

Supervised novelty detection in brain tissue classification; with an application to white matter hyperintensities

Hugo J. Kuijf^a, Pim Moeskops^a, Bob D. de Vos^a, Willem H. Bouvy^b, Jeroen de Bresser^c,
Geert Jan Biessels^b, Max A. Viergever^a, and Koen L. Vincken^a

^aImage Sciences Institute, University Medical Center Utrecht, Utrecht, the Netherlands;

^bDepartment of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht,
Utrecht, the Netherlands;

^cDepartment of Radiology, University Medical Center Utrecht, Utrecht, the Netherlands

ABSTRACT

Novelty detection is concerned with identifying test data that differs from the training data of a classifier. In the case of brain MR images, pathology or imaging artefacts are examples of untrained data. In this proof-of-principle study, we measure the behaviour of a classifier during the classification of trained labels (i.e. normal brain tissue). Next, we devise a measure that distinguishes normal classifier behaviour from abnormal behaviour that occurs in the case of a novelty. This will be evaluated by training a k NN classifier on normal brain tissue, applying it to images with an untrained pathology (white matter hyperintensities (WMH)), and determine if our measure is able to identify abnormal classifier behaviour at WMH locations.

For our k NN classifier, behaviour is modelled as the mean, median, or q1 distance to the k nearest points. Healthy tissue was trained on 15 images; classifier behaviour was trained/tested on 5 images with leave-one-out cross-validation. For each trained class, we measure the distribution of mean/median/q1 distances to the k nearest point. Next, for each test voxel, we compute its Z-score with respect to the measured distribution of its predicted label. We consider a Z-score ≥ 4 abnormal behaviour of the classifier, having a probability due to chance of 0.000032.

Our measure identified $>90\%$ of WMH volume and also highlighted other non-trained findings. The latter being predominantly vessels, cerebral falx, brain mask errors, choroid plexus. This measure is generalizable to other classifiers and might help in detecting unexpected findings or novelties by measuring classifier behaviour.

Keywords: novelty detection, brain, MRI, white matter hyperintensities, abnormalities, segmentation, detection

1. INTRODUCTION

Supervised voxel classification is a commonly used technique to segment brain MR images.¹⁻⁴ The normal human brain roughly consists of grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). In a clinical setting, brain MR images often show pathologies and imaging artefacts.⁵ A voxel classifier, trained on normal tissue, will incorrectly assign one of the trained labels to a voxel containing pathology or an artefact.

A possible solution is to include pathologies in the training data.⁶⁻⁸ However, it is generally hard to acquire a sufficient number of labelled samples to fill all relevant parts of the classifier feature space.⁹ Next to this, unexpected findings, rare pathologies, and imaging artefacts are difficult to train.

Novelty detection tries to identify test data that somehow differs from the training data.^{10,11} Thus, for a classifier trained on normal brain tissue, a novelty detection algorithm should identify pathologies, unexpected findings, and imaging artefacts.

Corresponding author:

Hugo J. Kuijf, Image Sciences Institute, University Medical Center Utrecht, Heidelberglaan 100, Room Q.02.445, 3584 CX, Utrecht, the Netherlands. E-mail: hugok@isi.uu.nl. Phone: +31 88 75 58562.

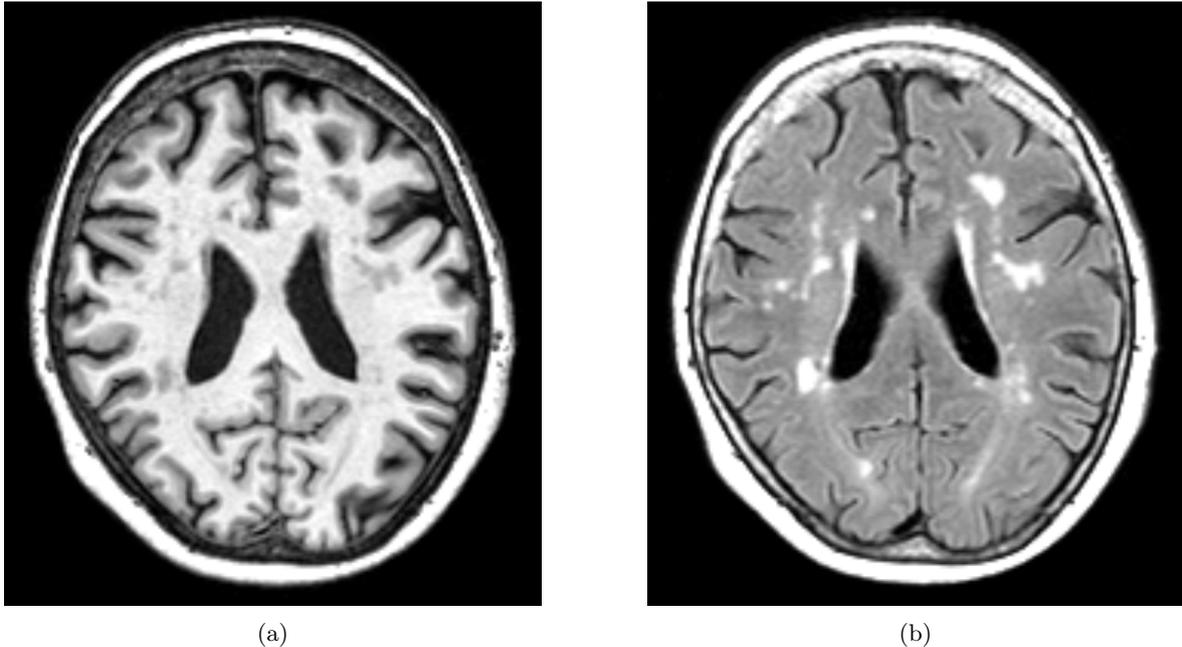


Figure 1: Example T1 (left) and FLAIR (right) sequence. Note the WMH appearing hypointense on T1 and hyperintense on FLAIR.

In this proof-of-principle study, we propose a novelty detection algorithm that relies on the behaviour of a classifier during classification. The behaviour of a classifier is measured during the classification of normal data. When classifying unknown test data, we determine if the classifier is behaving abnormally (as defined by a Z-score ≥ 4 to the measured normal behaviour) and thus encountered a novelty. This approach will be evaluated with a k NN classifier trained on normal brain tissue and applied to test data with white matter hyperintensities (WMHs).

2. METHODS AND MATERIALS

2.1 Participants and MRI

Twenty participants (age: 71 ± 4 years, 10 men) were recruited at the University Medical Center Utrecht, Utrecht, the Netherlands. Participants included patients with diabetes and matched controls, all with varying degrees of cerebral atrophy and WMH. The study was approved by the medical ethics committee of the University Medical Center Utrecht and all participants gave written informed consent.¹² The data has been previously used in the MRBrainS13 challenge (for full details, see <http://mrbrains.isi.uu.nl/>).⁴

All participants underwent a standardized MR exam on a 3 T Philips Achieva MR scanner using an eight-channel head coil. The exam included, amongst others, a 3D T1-weighted turbo field echo sequence (TR: 7.9 ms, TE: 4.5 ms) and a multi-slice fluid-attenuated inversion recovery (FLAIR) sequence (TR: 11 000 ms, TE: 125 ms, TI: 2800 ms). See figure 1 for an example. Using the `elastix` toolbox for medical image registration,¹³ the T1 image were aligned with the FLAIR image of each participant. All images were manually segmented, including the normal brain tissue and WMH, using a freehand spline drawing tool based on MeVisLab.^{14,15} The five participants with the most WMH volume were used for testing; the remaining fifteen participants were used to train the classifier.

2.2 Procedure

We trained a k NN classifier on the normal brain anatomy of the fifteen training participants. An example simulated k NN feature space, with just the two intensity features, is shown in Figure 2a.

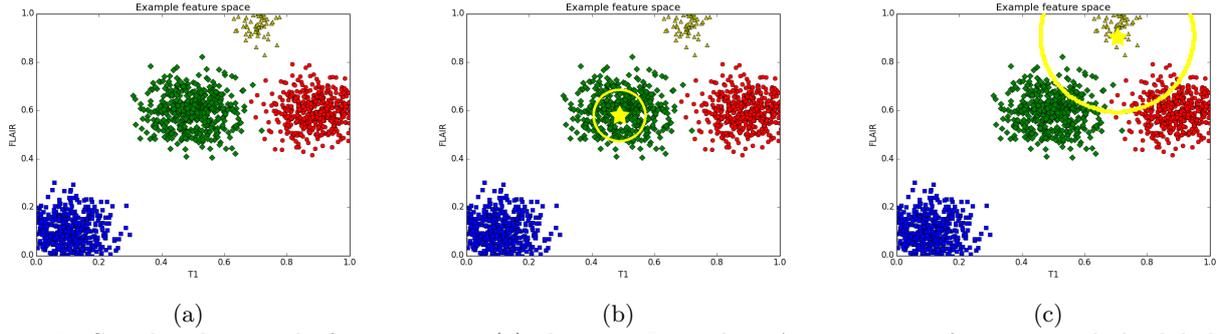


Figure 2: Simulated example feature space (a) showing T1 and FLAIR intensity features, and the labels for CSF (blue squares), GM (green diamonds), WM (red circles), and WMH (yellow triangles). A k NN classifier is trained on CSF+GM+WM and not on WMH, which is the novelty. When classifying an unknown voxel (yellow star) whose true label is GM (b), the k nearest neighbours are within a certain radius (yellow circle). When classifying an unknown voxel whose true label is WMH (c), the k nearest neighbours in the trained data are further away. The required search radius for trained classes can be learned and novelties can then be identified because of an exceptionally large search radius.

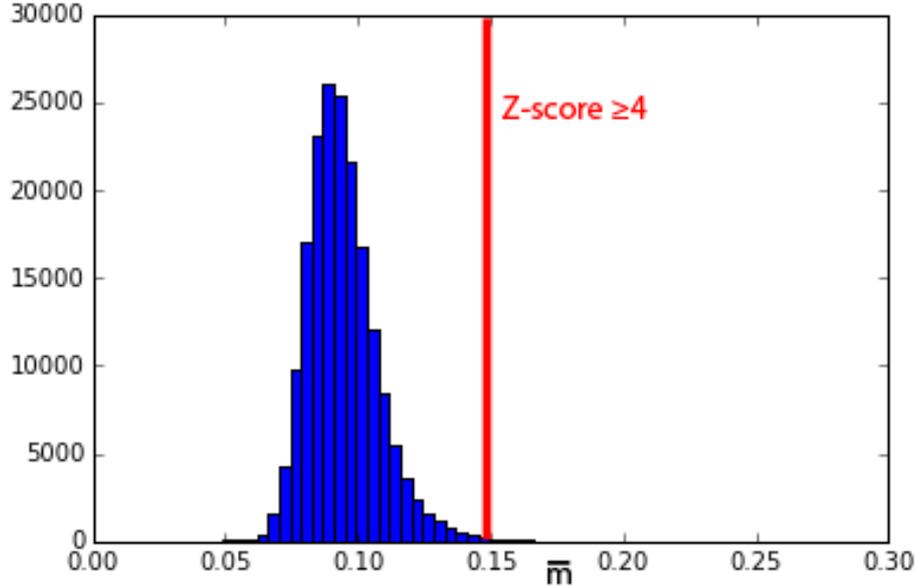


Figure 3: Learned distribution of \bar{m} for normal WM. If an unknown voxel is labelled as WM, its \bar{m} is compared to this distribution. If the Z-score ≥ 4 , the unknown voxel is considered a novelty.

The classifier searches for the k nearest neighbours, reports their labels, and from that derives a prediction for the test voxel. When the true label of a test voxel is amongst the training data, the search radius in feature space to the k neighbours is much smaller than when the test voxel is a novelty (e.g. a WMH, see Figure 2).

Hence, the behaviour of the k NN classifier (modelled by the search radius) might be used for novelty detection. We inspect three different measures M : the mean (\bar{m}), median (\tilde{m}), and Q1 (mean in the first quartile) distance to the k nearest neighbours.

The five test cases are used in a leave-one-out cross-validation, where M is learned on four cases and next applied on the remaining case. Each normal voxel in the four cases (i.e. each non-WMH voxel) is classified by the k NN classifier and the value for M is recorded per true label. This results in a (normal) distribution of M ; for example in Figure 3, where we show the mean (\bar{m}) search radius for WM.

Next, all voxels in the remaining test case are classified. For each voxel and its predicted label, we compare

Table 1: The percentage of identified WMH volumes having a Z-score ≥ 4 .

M	Identified WMH volume
\bar{m}	91.03 %
\tilde{m}	90.64 %
Q1	91.64 %

the behaviour of the classifier as measured by M to the learned distribution of M . For example, if a test-voxel is assigned the label WM, we compare the mean (\bar{m}) search radius against the distribution of \bar{m} for true WM voxels (given in Figure 3). This comparison is done by computing the Z-score: $z = \frac{X-\mu}{\sigma}$, where X is M for the test-voxel, μ is the average M of its predicted label and σ the corresponding standard deviation.

The Z-score is the number of standard deviations a sample is away from its mean. A Z-score ≥ 4 indicates abnormal behaviour of the classifier, having a probability due to chance of 0.000032. So if a voxel has a Z-score ≥ 4 for the behaviour of the classifier measured by M , we have detected a novelty.

2.3 Implementation

2.3.1 Features

The T1 and FLAIR image intensities were rescaled to $[0, 1]$, with the 0.5th and 99.5th percentile of the histogram as lower/upper bound, respectively. Spatial information was obtained by transforming the MNI152 atlas to the image with an affine registration, using elastix^{13,16}. Features for the classifier included: T1 and FLAIR intensities, MNI x/y/z spatial location normalized to $[0, 1]$, T1 and FLAIR smoothed with a 2D Gaussian kernel of $\sigma = 1.0$ mm.

2.3.2 Classifier

We used the k NN classifier implemented in scikit-learn,¹⁷ with $k = 50$ and all other settings were kept default. For the fifteen training images, 10 % of the normal tissue was randomly selected for training.

2.4 Evaluation

In a leave-one-out fashion, the label-specific values for M were measured from four participants. The remaining participant was classified and a Z-score map was computed. See Figure 4 for an example. Next, we recorded the volume of WMH that was identified on the Z-score map (i.e. individual WMHs having a cluster of at least five voxels with a Z-score ≥ 4). Finally, we visually examined other clusters in the Z-score map ≥ 4 , to identify other possible novelties.

3. RESULTS

The identified WMH volume percentage is given in Table 1. Overall, $>90\%$ of the WMHs had a Z-score ≥ 4 after classification by the classifier trained on normal tissue.

Besides WMH, other regions with a Z-score ≥ 4 were identified. Most errors occurred on the edge of the used brain mask, where non-brain structures were included in the mask, either fully or as partial volume voxels. Next, some normal anatomy that was not present in the training data (given the used features) was identified, including: vessels in the sulci, vessels outside the brain tissue (e.g. venous sinuses), the choroid plexus, cerebral falx, tentorium cerebelli, and calcium or increased iron contents in the basal ganglia. Some motion artefacts were identified. For one subject, the applied image intensity normalization was not successful. The normal appearing white matter had different intensity values compared to the training data, resulting in some voxels having a Z-score ≥ 4 . An example Z-score map is shown in Figure 4.

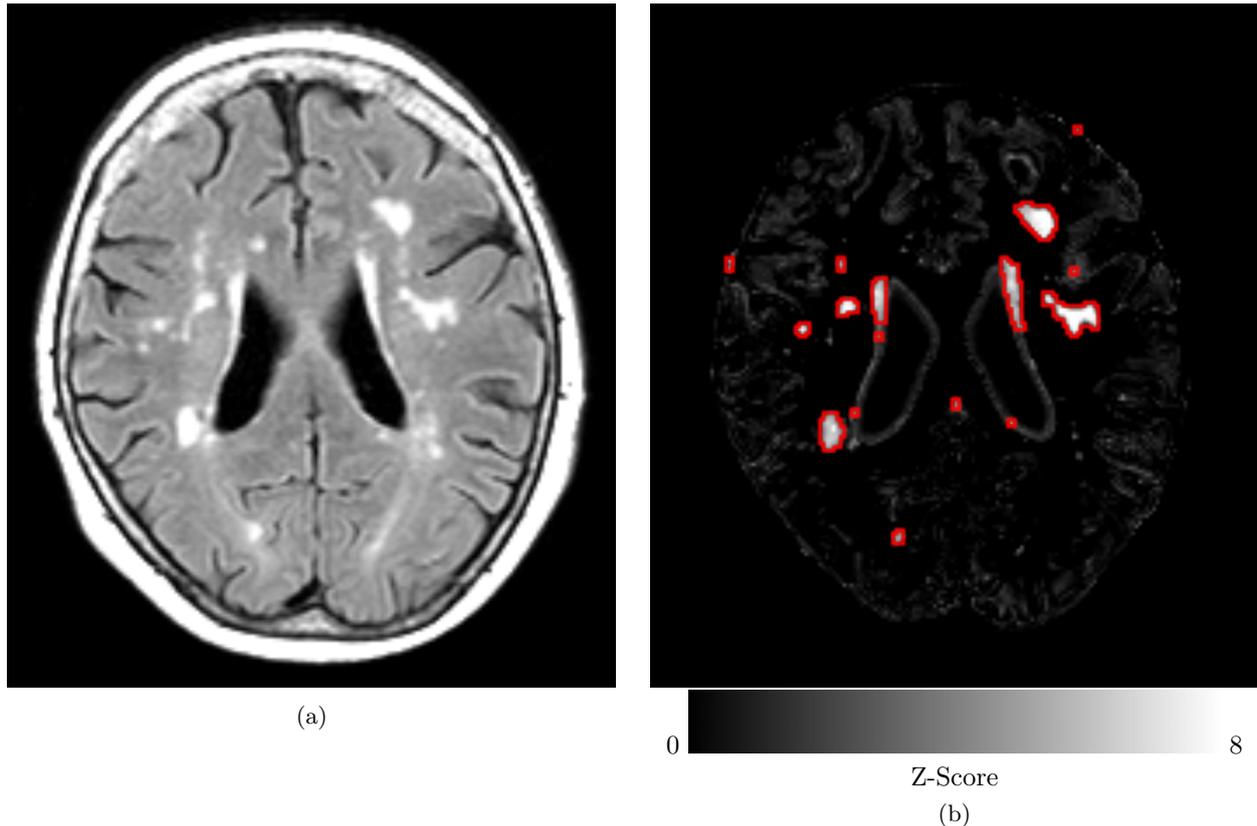


Figure 4: left a FLAIR image, right the Z-score map with regions ≥ 4 outlined in red. In the Z-score map, the non-outlined WMH in the right-bottom-corner was identified on a subsequent slice. Additional findings at the border of the brain indicate an oversegmentation of the used brain mask.

4. DISCUSSION

The presented method is able to identify $>90\%$ of the WMH volume in brain MR images, with a k NN classifier that is not trained on WMH. To our knowledge, this is one of the first applications of novelty detection in medical image analysis. This approach circumvents one of the largest issues in WMH segmentation: the underrepresentation of WMH examples in the training data.

A limitation of the presented approach is that novelty detection is aspecific. It detects all locations in an image that have a Z-score ≥ 4 , which are not necessarily WMH. In our evaluation, we chose to measure the recovered WMH volume and provide a description of the other findings in the images. For a fully automated application of this method, additional image analysis is needed to identify the WMHs amongst the novelties and censor the other findings.

Another limitation is the additional training stage, to create the distribution of M on normal data. This increases the required training time significantly. However, this can be performed offline and does not influence the computation time for test images.

The potential applications of the proposed method are diverse. Besides detecting WMH, the method can be employed to detect various abnormalities in brain MR images, e.g. tumours or infarctions. Future work will explore these applications. Furthermore, novelties can be classified or clustered into known pathologies based on characteristics derived from dedicated training data or knowledge from literature.

5. CONCLUSIONS

Computing Z-Score maps of classifier behaviour helps to identify unexpected findings in brain MR images. This work presents a supervised way of novelty detection, where classifier behaviour on normal tissue is measured

from training data. For a k NN classifier, measures derived from the required search radius and distances to the k nearest neighbours have proven to be effective in this study. The proposed technique and measures seem to be generalizable to other classifiers, hence enabling a broad range of applications.

6. ACKNOWLEDGMENTS

This study was financially supported by the project Brainbox (Quantitative analysis of MR brain images for cerebrovascular disease management), funded by the Netherlands Organisation for Health Research and Development (ZonMw) in the framework of the research programme IMDI (Innovative Medical Devices Initiative); project 104002002. This work was also supported by a Vidi grant from ZonMw, The Netherlands Organisation for Health Research and Development [91711384], to GJB.

REFERENCES

- [1] Cocosco, C. A., Zijdenbos, A. P., and Evans, A. C., “A fully automatic and robust brain MRI tissue classification method.” *Medical Image Analysis* **7**, 513–527 (dec 2003).
- [2] Anbeek, P., Vincken, K. L., van Bochove, G. S., van Osch, M. J. P., and van der Grond, J., “Probabilistic segmentation of brain tissue in MR imaging.” *NeuroImage* **27**, 795–804 (oct 2005).
- [3] Vrooman, H. A., Cocosco, C. A., van der Lijn, F., Stokking, R., Ikram, M. A., Vernooij, M. W., Breteler, M. M. B., and Niessen, W. J., “Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification.” *NeuroImage* **37**, 71–81 (aug 2007).
- [4] Mendrik, A., Vincken, K., Kuijf, H., Breeuwer, M., Bouvy, W., de Bresser, J., Alansary, A., de Bruijne, M., Carass, A., El-Baz, A., Jog, A., Katyal, R., Khan, A., van der Lijn, F., Mahmood, Q., Mukherjee, R., Van Opbroek, A., Paneri, S., Pereira, S., Persson, M., Rajchl, M., Sarikaya, D., Smedby, Ö., Silva, C., Vrooman, H., Vyas, S., Wang, C., Zhao, L., Biessels, G. G., Viergever, M., Smedby, O., Silva, C., Vrooman, H., Vyas, S., Wang, C., Zhao, L., Biessels, G. G., and Viergever, M., “MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans,” *Computational Intelligence and Neuroscience* **2015** (2015).
- [5] Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R. I., O’Brien, J. T., Barkhof, F., Benavente, O. R., Black, S. E., Brayne, C., Breteler, M., Chabriat, H., Decarli, C., de Leeuw, F.-E., Doubal, F., Duering, M., Fox, N. C., Greenberg, S., Hachinski, V., Kilimann, I., Mok, V., van Oostenbrugge, R., Pantoni, L., Speck, O., Stephan, B. C. M., Teipel, S., Viswanathan, A., Werring, D., Chen, C., Smith, C., van Buchem, M., Norrving, B., Gorelick, P. B., and Dichgans, M., “Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration.” *The Lancet. Neurology* **12**, 822–38 (aug 2013).
- [6] Anbeek, P., Vincken, K. L., van Osch, M. J. P., Bisschops, R. H. C., and van der Grond, J., “Probabilistic segmentation of white matter lesions in MR imaging.” *NeuroImage* **21**, 1037–44 (mar 2004).
- [7] de Boer, R., Vrooman, H. A., van der Lijn, F., Vernooij, M. W., Ikram, M. A., van der Lugt, A., Breteler, M. M. B., and Niessen, W. J., “White matter lesion extension to automatic brain tissue segmentation on MRI.” *NeuroImage* **45**, 1151–61 (may 2009).
- [8] Steenwijk, M. D., Pouwels, P. J., Daams, M., van Dalen, J. W., Caan, M. W., Richard, E., Barkhof, F., and Vrenken, H., “Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs),” *NeuroImage: Clinical* **3**, 462–469 (jan 2013).
- [9] Chawla, N. V., Japkowicz, N., and Kotcz, A., “Editorial: Special Issue on Learning from Imbalanced Data Sets,” *ACM SIGKDD Explorations Newsletter* **6**, 1 (jun 2004).
- [10] Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L., “A review of novelty detection,” *Signal Processing* **99**, 215–249 (jun 2014).
- [11] Chandola, V., Banerjee, A., and Kumar, V., “Anomaly detection: A survey,” *ACM Computing Surveys (CSUR)* **41**(September), 1–58 (2009).
- [12] Reijmer, Y. D., Brundel, M., de Bresser, J., Kappelle, L. J., Leemans, A., and Biessels, G. J., “Microstructural white matter abnormalities and cognitive functioning in type 2 diabetes: a diffusion tensor imaging study.” *Diabetes care* **36**, 137–44 (jan 2013).

- [13] Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. W., “Elastix: a Toolbox for Intensity-Based Medical Image Registration,” *IEEE transactions on medical imaging* **29**, 196–205 (jan 2010).
- [14] Ritter, F., Boskamp, T., Homeyer, A., Laue, H., Schwier, M., Link, F., and Peitgen, H.-O., “Medical Image Analysis,” *IEEE Pulse* **2**, 60–70 (nov 2011).
- [15] Kuijf, H. J., *Image processing techniques for quantification and assessment of brain MRI*, PhD thesis, Utrecht University (2013).
- [16] Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., and Collins, D. L., “Unbiased average age-appropriate atlases for pediatric studies,” *NeuroImage* **54**, 313–27 (jan 2011).
- [17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É., “Scikit-learn: Machine Learning in Python,” *The Journal of Machine Learning Research* **12**, 2825–2830 (feb 2011).