useful commands for human-computer interaction [1].

The motor imagery (MI) paradigm in BCI, which uses EEG data to identify the cerebral neural activities related to imagined limb movement [2], has been extensively studied in the literature. Decoding movement imagination from EEG is specially useful in various applications such as neurorehabilitation [3], [4], neuroprosthetics [5], and gaming [6]. However, the non-stationarity of EEG signals and presence of considerable variability in MI data pose significant challenges in the design of accurate BCI systems [7]. The non-stationarity of an EEG recording is an intrinsic property of the signal that can significantly vary from one record to another, even for a particular subject. Emotional and mental processes, as well as other cognitive and neurological factors, give rise to the intra- and inter-subject variability of the EEG [8]. Therefore, the EEG features learned from one subject do not necessarily correlate well with the features learned from another subject.

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

As a result, most of the existing BCI systems use subject-specific training schemes, which require extensive calibration of the system to individual users. Such a user-specific training process could be quite time-consuming and inconvenient, especially for people with disabilities, and can cause subject fatigue that affects the quality of the signal *per se* [9]. Therefore, the design of subject-independent (SI) BCI systems, which are calibration-free for new users, is of great interest.

First attempts to design calibration-free BCI systems used the idea of *transfer learning*, where the training data from the previous session of the same individual is used in the new session, removing the necessity of prior calibration per session [10]. An example of SI zero-training methods for event-related desynchronization based BCI was demonstrated by Fazli *et al.* [11], where the authors applied a combination of Linear Discriminant Analysis (LDA) and Common Spatial Pattern (CSP) filtering. Lotte *et al.* proposed the filter bank CSP feature extraction method based on multi-resolution frequency decomposition [12]. Ray *et al.* demonstrated another example of using CSP for SI BCI application with fused classification model [13]. SI training for detection of movement-related cortical potentials was performed by Niazi *et al.* [14].

Deep learning revolutionized many areas, including BCI, via efficiently exploiting the distributed features through multiple layers and demonstrating an improved performance compared with traditional methods [15], [16]. Convolutional Neural Network (CNN) belongs to a type of deep neural networks (DNNs) that has been widely used in wide range of applications, including character recognition [17], machinery fault diagnosis [18], [19], emotion classification [20], and many others. In the contexts of this research, the interest in CNNs is due to their ability to capture complex patterns from raw EEG data with efficient training time and high decoding accuracy [16], [21]–[23]. Deep CNN models such as DeepConvNet [24] and EEGNet [25] are examples of the well-known CNN-based architectures that demonstrated state-of-the-art performance in BCI applications. Other examples of the application of CNNs for MI classification problems include architectures with convolution across channels, time, and two-dimensional based kernels by Sakhavi *et al.* [26], and hybrid-scale CNN by Dai *et al.* [27]. Hybrid integration of CNN and other machine learning/deep models were also investigated: CNN and stacked autoencoders [28], [29], CNN and gated recurrent units [30], and CNN with the long-term, short-term memory networks [31]. Recently, one of the largest databases with 54 subjects for three BCI paradigms, including MI, has been collected by Lee *et al.* [32]. Several subject-independent methods have already been suggested and assessed based on the MI dataset collected therein. Kwon *et al.* have proposed a method based on spectral-spatial feature representation with deep CNNs [33]. Another research group took advantage of this large-scale dataset and presented the Convolutional Common Spatial Pattern Network (CCSPNet) [34]. Jeon *et al.* [35] proposed a framework for learning the class-relevant and subject-invariant feature representations. Although the proposed design allows the use of any decoding model as a feature extractor, the authors demonstrated promising results using EEGNet and DeepConvNet. The most recent work is by Zhang *et al.* [36], who achieved an average decoding accuracy of 84.19±10.08% by using the DeepConvNet architecture.

Ensemble systems can achieve better predictive performance than a single predictive model by reducing the variance of the less robust methods [37], [38]. In contrast to a single model that is prone to overfitting, an appropriate combination of predictions from several models potentially reduces the risk of overfitting and making incorrect decisions [38]. The ability of the ensemble to boost the performance of weak learners and its potential to deal with the extreme variability of the neural response data makes it an attractive approach for BCI applications. Thus, in this study, to further improve the performance of the CNN in SI classification of MI tasks, a *multi-subject* ensemble CNN model is designed; i.e., an ensemble of CNN classifiers where each base classifier is constructed using data collected from multiple subjects. We evaluate the predictive performance of the proposed technique on a large dataset [32].

The proposed methodology is described in Section II, where the details about the dataset, the type of the ensemble scheme, the training method, and the architecture for the base classifiers are provided. Section III presents the results followed by discussion in Section IV. Concluding remarks are presented in Section V.

II. METHODS

A. DATASET

We use one of the largest publicly available EEG-based MI dataset collected by Lee *et al.* from 54 participants (25 females and 29 males) [32]. The subjects had no pertinent disease and were aged between 24 and 35. Sixteen participants have been previously involved in BCI experiments, while others were naive BCI users. The dataset contains EEG recordings for two sessions of BCI experiments conducted with the same subjects and under the same protocol. Each experiment session consisted of training (offline) and testing (online) phases, with 100 trials in each phase, half of which were specific to the right-hand imagery tasks and the rest to the left-hand. EEG data were recorded using 62 Ag/AgCl electrodes at a sampling rate of 1000 Hz. The full description of the BCI experiment is available in [32]. For this study, the EEG recordings were down-sampled to 250 Hz, with an 8^{th} order Chebyshev anti-aliasing filter. Hereafter, this dataset is referred to as the MI-Dataset.

B. ENSEMBLE LEARNING

We use the *ensemble learning* scheme, as it has the potential to improve the overall predictive performance by reducing the variance via combining the decisions from several base models into a single prediction [37], [38]. In this regard,

the decisions of several base predictive models are combined via a voting mechanism. The framework is formalized in the sequel.

For a binary classification problem, a classifier is a mapping $\psi : \mathbb{R}^p \rightarrow \{0, 1\}$, such that $\psi(\mathbf{x}) = 0$ if $\mathbf{x} \in R_0$ and $\psi(\mathbf{x}) = 1$ if $\mathbf{x} \in R_1$, where $\mathbf{x}$ is a p-dimensional feature vector, and R_0 and R_1 are measurable sets partitioning the sample space. To form an ensemble classifier $\psi^{\mathrm{E}}(\mathbf{x})$, K base classifiers $\psi_i(\mathbf{x}), i = 1, \ldots, K$, are combined using a combination rule (combiner) $\mathcal{C}(.)$ as:

$$\psi^{\mathrm{E}}(\mathbf{x}) = \mathcal{C}\left(\bigcup_{i=1}^{K} \psi_i(\mathbf{x})\right). \tag{1}$$

The weighting combination rule is an important family of combiners in which for each base classifier $\psi_i(\mathbf{x})$, a weight w_i, which encodes the reliability of each base classifier, is calculated. Therefore, the ensemble classifier $\psi^{\mathrm{E}}(\mathbf{x})$ is formed as:

$$\psi^{\mathrm{E}}(\mathbf{x}) = \underset{y \in \{0,1\}}{\arg\max} \sum_{i=1}^{K} w_i \mathrm{I}_{\{\psi_i(\mathbf{x})=y\}}, \tag{2}$$

where $\mathrm{I}_{\{S\}}$ is 1 if statement S is true, zero otherwise. Herein, we use a majority vote (MV) combiner with equal weights for each base classifier ($w_i = 1, i = 1, \ldots, K$). Therefore, the output of the ensemble classifier is determined based on the decisions of the majority of the base classifiers. We choose K to be an odd number to avoid ties.

C. THE BASE CLASSIFIERS

As for the base classifier $\psi_i(\mathbf{x})$ used in (2), we choose the same CNN architecture utilized in [36], which is in fact the deep ConvNet proposed in [24] (here referred to as *DeepConvNet*). Fig. 1(a) shows a schematic representation of the DeepConvNet architecture, which consists of 4 convolutional layers and one dense *softmax* classification layer. The first layer is a special type of convolutional layer that performs splitted temporal and spatial convolution. The number of filters follows an expanding pattern with a factor of 2, starting from 25 filters in the first layer (for both temporal and spatial convolution), followed by 50, 100, and 200 filters in the second, third, and fourth convolutional layers, respectively. The detailed configuration of DeepConvNet that is used in this work is presented in Table 1. Accordingly, each convolutional step, except for the temporal convolution in the first layer, is followed by batch normalization, exponential linear unit (ELU) activation function, max-pooling, and dropout with a rate of 0.5. The Adam optimization algorithm with decoupled weight decay was used to train the model [39]. We trained the network for a maximum of 200 epochs with no early stopping. The learning rate was set to 0.01 with a weight decay of 0.0005. A batch size of 16 was used.

D. TRAINING AND TUNING BASE CLASSIFIERS, AND ASSESSING THE MS-EN-CNN

Similar to [36], for assessing the performance of our classifier in a SI classification context, we use leave-one-subject-out cross-validation (LOSO-CV) to successively hold out the observations for each subject, apply training, and model validation on the observations for remaining subjects (referred to as training subjects), and assess the performance of the constructed ensemble classifier on the held-out subject. To train the K base classifiers (i.e., K DeepConvNets) used in the ensemble rule (2) based on the training subjects, a K-fold CV is performed. In particular, once a subject is held out for testing, the remaining subjects (53 subjects) are randomly divided into K-folds. Each DeepConvNet is then trained on the data from $\lceil 53 - 53/K \rceil$ subjects (i.e., the subjects in $K-1$ folds) with the hyperparameter tuning being performed on the remaining $\lfloor 53/K \rfloor$ subjects (i.e., the subjects in the remaining fold); where $\lceil . \rceil$ and $\lfloor . \rfloor$ denote the ceiling and floor operators, respectively. We treat the epoch as the only hyperparameter to tune for the base models. In this regard, each base model is trained for a maximum of 200 epochs, and the epoch with the lowest validation loss is chosen. Thus, considering K-fold CV, K base DeepConvNet classifiers are trained and used subsequently in (2) to construct the ensemble classifier using the majority vote combination rule. Although, in general, K is one more hyperparameter to tune, we set $K = 13$, which was the maximum odd number that we could afford to achieve the results based on our computational resources (also see simulations conducted in Section IV-A). As each base CNN architecture is trained and tuned on a different subset of training subjects, we refer to the $\psi^{\mathrm{E}}(\mathbf{x})$ classifier as Multi-Subject Ensemble CNN (MS-En-CNN). A schematic representation of the MS-En-CNN architecture is shown in Fig. 1(b). A detailed flowchart of the methodology is shown in Fig. 2.

It should be noted that the whole data (combined from both phases and sessions, totalling 400 trials for each subject)

TABLE 1. Configuration of DeepConvNet that is used as a base classifier within MS-En-CNN.

Layers	Configuration details
Temporal 2D Convolution	25 filters, 10 × 1 kernel
Spatial 2D Convolution	25 filters, 1 × 62 kernel
Batch normalization	number of features =25
Non-linearity	ELU
Max-Pooling	3 × 1 kernel, 3 × 1 stride
Non-linearity	ELU
Dropout	50%
2D Convolution	50 filters, 10 × 1 kernel
Batch normalization	number of features =25
Non-linearity	ELU
Max-Pooling	3 × 1 kernel, 3 × 1 stride
Non-linearity	ELU
Dropout	50%
2D Convolution	100 filters, 10 × 1 kernel
Batch normalization	number of features = 50
Non-linearity	ELU
Max-Pooling	3 × 1 kernel, 3 × 1 stride
Non-linearity	ELU
Dropout	50%
2D Convolution	200 filters, 10 × 1 kernel
Batch normalization	number of features = 200
Non-linearity	ELU
Max-Pooling	3 × 1 kernel, 3 × 1 stride
Non-linearity	ELU
Fully connected	2 neurons
Activation	LogSoftmax

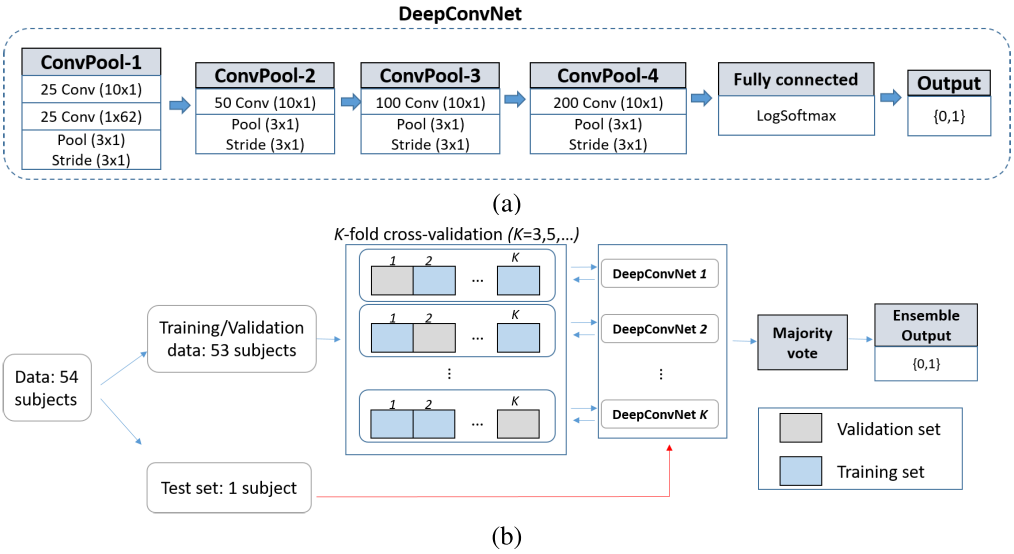


FIGURE 1. (a) Architecture of DeepConvNet as proposed by Schirrmeister *et al.* [24]; (b) Architecture for multi-subject DeepConvNet based ensemble (MS-Ens-DeepConvNet).

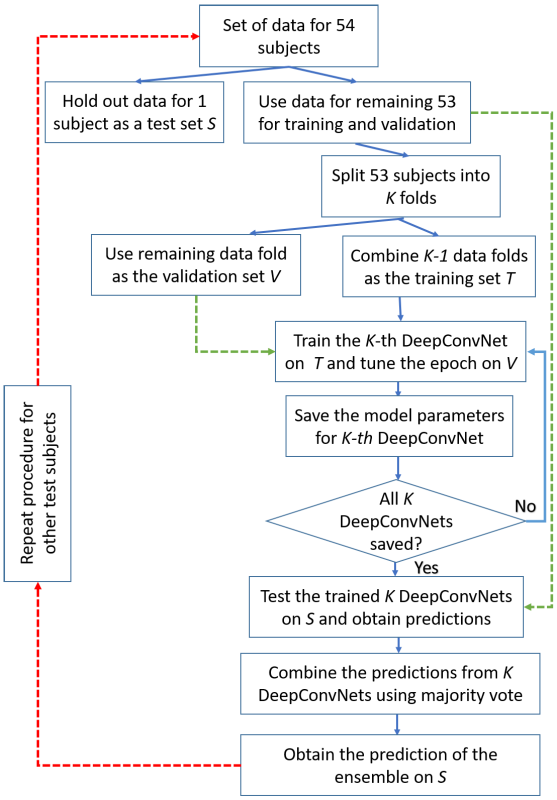


FIGURE 2. A flowchart describing the process of training and evaluating MS-En-DeepConvNet.

from all but the target subject (set aside for testing) is used for training. However, for the performance evaluation of the ensemble classifier, we have considered various cases of the target's data split (discussed in Section III-A). This, however, does not contradict to the fact that the training and test data are

TABLE 2. Subject-independent (LOSO-CV) performance of MS-En-CNN for various phases and sessions available in the MI-Dataset. s_i, $i = 1, 2$ and "on"/"off" indicate the session and the online/offline phases, respectively.

Session/(Phase)	Mean ± SD %	Median %	Range (Min-Max) %
s_1 / (off)	84.28 ± 10.63	84.50	53.00 (47.00-100.00)
s_1 / (on)	82.26 ± 12.33	83.50	42.00 (58.00-100.00)
s_2 / (off)	85.28 ± 10.57	86.50	49.00 (51.00-100.00)
s_2 / (on)	85.56 ±11.63	89.50	45.00 (55.00-100.00)
s_1 / (off+on)	83.27 ±9.92	83.50	43.00 (57.00-100.00)
s_2 / (off+on)	85.42 ± 10.16	86.50	46.50 (53.00-99.50)
$s_1 + s_2$	84.34 ± 9.42	84.88	44.25 (55.00-99.25)

of the same type, as both sessions had the same experimental protocol and subjects [32]. Training the classifier on one chunk of the data (taken from a particular phase or session) and testing it on a different subset of data does not create any problems. On the contrary, a session-free classifier, which is applicable to both online and offline phases, could be developed. This practice expands the utility of the developed classifier.

The source code for implementation of MS-En-CNN is available on GitHub.[1]

III. RESULTS

A. RESULTS OF THE PROPOSED METHOD

We present the SI classification accuracy of the MS-En-CNN on the MI-Dataset. Previously, several methods have been applied and evaluated on this dataset; however, different authors used different sessions and/or phases of the dataset for assessment (see Section III-B). To make a fair comparison with the previously reported results and in order to set a benchmark for future comparisons, here we report the performance for: 1) each phase (offline and online) and session

[1] https://github.com/irinadolzhikova/MS-En-CNN

TABLE 3. SI classification accuracy of different methods on the MI-Dataset. A "–" sign indicates a measure that is not reported in the reference.

Method	Session (Phase)	Mean ± SD %	Median %	Range (Min-Max) %
Pooled CSP [33]	s_2 (on)	65.65 ±16.11	58.00	55.00 (45.00- 100.00)
MR FBCSP [33]	s_2 (on)	68.59 ± 15.28	63.00	49.00 (48.00-97.00)
Fused model [33]	s_2 (on)	67.37 ± 16.01	62.50	57.00 (41.00-98.00)
CNN based fusion technique [33]	s_2 (on)	74.15 ± 15.83	75.00	59.00 (41.00-100.00)
CCSPNet [34]	s_2 (on)	74.11 ± 15.42	73.50	50.00 (50.00-100.00)
Jeon et al. [35] with DeepConvNet	s_2 (off+on)	73.32±13.55	–	–
Jeon et al. [35] with EEGNet	s_2 (off+on)	72.16±13.51	–	–
DeepConvNet [36]	s_2 (off+on)	84.19±10.08	84.50	47.50 (52.00-99.50)
MS-En-CNN	s_2 (on)	85.56 ±11.63	89.50	45.00 (55.00-100.00)
MS-En-CNN	s_2 (off+on)	85.42 ± 10.16	86.50	46.50 (53.00-99.50)

(s_1 and s_2) separately, 2) for pooled data from the online and the offline phases while keeping sessions separate, and 3) for a pooled data across all phases and sessions. Table 2 shows the mean ± standard deviation and the median LOSO-CV accuracy of MS-En-CNN in these scenarios (see Supplementary Materials for the subject-specific classification accuracies).

B. COMPARISON WITH THE STATE-OF-THE-ART METHODS

Recently, several algorithms have been tested on the MI-Dataset. Kwon *et al.* demonstrated the method of combining individually trained CNN models using concatenation fusion [33]. In addition the authors demonstrated an improved performance vs. the following methods: CSP, a fused model designed by Ray *et al.* [13] and multiresolution filter bank CSP (MR FBCSP). For the SI classification assessment using LOSO-CV, they used the online phase from Session 2 and reported an average accuracy of 74.15±15.83%. Another recent investigation proposed a hybrid architecture, called CCSPNet, which combines a wavelet kernel CNN, a temporal CNN, a CSP, and a dense neural network. The combined architecture achieved a LOSO-CV accuracy of 74.11±15.42% on the online phase from Session 2 [34]. Jeon *et al.* proposed a deep neural network that learns subject-invariant and class-relevant representations [35]. Using DeepConvNet and EEGNet as a feature extractor, the authors achieved on average 73.32±13.55% and 72.16±13.51%, respectively, in a zero-training scenario. The most recent work that reported the highest LOSO-CV accuracy of 84.19±10.08%, was based on a single DeepConvNet architecture and used the pooled data across both the online and the offline phases from Session 2 for assessment[2] [36].

These state-of-the-art results are shown in Table 3, where for the ease of comparison, we also add our results from Table 2 for both the online phase and the pooled data from the online and the offline phases from Session 2. Accordingly, MS-En-CNN outperforms previously reported results, both in terms of the average and the median accuracies. Considering the large number of subjects and trials, the observed improvements with respect to the runner-up accuracy reported in [36] is statistically significant. This is confirmed by a one-sided paired Wilcoxon signed rank test where the P-value is 2.0×10^{-4}. To have a better picture of the achieved improvement, we also present the scatter plot of the SI classification accuracy obtained using MS-En-CNN vs. those obtained in [36] for each subject when it is held out for testing (see Fig. 3). The points above the identity line in Fig. 3 indicate the outperformance of MS-En-CNN. MS-En-CNN achieved a higher SI classification accuracy for 35 out of 54 subjects in the MI-Dataset.

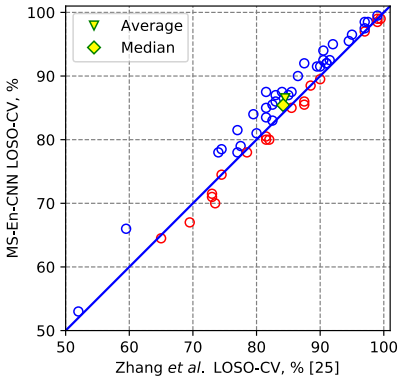


FIGURE 3. The scatter plot of the SI classification accuracy obtained using MS-En-CNN (in blue) vs. those obtained in [36] (in red) for each subject when it is held-out for testing. The vertical and horizontal axes are the accuracies achieved by the herein proposed MS-En-CNN algorithm and those reported by Zhang *et al.* [36], respectively. The points above the identity line indicate the outperformance of MS-En-CNN.

IV. DISCUSSION

A. THE EFFECT OF K ON THE PERFORMANCE OF MS-EN-CNN

Although in presenting the results shown in the previous section we used $K = 13$, which was the maximum odd number that we could afford based on our computational resources, K is a hyperparameter and naturally affects the performance of MS-En-CNN classifier. As a result, here we study the effect of K on the performance of MS-En-CNN. In this regard, we assumed $K \in \{3, 5, 7, 9, 11, 13\}$ and repeated the entire process that was described in Section II-D for each K (the results of $K = 13$ was already achieved). Table 4 shows the LOSO-CV accuracy of MS-En-CNN for different K obtained for the pooled data across both the online and the offline phases from Session 2. We further examined whether these results exhibit a statistically significant improvement with respect to the performance of the single

[2]based on the source codes for the implementation available at https://github.com/zhangks98/eeg-adapt

TABLE 4. Subject-independent (LOSO-CV) performance of MS-En-CNN as a function of K obtained on the pooled data across both the online and the offline phases from Session 2.

K-fold CV	Mean ± SD %	Median %	Range (Min-Max) %
$K = 3$	84.41 ± 10.27	85.25	46.00 (54.00-100.00)
$K = 5$	84.91 ± 10.12	85.50	46.50 (53.00-99.50)
$K = 7$	85.38 ± 9.93	87.00	46.00 (53.50-99.50)
$K = 9$	85.30 ± 9.87	87.50	46.00 (53.50-99.50)
$K = 11$	85.31 ± 10.14	87.00	47.50 (52.00-99.50)
$K = 13$	85.42 ± 10.16	86.50	46.50 (53.00-99.50)

TABLE 5. *P*-values calculated using one-sided Wilcoxon signed rank test for pairwise comparison between the classification accuracy achieved by an ensemble with the DeepConvNet based base classifiers (MS-Ens-DeepConvNet) trained using K-fold CV vs. a single model of DeepConvNet.

K-fold CV	P-values
$K = 3$	0.243
$K = 5$	0.023
$K = 7$	9×10^{-4}
$K = 9$	10^{-4}
$K = 11$	0.001
$K = 13$	2×10^{-4}

DeepConvNet that was trained in [36]. In this regard, the P-values of a one-sided paired Wilcoxon signed rank test were calculated and are presented in Table 5. We make a few observations/recommendations based on these results: 1) for $K \geq 7$, the MS-En-CNN shows a statistically significant performance improvement over the single DeepConvNet that was trained in [36]; 2) although a larger K such as 13 would potentially lead to a better performance than a relatively small K such as 3, we might be better off with a moderate K such as $K = 7$ or 9 because they virtually show a comparable performance to that of $K = 13$ and, at the same time, are computationally less expensive (a lower number of base classifiers should be trained).

B. THE EFFICACY OF MS-EN-CNN

1) A PRACTICAL PERSPECTIVE

From a practical standpoint, in this work, we have deployed the capabilities of a CNN-based ensemble framework for zero-calibration classification of MI EEG. The demonstrated remarkable performance observed by the proposed approach is the consequence of using the theoretically justifiable proposed ensemble learning scheme (see the following subsection on the analytical perspective) along with convolutional neural networks that on their own, have been proven to be powerful tools for decoding complex patterns. Leveraging the advantages of both learning realms, we were able to statistically significantly outperform previous research on the herein studied large MI dataset.

With a fixed base CNN architecture and one algorithmic hyperparameter (the epoch), we constructed K base models and aggregated their decision using majority voting. We further investigated the effect of K on the performance of MS-EN-CNN and examined its effect up to $K = 13$ based on our available computational resources. Depending on computing power capacity, in the future one may even examine a much larger K (with a maximum of $n-1$, where n is the number of subjects in the dataset) to potentially improve the classification performance. As a matter of fact, the theoretical justification outlined in the next section is based on setting $K = n - 1$; however, given the large number of subjects in our dataset (54 subjects), this is computationally challenging. Nonetheless, as shown in Section IV-A, with a moderate value of K ($K = 7$ or 9) the performance improvement is already significant.

2) AN ANALYTICAL PERSPECTIVE

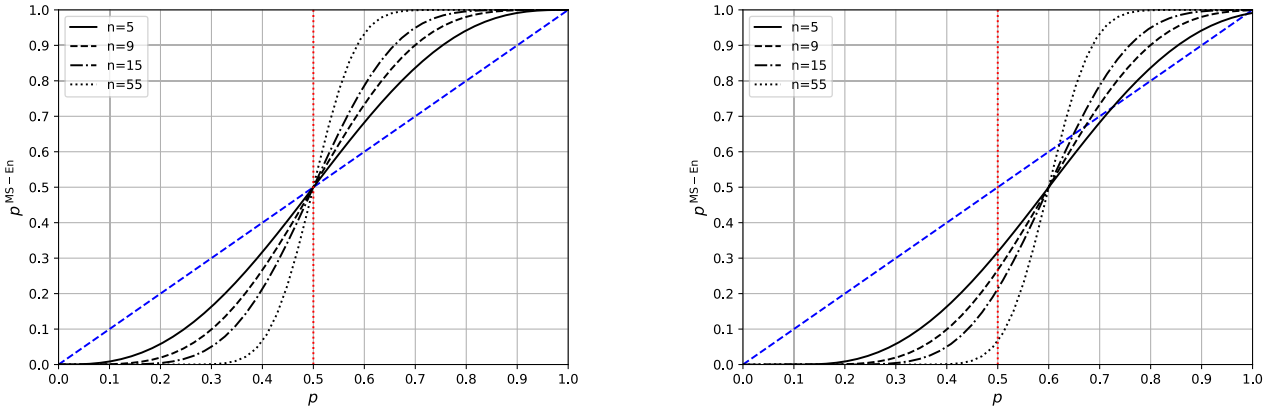
Our results presented in Section III show that the performance improvement achieved by MS-En-CNN with DeepConvNet base classifiers is statistically significant as compared with the performance of a single DeepConvNet that is trained using the data collected from all training subjects. This interesting observation raises the question of whether we can justify these results from a general machine learning standpoint.

Suppose there are n subjects (for simplicity, an odd number) in a binary classification problem (e.g., target vs. non-target), from which an equal number of trials are collected. The goal is to use the data from all these training subjects to construct a subject-independent classifier. The usual approach is to pool data from all these subjects to train one single classifier $\psi(\mathbf{x})$. In our application, this corresponds to the single DeepConvNet classifier trained in [36]. Hereafter, we refer to such a classifier as *pooled-data classifier*. Let p denote the accuracy of the pooled data classifier; i.e., the probability of correctly classifying a given $\mathbf{x}$. Furthermore, assume $p > 0.5$. This is not a restrictive assumption in practice because in many BCI experiments, it is fairly reasonable to assume that the pooled-data classifier has some predictability compared to random classification; after all, this is the underlying hypothesis justifying the process of data collection in the first place.

Although in constructing MS-En-CNN, we used the data from $\lceil n - n/K \rceil$ subjects (i.e., subjects in $K - 1$ folds) to train each base classifier, here for simplicity (to avoid the randomness involved in dividing subjects to K folds), we assume $K = n$; that is, each base classifier is trained on the data from all subjects except for one. Hereafter, by referring to a subject we indeed refer to the entire data collected from that subject.

We further assume that the training data is also used to construct an ensemble classifier $\psi^{\text{MS-En}}(\mathbf{x})$, obtained by using the majority vote combination rule (i.e., setting $w_i = 1$ in (2)) among n base classifiers $\psi_i(\mathbf{x}), i = 1, \ldots, n$, which are obtained by removing one subject from n training subjects.

Case 1. *Assumption*: Training a base classifier with any subset of $n - 1$ subjects has the same accuracy as training a classifier with all n subjects; in other words, each $\psi_i(\mathbf{x})$ has an accuracy p. This assumption ideally approximates a situation where having one less subject in our study does not



(a) Case 1

(b) Case 2

FIGURE 4. $p^{\text{MS-En}}$ as a function of p for different values of n.

considerably affect the predictability of a classification rule—this is the case especially if n is relatively large.

Under this assumption, the majority vote ensemble classifier correctly classifies $\mathbf{x}$ if and only if $(n+1)/2$ base classifiers $\psi_i(\mathbf{x})$ correctly classify a given $\mathbf{x}$. Let $p^{\text{MS-En}}$ denote this probability. It follows that

$$p^{\text{MS-En}} = \sum_{i=0}^{\frac{n-1}{2}} \binom{n}{n-i} p^{(n-i)}(1-p)^i \tag{3}$$

The question is whether $p^{\text{MS-En}} > p$. Fig. 4(a) shows $p^{\text{MS-En}}$ as a function of p. As observed in this figure, for $p > 0.5$, we have $p^{\text{MS-En}} > p$ for different values of n. At the same time, when p is neither close to 0.5 nor 1, the larger n is, the higher the improvement achieved by using $p^{\text{MS-En}}$ with respect to the pooled-data classifier.

Case 2: *Assumption*: Removing any subject from n subjects has an equal impact and drops p by $q < p$; that is, the accuracy of each $\psi_i(\mathbf{x})$ becomes $p - q$. Assuming a similar number of trials collected from each subject and the fact that all subjects go through the same data collection process would make this assumption fairly reasonable. In this case, $p^{\text{MS-En}}$ is obtained from (3) by replacing p with $p - q$. To be able to examine how q affects $p^{\text{MS-En}}$, we assume $q = p/n$ (just a working assumption to imply that the drop in accuracy is directly and inversely proportional to p and n, respectively). Fig. 4(b) shows $p^{\text{MS-En}}$ as a function of p for $q = p/n$. Accordingly, if p is relatively low for the pooled-data classifier (for example $p = 60\%$), $\psi^{\text{MS-En}}(\mathbf{x})$ would not be helpful even if we have a relatively large n (all curves are below the diagonal line for $p = 60\%$). The situation improves for larger p. For example, when $p = 70\%$, having $n = 9$ subjects leads to around 4% of improvement with respect to the performance of the pooled-data classifier, while having $n = 55$ subjects leads to 24% of improvement.

The analytical justification presented here can help shed light on how $\psi^{\text{MS-En}}(\mathbf{x})$ could possibly lead to performance improvement with respect to the usual pooled-data classification. This performance improvement could be seen as a consequence of voting among many base classifiers that all have approximately the same performance as the pooled-data classifier. However, we emphasize that the analytical results shown in this section are valid under ideal situations dictated by the aforementioned assumptions presented in Case 1 and Case 2. Nevertheless, the analytical results are independent from the particular choice of the base classifier (for example, the DeepConvNet that was used in our application due to its remarkable performance in EEG-based classification of MI tasks). These results call for future investigations on the utility of MS-En classification using other base classifiers and in other settings where the pooled-data classification exhibits a varying degree of accuracy.

V. CONCLUSION

The aim of this study was to develop a simple yet effective method to improve the decoding accuracy of SI BCI systems. With this aim, we have utilized the advantages of both deep CNNs and ensemble learning to propose a multisubject ensemble CNN classification rule (MS-En-CNN). We showed that using MS-En-CNN leads to an overall increase in classification accuracy. In particular, the proposed technique achieved an average SI classification accuracy of 85.42±10.16% on one of the largest open-access MI databases with 54 subjects, which outperformed the state-of-the-art methods on the same dataset. Not only is the work significant in terms of reporting the highest overall SI classification accuracies thus far reported on such a large MI dataset, but the proposed MS-En-CNN classification rule could also be potentially used in the future to achieve the state-of-the-art SI classification performance in other EEG-based paradigms.

ACKNOWLEDGMENT

The authors acknowledge the Institute of Smart Systems and Artificial Intelligence, Nazarbayev University, for providing access to their NVIDIA DGX-1 server for computational purposes.

REFERENCES

[1] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. Mcalpine, and Y. Zhang, "A survey on deep learning-based non-invasive brain signals: Recent advances and new frontiers," 2019, *arXiv:1905.04149*.

[2] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[3] Y. Bunno, "The application of motor imagery to neurorehabilitation," in *Evolving BCI TherapyEngaging Brain State Dynamics*. Kansai, Japan: Intech, 2018, pp. 53–71.

[4] 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.

[5] Y. Höller, A. Thomschewski, A. Uhl, A. C. Bathke, R. Nardone, S. Leis, E. Trinka, and P. Höller, "HD-EEG based classification of motor-imagery related activity in patients with spinal cord injury," *Frontiers Neurol.*, vol. 9, p. 955, Nov. 2018.

[6] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "EEG-based brain-computer interfaces using motor-imagery: Techniques and challenges," *Sensors*, vol. 19, no. 6, p. 1423, Mar. 2019.

[7] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, 2018, Art. no. 031005.

[8] S. Saha, K. A. Mamun, K. Ahmed, R. Mostafa, G. R. Naik, S. Darvishi, A. H. Khandoker, and M. Baumert, "Progress in brain computer interface: Challenges and opportunities," *Frontiers Syst. Neurosci.*, vol. 15, Feb. 2021, Art. no. 578875.

[9] M. Arvaneh, C. Guan, K. K. Ang, T. E. Ward, K. S. Chua, C. W. K. Kuah, G. J. E. Joseph, K. S. Phua, and C. Wang, "Facilitating motor imagery-based brain–computer interface for stroke patients using passive movement," *Neural Comput. Appl.*, vol. 28, no. 11, pp. 3259–3272, 2017.

[10] M. Krauledat, M. Schröder, B. Blankertz, and K.-R. Müller, "Reducing calibration time for brain-computer interfaces: A clustering approach," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 753–760.

[11] S. Fazli, F. Popescu, M. Danóczy, B. Blankertz, K.-R. Müller, and C. Grozea, "Subject-independent mental state classification in single trials," *Neural Netw.*, vol. 22, pp. 1305–1312, Nov. 2009.

[12] F. Lotte, C. Guan, and K. K. Ang, "Comparison of designs towards a subject-independent brain-computer interface based on motor imagery," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 4543–4546.

[13] A. M. Ray, R. Sitaram, M. Rana, E. Pasqualotto, K. Buyukturkoglu, C. Guan, K.-K. Ang, C. Tejos, F. Zamorano, and F. Aboitiz, "A subject-independent pattern-based brain-computer interface," *Frontiers Behav. Neurosci.*, vol. 9, p. 269, Oct. 2015.

[14] I. K. Niazi, N. Jiang, M. Jochumsen, J. F. Nielsen, K. Dremstrup, and D. Farina, "Detection of movement-related cortical potentials based on subject-independent training," *Med. Biol. Eng. Comput.*, vol. 51, no. 5, pp. 507–512, May 2013.

[15] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review," *J. Neural Eng.*, vol. 16, no. 5, Oct. 2019, Art. no. 051001.

[16] B. Abibullaev, I. Dolzhikova, and A. Zollanvari, "A brute-force CNN model selection for accurate classification of sensorimotor rhythms in BCIs," *IEEE Access*, vol. 8, pp. 101014–101023, 2020.

[17] M. B. Bora, D. Daimary, K. Amitab, and D. Kandar, "Handwritten character recognition from images using CNN-ECOC," *Proc. Comput. Sci.*, vol. 167, pp. 2403–2409, 2020.

[18] Z. Yang, J. Long, Y. Zi, S. Zhang, and C. Li, "Incremental novelty identification from initially one-class learning to unknown abnormality classification," *IEEE Trans. Ind. Electron.*, vol. 69, no. 7, pp. 7394–7404, Jul. 2022.

[19] J. Long, Y. Chen, Z. Yang, Y. Huang, and C. Li, "A novel self-training semi-supervised deep learning approach for machinery fault diagnosis," *Int. J. Prod. Res.*, pp. 1–14, Feb. 2022.

[20] J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li, and Y. Bi, "EEG-based emotion classification using a deep neural network and sparse autoencoder," *Frontiers Syst. Neurosci.*, vol. 14, p. 43, Sep. 2020.

[21] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. McAlpine, and Y. Zhang, "A survey on deep learning-based non-invasive brain signals: Recent advances and new frontiers," *J. Neural Eng.*, vol. 18, no. 3, Jun. 2020, Art. no. 031002.

[22] X. Lun, Z. Yu, T. Chen, F. Wang, and Y. Hou, "A simplified CNN classification method for MI-EEG via the electrode pairs signals," *Frontiers Human Neurosci.*, vol. 14, p. 338, Sep. 2020.

[23] N. Mammone, C. Ieracitano, and F. C. Morabito, "A deep CNN approach to decode motor preparation of upper limbs from time–frequency maps of EEG signals at source level," *Neural Netw.*, vol. 124, pp. 357–372, Apr. 2020.

[24] 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.

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

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

[27] G. Dai, J. Zhou, J. Huang, and N. Wang, "HS-CNN: A CNN with hybrid convolution scale for EEG motor imagery classification," *J. Neural Eng.*, vol. 17, no. 1, Jan. 2020, Art. no. 016025.

[28] X. Tang, J. Yang, and H. Wan, "A hybrid SAE and CNN classifier for motor imagery EEG classification," in *Proc. Comput. Sci. On-Line Conf.* Cham, Switzerland: Springer, 2018, pp. 265–278.

[29] M. Dai, D. Zheng, R. Na, S. Wang, and S. Zhang, "EEG classification of motor imagery using a novel deep learning framework," *Sensors*, vol. 19, no. 3, p. 551, 2019.

[30] W. Qiao and X. Bi, "Deep spatial-temporal neural network for classification of EEG-based motor imagery," in *Proc. Int. Conf. Artif. Intell. Comput. Sci.*, Jul. 2019, pp. 265–272.

[31] 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, Oct. 2019, Art. no. 066004.

[32] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, and S.-W. Lee, "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, May 2019.

[33] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 3839–3852, Oct. 2019.

[34] M. Nouri, F. Moradi, H. Ghaemi, and A. M. Nasrabadi, "Towards real-world BCI: CCSPNet, a compact subject-independent motor imagery framework," 2020, *arXiv:2012.13567*.

[35] E. Jeon, W. Ko, J. S. Yoon, and H.-I. Suk, "Mutual information-driven subject-invariant and class-relevant deep representation learning in BCI," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 6, 2021, doi: 10.1109/TNNLS.2021.3100583.

[36] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network," *Neural Netw.*, vol. 136, pp. 1–10, Apr. 2021.

[37] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[38] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1249, Jul. 2018.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

IRINA DOLZHIKOVA received the bachelor's (Hons.) and master's degrees in electrical and electronic engineering from Nazarbayev University, Kazakhstan, in 2016 and 2018, respectively. She is currently pursuing the Ph.D. degree in the area of machine learning with the focus on the EEG-based brain–computer interfaces. From 2017 to 2018, she was a Teaching Assistant at Nazarbayev University.

BERDAKH ABIBULLAEV (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electronic engineering from Yeungnam University, South Korea, in 2006 and 2010, respectively. He held research scientist positions at the Daegu Gyeongbuk Institute of Science and Technology (2010–2013) and at the Samsung Medical Center, Seoul, South Korea (2013–2014). In 2014, he received the National Institute of Health Postdoctoral Research Fellowship II to join a multi-institutional research project between the University of Houston BMI Systems Team and the Texas Medical Center in developing neural interfaces for rehabilitation. He is currently an Assistant Professor at the Robotics Department, Nazarbayev University, Kazakhstan. His research interests include machine learning algorithms, neural signal processing, and brain–computer/machine interfaces.

REZA SAMENI (Senior Member, IEEE) received the bachelor's degree in electronics engineering from Shiraz University, Iran, in 2000, the master's degree in biomedical engineering from the Sharif University of Technology, Iran, in 2003, and the dual Ph.D. degrees in signal processing and biomedical engineering from the Institut National Polytechnique de Grenoble (INPG), France, and the Sharif University of Technology, in 2008. He was a Tenured Associate Professor with the School of Electrical and Computer Engineering, Shiraz University (2008–2018) and an Invited Senior Researcher at the GIPSA-Laboratory, Grenoble, France (2018–2020). He has been an Associate Professor of biomedical engineering at Emory University, Atlanta, GA, USA, since 2020. His research interests include statistical signal processing with special interest in mathematical modeling and analysis of biomedical systems and signals.

AMIN ZOLLANVARI (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Shiraz University, Iran, and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2010. He held a postdoctoral position at Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA (2010–2012), and then joined the Department of Statistics, Texas A&M University, as an Assistant Research Scientist (2012–2014). Since 2015, he has been with Nazarbayev University, where he is currently an Associate Professor with the Department of Electrical and Computer Engineering. His research interests include machine learning, statistical signal processing, and biomedical informatics. He is an Associate Editor of IEEE ACCESS.