



Automatic Brain White Matter Hypertensities Segmentation using Deep Learning Techniques

José A. Viteri¹, Francis R. Loayza²^a, Enrique Pelaez¹^b
and Fabricio Layedra¹

¹Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador

²Facultad de Ingeniería en Mecánica y Ciencias de la Producción, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador

Keywords: Convolutional Neural Network, U-Net, WMH Segmentation.

Abstract: White Matter Hyperintensities (WMH) are lesions observed in the brain as bright regions in Fluid Attenuated Inversion Recovery (FLAIR) images from Magnetic Resonance Imaging (MRI). Its presence is related to conditions such as aging, small vessel diseases, stroke, depression, and neurodegenerative diseases. Currently, WMH detection is done by specialized radiologists. However, deep learning techniques can learn the patterns from images and later recognize this kind of lesions automatically. This team participated in the MICCAI WMH segmentation challenge, which was released in 2017. A dataset of 60 pairs of human MRI images was provided by the contest, which consisted of T1, FLAIR and ground-truth images per subject. For segmenting the images a 21 layer Convolutional Neural Network-CNN with U-Net architecture was implemented. For validating the model, the contest reserved 110 additional images, which were used to test this method's accuracy. Results showed an average of 78% accuracy and lesion recall, 74% of lesion f1, 6.24mm of Hausdorff distance, and 28% of absolute percentage difference. In general, the algorithm performance showed promising results, with the validation images not used for training. This work could lead other research teams to push the state of the art in WMH images segmentation.

1 INTRODUCTION


White Matter Hyperintensities (WMHs) are lesions observed in the brain, which stand out as areas of increased brightness when commonly observed as signal hyperintensity on FLAIR (Fluid Attenuation Inversion Recovery) sequences of Magnetic Resonance Imaging (MRI). The WMH lesions are presumed to be of vascular origin and have been associated with cognitive impairment, risk of stroke, dementia, and geriatric disorders (Breteler et al., 1994). Studying the WMHs lesions on these types of images, through a correct and precise segmentation process, would provide the means for improving the understanding of the brain damage and the associated cognitive and physical problems and the supporting benchmarks for diagnosing in early stages of the disease.


Recent research (Giorgio and De Stefano, 2013) has shown the importance of quantifying the WMH,

especially when analyzing diseases related to neurovascular and neurodegenerative disorders. The importance lies in the diagnosis, progression, and treatment monitoring of the neurological conditions, and it correlates with different WMH features.

Image analysis plays an essential role during clinical diagnosis. Recent research shows that image segmentation used to study the brain structure revealed promising results, particularly to follow-up patients or visualizing tissue abnormalities and tumors (Daliri, 2012). These results allow tracking relevant features of the segments; such as changes in volume, shape, or distribution of the abnormalities during patients' follow-up.

The MICCAI WMH Segmentation Challenge¹ was created to directly compare automatic segmentation techniques for the White Matter Hyperintensities (WMH). Since its launch, several methods have pushed the models' performance based on Convolutional Neural Networks-CNN, in particular the U-Net

^a <https://orcid.org/0000-0002-6283-3679>

^b <https://orcid.org/0000-0001-9355-5440>

¹<https://wmh.isi.uu.nl/>

architecture (Ronneberger et al., 2015a).

In this work, we propose a model based on a Fully CNN architecture, tailored for segmentation. This paper is organized as follows: Section 2 describes the related work about segmentation procedures using various techniques. Section 3 presents the methodology proposed in this research. Section 4 discusses the results and findings, and Section 5 contains some conclusions and future work.

1.1 Related Work

Segmentation techniques of brain images, such as the Hidden Markov Random Fields (Zhang et al., 2001), or through Probabilistic Methods (Ashburner and Friston, 2005), or K-Nearest Neighbors-KNN (Cocosco et al., 2003; Vrooman et al., 2007), are mostly related to understanding the main brain structures, like the gray and white matter, cerebrospinal fluid, and the surrounding tissues. However, the obtained results showed a need for automatic WMH segmentation; several techniques based on thresholds have been proposed with modest results, most of them still waiting for clinical trials (Chancay et al., 2015; Zijdenbos and Dawant, 1994). Contemporary advanced techniques using artificial intelligence for pattern recognition have been proposed in numerous studies (Li et al., 2018c; Jin et al., 2018; Li et al., 2018a; Xu et al., 2017b); these techniques use different deep learning architectures; such as CNN models, and pattern recognition based on texture classification (Bento et al., 2017). Several of these techniques have been mainly derived from the MICCAI challenge (Berseth, 2017). However, the accuracy of the segmentation, including detecting false positives or true negatives, is still the most significant challenge. Even though most of these deep learning techniques based on similar approaches, the pre-processing procedures, hyper-parameter calibration, and optimization techniques applied to the models' basic architecture provide outstanding segmentation results. Therefore, in this work, we propose a revised architecture based on a Fully Convolutional Neural Network for segmentation (Long et al., 2014) to push the current performance of the models considered state of the art.

Segmenting images to localized WMH lesions have been tackled through several machine learning approaches, and the analysis of FLAIR images is a common technique used for this kind of segmentation. Jack, C. et al. (Jack et al., 2001) shows that segmentation can be performed by analyzing the FLAIR hyperintensities histograms and establishing an intensity threshold. Morel B. et al. (Morel et al., 2016), used morphological operators to segment the WMH le-

sions on Transverse Relaxation Time, or T2 brain images. Ghafoorian, M. (Ghafoorian et al., 2017a) have proposed the use of CNN combined with anatomical location data, and more recent techniques, proposed by the same authors (Ghafoorian et al., 2017b), made use of transfer learning with a personalized top segment, based on a dense convolutional architecture as the output. Xu Y. et al. (Xu et al., 2017a) also proposed to use transfer learning from the Visual Geometry Group-VGG architecture, pre-trained from the ImageNet dataset, with a dense convolutional network for segmenting 3D brain images.

In this work, we propose an architecture based on a fully connected convolutional network, taking advantage of one of its characteristics, the shifting invariance, aimed at preserving the spatial relationships of relevant patterns, such as lesions, need to be propagated deep and up to the output layer.

The architecture of the model, as shown in the next section, uses an end-to-end technique for segmenting T1 and FLAIR images sequences. The segmentation process takes about 12 seconds to complete, and this proposed architecture reached the ninth place at the MICCAI WMH Challenge up to November 2020.

2 METHODOLOGY

The methodology was divided into three main phases. The first was data preparation and pre-processing; the second was data modeling using deep learning techniques; and, the third was evaluating the trained model, which was performed locally by the contest organizers. All this work was developed with python as the programming language (Van Rossum and Drake Jr, 1995). For data preparation and pre-processing, the following libraries were used: nipy (Gorgolewski et al., 2016), numpy (Oliphant, 2006), scipy (Virtanen et al., 2020) and simpleITK (Lowekamp et al., 2013). Keras (Chollet et al., 2015), with tensorflow (Abadi et al., 2016) as engine, was used for the deep learning model. Additionally, pandas (McKinney et al., 2010) and seaborn (Waskom et al., 2017) were utilized during the data evaluation step. Data preparation, validation, and evaluation of the model were primarily performed in a computer with Ubuntu 18, 16 GB of RAM, and an Nvidia 1060 GPU with 6GB of GDDR5 memory. For more demanding tasks, a Microsoft Azure virtual machine instance was used. The instance used was a Standard_NC6 Ubuntu 18, 56 GB of RAM, an Nvidia Tesla K80 with 12 GB of GPU memory (Microsoft, 2020).

2.1 The Dataset

The images used for training and validation came from a dataset provided by the MICCAI WMH Challenge (Kuijf et al., 2019), which consisted of images of 60 subjects, acquired from three different 3T scanners and places: Amsterdam (AMS GE3T), Utrecht, and Singapore. Each scanner contributed with a set of 20 pairs of images per subject. Fig. 1, shows a sample of the images.

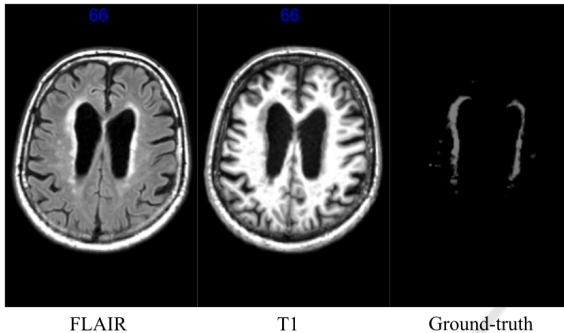


Figure 1: Sample of one slice of the dataset used as input for training. T1-weighted, FLAIR and a manually segmented mask of the WMH (Ground-truth).

It is important to note that scanners and volunteers came from three different hospitals and MRI scanners: two from the Netherlands and one from Singapore. As described in (Kuijf et al., 2019), the parameters' settings of the acquisition images like voxel size, slice number, and echo time were different for each scanner. The images acquired per subject were: 3D-T1-weighted images and 2D-multi-slice FLAIR images. Further, the provided images were pre-processed previously to correct the bias field inhomogeneities, re-sampled, and coregistered between them using SPM12 software. Additionally, the contest provided the ground-truth images obtained from each FLAIR image, which were manually segmented. The segmentation was done by two expert observers, *O1* and *O2*. The process was made following the STandards for ReportIng Vascular changes on nEuroimaging (STRIVE) conventions (Wardlaw et al., 2013) — the observer *O1* segmented all images using a contour drawing technique delineating the outline of each WMH. The second observer *O2* performed a peer review over the manual delineations of *O1*, following a peer project methodology. A detailed description of the manual segmentation can be found in (Kuijf et al., 2019). These images contained binary masks of the WMH lesions, which correspond to the ground-truth.

Additionally, the organizers kept in reserved 110 cases from five different scanners; these images were

not provided to the participants. 30 out of the 110 cases, were from each of the scanners mentioned before, and 20 from two additional scanners (from Amsterdam, but with different characteristics, such as less magnetic field strength). All of these images were also pre-processed using the same procedure described before. The contest reserved these images for testing purposes and metrics calculation.

2.2 Data Preparation and Further Pre-processing

Before training the model, data from all sources were merged into one dataset, consisting of 60 pairs of images: one T1-weighted image and FLAIR per subject. All slices for each image were resized to 200x200 pixels across the y and z axis, using the numpy library in python. Further, we selected a field of view from all images that contained the brain, cropping the volume to discard the neck. We also performed additional pre-processing procedures on all images; such as, a Gaussian normalization to reduce noise, highlight small brightness spots and smooth the images, and a morphological normalization to reduce the black and low-intensity brain regions produced by cerebral atrophy. This morphological normalization was performed with the scipy python library.

2.3 Data Augmentation

To prevent the model from overfitting and to increase the size of the training and testing datasets, we applied some data augmentation strategies with two approaches. The first includes standard operations, like rotation, scaling, and shearing to all images. Each transformation increased up to 60 additional images to the original dataset, increasing from 60 to 240 images for each MRI channel. We also included more complex data augmentation procedures; such as, linear and nonlinear transformations over the original data. For the linear transformations, we used a pointwise product between the FLAIR and T1 images. And, we applied a diffeomorphic transformation for the nonlinear data augmentation, normalizing from the native space to the MNI 152 standard space (Fonov et al., 2011; Fonov et al., 2009), using the nipy library. After these procedures, we created a separate dataset with the linear and nonlinear transformations.

2.4 Network Architecture

The proposed model's architecture was designed to use two types of images as input per subject: T1-

weighted and FLAIR images with the corresponding pre-processed procedures explained before. Additionally, both images need to be coregistered between them, including a field bias inhomogeneities correction.

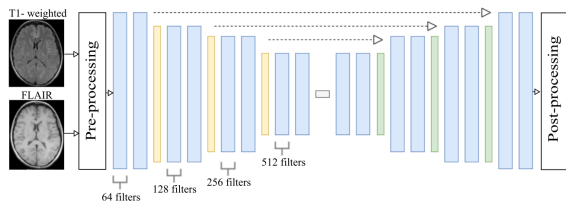


Figure 2: U-Net Neural Network Architecture.

For this research, we designed three different segmentation models based on a fully CNN architecture, as proposed by (Long et al., 2014; Milletari et al., 2016); and, in particular the U-net architecture (Ronneberger et al., 2015b), which had proven to be highly useful in biomedical image segmentation. In this work, the U-Net architecture performed the best, a model with 21 layers, including 15 convolutional layers, three upsampling layers, and three pooling layers. Fig. 2 shows a general representation of this model's architecture. For the first two layers we used 5x5 convolutional filters, while for the rest a set of 3x3 convolutional filters were used. After each convolutional layer, a Rectified Linear Unit (RELU) activation function was applied (Agarap, 2018). The yellow boxes in Fig. 2 represent the max pooling or downsampling procedures, and the green boxes the upsampling operation, both with 2x2 filters. The number of filters in each layer goes from 64 in the two first convolutional layers to 128, 256, and 512 filters in the left side of the U-Net. We use Adam Stochastic Gradient Descent for the learning process of the model (Kingma and Ba, 2014). The learning rate was set to 0.0002. The training was performed with 30 batches, and the parameter's search was performed during 50 epochs. The other two models were designed with a similar configuration, but with some modifications of the hyper-parameters, as discussed in the next section.

Once the models were trained, cross-validated and tested, the models were put into an inference stage, where the architectures were also tested with new T1 and FLAIR images, which were not seen during the previous phases, to let them recognize the WMH lesions, as well as to perform the segmentation in the images.

2.5 Evaluation

The models were evaluated using six metrics as defined by the MICCAI WMH segmentation challenge. Those metrics were: Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), Average Volume Difference (AVD), Sensitivity for detecting individual lesions, (Recall), and the F1-score. The Dice similarity metric measures the overlap between the manual segmentation and the model segmentation. The Hausdorff Distance measures how far two subsets of a metric space are from each other. As used in this challenge, the Hausdorff Distance is modified as to obtain the most robust version using the 95th percentile instead of the maximum 100th percentile distance. The Average Volume Difference metric measures the percentage difference in the volume of the manual segmentation lesions compared to the model segmentation. As for the Recall metric, this measures the ability of a model to find all relevant cases within the data. And, the F1 index is a way to combine the recall with its precision, which is defined as the harmonic mean of both, as defined in (Li et al., 2018b).

The model evaluation was performed in two parts. First, a local testing was made using the three different U-nets architectures, tuning the hyper-parameters to push for the ground-truth masks and for the metrics obtained by the 2017's challenge winner (Li et al., 2018b). Then, once our best model was tuned for the best performance, it was evaluated by the WMH challenge organizers, using their own additional test images, which placed our architecture ninth on the overall challenge up to November 2020.

3 RESULTS AND DISCUSSION

3.1 Local Results

Before configuring our three U-Net based models, we evaluated the *Reference* models, to set the basis for comparing our models. We evaluated the model presented by (Li et al., 2018b), which was taken as reference. Then, the challenge's *Reference* model, which we use it to obtain the metrics as described by the challenge. Based on these baseline architectures, our first proposed model was created based on the U-net architecture as shown in Fig. 2; this first model was called (*U-net#1*), and its architecture was configured to take two channels as input: a FLAIR image and an augmented image obtained by a dot product of the T1 and the FLAIR. This architecture produced low performance as compared to the reference models and did not require further analysis.

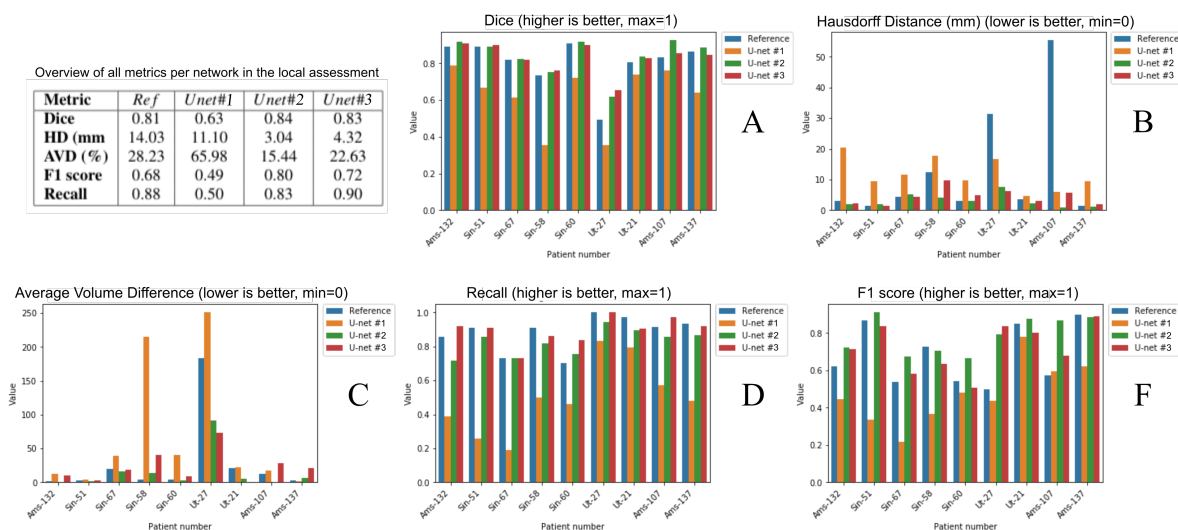


Figure 3: General results of the internal testing.

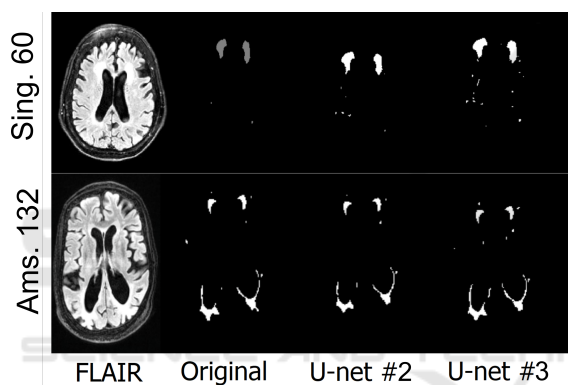


Figure 4: Figure shows a comparison of the same slice of two images between the ground-truth (Original) and automatic segmentation performed by $U-net\#2$ and $U-net\#3$.

Then a second model was configured, which we called ($U-net\#2$), and it was tune-designed to accept our pre-processed FLAIR and T1 images as inputs, which was defined in our pre-processing phase, with promising results. Fig. 3 shows the segments obtained by this models as compared to the original or ground-truth.

A third model then was evaluated, a model called ($U-net\#3$), which was designed to use an Exponential Linear Unit ELU (Clevert et al., 2015), as activation function in the convolutional steps, and the $he-normal$ as kernel initializer (He et al., 2015) in the last step. This model performed better than the reference, but not as good as the model called $U-net\#2$. Fig. 4 also shows the segments obtained by this models as compared to the original or ground-truth.

Therefore, our second model was uploaded to the WMH challenge platform. Figures 3 and 5 show the

results of the testings performed locally and the results assessed by the WMH organizers.

For local testing purposes, eight images were selected randomly and used to compare the resulting segments from the four models. The eight images were obtained from the three scanners proportionally, making sure to have at least two subjects for the input scanner. The metrics used in the local testing were the same as they were defined by the challenge organizers; that is, the Dice, Hausdorff Distance, Average Volume Difference, F1 score, and Recall. The results could be seen in Fig. 3. This figure shows a general summary of all metrics assessed locally in our experiments. The *Ams* prefix are patients from the Amsterdam scanner, *Sin* is from the Singapore scanner, and *Ut* is from the Utrecht scanner. Additionally, the table in Fig. 3, shows the average results for all metrics from each model. Fig. 4 shows the segments from two MRI images: Singapore (subject #60) and Amsterdam (subject #132). The *original* label is the manual segmentation done by experts and provided by the challenge organizers. As it is shown in Fig. 4, our $U-net\#2$ and $U-net\#3$ models obtained comparable segmentation results as to the experts' defined segments.

As it is seen in Fig. 3A, the Dice metric, in our $U-net\#2$ and $\#3$ models, performed better than the *Reference* model. Also, the performance gain was in general better. These metrics were 4.84% and 3%, respectively better than the *Reference* model.

On the other hand, the Hausdorff distance, as seen in Fig. 3B, was significantly better. For example, a subject from the Amsterdam group has a Hausdorff Distance of 65.52 mm in the *Reference* segmentation. While in model $U-net\#2$, such difference reached

0.98 mm, which represents an improvement of 67.10 times over the actual *References*. On average, the Hausdorff Distance between the manual segmentation and the automatic segmentation, obtained in this work, was 4.62 times better than the *Reference*.

The Average Volume Difference is observed in Fig. 3C. In this metric, we obtained more variation. For example, in two patients from the Amsterdam scanner, the *U-net#2* performed better than the rest of the models. The *U-net#3* worked very well on images from the Utrecht scanner. However, in general, both the *U-net#2* and #3 performed better segmenting than the *Reference*. While the average *AVD* of the validation patients in the *Reference* model was 28.23 % in our model *U-net#2* was 15.44 % and 22.63 % in the *U-net#3*. Although, Both methods performed better compared to the *Reference* model.

The F1 score can be seen in Fig. 3F. The average value of the eight validation subjects' images was 0.68, while our score was 0.80. In absolute terms, the *U-net#2* improved 17.3 % and the *U-net#3*, 6.16% as compared to the *Reference*.

Finally, the recall metric can be seen in Fig. 3D. This metric was the only one which did not improve, as compared to the *Reference* models using the *U-net#2* model. In average the value obtained with our model was 0.83, while the *Reference* model was 0.88. However, the *U-net#3* performed better in the recall metric, we obtained 0.90, representing 1.3% improvement as compared to the *reference* model.

3.2 WMH Validations

In this section, we present a summary of the evaluations made by the WMH challenge organizers. The model submitted to evaluation was the *U-net#2*. The evaluation method is based on the rankings of each metric, as described before, using a score from 0 to 1. The best performing team is ranked with 0 and the worst with 1. All other teams were ranked in between relative to their performance within that metric range. Then to compute the final score, the five ranks were averaged in an overall final score, as described in (WMH Segmentation Challenge, 2020). In general, this work was placed ninth out of 43 participant teams, and as reported in November 2020, with a score of 0.0596. Even though the metrics measured locally favored us, we could not improve the *Reference* performance with the organizers' assessment, as observed in the results section of the web site: (<https://wmh.isi.uu.nl>). A summary of this results can be seen in Fig. 5. Each figure includes a box plot for every metric against the scanners tested by the organizers. It is important to note, that the organizers

tested with images acquired from scanners with a different technology than the used for our training process, as it is seen in the table, in Fig. 5. All the images provided by the contest were obtained by different brands of 3 Tesla scanners. However for validation purposes, the contest also included images obtained from 1.5 Tesla (AMS GE1.5T) and a hybrid Positron Emission Tomography/MRI (AMS PETMR). The organizer's results includes the average for each scanner and the ranking of each metric against the other teams, our worst results are observed precisely for the AMS GE1.5T and AMS PETMR, images from scanners with different technology not used for training.

The Dice metric performed best for Singapore subjects. All validation results seem to have some standard deviation and outliers, considering the images' origin. In our case, the Singapore patients and the AMS GE3T patients have a Dice score of 0.80 and 0.79, respectively, demonstrating good performance. The Dice score, however, for the AMS PETMR was slightly lower, with a 0.71 value, than the rest of the scanners.

As for the Hausdorff distance metric, our model performed particularly good; in this metric our model was placed in fifth place with an overall rank of 0.017. We also obtained a slightly better result in this metric than the *Reference*, with an averaged distance of 6.30 mm, as compared to ours of 6.24 mm. The image shows that the distance is almost always below the 20 mm mark, with some outliers reaching 40 mm.

In the average volume difference metric, the average value with all the testing patients was 28.26 %. The performance was good in nearly all the scanners except the AMS PEMTR scanner, in that case, the average volume difference was 60.79 %.

Considering the recall, the results obtained from the Amsterdam images were above 0.80. However, with the Utrecht images, we obtained a score of 0.71. Also, it is noticeable that in this work, the performance was better for the exclusive training scanners, in which the other metrics did not perform well. Overall, for the local testing, the recall was the weakest metric obtained in our model, with an overall ranking of 0.137 compared to the other teams. However, the score of 0.78 was not as far from the 0.87, which is the current higher score obtained up to November 2020.

Finally, the F1 score had an average value of 0.74. This metric was the one with a more standard deviation. All the scanners had averaged performances between 0.7 and 0.8, except for the AMS PETMR scanner with 0.65.

Team: bioengineering_espol_team, rank: 0.060 (9th place)

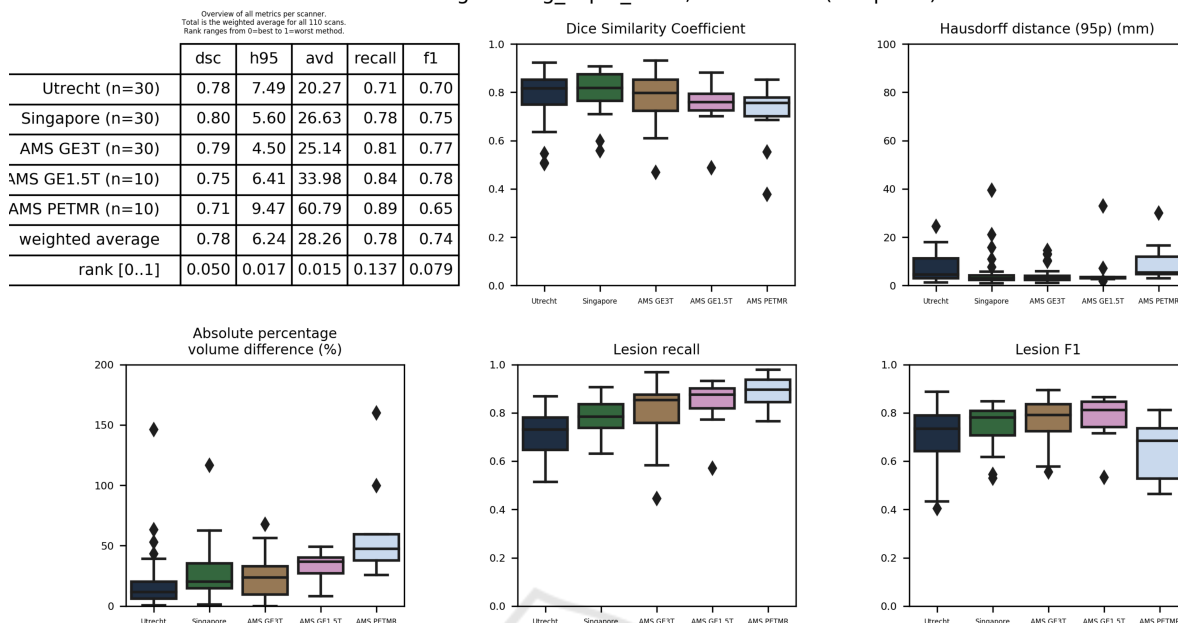


Figure 5: Overall results assessed by WMH challenge organizers.

3.3 Conclusions

We developed and evaluated a Convolutional Network architecture for segmenting automatically WMH, from FLAIR and T1 images. In our local evaluations, we obtained an improvement in 4 out of 5 metrics, as defined by the WMH segmentation challenge. That includes Dice, Hausdorff Distance, F1 score, and Average Volume Difference. We were ranked ninth place overall in the organizers’ assessment, obtaining the best metric for the Hausdorff distance. As it can be seen, our worst performance was with the images coming from scanners not used during training, which could be interpreted as over-fitting the data. However, considering a good performance obtained with this proposed architecture, the hyper-parameter tuning and the re-training of the algorithm, with images from additional scanners technology could improve the algorithm performance. Therefore, the work presented here could lead to other researchers to improve the state of the art, for all society’s benefit.

ACKNOWLEDGEMENTS

We thank the MICCAI-2017 WMH Challenge organizer Dr. Hugo J. Kuijf to share us all the datasets of images and making all the validation results. We also thank Phd. Luis Mendoza for the supervision of the project work and Kevin Cando as this work was

initialized by him as his final degree project.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu).

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3):839–851.

Bento, M., de Souza, R., Lotufo, R., Frayne, R., and Rittner, L. (2017). Wmh segmentation challenge: A texture-based classification approach. In *International MICCAI Brainlesion Workshop*, pages 489–500. Springer.

Berseth, M. (2017). Wmh segmentation challenge, miccai 2017.

Breteler, M., van Swieten, J. C., Bots, M. L., Grobbee, D. E., Claus, J. J., van den Hout, J. H., van Harskamp, F., Tanghe, H. L., de Jong, P. T., van Gijn, J., and Hofman, A. (1994). Cerebral white matter lesions, vascular risk factors, and cognitive function in a population-based study. *44(7):1246–1246*.

Chancay, O., Haro, T., Yapur, M., Alvarado, R., Pastor, M., and Loayza, F. (2015). Nuevo biomarcador en la enfermedad de parkinson mediante el análisis y cuantificación de lesiones cerebrales en secuencias flair obtenidas por resonancia magnética (acl-tool). *Revista Tecnológica-ESPOL*, 28(5).

- Chollet, F. et al. (2015). Keras. <https://github.com/fchollet/keras>, Sidst set 30/01/2020.
- Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus).
- Cocosco, C. A., Zijdenbos, A. P., and Evans, A. C. (2003). A fully automatic and robust brain mri tissue classification method. *Medical image analysis*, 7(4):513-527
- Daliri, R., M. (2012). Automated diagnosis of alzheimer disease using the scale-invariant feature transforms in magnetic resonance images. *J Med Syst*, 36:995-1000.
- Fonov, V., Evans, A., McKinstry, R., Almlí, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102. Organization for Human Brain Mapping 2009 Annual Meeting.
- Fonov, V., Evans, A. C., Botteron, K., Almlí, C. R., McKinstry, R. C., and Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313 - 327.
- Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, W. I., Sanchez, I. C., Litjens, G., de Leeuw, E. F., van Ginneken, B., Marchiori, E., and Platel, B. (2017a). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Science Reports*, 7.
- Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, R. C., de Leeuw, F., Tempny, C., van Ginneken, B., Fedorov, A., Abolmaesumi, P., Platel, B., and Wells, M. W. (2017b). Transfer learning for domain adaptation in mri: application in brain lesion segmentation. *MICCAI 2017, Part III. LNCS*, 10435:516-524.
- Giorgio, A. and De Stefano, N. (2013). Clinical use of brain volumetry. *Journal of Magnetic Resonance Imaging*, 37(1):1-14.
- Gorgolewski, K. J., Esteban, O., Burns, C., Ziegler, E., Pinsard, B., Madison, C., Waskom, M., Ellis, D. G., Clark, D., Dayan, M., Manhães-Savio, A., Notter, M. P., Johnson, H., Dewey, B. E., Halchenko, Y. O., Hamalainen, C., Keshavan, A., Clark, D., Huentburg, J. M., Hanke, M., Nichols, B. N., Wassermann, D., Eshaghi, A., Markiewicz, C., Varoquaux, G., Acland, B., Forbes, J., Rokem, A., Kong, X.-Z., Gramfort, A., Kleesiek, J., Schaefer, A., Sikka, S., Perez-Guevara, M. F., Glatard, T., Iqbal, S., Liu, S., Welch, D., Sharp, P., Warner, J., Kastman, E., Lampe, L., Perkins, L. N., Craddock, R. C., Küttner, R., Bielevietsov, D., Geisler, D., Gerhard, S., Liem, F., Linkersdörfer, J., Margulies, D. S., Andberg, S. K., Stadler, J., Steele, C. J., Broderick, W., Cooper, G., Floren, A., Huang, L., Gonzalez, I., McNamee, D., Papadopoulos Orfanos, D., Pellman, J., Triplett, W., and Ghosh, S. (2016). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. 0.12.0-rc1.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.
- Jack, R. C., O'Brien, P. C., Rettman, D. W. and, S. M. M., Xu, Y., Muthupillai, R., Manduca, A., Avula, R., and Erickson, B. J. (2001). Flair histogram segmentation for measurement of leukoaraiosis volume. *J. Magn. Reson. Imaging*, 14(6):668-676.
- Jin, D., Xu, Z., Harrison, A. P., and Mollura, D. J. (2018). White matter hyperintensity segmentation from t1 and flair images using fully convolutional neural networks enhanced with residual connections. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1060-1064. IEEE.
- Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Kuijff, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M. J., Casamitjana, A., et al. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556-2568.
- Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., and Menze, B. (2018a). Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. *NeuroImage*, 183:650-665.
- Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., and Menze, B. (2018b). Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. *NeuroImage*, 183:650 - 665.
- Li, H., Zhang, J., Muehlau, M., Kirschke, J., and Menze, B. (2018c). Multi-scale convolutional-stack aggregation for robust white matter hyperintensities segmentation. In *International MICCAI Brainlesion Workshop*, pages 199-207. Springer.
- Long, J., Shelhamer, E., and Darrell, T. (2014). Fully convolutional networks for semantic segmentation.
- Loweckamp, B., Chen, D., Ibanez, L., and Blezek, D. (2013). The design of simpleitk. *Frontiers in neuroinformatics*, 7:45.
- McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51-56. Austin, TX.
- Microsoft (2020). Nc-series. <https://docs.microsoft.com/en-us/azure/virtual-machines/nc-series>, Sidst set 02/03/2020.
- Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation.
- Morel, B., Xu, Y., Virzi, A., G'eraud, T., Adamsbaum, C., and Bloch, I. (2016). A challenging issue: detection of white matter hyperintensities on neonatal brain mri. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, page 93-96.
- Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.
- Ronneberger, O., Fischer, P., and Brox, T. (2015a). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical*

- image computing and computer-assisted intervention*, page 234–241.
- Ronneberger, O., Fischer, P., and Brox, T. (2015b). U-net: Convolutional networks for biomedical image segmentation.
- Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*.
- Vrooman, H. A., Cocosco, C. A., van der Lijn, F., Stokking, R., Ikram, M. A., Vernooij, M. W., Breteler, M. M., and Niessen, W. J. (2007). Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. *Neuroimage*, 37(1):71–81.
- Wardlaw, J., Smith, E., Biessels, G., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R., O'Brien, J., Barkhof, F., Benavente, O., Black, S., Brayne, C., Breteler, M., Chabriat, H., DeCarli, C., Leeuw, F.-E., Doubal, F., Duering, M., Fox, N., and v, S. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12:822–838.
- Waskom, M., Botvinnik, O., O’Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Vilalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., and Qalieh, A. (2017). mwaskom/seaborn: v0.8.1 (september 2017).
- WMH Segmentation Challenge (2020). Evaluation. <https://wmh.isi.uu.nl/evaluation/>, Sidst set 30/11/2020.
- Xu, Y., Geraud, T., and Bloch, I. (September 2017a). From neonatal to adult brain mr image segmentation in a few seconds using 3d-like fully convolutional network and transfer learning. *Proceedings of the 23rd IEEE International Conference on Image Processing (ICIP)*, page 4417–4421.
- Xu, Y., Geraud, T., Puybureau, É., Bloch, I., and Chazalon, J. (2017b). White matter hyperintensities segmentation in a few seconds using fully convolutional network and transfer learning. In *International MICCAI Brainlesion Workshop*, pages 501–514. Springer.
- Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57.
- Zijdenbos, A. P. and Dawant, B. M. (1994). Brain segmentation and white matter lesion detection in mr images. *Critical reviews in biomedical engineering*, 22(5-6):401–465.