



Universidade de Aveiro
Ano 2021

GIUDITTA TODESCO

**DETEÇÃO AUTOMÁTICA DE LESÕES DE
ESCLÉROSE MÚLTIPLA EM IMAGENS DE
RESSONÂNCIA MAGNÉTICA CEREBRAL
UTILIZANDO BIANCA**

**AUTOMATIC DETECTION OF MULTIPLE
SCLEROSIS LESIONS IN BRAIN MAGNETIC
RESONANCE IMAGING USING BIANCA**



Universidade de Aveiro
Ano 2021

GIUDITTA TODESCO

**DETEÇÃO AUTOMÁTICA DE LESÕES DE
ESCLEROSE MÚLTIPLA EM IMAGENS DE
RESSONÂNCIA MAGNÉTICA CEREBRAL
UTILIZANDO BIANCA**

**AUTOMATIC DETECTION OF MULTIPLE SCLEROSIS
LESIONS IN BRAIN MAGNETIC RESONANCE
IMAGING USING BIANCA**

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Tecnologias da Imagem Médica, realizada sob a orientação científica da Doutora Sílvia De Francesco, Professora da Escola Superior de Saúde da Universidade de Aveiro, e sob a co-orientação do Doutor Stefano Tambalo, Center for Mind and Brain Sciences, Universidade de Trento, Docente Convidado da Escola de Medicina e Cirurgia da Universidade de Verona.

o júri

presidente

Professor Doutor Augusto Marques Ferreira da Silva
Prof. Associado, Dep. de Eletrónica e Telecomunicações e Informática da Universidade de Aveiro

Professora Doutora Otilia da Anunciação Cardoso d'Almeida
Profª. Auxiliar, Faculdade de Medicina da Universidade de Coimbra

Professora Doutora Sílvia De Francesco
Profª. Adjunta, Escola Superior de Saúde da Universidade de Aveiro

palavras-chave

Ressonância Magnética, Esclerose Múltipla, Classificador k-NN, BIANCA, Machine Learning

resumo

Este trabalho teve como objetivo a concepção e otimização de um procedimento para aplicação de um algoritmo de *Machine Learning*, o classificador BIANCA (*Brain Intensity AbNormalities Classification Algorithm*), para detecção de lesões caracterizadas por hiperintensidade em T2 da matéria branca em estudos clínicos de Esclerose Múltipla por Ressonância Magnética.

O procedimento concebido inclui pré-processamento, identificação das lesões e otimização dos parâmetros do algoritmo BIANCA.

O classificador foi treinado e afinado utilizando os 15 casos clínicos que constituíam o conjunto de treino do desafio MICCAI 2016 (*Medical Image Computing and Computer Assisted Interventions*) e posteriormente testado em 30 casos clínicos de uma base de dados pública (Lesjak et al.).

Os resultados obtidos são em concordância com os alcançados pelas 13 equipas que concluíram o desafio MICCAI 2016, confirmando que este algoritmo pode ser uma ferramenta válida para a detecção e classificação de lesões de Esclerose Múltipla em estudos de Ressonância Magnética.

keywords

Magnetic Resonance, Multiple Sclerosis, k-NN classifiers, BIANCA, Machine Learning

abstract

The aim of this work was to design and optimize a workflow to apply the Machine Learning classifier BIANCA (Brain Intensity AbNormalities Classification Algorithm) to detect lesions characterized by white matter T2 hyperintensity in clinical Magnetic Resonance Multiple Sclerosis datasets.

The designed pipeline includes pre-processing, lesion identification and optimization of BIANCA options.

The classifier has been trained and tuned on 15 cases making up the training dataset of the MICCAI 2016 (Medical Image Computing and Computer Assisted Interventions) challenge and then tested on 30 cases from the Lesjak et al. public dataset.

The results obtained are in good agreement with those reported by the 13 teams concluding the MICCAI 2016 challenge, thus confirming that this algorithm can be a reliable tool to detect and classify Multiple Sclerosis lesions in Magnetic Resonance studies.

Content Index

Acronyms	3
1. Introduction	5
2. State of the art	9
3. Materials and methods	15
3.1 BIANCA	15
3.2 Datasets	16
3.3 Pre-processing	18
3.3.1 Brain Extraction Tool	18
3.3.2 Linear Registration	18
3.4 BIANCA optimization, training and test	19
3.4.1 Thresholding	19
3.4.2 Spatial weighting	19
3.4.3 Patches	19
3.4.4 Location of training points	19
3.4.5 Number of lesion and non-lesion points	19
3.4.6 Lesion load	20
3.5 Post-processing and performance evaluation	22
4. Results	25
4.1 Threshold	25
4.2 Spatial weighting	26
4.3 Patches	26
4.4 Location of training points	26
4.5 Number of lesion and non-lesion points	28
4.6 Lesion load	28
4.7 Optimization	29
4.8 BIANCA test results	30
5. Discussion	33
6. Conclusion	39
7. Bibliography	41
Acknowledgements	45

Acronyms

ACC: Accuracy

AD: Average Distance

BBB: Brain Blood Barrier

BET: Brain Extraction Tool

BIANCA: Brain Intensity AbNormalities Classification Algorithm

CAD: Computer Aided Diagnosis

CDR: Correct Detection Rate

COG: Centre Of Gravity

DER: Detection Error Rate

DICOM: Digital Imaging and COmmunications in Medicine

DIR: Double Inversion Recovery

EF: Extra Fraction

FALL: Fallout

FAST: FSL Automated Segmentation Tool

FDR: False Detection Rate

FLAIR: Fluid Attenuated Inversion Recovery

FLIRT: FMRIB's Linear Image Registration Tool

FMRIB: Functional Magnetic Resonance Imaging of the Brain

FN: False Negative

FOV: Field of View

FP: False Positive

FSL: FMRIB Software Library

Gd: Gadolinium

GM: Grey Matter

HD: Hausdorf Distance

ICC: Intra Class Correlation

JJ: Jacard Index

k-NN: k-Nearest Neighbour
LOCATE: LOcally Adaptive Thresholds Estimation
MICCAI: Medical Image Computing and Computer Assisted Interventions
ML: Machine Learning
MNI: Montreal Neuroimaging Institute
MPRAGE: Magnetisation PRepared RAPid Gradient Echo
MRI: Magnetic Resonance Imaging
MS: Multiple Sclerosis
NAWM: normal appearing white matter
NifTI: Neuroimaging Informatics Technology Initiative
OER: Outline Error Rate
PD: Proton Density
POF: Probabilistic Overlap Fraction
PPMS: primary-progressive Multiple Sclerosis
PPV: Positive Predictive Value
PrC: Pearson's Coefficient
PSI: Probabilistic Similarity Index
RAE: Relative Area Error
RRMS: relapsing-remitting Multiple Sclerosis
SEN: Sensitivity
SI: Similarity Index
SPE: Specificity
SPMS: secondary-progressive Multiple Sclerosis
sw: spatial weighting
TN: True Negative
TP: True Positive
VD: Volume Difference
WM: White Matter
WMH: White Matter Hyperintensities

1. Introduction

Multiple Sclerosis (MS) is a chronic disease that affects the central nervous system and changes its morphology and structure. Its pathological hallmarks include demyelination, inflammation, gliosis, axonal damage and brain atrophy (1,2). Lesions can affect all the tissue that contains myelin, so it can occur in the spinal cord, mainly in the cervical segment and usually on the posterior and lateral regions, in nerves, most commonly in the optic nerve, and in the brain, both in the Grey Matter (GM) and in the White Matter (WM). In the brain, lesions are located mostly in the periventricular and in the juxtacortical WM regions, in the corpus callosum and in infratentorial areas, mainly pons and cerebellum, and usually have oval or elliptical shapes (3).

The loss of myelin interrupts the transmission of the signals through the axons of the nervous system, resulting in a disruption of the body functions connected to this damage. Moreover, MS causes multiple inflammations that are reversible. Given the reversibility of the inflammation process, the axonal losses are considered the appropriate markers for the progression of the disease (2).

The MS patients can be divided in three clinical groups: relapsing–remitting MS (RRMS), secondary-progressive MS (SPMS), and primary-progressive MS (PPMS). The RRMS consists of stable periods interspersed by relapses followed by partial or whole recovery.

In SPMS, the steady progression stage, which differs from the previous one by the degree of disability, there is a lack of basic recovery after subsequent relapses. The PPMS patients are affected by progressive disease with occasional stability and temporary improvements (2).

The diagnosis can be achieved with Magnetic Resonance Imaging (MRI) because the loss of myelin creates a more hydrophilic environment and increases the water content in the lesions. Therefore, there is an increase of proton density and a prolongation of T1 and T2 relaxation times, which results in an increased MR signal intensity of lesions on PD/T2 weighted and decreased intensity in T1 weighted sequences (4).

Thanks to that, MRI is highly sensitive in detecting MS plaques and can also provide quantitative assessment of inflammatory activity and lesion load (5). Moreover, it can give quantitative estimation of the brain atrophy due to MS.

For these reasons, MRI is considered the most important modality to study the progression of this disease.

Despite their high sensitivity in the detection of MS lesions, conventional MRI sequences have difficulties to disclose the actual burden of GM and mixed GM-WM lesions because of their reduced dimensions and the low difference in their relaxation times versus the normal-appearing GM than that between WM lesions and normal-appearing WM (NAWM) (3).

The MS lesions can be divided into three groups based on their characteristics in different MRI sequences:

- T2w lesions: they are hyperintense when compared to the surrounding WM in T2w, PDw and T2 FLAIR sequences. They may be iso- or hypointense in T1w images. T2w lesions are not pathologically specific and can result from inflammation, edema, demyelination, or axonal loss. These lesions are shown in green in figure 1.
- Gadolinium (Gd)-enhanced lesions: these lesions show an increased signal intensity on T1w images after injection with Gd, and are usually associated with hyperintensity in T2w, PDw, and T2 FLAIR images. Some lesions that appear hypointense compared to normal-appearing WM (NAWM) on T1w images before Gd injection may only become isointense with NAWM after its administration. These lesions can often be missed if a pre-injection T1w image is not acquired for comparison. Gd enhancement is associated with active inflammatory activity and breakdown of the blood–brain barrier (BBB). These types of lesion are shown in blue in figure 1.
- Black holes: This term is used to refer to chronic T1w hypointense lesions. These lesions usually appear hyperintense on T2w, PDw, and T2 FLAIR images. Since transient inflammation may be associated with hypointensity in T1w images, some hypointensities in T1w images may disappear after a month or two. Thus, to qualify as a “black hole”, a T1w lesion should not change its signal intensity upon Gd injection and should generally have been present for at least several months. Such lesions are usually associated with relatively more severe tissue injury and axonal loss (6). The black holes are shown in red in figure 1.

Nowadays, the study and the detection of the lesions are performed manually by experts that use their high level of anatomical knowledge to identify the lesions’ evolution. Since this process is time consuming and prone to intra-observer and inter-observer variability (7), an automated lesion detection technique could reduce both the disagreement and the time involved in the process.

Resorting to modern Machine Learning (ML) algorithms, many automated procedures have been proposed for the detection and analysis of MS lesions, but there is no one commonly accepted in the clinical practice (5).

The aim of this work is to test an automated identification of the MS lesions in the brain white matter using the new algorithm, Brain Intensity AbNormality Classification Algorithm (BIANCA), developed for the Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library (FSL, Oxford, UK) package, to segment the hyperintense abnormalities in brain MRI.

In chapter two we introduce the state of the art of MS lesion segmentation with ML algorithms, with a particular attention to the k-Nearest Neighbour (k-NN) classifier method. In chapter three BIANCA is presented, and the pipeline of pre-processing described. Moreover, in chapter three and four the optimization of BIANCA options and of the results are explained.

In chapter five the results and some considerations about the work done are presented.

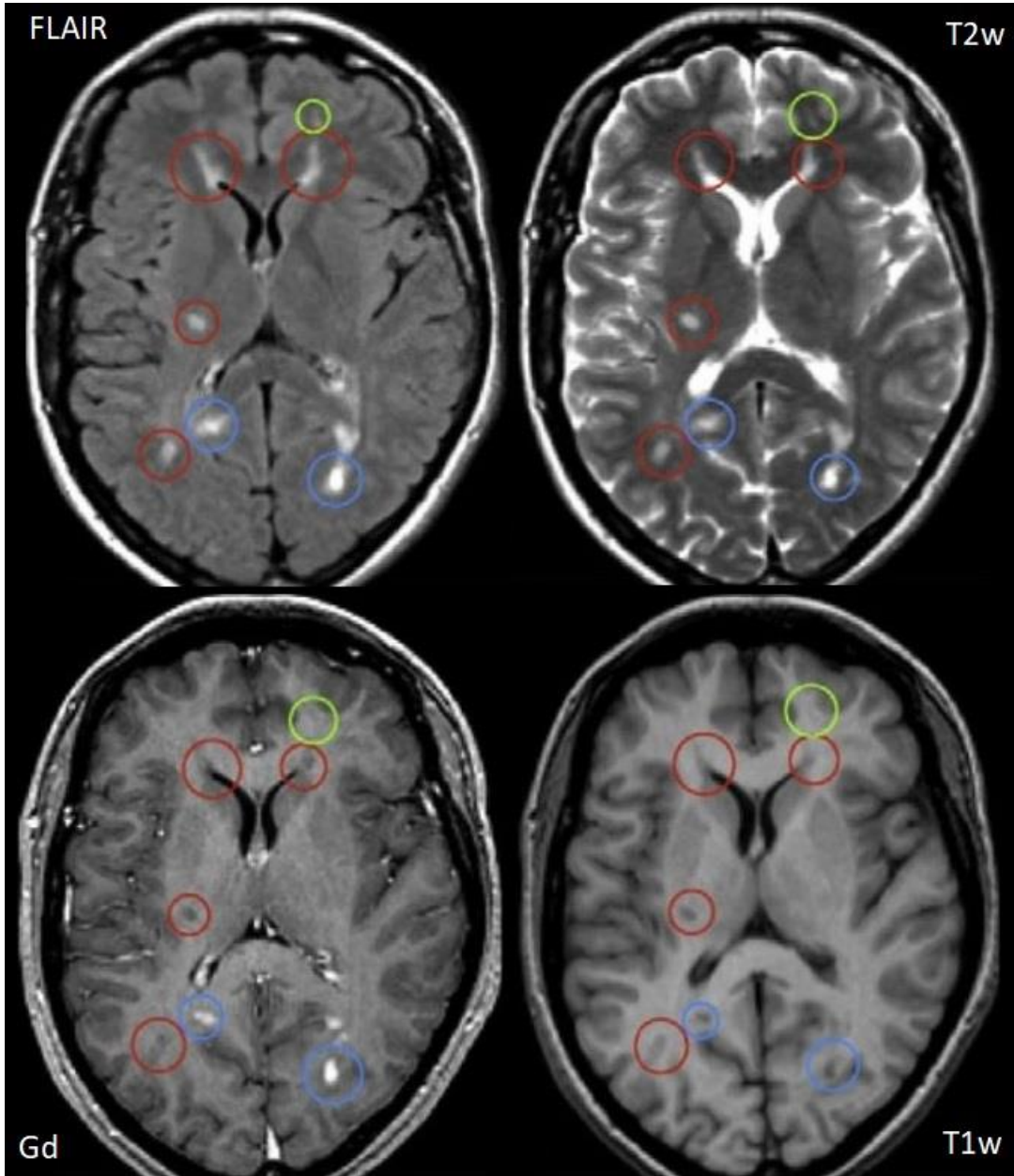


Figure 1: Example of MS lesions on MRI: FLAIR, T2w, Gd-enhanced T1w and T1w images. Lesion types that can be observed: in blue, enhancing lesions, in green, lesions visible only on T2w, in red, black holes (6).

2. State of the art

Since MRI is highly sensitive in detecting MS plaques and MS correlated problems in the central nervous system, many different automated algorithms have been presented in the last years. To the best of our knowledge, six reviews have been published about the topic: Mortazavi et al. (2012) (2), two reviews of Lladó et al. (2012) (7,8), García-Lorenzo et al. (2013) (6) Danelakis et al. (2018) (5) and Balakrishnan et al. (2021) (9). They present all the methods of lesion detection and segmentation used until 2021. Moreover, they give information about the image preparation and the evaluation metrics of the algorithms.

A Computed Aided Diagnosis (CAD) require some common steps (2):

- Image acquisition: the most common MRI protocols used for MS patients include T1-weighted (T1w), T2-weighted (T2w), PD-weighted (PDw) and fluid attenuated inversion recovery T2 (FLAIR) sequences. Usually also T1w sequences after the injection of Gadolinium are acquired (5);
- Pre-processing: In case of automated detection and segmentation of MS lesions, the most common steps include:
 - Registration: the multi-modality images used need to be co-registered to correct patient motion and to obtain a unique map of the brain.
 - Brain extraction: the non-brain tissues are removed from the image.
 - Inhomogeneity correction and noise reduction: it is necessary to remove the inhomogeneity due to the random noise, which can bring to misclassifications.
 - Intensity normalization.
- Feature extraction and transformation: since the features have different ranges, a normalization is needed to obtain meaningful distances in feature space for selecting the k “nearest” neighbours.
- Classification with the chosen model.
- Post-processing to reduce the misclassified elements.

Due to the high variety of MS lesions, most of the approaches for automated segmentation combine different characteristics of the lesions in the images obtained with different sequences (8).

The aforementioned reviews point out the difficulty of comparing the existing techniques which aren't tested on the same dataset and with the same measures. Therefore, the authors propose some common evaluation measures that should be used for the future works. These metrics are divided in:

- deterministic metrics, in which each voxel is assigned to only one tissue type (table 1).
- area and volume measures (table 2).
- probabilistic measures, that use probability maps which attribute to each voxel the probability to belong to a class (table 3).
- distance measures, which evaluate border distances between the segmentation and the ground truth (table 3).

The deterministic metrics are based on the confusion matrix (figure 2), which is the result of the machine learning classifiers. From a clinical point of view, they can be explained as:

- True positive (TP): Refers to correctly segmented MS lesions areas.
- True negative (TN): Refers to correctly rejected MS lesions areas.
- False positive (FP): Refers to incorrectly segmented MS lesions areas.
- False negative (FN): Refers to incorrectly rejected MS lesions areas.

The evaluation measurements presented in table 1, 2 and 3 are taken from Danelakis review (5).

Table 1: Deterministic evaluation measurements for automated segmentation (5,10).

DETERMINISTIC MEASUREMENTS		
MEASURE	CALCULATION	DESCRIPTION
Sensitivity (SEN) Also called: Overlap Fraction (OF), True Positive Rate (TPR), Recall	$SEN = \frac{TP}{FN + TP}$	TP: True Positive FN: False Negative
Specificity (SPE) Also called: True Negative Rate (TNR)	$SPE = \frac{TN}{TN + FP}$	TN: True Negative FP: False Positive
Accuracy (ACC)	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$	
Similarity Index (SI) Also called: F ₁ Score, Dice Similarity Coefficient (DSC)	$SI = \frac{2TP}{2TP + FP + FN}$	
Positive Predictive Value (PPV) Also called: Precision, Reliability	$PPV = \frac{TP}{TP + FP}$	
Fallout (FALL) Also called: False Positive Rate (FPR), False Alarm Ratio	$FALL = 1 - SPE = \frac{FP}{FP + TN}$	
Extra Fraction (EF)	$EF = \frac{FP}{TP + FN}$	
Jacard Index (JI)	$JI = \frac{TP}{TP + FP + FN}$	

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 2: Visual representation of the confusion matrix.

Table 2: Area and volume evaluation measurements for automated segmentation (5,10).

AREA AND VOLUMES MEASUREMENTS		
MEASURE	CALCULATION	DESCRIPTION
Detection Error Rate (DER)	$DER = \frac{DE}{MTA}$	DE: Detection Error MTA: Mean Total Area
Outline Error Rate (OER)	$OER = \frac{OE}{MTA}$	OE: Outline Error
Correct Detection Ratio (CDR)	$CDR = \frac{TP}{MS}$	MS: Manually Segmented Area
False Detection Ratio (FDR)	$FDR = \frac{AS - TP}{MS}$	AS: Automatically Segmented Area
Relative Area Error (RAE)	$RAE = \frac{AS - MS}{MS}$	
Volume Difference (VD)	$VD = \frac{FN - FP}{2TP + FP + FN}$	

Table 3: Probabilistic and distance evaluation measurements for automated segmentation (5,10).

PROBABILISTIC AND DISTANCE MEASUREMENTS		
MEASURE	CALCULATION	DESCRIPTION
Intra-Class Correlation (ICC)	$ICC = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\epsilon^2}$	σ_s^2 : differential variance between the segmentations σ_ϵ^2 : differential variance between the points in the segmentations
Pearson's r Coefficient (PrC)	$PrC = \frac{\sum_{i=1}^N (x_i + \bar{x})(y_i + \bar{y})}{\sqrt{\sum_{i=1}^N (x_i + \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i + \bar{y})^2}}$	x_i, y_i : volumes of the ground truth and the automatic segmentation, \bar{x}, \bar{y} : respective means of the absolute volumes N: number of time points.
Probabilistic Similarity Index (PSI)	$PSI = \frac{2 \times \sum P_{x,gs=1}}{\sum 1_{x,gs=1} + \sum P_x}$	$\sum P_{x,gs=1}$: sum over all voxel probabilities, where in the manual segmentation (gold standard, gs) the voxel value=1; $\sum 1_{x,gs=1}$: sum over all the voxel of in the gold standard; $\sum P_x$: sum over all the probabilities in the probability map.
Probabilistic Overlap Fraction (POF)	$POF = \frac{\sum P_{x,gs=1}}{\sum 1_{x,gs=1}}$	
Probabilistic Extra Fraction (PEF)	$PEF = \frac{\sum P_{x,gs=0}}{\sum 1_{x,gs=0}}$	$\sum P_{x,gs=0}$: sum over all voxel probabilities where in the gold standard the intensity value=0
Hausdorf Distance (HD)	$HD(A, B) = \max(h(A, B), h(B, A))$	$h(A, B) = \max_{a \in A} \min_{b \in B} \ a - b\ $ $\ a - b\ $: Euclidean distance A, B: two finite sets
Average Distance (AD)	$AD(A, B) = \max(d(A, B), d(B, A))$	$d(A, B) = \frac{1}{N} \sum_{a \in A} \min_{b \in B} \ a - b\ $ $\ a - b\ $: the Euclidean distance; A, B: two finite sets; N: number of elements of the finite sets;

The ML algorithms used to segment the MS lesions can be divided in supervised and unsupervised methods. The former “learn” the definition of lesions from example images that have been previously segmented by another method, usually manual segmentation. The latter do not require labelled training data to perform the segmentation (6).

According to Danelakis et al. (5), the supervised methods have a wide list of classifiers that can be used, and they have a strong accuracy due to the training process. On the other hand, manual segmentation and the training process are time consuming.

The unsupervised methods are faster to set up because no training process is needed, but it's more difficult for them to achieve high accuracy. Moreover, the vast majority of them are built on parametric distributions of signal intensities in structural neuroimaging data.

A recent systematic review on automatic segmentation of WMH (9), shows that at the moment there is no evidence to favour the application of one method rather than other.

BIANCA, explained in detail in the next paragraph, uses the k-Nearest Neighbour (k-NN) classifier, which is a nonparametric procedure for estimation of local class conditional probability density functions from sample patterns (11).

In this method each voxel is treated as a separate sample, and it is associated to a feature space. A list of voxels and therefore features is collected during the training phase.

In the testing phase, the voxel is located in the feature space considering its features, and then is classified according to the k closest training examples in the feature space (10).

The output of the classification is the probability of a voxel being part of a lesion, calculated as the proportion of k neighbours. The algorithm is a supervised method, so a pre-classified dataset is required as training data. The k-NN method requires both high memory capacity for storing the model parameters and long training time.

The k-NN classifier has been already used to segment the MS lesions: Vinitiski et al. (12), Mohamed et al. (13) and Wu et al. (14) have used this method only with MRI features; Mohamed et al. and Vinitiski et al. used T1w, T2w and PDw images, while Wu et al. used T2w, PDw and T1w post-Gd sequences.

Most recently, Steenwijk et al. (15) and Fartaria et al. (16) added spatial information as an additional feature of the k-NN.

Steenwijk et al., after the pre-processing (consisting in brain extraction, RF inhomogeneities correction and linear registration of the sequences), used as features the T1w, the T2 FLAIR, the spatial coordinates in the MNI (Montreal Neuroimaging Institute) space obtained through the registration, and the tissue type probability obtained with FMRIB Automated Segmentation Tool (FAST) of the FSL software. The dataset consisted in 20 MS patients (15).

Fartaria et al. uses the voxel intensity of the MRI sequences MPRAGE (Magnetization Prepared RApid Gradient Echo), T2 FLAIR, MP2RAGE and DIR (Double Inversion Recovery) as features, together with the spatial location in MNI space, obtained from the pre-processing, and the tissue type probability map. The dataset consisted in 39 early-stage MS patients (16).

The k-NN classifier returns the probability about the classification of the element, based on the label of the nearest neighbours. For this reason, the chosen k number affects the final result. In literature, the k number of brain lesion segmentation algorithm is settled between 15 and 100 (10,16). To obtain deterministic measures the probability maps are thresholded and transformed binary maps. The chosen threshold value is of critical importance, since it can affect the final classification. Many values can be found in different articles, threshold values between 0.26 and 0.7 are commonly accepted (2,10,16).

3. Materials and methods

3.1 BIANCA

BIANCA is an algorithm included in the FSL package (FMRIB Software Library, Oxford, UK) to classify white matter hyperintensities. It's a fully automated, supervised method that uses the k-Nearest Neighbour (k-NN) classifier (17).

The already segmented training dataset is divided in White Matter Hyperintensities (WMH) and non-WMH.

BIANCA automatically applies the leave-one-out cross-validation method: a reduced training set is used for the segmentation of a subject from the training dataset, where the reduced training set excludes this subject and is built from the voxels of the remaining training subjects (17).

The main options available in the algorithm are (figure 3):

- Multiple MRI modalities – it is possible to work with many of them, both 2D and 3D. The images need to be registered to a consistent reference MRI modality to allow the algorithm to work in the subject's space.
Intensity normalization using variance scaling (subtraction of the mean from the feature values and division of the outcome by standard deviation (10)) is automatically applied to all the images.
- Spatial weighting (sw) – can be applied to the spatial coordinates obtained after the registration. The spatial information increases the accuracy of the segmentation, since some regions are more likely to be affected by the MS lesions than others. Weighting allows to emphasize or de-emphasize the role of the coordinates, with higher value for spatial weighting leading to the neighbouring feature vectors being more likely to come from similar spatial locations.
If $sw=1$ the data is simply variance normalised, if $sw=0$ the spatial coordinates will be ignored, if sw is large the spatial features have a prior role to define the probability map and the intensity is ignored. To use this option the standard MNI space is needed.
- Patch – it's possible to add an intensity feature containing the local average intensity. It's possible to select one or more patch size, by setting their edge size in voxels (D), and they can be 2D or 3D.
- Number and location of training points – It is possible to select the number of training points belonging to the manual segmented lesion and to the non-segmented image. There are three different option for that:
 - Fixed + Equal (FE), in which a fixed value N is set, and the algorithm uses both for the WMH and the non-WMH;
 - All WMH + Equal (AE), in which all the points classified as WMH are used and an equal number of points is used in the non-WMH;

- Fixed + Unbalanced (FU), in which it's possible to specify different numbers of training points for WMH and non-WMH.

BIANCA allows also to restrict the selection of the location of the points in the non-WMH so that points close to the borders between the two labelled areas are preferentially selected as non-WMH points ("surround" option) or excluded from the training set ("no border" option). The default option considers all the training points inside the brain that are not classified as WMH ("all" option).

When the probability map is obtained, it is possible to threshold it to obtain a binary map. To obtain the proper thresholding it is necessary to try several values and check the consequences on the confusion matrix to see how the False Positive (FP) and False Negative (FN) change.

It is possible to use an exclusion mask to the BIANCA's output, to reduce the FP. These masks can be obtained by the class segmentation of FSL-FAST (17).

At the current date, the algorithm it's still in beta version as part of FSL. In its presenting article (17) it has been tested on a neurodegenerative cohort and a vascular cohort.

3.2 Datasets

The dataset used throughout the training and optimization process comes from the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2016 challenge. It is composed of 15 MR studies of MS patients, acquired in three different scanners (5 Philips Ingenia 3T, 5 Siemens Aera 1,5T, 5 Siemens Verio 3T). For each patient five sequences are provided: a 3D T2w, a 3D FLAIR, a 3D T1w, a 3D T1w post Gadolinium injection and a 2D DP/T2w, all given as raw data in NIFTI (Neuroimaging Informatics Technology Initiative) format. Moreover, for each patient seven manual segmentations and a consensus segmentation made from the manual segmentations are available.

It is possible to also access pre-processed images (18), but it has been chosen to work with the raw dataset, in order to develop and optimize the ideal pipeline to work with BIANCA.

In order to test the optimized algorithm, the public dataset of Lesjak et al. was used (19). It consists of 30 MR studies of MS patients acquired on 3T Siemens scanners using 2D T1w, 2D T1w post Gd injection, 2D T2w, 3D FLAIR sequences. Moreover, the consensus white matter segmentations are given. Also, in this dataset it is possible to access pre-processed images, but we choose to apply the pipeline used for the previous dataset.

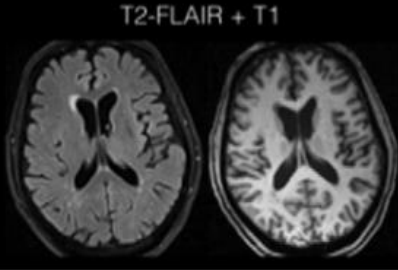
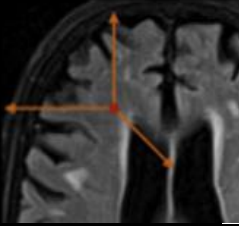
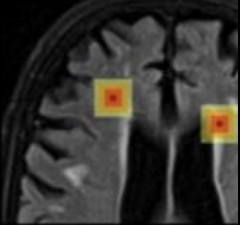
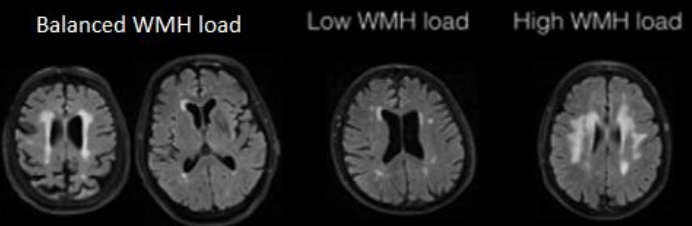
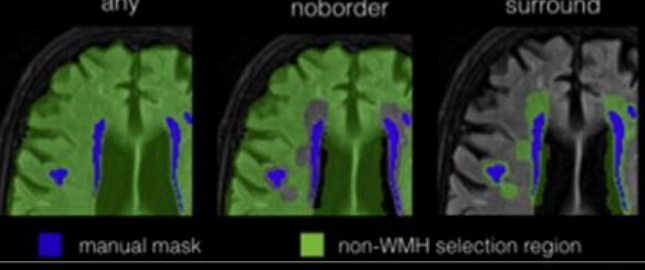
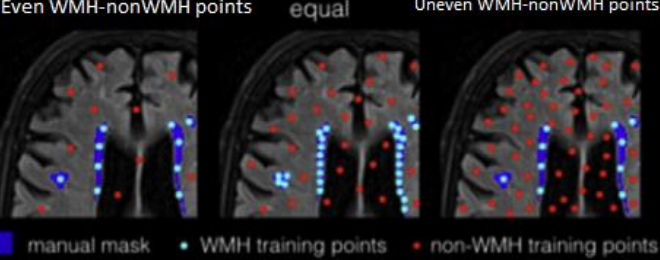
BIANCA OPTION	DESCRIPTION	OPTION TESTED	SCHEMATIC REPRESENTATION
MRI Modality	Intensity features	T1+FLAIR	
Spatial Weighting	Spatial coordinates	0 – 5	
Patch	Local Intensity Averages	2D and 3D, 1 - 10	
Lesion Load	Subject included in the training dataset in terms of WMHs	High, Low, balanced, Leave one out	
Location of training points	Where to select the non lesion area	Any, noborder, surround	 <p> ■ manual mask ■ non-WMH selection region </p>
Number of training points	Maximum number of training points to use	Even 10k-100k, Uneven 2k/10k-10k/50k - equal	 <p> ■ manual mask ● WMH training points ● non-WMH training points </p>

Figure 3: BIANCA options (17).

3.3 Pre-processing

The first step of the pre-processing requires a DICOM (Digital Imaging and COmmunications in Medicine) to NIfTI conversion to make the data compliant with the FSL suite. After that, two additional steps are required: the brain extraction, needed in at least one MR modality, and the linear registration between modalities and with a reference, to be able to use spatial information (figure 5).

The computer used for this work is a personal computer (Intel i7 2.20GHz x 8 core CPU, operating system Ubuntu 18.04.5 LTS, RAM 8GB).

3.3.1 Brain Extraction Tool

Removal of the non-brain structures is performed using the Brain Extraction Tool (BET) (20) included in the FSL package.

This method estimates the lower and upper threshold between brain and background on a histogram-based heuristic. From the pixels pre-classified as brain, a rough Centre of Gravity (COG) is estimated. A tessellated sphere is then generated centred in the COG with a radius set half of the estimated brain radius. The final shape is obtained by iteratively subdividing each triangle in smaller ones and adjusting each vertex.

Since some of the data available include also neck in the field of view, to have a better estimation of the centre of the head, an automatic cropping of the Field Of View (FOV) was done using the FSL tool *robustfov* (20).

3.3.2 Linear Registration

The chosen algorithm is FMRIB's Linear Image Registration Tool (FLIRT) (21,22), included in the FSL package, which is a fully automated method for linear inter and intra-modal brain image registration. The algorithm is based on the optimization of a cost function to maximize the similarities between the floating image and the reference image. To do so it uses a multiresolution local registration called repeatedly, preceded by an initial search that is focused on the rotational part of the transformation space (21,22).

To ensure a consistent localization of the spatial features in BIANCA, all the images were transformed to a standard MNI space, using a 1mm resolution template provided with FSL. For the registration, the default settings were used, as suggested in the FLIRT User Guide (23).

It was decided first to work at a coarse resolution to reduce the number of voxels to classify and thus the running time. Once the pipeline was optimized on the MICCAI dataset at 2mm resolution, it was applied to the higher resolution dataset (1mm isotropic voxel size) to improve both resolution and accuracy of the classification.

With MNI1mm registration the processing time is about doubled when compared with the MNI2mm registration.

3.4 BIANCA optimization, training and test

The input data used for the algorithm are the brain extracted 3DT1w images, the binary brain masks obtained with BET, the 3D FLAIR images and the consensus lesion masks.

BIANCA default options are $sw = 1$, no patch, location of training points = any, number of training points = fixed + equal (2000 lesion and non-lesion points).

The features to be optimized are: lesion load distribution, thresholds, spatial weighting, use of patches, location and number of training points.

These parameters were tested independently starting from the default options, and then the combination of the best option was applied.

To test the different features, except for lesion load, all the subjects have been used as part of the training dataset thanks to the leave-one-out characteristic of the algorithm that automatically excludes the voxels of the testing subject from the classifier.

3.4.1 Thresholding

The thresholds have been tested with values ranging from 0,1 to 1 in steps of 0,1. The next measures have been carried out using the threshold with the best evaluation measurements.

On the probability maps obtained with the optimized options, a spatially optimized thresholding was tested, LOcally Adaptive Threshold Estimation (LOCATE, algorithm explained in more detail in the post processing paragraph 3.5) to obtain the best results.

3.4.2 Spatial weighting

The spatial coordinates feature weight was tested from 0 to 5 with a step of 0,5, using the best threshold and all the other features set to default.

3.4.3 Patches

The intensity local average was tested from 1 to 10, both 2D and 3D.

3.4.4 Location of training points

The available option considered for location of the non-lesion points are: any, without the borders of the lesions, and only the surroundings of them.

3.4.5 Number of lesion and non-lesion points

The default option is an equal number of lesion and non-lesion points, and this option was tested with values from 1000 to 20000 with a step of 1000 and then from 25000 to 100000 with a step of 5000. This was done because in the first optimization, with the linear registration to MNI2mm, the best result was obtained with 4000 lesion and non-lesion points. The optimization with the improved resolution required higher values, so the step was moved from 1000 to 5000.

Moreover, the option that considers all the lesion areas and an equal number of non-lesion areas (equal points option) was tested.

As a final step, the unbalanced option was tested with value of lesion points from 2000 until 10000 with a step of 2000 and non-lesion points from 10000 until 50000 with a step of 10000.

3.4.6 Lesion load

The lesion load in MS patients is highly variable. In our dataset, the load spans a range of $[4 \times 10^3, 167 \times 10^3]$ voxels, with a mean lesion load of 58×10^3 voxels (figure 4). It is known that the load influences the reliability of the results, giving higher performances with higher lesion load (17) so the training dataset was divided in three different groups, in order to obtain a high, low and balanced lesion load with respect to the average value. This was carried out with the default option of the algorithm and the optimal threshold, with images co-registered to MNI2mm.

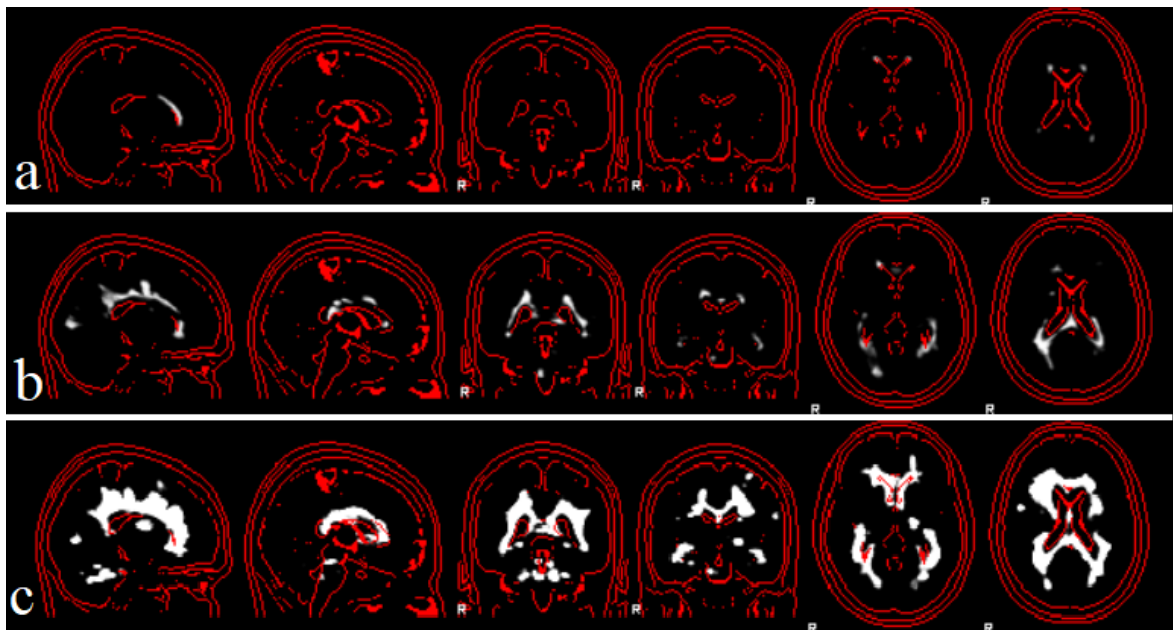


Figure 4: Different lesion load of dataset cases: a) minimum load, b) intermediate load, c) maximum lesion load. In red it is shown the brain outline of the MNI template and in white the WMH binary manual mask.

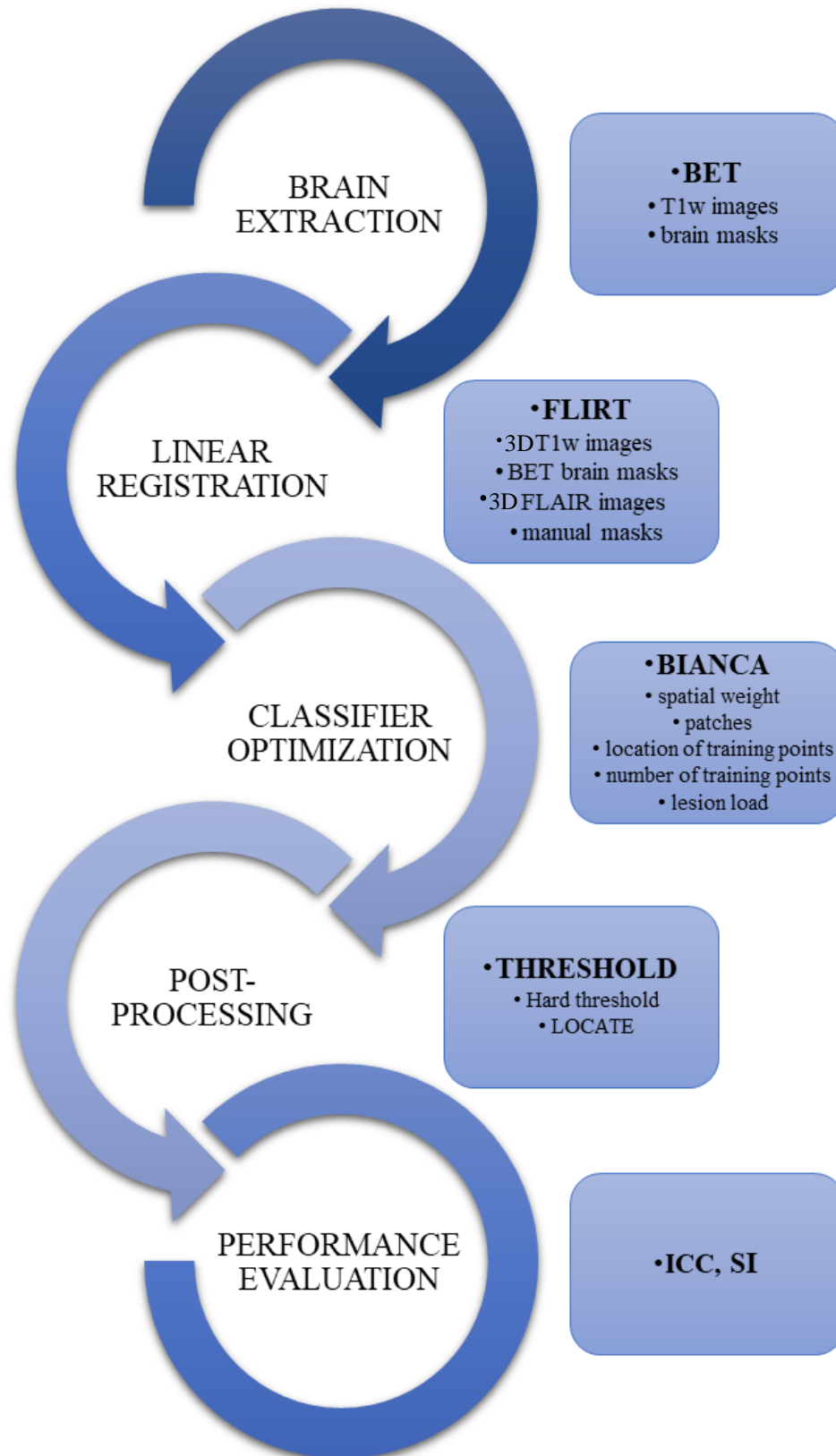


Figure 5: Pipeline of this work: the pre-processing consists in brain extraction and linear registration, performed with BET and FLIRT. Then the classifier, BIANCA, is trained and its options optimized. The probability maps obtained are thresholded, using hard thresholds and LOCATE, a spatially optimized threshold, and the binary maps are evaluated through ICC and SI.

3.5 Post-processing and performance evaluation

The probability maps obtained by BIANCA need to be thresholded to obtain binary maps to be compared to the manual lesion masks.

To do so, FSL offers an automatic tool, `BIANCA_overlap_measures`, that returns a number of metrics described as follows:

- Dice Similarity Index (SI):

$$SI = \frac{2 \times (\text{manual mask} \cap \text{BIANCA mask})}{\text{manual mask} + \text{BIANCA mask}}$$

- Voxel-level false detection rate (FDR):

$$FDR = \frac{\text{false positives}}{\text{total number of voxels labelled as lesion in BIANCA mask}}$$

- Voxel-level false detection rate (FNR):

$$FNR = \frac{\text{false negatives}}{\text{total number of voxels labelled as lesion in manual mask}}$$

- Cluster-level FDR:

$$clFDR = \frac{\text{clusters incorrectly labelled as lesion}}{\text{total number of lesion in BIANCA mask}}$$

- Cluster-level FNR:

$$clFNR = \frac{\text{clusters incorrectly labelled as non lesion}}{\text{total number of lesion in manual mask}}$$

- Mean Total Area (MTA):

$$MTA = \frac{\text{true voxels} + \text{positive voxels}}{2}$$

- Detection error rate (DER):

$$DER = \frac{FP + FN}{MTA}$$

- Outline error rate (OER):

$$OER = \frac{\text{voxels of true positive clusters of manual and BIANCA masks without the overlapped one}}{MTA}$$

- Volume of BIANCA segmentation (after applying the specified threshold)
- Volume of manual mask

For the thresholding, BIANCA User Guide (24) suggests an alternative algorithm based on MATLAB, named LOCATE.

This method estimates local thresholds in BIANCA's lesion probability map by segmentation with Voronoi tessellation, extraction of local features and estimation of the optimal local threshold using a supervised learning method.

The processing starts by identifying local maxima on Gaussian filtered probability map to avoid spurious fluctuation due to isolated voxels. Then the map is tessellated into Voronoi polygons and different thresholds are applied considering the mean intensity value in the corresponding area of the base image, the distance between ventricles (optional) and the volume of the thresholded region. The optimal local threshold for each area is obtained using a random forest regression model (25). This method was applied to the probability maps obtained with the best combination of all the parameters (figure 6).

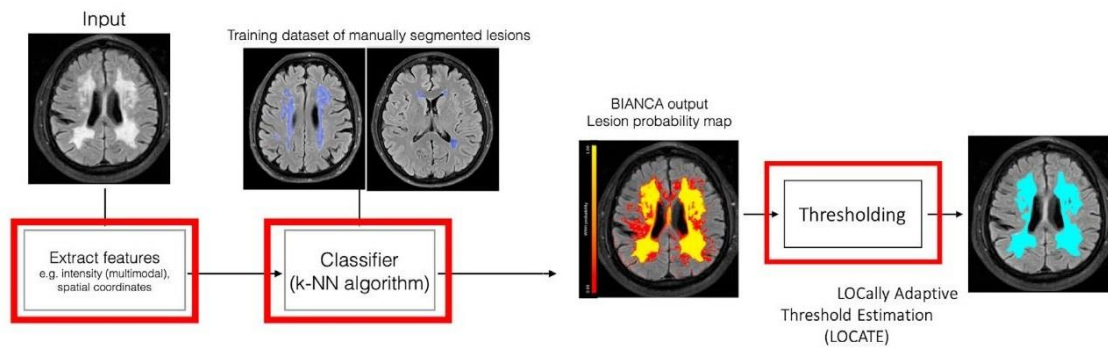


Figure 6: Application of LOCATE to improve BIANCA performance (25).

4. Results

Each option was evaluated considering both the Similarity Index (SI) obtained with BIANCA Post-Processing and the Intra-Class Correlation coefficient (ICC), obtained using MATLAB (release R2020a) and estimated with 95% confidence intervals based on 2-way mixed effects model, considering the consistency and single rater (ICC (3,1)) (26), between the total lesion volume from BIANCA output and the one from the manual masks. The best combination was selected mostly considering the ICC value (17). In fact, higher ICC values indicate a better correlation and concordance between the two measures, which are the number of voxels labelled as lesion by BIANCA and of those obtained from the manual masks.

The trained and optimized classifier was then tested with Lesjak et al. dataset. Before the application of BIANCA, the pipeline of pre-processing described in paragraph 3.3 was applied.

4.1 Threshold

The best resulting threshold (figure 7) is 0.9, with an ICC=0.70 and a SI=0.43. All the other optimizations used this value to binarize the probability mask obtained by BIANCA. The running time of this optimization was about 10 hours.

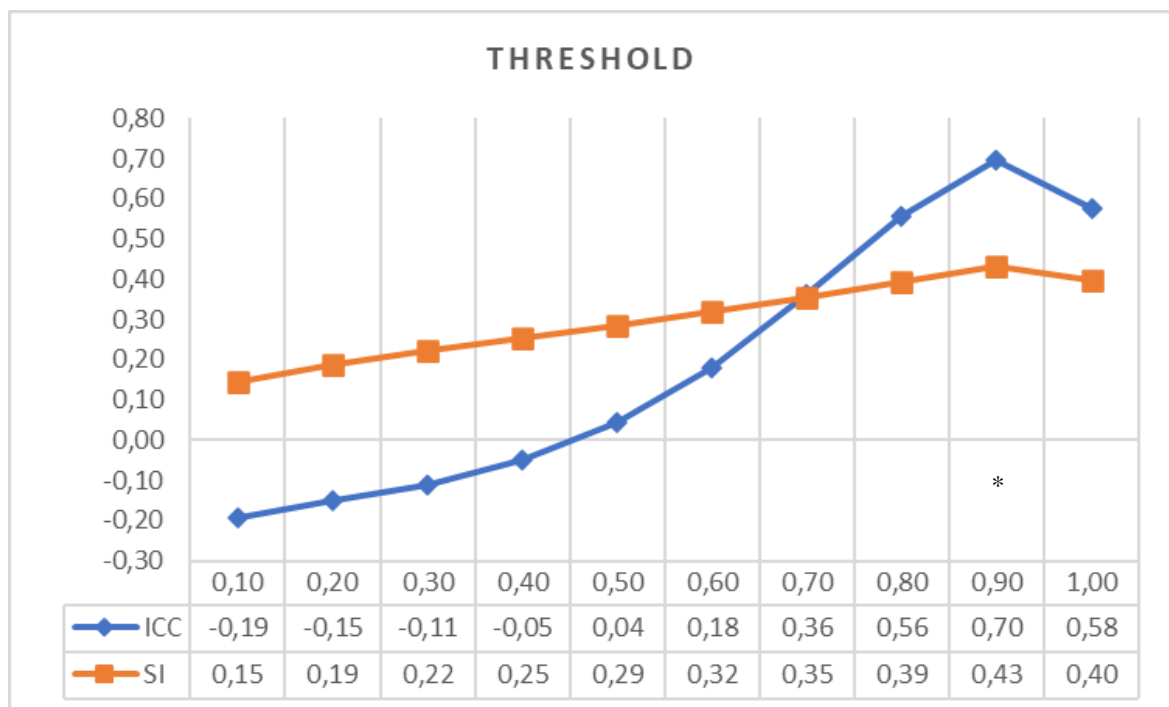


Figure 7: Distribution of the SI and ICC with the application of different thresholds. The “*” indicates the option selected for the optimization. ICC and SI values are rounded to the second decimal place.

4.2 Spatial weighting

For the weighting of the spatial features (figure 8), similar results were obtained with 1 and 1.5, however, 1.5 was chosen given the comparable absolute values of SI and ICC. This optimization took about 12 hours.

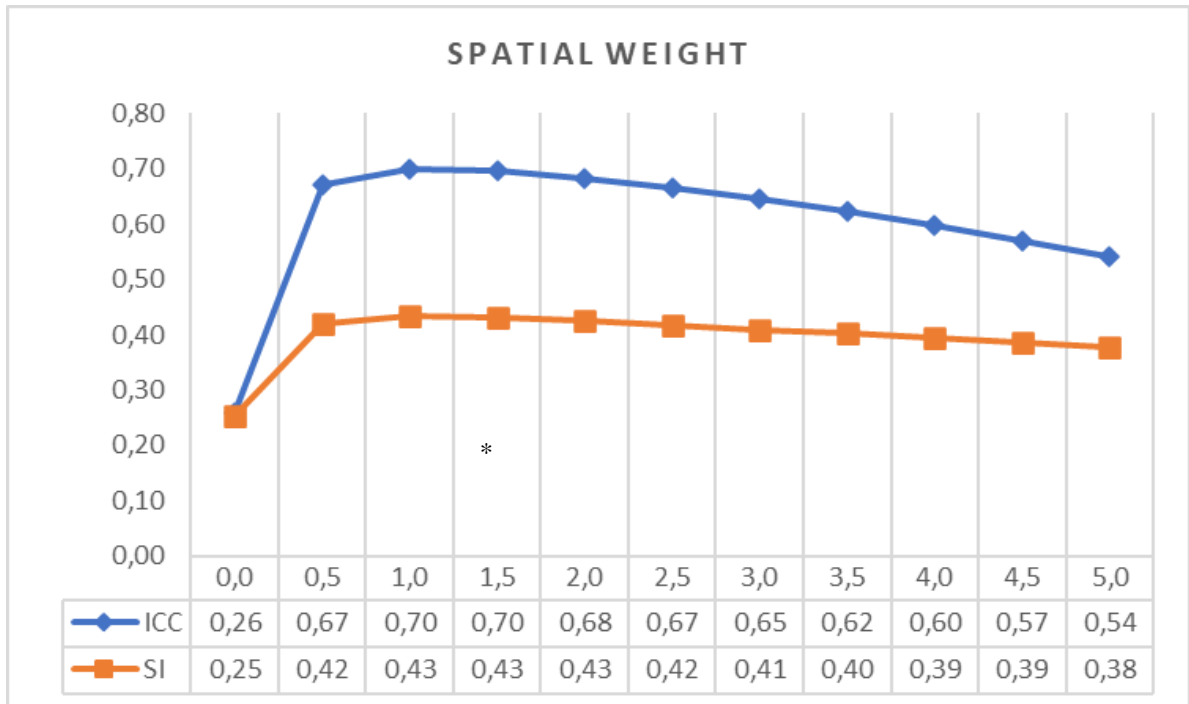


Figure 8: Distribution of the SI and ICC with the application of different values of spatial weighting. The “*” indicates the option selected for the optimization. ICC and SI values are rounded to the second decimal place.

4.3 Patches

In the patches optimization (figure 9), the mean SI obtained with 2D and 3D patches were similar. The best result was obtained with 3D patch=4. This optimization took about 16 hours.

4.4 Location of training points

Similar results were obtained using all the non-lesion points and without the area surrounding the lesions (figure 10). The option considered the best one was “no border”, with a mean SI=0.44 and an ICC=0.72. This optimization took about 6 hours.

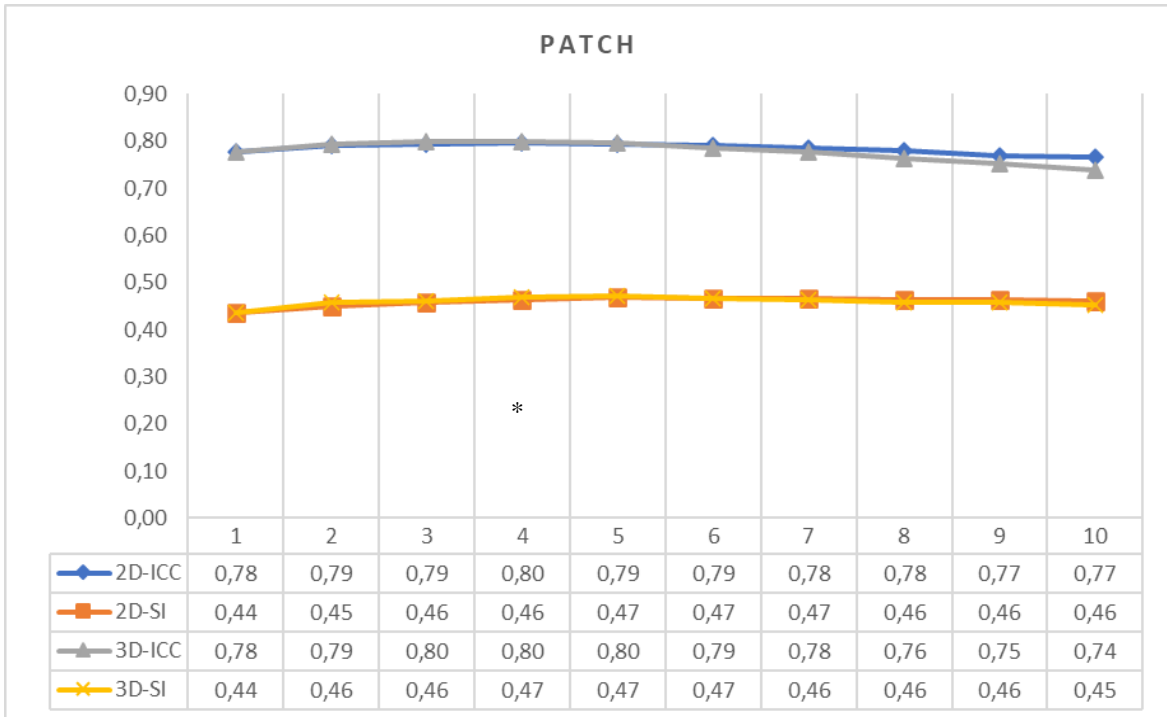


Figure 9: Distribution of the SI and ICC with the application of different patches, 2D and 3D. The “*” indicates the option selected for the optimization. ICC and SI values are rounded to the second decimal place.

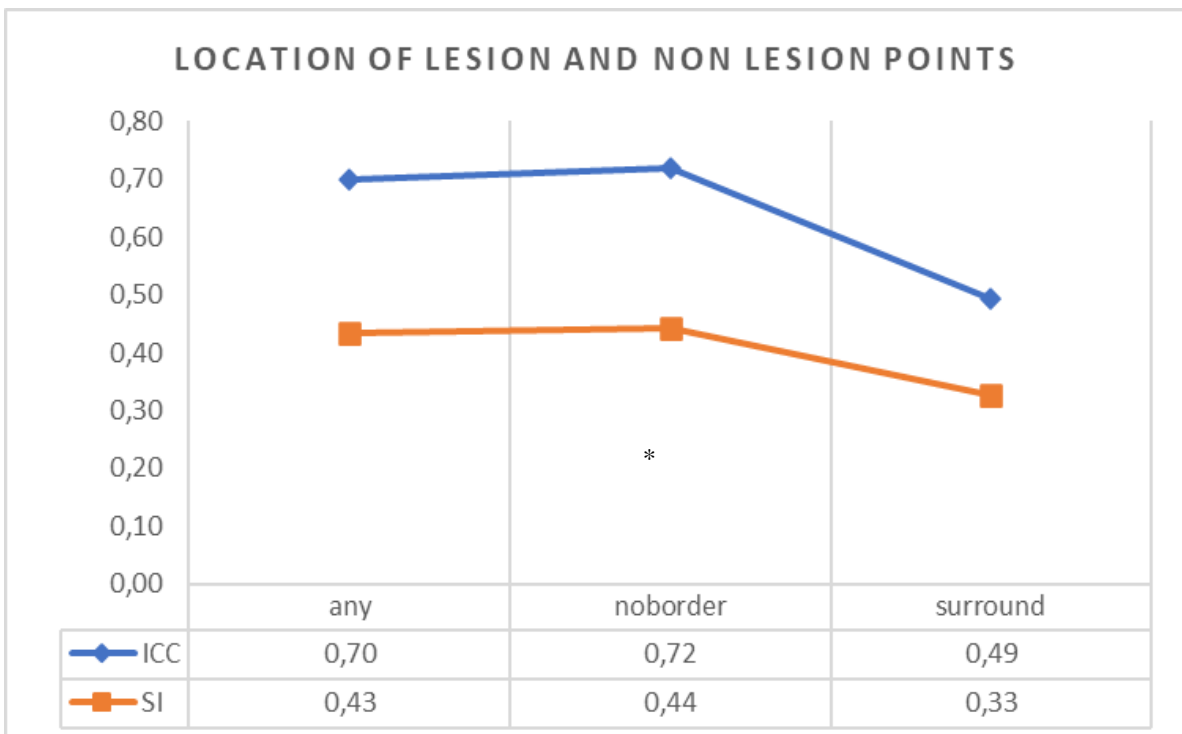


Figure 10: Distribution of the SI and ICC with the application of different option regarding the location of non-lesion points. The “*” indicates the option selected for the optimization. ICC and SI values are rounded to the second decimal place.

4.5 Number of lesion and non-lesion points

An equal number of lesion and non-lesion points to train the classifier gives better correlation and better mean SI (figure 11).

The option chosen is 60000 because the absolute value of both SI and ICC are the highest. The option with an unbalanced number of lesion and non-lesion points returned lower correlation. All the optimization of this option took about 48 hours.

4.6 Lesion load

The use of a training set with lesion load lower than the test set (mean training load= 64×10^3 , mean test load= 205×10^3), yield higher mean SI (0.53) but lower ICC (0.2) (figure 12).

The use of a higher load in the training set than the test one (mean training load= 111×10^3 , mean test load= 20×10^3), results in high ICC (0.74) and a low mean SI (0.16).

Balancing the loads in the training and test sets (mean training load= 93×10^3 , mean test load= 92×10^3) return a mean SI=0.37 and ICC=0.72.

The best combination of ICC and SI was considered to be the latter. Even if this option was tested, the division of the training subject in three groups wasn't used for the final optimization. Instead, the leave one out option was used, as it allows for a higher number of subjects included in the training set that ultimately lead to a more robust classifier.

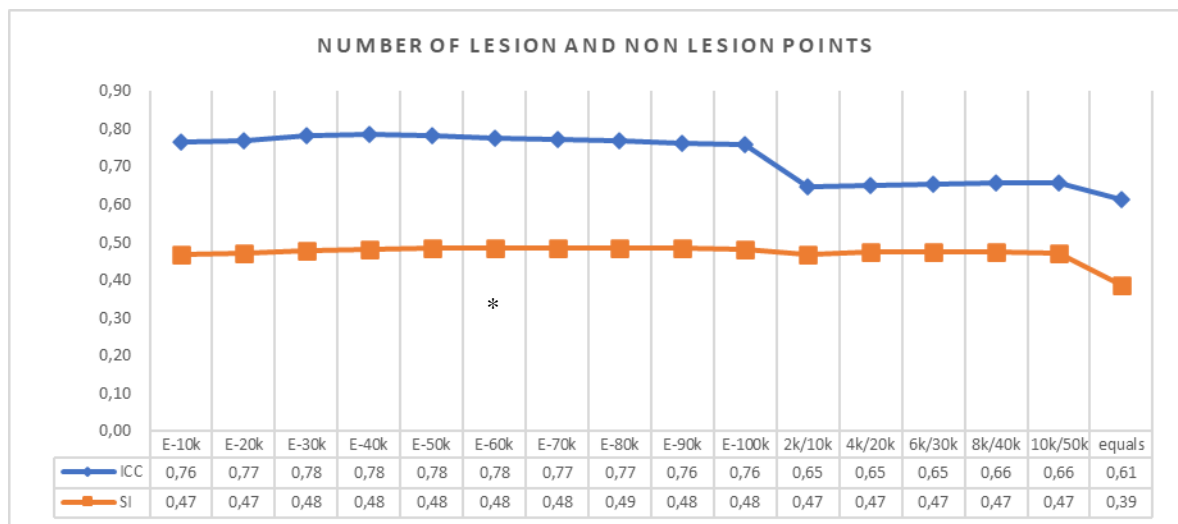


Figure 11: Distribution of the SI and ICC with the application of different number of lesion and non lesion points. The “*” indicates the option selected for the optimization. ICC and SI values are rounded to the second decimal place.

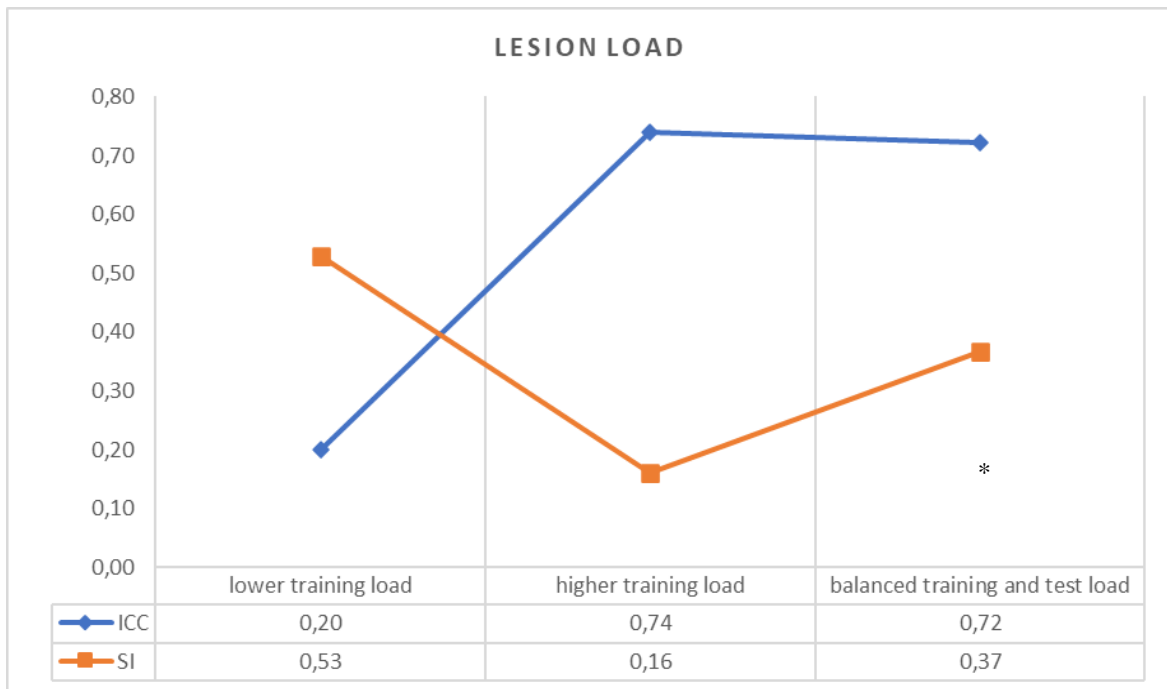


Figure 12: Distribution of the SI and ICC with different lesion load. The “*” indicates the option selected for the optimization. ICC and SI values are rounded to the second decimal place.

4.7 Optimization

The options used (table 4) in the final optimization are $sw = 1.5$, 3D patch = 4, location of training points = no border, number of training points = fixed + equal (60000 lesion and non-lesion points). The comparison has been done (figure 13) considering the SI and ICC in three different cases: the default options with default threshold = 0.9, the optimized probability map thresholded with the optimized threshold = 0.9, and finally the optimized probability map thresholded using LOCATE.

The best results were obtained using LOCATE, with an improvement of ICC from 0.85 to 0.92 when comparing the probability map thresholded with a hard threshold and with LOCATE.

In figure 14 there are three examples of the binary maps obtained with this optimization.

OPTIONS	CHOICES
Spatial weighting	1,5
Patch	3D – 4
Location of non-lesion points	No border
Number of training points	Fixed + equal – 60000
Threshold	LOCATE

Table 4: Optimization choices for BIANCA options.

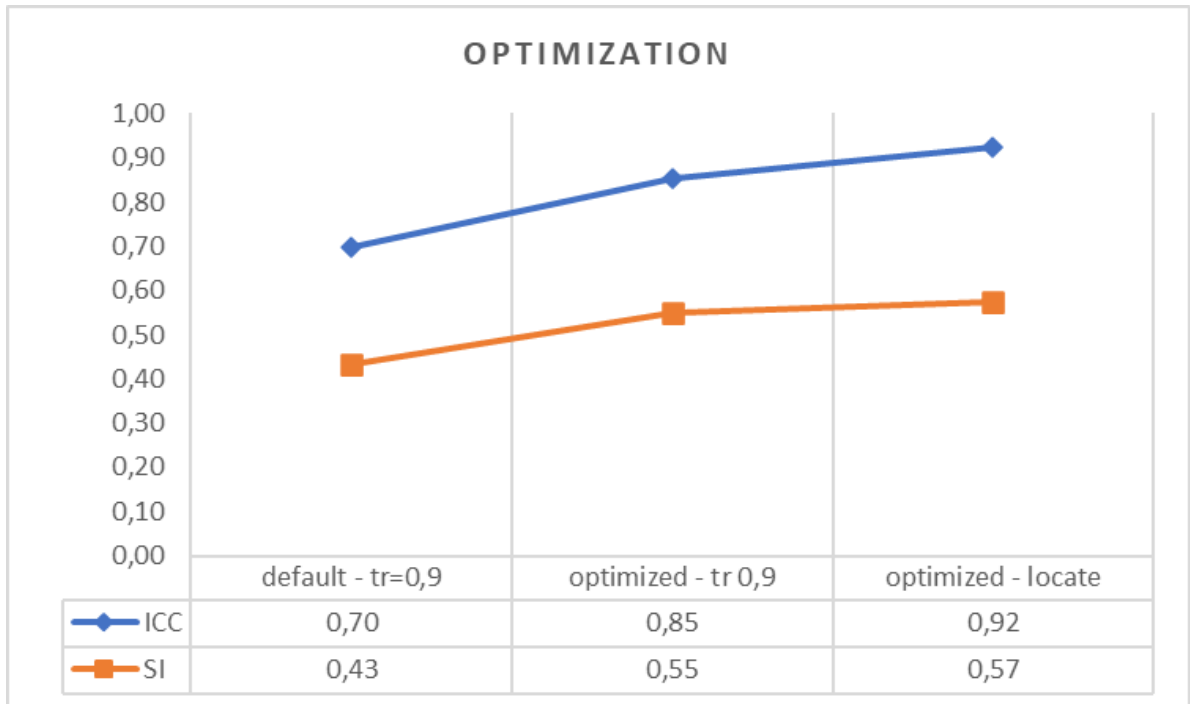


Figure 13: Distribution of the SI and ICC of the optimized dataset with the application of threshold=0.9 (first column) and using LOCATE (second column). The third column are the ICC and SI of the dataset with the application of default BIANCA options and a threshold of 0.9. ICC and SI values are rounded to the second decimal place.

4.8 BIANCA test results

The 30 patients of the test dataset were evaluated using the optimized classifiers trained with the MICCAI dataset. The options used are sw=1.5, 3D patch=4, location of training points = no border, number of lesion and non-lesion points = Fixed + Equal with 60000 points, LOCATE to create the binary map.

The results on the test dataset are ICC=0.27 and SI=0.26 (table 5).

Table 5: Comparison between SI and ICC of the training and testing dataset using the optimized options.

	SI	ICC
Training dataset	0.57	0.92
Testing dataset	0.26	0.27

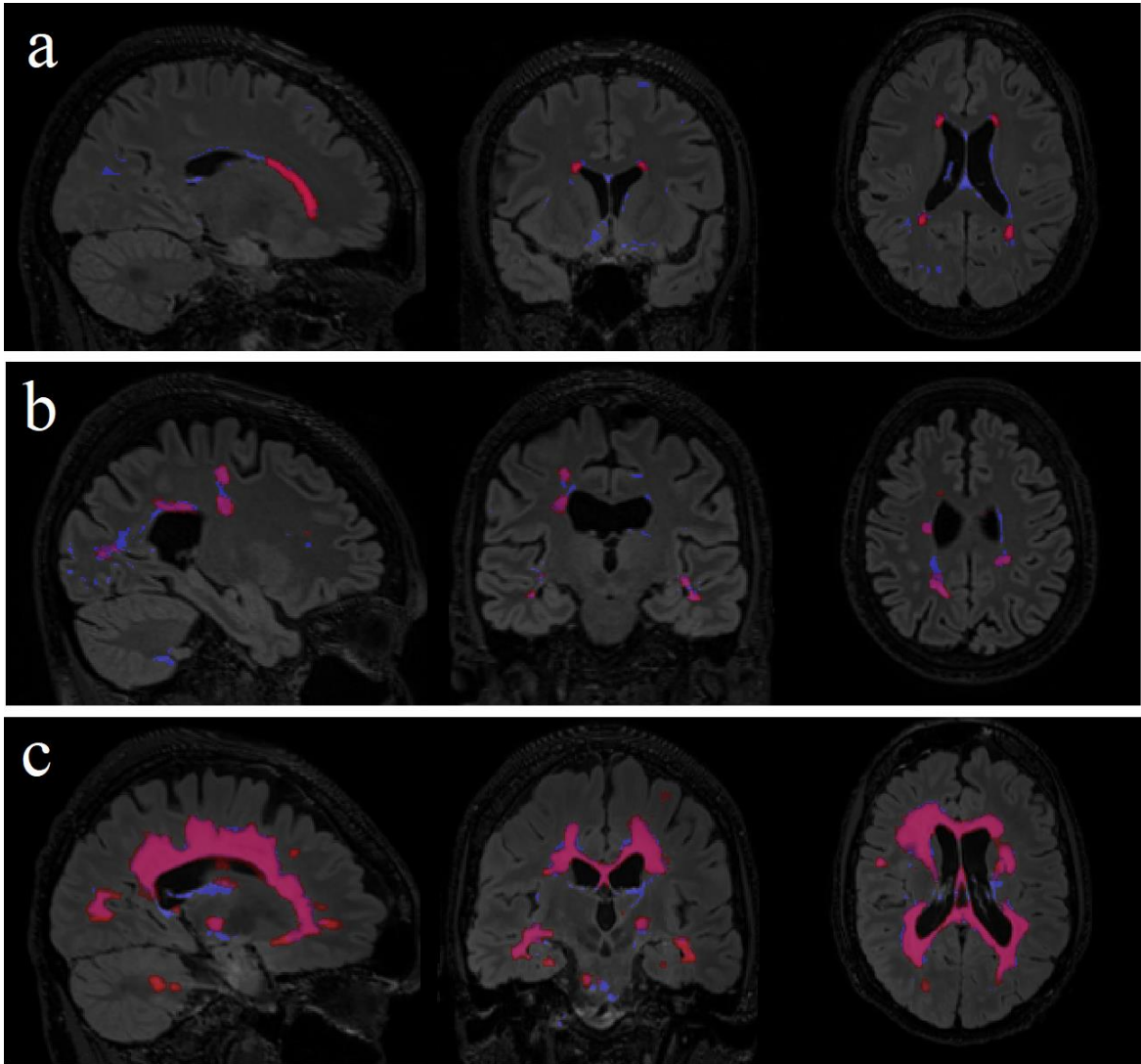


Figure 14: Three different examples on BIANCA results over FLAIR images: a) low lesion load, b) intermediate load, c) high lesion load. In red are shown the manual masks, in blue the LOCATE thresholded binary maps. The pink area is the overlap of the manual mask and BIANCA result.

5. Discussion

In this work, an algorithm developed to perform automatic segmentation of white matter hyperintensities using a k-NN classifier, BIANCA, was tested in an open MS dataset.

The dataset used for the optimization and training was taken from the MICCAI 2016 challenge and consists of 15 patients acquired with three different MRI scanners, both at 1.5 T and 3T. For each patient, only the 3DT1 and FLAIR images were considered, starting from the raw files to create a pipeline of pre-processing that could be considered for clinical application. To do so, FSL tools (BET, FLIRT) were used. With the pre-processed images, the tuning of BIANCA consisted in an iterative optimization of five main parameters: spatial weighting, uses of intensity patches, number and location of training points, optimization of the threshold and the effect of lesion load on BIANCA output. Moreover, the robustness of the manual thresholding was compared against a data-driven local-thresholding algorithm, LOCATE.

After the optimization, the Lesjak et al. public dataset, consisting of 30 subjects acquired at 3T, was used to test the classifier.

It was decided to not use exclusion masks for the variable distribution of MS lesions (3). To evaluate the output, two indices were considered: SI and ICC. Both the metrics evaluate the correspondence between WMH-maps found by BIANCA and the manual masks, from a spatial and quantitative point of view, respectively. The joint evaluation of the two indices is necessary to explain the performance of the segmentation.

The best combination was obtained with a spatial weighting of 1.5, 3D intensity patches with $D=4$ and with a fixed number of 60×10^3 training points in the lesion area and an equal number of the non-lesion one, avoiding the edges of the latter. The probability map obtained with the previous parameters and LOCATE for spatially optimized thresholding, proved to return the best results, with a $SI=0.57$ and an $ICC=0.92$.

BIANCA is based on k-NN classifier, that builds a voxel-wise feature space. This characteristic allows the leave-one-out option, that automatically excludes all the voxels belonging to the testing subject from the training set, so that the optimization and the test can be done with the same dataset not influencing the final result. The optimization process of this work has been done using this option.

With the optimized probability maps and spatially optimized thresholding, it is possible to obtain the plot of mean SI as a function of lesion load, reported in figure 15. In the plot, it is possible to notice how the SI increases with the increase of the lesion load, with a logarithm as the best fitting curve, as it is for the MICCAI challenge results. A possible hypothesis for this could be a threshold-dependent performance of the ML methods, where sparse lesion distributions lead to a suboptimal classification. This confirms that the lesion load of the dataset has an influence of the final result of the classification (17).

The dataset used in this work is the training dataset used in the MICCAI 2016 Challenge whose results are publicly available. In table 6 are summarized the results of all the teams that successfully concluded the challenge (18,27,28).

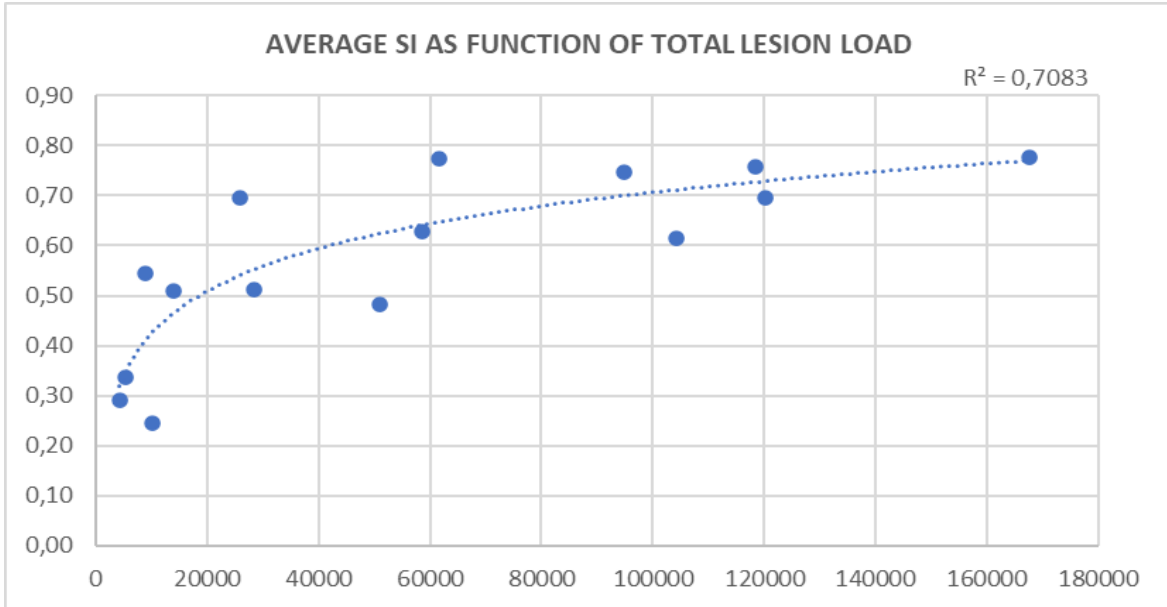


Figure 15: Distribution of mean SI, obtained with the optimized BIANCA options, as function of the lesion load.

Table 6: Participants to the MICCAI 2016 Challenge, their segmentation approach, the sequences they used and the average SI they obtained (18,27,28).

Authors	Segmentation Approach	Sequences used	Average SI
R. McKinley T. Gundersen	Ensemble of three 2D fully Convolutional Neural Networks with skip connections	FLAIR (pre-processed)	0.591
E. Roura X. Lladó	Outlier segmentation based on brain tissue labelling and post-processing rules	T_1 -w, FLAIR (raw)	0.572
S. Valverde M. Cabezas	Cascade of two 7-layer convolutional neural networks of 3D patches	T_1 -w, T_2 -w, PD, FLAIR (pre-processed)	0.541
F.J. Vera-Olmos N. Malpica	Grey matter filter as input to a RF classifier corrected with Markov Random Field processing	T_1 -w, T_2 -w, PD, FLAIR (pre-processed)	0.521
J. Knight A. Khademi	Segmentation by edge-based model of partial volume/pure tissue grey levels	FLAIR (raw)	0.490
J. Beaumont O. Commowick	Multi-modal abnormalities detection from normalized images on an atlas	T_1 -w, T_2 -w, FLAIR (pre-processed)	0.485
S. Doyle F. Forbes	HMRF segmentation framework with a weighted data model	T_1 -w, FLAIR (raw)	0.489
J. Beaumont O. Commowic	Graph cut segmentation initialized by a robust EM	T_1 -w, T_2 -w, FLAIR (pre-processed)	0.453
A. Mahbod C. Wang	Supervised artificial neural network with intensity and spatial based features	FLAIR (pre-processed)	0.430
H. Urien I. Bloch	Hierarchical segmentation using max-tree, spatial context and anatomical constraints	T_1 -w, T_1 -w Gd, T_2 -w, PD, FLAIR (raw, pre-processed)	0.347
M. Santos A. Silva-Filho	Multilayer perceptron with cost functions oriented to competition evaluation metrics	T_1 -w, T_2 -w, FLAIR (pre-processed)	0.340
J. Muschelli E. Sweeney	Random Forest (RF) on normalized multi-modal features	T_1 -w, T_2 -w, PD, FLAIR (raw)	0.341
X. Tomas-Fernandez S.K. Warfield	Lesions and brain tissue segmentation through simultaneous estimation of spatially and population varying intensity distributions	T_1 -w, T_2 -w, FLAIR (raw)	0.228

The MICCAI 2016 training dataset was composed of 15 training subjects, considered also in this work, and n=38 cases to test the algorithm.

Although it would be important to test our optimized BIANCA pipeline on the MICCAI testing dataset, it wasn't possible (since it is not publicly available). Anyway, the results obtained are in good agreement with the results reported by the 13 teams that successfully concluded the challenge (figure 16).

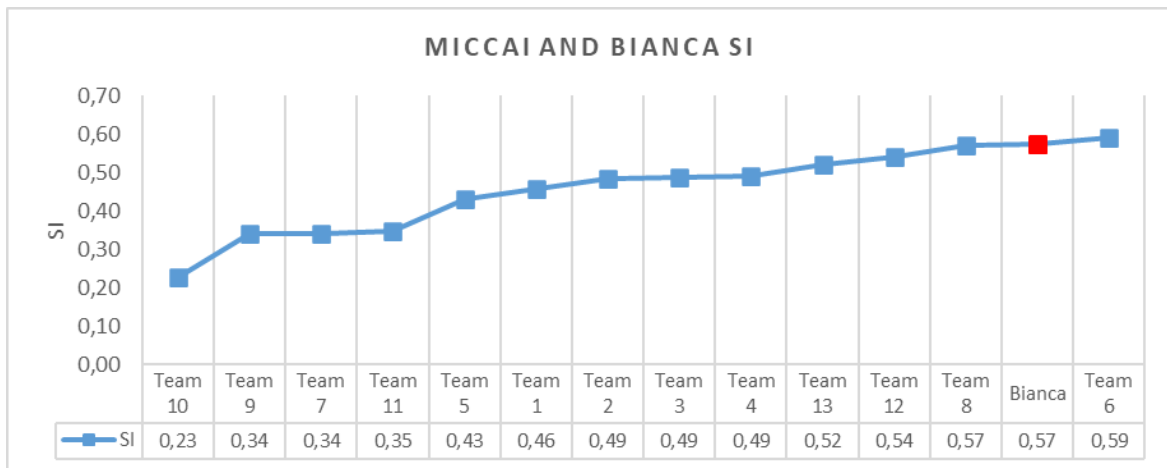


Figure 16: Distribution of SI obtained with the MICCAI challenge dataset by BIANCA and the challenge participants.

It's worth to notice that those results share with ours the same dataset however the test set is different. Therefore, our results are not strictly comparable with the challenge ones.

Different conditions were tested in the definition of the training set, trying to balance the lesion load between training and test subjects. All these combinations lead to sub optimal results, thus corroborating the hypothesis that a larger sample with leave-one-out could be an effective way to produce a robust training procedure.

On the other side, the test of the optimized classifier carried out on the Lesjak et al. public dataset returned SI=0.26 and an ICC=0.27.

The highly different results obtained with Lesjak et al. testing dataset can be due to a number of reasons.

First, in the testing dataset the sequences acquired for each patient consist of 2D T1w and 3D FLAIR, while in the training dataset the sequences are 3D T1w and 3D FLAIR. In our classifier the spatial feature and the intensity features are taken both from T1w and FLAIR images. Even if the images are all co-registered to MNI1mm, the spatial resolution of the testing dataset is lower than the training dataset, and this may influence the precision of the results.

The training dataset is heterogeneous, in terms of scanner properties: 5 patients acquired with a 1.5T Siemens Aera, 5 patients acquired with a 3T Siemens Verio and 5 patients acquired with a 3T Philips Ingenia. On the other hand, the testing dataset (n=30) was entirely acquired on a 3T Siemens Magnetom Trio. In figure 17 are reported the SI results of the MICCAI dataset, distributed for lesion load and scanner type. It is possible to observe that the 1.5T acquisition provides slightly better results compared to 3T scanners,

given the same lesion load. This could be explained by the expected higher sensitivity of the 3T MR in detecting smaller lesions, that could possibly result in a baseline shift between the two field strengths.

Moreover, there is a difference in the results obtained with the two high field scanners.

The two groups have a different image resolution: in the 3T Siemens T1 voxel size is 1x1x1 mm, while in the FLAIR is 1.1x0.5x0.5 mm, in the 3T Philips T1 voxel size is 0.85x0.74x0.74 mm, while in the FLAIR images is 0.7x0.74x0.74 mm.

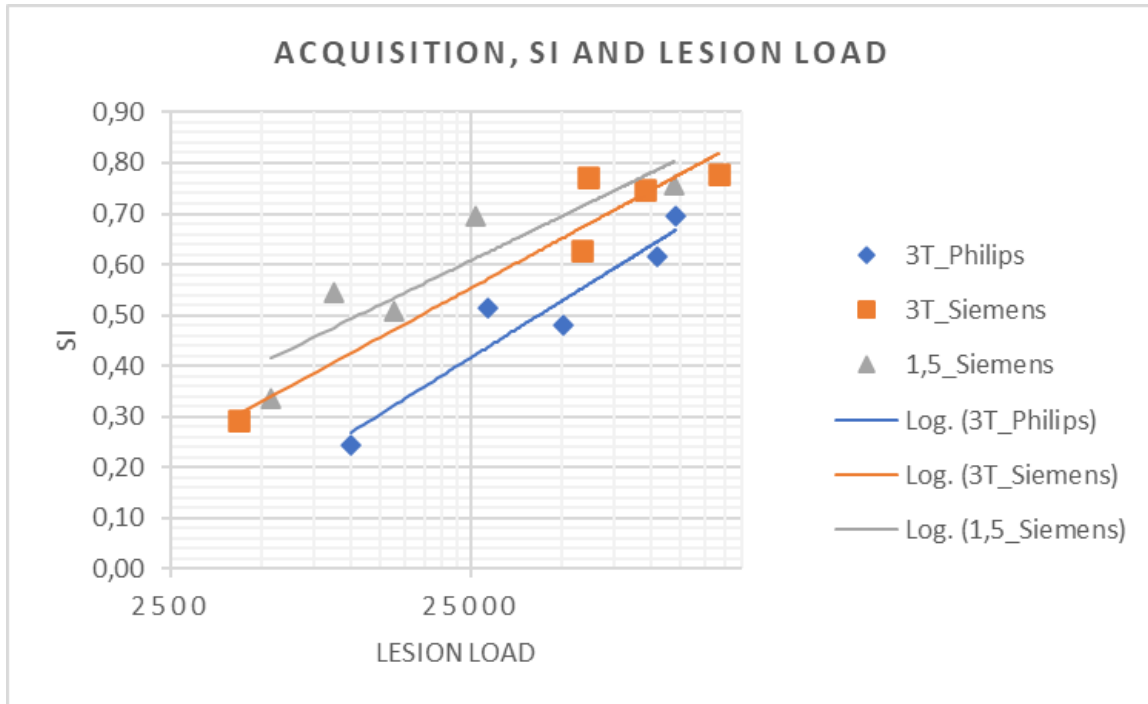


Figure 17: SI MICCAI challenge dataset results as function of the lesion load (in logarithmic scale) divided per acquisition scanners.

The distribution of lesion load in the two dataset is shown in the box plot in figure 18. The MICCAI dataset has a mean lesion load higher than the Lesjak et al. dataset. Even if the second quartile is similar, the third quartile is higher in the MICCAI dataset, meaning that the distribution of the lesion load of the Lesjak dataset is skewed towards lower values. Since the lesion load influences the final result of the classifier, the different distribution of lesion load also contributes to the different results obtained with the two datasets.

Publicly available MS datasets are rare, so it wasn't possible to obtain a wide number of cases to train and test BIANCA without the differences we had between our sets. Future works could be focused on studying the impact of different sequences, different fields and different scanners on the training and test of this classifier.

Even with differences between the training and testing datasets, the results obtained with Lesjak et al. dataset are included in the range of results obtained in the MICCAI challenge, where the testing dataset is composed of images acquired with the same scanners and the same protocols used in the training set.

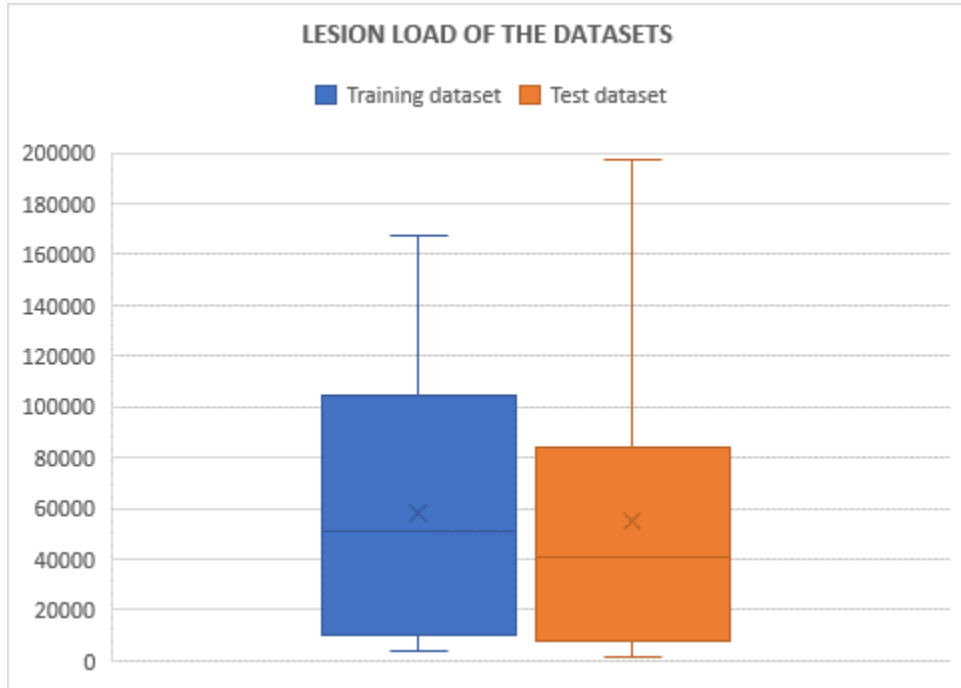


Figure 18: Boxplot of the lesion load of the two dataset used: in blue the MICCAI challenge dataset, used for the training of the classifier, and in orange the Lesjak et al. dataset, used for the test of the classifier.

The processing time required by the optimization and training of the classifier is highly influenced by the computational power available, also considering that only few steps of the FSL pipeline could benefit from advanced computational platforms such as parallel or GPU accelerated systems.

Anyhow, the optimized pipeline itself showed interesting results that suggest a possible application on different datasets.

6. Conclusion

The aim of this work was to design and optimize a workflow to automate WMH lesion detection with a ML approach.

The optimization was conducted on 15 cases from the MICCAI 2016 Challenge, pre-processed using FSL tools BET and FLIRT.

The results obtained with the optimized classifier are $SI=0.57$ and $ICC=0.92$, indicating a good correlation between the results and the ground truth, available as a subject-specific manual labelling of the lesions.

After the optimization, the test of the classifier has been carried out on 30 cases of the Lesjak public dataset, obtaining an $SI=0.26$ and an $ICC=0.27$.

The results were probably influenced by the differences between the datasets in the acquisition and in the lesion load distribution.

The small number of subjects for training and test is the main limitation of this study. A higher number of training subjects would increase the reliability of the features collected by the classifier and would improve the performance of the algorithm.

A limited availability of public MS neuroimaging data and the pandemic of the last year represented the main obstacles in finding a bigger and homogeneous dataset to use for the present work.

Future developments may include the use of a more extended training set, to increase the feature space, and a more homogeneous test set with clinically relevant features to evaluate.

Other improvements may be achieved with the inclusion of different MR modalities, such as T2w and T1w sequences after injection of Gadolinium.

A robust identification of MS lesions is crucial, since it could help the radiologist in the longitudinal assessment of MS disease progression as a function of time.

A highly desirable evolution of this method is the extraction of a wider set of features from the segmented WMH to exploit a radiomics approach to ultimately define the imaging fingerprint of the disease, thus pushing forward the knowledge in this field.

7. Bibliography

1. Metz I, Weigand SD, Popescu BFG, Frischer JM, Parisi JE, Guo Y, et al. Pathologic heterogeneity persists in early active multiple sclerosis lesions. *Ann Neurol*. 2014 May;75(5):728–38.
2. Mortazavi D, Kouzani AZ, Soltanian-Zadeh H. Segmentation of multiple sclerosis lesions in MR images: a review. *Neuroradiology*. 2012 Apr 17;54(4):299–320.
3. Filippi M, Agosta F. Imaging biomarkers in multiple sclerosis. *J Magn Reson Imaging*. 2010 Apr;31(4):770–88.
4. Traboulsee A, Li DKB, Zhao G, Paty DW. Conventional MRI Techniques in Multiple Sclerosis. In: *MR Imaging in White Matter Diseases of the Brain and Spinal Cord*. Berlin/Heidelberg: Springer-Verlag; 2005. p. 211–23.
5. Danelakis A, Theoharis T, Verganelakis DA. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Comput Med Imaging Graph*. 2018 Dec;70:83–100.
6. García-Lorenzo D, Francis S, Narayanan S, Arnold DL, Collins DL. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med Image Anal*. 2013 Jan;17(1):1–18.
7. Lladó X, Ganiler O, Oliver A, Martí R, Freixenet J, Valls L, et al. Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology*. 2012 Aug 20;54(8):787–807.
8. Lladó X, Oliver A, Cabezas M, Freixenet J, Vilanova JC, Quiles A, et al. Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches. *Inf Sci (Ny)*. 2012 Mar;186(1):164–85.
9. Balakrishnan R, Valdés Hernández M del C, Farrall AJ. Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data – A systematic review. *Comput Med Imaging Graph*. 2021 Mar;88(December 2020):101867.
10. ANBEEK P, VINCKEN K, VANOSCH M, BISSCHOPS R, VANDERGROND J. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. *Med Image Anal*. 2004 Sep;8(3):205–15.
11. Anbeek P, Vincken KL, van Osch MJP, Bisschops RHC, van der Grond J. Probabilistic segmentation of white matter lesions in MR imaging. *Neuroimage*. 2004 Mar;21(3):1037–44.
12. Vinitiski S, Gonzalez C, Mohamed F, Iwanaga T, Knobler RL, Khalili K, et al. Improved intracranial lesion characterization by tissue segmentation based on a 3D

- feature map. *Magn Reson Med.* 1997 Mar;37(3):457–69.
13. Mohamed FB, Vinitiski S, Gonzalez CF, Faro SH, Lublin FA, Knobler R, et al. Increased differentiation of intracranial white matter lesions by multispectral 3D-tissue segmentation: preliminary results. *Magn Reson Imaging.* 2001 Feb;19(2):207–18.
 14. Wu Y, Warfield SK, Tan IL, Wells WM, Meier DS, van Schijndel RA, et al. Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI. *Neuroimage.* 2006 Sep;32(3):1205–15.
 15. Steenwijk MD, Pouwels PJW, Daams M, van Dalen JW, Caan MWA, Richard E, et al. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage Clin.* 2013;3:462–9.
 16. Fartaria MJ, Bonnier G, Roche A, Kober T, Meuli R, Rotzinger D, et al. Automated detection of white matter and cortical lesions in early stages of multiple sclerosis. *J Magn Reson Imaging.* 2016 Jun;43(6):1445–54.
 17. Griffanti L, Zamboni G, Khan A, Li L, Bonifacio G, Sundaresan V, et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *Neuroimage.* 2016 Nov;141:191–205.
 18. Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, et al. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. *Sci Rep.* 2018 Dec 12;8(1):13650.
 19. Lesjak Ž, Galimzianova A, Koren A, Lukin M, Pernuš F, Likar B, et al. A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus. *Neuroinformatics.* 2018 Jan 4;16(1):51–63.
 20. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp.* 2002 Nov;17(3):143–55.
 21. Jenkinson M, Bannister P, Brady M, Smith S. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *Neuroimage.* 2002 Oct;17(2):825–41.
 22. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. *Med Image Anal.* 2001 Jun;5(2):143–56.
 23. FLIRT [Internet]. Available from: <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT>
 24. BIANCA [Internet]. Available from: <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA/Userguide>
 25. Sundaresan V, Zamboni G, Le Heron C, Rothwell PM, Husain M, Battaglini M, et al. Automated lesion segmentation with BIANCA: Impact of population-level features, classification algorithm and locally adaptive thresholding. *Neuroimage.*

2019 Nov;202:116056.

26. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979 Mar;86(2):420–8.
27. Commowick O, Cervenansky F, Ameli R. MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure. *Miccai.* 2016;
28. Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, et al. MICCAI 2016 MS lesion segmentation challenge: supplementary results. 2018 Jul 9;

Acknowledgements

This work is the final chapter of my experience in the University of Aveiro. It has been great and difficult, and it has led both to a personal and professional growth.

All this experience wouldn't be possible without professora Silvia De Francesco, who made me discover the Medical Imaging Technologies course in Verona, and helped me throughout the experience, from the house search in Aveiro to my problems with the Portuguese language, until being my thesis supervisor. To her goes my gratitude for her support.

I also want to thank Stefano Tambalo, the co-supervisor of this work. With his patience, knowledge and experience he has made this work possible. Moreover, he is responsible for my passion for MRI and image processing along with Diego Cavalli and Andrea Spagnolo. I thank the three of them for the support and for involving me in the MR research world.

Thanks to all my family and friends, that have always been loving and supportive.

In particular thanks to my parents, Lucia and Daniele, to my sisters, Marta, Marika and Noemi, and to my boyfriend, Francesco. I know that without them on my side I wouldn't be the person I am and I wouldn't reach the goals I've reached.