BIOCYBERNETICS AND BIOMEDICAL ENGINEERING 41 (2021) 946–959

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/bbe

Original Research Article

Prediction of drug response in major depressive disorder using ensemble of transfer learning with convolutional neural network based on EEG

Mohsen Sadat Shahabi [a], *Ahmad Shalbaf* [a,*], *Arash Maghsoudi* [b]

[a] *Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[b] *Department of Biomedical Engineering, Science and Research Brach, Islamic Azad University, Tehran, Iran*

ARTICLE INFO

Article history:
Received 13 January 2021
Received in revised form
10 June 2021
Accepted 12 June 2021
Available online 20 June 2021

Keywords:
Depression
Electroencephalogram
Transfer learning
Convolutional neural network
Continuous wavelet transform

ABSTRACT

Major Depressive Disorder (MDD) is one of the leading causes of disability worldwide. Prediction of response to Selective Serotonin Reuptake Inhibitors (SSRIs) antidepressants in patients with MDD is necessary for preventing side effects of mistreatment. In this study, a deep Transfer Learning (TL) strategy based on powerful pre-trained convolutional neural networks (CNNs) in the big data datasets is developed for classification of Responders and Non-Responders (R/NR) to SSRI antidepressants, using 19-channel Electroencephalography (EEG) signal acquired from 30 MDD patients in the resting state. Multiple time–frequency images are obtained from each EEG channel using Continuous Wavelet Transform (CWT) for feeding into pre-trained CNN models that are VGG16, Xception, DenseNet121, MobileNetV2 and InceptionResNetV2. Our plan is to adapt and fine-tune the weights of networks to the target task with the small-sized dataset. Finally, to improve the recognition performance, an ensemble method based on majority voting of outputs of five mentioned deep TL architectures has been developed. Results indicate that the best performance among basic models achieved by DenseNet121 with accuracy, sensitivity and specificity of 95.74%, 95.56% and 95.64%, respectively. An Ensemble of these basic models created to surpass the accuracy obtained by each individual basic model. Our experiments show that ensemble model can gain accuracy, sensitivity and specificity of 96.55%, 96.01% and 96.95%, respectively. Therefore, proposed ensemble of TL strategy of pre-trained CNN models based on WT images obtained from EEG signal can be used for antidepressants treatment outcome prediction with a high accuracy.

© 2021 Nalecz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences. Published by Elsevier B.V. All rights reserved.

* Corresponding author at: Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
E-mail address: shalbaf@sbmu.ac.ir (A. Shalbaf).

https://doi.org/10.1016/j.bbe.2021.06.006

0168-8227/© 2021 Nalecz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences. Published by Elsevier B.V. All rights reserved.

1. Introduction

Major Depressive Disorder (MDD) is one of the leading causes of disability worldwide and a life threatening disease that increases suicidal thoughts in patients and must be treated in time [1]. Most common treatment option for MDD patients in clinics are prescribed antidepressants that used as first-line treatment [2]. However, only half of MDD patients respond to treatment using antidepressants [3]. Due to heterogeneity of the condition, choosing the right antidepressant is based on experience and sometimes based on trial and error, thus its effectiveness is limited [4]. If treatment failed, a sufficient period of 2 to 4 weeks is wasted. The limitation in treatment effectiveness, necessitates for prediction of response to a specific treatment. Many researchers tried to discriminate MDD patients that respond to antidepressants using various predictors including neuroimaging, sociodemographic, symptom profile, genetics, clinical comorbidities, and peripheral markers [5]. Among these methods, neuroimaging methods based on Electro-Encephalogram (EEG) signal obtained from scalp electrodes is a popular choice and has special advantages due to its easy-to-setup, low cost, high temporal resolution and clinical acceptance [6–11].

Some researchers tried to predict treatment outcome for antidepressants prior to treatment using EEG signal. Wajid et al. extracted time–frequency features from different frequency bands of EEG signal and classified them using a logistic regression model [12]. Three time–frequency decomposition techniques including wavelet transform (WT) analysis, short-time Fourier transform (STFT) and empirical mode decomposition (EMD) used for classifying MDD patients as responder and non-responder. Best classification accuracy of 91.6% achieved using combination of best features from there mentioned decomposition methods. Khodayari-Rostamabad et al. extracted a set of candidate features from pre-treatment resting-state EEG signal, including power spectral density, squared spectral coherence, mutual information, left-to-right hemisphere and anterior/posterior power ratios of all channels and best features selected based on Fisher discriminant ratio [13]. Then a Mixture of Factor Analysis (MFA) classifier used to classify treatment outcome based on multiple features and accuracy of purposed model reached to 87.9%. Zhdanov et al. predicted antidepressants treatment outcome using clinical and pre-treatment EEG data of 122 participants from Canadian Biomarker Integration Network in Depression (CAN-BIND-1) study [14]. A Support Vector Machine (SVM) classifier was able to identify responders with 79.2% accuracy. Jaworska et al. used multiple machine learning algorithms including Random Forest classifier to predict response to antidepressant using multiple features extracted from pre-treatment EEG signal, clinical information and source-localized current density [15]. When most important features of three methods combined together, classification model accuracy raised to 88%. Rajpurkar et al. developed a new machine learning algorithm to predict outcomes of antidepressant treatment using pretreatment EEG that uses Gradient Boosted Decision Tree (GBDT) classifier [16]. EEG features including absolute/relative power of EEG frequency bands in occipital and frontal regions, and beta-alpha ratios are considered. Results indicated that most important EEG features were absolute delta band power at the occipital electrode sites (O1 and Oz) with highest C-index score of 0.963. Moreover, Van der Vinne et al. investigated Frontal Alpha Asymmetry (FAA) feature stability in 423 MDD patients and concluded that FAA is a stable EEG feature over time and type of antidepressant [17]. Actually FAA is asymmetry of EEG signal power between left and right hemispheres in alpha band power of frontal electrodes. In another studies EEG biomarker for prescription of antidepressants in MDD are investigated [18].

As mentioned above, EEG signal classification of responders and non-responders using traditional machine learning methods needs feature extraction, perform feature selection and finally conventional classification methods [19]. But in recent years, deep neural networks that mimics human brain hierarchical structure of neurons, has gained so much attention in the field of image and signal classification, natural language processing and many other fields as a disruptive alternative to the aforementioned feature based methods [20–25]. By using deep neural networks and especially convolutional neural networks (CNNs), best classification accuracies acquired without feature extraction in medical image classification problems [26,27]. These algorithms are able to automatically extract significant features and classify them directly from the data. CNN models has been utilized for EEG classification by converting EEG from signal space to image space [28,29]. For example Garg et al used WT to create 2-D scalogram images from EEG signals as input of a CNN model for emotion recognition problem [30]. Li combined WT images and a simplified CNN model for classification of motor imagery EEG signals to the left and right hand tasks [32]. Shovon et. al utilized STFT images of three EEG channels as inputs to a multi-input CNN model for motor imagery EEG signal classification [31]. Also, some works have done for EEG classification of depressed patients from healthy subjects using CNNs [6,8,32–34]. Mumtaz et al used one-dimensional CNN as their base model and achieved 98% accuracy in classification of healthy and MDD subjects [32]. In another study, Dang et al used time–frequency representation of multichannel EEG signals in different frequency bands as inputs to a multilayer CNN network for classification of healthy and MDD subjects [33].

However, CNN models have tremendous number of parameters and hyper-parameters and need a lot of training data to optimize all of them [35]. Since there are many obstacles against obtaining proper clinical EEG data, then gathering huge datasets might be very problematic. This problem can be minimized using deep Transfer Learning (TL) strategy so that CNN models have already been pre-trained on huge number of samples in a specific domain and then to be used for a slightly different domain with limited training data [36]. TL can significantly reduce number of parameters and resources needed for model development. Zhang et al proposed a deep TL model based on VGG16, VGG19 and ResNet50 models for cross-subject seizure detection [37]. They used 'ImageNet' initial weights and fine-tuned top layers with two neurons as output layer for binary classification. Their model accuracy reached to 98% for seizure/non-seizure classification. Cimtay

Table 1 – Summary of MDD patient's clinical characteristics [12]

Information	Responder	Non-Responder	Total
Age [years]	40.7 (±13.0)	41.1 (±12.5)	40.3 (±12.9)
Gender (female/male)	8/8	9/9	17/17
Pretreatment BDI-II	18.4 (±7. 4)	22.8 (±12.5)	20.6 (±8.6)
Post-treatment BDI-II	9.1 (±6.3)	22.1 (±3.3)	15.6 (±4.5)

et al used InceptionResNetV2 as a feature extractor for emotion recognition using 3-D images created from EEG signals of multiple datasets [38]. Classification capability of pre-trained CNN model is enhanced using extra pooling and dense layers added to the top of basic model. Shalbaf et al used AlexNet, ResNet-18, VGG-19 and Inception-v3 models for EEG-based diagnosis of schizophrenia patients from healthy controls [39]. Continuous WT utilized to create time–frequency representations from EEG signal. Features extracted by these pre-trained networks are fed to a SVM classifier for classification and best accuracy of 98.6% gained for ResNet-18 pre-trained network. Raghu et al investigated ten pre-trained networks for the recognition of epileptic seizure types as an 8-class classification problem [40]. In this study, Alexnet, Vgg16, Vgg19, Squeezenet, Googlenet, Inceptionv3, Densenet201, Resnet18, Resnet50, and Resnet101 used as pre-trained networks or as feature extraction models for a SVM classifier. Stacks of spectrogram images were fed to each network and finally InceptionV3 and Googlenet achieved highest classification accuracies.

Aim of this work is developing a deep TL strategy based on powerful pre-trained CNNs, to predict antidepressant treatment outcome of MDD patients, using time–frequency images obtained from WT of pre-treatment resting-state EEG signal. Our plan is to learn a powerful CNN with previous training in the big data datasets and then adapt its network weight to the target task with the small-sized dataset. We use ensemble of different CNNs based TL methods for obtaining best performance for classification of MDD patients as responder and non-responders to antidepressants.

2. Methods

2.1. Dataset

In this study, the resting-state 19-channel EEG recording of 60 subjects in eye-close (EC) situations is utilized [12][1]. 30 subjects diagnosed as MDD based on DSM-IV criteria and received prescription for treatment using Selective Serotonin Reuptake Inhibitors (SSRIs) antidepressants and 30 remaining subjects were healthy controls. After 4 weeks of regular dose, 12 patients responded to treatment based on Beck Depression Inventory (BDI) measure while 18 patients did not show significant improvement. Response to SSRI treatment is defined as 50% improvement in clinical symptoms assessed with the pre- vs. post-treatment BDI scores [41–43]. In other words, response to SSRI treatment means that the score of questionnaires of BDI after a treatment period shows at least 50 percent decrease in depression severity. In this work, 12 MDD patients are present as responder and 18 patients are as non-responder. Clinical characteristics are summarized in Table1. EEG data were recorded for 5-minutes using19-electrode EEG cap with linked-ear reference from the scalp. All electrodes are placed according to international 10–20 system. Five pack of electrodes assigned to five regions of brain. Frontal region includes Fp1, F3, F7, Fz, Fp2, F4, and F8 electrodes, parietal lobe includes P3, Pz and P4 electrodes, occipital region contains O1 and O2 electrodes, left and right temporal regions include T3, T4, T5, T6 electrodes and finally C3, C4 and Cz electrodes assigned to the central region of brain. We have done the pre-processing steps used by Mumtaz et al. in the same data [12]. The sample rate was 256 Hz and signals were band-pass filtered in 0.5–70 Hz [12]. Notch filter in 50 Hz applied to reject power line noise. In this study, 126 s of EEG data of each subject was utilized and remaining data were discarded.

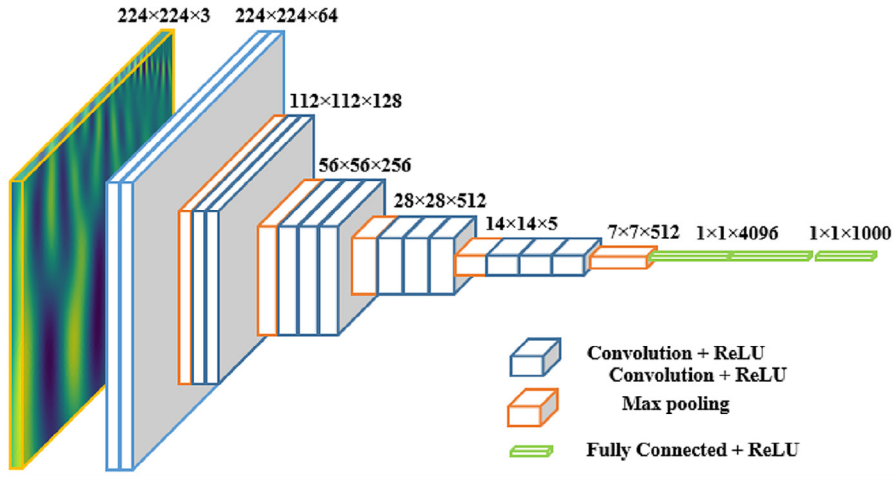
2.2. EEG signal to image conversion

CNN models can be utilized for EEG classification by converting EEG from signal space to image space. In this study, we used WT to create 2-D scalogram images as 2D time–frequency maps from EEG signals as input of a CNN model. Continuous WT uses multiple scales of a basis function to decompose a signal and provides detailed analysis of a signal [12]. WT has the ability to illustrate EEG features by decomposing time–frequency patterns. Because of non-stationary nature of EEG signal, wavelet decomposition yields a detailed and robust representation of signal. WT of each 3 s of EEG signal calculated using 'morlet' mother function and scale parameter between 1 and 256, thus the size of each 2D time–frequency map will be 256*768. Three sequential images from 9 s of EEG signal are concatenated to create three channels of a RGB image and finally, this 3D image is resized to feed into the CNN model. So from each subject with 126 s, we created 14 (126/9) input 3D image. Finally, with 126 s of 19 EEG channels for each of 30 MDD patients, then 7980 (14*19*30) input 3D images are produced.

2.3. Cnn

CNN models constructed from multiple convolutional layers that extract spatial dependencies in the input image and create different abstract representations in each layer using filters with various sizes. By stacking multiple of these convolutional layers, a CNN model can automatically produce best features

[1] Data is available from: https://figshare.com/articles/EEG_Data_New/4244171




```
def create_model():
  conv_base = VGG16(weights='imagenet',include_top=False,input_shape=(224,224,3))
model = models.Sequential()
....
model.compile(loss='binary_crossentropy', optimizer=Adam(lr=lr),metrics=['acc'])
```

Fig. 1 – VGG-16 architecture and its code.

for feeding to a classifier. Different pre-trained CNN models are used for EEG classification. Most of these architectures built upon simple topology of VGG model [44] that actually is a stack of multiple convolutional layers. Some new concepts like inception blocks [45], residual learning and skip connection [46] combined to create models with more capabilities. These networks have a lot of weights that should be trained and have hyper parameters, including batch size and learning rate that shall be optimized for each model separately. In this work, five different pre-trained CNN models used that are VGG16, DenseNet121, InceptionResNetV2, Xception and MobileNetV2. These pre-trained CNN models are trained on ImageNet dataset that consider 1000 classes as classification output. In this work, Adaptive moment estimation (Adam) optimizer using binary cross-entropy loss function used for fine-tuning of the models through 50 epochs. Learning rate and batch size were optimized for each model in the validation process to reach the maximum performance. Classification layer jointly trained with pre trained CNN model to predict responders and non-responders categories. Keras[2] package used for implementation of these models and training phase performed on Nvidia K80 GPU with 12 GB RAM from Google Colaboratory. Pre-trained models utilized in this paper and also implementation and training of each model is done in keras package. Python programming language is used to write functions to utilize keras methods for model creation and training. These keras methods are included 'applications' for model creation and 'fit' and 'compile' for training.

2.3.1. *Vgg 16*

One type of pre-trained CNN models that is used in this study is VGG-16 model. Simplicity of VGG architecture and its modular pattern in layers made it famous and widely popular for image classification [47]. VGG-16 uses small convolution filters for decreasing computational complexity and a deeper network architecture with 16 layers [44]. Fig. 1 shows a schematic of VGG-16.

2.3.2. *Xception*

Xception model derived from the idea of inception module [48]. This architecture uses 1×1 convolutions followed by 3×3 convolutions to firstly reduce dimensionality in filter space and secondly save spatial information [45]. This approach can make network efficient by decoupling spatial and channel correlation [47]. Fig. 2 shows the architecture of Xception module.

2.3.3. *DenseNet*

DenseNet [49] used cross-layer connectivity between each preceding layer to all subsequent layers in feed-forward fashion. Direct connection between layers improves flow of information in the network through gradients shared among layers in back propagation algorithm, therefore, this architecture can minimize vanishing gradient problem [47]. This architecture is shown in Fig. 3.

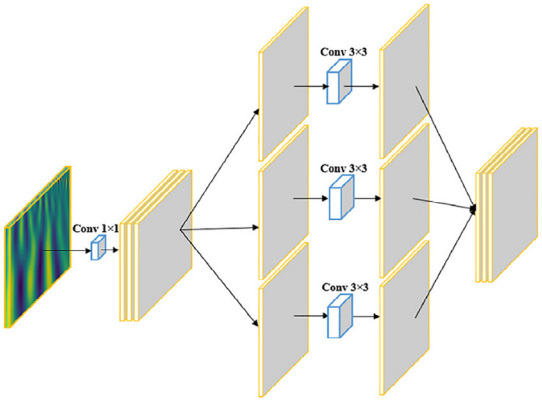
2.3.4. *MobileNetV2*

MobileNet model is introduced for embedded systems and mobile applications to smoothly do the computer vision tasks on devices with low processing capacity. It's based on depth-wise separable convolutions to build light weight deep neural networks [50]. A newer version of this architecture called MobileNetV2 is used in this study (Fig. 4) [51].

2.3.5. *Inception-ResNetV2*

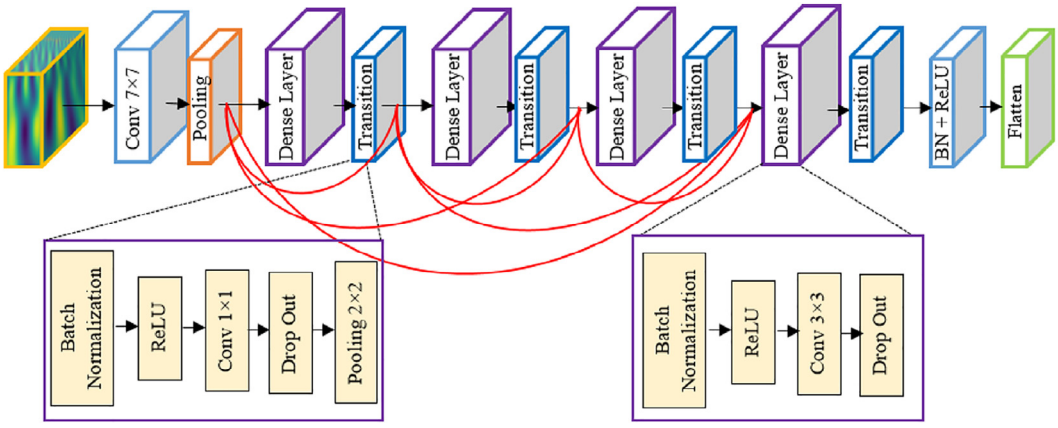
Inception-ResNetV2 architecture is based on Inception blocks that use split, transform and merge functions for feature extraction. Using small and asymmetric filters in inception

[2] https://keras.io/




```
def create_model():
  conv_base = Xception(weights='imagenet',include_top=False,input_shape=(299,299,3))
model = models.Sequential()
....
model.compile(loss='binary_crossentropy', optimizer=Adam(lr=lr),metrics=['acc'])
```

Fig. 2 – Xception building block and its code.




```
def create_model():
  conv_base = DenseNet121(weights='imagenet',include_top=False,input_shape=(224,224,3))
model = models.Sequential()
....
model.compile(loss='binary_crossentropy', optimizer=Adam(lr=lr),metrics=['acc'])
```

Fig. 3 – DenseNet121 Architecture and its code.

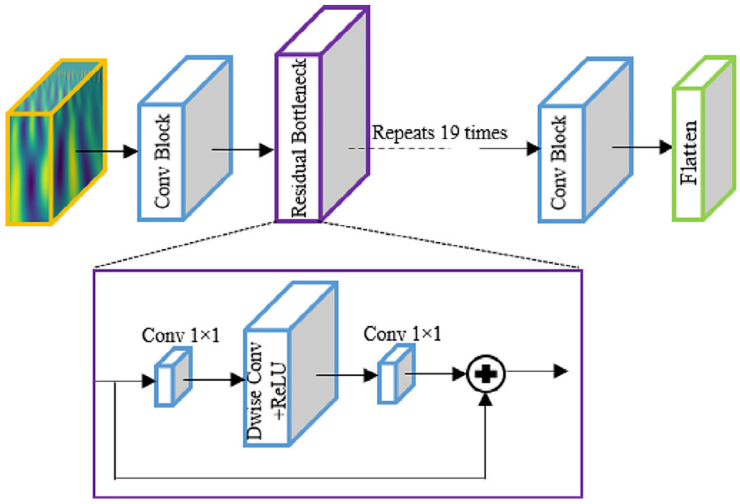
module and combining them by 1*1 convolutions as a bottleneck, produces a cross-channel correlation between different paths in model. Inception-ResNetV2 combines the power of residual learning and inception block (Fig. 5.) [47].

2.4. *Fine-Tune*

For fine-tuning of each model, 'imagenet' weights used as initial weights of networks and then top fully-connected layers discarded and replaced by a task-specific classifier and finally whole layers of models leaved to be trainable. Top fully connected layer/layers should be replaced by two other fully connected layers for the purpose of binary classification. A fully connected layer with 128 or 256 neurons and ReLu activation function followed by a single neuron with a sigmoid activation function used for binary classification of abstract features extracted from EEG images. After each fully-connected layer, a batch normalization layer used to normalize extracted features against their mean and standard deviation. The first fully connected layer is equipped by L1 and L2 regularizers for each neuron weight, bias and activation function to control parameter changes and prevent overfitting of network.

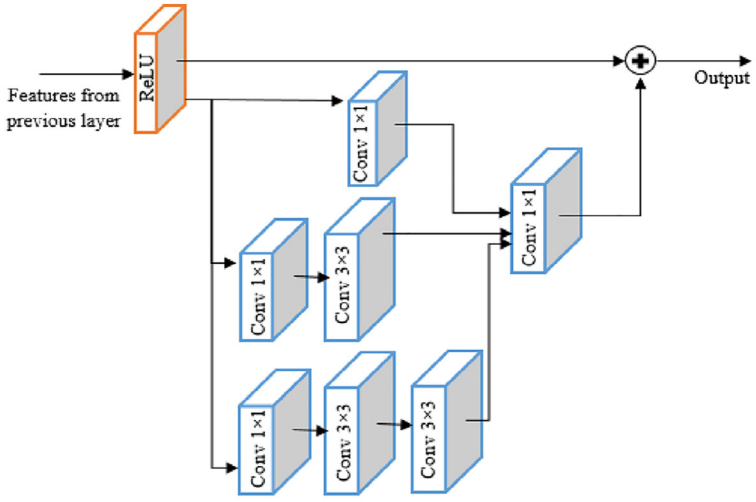
2.5. *Performance evaluation*

Performance of different TL models are compared using accuracy, sensitivity, specificity, F1-score measures, Receiver Operating Curve (ROC) and Precision-Recall Curve (PRC). ROC curves used for model performance measurement and high




```
def create_model():
  conv_base = MobileNetV2(weights='imagenet',include_top=False,input_shape=(224,224,3))
model = models.Sequential()
....
model.compile(loss='binary_crossentropy', optimizer=Adam(lr=lr),metrics=['acc'])
```

Fig. 4 – MobileNetV2 Architecture and its code, Dwise Conv stands for Depth wise Convolution.




```
def create_model():
  conv_base = InceptionResNetV2(weights='imagenet',include_top=False,input_shape=(299,299,3))
model = models.Sequential()
....
model.compile(loss='binary_crossentropy', optimizer=Adam(lr=lr),metrics=['acc'])
```

Fig. 5 – Building Block of Inception-ResNetV2 and its code.

value of AUC (Area Under ROC Curve) shows that model works well. ROC curve shows True Positive Rate (TPR) against False Positive Rate (FNR) when classification threshold changes between 0 and 1. A skillful model would bow up to top left of plot. 12 responders and 18 non-responders in our dataset, makes a 50% imbalance for class distribution in this study. This is not a very bad case of data imbalance, but should be considered in model development and evaluation. Therefore PRC curve used here to investigate each model's performance against majority and minority classes. PRC curve plots precision against recall for thresholds between 0 and 1. Precision and recall do not use true negative in their calculations, therefore PRC is not affected by majority class predictions. Along with these measures, confusion matrix also shows details of a model performance using, the proportion of responders that were correctly classified (i.e., true positives [TP]), the proportion of non-responders that were correctly classified (i.e., true negatives [TN]), the proportion

of responders that were misclassified as non-responders (i.e., false positive [FP]) and the portion of non-responders that were misclassified as responders (i.e., false negative [FN]) [15].

Totally 7980 images (14*19*30) were extracted from EEG signal of 30 MDD patients and then all images uniformly were shuffled and from these images, 1596 images are separated as final test data and 6384 images are used for train and validation of model. In this study, for train and validation of model, 10-fold cross validation method used for calculating model generalizability. In each fold of a 10-fold cross validation, 10% of train data hold out for validation and tuning hyper-parameters. Hyper-parameters of each CNN model tuned separately for exceling accuracy of the network. Performance measures of each fold calculated on the validation set and averaged on different folds. When hyper-parameters tuned and best model configuration with highest validation accuracy selected, it will be trained on whole train data (6384 images) and used to predict new unseen test data (1596 images). Comparison of these models for WT input images have been done using accuracy, sensitivity, specificity, F1-score and AUC. Then in second step all models from the first step are used to create an ensemble of basic models.

3. Results

Prediction of response to antidepressants is investigated using pretreatment EEG signals recorded in resting state. The procedure is shown in Fig. 6. After pre-processing of acquired EEG signal, we segmented EEG signal of each single channel to constant length non-overlapping windows (nine second) and converted to 3D color images as explained in previous section using WT transform. WT decomposes original signal to combinations of multiple scales and translations of mother function. Time-frequency representation of EEG signal can magnify specific patterns that highlights responders and non-responders differences. Independently, 5 versions of pre-trained CNNs model named VGG16, Xception, DenseNet121, MobileNetV2 and InceptionResNetV2 on ImageNet datasets were fine-tuned on our dataset of images, with the goal of transferring the information into our task that has limited training data. Each CNN model can handle a specific input image size, then input images created using WT (256*768*3) should be resized for feeding into each CNN model. VGG-16, MobileNetV2 and DensNet121 accepts input images of size 224*224*3. While, InceptionResNetV2 and Xception uses input image size of 299*299*3. Totally 7980 (14*19*30) images were extracted from all 19 channels of resting-state EEG of 30 MDD patients. Then 1596 images are separated as final test data and 6384 images are used for train and validation of model. 10-fold cross validation used for training the basic TL models. Because of non-linear behavior of EEG data and high variance of CNN models, training a convolutional model will suffer from high variability. This problem can be minimized by averaging the accuracy from different folds of cross-validation. In this manner, the basic model can learn different patterns laid in different parts of data without missing a portion of training data for model development process. Learning rate and batch size were optimized for each model in the validation process to reach the maximum performance (as mentioned in table 2).

Average learning curves including accuracy and loss, for training and validation data in all 10 folds are presented in Figs. 7 and 8. Shaded area around each plot shows maximum and minimum values that were achieved by each curve in all folds of cross-validation. In other words, shaded area shows maximum fluctuation that a model had in all folds. Performance of all models in Fig. 7 exponentially grows up to reach a plateau after 50 epochs. This indicates that continuous WT with "morlet" mother function is a reliable tool for extracting time–frequency features of EEG signal. If we keep training the models, it is possible to lose model performance and the accuracy may decrease. As depicted in Figs. 7 and 8, training accuracy is higher than the accuracy of validation and training loss is lower than validation loss. That is because of model ability to learn various patterns in training data that are not necessarily discriminating patterns of validation data. In these figures, some models are more stable than others which have more fluctuations, but finally all models converge. Stability of model performance in all folds shows reliability of trained model, reproducibility of results and generalizability of the model. There is negligible fluctuation in training phase but a more considerable fluctuation in validation phase.

When models passed through 10 fold cross validation and hyper-parameters are tuned for best performance (table 2), we trained each model on all 6384 train images and calculate performance measures on 1596 images of new unseen test data

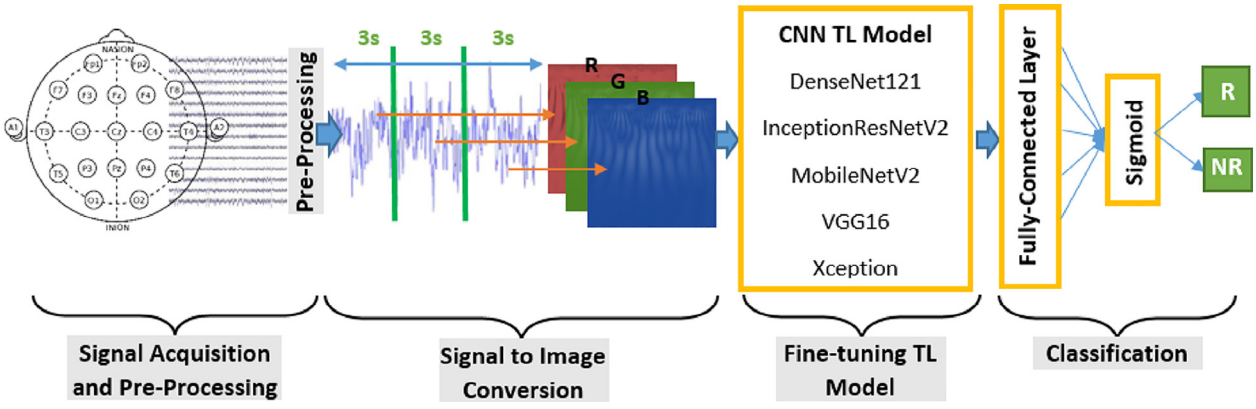


Fig. 6 – Different stages passed for prediction of treatment outcome using EEG signals.

Table 2 – Main hyper-parameters of TL models.

Model Name	batch_size	Learning Rate	FC_layer Size	Input image
VGG16	64	8E-06	256	224*224*3
InceptionResNetV2	64	3E-04	128	299*299*3
Xception	64	2E-04	128	299*299*3
MobileNetV2	16	9E-05	256	224*224*3
DensNet121	64	3E-04	128	224*224*3

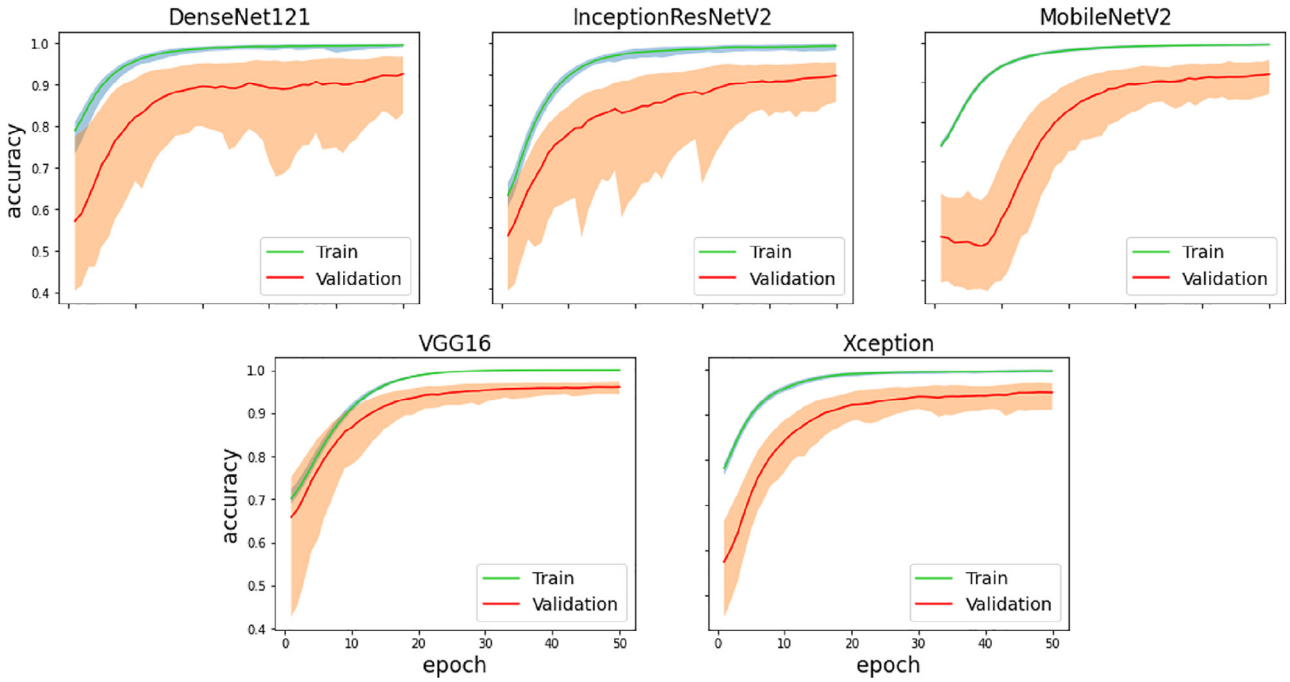


Fig. 7 – Accuracy curves of train and validation of cross-validation process.

to evaluate of the models. Fig. 9 shows accuracy and loss curves for training of final models on whole train data. All performance metrics for evaluation of final models on test set are summarized in table 3. Results indicate that DenseNet121 has the best performance in the test dataset regarded to accuracy, sensitivity, specificity, F1-score and AUC that are respectively 95.74%, 95.56%, 95.64%, 95.6% and 0.988. Also, VGG16 model has the next top performance by accuracy of 94.74%. ROC and PRC curves for evaluating the performance of all models are also shown in Fig. 10. ROC curves in Fig. 10 are sticking to left top corner that means AUC values are near 1 as shown in table 3. In addition, ROC curves of all models show very high values of AUC that means models are stable. Also, PRC curves show very small perturbations and keep having high AUC values of ROC. PRC curve compares each model's precision against its recall and is more informative than ROC curve especially in the case of imbalanced data that is the case of this study. Finally, a popular and effective way to generalize and enhance the performance of a classification model is using ensemble methods. Therefore, an ensemble model created using majority voting among all basic TL models. The results in table 3 showed that this combination of five basic models improved overall accuracy up to 96.55%. Sensitivity, specificity and F1-score of ensemble model are 96.01%, 96.95% and 96.41% respectively that shows an improvement compared to each single model. Confusion matrixes of all TL models and final ensemble model are presented in Fig. 11. Here, true labels and predicted labels are shown respectively in rows and columns of confusion matrix. Highest value of TP belongs to MobileNetV2 and highest value of TN belongs to ensemble model. On the other hand, lowest values of FP and FN belong to the ensemble model and MobileNetV2 respectively.

4. Discussion

In the present study, we investigated EEG classification for responders and non-responders to SSRI antidepressants using ensemble of different pre-trained CNN models based on TL. We used WT images obtained from EEG signal as input to the CNN models to compare the effect of different representations of EEG signals on the classification accuracy. We explored 5 state-of-the-art pre-trained CNN architectures, namely VGG16, Xception, DenseNet121, MobileNetV2 and InceptionResNetV2. These pre-trained networks are fine-tuned on the target task and DenseNet121 model is the best

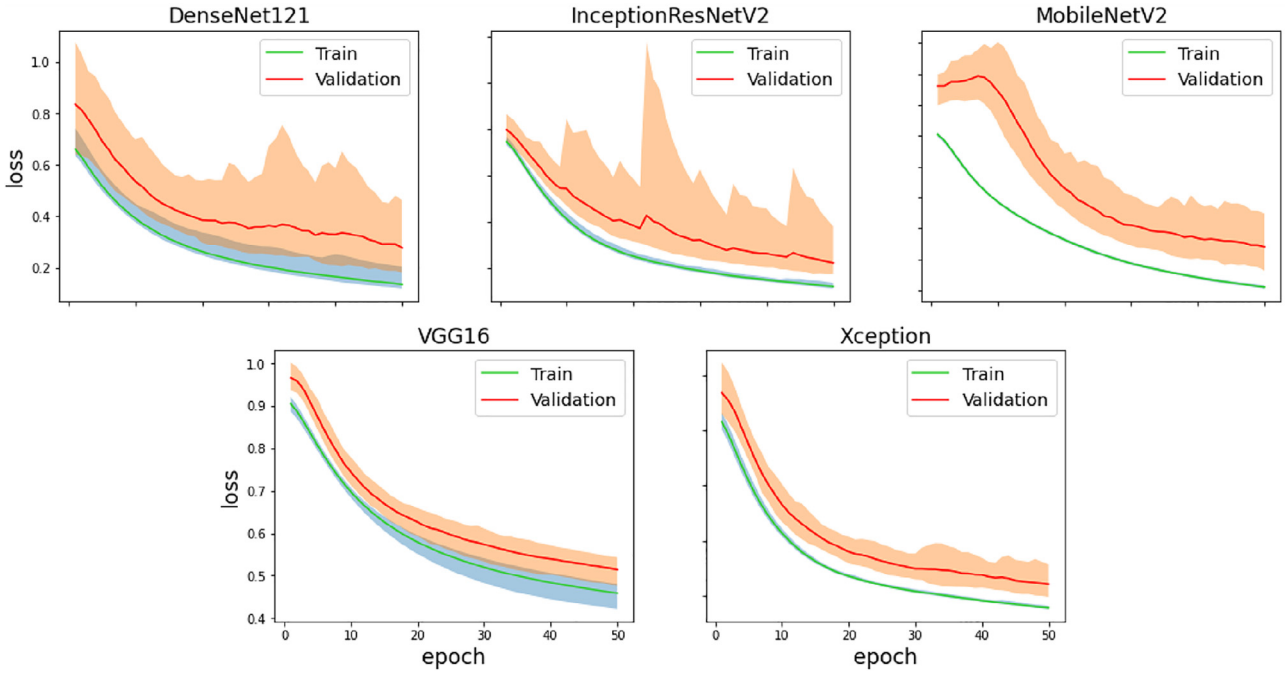


Fig. 8 – Loss curves of train and validation of cross-validation process.

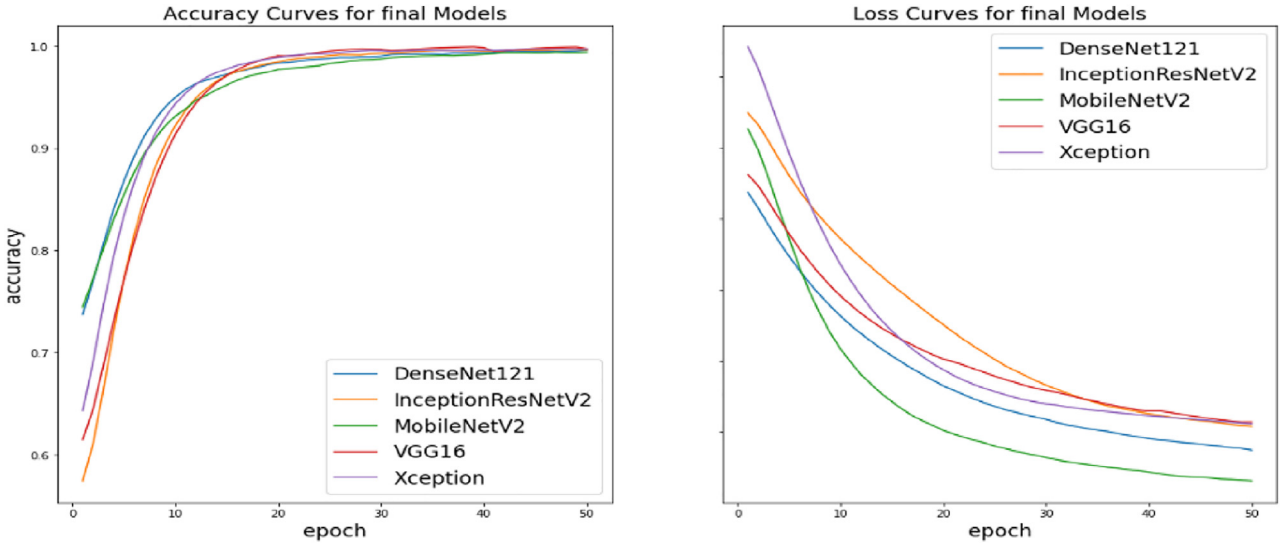


Fig. 9 – Accuracy (left) and loss (right) of training final model on all train set.

Table 3 – Comparison of different metrics for performance measures of all models.

	Accuracy	Sensitivity	Specificity	F1-score	AUC
Model Name					
VGG16	94.74	94.55	94.59	94.57	0.983
IncResNetV2	94.42	94.37	94.16	94.26	0.978
Xception	93.17	92.88	93.02	92.95	0.976
MobileNetV2	94.55	94.68	94.2	94.41	0.983
DensNet121	95.74	95.56	95.64	95.6	0.998
Ensemble Model	96.55	96.01	96.95	96.41	0.96

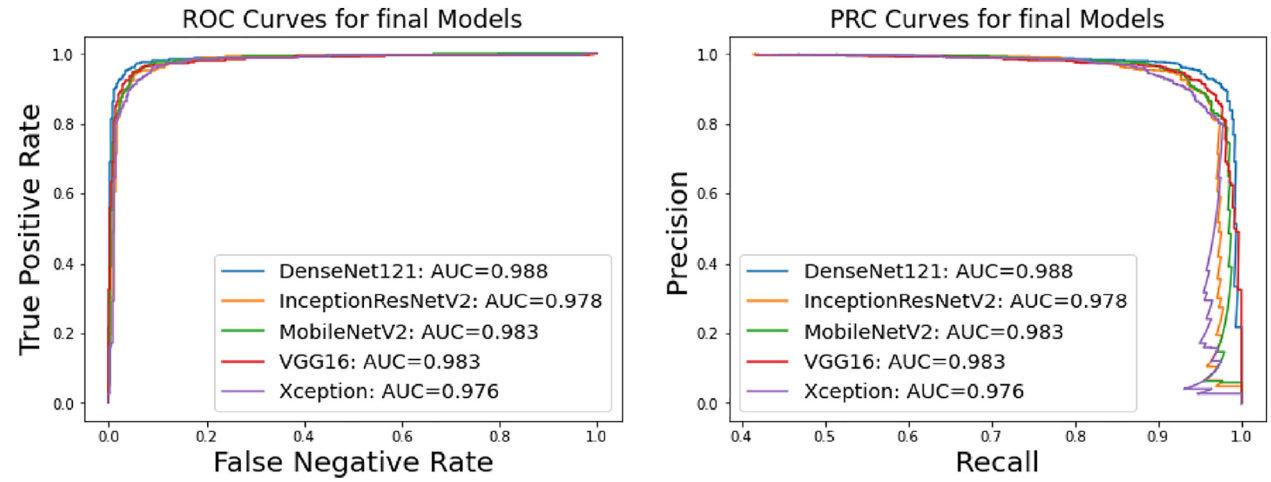


Fig. 10 – ROC and PRC of final models.

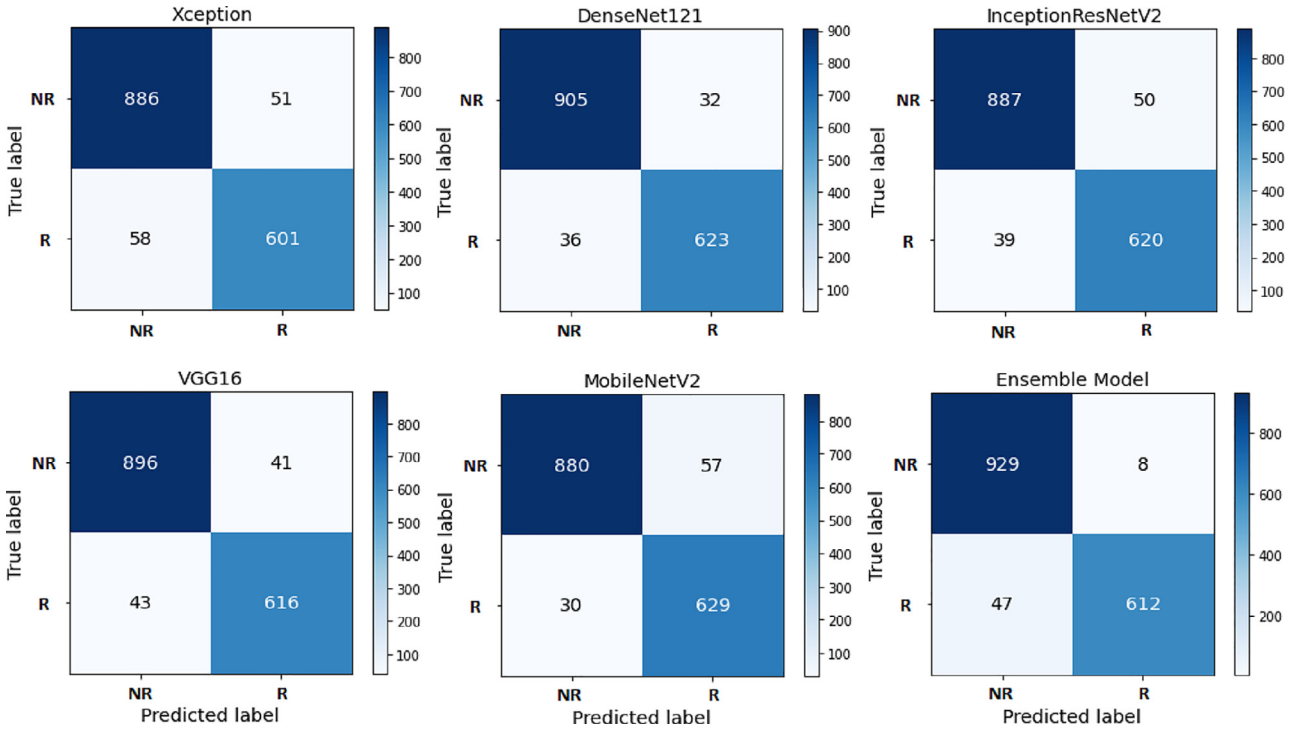


Fig. 11 – Confusion matrixes of all final TL models and ensemble model.

among the other models based on the accuracy (95.74%), sensitivity (95.56%), specificity (95.64%), F1-score (95.6%) and AUC (0.988) metrics. Finally, to improve the classification accuracy, an ensemble method based on majority voting of outputs of 5 deep TL architectures has been developed with higher accuracy (96.55%), sensitivity (96.01%), and specificity (96.95%) than the individual TL models. Therefore, we created a combination of multiple TL models that can accurately classify responders and non-responders to specific treatment using antidepressants.

Our proposed well-trained deep learning model can extract some specialized features from input EEG that simultaneously are stable and predictive. Bag of convolution filters in hierarchical convolutional layers of a deep learning model yields some complex features that are optimized for robustness through multiple subjects. Several studies mentioned that absolute/relative power of EEG frequency bands (such as FAA [17]) and complexity measures of EEG that represent non-linear dynamics of brain (such as Sample Entropy [43]) are predictive biomarkers for treatment outcome prediction. These predictive biomarkers depend on location, power and frequency distribution of EEG signal. However, in our CNN models, global characteristics of EEG signal are considered, comprehensively. WT images that are decomposition of EEG signal through time and frequency, are separately provided for each of 19 channels of EEG signal. Then, our deep learning

model extracts optimal features to simultaneously consider power of frequency bands (because of frequency decomposition in WT), asymmetry of EEG powers (because of channel dependence of each WT image) and complexity of EEG signal (because of non-linear nature of CNN model). So, we can expect this superior performance of our deep learning model. This accurate prediction capability from pre-treatment EEG results in timely decision for tailoring or switching antidepressants. Because changes in clinical symptoms after initiation of antidepressants often take some time to be apparent that prolongs decision for switching or tailoring antidepressants [52,53]. Consequently, our precise deep learning model can help in designing an efficient treatment plan.

CNN models are known to have high variance in performance because of high number of parameters and sensitivity to initial conditions [54]. Besides, EEG signals have shown non-linear behavior especially in the case of multi-channel recordings [37] and depression can induce functional and even anatomical changes in brain [55]. So each model's performance can highly alter in each part of dataset and for each set of hyper-parameters used for TL model. This problem can be minimized using ensemble learning that combines different models to construct a good model with best achievable performance. This can be done in different ways to save the model performance and the generalization for new unseen data [54]. For example Pei et al. [56] used the ensemble of multiple basic SVM models using features extracted from different MRI regions in combination with genetic biomarkers for prediction of response to MDD treatment that led to improvement of the accuracy. Here we have used different TL models trained on the same data as the basic models. Combining different models can increase final model's performance by utilizing the capabilities of each basic model for classification of some part of data. That means that each basic model may have a skill in some part of data and combination of these models using majority voting would yield a better performance.

Demographics and clinical variables can be used for MDD treatment prediction with some success [59]. In the present study, patients with age of 18–65 years are included and the average age of responders and non-responders are 40.7 and 41.1, respectively. Non-responders have slightly higher age compared to responders that is not significant. Indeed, there is same number of male and female subjects in responder and non-responder groups. So, there is no significant difference in demographic variables. However, average pre-treatment BDI-II scores of responders and non-responders are 18.4 (±7. 4) and 22.8 (±12.5), respectively. As table 1 show, pre-treatment BDI-II score of responders is a little lower than BDI-II score of non-responders. This means that non-responders are experiencing higher symptom severity than responders.

A recent meta-analysis reported that only a few studies on antidepressant outcomes were done with cross-validation or another method for out-of-sample verification [16,57]. In this work each model's performance and stability is validated through k-fold cross-validation that k selected to be 10. As demonstrated in learning curves of Figs. 7 and 8, acceptable variation of model performance achieved for all models. The value of k selected to be 10 that is acceptable and high enough to measure model's stability and generalizability and to preserve computational efficacy [58]. Hyper-parameter tuning using cross-validation, resulted in models that generalized very well on WT images of test set. Hyper-parameters are listed in table 2 and other hyper-parameters like regularizes, activation function type, dropout rate, number and size of dense layers and optimizer type were wisely selected and tuned for each model to obtain best performance on validation set by using grid searching method. This tuning process has been done using accuracy and loss curves for each model. Training and validation curves in Figs. 7 and 8 show that TL models used in this study, have the capability to learn patterns for classification of responders and non-responders. Here, we used imagenet weights for initialization of model weights instead of random initialization. It is worthy to mention that using random initialization, the model performance in this study is decreased. This shows that imagenet weights have huge capabilities and this is because of large dataset that used for reaching to these weights through CNN models. Very large dataset can provide very good representation of low-level features from each type of image [59], including edges, curvatures and other basic elements of an image that is common between images in imagenet dataset and WT images of present study.

Ensemble of basic models is a common way for increasing model generalization and prediction accuracy. We highlighted results of the proposed ensemble model in table 4. All performance metrics of the ensemble model are improved against all single TL models. AUC of ensemble model decreased against basic models but is not out of acceptable range for a well-performing model. Fig. 11 shows confusion matrixes of all final models and ensemble model. Confusion matrix of ensemble model shows that combining all models together, highly decreased number of FP and increased number of TN predictions. Highest TN and lowest FP among all models belongs to the proposed ensemble model. It is interesting because of increased power of ensemble model to predict non-responders. This finding yields a model that can accurately classify non-responders to a specific treatment that can prevent clinicians from beginning a mistreatment. Finding non-responders to a specific treatment with high confidence, leads physicians to select another possible choice of treatment.

Table 4 compares results of our work with new best other studies in the prediction of antidepressant treatment outcome from EEG signal. As it is observed, the performance achieved in this study is promising for automatic treatment outcome prediction. Compared to other similar studies, this work has the advantage of combination of the WT image converted from EEG signals and ensemble of powerful pre-trained CNNs based on TL methods comparing the other methods that caused higher accuracy. Relatively small number of subjects used in this study can make models vulnerable to bias and lack of generalizability. We tried to minimize this problem using 10-fold cross validation and TL models. In the future, for further increasing model performance and generalization, Long short-term memory (LSTM) neural network can be used, that is a recurrent cell with learning capability, upon CNN models to extract best temporal, spatial and frequency features from EEG signal.

Table 4 – Comparison of classification results of our work and new existing works in the prediction of antidepressant treatment outcome from EEG signal. ACC, SEN and SPE stand for Accuracy, Sensitivity and Specificity respectively.

Study	Dataset	Methods or features	Classifier/Analyzer	Results
Mumtaz et al., 2017 [12]	16 R/18 NR	Combination of Wavelets + STFT + EMD features	Logistic Regression	ACC = 91.6% SEN = 90.0% SPE = 90.0%
Khodayari-Rostamabad et al., 2013 [13]	11 R/11 NR	power spectral density, squared spectral coherence, mutual information, left-to-right hemispheres and anterior/posterior power ratio	MFA	ACC = 87.4% SEN = 94.9% SPE = 80.9%
Zhdanov et al., 2020 [14]	CAN-BIND-155 R/67 NR	power spectral, spatiotemporal complexity features	SVM	ACC = 82.4% SEN = 72.9% SPE = 85.5%
Jaworska et al., 2019 [15]	27 R/24 NR	Demographic data, EEG power features, source-localized current density	RF	ACC = 88% SEN = 77% SPE = 99%
Rajpurkar et al., 2020 [16]	528 R & NR	Bag of features (including O1-Oz delta band power)	GBDT	C-Index = 0.963
Van der Vinne et al., 2019 [17]	453 R & NR	FAA Stability	ANOVA	Response Rate = 66.4%
Van der Vinne et al., 2021 [18]	122 R & NR	EEG-informed prescription (using Frontal Alpha Asymmetry, Alpha Peak Frequency and EEG abnormality)	ANOVA	–
Van der Vinne et al., 2019 [60]	57 R & NR	EEG Abnormality (Diffuse Slowing, Focal Slowing, Paroxysmal and Non-Paroxysmal Activity)	ANOVA	Response Rate = 74%
Our work	12 R/18 NR	WT images	ensemble of CNN TL models	ACC = 96.55% SEN = 96.01% SPE = 96.95%

5. Conclusion

We performed a comprehensive study to investigate the effects of new ensemble of powerful pre-trained CNN models based on TL to classify responders and non-responders to treatment using SSRI antidepressants based on EEG signal and provided insightful findings. This work showed that TL strategy based on pre-trained CNNs developed through cross-validation can overcome the limitation in number of training samples and provides generalizability for deep learning models. Our findings demonstrate that WT images obtained from pre-treatment resting-state EEG that contain power spectrum patterns of different frequency bands can capture non-stationary nature of EEG signal and have significant capabilities in the prediction of treatment outcome. Images created using WT consider all frequency bands of EEG by implementing multiple scales of basis function to envelope whole frequency domain. This means that intuitively WT images contain information of all EEG bands and put them next to each other. Because of using morlet mother function as basis function for WT calculations, there is smooth transitions between different parts of an image and therefore helps deep learning model to learn most reliable patterns in the WT image. Finally, results indicate that the best performance among basic models achieved by DenseNet121 with accuracy, sensitivity and specificity of 95.74%, 95.56% and 95.64%, respectively. An ensemble of these basic TL models can perform better than each individual model by utilizing capabilities of each model in classification of each WT image. It was observed that the majority voting among the results of 5 deep transfer learning architectures namely VGG16, Xception, DenseNet121, MobileNetV2 and InceptionResNetV2 has the highest performance based on accuracy (96.55%), sensitivity (96.01%), and specificity (96.95%) metrics. Superior results in this paper compared to new best other similar studies shows the power of obtained WT images of EEG signal and ensemble of powerful CNN architectures based on TL strategy for classification of responders and non-responders to SSRI antidepressants.

CRediT authorship contribution statement

Mohsen Sadat Shahabi: Methodology, Software, Investigation, Validation, Writing - original draft. **Ahmad Shalbaf:** Conceptualization, Methodology, Investigation, Supervision, Validation, Writing- Reviewing and Editing. **Arash Maghsoudi:** Conceptualization, Methodology, Investigation, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Isometsä E. Suicidal behaviour in mood disorders—who, when, and why? Can. J. Psych. 2014;59(3):120–30.

[2] Mehltretter J, Rollins C, Benrimoh D, Fratila R, Perlman K, Israel S, et al. Analysis of features selected by a deep learning model for differential treatment selection in depression. Front. Artif. Intell. 2020;2. https://doi.org/10.3389/frai.2019.0003110.3389/frai.2019.00031.s001.

[3] Sinyor M, Schaffer A, Levitt A. The sequenced treatment alternatives to relieve depression (STAR* D) trial: a review. Can. J. Psych. 2010;55(3):126–35.

[4] Leuchter AF, Cook IA, Hunter AM, Korb AS. A new paradigm for the prediction of antidepressant treatment response. Dialogues Clin. Neurosci. 2009;11:435.

[5] Perlman K, Benrimoh D, Israel S, Rollins C, Brown E, Tunteng J-F, et al. A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. J. Affect. Disord. 2019;243:503–15.

[6] Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H, Subha DP. Automated EEG-based screening of depression using deep convolutional neural network. Comput. Methods Programs Biomed. 2018;161:103–13.

[7] Lebiecka K, Zuchowicz U, Wozniak-Kwasniewska A, Szekely D, Olejarczyk E, David O. Complexity analysis of EEG data in persons with depression subjected to transcranial magnetic stimulation. Front. Physiol. 2018;9:1385.

[8] Ay B, Yildirim O, Talo M, Baloglu UB, Aydin G, Puthankattil SD, et al. Automated depression detection using deep representation and sequence learning with EEG signals. J. Med. Syst. 2019;43(7). https://doi.org/10.1007/s10916-019-1345-y.

[9] Radenkovic MC. Machine learning approaches in Detecting the Depression from Resting-state Electroencephalogram (EEG). A Review Study. arXiv preprint arXiv:190311454 2019.

[10] Liao S-C, Wu C-T, Huang H-C, Cheng W-T, Liu Y-H. Major depression detection from EEG signals using kernel eigen-filter-bank common spatial patterns. Sensors. 2017;17(6):1385. https://doi.org/10.3390/s17061385.

[11] Afshani F, Shalbaf A, Shalbaf R, Sleigh J. Frontal–temporal functional connectivity of EEG signal by standardized permutation mutual information during anesthesia. Cogn. Neurodyn. 2019;13(6):531–40.

[12] Mumtaz W, Xia L, Yasin MAM, Ali SSA, Malik AS. A wavelet-based technique to predict treatment outcome for major depressive disorder. PLoS One. 2017;12.

[13] Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, MacCrimmon DJ. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. Clin. Neurophysiol. 2013;124(10):1975–85.

[14] Zhdanov A, Atluri S, Wong W, Vaghei Y, Daskalakis ZJ, Blumberger DM, et al. Use of machine learning for predicting escitalopram treatment outcome from electroencephalography recordings in adult patients with depression. JAMA network open. 2020;3(1):e1918377. https://doi.org/10.1001/jamanetworkopen.2019.18377.

[15] Jaworska N, de la Salle S, Ibrahim M-H, Blier P, Knott V. Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. Front. Psychiatry 2019;9:768.

[16] Rajpurkar P, Yang J, Dass N, Vale V, Keller AS, Irvin J, et al. Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: a prespecified secondary analysis of a randomized clinical trial. JAMA network open. 2020;3(6):e206653. https://doi.org/10.1001/jamanetworkopen.2020.6653.

[17] van der Vinne N, Vollebregt MA, van Putten MJAM, Arns M. Stability of frontal alpha asymmetry in depressed patients during antidepressant treatment. NeuroImage: Clin. 2019;24:102056. https://doi.org/10.1016/j.nicl.2019.102056.

[18] van der Vinne N, Vollebregt MA, Rush AJ, Eebes M, van Putten MJAM, Arns M. EEG biomarker informed prescription of antidepressants in MDD: a feasibility trial. Eur. Neuropsychopharmacol. 2021;44:14–22.

[19] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, et al. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. J. Neural Eng. 2018;15(3):031005. https://doi.org/10.1088/1741-2552/aab2f2.

[20] Major TC, Conrad JM. A survey of brain computer interfaces and their applications. IEEE SoutheastCon 2014: Institute of Electrical and Electronics Engineers Inc, 2014.

[21] Sengupta S, Basak S, Saikia P, Paul S, Tsalavoutis V, Atiah F, et al. A review of deep learning with special emphasis on architectures, applications and recent trends. Knowl.-Based Syst. 2020;194:105596. https://doi.org/10.1016/j.knosys.2020.105596.

[22] Alshemali B, Kalita J. Improving the reliability of deep neural networks in NLP: A review. Knowl.-Based Syst. 2020;191:105210. https://doi.org/10.1016/j.knosys.2019.105210.

[23] Zhu H. Big data and artificial intelligence modeling for drug discovery. Ann. Rev. Pharmacol. Toxicol. Appl. Pharmacol. 2020;60(1):573–89.

[24] Shalbaf A, Shalbaf R, Saffar M, Sleigh J. Monitoring the level of hypnosis using a hierarchical SVM system. J. Clin. Monitor. Comput. 2020;34(2):331–8.

[25] Saeedi A, Saeedi M, Maghsoudi A, Shalbaf A. Major depressive disorder diagnosis based on effective connectivity in EEG signals: a convolutional neural network and long short-term memory approach. Cogn. Neurodyn. 2021;15(2):239–52.

[26] Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. Deep learning in medical image registration: a review. Physics in Medicine. Biol. Philos. 2020;65(20):20TR01. https://doi.org/10.1088/1361-6560/ab843e.

[27] Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image classification and segmentation. Ann. Transl. Med. 2020;8.

[28] Merlin Praveena D, Angelin Sarah D, Thomas GS. Deep Learning Techniques for EEG Signal Applications–a Review. IETE Journal of Research. 2020;1–8.

[29] Craik A, He Y, Contreras-Vidal JL. Deep learning for electroencephalogram (EEG) classification tasks: a review. J. Neural Eng. 2019;16(3):031001. https://doi.org/10.1088/1741-2552/ab0ab5.

[30] Garg D, Verma GK. Emotion Recognition in Valence-Arousal Space from Multi-channel EEG data and Wavelet based Deep Learning Framework. Procedia Comput. Sci. 2020;171:857–67.

[31] Shovon TH, Al Nazi Z, Dash S, Hossain MF. Classification of motor imagery eeg signals with multi-input convolutional neural network by augmenting stft. In: 2019 5th International Conference on Advances in Electrical Engineering (ICAEE): IEEE. p. 398–403.

[32] Mumtaz W, Qayyum A. A deep learning framework for automatic diagnosis of unipolar depression. Int J Med Inform. 2019;132:103983. https://doi.org/10.1016/j.ijmedinf.2019.103983.
[33] Dang W, Gao Z, Sun X, Li R, Cai Q, Grebogi C. Multilayer brain network combined with deep convolutional neural network for detecting major depressive disorder. Nonlinear Dyn. 2020;102(2):667–77.
[34] Wan Z, Huang J, Zhang H, Zhou H, Yang J, Zhong N. HybridEEGNet: A convolutional neural network for EEG feature learning and depression discrimination. IEEE Access 2020;8:30332–42.
[35] Haibo He, Garcia EA. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 2009;21(9):1263–84.
[36] Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: International conference on artificial neural networks. Springer; 2018. p. 270–9.
[37] Zhang B, Wang W, Xiao Y, Xiao S, Chen S, Chen S, et al. Cross-subject seizure detection in EEGs using deep transfer learning. Comput. Math. Methods Med. 2020;2020:1–8.
[38] Cimtay Y, Ekmekcioglu E. Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition. Sensors. 2020;20(7):2034. https://doi.org/10.3390/s20072034.
[39] Shalbaf A, Bagherzadeh S, Maghsoudi A. Transfer learning with deep convolutional neural network for automated detection of schizophrenia from EEG signals. Phys. Eng. Sci. Med. 2020:1–11.
[40] Raghu S, Sriraam N, Temel Y, Rao SV, Kubben PL. EEG based multi-class seizure type classification using convolutional neural network and transfer learning. Neural Netw. 2020;124:202–12.
[41] Macher J-P, Crocq M-A. Treatment goals: response and nonresponse. Dialogues Clin Neurosci. 2004;6:83.
[42] Nierenberg AA, DeCecco LM. Definitions of antidepressant treatment response, remission, nonresponse, partial response, and other relevant outcomes: a focus on treatment-resistant depression. J. Clin. Psych. 2001;62:5–9.
[43] Mahato S, Goyal N, Ram D, Paul S. Detection of depression and scaling of severity using six channel EEG data. J. Med. Syst. 2020;44:1–12.
[44] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Very Deep Convolutional Networks for Large-scale Image Recognition. 2014.
[45] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition2015. p. 1-9.
[46] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition2016. p. 770-8.
[47] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. arXiv preprint arXiv:190106032. 2019.
[48] Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition2017. p. 1251-8.
[49] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition2017. p. 4700-8.
[50] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:04861. 2017.
[51] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition2018. p. 4510-20.
[52] Boyce P, Hopwood M, Morris G, Hamilton A, Bassett D, Baune BT, et al. Switching antidepressants in the treatment of major depression: When, how and what to switch to? J. Affect. Disord. 2020;261:160–3.
[53] Rafeyan R, Papakostas GI, Jackson WC, Trivedi MH. Inadequate response to treatment in major depressive disorder: augmentation and adjunctive strategies. J. Clin. Psych. 2020;81(3). https://doi.org/10.4088/JCP.OT19037BR3.
[54] Sun S, Zhang C, Zhang D. An experimental evaluation of ensemble methods for EEG signal classification. Pattern Recogn. Lett. 2007;28(15):2157–63.
[55] Srivastava K, Prakash J, Shashikumar R, Ryali VSSR, Bhat PS, Khan S. Neuropsychophysiological correlates of depression. Industrial psychiatry journal. 2010;19(2):82. https://doi.org/10.4103/0972-6748.90336.
[56] Pei C, Sun Y, Zhu J, Wang X, Zhang Y, Zhang S, et al. Ensemble Learning for early-response prediction of antidepressant treatment in major depressive disorder. J. Magn. Reson. Imaging 2020;52(1):161–71.
[57] Widge AS, Bilge MT, Montana R, Chang W, Rodriguez CI, Deckersbach T, et al. Electroencephalographic biomarkers for treatment response prediction in major depressive illness: a meta-analysis. Am. J. Psychiatry 2019;176(1):44–56.
[58] Marcot BG, Hanea AM. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Comput. Statistics 2020;1–23.
[59] Kornblith S, Shlens J, Le QV. Do better imagenet models transfer better? Proceedings of the IEEE conference on computer vision and pattern recognition2019. p. 2661-71.
[60] van der Vinne N, Vollebregt MA, Boutros NN, Fallahpour K, van Putten MJAM, Arns M. Normalization of EEG in depression after antidepressant treatment with sertraline? A preliminary report. J. Affect Disord. 2019;259:67–72.