Limitations of machine learning in psychiatry: Participation in the PAC 2018 depression challenge

Fabian Eitel [1], **Sebastian Stober** [2], **Lea Waller** [1], **Lena Dorfschmidt** [1,3], **Henrik Walter** [1], **and Kerstin Ritter** [1,*]

[1] *Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health; Department of Psychiatry and Psychotherapy, Bernstein Center for Computational Neuroscience; 10117 Berlin, Germany.*
[2] *Institut für Intelligente Kooperierende Systeme (IKS), Artificial Intelligence Lab; Otto-von-Guericke-Universität Magdeburg, Germany.*
[3] *Collaborative Research Centre (SFB 940) "Volition and Cognitive Control", Technische Universität, Dresden, Germany.*

Correspondence*:
Kerstin Ritter; Charitéplatz 1, 10117 Berlin, Germany
kerstin.ritter@charite.de

ABSTRACT

Classifying subjects with psychiatric diseases based on structural magnetic resonance imaging (MRI) data is a highly challenging task. Nevertheless, a number of studies report decoding accuracies of up to 90% using relatively small sample sizes. Here, we will present contradicting results on a large scale ($N = 2240$) dataset of subjects with and without depression as part of our participation in the Predictive Analytics Competition (PAC) 2018, in which we achieved the 3rd place. Contrary to our expectations, classification accuracies varied only little between a variety of simple and highly-complex classifiers and did not exceed 67% based on an internal validation set. We discuss our results in light of these opposing results and conclude that well designed challenges and large sample sizes provide a good way to get unbiased estimates of prediction performance.

KEYWORDS

depression, psychiatric disease, structural MRI, neuroimaging, machine learning, convolutional neural networks, deep learning, PAC2018 challenge

1 INTRODUCTION

In the last two decades, neuroimaging data including structural and functional magnetic resonance imaging (MRI) data has become one of the cornerstones for studying not only neurological but also psychiatric diseases [1; 2]. Whereas traditional statistical analyses focused on general group comparisons, machine learning methods are intended to draw conclusions about individual subjects, e.g. in terms of diagnosis, prognosis or treatment. Although psychiatric diseases are considered to be complex and heterogeneous, a number of studies reported decoding accuracies of up to 90 % for various tasks [2; 3; 4; 5; 6; 7]. Most of these studies employed relatively small sample sizes and it has been questioned whether results reflect true effects or might be an artifact of large variation in decoding accuracies typically depicted in small sample sizes [4; 8].

Machine learning challenges provide a good opportunity to address this issue by using a two-step procedure. First, a public dataset is published, on which the different teams can develop their models. Second, only the examples (but not the labels) from a holdout dataset is published, on which teams can make their predictions. Finally, the organizers can evaluate the predictions using the labels of the holdout dataset. Here, we will describe our experiences with the PAC 2018 challenge[1] organized by the Translational Psychiatry Group at University Muenster, Germany[2], in which we got the 3rd place out of 49 teams. In this challenge, structural MRI data sets and additional covariates from $N = 2240$ subjects with and without depression were provided. The task was to train a classifier to detect depression. In this work we show that no algorithm from a variety of simple and highly-complex classifiers was able to outperform a modest balanced accuracy of around 67% on the internal validation set. Our final submission on the holdout data set ($N = 448$) resulted in a balanced

[1] https://www.photon-ai.com/pac
[2] https://www.medizin.uni-muenster.de/translap/forschung/

accuracy of 64 %, whereas the winning team has achieved 65%.

2 METHODS

2.1 Data

The data set included spatially normalized voxel-based morphometry (VBM) images (i.e. gray matter probability maps) of $N = 1792$ subjects with and without depression (759 vs. 1033). Raw data were preprocessed by the challenge organizers using the SPM toolbox CAT-12 (Matlab 9.0 / SMP12 rev. 6685 / CAT12 v.1184) and quality checked in-house. Additionally, for each subject the following covariates were provided: Age (mean 35.55 ± 12.70), gender (59% females), imaging site (3 different sites) and total intracranial volume (TIV, mean 1571 ± 167). At the end of the challenge, the same data were released for the holdout samples ($N = 448$) excluding the disease labels.

2.2 Classification methods

Based on the public data set, we performed a number of different classification analyses using (1) only covariates (2) only VBM images and (3) both, covariates and VBM images. For classification we employed four machine learning algorithms, namely support vector machines (SVM), random forests (RF), gradient boosting (GB) and convolutional neural networks (CNNs). SVM, RF and GB were applied to (i) covariates only, (ii) mean values of VBM images for all cortical regions ($N = 48$) contained in the Harvard Oxford atlas[3] and (iii) whole-brain based principal component analysis (PCA) projections. For CNNs, we computed the following variants: a) vanilla CNN b) inclusion of additional data in the training base ($N = 372$, 181 with depression)[4], c) transfer learning from Alzheimer's Disease Neuroimaging Initiative[5] (ADNI, $N = 747$) data with fine-tuning on the PAC2018 data and d) multi-task learning using covariates as additional outputs which the network needs to predict. In total we present the results of 13 different configurations as shown in Table 1.

We split the public data set into a pure (balanced) training data set (90 % of subjects with depression, same number of subjects without depression, total $N = 1366$) and a validation set ($N = 426$). Hyperparameters were tuned on the training data using grid search in a 5-fold cross-validation (PCA: number of components $= [10, 100, 1000]$; SVM: linear and radial basis function [RBF], $C, \gamma = [0.01, 0.1, 1, 10, 100]$; RF: number of estimators $= [10, 100, 1000]$, number of features considered for split$= [5, 10, 15]$, maximum depth $= [\text{None}, 0.1, 1, 10]$); GB: number of estimators $= [10, 100, 1000]$, learning rate $= [0.01, 0.1, 1]$ and fraction of samples used for fitting $= [0.01, 0.1, 1]$). Due to the large number of configurations the CNN network architectures and hyperparameters have been tuned manually. The respective classification algorithm was then trained again with best parameters on the full training data set and tested on the validation set, for which we report the balanced accuracy, sensitivity and specificity. For our final model, we additionally report the accuracy measures for the externally provided holdout data from the PAC2018 challenge.

Table 1. Performance (in %) for different classification methods on the internal validation set.

Data	Extr.	Classifier	Bal. acc.	Sens.	Spec.
Cov.	-	SVM	64.52 %	69.33 %	59.71 %
Cov.	-	RF	63.62 %	62.67 %	64.57 %
Cov.	-	GB	62.10 %	61.33 %	62.86 %
VBM	Atlas	SVM	66.58 %	63.17 %	70.00 %
VBM	Atlas	RF	63.62 %	62.67 %	64.57 %
VBM	Atlas	GB	60.75 %	65.79 %	55.71 %
VBM	PCA	SVM	65.04 %	65.79 %	64.29 %
VBM	PCA	RF	57.54 %	47.37 %	67.71 %
VBM	PCA	GB	62.29 %	56.58 %	68.00 %
VBM	-	CNN	64.52 %	53.33 %	75.71 %
VBM	-	CNN + ext. data	66.52 %	64.47 %	68.57%
VBM	-	CNN + transfer	64.33 %	70.66 %	58.00 %
VBM	-	CNN + multitask	62.29 %	64.00%	60.57%

Abbreviations: Cov., covariates; VBM, voxel-based morphometry; Extr., extraction; PCA, principal component analysis; SVM; support vector machine; RF, random forest; GB, gradient boosting; Bal. acc., balanced accuracy; Sens., sensitivity; Spec., specificity

3 RESULTS

Classification results for our internal validation set are depicted in Table 1. For only covariates, the highest balanced accuracy was 64.52 % using a SVM with a RBF kernel ($C = 1, \gamma = 0.1$). For the VBM data in combination with classical machine learning techniques, a combination of atlas-based feature extraction and a SVM with RBF kernel ($C = 1, \gamma = 0.01$) was best, resulting in a balanced

[3] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/data/atlas-descriptions.html

[4] 166 from the Human Connectome Project (http://www.humanconnectomeproject.org/) and 206 from our local data base

[5] http://adni.loni.usc.edu/

accuracy of 66.58 %. For the CNNs, we obtained the best result (66.52 %) using additional external data in a model consisting of 6 convolutional blocks (Conv-BatchNorm-ReLU) and 3 fully-connected layers (drop out of 0.2 before the first two fully-connected layers).

For evaluation on the external holdout data provided in the PAC 2018 challenge, we had to choose one algorithm. Although the atlas-based SVM gave us slightly higher results, we decided to use the WB-based CNN + external data, since sensitivity and specificity were more balanced and we anticipated that the inclusion of external data might be advantageous on the holdout data. On the holdout data, we got a balanced accuracy of 64 % (sensitivity 58 %, specificity 69 %).

4 DISCUSSION

Based on the PAC 2018 depression challenge data set, we explored a number of classical and more advanced hierarchical techniques for the detection of depression. Notably, the classification accuracy did not exceed 67 % for any of the particular analyses based on the validation set. Additionally, the classification accuracies were in a similar range, irrespective of the underlying data (covariates, VBM data or both) and types of dimensionality reduction or machine learning algorithm. Results on the external data set provided in the challenge were in the same range, not exceeding 65 %.

Given a number of studies that reported classification accuracies of up to 90 % for the classification of depression based on structural MRI data (for an overview, see Wolfers et al. [2]), not only our team were surprised in light of these results. We would like to discuss the following potential explanations. First, most studies that have found high classification accuracies employed relatively small sample sizes (below 100 subjects per group) and therefore results might be explained by a combination of large error bars characteristic for small sample sizes and a positive publication bias [4; 8].

Second, the data at hand might not be suitable for this task. Within the PAC 2018 challenge, only highly preprocessed VBM images and a small set of covariates were released. While this leads on the one hand to a clearer classification task, it strongly limits the ways of data analysis, e.g. with respect to different preprocessing pipelines or further stratification of patients according to some clinical variables such as symptom severity. Additionally, other data domains, such as functional MRI, might be more relevant for depression [3].

Third, the labels in the psychiatric field might not be reliable enough for being employed in a (supervised) machine learning framework. Although psychiatric research is not imaginable without clinical labels such as depression, those diagnostic categories have been severely criticized for not incorporating underlying neurobiological correlates and their limited ability to account for heterogeneity as well as comorbidities within and across clinical categories [9; 10; 11; 12].

And finally, there might exist better methods for analyzing this specific data set. However, since we sampled our techniques from a wide range of traditional and advanced machine learning, we believe that our analyses reflect at least some current state-of-the-art of data analysis in the neuroimaging field. Additionally, the risk for missing a suitable data pipeline is further alleviated by the participation of 49 teams applying their best algorithms to the same data set.

In conclusion, we believe that such machine learning challenges provide a huge value for the scientific community, especially in fields with very high-dimensional data and complex categories as given in clinical neuroimaging. They are a great opportunity to make large data sets available to a broader range of people, who can benchmark their specific algorithms and pipelines. Since results will be evaluated on an external data set, the chance for p-hacking and publishing false positives is strongly reduced [13] and realistic estimates of performance measures can be obtained.

Author contributions

FE, SS, LW, LD, HW and KR designed the study. FE, LW and KR analyzed the data. FE, SS, HW and KR wrote the manuscript.

Competing interests

The authors declare no conflict of interest.

Grant information

We acknowledge support from the German Research Foundation (KR: DFG, 389563835), the Manfred and Ursula-Müller Stiftung (KR), Charité – Universitätsmedizin Berlin (KR: Rahel-Hirsch scholarship) and Studienstiftung des Deutschen Volkes (LD).

REFERENCES

[1]Graziella Orrù, William Pettersson-Yeo, Andre F Marquand, Giuseppe Sartori, and Andrea Mechelli. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease:

a critical review. *Neuroscience & Biobehavioral Reviews*, 36:1140–1152, 2012.

[2]Thomas Wolfers, Jan K Buitelaar, Christian Beckmann, Barbara Franke, and Andre F Marquand. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience & Biobehavioral Reviews*, pages –, 2015.

[3]Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, Alan F Schatzberg, Keith Sudheimer, Jennifer Keller, Helen S Mayberg, Faith M Gunning, George S Alexopoulos, Michael D Fox, Alvaro Pascual-Leone, Henning U Voss, BJ Casey, Marc J Dubin, and Conor Liston. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, 23(1):28–38, 1 2017.

[4]Choong-Wan Woo, Luke J Chang, Martin A Lindquist, and Tor D Wager. Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, 20(3):365–377, 3 2017.

[5]Carles Soriano-Mas, Jesús Pujol, Pino Alonso, Narcís Cardoner, José M Menchón, Ben J Harrison, Joan Deus, Julio Vallejo, and Christian Gaser. Identifying patients with obsessive-compulsive disorder using whole-brain anatomy. *NeuroImage*, 35(3):1028–1037, 2007.

[6]Sandra Vieira, Walter H.L. Pinaya, and Andrea Mechelli. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75, 3 2017.

[7]Benson Mwangi, Klaus P. Ebmeier, Keith Matthews, and J. Douglas Steele. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. *Brain*, 135(5):1508–1521, 5 2012.

[8]Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 10 2018.

[9]Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, 167(7):748–751, 2010.

[10]Thomas Insel and Bruce N Cuthbert. Brain disorders? Precisely. *Science*, 348(6234):499–500, 2015.

[11]Bruce N Cuthbert and Thomas R Insel. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, 11(1):126, 2013.

[12]Sarah L Karalunas, Damien Fair, Erica D Musser, Kamari Aykes, Swathi P Iyer, and Joel T Nigg. Subtyping attention-deficit/hyperactivity disorder using temperament dimensions: toward biologically based nosologic criteria. *JAMA psychiatry*, 71(9): 1015–1024, 2014.

[13]John P A Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 8 2005.