How and why does deep ensemble coupled with transfer learning increase performance in bipolar disorder and schizophrenia classification?

Sara Petiton, Antoine Grigis, Benoit Dufumier, Edouard Duchesnay

HOW AND WHY DOES DEEP ENSEMBLE COUPLED WITH TRANSFER LEARNING INCREASE PERFORMANCE IN BIPOLAR DISORDER AND SCHIZOPHRENIA CLASSIFICATION?

Sara Petiton[1], *Antoine Grigis*[1], *Benoit Dufumier*[1,2], *Edouard Duchesnay*[1]

[1] NeuroSpin, CEA Saclay, Université Paris-Saclay, France
[2] EPFL, Lausanne, Switzerland

ABSTRACT

Transfer learning (TL) and deep ensemble learning (DE) have recently been shown to outperform simple machine learning in classifying psychiatric disorders. However, there is still a lack of understanding as to why that is. This paper aims to understand how and why DE and TL reduce the variability of single-subject classification models in bipolar disorder (BD) and schizophrenia (SCZ). To this end, we investigated the training stability of TL and DE models. For the two classification tasks under consideration, we compared the results of multiple trainings with the same backbone but with different initializations. In this way, we take into account the epistemic uncertainty associated with the uncertainty in the estimation of the model parameters. It has been shown that the performance of classifiers can be significantly improved by using TL with DE. Based on these results, we investigate i) how many models are needed to benefit from the performance improvement of DE when classifying BD and SCZ from healthy controls, and ii) how TL induces better generalization, with and without DE. In the first case, we show that DE reaches a plateau when 10 models are included in the ensemble. In the second case, we find that using a pre-trained model constrains TL models with the same pre-training to stay in the same basin of the loss function. This is not the case for DL models with randomly initialized weights. [1]

***Index Terms*—** deep learning, brain anatomical MRI, transfer learning, deep ensemble learning, bipolar disorder, schizophrenia

1. INTRODUCTION

Deep Learning (DL) has been shown to be an efficient way to classify medical images from MRI scanners [1]. However, machine learning (ML) algorithms tend to perform as well, if not better, than DL when applied to some psychiatric disorders [2]. It has recently been shown that deep ensemble (DE) and transfer learning (TL) paradigms outperform ML for single-subject prediction using 3D whole-brain anatomical MRIs [2]. These newly proposed paradigms offer better performance on psychiatric classification tasks. Nevertheless, it isn't clear how TL enables this gain or to what extent DE improves predictions.

In the case of psychiatric disorder classification, as for other medical prediction tasks, the reliability and robustness of predictions are very important. However, DL models whose weights have been randomly initialized (RI-DL) have multiple sources of variability [3]: aleatoric uncertainty, inherent to the data distribution, and epistemic uncertainty, also known as knowledge or model uncertainty [4]. In this study, we focus on the epistemic uncertainty associated with the random initialization of the model weights. Each time RI-DL models are trained, they may find a different set of weights, which in turn produce different predictions and different latent representations. A successful approach to reduce the variance of these models is to train multiple models and combine their predictions. This is known as ensemble learning [5]. It reduces prediction variance and can lead to better prediction performance. Unfortunately, training multiple models for DE can be time-consuming and computationally expensive. Therefore, finding a threshold for the number of models that need to be trained to see a significant performance improvement can save both time and resources.

On the other hand, TL can improve predictions by using a pre-trained model before fine-tuning previously acquired knowledge to the desired classification task [6]. Here, contrastive learning was used for pre-training using age as a weak supervision [7]. The resulting predictions have been shown to outperform both ML and RI-DL. To illustrate and understand how TL works, a recent study proposed to compare models trained from different weights initializations using performance barrier curves [8]. The authors compared TL and RI-DL models on natural images, drawings, and chest X-ray classification tasks. These comparisons were made by studying the effect of linear interpolation between the weights of any pair of models on a surrogate prediction task.

This paper aims to understand how and to what extent DE and TL outperform RI-DL for bipolar disorder (BD) and

[1] The scripts related to this study can be found at : `https://github.com/SaraMPetiton/DE_with_TL_study`

schizophrenia (SCZ) classification tasks. The proposed contributions are two-fold :

- First, we show that 10 trained models are sufficient for the best performance improvement with DE for SCZ and BD.
- Secondly, we compare the loss landscapes of RI-DL and TL models using barrier curves. To the best of our knowledge, this is the first time that this method, proposed in [8], has been applied to whole-brain MRI datasets. We show that using TL with 3D MRIs for psychiatric classification tasks enables models to stay in the same basin of the loss landscape and presents a more robust approach than RI-DL.

2. MATERIALS AND METHODS

2.1. Datasets

For SCZ classification, the datasets used are SCHIZCONNECT-VIP, CNP, PRAGUE, BSNIP, and CANDI, with 933 subjects used for training, 116 for validation, and 133 for testing. For BD classification, the datasets are BIOBD, BSNIP, CNP, and CANDI, with 832 subjects for training, 103 for validation, and 131 for testing. All splits are stratified on age, sex, site, and diagnosis, and the test sets include sites never seen during training to prevent overfitting on acquisition sites [9]. CAT12 is used to compute voxel-based morphometry (VBM) gray matter maps [10]. These maps are used as input to the proposed TL and RI-DL models.

2.2. Learning strategy

In this study, we use a DenseNet 121 backbone[2] from [2]. This backbone, while limiting the number of parameters to be estimated, has been shown to give the best results on the psychiatric disorder classification tasks considered [2]. The pre-trained model we used for TL was trained on healthy brains from the OpenBHB, HCP, OASIS 3, and ICBM datasets using a weakly-supervised contrastive learning method[3] [7]. This pre-trained model uses the age-aware InfoNCE loss based on the hypothesis that capturing the biological variability in the healthy population related to non-specific variables (in this case, age) with large datasets allows easier discovery of specific pathological variability. The pre-trained weights are then used as a starting point for the training of TL models with the same architecture as the RI-DL models.

During training, the models' learning rates decrease by a factor of γ every 10 epochs. This learning rate decay strategy aims to gradually take smaller steps during gradient descent as we get closer to a minimum of the loss function. We found that the optimal value of γ should not be the same depending on the classification task. To tune this hyperparameter, we trained the considered TL and RI-DL models for 200 epochs with γ equal to 0.2, 0.4, 0.6, and 0.8. The ROC-AUC metric, which is well suited for the binary classification tasks considered (Healthy Controls vs. BD and Healthy Controls vs. SCZ), is used to evaluate model performance.

[2] https://github.com/Duplums/SMLvsDL

[3] https://github.com/Duplums/yAwareContrastiveLearning

2.3. Deep ensemble learning

For each sample, we grouped T models (either TL or RI-DL) and computed the average of their predicted labels, viewing it as a distribution estimation of $p(y|x, \mathcal{D})$, with x the input image, $\mathcal{D}$ the training set, and y the clinical status:

$$\hat{p}^T(y|x, \mathcal{D}) = \frac{1}{T}\sum_{t=1}^{T} p(y|x, \theta^{(t)}) \approx \frac{1}{T}\sum_{t=1}^{T} \hat{y}_{\theta_t} = \hat{y}^{T-DE} \tag{1}$$

where T is the number of trained models, θ the model's weights, and $\hat{y}^{T-DE}$ corresponds to the predicted labels from DE averaging. This averaging minimizes the epistemic uncertainty of the models [11]. It has already been shown in the literature that the use of DE with TL leads to better results compared with RI-DL or TL alone [2]. Here, we investigate how many models are needed to benefit from the performance improvement brought by DE and how the number of models influences performance variability.

From $N = 90$ trained TL models or $N = 90$ trained RI-DL models $\{f_{\theta_1}, ..., f_{\theta_N}\}$, we get individual predictions $\hat{y}_{\theta_i}$, where $i \in [\![1, N]\!]$. Then, we draw with replacement (bootstrap) P subsets of T models with $T \in \{2, 5, 10, 15, 20, 30, 40, 50, 60\}$, from which we compute an ensemble score $\hat{y}_p^{T\text{-DE}}$ using Eq. 1, where $p \in [\![1, P]\!]$. There is no significant computational overhead in using a large value of P, since we are only bootstrapping the predictions. After testing several values of P, we chose $P = 10^5$.

2.4. Linear interpolation of TL and RI-DL models

To understand why TL outperforms RI-DL for single subject classification of SCZ and BD, we applied the linear interpolation method proposed in [8], which linearly interpolates pairs of TL and RI-DL model weights to look for barriers in the loss landscape. The choice of using linear interpolation as a way to study the flatness of the loss landscape near a solution was discussed in [8]. In [12], [13], and [14], authors demonstrate that two minima in any DL model loss landscape can always be connected by a non-linear path maintaining a low loss. By contrast, finding whether the linear interpolation path between two DL models maintains a low loss or not enables us to decipher whether our trained models lie in the same local minimum of the loss function.

We performed linear interpolations between TL and RI-DL models. The weights of the interpolated models along the linear interpolation path are calculated as follows:

$$\theta_\lambda = (1 - \lambda)\theta_1 + \lambda\theta_2 \tag{2}$$

Task	Strategy	Baseline	Deep Ensemble		
			T=3	T=10	T=40
BD classification ↑	TL	$74.68_{\pm 1.96}$	$76.24_{\pm 1.26}$	**77.06**$_{\pm 0.74}$	$77.53_{\pm 0.41}$
	RI-DL	$71.19_{\pm 2.8}$	$73.36_{\pm 1.76}$	$74.55_{\pm 1.04}$	$75.07_{\pm 0.55}$
SCZ classification ↑	TL	$72.76_{\pm 1.65}$	$73.56_{\pm 1.05}$	**73.94**$_{\pm 0.63}$	$74.12_{\pm 0.35}$
	RI-DL	$72.51_{\pm 2.1}$	$73.76_{\pm 1.35}$	$74.16_{\pm 0.79}$	$74.3_{\pm 0.42}$

Table 1. ROC-AUC (in %) with standard deviations for both BD and SCZ classification tasks. Randomly initialized DL (RI-DL) models are compared with transfer learning (TL) models. In both cases, we evaluate the benefit of using deep ensemble (DE) learning. In our setting, "Baseline" corresponds to "no-DE", and T is defined in Eq. 1.

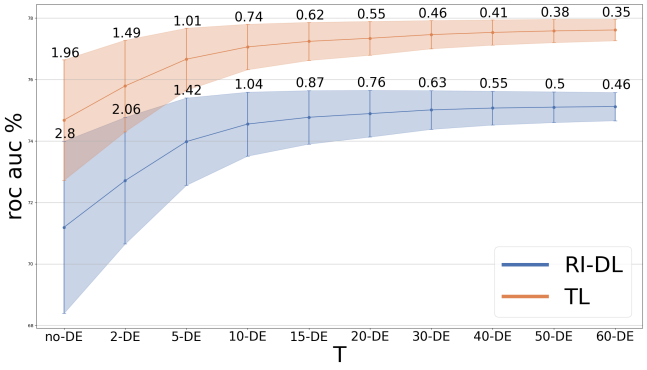

Fig. 1. Learning curves obtained by monitoring the ROC-AUC performance of BD classification as a function of the number of models T considered in the deep ensemble (DE) strategy. The obtained standard deviations are shown directly in the figure for each T-DE value examined on the x-axis. The "x=no-DE" configurations correspond to the means and standard deviations of the 90 trained models without DE.

where $\lambda \in [0, 1]$ is the linear interpolation coefficient, θ_1 the weights of the first model, and θ_2 the weights of the second. In practice, we used 30 values of λ, uniformly distributed between 0 and 1.

Given a pair of trained models, we examine the behavior of the models obtained along such a linear interpolation path. If the chosen performance metrics remain good along this path, then no performance barrier is met, meaning that the two input models rest in the same basin of the loss landscape. Conversely, if the performance metric drops or is highly irregular along the path, it means that a performance barrier was encountered and that the two input models do not lie in the same basin of the loss landscape. For our classification tasks, this barrier will materialize as a decrease in the chosen performance metric, i.e., the ROC-AUC.

The experimental setup compares two RI-DL and two TL models initialized with the same pre-trained weights. Interestingly, in [8], the authors also looked at the linear interpolation between models at their last training epoch and at the epoch at which they perform best. We replicated this experiment to see if the TL models converge faster than the RI-DL models. We linearly interpolated TL and RI-DL models at their last training epoch (we trained them for 200 epochs) and at their best-performing epoch. We will refer to the former models as RI-DL and TL, and to the latter as RI-DL* and TL*. Finally, we propose to study the following scenarios for the two classification tasks considered: TL to TL, RI-DL to RI-DL, TL to TL*, and RI-DL to RI-DL*.

3. RESULTS

3.1. DE performance improvement reaches a plateau

To improve SCZ and BD classification using DE, we searched for an optimal number of models to train. From two sets of 90 RI-DL models and 90 TL models, we investigated the performance of DE learning as a function of the number of models T considered (see Eq. 1). The results are shown in the Table 1 and Figure 1. Overall, we found that DE performance with TL reaches a plateau when using $T = 10$ models (10-DE) for both BD and SCZ classification tasks and that the most robust and accurate predictions were obtained by using TL with DE.

More specifically, from Table 1 and for the BD classification task, the mean gain in ROC-AUC from TL with no-DE to TL with 40-DE is 2.85%. Similarly, the gain in ROC-AUC from TL with no-DE to TL with 10-DE is 2.38%. Therefore, the gain from using 40 instead of 10 models for DE is only 0.47%. We witnessed similar trends when using RI-DL models. The ROC-AUC increases by 3.88% with 40-DE compared with no-DE, and by 3.36% with 10-DE. For the SCZ classification task, the gain from TL with no-DE to TL with 40-DE is 1.36%, compared with 1.18% for TL with 10-DE. For the RI-DL models, the ROC-AUC gain from no-DE is 1.79% for 40-DE and 1.65% for 10-DE. We can see that the improvements of the ROC-AUC from no-DE to 10-DE and from no-DE to 40-DE are very similar. Looking at the learning curve in Figure 1, we confirm this observation. We can see that the ROC-AUC starts to reach a plateau after 10-DE.

From Table 1, we can also see that TL with DE outperforms RI-DL with DE only in the case of BD. For SCZ classification, the TL and RI-DL models have very close ROC-AUC performances. TL with 40-DE gives 2.46% higher ROC-AUC values than RI-DL with 40-DE for BD. For SCZ, TL with 40-DE gives 0.18% lower ROC-AUC values than RI-DL with 40-DE.

In all cases, we see that as the number of models used in DE increases, the ROC-AUC increases, and the associated standard deviation decreases (see Figure 1 for the BD classification task). Note that the standard deviation is also always lower with TL models. The most robust predictions are therefore obtained by using TL with DE.

3.2. Transfer learning minimizes variability of trained models

In Figure 2, we have plotted the evolution of the ROC-AUC performance metric along the linear interpolation path between two selected models in the two classification tasks considered (BD and SCZ). In both cases, the ROC-AUC remains high and resembles an almost straight line when linearly interpolating between the weights of two models trained with TL (blue and green curves in Figure 2). This means that the TL and TL* models remain in the same basin of the loss landscape, as they do not encounter a barrier that would cause the ROC-AUC to drop along the x-axis. However, the ROC-AUC along the x-axis when two RI-DL models are interpolated (orange and red curves in Figure 2) decreases when λ (the linear interpolation coefficient) is close to zero, and increases again when λ is close to 1. This means that the considered input RI-DL models encounter a barrier in the loss landscape and thus fail to complete their training in the same loss basin. This shows that the TL models do not tend to fall into different local minima of loss during the fine-tuning process. The TL models are, therefore, more reliable than the RI-DL models, as they predict results with higher consistency for the two classification tasks considered.

The interpolation of TL to TL* models (the green curves in Figure 2) shows that TL models remain in the same basin of the loss landscape from their best-performing epoch to their last training epoch. This is not the case for RI-DL models (the red curves in Figure 2), where a decrease in the performance metric along the x-axis is observed for both BD and SCZ classification tasks. This means that once TL models reach their best ROC-AUC performance metric, they will remain in the same loss basin for the rest of the training. Therefore, their performance capability will remain the same once their best-performing epoch has been reached. Therefore, the TL models we studied not only perform better and with less variability than the RI-DL models, but they also require fewer training epochs.

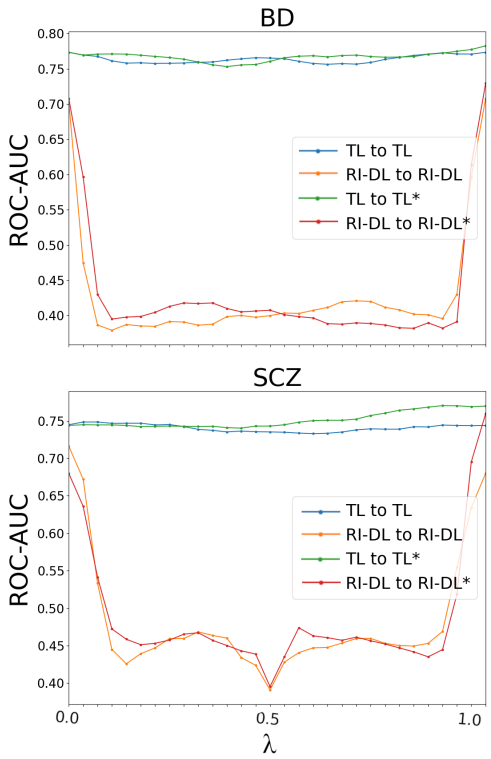

Fig. 2. Linear interpolation between RI-DL and TL models at the last and best training epochs on both BD and SCZ datasets. $\lambda \in [0, 1]$ is the linear interpolation coefficient (see Eq. 2).

4. CONCLUSION

In this paper, we explore how TL and DE can improve the performance of single-subject classification of BD and SCZ. We show how both techniques can reduce model variability. In particular, the variability reduction achieved by DE learning depends on how many trained model predictions are averaged. In our two applications, we have shown that ten models are sufficient for this averaging to be beneficial both in terms of performance and variability reduction, as well as model robustness. Furthermore, we show that TL maintains BD and SCZ classification models in the same basin of the loss landscape. Indeed, TL prevents the trained models from moving to different basins during fine-tuning. As a result, these models produce similar predictions. Compared with RI-DL, TL provides better, more robust predictions, and requires fewer training epochs.

Overall, this work sheds light on the underlying mechanisms of performance improvements when using TL and DE in psychiatric disorder classification. We have shown that (i) 10 trained models are sufficient to achieve excellent and robust predictions when using DE and (ii) that TL models using 3D whole-brain MRI data provide coherent results by staying in the same basin of the loss landscape.

Further work could investigate why TL using age-aware contrastive learning [7] as pre-training benefits some psychiatric disorders more than others in comparison with RI-DL. In [2], it is suggested that TL might not perform as well on the SCZ classification task as BD due to a simplicity bias [15] hindering model generalizability. Indeed, SCZ subjects have been shown to display stronger deviations from healthy controls than BD subjects [16], making the classification of SCZ against healthy controls an easier task. Moreover, some studies [17] [18] have shown that SCZ is associated with accelerated brain aging (much more so than BD), indicating that there might be an overlap between the pre-training model and the classification task.

5. ACKNOWLEDGMENTS

This work was funded by Big2Small (ANR-19-CHIA-0010-01), RHU-PsyCARE (ANR-18-RHUS-0014), and R-LiNK (H2020-SC1-2017, 754907).

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data under the local ethics policy.

7. REFERENCES

[1] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.

[2] B. Dufumier, "Representation learning in neuroimaging: Transferring from big healthy data to small clinical cohorts," Ph.D. dissertation, Université Paris-Saclay, 2022. [Online]. Available: `https://theses.hal.science/tel-03963547`.

[3] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, 2016. [Online]. Available: `https://www.cs.ox.ac.uk/people/yarin.gal/website/thesis/thesis.pdf`.

[4] M. Abdar, F. Pourpanah, S. Hussain, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, May 2021.

[5] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS*, vol. 30, 2017.

[6] J. M. Valverde, V. Imani, A. Abdollahzadeh, *et al.*, "Transfer learning in magnetic resonance brain imaging: A systematic review," *J. Imaging*, vol. 7, no. 4, p. 66, Apr. 2021.

[7] B. Dufumier, P. Gori, J. Victor, *et al.*, "Contrastive learning with continuous proxy meta-data for 3d mri classification," in *MICCAI*, 2021, pp. 58–68.

[8] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?" *arXiv*, 2021.

[9] C. Wachinger, A. Rieckmann, S. Pölsterl, and Alzheimer's Disease Neuroimaging Initiative and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing, "Detect and correct bias in multi-site neuroimaging datasets," *Med. Image Anal.*, vol. 67, p. 101 879, Jan. 2021.

[10] C. Gaser, R. Dahnke, P. M. Thompson, F. Kurth, E. Luders, and A. D. N. Initiative, "Cat – a computational anatomy toolbox for the analysis of structural mri data," *bioRxiv*, 2022.

[11] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS*, vol. 30, Nov. 2017.

[12] S. Fort and S. Jastrzebski, "Large scale structure of neural network loss landscapes," *arXiv*, 2019.

[13] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson, "Loss surfaces, mode connectivity, and fast ensembling of dnns," in *NeurIPS*, Oct. 2018.

[14] F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht, "Essentially no barriers in neural network energy landscape," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, Jul. 2018, pp. 1309–1318.

[15] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli, "The pitfalls of simplicity bias in neural networks," Jun. 2020.

[16] T. Wolfers, N. T. Doan, T. Kaufmann, *et al.*, "Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models," *JAMA Psychiatry*, vol. 75, no. 11, pp. 1146–1155, Nov. 2018.

[17] C. Constantinides, L. K. M. Han, C. Alloza, *et al.*, "Brain ageing in schizophrenia: Evidence from 26 international cohorts via the enigma schizophrenia consortium.," *Mol Psychiatry*, vol. 28, pp. 1201–1209, Mar. 2023.

[18] S. Shahab, B. H. Mulsant, M. L. Levesque, *et al.*, "Brain structure, cognition, and brain age in schizophrenia, bipolar disorder, and healthy controls.," *Neuropsychopharmacol*, vol. 44, pp. 898–906, Apr. 2019.