Patch-based abnormality maps for improved deep learning-based classification of Huntington's disease

Kilian Hett[1], Rémi Giraud[2], Hans Johnson[3], Jane S. Paulsen[4,5], Jeffrey D. Long[5,6], and Ipek Oguz[1]

[1] Vanderbilt University, Dept. of Electrical Engineering and Computer Science, Nashville, TN, USA
[2] Bordeaux INP, University of Bordeaux, CNRS, IMS, UMR 5218, Talence, France
[3] University of Iowa, Dept. of Electrical and Computer Engineering, Iowa City, IA, USA
[4] University of Iowa, Dept. of Neuroscience, Iowa City IA, USA
[5] University of Iowa, Dept. of Psychiatry, Iowa City IA, USA
[6] University of Iowa, Dept. of Biostatitsics, Iowa City IA, USA

Abstract. Deep learning techniques have demonstrated state-of-the-art performances in many medical imaging applications. These methods can efficiently learn specific patterns. An alternative approach to deep learning is patch-based grading methods, which aim to detect local similarities and differences between groups of subjects. This latter approach usually requires less training data compared to deep learning techniques. In this work, we propose two major contributions: first, we combine patch-based and deep learning methods. Second, we propose to extend the patch-based grading method to a new patch-based abnormality metric. Our method enables us to detect localized structural abnormalities in a test image by comparison to a template library consisting of images from a variety of healthy controls. We evaluate our method by comparing classification performance using different sets of features and models. Our experiments show that our novel patch-based abnormality metric increases deep learning performance from 91.3% to 95.8% of accuracy compared to standard deep learning approaches based on the MRI intensity.

Keywords: Patch-based method · Deep learning · Huntington's disease

1 Introduction

Huntington's disease (HD) is a fatal autosomal dominant neurodegenerative disorder that causes motor, behavioral and cognitive abnormalities. The pathological mutation consists of an abnormal cytosine-adenine-guanine (CAG) repeat in the huntingtin gene (HTT) [4]. Gene modification leads to pathological brain changes, and imaging studies have shown structural changes in the striatum [15]. Unlike many other degenerative diseases, a genetic test can determine the presence of the mutated gene well before the onset of symptoms, which makes HD a good candidate for the evaluation of new imaging-based methods.

Recently, deep learning methods have reached state-of-the-art performances in many medical imaging problems. However, such approaches have only obtained similar performance to methods using a combination of good feature engineering and machine learning methods for the detection and prediction of neurodegenerative diseases [1]. Although deep-learning approaches have shown promising results, one of the main limitations is a lack of large annotated training datasets, combined with the high dimensionality of the medical image data. Recent methods proposed to combine deep learning with pre-processed features [16,11]. These approaches proposed to use as input of deep-learning networks gray matter (GM) volumes to learn abnormalities over each subject. This results in normalized features that not only substantially reduce the problem dimensionality but also have lower variability compared to the raw MRI intensities. More advanced features have also been proposed for classification tasks to describe subtle anatomical changes. Among them, the patch-based grading framework has shown state-of-the-art performances [2,17,18,8,7]. This framework aims to detect local similarities of a given test image to two template libraries representing two different populations (e.g., diseased vs. healthy). The general idea of patch-based grading methods is to compare structural patterns of a local patch in a test image with the images in the template libraries. However, such methods are highly dependent on the two template libraries, which can be affected by many factors (differences in distributions of MR protocols, sex, age, etc). A proposed solution is to use a patch-based framework to detect abnormal patterns by using the coefficients of sparse coding to assess the subject under study [3].

Just like the pre-processed features proposed by [16,11], we hypothesize that using a normalized patch-based grading maps as an input feature will lower the variability that limits deep learning models. In addition, deep neural networks can efficiently learn abnormality patterns over the brain without requiring segmentation maps. This is the **first contribution** of our paper. While the patch-based grading method is very powerful when appropriate template libraries can be identified for a given task, this can sometimes be difficult to achieve. For example, a real dataset may contain more than two distributions: a movement disorders clinic may deal with not only healthy controls and Parkinson's patients, but also Lewy body disease patients, and essential tremor patients. Even within a single disease, disease heterogeneity may make it impractical to build a representative template library. In such scenarios, it is much easier to rely on a single template library distribution representing healthy controls, and report the local deviation of a given test subject from the healthy control distribution. The **second contribution** of our paper is a new patch-based abnormality metric which achieves this goal.

2 Materials and Methods

2.1 Dataset

All T1-weighted (T1w) MRIs come from PREDICT-HD [12], which is a multi-site longitudinal study of HD. The MRIs have been acquired using 3 Tesla MRI

Table 1. Demographic description of the dataset used in our experiments. Subjects are divided into three populations: healthy control (HC), Pre-Manifest HD (*i.e.*, abnormal CAG repetition in the HTT gene without motor impairment), and Manifest HD (*i.e.*, abnormal CAG with motor impairment).

	Healthy control	Huntington's disease	
		Pre-Manifest	Manifest
Number of MRIs	327	300	117
Age (years)	49.3±11.9	43.4±10.1	56.8±6.5
Sex (F/M)	206/121	199/101	28/28
CAG length	15-35	37-57	37-57

scanners from different vendors (*e.g.*, GE, Phillips, and Siemens). The cohort used in the study includes 683 MPRAGE images from subjects representing three populations: control subjects (HC), pre-manifest HD that is composed of subjects with the expanded CAG repeat but who have not yet received a motor diagnosis at the time of the scan, and manifest HD which refers to patients who already have a motor diagnosis by the time of the scan (see Table 1).

2.2 Preprocessing

The preprocessing was conducted with the BRAINSAutoWorkup pipeline [14]. This pipeline is composed of the following steps: (1) denoising with non-local means filter, (2) anterior/posterior commissure and intra-subject alignments with rigid transformation, (3) bias field correction, and (4) regional segmentation with a multi-atlas method using atlases from Neuromorphometrics[7]. Finally, MRI intensities have been standardized using a piece-wise linear histogram normalization technique.

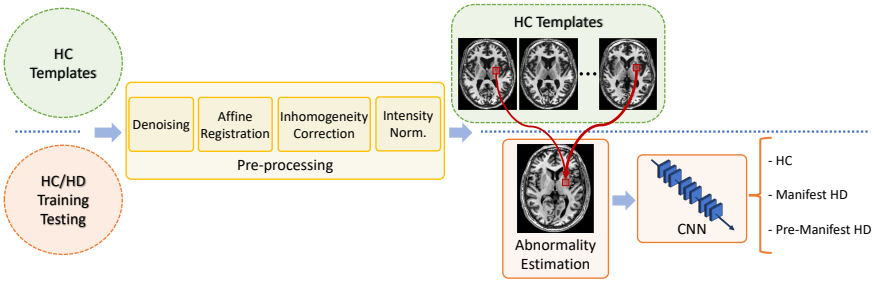

Fig. 1. Pipeline of the proposed method. First, HC from the dataset is separated into two subset, the HC templates used to estimate the local abnormality, the second is the set of HC for the evaluation of our method. Once all MRIs are preprocessed, we estimate the local abnormality using the HC template library. Finally, a convolutional neural network with softmax is used to obtain final classification.

[7] `http://www.neuromorphometrics.com`

2.3 Patch-based metrics

We consider two patch-based metrics (*i.e.*, patch-based grading and our proposed patch-based abnormality) to detect structural changes. Both methods rely on the accurate selection of closest patches in terms of intensity differences. In our work, closest patch selection is done using a version of PatchMatch that ensures uniqueness of patches extracted for the template library [5]. PatchMatch is an optimized algorithm that enables us to extract the most similar patches from the template library in a sparse selection fashion.

Patch-based grading (PBG). Patch-based grading has been introduced to detect local similarity between two populations of subjects [2]. At each voxel x of the subject under study, this method estimates a grading value $g(x)$ based on the following equation:

$$g(x) = \frac{\sum_{T \in K_x} \exp\Big(- \frac{||S(x) - T(y)||_2^2}{h(x)} \Big) p_T}{\sum_{T \in K_x} \exp\Big(- \frac{||S(x) - T(y)||_2^2}{h(x)} \Big)}, \tag{1}$$

where $S(x)$ is the patch surrounding the voxel x of the test subject image. $T(y)$ is a patch from the set K_x, which contains the most similar patches to $S(x)$ from the training library, as determined by the PatchMatch algorithm. $h(x) = min||S(x) - T(y)||_2^2$ is a normalization factor. Finally, p_T is the pathological status set to -1 for patches extracted from HD patients and to 1 for those extracted from HC subjects.

Patch-based abnormality (PBA). Our method derives from the patch-based grading framework. To address PBG's dependence on the two template libraries, we estimate the local differences from a single template library composed only of HC subjects (see Fig. 1). The abnormality $a(x)$ for each voxel x of the MRI under study, is defined as:

$$a(x) = \frac{\sum_{T \in K_x} ||S(x) - T(y)||_2^2}{\sigma_{S(x)}}, \tag{2}$$

where $\sigma_{S(x)}$ is standard deviation of intensities over the patch $S(x)$, which normalizes the differences of signal intensity contained in each patch $S(x)$. Similar to Eq. 1, K_x is the set of closest patches provided by the PatchMatch algorithm. This results in a low abnormality metric if the current patch is similar to age-matched control subjects, and in a high abnormality metric if the patch does not fit well within the distribution of age-matched control subjects (see Fig. 2).

2.4 Network architecture

In order to model the spatial disease signature and perform the subject-level classification, we used a convolutional neural network (CNN) approach. In recent

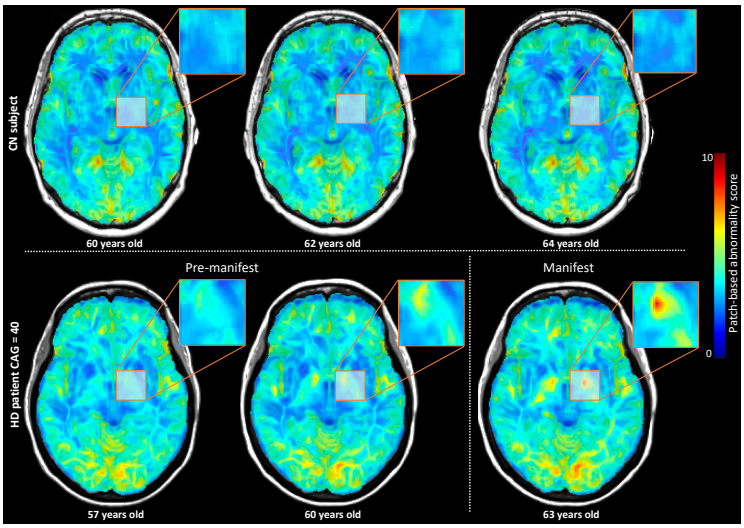

Fig. 2. Illustration of patch-based abnormality maps for **(top)** a healthy control subject and **(bottom)** an HD patient with 40 CAG repeats. **From left to right**, 3 different time points are shown for each subject. The HD subject is in the pre-manifest stage for the first time points, but converts to clinical diagnosis by the third time point. **Blue** represents areas with a low abnormality score $a(x)$, whereas **red** represents areas with high abnormality score $a(x)$. We note a progressive increase of abnormality near the basal ganglia during the course of the disease which is consistent with HD pathology, while the abnormality map for the HC subject remains stable.

years, many different architectures have been proposed in the pattern recognition field. Among them, deep residual neural network (ResNet) has shown competitive performances [6]. This architecture is characterized by skipped connections of different blocks of layers (see Fig. 3). ResNet has demonstrated a reduced training error compared to other networks with similar depth. Indeed, the residual mapping enables to reduce the training error, which is generally correlated with the network depth for classic stacked architectures. In addition, to address the problem of GPU memory limitation, we used a 3D patch approach. Thus, both networks have as input 8 channels that represent non-overlapping patches from the input data (*i.e.*, T1w MRI or PBA maps).

2.5 Implementation and evaluation

We evaluate our proposed method on two classification tasks: HC versus manifest HD patients, and HC versus pre-manifest HD patients. MRIs from HC subjects have been divided in two disjoint subsets, one to construct the patch-based abnormality maps and the others to train the model. The data partition results in 163 MRIs from HC subjects for the computation of PBA, 468 images for HC (164) vs. Pre-manifest HD (300), and 280 MRI for CN (164) vs. Manifest

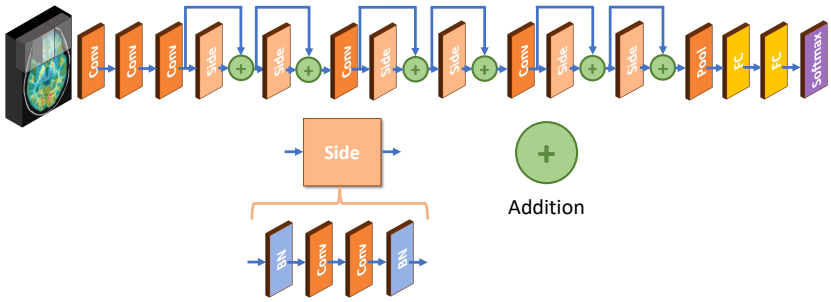

Fig. 3. Illustration of the convolutional neural network architecture used to validate our work. The architecture consist of a combination of convolutional layer (Conv), batch normalization (BN), Skipped connection layer (Side), pooling layer (Pool), and fully connected layer (FC). A softmax layer estimates the probability for each class.

HD (117). We tailored the template library for the estimation of PBA maps by selecting a set of 30 HC MRI using an age-matching technique. The patch size for the abnormality metric computation has been set to $7 \times 7 \times 7$ voxels.

The network was trained using Adam optimizer with a constant learning rate equals to 0.0001. We used cross-entropy as loss function and a batch size of 2 over 20 epochs. For the deep learning network, the patch size was set to $64 \times 64 \times 64$ voxels. A stratified 5-fold cross-validation was conducted to obtain training and testing sets. Because of the longitudinal nature of our dataset, care was taken to ensure different timepoints from the same patient were either all in the template library, all in the training set or all in the testing set, to avoid any leakage. We estimated classification performance in terms of precision, accuracy, specificity, sensitivity, and area under the ROC curve. All code can be found online[8].

To evaluate our methods, we compare the results obtained using 6 different classification methods: LDA classifier with putamen volume, LDA classifier with classic PBG feature [2] averaged over the putamen using the same parameters described in [9], LDA classifier with proposed PBA feature averaged over the putamen, ResNet with T1w input, ResNet with proposed PBA input, and ResNet with the concatenation of T1w and PBA inputs. For the first three experiments, we use the multi-atlas based putamen segmentation mask (Sec. 2.2).

3 Results and Discussion

The results of the HC vs. Manifest-HD classification task are shown in Table 2. The ResNet classifier using PBA features as input has the best performance for all classification metrics.

We note that the putamen volume, which is well known to be a crucial marker of HD progression [13,10], dramatically outperforms (by about 15 AUC points) the patch-based approaches if a simple LDA classifier is used (PBG and

[8] `https://github.com/hettk/patch-based_abnormality`

PBA maps are each summarized by averaging within the putamen ROI in this setup). This may be due to inability of the simple classifier to model the complex patterns of neurodegeneration, as well as imperfections in the putamen segmentation. Using the ResNet classifier and the whole feature maps instead of the average within putamen, we see a big improvement in classification performance, consistent with our hypothesis. Among the ResNet classifiers, the model using the proposed PBA feature substantially outperforms the model using the raw T1w intensities. Indeed, the ResNet with PBA feature input is the only model that outperforms the simple LDA classifier on putamen volumes. Finally, we note that concatenating the T1w and PBA features within ResNet does not improve classification performance. This may potentially be because PBA extracts the same information from T1w by highlighting regions impacted by changes due to the neurological disease. Consequently, although PBA makes it easier to detect controls from HD patients, our experiments are inconclusive about any complementarity between PBA and T1w.

Table 3 shows the results from the same experiments on the HC vs. Pre-Manifest HD classification task. We note that most of the observations from Table 2 similarly hold for this experiment: putamen volume outperforms patch-based methods with simple LDA classifier; ResNet substantially improves upon the LDA based performance for any of the patch-based grading metrics; and the ResNet performance tends to be best when using the proposed PBA and grading metrics. PBA obtains similar results to grading but without requiring a second template library, which is problematic. Indeed, we did not perform HC vs. Manifest HD to avoid double-dipping, since the Manifest HD group has been used to compute PBG. We note that, unlike the HC vs. Manifest HD task, the concatenation of T1w and PBA features in ResNet does improve some of the performance metrics, suggesting some potential complementarity. Additionally, again unlike the HC vs. Manifest HD task, no patch-based method reaches the performance of the putamen volume. This may suggest that at this stage of the disease progression, the local abnormalities are more subtle and the global volume is a more robust marker of pathology. In future work, we will explore

Table 2. Comparison of different methods for HC versus manifest HD classification. Results represent the average of classification performances for 5 xval folds. All results are expressed in terms of percentage. Best and second scores for each metric are expressed in bold and underlined font, respectively.

	Classifier	AUC	Precision	Accuracy	Specificity	Sensitivity
Putamen volume	LDA	<u>96.1±0.3</u>	<u>89.1±0.3</u>	<u>93.5±0.4</u>	<u>96.6±0.2</u>	<u>89.1±0.3</u>
Putamen grading	LDA	81.0±0.5	75.9±1.2	75.9±1.6	87.5±1.4	60.0±1.0
Putamen PBA	LDA	81.6±0.5	73.2±1.1	76.2±1.6	89.5±1.6	58.5±1.2
ResNet T1w	Softmax	90.7±1.8	87.7±2.4	91.3±1.7	95.2±0.9	85.0±3.0
ResNet PBA	Softmax	**96.3±0.7**	**96.9±0.6**	**95.8±0.8**	**98.8±0.2**	**91.0±1.8**
ResNet T1w + PBA	Softmax	91.6±1.6	85.0±3.0	90.9±1.8	96.4±0.7	82.0±3.6

Table 3. Comparison of different methods for HC versus pre-manifest HD classification. Results represent the average of classification performances for 5 xval folds. All results are expressed in terms of percentage. Best and second scores for each metric are expressed in bold and underlined font, respectively.

	Classifier	AUC	Precision	Accuracy	Specificity	Sensitivity
Putamen volume	LDA	**93.9±0.3**	**93.1±0.6**	85.7±0.5	72.5±0.3	93.1±0.6
Putamen grading	LDA	74.5±1.6	87.2±1.9	72.6±1.7	45.7±1.9	87.2±1.8
Putamen PBA	LDA	77.6±1.2	87.7±1.8	72.5±1.5	45.6±1.6	87.7±1.8
ResNet T1w	Softmax	89.8±2.0	87.7±1.4	88.8±1.3	68.1±3.9	98.6±1.5
ResNet grading	Softmax	92.1±1.6	91.4±1.3	90.5±1.3	75.9±3.8	97.8±1.2
ResNet PBA	Softmax	91.9±1.6	90.7±1.3	91.4±1.3	75.7±3.8	**98.6±1.2**
ResNet T1w + PBA	Softmax	89.0±2.2	92.3±1.4	**91.6±1.6**	**80.5±3.7**	97.6±1.1

combining the putamen volume with the patch-based deep learning model to better capture the disease pathology.

It is worth noting that the performance of the LDA classifier with the patch-based methods is lower than what has been reported in other similar tasks (e.g., hippocampus analysis for Alzheimer's classification). However, the results we present here are consistent with previous work that reports lower detection performance for patch-based methods comparing intensity within the putamen for HD classification [9]. One of the main strengths of our method that combines patch-based and deep learning approaches is its independence from segmentation of a region of interest. This addresses the dependence of current patch-based grading methods on accurate segmentation maps to aggregate grading values into a final ROI-based feature [2,8].

4 Conclusion

In this paper we proposed a new patch-based framework to estimate local abnormalities to improve classification performance of a deep learning method in the context of HD detection. The distance from a distribution of healthy controls is estimated using a patch-match scheme preserving the uniqueness of patches extracted from the control library. Our experiments demonstrated superior classification performance of convolutional neural network when the patch-based abnormality maps are used as input of the network compared to the straightforward use of T1w intensities. In future work, we will investigate the combination of pre-trained and data augmentation techniques with our novel approach, as well as the incorporation of putamen volume into our model.

Acknowledgements. This work was supported, in part, by the NIH grant R01-NS094456. The PREDICT-HD study was funded by the National Center for Advancing Translational Sciences, and the National Institutes of Health (NIH; NS040068, NS105509, NS103475) and CHDI.org. Vanderbilt University Institutional Review Board has approved this study.

References

1. Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D.: Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage **145**, 137–165 (2017)
2. Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Allard, M., Collins, D.L., Initiative, A.D.N., et al.: Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. NeuroImage: Clinical **1**(1), 141–152 (2012)
3. Coupé, P., Deledalle, C.A., Dossal, C., Allard, M., Initiative, A.D.N., et al.: Sparse-based morphometry: Principle and application to alzheimer's disease. In: International Workshop on Patch-based Techniques in Medical Imaging. pp. 43–50. Springer (2016)
4. Dayalu, P., Albin, R.L.: Huntington disease: pathogenesis and treatment. Neurologic clinics **33**(1), 101–114 (2015)
5. Giraud, R., Ta, V.T., Papadakis, N., Manjón, J.V., Collins, D.L., Coupé, P., Initiative, A.D.N., et al.: An optimized patchmatch for multi-scale and multi-feature label fusion. NeuroImage **124**, 770–782 (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Hett, K., Ta, V.T., Catheline, G., Tourdias, T., Manjón, J.V., Coupe, P.: Multimodal hippocampal subfield grading for Alzheimer's disease classification. Scientific Reports **9**(1), 1–16 (2019)
8. Hett, K., Ta, V.T., Manjón, J.V., Coupé, P., Initiative, A.D.N., et al.: Adaptive fusion of texture-based grading for Alzheimer's disease classification. Computerized Medical Imaging and Graphics **70**, 8–16 (2018)
9. Hett, K., Johnson, H., Coupé, P., Paulsen, J.S., Long, J.D., Oguz, I.: Tensor-based grading: A novel patch-based grading approach for the analysis of deformation fields in huntington's disease. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1091–1095. IEEE (2020)
10. Kim, E.Y., Lourens, S., Long, J.D., Paulsen, J.S., Johnson, H.J.: Preliminary analysis using multi-atlas labeling algorithms for tracing longitudinal change. Frontiers in neuroscience **9**, 242 (2015)
11. Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D.: Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. Medical image analysis **48**, 117–130 (2018)
12. Paulsen, J.S., Langbehn, D.R., Stout, J.C., Aylward, E., Ross, C.A., Nance, M., Guttman, M., Johnson, S., MacDonald, M., Beglinger, L.J., et al.: Detection of Huntington's disease decades before diagnosis: the Predict-HD study. Journal of Neurology, Neurosurgery & Psychiatry **79**(8), 874–880 (2008)
13. Paulsen, J.S., Long, J.D., Johnson, H.J., Aylward, E.H., Ross, C.A., Williams, J.K., Nance, M.A., Erwin, C.J., Westervelt, H.K., Harrington, D.L., et al.: Clinical and biomarker changes in premanifest huntington disease show trial feasibility: a decade of the PREDICT-HD study. Frontiers in aging neuroscience **6**, 78 (2014)
14. Pierson, R., Johnson, H., Harris, G., Keefe, H., Paulsen, J.S., Andreasen, N.C., Magnotta, V.A.: Fully automated analysis using BRAINS: AutoWorkup. NeuroImage **54**(1), 328–336 (2011)

15. Ross, C.A., Aylward, E.H., Wild, E.J., Langbehn, D.R., Long, J.D., Warner, J.H., Scahill, R.I., Leavitt, B.R., Stout, J.C., Paulsen, J.S., et al.: Huntington disease: natural history, biomarkers and prospects for therapeutics. Nature Reviews Neurology **10**(4), 204 (2014)
16. Suk, H.I., Lee, S.W., Shen, D., Initiative, A.D.N., et al.: Deep ensemble learning of sparse regression models for brain disease diagnosis. Medical image analysis **37**, 101–113 (2017)
17. Tong, T., Gao, Q., Guerrero, R., Ledig, C., Chen, L., Rueckert, D., Initiative, A.D.N., et al.: A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease. IEEE Transactions on Biomedical Engineering **64**(1), 155–165 (2016)
18. Tong, T., Ledig, C., Guerrero, R., Schuh, A., Koikkalainen, J., Tolonen, A., Rhodius, H., Barkhof, F., Tijms, B., Lemstra, A.W., et al.: Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting. NeuroImage: Clinical **15**, 613–624 (2017)