

A single-layer network unsupervised feature learning method for white matter hyperintensity segmentation

Koen Vijverberg^{1,2}, Mohsen Ghafoorian^{1,2}, Inge W.M. van Uden³, Frank-Erik de Leeuw³,
Bram Platel² and Tom Heskes¹

¹*Radboud University, Institute for Computing and Information Sciences, Nijmegen, the Netherlands*

²*Radboud University Medical Center, Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Nijmegen, the Netherlands*

³*Radboud University Medical Center, Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Nijmegen, the Netherlands*

ABSTRACT

Cerebral small vessel disease (SVD) is a disorder frequently found among the old people and is associated with deterioration in cognitive performance, parkinsonism, motor and mood impairments. White matter hyperintensities (WMH) as well as lacunes, microbleeds and subcortical brain atrophy are part of the spectrum of image findings, related to SVD. Accurate segmentation of WMHs is important for prognosis and diagnosis of multiple neurological disorders such as MS and SVD. Almost all of the published (semi-)automated WMH detection models employ multiple complex hand-crafted features, which require in-depth domain knowledge. In this paper we propose to apply a single-layer network unsupervised feature learning (USFL) method to avoid hand-crafted features, but rather to automatically learn a more efficient set of features. Experimental results show that a computer aided detection system with a USFL system outperforms a hand-crafted approach. Moreover, since the two feature sets have complementary properties, a hybrid system that makes use of both hand-crafted and unsupervised learned features, shows a significant performance boost compared to each system separately, getting close to the performance of an independent human expert.

1. INTRODUCTION

Studies show that cerebral small vessel disease (SVD) is correlated with several problems, such as: cognitive decline, dementia, Parkinsonism and motor- and mood impairments^{2,3}. SVD presents with a spectrum of findings on images of the brain that includes: white matter hyperintensities (WMH), lacunes, microbleeds and brain atrophy⁴. Correctly identifying WMHs is valuable for the prognosis and understanding of etiology and progression of SVD.

Clinically the presence of WMHs is sometimes very coarsely scored by radiologists. However accurate segmentation is done in research settings either by neurology experts, or (semi-)automatically using classical machine learning techniques employing complex, hand-crafted features⁵. Manual annotation has several drawbacks: It is usually very time consuming that makes it infeasible in practice for large datasets, and is subject to significant inter- and intra-rater variability. Classical machine learning techniques are often faster than a human in segmentation but encompass complex features, which require in-depth insight into the domain. Put another way, in the traditional machine learning algorithms, the overall performance strictly depends on careful domain-dependent choice of features.

Earlier studies on automated detection of SVD can be divided into two major categories⁵: supervised and unsupervised approaches. Supervised approaches include systems based on k-nearest neighbors²⁰, support vector machines²¹, Bayesian methods^{22,23}, etc. Unsupervised approaches include: intensity threshold based segmentation²⁴, adaptive threshold based segmentation²⁵, and more recent studies which also include spatial features and neighborhood information²⁶.

Alternatively it is possible to learn an appropriate set of features with an unsupervised feature learning (USFL) framework, in which, as the name suggests, the algorithm independently learns specific features that optimally represent the data. Different USFL methods have been widely proposed and used in the past few years, resulting in excellent performance on various well-known datasets¹⁸. However methods that automatically learn the features for the WMH segmentation task are rare²⁹.

In this paper, we propose to use k-means for feature learning and we use the learned features in a sliding window manner to segment the WMHs. We also compare the acquired computer aided detection (CAD) system to a similar model but trained on hand-crafted features. Finally, a hybrid model trained on both the hand-crafted and learned features is presented.

2. MATERIALS AND METHODS

2.1 Data

Data used in this paper is provided by the RUN DMC study¹, which investigates the relation between SVD and cognitive and motor status deterioration. The dataset consists of 503 MRI scans of individuals diagnosed with SVD. MR images together with other motor and cognitive tests have been acquired.

Scans were conducted using a 1.5 Tesla Magnetom Sonata MRI scanner by Siemens AG. The process for each patient included taking three separate MR scans: T1, T2* and FLAIR. In our research we use only the information provided by the FLAIR scans, because WMHs are most clearly –but not exclusively- visible on this imaging modality. The FLAIR scans used in this study were made axially, at a resolution of 1.2×1.0×5.0 mm with an inter-slice gap of 1.0 mm.

WMHs were manually annotated on the FLAIR scans by a trained neurology resident. 50 cases were also annotated by another expert to check if they consistently marked the same parts as a WMH. From the total population of patients, 300 were selected for training, and 37 patients for testing. The 37 test cases were randomly taken from the subset of double annotated cases.

2.2 Unsupervised feature learning

Within the USFL framework, we automatically learn a number of discriminative features from a set of images, then we use those features to classify voxels in a given image. There are many variations of USFL algorithms¹⁸. The one we utilize and describe in this paper is a simple but commonly used algorithm that has shown good performance in different applications^{9,12}. There are several deviations from the common method to accommodate for the specifics of our dataset. The USFL algorithm consists of two main steps: 1) learning a set of optimal dictionary terms from the data and 2) applying the learned dictionary terms to compute a new set of features for a given sample. In this research we will use only a single-layer network to transform the features. Research by Coates et al., 2011⁹ shows that the number of dictionary terms is more important than the number of layers for prediction power.

2.2.1 Learning the dictionary terms

Learning the feature mapping can be done in multiple ways. For example using sparse auto-encoders¹³, sparse restricted Boltzmann machines¹⁴, k -means clustering¹⁵ and Gaussian mixtures¹⁶. In this research k -means clustering is used, which is easy to implement and gives good results⁹.

For k -means there is only one parameter to be chosen: k as the number of centroids. A greater k will usually improve classification results⁹, but at a higher computational cost. As input for the k -means algorithm, we take patches of 5×5 voxels, flattened to a single vector per patch. The patches are taken for every voxel marked by the readers as WMHs and an equal number of random voxels from the normal tissues of the brain. Subsequently the k -means algorithm is performed to obtain a clustering of the feature vectors in the 25 dimensions space. Centroids of each cluster are then computed and considered as the k learned dictionary terms. Figure 1 shows a visual representation of the learned dictionary terms for our dataset.



Figure 1: The set of learned dictionary terms for $K = 200$

2.2.2 Extracting the feature vectors

To learn the discrimination between WMHs and normal brain tissue, we consider the appearances in a local 16×16 neighborhood of each voxel. Local neighborhoods are sampled the same way as patches were taken for k -means: All of the WMH positive voxels plus an equal number of random voxels are selected. In order to train a classifier we need to map each local neighborhood to a feature vector. The procedure is depicted in Figure 2 and is as follows: First, all possible 5×5 patches (i.e. $s=1$) are sampled and the corresponding feature vector of length k is calculated for each by computing the patches' distance to each learned dictionary term. Then each neighborhood is divided into 4 pools and the average response of all the patches inside a pool to each dictionary term is averaged. This way each pool is represented by a feature vector of length k . Finally concatenating the features obtained from each pool together, a final representation of the whole local neighborhood of length $4 \times k$ is acquired.

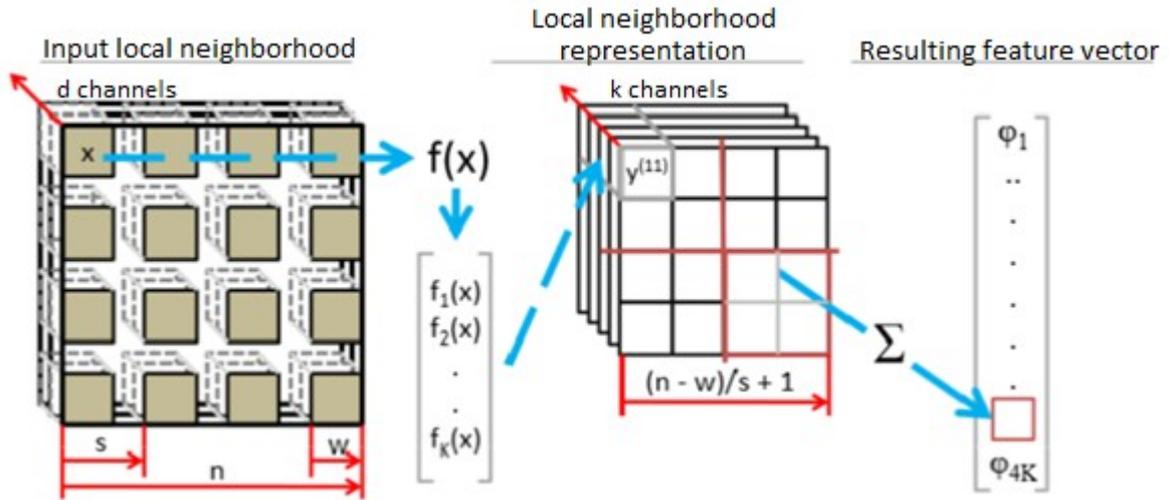


Figure 2: An Illustration of the feature extraction process on an input local neighborhood⁹, where n and w are sizes of local neighborhood and patches respectively, s represents the stride for patch sampling, d is the number input channels, k is number of dictionary terms in k-means and f stands for the transformation function from patches to vectors of dictionary term distances.

2.3 Location features

Previous studies on detection of WMHs^{17, 5} show that spatial location information is very important for an accurate detection. This means that a CAD system that perfectly captures the local appearance, still does not perform optimally. For this reason we augment the automatically learned features with a set of eight spatial location features. The location features used are: x -, y - and z -coordinates of the corresponding voxel in MNI¹⁹ space, location-based WMH prior probability, and a set of Euclidian distances to several landmarks in the brain, which are distances to brain cortex, left and right ventricles and midsagittal brain surface. WMH prior probability is the likelihood of occurrence of hyperintensities on the corresponding MNI atlas location that is obtained from the large population of 503 patients from the RUN DMC dataset.

2.4 Hand-crafted features

Features in an earlier research¹⁷, engineered and optimized on the same dataset, are used as a reference material to benchmark our USFL algorithm. The set of hand-crafted features consists of 22 features in total that are: A group of intensity features including FLAIR and T1 intensities, second order derivative features including multiscale Laplacian of Gaussian ($t=1,2,4$ mm), multiscale determinant of Hessian ($t=1,2,4$ mm), vesselness filter ($\sigma=1$ mm), a multiscale annular filter ($t=1,2,4$ mm), FLAIR intensity mean and standard deviation in a 16×16 neighborhood, as well as the 8 location features as described in the previous subsection.

2.5 Model training

Once the two datasets for the automatically learned and hand-crafted features are created, we train a random forest classifier with 50 subtrees, with a cross-validated max-depth to prevent overtraining. We ran feature learning our algorithm twice for $k=100$ and 200 , with the location features augmented in both cases. Exactly the same classifier was used to train on the hand-crafted features. We also combined the two feature sets to check if the features from the two groups could possibly complement each other and result in a boosted performance of a hybrid system.

2.6 Segmentation

For the segmentation of the test images, we consider local neighborhoods for all voxels that are to be classified in a sliding window fashion. For each local neighborhood, we compute the corresponding feature vectors and assign the central voxel with a likelihood resulted by the model. We then threshold the resulting likelihood map on all test images with a unique optimized threshold for the Dice score, to obtain binary segmentation masks.

3. RESULTS

We evaluate and compare the performance of the proposed algorithm based on the Dice coefficient, which is the most widely used evaluation for WMH segmentation task⁵. Table 1 demonstrates and compares the performance of the different algorithms and the human experts, considering expert 1, expert 2 and the logical OR of the two experts as the reference standards. We also perform a statistical significance test, by means of comparing the performances of different methods on 100 bootstraps of test cases and reporting the resulting p -values. Results are presented in table 2. Figure 3 shows a comparison of ROC curves for different methods at low false positive

rates, where the highest agreement between the segmentations of the system and the reference standard is achieved. Figure 4 displays segmentation for three cases.

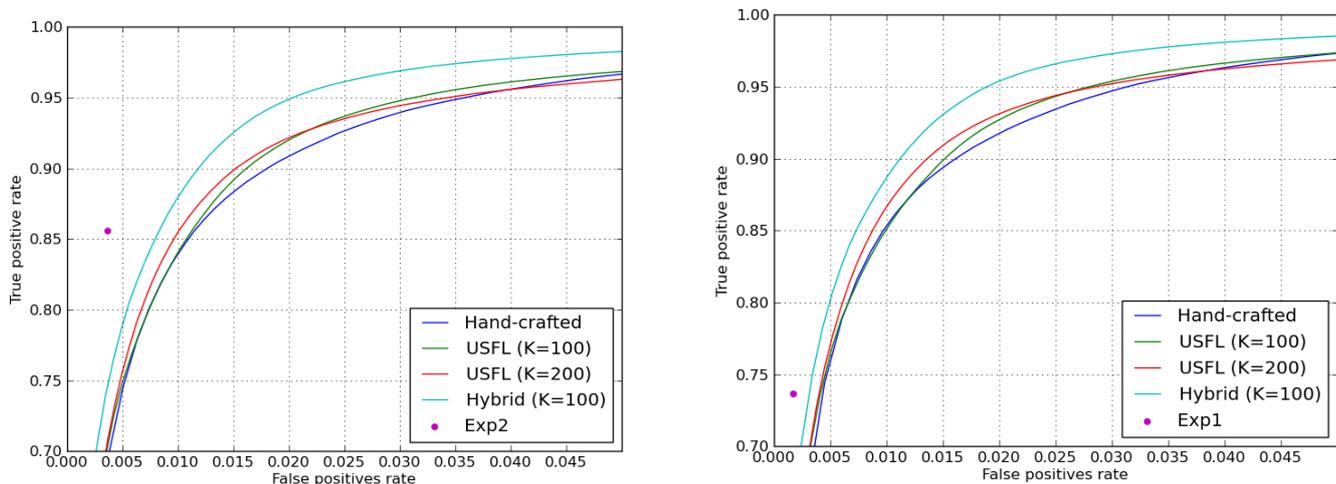


Figure 3: ROC curves comparison for different methods using expert 1 (left). and expert 2 (right) as the reference standard.

Table 1: Unsupervised feature learning and handcrafted feature results

Reference standard	Hand-crafted	USFL (k = 100)	USFL (k = 200)	Hybrid (k=100)	Exp1	Exp2	4.
Dice (exp1)	0.709	0.718	0.718	0.744	-	0.792	
Dice (exp2)	0.702	0.711	0.712	0.737	0.792	-	
Dice (exp1 exp2)	0.718	0.724	0.727	0.752	-	-	

Table 2: P-values for statistical significance comparison of different methods. p_{ij} represents the probability for the null hypothesis that the method in row i is better than the method in column j .

Method	USFL (K = 100)	USFL (K = 200)	Hybrid (k=100)
Hand-crafted	0.20	0.14	<0.01
USFL (K=100)	-	0.15	<0.01
USFL (K=200)	-	-	<0.01

4. DISCUSSION AND CONCLUSION

Considering the experimental results presented in Table 1 and Figure 2, the automatically learned features, even though do not require any knowledge from domain experts, appear to be comparably reliable for the segmentation of WMHs. Referring to Table 2, a significantly better performing hybrid CAD system can be achieved by combining the two sets of features that are apparently complementary to each other.

The number of dictionary terms can be indicative in the performance of the k-means USFL algorithm; a larger set of learned dictionary terms, leads to a slightly better discrimination between WMHs and normal appearing brain tissue. This is in accordance with results obtained in other studies⁹. The better performance is obviously achieved at a higher computational cost.

Further investigation of the algorithm parameter space is possible by changing the local neighborhood size, which was kept 16×16 voxels in this study. Although larger local neighborhood sizes, will help capture more contextual information, they might result in lower localization accuracy. An optimal value for this could be found considering this trade-off. The dictionary patch size of 5×5 voxels can also be tweaked to possibly obtain a better result. Changing the pooling shape is also another option.

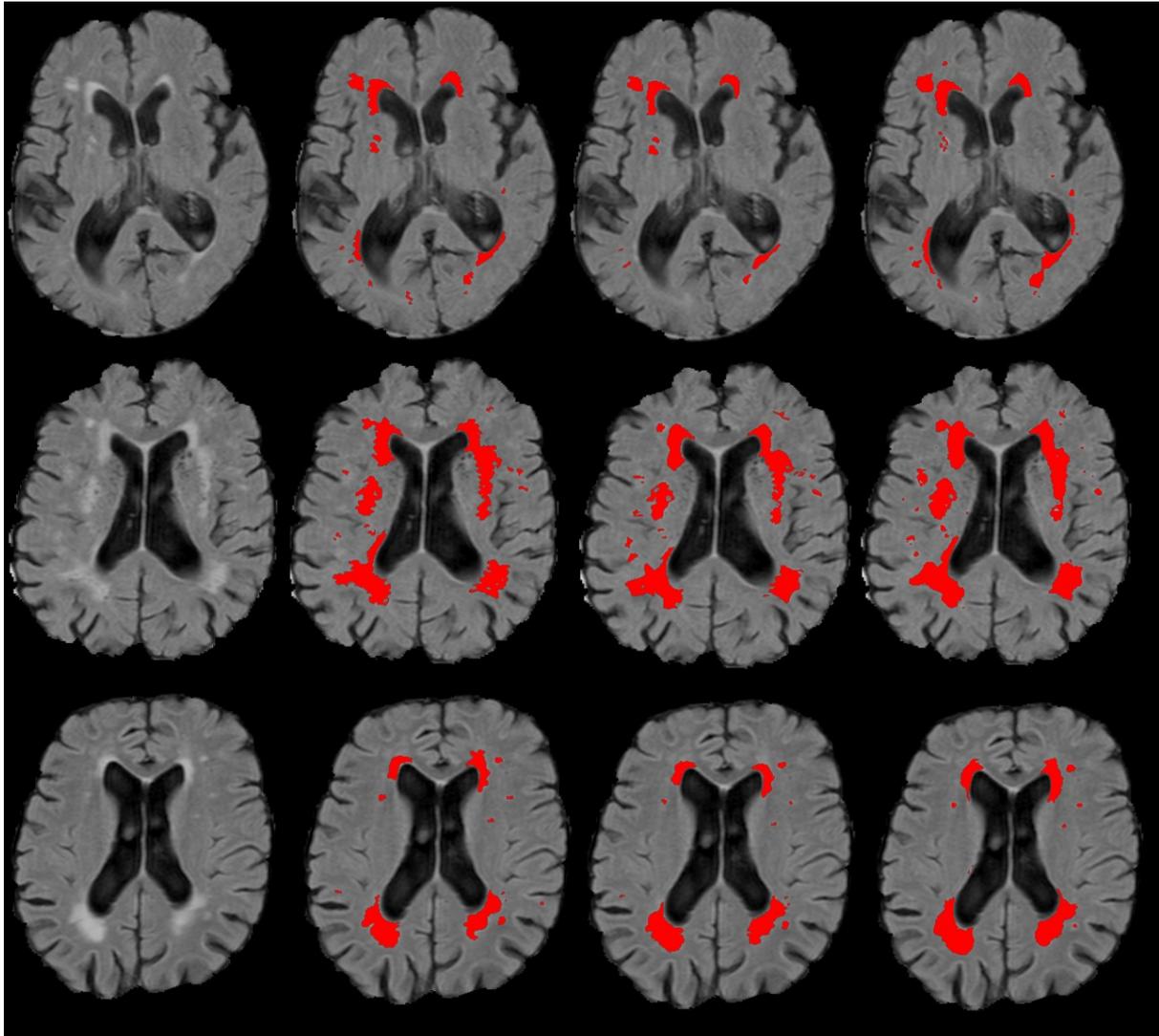


Figure 4: The performance of the hybrid model on three sample slices. The columns represent original FLAIR, annotations of human expert 1, expert 2 and segmentations of the hybrid model ($k=100$) from left to right respectively.

References

- [1] van Norden, A. G., van den Berg, H. A., de Laat, K. F., Gons, R. A., van Dijk, E. J., & de Leeuw, F. E. (2011). Frontal and Temporal Microbleeds Are Related to Cognitive Function The Radboud University Nijmegen Diffusion Tensor and Magnetic Resonance Cohort (RUN DMC) Study. *Stroke*, 42(12), 3382-3386.
- [2] Prins, N. D., & Scheltens, P. (2015). White matter hyperintensities, cognitive impairment and dementia: an update. *Nature Reviews Neurology*, 11(3), 157-165.
- [3] de Laat, K. F., Tuladhar, A. M., van Norden, A. G., Norris, D. G., Zwiers, M. P., & de Leeuw, F. E. (2010). Loss of white matter integrity is associated with gait disorders in cerebral small vessel disease. *Brain*, awq343.
- [4] Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, Dichgans, M. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8), 822-838.
- [5] Caligiuri, M. E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., & Cherubini, A. (2015). Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review. *Neuroinformatics*, 1-16.
- [6] Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2), 143-156.

- [7] Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2), 825-841.
- [8] Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2013). Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8), 1915-1929.
- [9] Coates, A., Ng, A. Y., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *International conference on artificial intelligence and statistics* (pp. 215-223).
- [10] Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, 17(3), 143-155.
- [11] Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on*, 20(1), 45-57.
- [12] Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004, May). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (Vol. 1, No. 1-22, pp. 1-2).
- [13] Coates, A., & Ng, A. Y. (2011). The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 921-928).
- [14] Hinton, G. (2010). A practical guide to training restricted Boltzmann machines. *Momentum*, 9(1), 926.
- [15] Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In *ICML* (Vol. 1, pp. 577-584).
- [16] Lee, D. S. (2005). Effective Gaussian mixture learning for video background subtraction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5), 827-832.
- [17] Ghafoorian, M., Karssemeijer, N., van Uden, I., de Leeuw, F. E., Heskes, T., Marchiori, E., & Platel, B. (2015, March). Small white matter lesion detection in cerebral small vessel disease. In *SPIE Medical Imaging* (pp. 941411-941411). International Society for Optics and Photonics.
- [18] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8), 1798-1828.
- [19] Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., ... & Mazoyer, B. (2001). A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association*, 8(5), 401-430.
- [20] Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.
- [21] Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., ... Davatzikos, C. (2008). Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Academic Radiology*, 15 (3), 300-13. doi:10.1016/j.acra.2007.10.012.
- [22] Herskovits, E. H., Bryan, R. N., & Yang, F. (2008). Automated Bayesian segmentation of microvascular white-matter lesions in the ACCORD-MIND study. *Advances in Medical Sciences*, 53(2), 182-90. doi:10.2478/v10039-008-0039-3
- [23] Maillard, P., Delcroix, N., Crivello, F., Dufouil, C., Gicquel, S., Joliot, M., ... Mazoyer, B. (2008). An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases. *Neuroradiology*, 50 (1), 31-42. doi: 10.1007/s00234-007-0312-3.
- [24] Jack, C. R., Brien, P. C. O., Rettman, D. W., Shiung, M. M., Xu, Y., Muthupillai, R., ... Erickson, B. J. (2001). FLAIR Histogram Segmentation for Measurement of Leukoaraiosis Volume. *Journal of Magnetic Resonance Imaging*, 14 (6), 668-676. doi:10.1002/jmri.10011.
- [25] Admiraal-Behloul, F., van den Heuvel, D. M. J., Olofsen, H., van Osch, M. J. P., van der Grond, J., van Buchem, M., & Reiber, J. H. C. (2005). Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *NeuroImage*, 28(3), 607-17. doi:10.1016/j.neuroimage.2005.06.061.
- [26] Anitha, M., Selvy, P. T., & Palanisamy, V. (2012). WML detection of brain images using fuzzy and possibilistic approach in feature space, *WSEAS TRANSACTIONS on COMPUTERS*, E-ISSN, 2224-2872.
- [27] Shi, L., Wang, D., Liu, S., Pu, Y., Wang, Y., Chu, W. C. W., ... Wang, Y. (2013). Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction. *Journal of Neuroscience Methods*, 213 (1), 138-46. doi:10.1016/j.jneumeth.2012.12.014.
- [28] De Leeuw, F. E., de Groot, J. C., Achten, E., Oudkerk, M., Ramos, L. M. P., Heijboer, R., ... & Breteler, M. M. B. (2001). Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. The Rotterdam Scan Study. *Journal of Neurology, Neurosurgery & Psychiatry*, 70(1), 9-14.
- [29] Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I., de Leeuw, F. E., van Ginneken, B., and Platel, B. "Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation", in: *IEEE International Symposium on Biomedical Imaging*, 2016