POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

Assessment of White Matter Hyperintensities segmentation and their correlation with Dementia

TESI MAGISTRALE IN BIOMEDICAL ENGINEERING – INGEGNERIA BIOMEDICA

AUTHOR: Stefano Calafà

ADVISOR: Prof. Giuseppe Baselli

CO-ADVISOR: Valentina Bordin

ACADEMIC YEAR: 2021-2022

1. Introduction

The increase of life expectancy occurred in the last decades contributed to the spreading diffusion of Neurodegenerative Diseases (NDs) in elder individuals. As a result, the scientific community is highly invested in the investigation of biological hallmarks of both ND presence and progression. In this context, White Matter Hyperintensities (WMHs) – a common finding on brain Magnetic Resonance Imaging (MRI) usually associated to normal aging – have recently gained increasing importance as neuroimaging sign for several neurological and cerebro-vascular conditions[1]. They have been associated to demyelination, axonal loss [2] and lesions of the small blood vessels, often causing micro-bleedings with ferritin and calcium deposits. Alongside this, the existing literature has reported their correlation with progressive cognitive impairment, and several NDs such as Alzheimer's Disease (AD), Parkinson's Disease and Multiple Sclerosis [3], [4]. The need for semi-automated and automated approaches allowing for WMH segmentation is useful to provide clinician and researchers with a deeper level understanding around their origin and progression. But, most importantly, to prove their relevance in the early-stage diagnosis of such conditions, since early therapy is the only way to slow neurodegeneration progression, so far. In order to automatically segment and quantify their volume, many Machine Learning (ML) algorithms have been developed over the last twenty years. However, the lack of generalized quantitative standards and of fully optimized performance have prevented their widespread diffusion to most clinical contexts.

The primary aim of this thesis is therefore to evaluate and improve the application of BIANCA, a fully automated and supervised tool developed by the Oxford University to segment WMHs. BIANCA is based on the K-Nearest Neighbors (k-NN) algorithm[5] and works by classifying the image's voxels based on both their local intensity and spatial features. Despite the widely known robustness of these algorithms, some aspects still remain unexplored and require further investigation.

Therefore, with this project, we aimed at evaluating BIANCA performance on a population

of subjects affected by AD, according to different parameters: (i) the number of subjects used for training BIANCA; (ii) the combination of MRI modalities involved in the process; (iii) the utilized training strategy (mixed vs single-site). The ultimate goal was to find the proper combination of values able to optimize results. While carrying out the third step, a harmonization training set derived from a previous study [6] was also validated.

Alongside these aspects, we also aimed at validating the role of WMHs as early-stage biomarker for AD dementia. This was carried out by feeding both imaging data (i.e., WMH volumes extracted using BIANCA) and clinical variables to different ML algorithms trying to predict the Clinical Dementia Rate (CDR) of the OASIS3 participants. Evaluating the importance exerted by WMHs on the final classification allowed us to get a sense of their relevance in diagnostic frameworks.

2. Materials

Dataset

The third release of the Open Access Series of Imaging Studies (OASIS3) is a longitudinal cohort of data which collected, over the course of thirty years, thousands of medical records across several different research projects [7].

The OASIS3 focuses on the effects of both normal aging and early-stage AD including a compilation of data from 1076 participants. Out of those, 605 were cognitively normal adults, while the remaining 493 were affected by various stages of AD cognitive decline.

As regards the imaging sub-part, the dataset includes over two thousand MRI session with a combination of various contrast: T1-weighted (T1-w), T2-weighted, FLAIR, Susceptibility Weighted Imaging (SWI), diffusion weighted imaging and many more. As for the non-imaging sub-part, there are over six thousand records present in OASIS3 containing information about the demographic, habits, medical history, and cognitive status of participants.

All the available OASIS3 data is hosted by the XNAT central repository (central.xnat.org).

3. Methods

The full details on the experiments conducted during this thesis are presented below. Two sections are outlined according to the aspects on which we focused: evaluation of BIANCA performance or the classification algorithm.

3.1. Evaluation of BIANCA performance

Data selection

After accessing the OASIS3 database, we identified the imaging sessions having a combination of the following MRI modalities: FLAIR, T1-w and SWI. We downloaded 206 sessions relative to 172 patients. Then, we decided to keep a single session for patients who had multiple ones and to use only images acquired with the 3T Siemens TrioTim 35248, which was the scanner utilized in the majority of cases. We therefore remained with a total of 159 sessions (from 159 patients).

Eventually, we further narrowed the dataset by visually inspecting the FLAIR images of each participant and considering the following inclusion criteria: high lesional loads for WMHs, no strong artifacts of any kind, and a "regular" brain anatomy. Accordingly, we selected a group of 40 patients on which we carried out the final analysis. Information about their demographics is reported as follows:

- Age = 69.83 ± 6.67;
- Female:Male ratio = 23:17.

Manual WMH mask creation

To derive the proper *ground truth* necessary to train BIANCA, we manually segmented WMHs from the FLAIR scans of all subjects involved in our analysis. At this purpose, we used the Jim8 software, a display package developed by Xinapse that allows for an easy viewing and easy analysis of MRI, CT and other types of medical images.

We first created *regions of interest* underlying WMH contours, and then used the *masker tool* to derive binary images from them. Eventually, we repeated the segmentations 4 months after the first round. The different available masks were referred to as "preliminary" and "expert", respectively. Indeed, the former were outlined in a time span of 4 weeks at the very beginning of the project, thereby being associated to little experience on both WMH morphology and Jim8 usage. On the other hand, the "expert" segmentations were performed in a time span of 5 days half of the way into the project. Thus, they reflected higher experience and improved rating abilities.

Image pre-processing

In order for BIANCA to perform optimally a thorough data preparation is required. Firstly, since the tool works in single subject's space, all the input images need to be registered to a common MRI scan (FLAIR in our case). In addition, the spatial inhomogeneities of the magnetic field should be corrected and at least one MRI modality has to be brain-extracted. Finally, a registration matrix from base image to standard MNI space needs to be derived, to allow for the extraction of spatial features without the presence of any bias.

Therefore, every MRI scan at our disposition underwent the following steps, carried out using tools from the FSL library [8]–[10]: i) brain extraction conducted using BET; ii) biasfield correction conducted using FAST; iii) registration between the current image space and the FLAIR space conducted using FLIRT (performed only on the T1-weighted and SWI images). A registration matrix from FLAIR to MNI space was also derived for each subject using FLIRT and combining its results with that of former pre-processing steps. Finally, an exclusion mask was applied to FLAIR scans to exclude anatomical structures that might be incorrectly classified as WMHs.

Running BIANCA

Once pre-processed, the images were ready to be fed to BIANCA. The algorithm could be trained and tested either separately or with a leave-one-out validation approach. In both cases, a *masterfile* containing all the required images and matrices had to be created. Specifically, the *masterfile* is a text file containing a row for every subject involved in the analysis and, for each row, a list of all the necessary files (i.e., their paths). These latter are written following a specific order, which needs to be maintained throughout the whole document.

The *masterfile* was the starting point to run BIANCA from Terminal and perform the analysis steps outlined below.

Incremental Training analysis

At first, we evaluated the performance of BIANCA obtained with an incremental number of training subjects (i.e., 10, 15, 20, 25, 30). The aim was to assess the existence of numerosity ranges that applicable in the BIANCA training to segmentation. The analysis was conducted for every combination of MRI modality described in the following section, using separate procedures for training and testing. Performance was evaluated by comparison with both the "preliminary" and the "expert" manual segmentations.

Multimodality analysis

Secondly, we evaluated how the different combination of MRI modalities impacted on the final performance. To do that we fixed the number of training subjects to 40 (using a leave-one-out validation approach), including the entire dataset at our disposition. Then, using the "expert" manual masks, we assessed the following MRI combinations:

- FLAIR only
- FLAIR + T1-weighted
- FLAIR + SWI
- FLAIR + T1-weighted + SWI
- SWI only

Harmonization pipeline

Then, we evaluated the performance reached by BIANCA when trained on a well-known population and tested on a different one (mixed training approach). In particular, we used the training set from [6] designed with the purpose of harmonizing data from heterogeneous cohorts. The result obtained on their testing set were overall satisfying. However, the training outcome had never been validated on different datasets with respect to the one involved in its development – Whitehall and UK Biobank. So, confirming the previous results on OASIS3 was done in the present study to show the widespread applicability of the previous training when presented completely new data coming from different protocols.

Finally, results from the mixed training approach were compared with those obtained from the incremental training set analysis (representing a single-site training approach) to evaluate the difference between different training strategies (mixed vs single-site). The analysis was conducted using the "expert" manual masks.

Performance evaluation

Performance was evaluated by means of the Dice Similarity Index (DICE), calculated as 2*(voxels in the intersection of manual and BIANCA

masks)/(manual mask lesion voxels + BIANCA lesion voxels). The statistical significance for the evaluated comparisons was assessed by means of ANOVA tests.

3.2. The Classification algorithm

Data selection and pre-processing

In order to build the classification algorithm, we extended the analysis to a larger fraction of the OASIS3 dataset. In particular, we selected 471 imaging sessions with available FLAIR and T1 scans, among which there were the 40 manually labeled used for the former evaluations. The MRI data had to be processed and fed to BIANCA after training the tool on data from the Whitehall and UK Biobank populations. This, in order to derive the volumetric amounts of WMHs used as imaging variable within the model. Secondly, we had to download all the clinical records available from the OASIS3 (which were higher in number with respect to their imaging counterpart) and to eventually select only the ones that matched our 471 images according to a time gap minimization criterion.

Eventually, the dataset had to be filtered and properly organized to be suitable for the step of model creation. First, the target variable (namely the CDR) was binarized, creating two groups: CDR = 0 and CDR = 1. After that, all NaNs values were eliminated from the remaining variables, categorical features were turned into binary ones and continuous features were normalized.

Model Creation

The dataset was split in *training*, *validation* and *testing sub-sets* which allowed to build and evaluate the following machine learning models: support vector machines (SMVs), random forest classifier (RFCs) and artificial neural networks (ANNs). The RFC was trained both with and without a step of principal component analysis (PCA) carried out on the continuous variables of the dataset. The model hyperparameters were tuned using a Grid Search approach with 5-fold Cross Validation for the SVM and RFC models. This step was instead carried out manually for the ANN. Each model was trained and tested using three different datasets, each with a specific ratio between the CDR = 0 and CDR = 1 records: 1:1, 2:1, 3:1. The last analysis aims at training performance evaluation in the presence of a decreased prevalence of positive AD cases. Results were evaluated separately for each case.

Performance evaluation

Performance was assessed throughout confusion matrices and of the following indexes: accuracy, precision, recall, F1-score. The importance exerted by the different input variable on the final classification was instead evaluated by means of the permutation importance score.

4. Results and discussion

Evaluation of BIANCA performance

As regards the incremental training analysis, two major information can be derived from results (displayed in Fig. 1). First, SWI alone does not carry any information as for the WMH segmentation, providing very poor outcomes. Second, there is an increase in performance when adding more subjects to the training which, however, reaches a plateau either around 20 or 25 subjects. This depends on to the segmentation round used to carry out the evaluation ("preliminary" or "expert" manual masks, respectively). This difference led us to assume that the "expert" segmentations were more accurate and therefore suitable to assess results, and further highlighted the need for accurate ground-truth (alias, gold-standard) in the training process.

As for the multi-modality analysis (see Fig. 2), a significant difference in BIANCA performance was demonstrated for the following:

- "FLAIR" against "FLAIR + SWI" ($0.01 < p\text{-value} < 0.05$);
- "FLAIR + T1w" against "FLAIR + SWI" ($p\text{-value} < 0.001$);
- "FLAIR + T1w" against "FLAIR + T1w + SWI" ($p\text{-value} < 0.001$);

In particular, the first combination was always better than the second. These results indicate "FLAIR" and "FLAIR + T1w" as the best combinations of MRI modalities to run BIANCA with. In addition, they acknowledge a worsening in performance caused by the addition of the SWI contrast, which seemed to act as a source of noise for the algorithm.

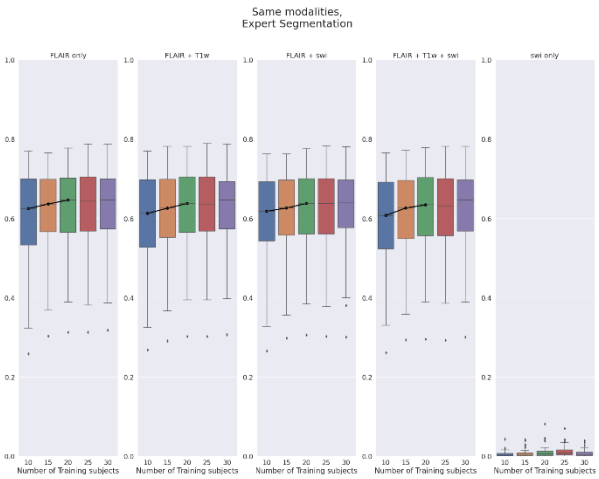

Figure 1. Boxplot of the DICE index (represented on the y axis) between BIANCA outputs obtained training with an increasing number of subjects (represented on the x axis) and the corresponding "expert" segmentations.

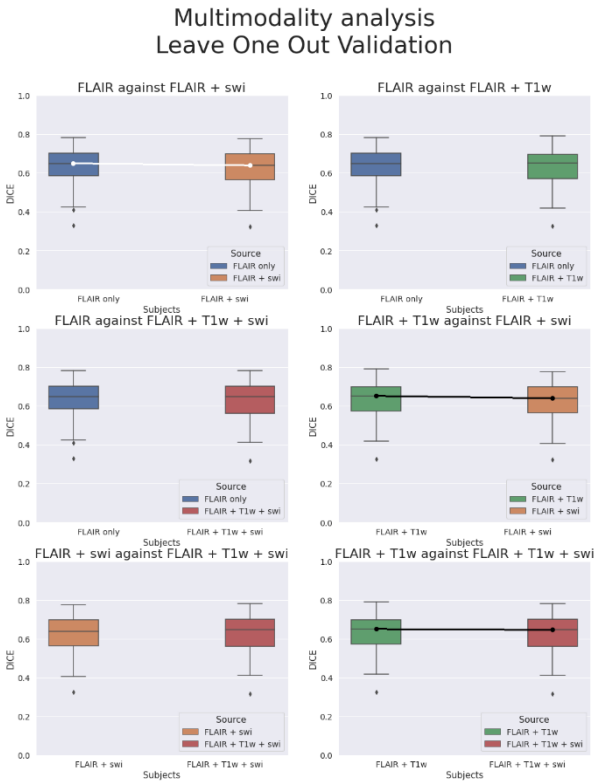

Figure 2 Pairwise comparisons of the distributions present in Fig. 3.6. Full lines indicate significant differences (p-value < 0.05) between distributions, calculated using a RANOVA test.

As for the harmonization pipeline, after testing it on data from the OASIS3, results were compared with the existing literature (in which the testing phase was conducted on the Whitehall and UK Biobank populations). We summarize the main descriptive statistics of both cases in Table 1. Results appeared being fully comparable and proved the achievement of a certain degree of validation for the training set of [6].

	Testing on OASIS3	Testing on Whitehall and UK Biobank
Mean	0.61	0.52 (for WH Scanner1); 0.47 (for WH Scanner2); 0.63 (for UK Biobank);
Median	0.63	0.54 (for WH Scanner1); 0.47 (for WH Scanner2); 0.65 (for UK Biobank);
Std	0.12	0.10 (for WH Scanner1); 0.05 (for WH Scanner2); 0.10 (for UK Biobank);

Table 1. Segmentation performance obtained testing with: i) data from the OASIS3; ii) data from the Whitehall and UK Biobank datasets. The training phase was conducted in both cases using the Whitehall + UK Biobank dataset from [6].

Additionally, in Fig. 3 we report the results obtained when comparing the mixed training strategy (i.e., training on Whitehall + UK Biobank and testing on OASIS3 – blue) with the single-site approach represented by the incremental training analysis (orange). A significant improvement was provided by the latter only when the number of involved subjects was higher than or equal to 20.

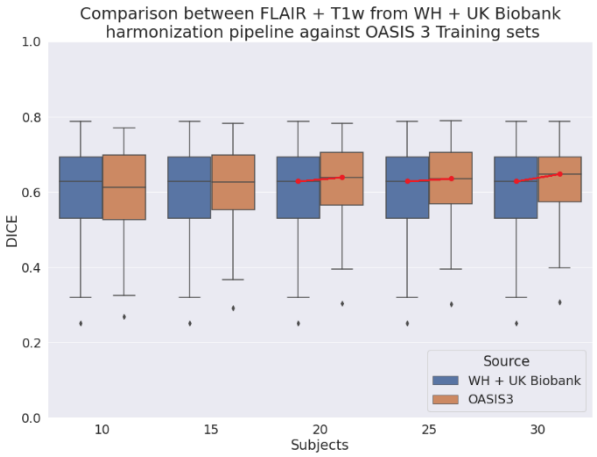

Figure 3. Boxplot of the DICE index (represented on the y axis) between BIANCA outputs obtained with both a mixed training strategy and a single-site training strategy.

Classification Model

As regards the implemented models, SVM was the one providing the best tradeoff between classification performance and explainability. Its quantitative evaluation metrics reported the following values: 80% accuracy, 54% precision, 85% recall, 66% F1-score. On the other hand, results from the permutation ranking (displayed in Fig. 4) indicated the WMH volume among the most relevant features as for the classification of CDR, thus confirming their importance as neuroimaging hallmarks for AD dementia. A great role was also

played by the subjects' age and by the number of years passed since they were first included into the study.

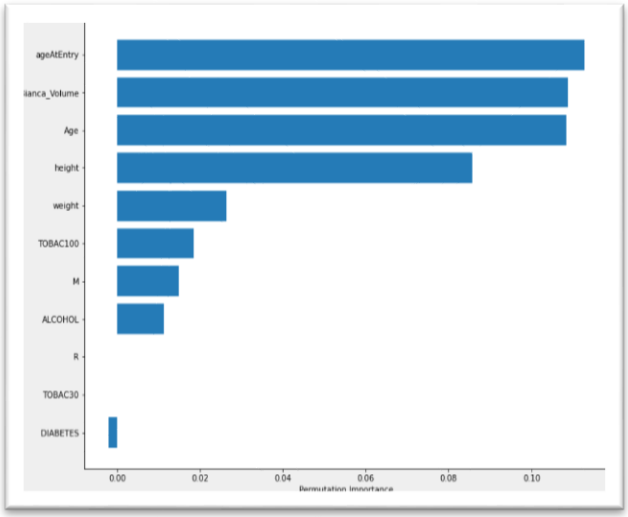

Figure 4. Permutation importance (reported on the x axis) of the different input variables (reported on the y axis) involved in the SVM model.

As for the RFC model, results obtained using PCA were slightly better with respect to the SVM (82% accuracy and 80% for precision, recall and F1-score). However, this model suffered from significant explainability limitations due to the presence of linear combinations of imaging and non-imaging features among the most important variables. On the other hand, the RFC model without PCA gave the worst performance (76% accuracy, 80% precision, 60% recall, 68% F1-score) and a great deal of variability in the permutation importance results.

Finally, the ANN model reached the best classification performance but, due to its lack of explainability, we only used it as gold standard for the other models. Its results were: 84% accuracy, 76% precision, 81% recall, 78% F1-score.

5. Conclusions

contrast and recognized its presence as potential source of noise for the segmentation. Finally, we found that the well-known superiority of the single-site training with respect to the mixed, is held only if a minimum number of subjects are used for training. This was again represented by 20. Furthermore, the results obtained when applying an external training set – developed with harmonization purposes – to our population gave a performance comparable with that of the corresponding literature. This reinforced the evidence of the widespread applicability of the previous training of BIANCA, based on harmonization techniques. Finally, the ML models we built, successfully confirmed the importance of WMHs in the assessment of AD Clinical Dementia. With these findings we have strengthened our former knowledge on the automatic segmentation strategy represented by BIANCA. In addition, we have validated a harmonization pipeline to derive integrated measures of WMH volumes. Eventually, we have brought further evidence about the role of WMHs as early-stage hallmark of AD dementia.

6. Bibliography

[1] Z. Morris *et al.*, "Incidental findings on brain magnetic resonance imaging: Systematic review and meta-analysis," *BMJ (Online)*, vol. 339, no. 7720, pp. 547–550, Sep. 2009, doi: 10.1136/bmj.b3016.

[2] A. J. Farrall and J. M. Wardlaw, "Blood–brain barrier: Ageing and microvascular disease – systematic review and meta-analysis," *Neurobiology of Aging*, vol. 30, no. 3, pp. 337–352, Mar. 2009, doi: 10.1016/j.neurobiolaging.2007.07.015.

[3] A. C. Birdsill *et al.*, "Regional white matter hyperintensities: aging, Alzheimer's disease risk, and cognitive function," *Neurobiology of Aging*, vol. 35, no. 4, pp. 769–776, Apr. 2014, doi: 10.1016/j.neurobiolaging.2013.10.072.

[4] M. Dadar *et al.*, "White matter hyperintensities are linked to future cognitive decline in de novo Parkinson's disease patients," *NeuroImage: Clinical*, vol. 20, pp. 892–900, 2018, doi: 10.1016/j.nicl.2018.09.025.

[5] L. Griffanti *et al.*, "BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities," *Neuroimage*, vol. 141, pp. 191–205, Nov. 2016, doi: 10.1016/J.NEUROIMAGE.2016.07.018.

[6] V. Bordin *et al.*, "Integrating large-scale neuroimaging research datasets: Harmonisation of white matter hyperintensity measurements across Whitehall and UK Biobank datasets,"

Neuroimage, vol. 237, Aug. 2021, doi: 10.1016/j.neuroimage.2021.118189.

[7] P. J. LaMontagne *et al.*, "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease," *medRxiv*, p. 2019.12.13.19014902, Jan. 2019, doi: 10.1101/2019.12.13.19014902.

[8] M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *Neuroimage*, vol. 62, no. 2, pp. 782–790, Aug. 2012, doi: 10.1016/j.neuroimage.2011.09.015.

[9] S. M. Smith *et al.*, "Advances in functional and structural MR image analysis and implementation as FSL," *Neuroimage*, vol. 23, pp. S208–S219, Jan. 2004, doi: 10.1016/j.neuroimage.2004.07.051.

[10] M. W. Woolrich *et al.*, "Bayesian analysis of neuroimaging data in FSL," *Neuroimage*, vol. 45, no. 1, pp. S173–S186, Mar. 2009, doi: 10.1016/j.neuroimage.2008.10.055.