Motor Imagery Classification for Brain Computer Interface Using Deep Convolutional Neural Networks and Mixup Augmentation

Haider Alwasiti and Mohd Zuki Yusoff, *Member, IEEE*

Abstract*—*Goal: **Building a DL model that can be trained on small EEG training set of a single subject presents an interesting challenge that this work is trying to address. In particular, this study is trying to avoid the need for long EEG data collection sessions, and without combining multiple subjects training datasets, which has a detrimental effect on the classification performance due to the inter-individual variability among subjects.** ***Methods:*** **A customized Convolutional Neural Network with mixup augmentation was trained with ~120 EEG trials for only one subject per model.** ***Results:*** **Modified ResNet18 and DenseNet121 models with mixup augmentation achieved 0.920 (95% Confidence Interval: 0.908, 0.933) and 0.933 (95% Confidence Interval: 0.922, 0.945) classification accuracy, respectively.** ***Conclusions:*** **We show that the designed classifiers resulted in a higher classification performance in comparison to other DL classifiers of previous studies on the same dataset, despite the limited training dataset used in this work.**

***Index Terms*—EEG, deep learning, BCI, stockwell transform.**

***Impact Statement*—Mixup augmentation with modified Convolutional Neural Networks could be trained with ~120 EEG trials for only one subject per model, effectively mitigating the inter-individual variability of brain computer interface classification.**

Manuscript received 5 March 2022; revised 23 September 2022 and 16 October 2022; accepted 23 October 2022. Date of publication 23 November 2022; date of current version 13 December 2022. This work was supported in part by the Ministry of Education Malaysia through Higher Institutional Centre of Excellence (HICoE) Scheme awarded to the Centre for Intelligent Signal and Imaging Research (CISIR), Universiti Teknologi PETRONAS (UTP), Malaysia and in part by Yayasan Universiti Teknologi PETRONAS (YUTP) Fund under Grant 015LC0-239. The review of this article was arranged by Editor P. Bonato. *(Corresponding author: Haider Alwasiti.)*

Haider Alwasiti is with the Helsinki Lab of Interdisciplinary Conservation Science, University of Helsinki, FI-00014 Helsinki, Finland (e-mail: haider.alwasiti@helsinki.fi).

Mohd Zuki Yusoff is with the Centre for Intelligent Signal and Imaging Research (CISIR), Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia (e-mail: mzuki_yusoff@utp.edu.my).

This article has supplementary downloadable material available at https://doi.org/10.1109/OJEMB.2022.3220150, provided by the authors.

Digital Object Identifier 10.1109/OJEMB.2022.3220150

I. Introduction

AFTER the invention of the Electroencephalogram by Hans Berger in 1924 [9], the interest in using the electrical brain signals (EEG) for control has been very popular for many decades. However, in the last three decades, the research proved that communication and control using brain waves is possible. EEG can be modulated by the thinking process [6], [38], which has been leveraged for brain computer interface applications (BCI) [5], [15]. With motor imagery BCI, by the imagination of the movement, the system is able to detect that imaginary process [38]. However, robust and highly accurate BCI systems are yet to be developed, due to the extremely weak and noisy brain waves that are correlated with the different thinking process of the human brain [7].

Recently, deep learning (DL) has been attempted to classify MI-BCI signals. DL classifiers, in comparison to the traditional shallow ML classifiers, are less affected by the curse of dimensionality. Tabar et al. [35] proposed a CNN with stacked auto-encoders (SAEs) that has been trained on time-frequency maps from Short Time Fourier Transform (STFT) of EEG signals. They reported 0.75 average classification accuracy on nine subjects with a training time of 0.3 h. Schirrmeister et al. [32] trained two-layers shallow CNN classifier, five-layers deep CNN and a 31-layers ResNet. The shallow CNN achieved higher performance (mean classification accuracy 0.74) by a few percents over the deep CNN methods on a public dataset with frequency range of 0–38 Hz. An improvement in the classification accuracy has been achieved by combining the BCI Competition IV-2a public dataset with their own dataset (frequency: 0–125 Hz; 20 subjects; 1000 trials per subject) to reach 0.84 average accuracy with 1 h training time. Despite that EEG shows 1/f power spectrum distribution, which makes it difficult to have a good SNR for frequencies more than 60 Hz, the study showed that gamma waves (40–125 Hz) were encoding useful information for MI-BCI and effectively enhanced the classification performance, since physiologically, 60 to 100 Hz EEG frequency band is known to be increased in amplitude with movement execution and is correlated with movement-related information [12], [16], [29].

Therefore, employing gamma wave features can potentially enhance the classification accuracy. However, only recently, with

the advent of deep learning methods, gamma waves have been incorporated more frequently. This is likely due to the tradeoff between the potentially small classification improvement due to gamma wave features utilization, and the detrimental effect of increasing the number of features without increasing the training dataset size. Therefore, since classical ML models are more susceptible to the curse of dimensionality, avoiding the gamma wave features resulted in better performance. However, DL models could mitigate the increase in feature space dimensionality even without increasing the training size.

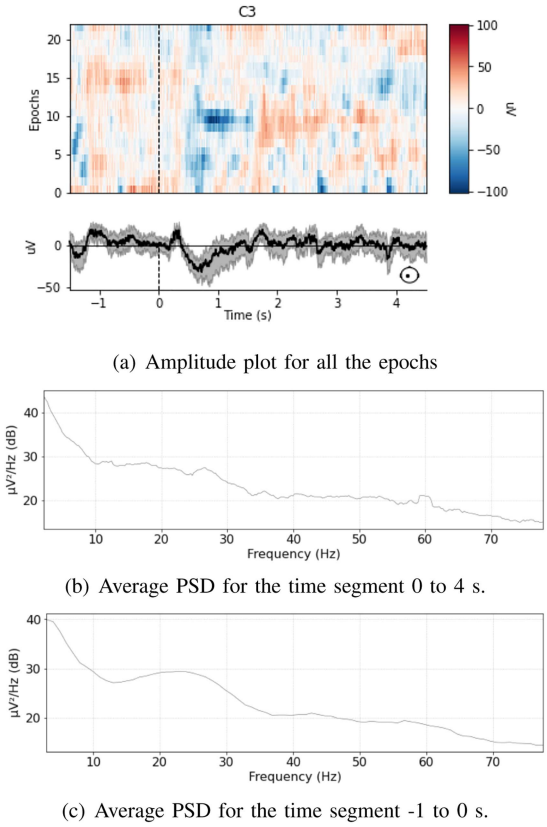
Our contributions in this work are the development of the first, to our knowledge, CNN classifier that is able to converge and classify EEG signals with a small number of training samples for only single subject per model, effectively solving the issue of inter-individual variability of MI-BCI EEG. Mixup augmentation has been utilized for the first time for MI-BCI classification. Others have attempted this task using CNNs, but needed substantially larger training dataset in comparison to what could be achieved in this work, in part because the augmentation techniques that are commonly used with CNNs are not applicable to EEG spectrograms. We show that with mixup augmentation, we can counteract the limited dataset issue, without sacrificing the performance of the classifier.

II. Materials and Methods

A. Data Preprocessing

The EEG signal of each user was segmented into 5 seconds epochs. Each epoch begins one second before starting the imagery event trial and lasts for four seconds. Fig. 1 demonstrates a plot with 22 epochs of the C3 channel of subject 22. In Fig. 1(a), the upper panel demonstrates the amplitude plot for all the epochs. The color bar indicates the EEG amplitude of the signal recorded in C3, y-axis stands for the epochs/trials, x-axis for the time in seconds. The lower panel shows the mean over trials (bold signal) and the gray shadow is the standard deviation over trials. The dotted line remarks the onset of the event where the target shown on the screen. Fig. 1(b) shows the average PSD a single epoch for the time segment 0 to 4 s, and Fig. 1(b) demonstrates the average PSD of the same epoch for the time segment -1 to 0 s (the baseline segment). Common average reference spatial filter has been applied to the raw EEG signals, where the average of all channels montage was subtracted from the EEG channel of interest.

Each epoch was converted into 64 spectrograms, which are showing the time-frequency representation of the EEG power over time for the frequency range of (2–78 Hz; y-log scaled) for each EEG channel. To increase the signal to noise ratio, we have adopted a baseline correction for all trials. Basically, this was a spectral subtraction method that is commonly used for reducing background noise in speech signals [11]. The stationary noise was estimated from 1 s period baseline before the beginning of the imagery movement. The spectral power plot was normalized by the logratio referred baseline method, where the power spectrum was divided by the baseline mean power and taking the log of the result.



(a) Amplitude plot for all the epochs

(b) Average PSD for the time segment 0 to 4 s.

(c) Average PSD for the time segment -1 to 0 s.

Fig. 1. C3 channel's plots of the subject 22. (a) Upper panel: Amplitude plot for all epochs. The color bar indicates the EEG amplitude, y-axis stands for the epochs, x-axis for the time in seconds. Lower panel: the bold signal is the mean over trials and the gray shadow is the standard deviation over trials. The dotted line remarks the onset of the event where the target shown on the screen. (b) Average PSD of a single epoch for the time segment 0 to 4 s. (c) Average PSD of the same epoch for the time segment -1 to 0 s (the baseline segment).

Stockwell Transform has been shown, in a previous study, as an effective EEG preprocessing method for MI-BCI classification [4], [5]. The discrete time Stockwell Transform is expressed as follows: Let $f = m\Delta_F, \alpha = p\Delta_F$ and $t = n\Delta_T$, where f is the frequency, t is the time, N is the total samples count, Δ_F is the sampling frequency, α is the Gaussian window width and Δ_T is the sampling interval, then:

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

The Gaussian window width α of the Stockwell Transform controls the tradeoff for the spectral and temporal resolution, and empirically we found that a width of 0.6 had the best performance for this study. Fig. 2 demonstrates a sample of a Stockwell Transform plot. The frequency axis has been log-transformed to put more weight on the *mu* (8–12 Hz) and the *beta rhythms* (18–25 Hz) which are physiologically correlated with movement or imagination to move the limbs [21].

B. Convolutional Neural Network

A modified ResNet18 model was used for this project. Other model architectures have been attempted. However,

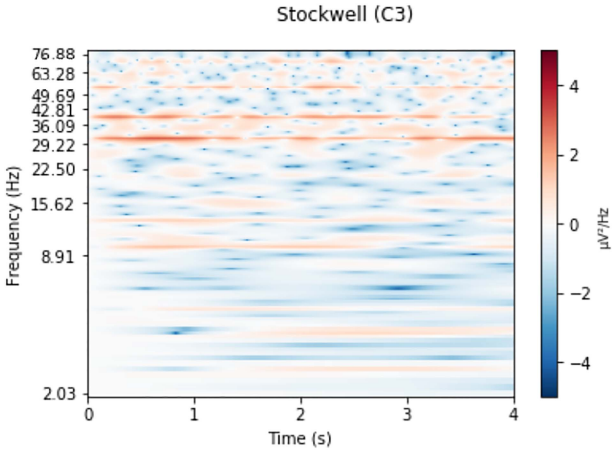


Fig. 2. EEG channel C3 log freq-scaled Stockwell power spectrogram of 1 trial (left hand).

only DenseNet121 performed better than ResNet18. The DenseNet121 model has been modified similarly, reaching a total number of 11,164,547 parameters. Both models performed better with transfer learning from ImageNet pretrained weights; therefore, all reported results used pretrained weights. The architecture of the DenseNet121 model is demonstrated in Fig. 3.

The network is optimizing the DataLoss 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. We have observed that a small part of the gradients during training will diverge, and clipping gradients into 0.1 helped to improve the accuracy. An optimum weight decay has been found by grid search and 0.1 yielded the best value for both CNN models. Weight decay is considered as one type of L2 regularization method which has been used to enhance generalization performance of neural networks and decrease overfitting [18]. With this regularization term, the loss function is expressed by:

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

where x is the mini-batch, w is the model weights and c is the weight decay constant. During gradient descent, the model weights were updated in each iteration by:

$$w := w(1 - \kappa c) - \kappa \frac{d\,\text{Loss}(w, x)}{dw} \qquad (3)$$

where $\frac{1}{2}\, c \,\|w\|^2$ is the L2 penalty term and κ is the learning rate. Therefore, the weight decay is encouraging all the model weights to be scaled down and proportionally decaying toward zero. However, recently with the common use of Batch Normalization in DL models [17], the effect of weight decay on training CNN models when used with Batch Normalization is poorly understood [40]. Batch Normalization is counteracting the effect of scaling down model's weights since the Batch Normalization is generally making its output invariant to the scaling effect of the previous layer's output. Nevertheless, it is still used as a useful trick to decrease overfitting and improving model's accuracy. It is speculated that instead of the L2 regularization effect, its

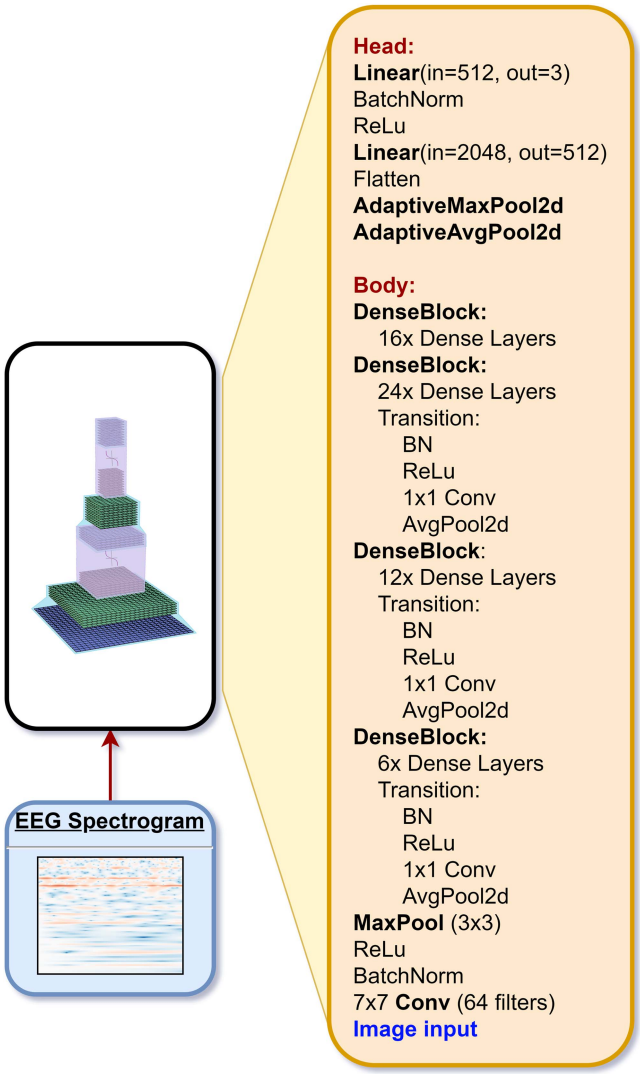


Fig. 3. Modified DenseNet121 model architecture.

action when used with BN is more likely on preventing the decay of effective learning rate over time. Also, by maintaining a higher effective learning rate over time, it leads to better generalization [40].

One of the main challenges in training the models was the limited dataset. DL models typically need a large amount of data for training. Data augmentation is commonly used to expand small and medium-sized datasets. It is a synthetic process that creates more data items by combining or changing the characteristics of the existing data items to encourage the classifier to correlate the unchanged characteristics with the trained classes and ignore those randomly changing characteristics in the dataset. For image data classification problems, this is performed by randomly rotating the image items, random light changes, random Gaussian blurring (to emulate lens effects), random resizing, random cropping, or copping out and other visual effects [25]. For audio spectrograms, typically random noise or echo is mixed with data items to emulate different environments, filtering or pitch shifting [30], [34]. However, for EEG spectrograms, all these

augmentation methods are not useful, not even those typically used for audio spectrograms since the EEG background noise cannot be easily replicated to be mixed with the EEG signal. Furthermore, the inter or intra individual variation of the EEG signals is highly stochastic and unpredictable. Therefore, we have adopted mixup augmentation, a powerful augmentation technique that alleviates all the limitations of the common augmentation methods. Mixup augmentation implemented at training time to generate augmented images by assigning a random weight λ for the 2 data items to be mixed up.

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \quad (4)$$

where (x_i, y_i) and (x_j, y_j) are 2 randomly chosen items from the training dataset, and $\lambda \sim \text{Beta}(\sigma, \sigma)$ for $\lambda \in [0, 1]$, $\sigma \in (0, \inf)$. We have chosen $\sigma = 0.4$ like what has been suggested in [20], to follow Beta distribution. Following this Beta distribution means that there is a high probability that λ is close to either 0 or 1, which means most of the mixed up image comes from one of the 2 data samples.

III. Results

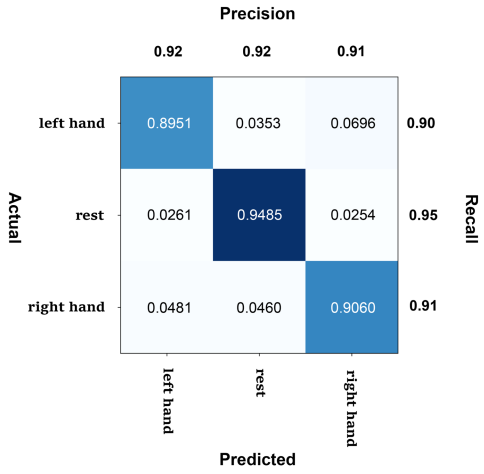
A. Performance of the Classifiers

The training time needed for training each model was more than 10 hours. In order to check as much as possible on refinements, the first 10 subjects have been combined and used for fast experimentations to estimate the refinements' improvement on validation accuracy. Later, after checking all the refinements, the test accuracy of both models were estimated on the entire dataset of 109 subjects and reported as the average accuracy of 109 subjects, by one subject per model method, on the best collection of refinements and model design.This approach worked as a protection against overfitting by holding out a substantial portion of the dataset during the experiments of refinement and parameter tuning.

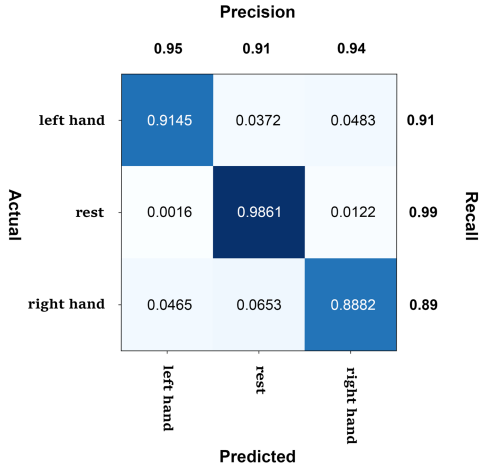
By using Mixup augmentation, the model's accuracy improved from 0.74 to 0.83. With 0.1 weight decay, the accuracy improved further in comparison to the model refinements without using any weight decay, resulting in 0.85 accuracy. Moreover, stacking the spectrograms and eliminating the background resulted in 0.88. Finally, to test the models by using only one subject per model on the entire dataset, the classifier could achieve 0.920 (95% CI 0.908, 0.933) and 0.933 (95% CI 0.922, 0.945) average classification accuracy with ResNet18 and DenseNet121 models, respectively. This is clearly showing that the model could mitigate the inter-individual variability and could achieve higher classification performance despite the smaller dataset. The normalized confusion matrices of both models are plotted in Fig. 4.

B. Ablation Study

Table I shows the refinements that have been carried out throughout the experiments. Each refinement has been added to the previous model settings. If the refinement improved the performance, it was used in all following steps. If the refinement effect was detrimental, the refinement was dropped and we proceeded to apply the following refinement to the best previous model settings. Finally, the highest classification accuracy was achieved by having the following: DenseNet121 with a customized head, mixup augmentation, 1 subject per model, frequency range 2–78 Hz, Adam optimizer, concatenated spectrograms, WD 0.1 and gradient clipping 0.1.



(a) ResNet18 confusion matrix



(b) DenseNet121 confusion matrix

Fig. 4. Normalized confusion matrix of ResNet18 and DenseNet121 models.

C. Performance Comparison

Fig. 5 shows the performance of both ResNet18 and DenseNet121 models on the entire 109 subjects. DenseNet121 model could classify more subjects with perfect accuracy. However, a statistical two-sample t-test between both model's performance showed no significant difference (p-value > 0.05). Table II shows the estimation of the difference of means and the descriptive statistics of the two prediction samples.

TABLE I
REFINEMENTS STACKING AND THEIR EFFECT ON CLASSIFICATION ACCURACY. THE REFINEMENT WAS DROPPED, WHENEVER THE CLASSICIFICATION WAS NOT IMPROVED, AND THE NEXT REFINEMENT WAS ADDED TO THE BEST PREVIOUS MODEL

Refinements	ΔAcc	Acc
ResNet18, freq.: 6-36 Hz, white background, 10 subj. per model, clip grads 0.1, pretrained		0.57
freq.: 2-78 Hz	0.11	0.68
no pretraining	-0.06	0.62
gray background	0.03	0.71
black background	0.02	0.73
random gray level background	-0.14	0.59
random colored background	-0.29	0.44
customized model head	0.01	0.74
no clipping gradients	-0.03	0.71
mixup augmentation	0.09	0.83
weight decay 0.1	0.02	0.85
stacking the spectrograms without background	0.03	0.88
15 central EEG channels	-0.13	0.75
15 central EEG channels, 2x resolution	-0.14	0.74
109 subjects trained on a single model	-0.14	0.74
ResNet18, 1 subj. per model (109 subj.)	**0.04**	**0.92**
DenseNet121, 1 subj. per model (109 subj.)	**0.01**	**0.93**

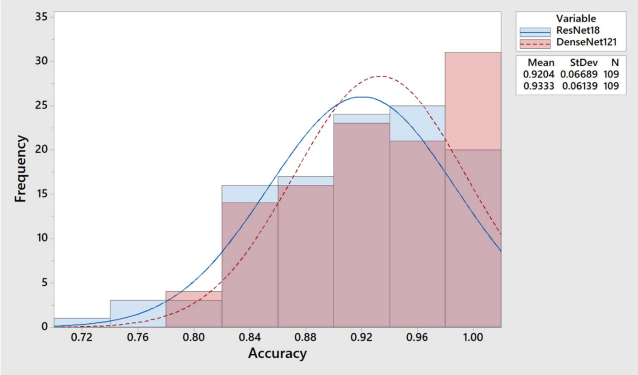


Fig. 5. Histogram of the classification accuracy of ResNet18 and DenseNet121 models.

TABLE II
TWO-SAMPLE T-TEST AND CONFIDENCE INTERVAL (CI) FOR RESNET18 AND DENSENET121 MODELS ACCURACY

	N	Mean	StDev	SE Mean
DenseNet121	109	0.9333	0.0614	0.0059
ResNet18	109	0.9204	0.0669	0.0064

Difference = μ (DenseNet121) - μ (ResNet18)
Estimate for difference: 0.01293
95.0% CI for difference: (0.00421, 0.03007)
t-Test of difference: t-Value = 1.49 P-Value = 0.139 DF = 214

Furthermore, Fig. 6 shows the comparison of DenseNet121 and ResNet18 performance on each subject. The graph shows multiple subjects where ResNet18 outperformed DenseNet121 model's performance, despite that the overall performance of DenseNet121 was better. The cubic polynomial fit in the figures does not refer to any trend, since the x-axis is the subject number, which is a nominal variable and there is no intrinsic ordering.

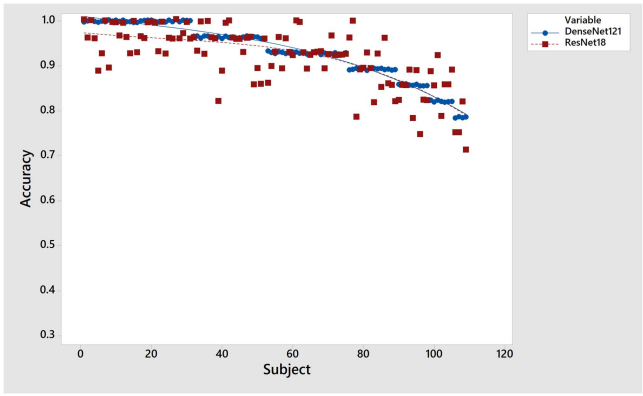


Fig. 6. DenseNet121 and ResNet18 performance with cubic polynomial fit with subjects sorted in descending order according to DenseNet121 performance for clarity.

Nevertheless, the fit makes it clear that the performance of DenseNet121 was strongly correlated with the performance of ResNet18.

The Pearson correlation between both models is considered a high degree correlation ($\rho = 0.735$) [24]. Moreover, the statistical significance test of the Pearson correlation between DenseNet121 and DML performance was estimated as highly significant ($p < 0.001$). The strong correlation is suggesting that the two models were classifying the EEG signals in a similar approach. This is further supported by the statistical insignificance of the two-sample t-test for the difference of means as shown in Table II. In light of these findings, ensembling between DenseNet121 and RenseNet18 classifiers can be considered for future work, since there were multiple subjects where the ResNet18 classifier was performing better, despite that the average accuracy of all subjects with DenseNet121 classifier was higher. Furthermore, the correlation study between DenseNet121 and ResNet18 demonstrates that there were consistently low performing subjects, which may support the previous studies that reported 10-30% of healthy subjects were not able to modulate their EEG for BCI control [1], [2], [3], [10], [22], [23], [27], [28], [31], [33], [37].

IV. DISCUSSION

EEG channels that are mostly associated with motor imagery in EEG studies are usually C3, C4, and Cz. Some studies also chose to add other central EEG channels such as FC3, FC4, C5, C6, CCP3, and CCP4 [26], [39]. Feature selection to decrease the number of features used for training a model is important in traditional ML methods due to the curse of dimensionality [13]. In particular, Hughes Phenomenon shows that when the number of features increases after a certain point, the performance of the model's classification decreases provided that the dataset size is fixed. Hence, most BCI research selected only the most correlated frequency range and EEG channels. DL models are less affected by the curse of dimensionality. Few theories show how neural networks in general and deep learning, in particular, are more immune to larger feature space [14].

Manifold hypothesis and Sparse coding theory suggest that the high dimensional feature space manifold actually sits on top of lower-dimensional feature space embedded in the higher dimensional manifold. Hence, DL neural networks are good at exploiting the pattern in the high dimensional feature space and reducing it to a lower-dimensional manifold. This made it possible for DL models to use more features than most of the previous classical machine learning BCI studies; hence all the 64 EEG channels and more frequency range spectrums have been included. In our experiments, including only 15 central EEG channel spectrograms either with the same resolution of the spectrogram images or with twice higher resolution resulted in 0.12 less accuracy and 0.11 respectively in comparison to including all the 64 channels. The improvement of accuracy shows clearly that the CNN models have trained useful features outside the traditional features that are used for BCI. These findings have been supported by several neurophysiological studies that reported the involvement of brain regions other than the regions covered by the central EEG channels [8], [19], [36].

The evidence to date demonstrates that the adaptation of CNS to control BCI using direct output from the cerebral cortex is indeed possible, albeit imperfect. BCI systems are less smooth, less accurate, and with more trial-to-trial variability. While robust BCI systems are desirable, it does not need to be perfectly reliable to attain wide adoption. There are few medical conditions like locked-in syndrome or other types of physical disabilities, where imperfect BCI is still helpful to enable them to attain some control in certain environments.

V. Conclusion

In this study, we propose a novel approach to classify MI-EEG signals using Stockwell Transform, two custom CNN classifiers and mixup augmentation trained on small training sets of only one subject per model, despite that DL models typically need a large amount of data for training. Building a DL model that can be trained on an extremely small EEG training set of a single subject presents an interesting challenge that this work is trying to address since neither long EEG session recording from a single user is feasible due to user fatigue, nor collecting EEG dataset from multiple users is desirable due to the inter-individual variability that leads to decrease of classification performance.

Stockwell Transform has been employed to preprocess EEG epochs due to its advantage in comparison to the other common EEG preprocessing methods with its implicit phase-normalized frequency bands. This renders the transformed EEG signals in the frequency domain distortion-free, which yields a better representation of the EEG features in the frequency domain.

Moreover, mixup augmentation has been utilized for the first time in BCI EEG classification. It has a unique approach of augmenting the dataset in a way that is more suitable for EEG signals in comparison to the other common augmentation methods. This is due to the inter and intra individual variations of the EEG signals which are highly stochastic and unpredictable that hinder the current commonly used augmentation methods to mimic such variability. We showed that mixup augmentation used with CNN models enhanced the performance of the classification significantly, suggesting that this augmentation technique is an important tool to be introduced for EEG classification methods.

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

Supplementary Materials

Additional figure and more details are included in supplementary materials. There is an elaboration on the current literature and details of the methods along with an extensive discussion of the results.

References

[1] B. Z. Allison et al., "Toward a hybrid brain–computer interface based on imagined movement and visual attention," *J. Neural Eng.*, vol. 7 no. 2, 2010, Art. no. 026007.
[2] B. Z. Allison et al., "Towards an independent brain–computer interface using steady state visual evoked potentials," *Clin. Neuriophysiol.*, vol. 119, no. 2, pp. 399–408, 2008.
[3] B. Z. Allison, E. W. Wolpaw, and J. R. Wolpaw, "Brain–computer interface systems: Progress and prospects," *Expert Rev. Med. Devices*, vol. 4, no. 4, pp. 463–474, 2007.
[4] H. Alwasiti, M. Z. Yusoff, and K. Raza, "Motor imagery classification for brain computer interface using deep metric learning," *IEEE Access*, vol. 8, pp. 109949–109963, 2020.
[5] H. Alwasiti and M. Z. Yusoff, "Shredded control of drones via motor imagery brain computer interface," *Compusoft*, vol. 9, no. 3, pp. 3606–3610, Mar. 2020.
[6] 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.
[7] 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.
[8] M. Bakker, F. P. De Lange, R. C. Helmich, R. Scheeringa, B. R. Bloem, and I. Toni, "Cerebral correlates of motor imagery of normal and precision gait," *Neuroimage*, vol. 41, no. 3, pp. 998–1010, 2008.
[9] H. Berger, "On the EEG in humans," *Arch. Psychiatr. Nervenkr*, vol. 87, pp. 527–570, 1929.
[10] 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, 2007.
[11] 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.
[12] N. E. Crone, D. L. Miglioretti, B. Gordon, and R. P. Lesser, "Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. event-related synchronization in the gamma band," *Brain: A J. Neurol.*, vol. 121, no. 12, pp. 2301–2315, 1998.
[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012.
[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT press, 2016, pp. 436–444.
[15] B. Graimann, B. Z. Allison, and G. Pfurtscheller. *Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction*. Berlin, Germany: Springer, 2010.
[16] J. Hammer, T. Pistohl, J. Fischer, P. Kršek, M. Tomášek, and P. Marusič, "Andreas schulze-bonhage, ad aertsen, and tonio ball. predominance of movement speed over direction in neuronal population signals of motor cortex: Intracranial EEG data and a simple explanatory model," *Cereb. Cortex*, vol. 26, no. 6, pp. 2863–2881, Jun. 2016.
[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. conf. mach. learn.*, pp. 448–456, 2015.
[18] A. Krogh and J. A. Hertz, "A. simple weight decay can improve generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 950–957, 1992.

[19] M. G. Lacourseac, E. L. R. Orr, S. C. Cramere, and M. J. Cohen, "Brain activation during execution and motor imagery of novel and skilled sequential hand movements," *Neuroimage*, vol. 27, no. 3, pp. 505–519, 2005.
[20] D. Liang, F. Yang, T. Zhang, and P. Yang, "Understanding mixup training methods," *IEEE Access*, vol. 6, pp. 58774–58783, 2018.
[21] 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 Topogr.*, vol. 12, no. 3, pp. 177–186, 2000.
[22] F. Nijboer et al., "An auditory brain–computer interface," *J. Neurosci. Methods*, vol. 167, no 1, pp. 43–50, 2008.
[23] A. Nijholt et al., "Brain-computer interfacing for intelligent systems," *IEEE Intell. Syst.*, vol. 23, no. 3, pp. 72–79, May/Jun. 2008.
[24] A. Nijholt et al., "Directory of statistical analyses. pearson's correlation coefficient - statistics solutions," 2020. Accessed: 06, 2020. [Online]. Available: https://www.statisticssolutions.com/pearsons-correlation-coefficient
[25] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Netw. Vis. Recognit.*, vol. 11, pp. 1–8, 2017.
[26] G. Pfurtscheller and F.H. Lopes Da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurio-physiol.*, vol. 110, no. 11, pp. 1842–1857, 1999.
[27] 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, 2008.
[28] G. Pfurtscheller et al., "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.
[29] 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, 2012.
[30] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.
[31] C. Sannelli, M. Braun, M. Tangermann, and K.-R. Müller, *Estimating Noise and Dimensionality in Bci Data Sets: Towards Illiteracy Comprehension*. 2008.
[32] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
[33] 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.
[34] D. Stowell, T. Petrusková, M. Šálek, and P. Linhart, "Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions," *J. Roy. Soc. Interface*, vol. 16, no. 153, 2019, Art. no. 20180940.
[35] 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, 2016, Art. no. 016003.
[36] B. C. M. van Wijk, V. Litvak, K. J. Friston, and A. Daffertshofer, "Nonlinear coupling between occipital and motor cortex during motor imagery: A dynamic causal modeling study," *Neuroimage*, vol. 71, pp. 104–113, 2013.
[37] Theresa M. Vaughan et al., "The wadsworth BCI research and development program: At home with BCI," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 229–233, 2006.
[38] J. Wolpaw and E. W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice*, OUP: USA, 2012.
[39] H. Yuan, A. Doud, A. Gururajan, and B. He, "Cortical imaging of event-related (de) synchronization during online control of brain-computer interface using minimum-norm estimates in frequency domain," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 5, pp. 425–431, Oct. 2008.
[40] G. Zhang, C. Wang, B. Xu, and R. Grosse, "Three mechanisms of weight decay regularization," in *Proc. Int. Conf. Learning Representations*, 2019.