3D Deformable Convolutions for MRI classification

1st Marina Pominova*
Skolkovo Institute of Science and Technology
Moscow, Russia
marina.pominova@skoltech.ru

2nd Ekaterina Kondrateva*
Skolkovo Institute of Science and Technology
Moscow, Russia
ekaterina.kondrateva@skoltech.ru

3rd Maksim Sharaev
Skolkovo Institute of Science and Technology
Moscow, Russia
m.sharaev@skoltech.ru

4th Sergey Pavlov
Moscow Institute of Physics and Technology
Moscow, Russia
sergey.pavlov@phystech.edu

5th Alexander Bernstein
Skolkovo Institute of Science and Technology
Moscow, Russia
a.bernstein@skoltech.ru

6th Evgeny Burnaev
Skolkovo Institute of Science and Technology
Moscow, Russia
e.burnaev@skoltech.ru

***Abstract*—Deep learning convolution neural networks have proved to be a powerful tool for MRI analysis. In current work, we explore the potential of the deformable convolution deep neural network layers for MRI data classification. We propose new 3D deformable convolutions (d-convolutions), implement them in VoxResNet architecture and apply for structural MRI data classification. We show that 3D d-convolutions outperform standard ones and are effective for unprocessed 3D MR images being robust to particular geometrical properties of the data. Firstly proposed *dVoxResNet* architecture exhibits high potential for the use in MRI data classification.**

***Index Terms*—neuroimaging, MRI, bipolar disorder, schizophrenia, biomarkers, deep learning, convolutional neural networks**

I. Introduction

There is a need for accurate prediction and diagnostics in neurology and psychiatry. MRI is considered as one of the most powerful diagnostic instruments applicable for multiple examinations both in adults and children, see [1]–[4].

There is a number of successful applications of convolutional neural networks (CNN) with different architectures for segmentation of MRI data. Many of these solutions are based on adapting existing approaches to analyzing 2D images for processing of three-dimensional data.

For example, for brain segmentation, an architecture similar to ResNet [5] was proposed, which expands the possibilities of deep residual learning for processing volumetric MRI data using 3D filters in convolutional layers. The model, called VoxResNet [6], consists of volumetric residual blocks (VoxRes blocks), containing convolutional layers as well as several deconvolutional layers. The authors demonstrated the potential of ResNet-like volumetric architectures, achieving better results than many modern methods of MRI image segmentation [7]. Convolutional neural networks also showed good classification results in problems associated with neuropsychiatric diseases such as Alzheimer's disease [2]. Results were more accurate or comparable to earlier feature based approaches that use extracted morphometrical lower dimensional brain characteristics [8]–[11].

Thus, convolutional networks can be applied directly to the raw neuroimaging data without loss of model performance and over-fitting, which allows skipping the feature extraction step.

However, there is a problem: the traditional convolutional networks are very sensitive to image size, scale and spatial orientation thus require thorough pre-processing specific to a clinical application. Because different locations in the input feature maps may correspond to objects with different scales or deformation, adaptive determination of scales or receptive field sizes is desirable for certain tasks, particularly in neuroimaging.

It is known that CNNs are inherently limited to model large unpredictable transformations because of the fixed geometric structure of the sampling matrix, and restricted receptive fields.

In the present study we carried out an extensive experimental evaluation of deep CNNs both with traditional and deformable convolutions for bipolar disorder and schizophrenia diagnostics based on structural MRI data. The article has the following structure: in Section 2 we review current deep learning approaches used for MRI data analysis and their drawbacks as well as a possible solution — deformable convolutions. Here we also present the training datasets and the classification task. We describe obtained results in Section 3, provide discussion of the network performance in Section 4 and draw conclusions in Section 5.

II. Materials and Methods

A. MRI data characteristics

In magnetic resonance imaging (MRI), strong magnetic fields are used to create images of biological tissues and physiological changes in them. MRI data, i.e. 3D brain images, are often noisy and high dimensional.

MR noise is associated with the scanning procedure (low-level hardware artefacts such as magnetic field inhomogeneity, radiofrequency noise, surface coil artefacts and others) and signal processing (chemical shift, partial volume, etc.); as well as with the scanned patient (physiological noise such as blood flow, movements, etc.) [12].

In addition to MRI data cleaning problem, there is another common challenge of the brain imaging analysis related to big data dimensionality, which mostly depends on resolution parameters of the scanner inductive detection coil. For instance, standard voxel sizes are within 0.5–$2\,\mathrm{mm}^3$ in case of structural imaging (resulting in 10^7 voxels for the whole brain volume). Thus an MRI image, composed of a huge number of small-sized voxels, has higher spatial resolution and, hence, high dimensionality.

A canonical approach to structural MRI data analysis is morphometry features extraction: brain structure is first segmented into anatomical regions or tissue types, and then different characteristics like region volumes, tissue thickness and many others are calculated. These features are then used in regression or classification algorithms.

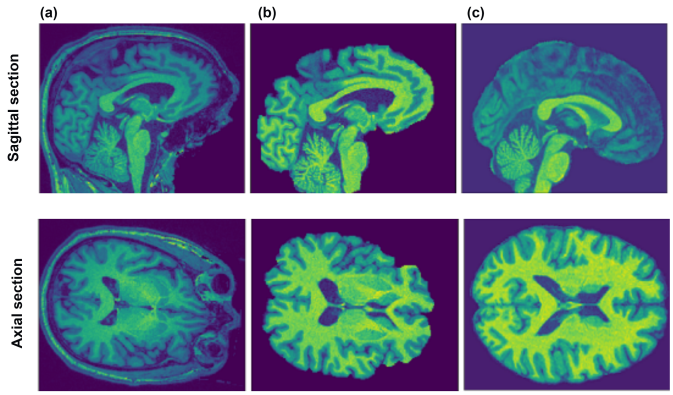

Fig. 1. Sagittal and Axial sections view of structural MRI images on different preprocessing stages: (a) imagery without preprocessing after anonymization, (b) skull-stripped imagery, (c) skull-stripped and MNI normalized images.

B. Deformable Convolutions

Even after the same preprocessing, MRI data from different scanners can vary significantly in size and aspect ratio of brain images.

There are two natural ways to make the model more stable with respect to such changes in spatial characteristics. First, we can augment the training data by scaling images by random values along different axes. Second, we can make the network itself invariant to transformations of the input image, introducing certain modification to its architecture. The latter can be implemented using deformable convolutions [13].

The key difference between regular convolution and its deformable counterpart is the ability to deform the sampling grid for the standard convolution by predicting offsets for sampling locations. The offsets are learned from the activations of preceeding layer using additional convolutional layer, separately for each point of the activation map where the kernel is applied. The Fig. 2 illustrates how the offsets for the deformation are learned from the previous layer activation maps in the case of two- and three-dimensional image.

The whole process for deformable convolution looks as follows. Suppose, R is a sampling grid for the regular convolution. For each location p_n on the input feature map x and location p_0 in the sampling grid, an additional convolutional layer, typically with kernel of the same size and dilation, predicts an offset $\Delta p_n,\ n = 1, \ldots, |R|$. Then the sampling grid for deformable convolution is augmented with predicted offsets, and kernel is applied to the values at locations $p_n + \Delta p_n$, which are computed via bilinear interpolation, since the obtained locations can be fractional. The final value, computed by deformable convolution at a point p_0 is $\sum_{p_n \in R} w_n x(p_0 + p_n + \Delta p_n)$, where $w_n, n = 1, \ldots, |R|$ are weights of the convolution kernel, instead of $\sum_{p_n \in R} w_n x(p_0 + p_n)$ in the case of standard one.

As a result, the receptive field of the convolutional kernel can change accordingly to the deformation of the input feature map and thus adapt to the variations of the size and scales of the learned distinguishing patterns for classification. On the Fig. 3 we can observe how the receptive field of deformable convolution differs from the regular one due to its adaptive sampling grid in case of brain imagery.

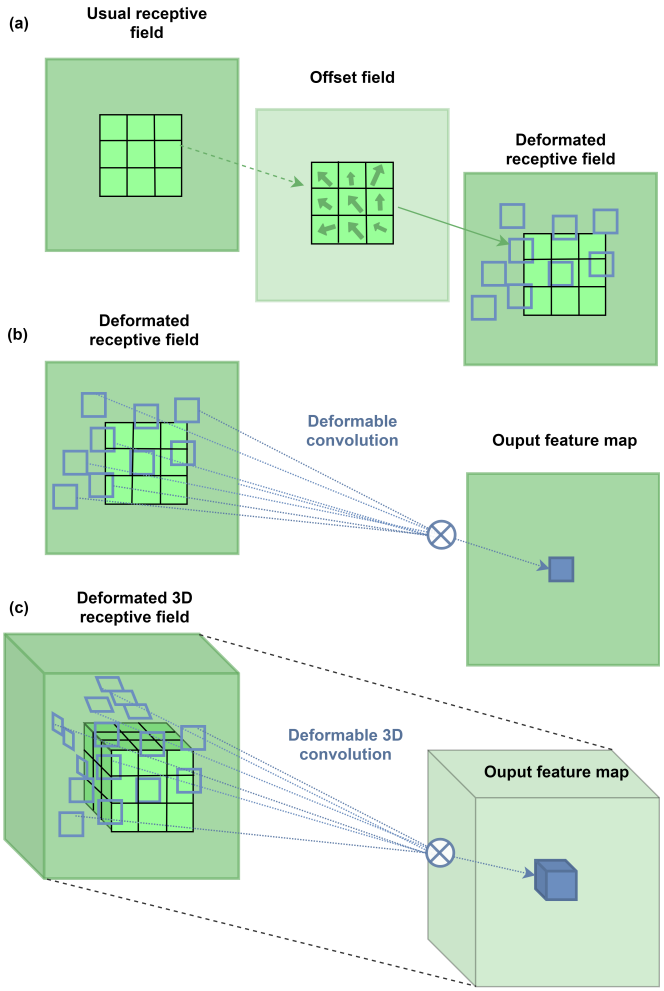

Fig. 2. Illustration of deformable convolution: (a) 2D d-convolution [3, 3], the arrow in the offset field corresponds to how the blue squares are shifted in the input feature map; (b) 2D d-convolution; (c) 3D d-convolution [3, 3, 3].

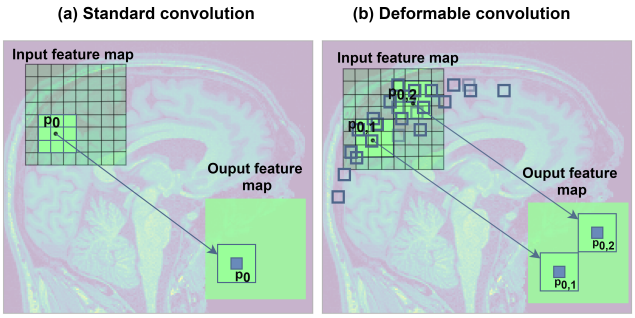

Fig. 3. Standard (a) and d-convolutions (b) applied to unprocessed brain imagery (Sagittal section). Grid represents pixel distribution within a sample in 2D case: (a) kernel with static receptive field; (b) d-convolution offsets with dynamic and learnable receptive field.

C. dVoxResNet architecture

For the structural MRI data classification, we applied a modification of VoxResNet architecture consisting of 6 convolutional layers and 4 VoxRes blocks with two convolutional layers in each. First, we obtained baseline classification performances using standard model with only regular convolutional kernels. Next, we replaced regular three-dimensional convolutions in one or several layers of the network with their deformable counterparts. The architecture of our VoxResNet model is shown on Fig. 4. We tested inserting deformable convolutions in the layers Conv3D from 4 to 6 and in both convolutional layers of VoxRes blocks from 2 to 4. Moreover, to study the effect of stacked deformable convolutions, we tried to apply them in several sequential layers and blocks of the network.

Due to the small sample sizes (172 subjects) we compare the classification results significance with paired non-parametric *ttest* on ROC/AUC scores for repeated 5-fold cross validation (see description of a general pipeline in [14]). In [15] it was shown, that stratified 5-fold CV had remarkably low bias compared to CV without repeats. Thus the variance could be reduced by repeating the n-fold error estimation over more than one random split of the data.

The models were implemented in `PyTorch` and trained on a single GPU [16].

D. Data and preprocessing

dVoxResNet performance was explored on a classification task between patients with Schizophrenia and Bipolar Disorder (as the most widespread psychiatric disorders) and healthy controls based on brain structural images. We aimed at finding a relatively big public dataset with multiple disorders collected at the same site.

Dataset: The data was collected from open databases from *OpenNeuro (OpenfMRI)*[1] platform.

The *Main* dataset is retrieved from *Consortium for Neuropsychiatric Phenomics study*[2] [17] for two pathologies:

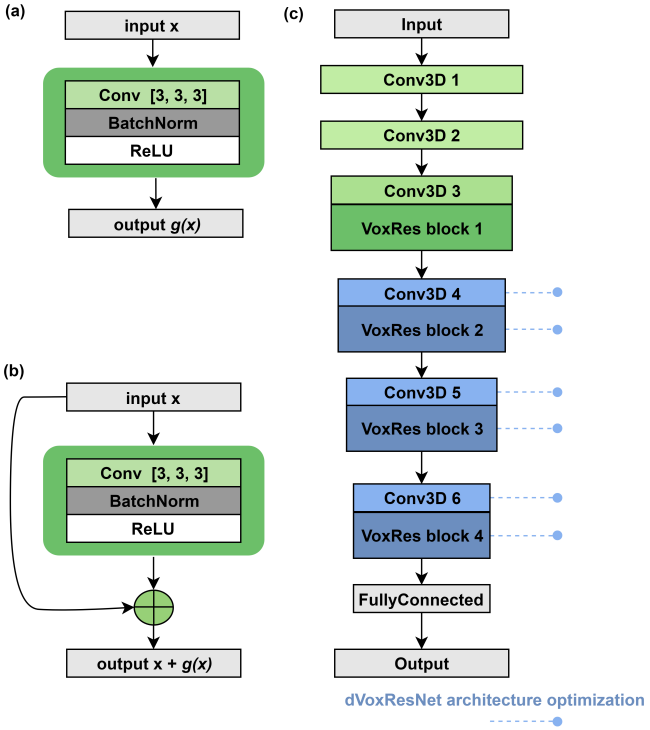

Fig. 4. Architecture of dVoxResNet model used for structural MRI data classification: (a) *Conv3D* unit structure; (b) *VoxRes* unit structure; (c) *dVoxResNet* model architecture, blue nodes represent chancing blocks for ablation study of design choices used for model optimisation.

Schizophrenia and Bipolar Disorder as one of most spread psychiatric disorders which are known to have corresponding bio-markers or pathology patterns in brain structure [18], [19].

The additional dataset Working memory in healthy and schizophrenic individuals [3] [20] was used as a reference dataset for *dVoxResNet* performance check.

We consider three classification problems: two main binary classifications from one source and one additional test dataset.

- *Dataset 1 (Main)*: 50 Schizophrenia patients vs 122 Healthy Controls;
- *Dataset 2 (Main)*: 49 Bipolar Disorder patients vs 122 Healthy Controls;
- *Dataset 3 (Additional)*: 47 Schizophrenia vs 34 Healthy Controls.

For each dataset the goal was to predict whether a subject is from pathology group or healthy control group.

Preprocessing: We performed different preprocessing steps included in regular structural MRI data preprocessing pipeline:

- data anonymization;
- skull-stripping;
- skull-stripped brain normalization.

First, `Pydeface`[4] was applied to all anatomical images to ensure deindentification of subjects. No additional preprocessing was applied on this stage, see Fig.1 (a).

[1]https://openneuro.org/, https://www.openfmri.org/
[2]https://www.openfmri.org/dataset/ds000030/
[3]https://www.openfmri.org/dataset/ds000115/
[4]https://github.com/poldracklab/pydeface

Second, defaced data was scull stripped with FSL/BET [5] brain extraction toolbox [21]. Thus the data from the second stage had no potentially uninformative features, see Fig.1 (b).

On the third step brain images were normalized to MNI space [22], which implies standardization and alignment to common space for brain volumetric analysis during standard preprocessing protocol *FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle* [23] in fmriprep [6] [24] toolbox. The result of the procedure can be seen at Fig.1 (c).

After normalization, MR images were segmented in order to extract structural morphometric features (brain volumes, thicknesses, curvatures, number of vertexes, average voxel intensities within the region, etc.) which together form a feature vector.

Structural features are calculated from $T1$ images using default processing pipeline in Freesurfer [25] toolbox. Thus, morphometrical characterinstics (volumes, surface areas, thicknesses, etc.) are calculated for 34 cortical regions according to Desikan-Killiany Atlas and for 45 subcortical areas according to Automatic subcortical segmentation [26] resulting in a vector of 927 features for each subject.

The ***Baseline performance*** was calculated with classification pipeline described in [27] with hyper-parameter grid search across classifiers: Logistic Regression, Support Vector Machine, Random Forest Classifier, k-Nearest Neighbours Classifier. The best model was SVC, C = 10, kernel = rbf, gamma = 0.01 implemented in Sklearn.

In order to test the dependence of classification performance on the level of images preprocessing, we ran classifiers on data after each preprocessing step, i.e. no preprocessing, skull-stripping, skull-stripping with normalization to MNI space.

III. Results

For both classification problems: Schizoprenia (50) vs Control (122) and Bipolar Disorder (49) vs Control (122) deep neural network classification performance is lower on MNI-normalized images.

For Schizophrenia classification all deep neural methods outperform the Baseline classifier on morphometry features, see TABLE I where baseline classification AUC is **0.739 (0.086)**. All considered convolutional network architectures outperform the baseline on imagery without normalization. Optimized *dVoxResNet* shows statistically significant performance increase according to the cross-validation scores.

For Bipolar Disorder classification deep neural methods show no significant difference from baseline classifier performance of **0.668 (0.074)**, see TABLE II. Yet the optimized architecture of *dVoxResNet* shows statistically significant performance increase yielding **0.687 (0.065)** on skull-stripped brain volumes.

The explicit description on model optimisation and corresponding cross validation scores for dVoxResNet model performance depending architecture studied on schizophrenia and Bipolar Disorder classification are shown in TABLE III, see APPENDIX.

[5]https://fsl.fmrib.ox.ac.uk/fslcourse/lectures/practicals/intro2/index.html

[6]https://github.com/poldracklab/fmriprep

TABLE I
SCHIZOPHRENIA CLASSIFICATION: SCHIZOPHRENIA (50) / CONTROL (122). VOXRESNET MODEL WITH AND WITHOUT D-CONVOLUTIONAL LAYERS ON DIFFERENT PREPOSSESSING STAGES. *Validated on 3-fold CV with 3 repeats, ROC/AUC.*

MRI data type and preprocessing.	3D CNN, Mean (STD)	dVoxResNet, Mean (STD)	dVoxResNet optimised, Mean (STD)
No prepossessing	0.788 (0.068)	0.808 (0.064)	**0.823 (0.058)**
skull-stripping	0.778 (0.075)	0.806 (0.055)	**0.823 (0.065)**
skull-stripping, MNI normalization	0.736 (0.070)	0.737 (0.063)	0.731 (0.063)
Baseline: morphometry features		0.739 (0.086)	

TABLE II
BIPOLAR DISORDER CLASSIFICATION: BIPOLAR DISORDER (50) / CONTROL (122). VOXRESNET MODEL WITH AND WITHOUT D-CONVOLUTIONAL LAYERS ON DIFFERENT PREPOSSESSING STAGES. *Validated on 5-fold CV with 3 repeats, ROC/AUC .*

MRI data type and preprocessing.	3D CNN, Mean (STD)	dVoxResNet, Mean (STD)	dVoxResNet optimised, Mean (STD)
No prepossessing	0.639 (0.088)	0.639 (0.088)	**0.676 (0.081)**
skull-stripping	0.648 (0.102)	0.651 (0.065)	**0.687 (0.065)**
skull-stripping, MNI normalization	0.639 (0.105)	0.629 (0.065)	0.631 (0.097)
Baseline: morphometry features		0.668 (0.074)	

IV. Discussion

It can be seen that insertion of deformable convolution layers instead of traditional ones in VoxResNet architecture yield statistically significant improvement in classification accuracy for both schizophrenia and bipolar disorder classification tasks. Moreover, d-convolutions perform well on MRI data without preprocessing and skull-stripping. Our experiments with *dVoxResNet* architecture also revealed decrease of classification accuracy of schizophrenia and bipolar disorder with increaing the level of preprocessing. This could potentially arise from the fact that data cleaning and preprocessing removes informative parts of data. Given that typical preprocessing pipeline for structural imagery is computationally expensive and takes up to 15 min per subject for skull-stripping and up to 5 hours per subject for normalization in FreeSurfer (on a single CPU core), applying d-convolutions seems beneficial and the need for thourough MR images preprocessing for classification with convolutional networks should be investigated.

Insertion of deformable convolution layers in VoxResNet architecture yield statistically significant improvement in classification accuracy for imagery without preprocessing and

skull-stripping of images. However, they do not improve classification performance on preprocessed data. This can be possibly explained by the fact that MNI normalization significantly reduces variability in small deformations of the brain image and deformable convolutions lack information for training the meaningful offset. But at the same time the number of trainable parameters increases, making the model with deformable convolutions less stable to overfitting.

In case of disease recognition based on unprocessed or skull-stripped data, the use of deformable convolutions in only one layer or block gives only a slight quality improvement, and thus is potentially not enough to provide the network with the necessary deformation modelling capability. However, the use of stacked deformable convolutions in several sequential layers already allows to obtain the effect of a statistically significant improvement of classification performance. These results are also consistent with the observations for deformation modelling with 2D deformable ConvNets [28].

We also applied deformable ConvNets for schizophrenia recognition problem on *Dataset 3 (Additional)* of smaller size, see TABLE IV in APPENDIX. We tested both training the model from scratch on this small sample and transferring and fine-tuning the pre-trained weights for Schizophrenia versus Healthy control classification from the main dataset. Yet it does not show any significant classification improvement, which can be due to small sample size causing in poor model generalizability.

A. Limitations

The current study has several limitations mostly resulting from the computational cost of deformable convolutions and 3D networks on MRI volumes. The d-convolutions are to be used with the data augmentation yet it more computationally expensive.

The use of convolutions with bigger than [3, 3, 3] kernels also was restricted due to *GPU* memory capacity.

Limited data: 122 control, 50 Schizophrenia and 49 Bipolar Disorder subjects; larger samples classification will allow more disperse estimation of model performance, without computational expenses on cross validation.

B. Further directions

Deformable convolution tries to learn how to predict the sampling locations in a way to make filters invariant to small deformations. Thus, augmentation of brain imagery data using small scales and affine transformations can add more variability and provide additional information for training networks with deformable convolutional modules. It could also be potentially helpful to get more stable results by increasing the training sample size. However, applying such augmentation is more computationally expensive.

Also wide learnable receptive field could be beneficial on first layers, kernels with size [7, 7, 7] or [11, 11, 11] can be more effective for global patterns then [3, 3, 3] kernels.

The use of deformable convolutions with bigger than $3\times3\times3$ kernels requires extended *GPU* memory capacity. Prediction of the offsets for each unit of the convolutional kernel of size k at each point of the input feature map assumes generation of the additional $3 \times k^3$ activation maps.

D-convolutions now explored for MRI classification problems were originally proposed for image segmentation tasks, and can be further utilized for brain segmentation as well.

Apart from deformable convolutions, transformable convolutions [29] can improve deep neural model performance. Besides dynamic sampling matrix in transformable convolutions there is a static global sampling matrix and they are used together for getting the output feature map. Thus defining global offset map depending on a brain structure we can introduce a domain specific and improve accuracy. Other lines of research could include usage of sparse convolutions [30] for computational efficiency and fusion of multi-fidelity data [31] to increase prediction accuracy.

The additional study would be a great of interest both for potential accuracy improvement and neural networks results interpretation.

V. CONCLUSION

We proposed new 3D deformable convolutions (d-convolutions) application for structural MRI classification task and implemented it in VoxResNet architecture — *dVoxResNet*.

The usage of deformable convolution layers yields statistically significant improvement in classification performance for unprocessed and skull-stripped brain images. Yet it was not effective for normalized brain images, which could potentially arise from the fact that data normalization removes informative parts of data and reduces variability. We showed that deformable convolutions could be competitive analogues of standard ones, and their usage is reasonable despite of the high computational cost. 3D d-convolutions significantly outperform standard ones in binary classification tasks and are effective for unprocessed 3D images.

Firstly proposed *dVoxResNet* architecture show a high potential for application to other psycho-neurological disorders diagnostics from different datasets.

ACKNOWLEDGMENT

The study was supported by the Russian Science Foundation under Grant 19-41-04109.

REFERENCES

[1] B. C. Bernhardt, S.-J. Hong, A. Bernasconi, and N. Bernasconi, "Magnetic resonance imaging pattern learning in temporal lobe epilepsy: classification and prognostics," *Annals of neurology*, vol. 77, no. 3, pp. 436–446, 2015.

[2] E. Hosseini-Asl, G. Gimel'farb, and A. El-Baz, "Alzheimer's disease diagnostics by a deeply supervised adaptable 3d convolutional network," *arXiv preprint arXiv:1607.00556*, 2016.

[3] M. Pominova, A. Artemov, M. Sharaev, E. Kondrateva, A. Bernstein, and E. Burnaev, "Voxelwise 3D convolutional and recurrent neural networks for epilepsy and depression diagnostics from structural and functional MRI data," in *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 2018-November, pp. 299–307, IEEE, nov 2019.

[4] S. Ivanov, M. Sharaev, A. Artemov, E. Kondratyeva, A. Cichocki, S. Sushchinskaya, E. Burnaev, and A. Bernstein, "Learning connectivity patterns via graph kernels for fmri-based depression diagnostics," in *Proc. of IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 308–314, 2018.

[5] V. Buchstaber and T. Panov, "Toric topology.," *Mathematical Surveys and Monographs. Amer. Mathematical Society*, vol. 205, 2015.

[6] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2018.

[7] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016.

[8] F. Milletari, S.-A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Bötzel, *et al.*, "Hough-cnn: deep learning for segmentation of deep brain regions in mri and ultrasound," *Computer Vision and Image Understanding*, vol. 164, pp. 92–102, 2017.

[9] L. Zou, J. Zheng, C. Miao, M. J. Mckeown, and Z. J. Wang, "3d cnn based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural mri," *IEEE Access*, vol. 5, pp. 23626–23636, 2017.

[10] A. Farooq, S. Anwar, M. Awais, and S. Rehman, "A deep cnn based multi-class classification of alzheimer's disease using mri," in *2017 IEEE International Conference on Imaging systems and techniques (IST)*, pp. 1–6, IEEE, 2017.

[11] L. Chen, Y. Wu, A. M. DSouza, A. Z. Abidin, A. Wismüller, and C. Xu, "Mri tumor segmentation with densely connected 3d cnn," in *Medical Imaging 2018: Image Processing*, vol. 10574, p. 105741F, International Society for Optics and Photonics, 2018.

[12] L. Erasmus, D. Hurter, M. Naude, H. Kritzinger, and S. Acho, "A short overview of mri artefacts," *SA J. Radiol.*, vol. 8, pp. 13–17, 2004.

[13] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.

[14] M. Sharaev, A. Andreev, A. Artemov, E. Burnaev, E. Kondratyeva, S. Sushchinskaya, I. Samotaeva, V. Gaskin, and A. Bernstein, "Pattern recognition pipeline for neuroimaging data," in *Artificial Neural Networks in Pattern Recognition* (L. Pancioni, F. Schwenker, and E. Trentin, eds.), (Cham), pp. 306–319, Springer International Publishing, 2018.

[15] C. Beleites, R. Baumgartner, C. Bowman, R. Somorjai, G. Steiner, R. Salzer, and M. G. Sowa, "Variance reduction in estimating classification error using sparse datasets," *Chemometrics and intelligent laboratory systems*, vol. 79, no. 1-2, pp. 91–100, 2005.

[16] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[17] K. J. Gorgolewski, J. Durnez, and R. A. Poldrack, "Preprocessed consortium for neuropsychiatric phenomics dataset," *F1000Research*, vol. 6, 2017.

[18] L.-L. Zeng, H. Shen, L. Liu, L. Wang, B. Li, P. Fang, Z. Zhou, Y. Li, and D. Hu, "Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis," *Brain*, vol. 135, no. 5, pp. 1498–1507, 2012.

[19] J. Dakka, P. Bashivan, M. Gheiratmand, I. Rish, S. Jha, and R. Greiner, "Learning neural markers of schizophrenia disorder using recurrent neural networks," *arXiv preprint arXiv:1712.00512*, 2017.

[20] G. Repovs and D. M. Barch, "Working memory related brain network connectivity in individuals with schizophrenia and their siblings," *Frontiers in human neuroscience*, vol. 6, p. 137, 2012.

[21] M. Jenkinson, M. Pechaud, S. Smith, *et al.*, "Bet2: Mr-based estimation of brain, skull and scalp surfaces," in *Eleventh annual meeting of the organization for human brain mapping*, vol. 17, p. 167, Toronto., 2005.

[22] L. Laitinen, "Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging. by jean talairach and pierre tournoux. translated by mark rayport. georg thieme verlag, stuttgart-new york, 1988. pp. 122, figs (coloured). isbn 313711701-1 (georg thieme verlag, stuttgart; isbn 0 86577 293 2 (thieme medical publishers, inc. new york)," 1989.

[23] A. Klein, S. S. Ghosh, F. S. Bao, J. Giard, Y. Häme, E. Stavsky, N. Lee, B. Rossa, M. Reuter, E. C. Neto, *et al.*, "Mindboggling morphometry of human brains," *PLoS computational biology*, vol. 13, no. 2, p. e1005350, 2017.

[24] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, *et al.*, "fmriprep: a robust preprocessing pipeline for functional mri," *Nature methods*, vol. 16, no. 1, p. 111, 2019.

[25] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[26] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, *et al.*, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[27] M. Sharaev, A. Artemov, E. Kondrateva, S. Sushchinskaya, E. Burnaev, A. Bernstein, R. Akzhigitov, and A. Andreev, "Mri-based diagnostics of depression concomitant with epilepsy: in search of the potential biomarkers," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 555–564, IEEE, 2018.

[28] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9308–9316, 2019.

[29] L. Xiao, H. Zhang, W. Chen, Y. Wang, and Y. Jin, "Transformable convolutional neural network for text classification.," in *IJCAI*, pp. 4496–4502, 2018.

[30] A. Notchenko, Y. Kapushev, and E. Burnaev, "Large-scale shape retrieval with sparse 3d convolutional neural networks," in *Analysis of Images, Social Networks and Texts* (W. M. van der Aalst, D. Ignatov, M. Khachay, and et al., eds.), (Cham), pp. 245–254, Springer International Publishing, 2018.

[31] E. Burnaev, A. Cichocki, and V. Osin, "Fast multispectral deep fusion networks," *Bull. Pol. Ac.: Tech.*, vol. 66, no. 4, pp. 875–880, 2018.

APPENDIX

TABLE III

dVOXRESNET MODEL OPTIMIZATION: MODEL PERFORMANCE DEPENDING ARCHITECTURE STUDIED ON SCHIZOPHRENIA AND BIPOLAR DISORDER CLASSIFICATION *Validated on 3-fold CV, ROC/AUC*. D-CONVOLUTIONS ARE INSERTED IN BLOCKS ACCORDING TO *idx*: CONV3D BLOCKS [1:6], VOXRES BLOCKS [1:4], WHERE COLOUR DEFINES **STATISTICALLY SIGNIFICANT PERFORMANCE INCREASE**, **STATISTICALLY SIGNIFICANT PERFORMANCE DECREASE**

Conv3D block idx; VoxRes block idx	Shizophrenia (50)/ Healthy Control (122)			Bipolar Disorder (49)/ Healthy Control (122)		
	unprocessed	skull - striped	MNI normalized	unprocessed	skull - striped	MNI normalized
Image input size	176x200x152	176x200x152	144x176x144	176x200x152	176x200x152	144x176x144
- ; -	**0.788 +/- 0.068**	**0.778 +/- 0.075**	**0.736 +/- 0.070**	**0.639 +/- 0.088**	**0.648 +/- 0.102**	**0.639 +/- 0.105**
4 ; -	0.808 +/- 0.064	0.806 +/- 0.055	0.737 +/- 0.063	0.675 +/- 0.096	0.651 +/- 0.065	0.629 +/- 0.065
- ; 2	0.790 +/- 0.070	0.797 +/- 0.064	0.734 +/- 0.091	0.654 +/- 0.084	0.654 +/- 0.104	0.658+/- 0.099
5 ; -	0.797 +/- 0.076	0.805 +/- 0.065	0.736 +/- 0.082	0.667 +/- 0.093	0.662 +/- 0.089	0.599 +/- 0.104
- ; 3	0.801 +/- 0.074	0.802 +/- 0.066	0.721 +/- 0.071	0.680 +/- 0.093	0.637 +/- 0.081	0.629 +/- 0.073
6; -	0.783 +/- 0.072	0.802 +/- 0.068	0.739 +/- 0.062	0.638 +/- 0.107	0.674 +/- 0.092	0.622 +/- 0.088
- ; 4	0.782 +/- 0.069	0.798 +/- 0.052	0.748 +/- 0.063	0.640 +/- 0.089	0.669 +/- 0.085	0.611 +/- 0.116
4 ; 2	**0.822 +/- 0.068**	0.801 +/- 0.071	0.734 +/- 0.071	0.671 +/- 0.095	0.668 +/- 0.110	**0.594 +/- 0.075**
4, 5 ; 2	**0.817 +/- 0.045**	0.802 +/- 0.059	0.716 +/- 0.086	0.672 +/- 0.081	0.638 +/- 0.078	0.604 +/- 0.100
4, 5 ; 2, 3	**0.823 +/- 0.058**	**0.823 +/- 0.065**	0.731 +/- 0.063	0.679 +/- 0.062	0.638 +/- 0.088	**0.571 +/- 0.091**
5 ; 2, 3	**0.815 +/- 0.072**	0.803 +/- 0.054	0.745 +/- 0.078	0.669 +/- 0.092	0.655 +/- 0.079	0.659 +/- 0.092
5 ; 3	0.805 +/- 0.077	**0.812 +/- 0.062**	0.720 +/- 0.089	**0.681 +/- 0.075**	0.659 +/- 0.069	0.659 +/- 0.092
5, 6 ; 3	0.793 +/- 0.067	0.807 +/- 0.066	0.727 +/- 0.087	0.676 +/- 0.064	0.661 +/- 0.083	0.626 +/- 0.091
5, 6 ; 3, 4	0.797 +/- 0.072	**0.816 +/- 0.060**	0.722 +/- 0.072	**0.676 +/- 0.081**	**0.687 +/- 0.065**	0.631 +/- 0.097
6 ; 3, 4	0.795 +/- 0.079	0.805 +/- 0.071	0.726 +/- 0.064	0.656 +/- 0.092	0.660 +/- 0.091	0.606 +/- 0.097

TABLE IV

SCHIZOPHRENIA CLASSIFICATION ON MAIN AND ADDITIONAL DATASETS. COMPARISON OF TRANSFER OF PRE-TRAINED WEIGHTS AND FINE-TUNING THE VOXRESNET MODEL FOR SCHIZOPHRENIA VERSUS CONTROL CLASSIFICATION FROM DATASET 1 (MAIN) TO SMALLER TEST DATASET 3 (ADDITIONAL)

	5-fold CV with 3 repeats ROC AUC score, Mean (STD)			
Shizophrenia / Healthy Control	**Main** unprocessed		**Additional** unprocessed	
Conv3D block idx; VoxRes block idx	dVoxResNet	dVoxResNet	Weights transfer from **Main**	Finetuning
8ine **- ; -**	**0.794 (0.068)**	**0.742 (0.150)**	**0.669**	**0.749 (0.128)**
4 ; 2	0.822 (0.068)	0.764 (0.113)	0.681	0.740 (0.124)
4, 5 ; 2	0.817 (0.045)	0.753 (0.143)	0.653	0.755 (0.126)
4, 5 ; 2, 3	0.823 (0.058)	0.746 (0.114)	0.621	0.676 (0.140)
5 ; 2, 3	0.815 (0.072)	0.712 (0.125)	0.646	0.759 (0.113)