UNIVERSITÀ DEGLI STUDI DELL'AQUILA
DIPARTIMENTO DI MEDICINA CLINICA, SANITÀ PUBBLICA, SCIENZE DELLA VITA E DELL'AMBIENTE

Dottorato di Ricerca in Scienze della Salute e dell'Ambiente

Curriculum: Imaging Molecolare ed Ultrastrutturale

XXXIV Ciclo

Artificial Intelligence Models for Neurological Multimodal Imaging Analysis

SSD: INF/01

Dottorando

Ing. Matteo Polsinelli

Coordinatrice del corso
Prof.ssa Maria Grazia Cifone

Tutor
Prof. Giuseppe Placidi

a.a. 2020/2021

Acknowledgements

The work reported in this thesis was carried at the A^{2}VI-Lab (Acquisition, Analysis, Visualization & Imaging Laboratory), Department of Life, Health & Environmental Sciences (MESVA), University of L'aquila.

My deepest gratitude goes to my supervisor, professor Giuseppe Placidi. Not only because this thesis would hardly have been completed without his encouragement and for all the things that he teached to me since I was bachelor student. But also for showing me, every day, that you can be passionate about your work like the first day you started it. I would like to express my gratitude to professor Marco Ferrari and professor Maria Grazia Cifone, Coordinators of the Ph.D. course in Health and Environmental Sciences, for the attention, the dedication and passion that they demonstrated to me and my course colleagues.

I would like to thank professor Guido Macchiarelli, Head of the Department of Life, Health & Environmental Sciences for making it possible to carry out this work in his Department.

I am grateful to all the professors and colleagues that shared the knowledge in their fields of research and allowed me to collaborate with them. I would like to thank, among all, professors Luigi Cinque, Filippo Mignosi, Sara Invitto, Lia Ginaldi, Alessandra Splendiani, Giovanni De Gasperis and Emanuele Tommasino.

My special thanks goes to the colleagues and friends who have been with me during this period of study and work and that have made it a stimulating and precious experience: prof. Sandra Cecconi, drs. Matteo Spezialetti, Valentina Di Nisio, Alessandro Di Matteo, Eleni Theodoridou, Daniele Lozzi. Carmelita Marinelli and Gianna Rossi.

I cannot forget to thanks all my friends of "Tavolo Coppito 1".

I want to thanks my family: mom, dad, my brother and Yara. To all of you is dedicated this work.

I want to thanks my grand parents, Giuseppe and Giovanna, for what they did to me when I needed the most.

Finally, I want to show my gratitude to the reviewers of the manuscript, Prof. Maria De Marsico and Prof. Giovanna Castellano for their precious suggestions.

List of Candidate's Publications

[1] G. Placidi, L. Cinque, M. Polsinelli, and M. Spezialetti, "Measurements by a leap-based virtual glove for the hand rehabilitation," *Sensors*, vol. 18, no. 3, p. 834, 2018.

[2] P. Di Giamberardino, D. Iacoviello, G. Placidi, M. Polsinelli, and M. Spezialetti, "A brain computer interface by eeg signals from self-induced emotions," in *European Congress on Computational Methods in Applied Sciences and Engineering*, pp. 713–721, Springer, Cham, 2017.

[3] G. Placidi, L. Cinque, A. Petracca, M. Polsinelli, and M. Spezialetti, "A virtual glove system for the hand rehabilitation based on two orthogonal leap motion controllers.," in *ICPRAM*, pp. 184–192, 2017.

[4] G. Placidi, L. Cinque, A. Petracca, M. Polsinelli, and M. Spezialetti, "Iterative adaptive sparse sampling method for magnetic resonance imaging.," in *ICPRAM*, pp. 510–518, 2017.

[5] G. Placidi, L. Cinque, F. Mignosi, M. Polsinelli, and M. Spezialetti, "Sparse sampling for magnetic resonance imaging,"

[6] G. Placidi, M. Polsinelli, M. Spezialetti, L. Cinque, P. Di Giamberardino, and D. Iacoviello, "Self-induced emotions as alternative paradigm for driving brain–computer interfaces," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2018.

[7] G. Placidi, L. Cinque, M. Polsinelli, and M. Spezialetti, "Characterization of a virtual glove for hand rehabilitation based on orthogonal leap controllers," in *International Conference on Pattern Recognition Applications and Methods*, pp. 190–203, Springer, Cham, 2017.

[8] G. Placidi, M. Polsinelli, M. Spezialetti, and L. Cinque, "Bci driven by self-induced emotions: a multi-class study," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6, IEEE, 2018.

[9] M. Polsinelli, P. A. Banchetti, A. Cacchio, V. Calvisi, C. Marini, G. Placidi, M. Spezialetti, and L. Cinque, "Hand movement parameters calculated by the leap based virtual glove," in *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6, IEEE, 2018.

[10] G. Placidi, L. Cinque, M. Polsinelli, and M. Spezialetti, "Forces calculation module for the leap-based virtual glove," in *Proceedings of the 2018 10th International Conference on Bioinformatics and Biomedical Technology*, pp. 67–72, 2018.

[11] G. Placidi, L. Cinque, and M. Polsinelli, "A web application for characterizing spontaneous emotions using long eeg recording sessions," in *Innovations in Big Data Mining and Embedded Knowledge*, pp. 185–202, Springer, Cham, 2019.

[12] G. Placidi, L. Cinque, M. Polsinelli, A. Splendiani, and E. Tommasino, "Automatic framework for multiple sclerosis follow-up by magnetic resonance imaging for reducing contrast agents," in *International Conference on Image Analysis and Processing*, pp. 367–378, Springer, Cham, 2019.

[13] S. Invitto, A. Grasso, D. D. Lofrumento, V. Ciccarese, A. Paladini, P. Paladini, R. Marulli, V. De Pascalis, M. Polsinelli, and G. Placidi, "Chemosensory event-related potentials and power spectrum could be a possible biomarker in 3m syndrome infants?," *Brain sciences*, vol. 10, no. 4, p. 201, 2020.

[14] M. Balconi, G. Fronda, D. De Filippis, M. Polsinelli, and G. Placidi, "A preliminary structured database for multimodal measurements and elicitations of emotions: M2e2mo," 2019.

[15] M. Polsinelli, L. Cinque, and G. Placidi, "A light cnn for detecting covid-19 from ct scans of the chest," *Pattern recognition letters*, vol. 140, pp. 95–100, 2020.

[16] M. De Martinis, M. M. Sirufo, M. Polsinelli, G. Placidi, D. Di Silvestre, and L. Ginaldi, "Gender differences in osteoporosis: A single-center observational study," *The world journal of men's health*, vol. 39, no. 4, p. 750, 2021.

[17] G. Placidi, D. Avola, L. Cinque, M. Polsinelli, E. Theodoridou, and J. M. R. Tavares, "Data integration by two-sensors in a leap-based virtual glove for human-system interaction," *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18263–18277, 2021.

[18] G. Placidi, L. Cinque, and M. Polsinelli, "Guidelines for effective automatic multiple sclerosis lesion segmentation by magnetic resonance imaging.," in *ICPRAM*, pp. 570–577, 2020.

[19] G. Placidi, L. Cinque, and M. Polsinelli, "A fast and scalable framework for automated artifact recognition from eeg signals represented in scalp topographies of independent components," *Computers in Biology and Medicine*, vol. 132, p. 104347, 2021.

[20] G. Placidi, L. Cinque, F. Mignosi, and M. Polsinelli, "Multiple sclerosis lesions identification/segmentation in magnetic resonance imaging using ensemble cnn and uncertainty classification," *arXiv preprint arXiv:2108.11791*, 2021.

[21] G. Placidi and M. Polsinelli, "Local contrast normalization to improve preprocessing in mri of the brain," in *International Conference on Bioengineering and Biomedical Signal and Image Processing*, pp. 255–266, Springer, Cham, 2021.

[22] G. Placidi, G. D. Gasperis, F. Mignosi, M. Polsinelli, and M. Spezialetti, "Integration of a bci with a hand tracking system and a motorized robotic arm to improve decoding of brain signals related to hand and finger movements," in *International Symposium on Visual Computing*, pp. 305–315, Springer, Cham, 2021.

Contents

Summary

Medical imaging (MI) refers to several technologies that provide images of organs and tissues of human body for diagnosis and scientific purposes. Furthermore, the technologies that allow us to capture medical images and signals are advancing rapidly, providing higher quality images of previously unmeasured biological features at decreasing costs. This has mainly occurred for highly specialized applications, such as cardiology and neurology.

Artificial Intelligence (AI), which to date has largely focused on non medical applications, such as computer vision, provides to be an instrumental toolkit that will help unleash the potential of MI. In fact, the significant variability in anatomy across individuals, the lack of specificity of the imaging techniques, the unpredictability of the diseases, the weakness of the biological signals, the presence of noise and artifacts and the complexities of the underlying biology often make it impossible to derive deterministic algorithmic solutions for the problems encountered in neurology.

Aim of this thesis was to develop AI models capable of carrying out quantitative, objective, accurate and reliable analyzes of imaging tools, EEG and MRI, used in neurology. Beyond the development of AI models, attention was focused on the quality of data which can be lowered by the "uncertainty" produced by the issues cited above. Further, the uncertainty affecting data was also described, discussed and addressed.

Main results have been the proposal of innovative AI-based strategies for signal and image improvement through artifact reduction and data stabilization both in EEG and in MRI. This has allowed to apply EEG for weak signals recognition and interpretation (infant 3M patients), to provide effective strategies for dealing MRI variability and uncertainty in multiple sclerosis segmentation, both for single source and multiple-source MRI. According to the used evaluation criteria, the obtained results are comparable with those obtained by human experts.

Future developments will regard the generalization of the proposed strategies to cope with different diseases or with different applications of MI. Particular attention will be paid to the optimization of the models and to understand the processes underlying their behavior. To this aim, specific strategies for checking the deep structures of the proposed architectures will be studied. In this way, besides model

optimization, it would be possible to get the functional relationships among the features generating from the model and use them to improve human knowledge (a sort of inverse transfer learning).

Sommario

Al MI fanno riferimento diverse tecnologie che forniscono immagini di organi e tessuti del corpo umano per scopi diagnostici e scientifici. Inoltre, tali tecnologie che ci consentono di acquisire immagini e segnali medici stanno avanzando rapidamente, fornendo immagini di qualità superiore di caratteristiche biologiche precedentemente non misurate e allo stesso tempo a costi ridotti. Ciò si è verificato principalmente per applicazioni altamente specializzate, come la cardiologia e la neurologia.

L'AI, che è stata principalmente utilizzata per applicazioni non mediche, come la visione artificiale, fornisce uno strumento che contribuirà allo sviluppo il potenziale del MI. Infatti, la significativa variabilità anatomica tra gli individui, la mancanza di specificità delle tecniche di imaging, l'imprevedibilità delle malattie, la debolezza dei segnali biologici, la presenza di rumore e artefatti e la complessità della biologia sottostante spesso rendono impossibile sviluppare soluzioni algoritmiche deterministiche per i problemi incontrati in neurologia.

Lo scopo di questa tesi è stato quello di sviluppare modelli di IA in grado di effettuare analisi quantitative, accurate e affidabili dei dati forniti dagli strumenti di imaging più utilizzati in neurologia: EEG e MRI.

Al di là dello sviluppo di modelli di IA, l'attenzione è stata focalizzata sulla qualità dei dati che può ridursi per via dell' "incertezza" prodotta dalle problematiche sopra citate. Tale incertezza è stata anche descritta, discussa e affrontata.

I risultati ottenuti sono stati la proposta di strategie innovative basate sull'AI per il miglioramento del segnale e dell'immagine attraverso la riduzione degli artefatti e la stabilizzazione dei dati sia nell'EEG che nella MRI. Ciò ha consentito di applicare l'EEG per il riconoscimento e l'interpretazione dei segnali deboli (pazienti infantili 3M), per fornire strategie efficaci per affrontare la variabilità e l'incertezza della MRI nella segmentazione della sclerosi multipla, sia per la MRI a sorgente singola che a sorgente multipla. Secondo i criteri di valutazione utilizzati, i risultati ottenuti sono confrontabili con quelli ottenuti da esperti umani.

Gli sviluppi futuri riguarderanno la generalizzazione delle strategie proposte per far fronte a diverse malattie o con diverse applicazioni del MI. Particolare attenzione sarà riservata all'ottimizzazione dei modelli e alla comprensione dei processi alla base del loro funzionamneto. A tal fine verranno studiate strategie specifiche per il

controllo delle strutture interne delle architetture proposte. In questo modo, oltre all'ottimizzazione del modello, sarà possibile ottenere le relazioni funzionali tra le caratteristiche generate dal modello e utilizzarle per migliorare la conoscenza umana (una sorta di transfer learning inverso).

List of Figures

List of Tables

Part I

Introduction

Medical imaging (MI) refers to the techniques and the processes implemented to make images of organs and tissues of human body, for clinical and scientific purposes, that formerly was thought as a tool for diagnosis. MI is currently also used for treating, managing and predicting the progression of diseases, in particular for applications in oncology, cardiology and neurology [4]. In fact, fast, precise and minimally invasive imaging tools have been structured to use MI for treating several diseases. Imaging tools allow to collect multimodal data necessary to make a complete clinical evaluation of the patient. In this sense, MI is playing an increasingly important role towards personalized therapy. This has mainly occurred for highly specialized applications, such as cardiology and in particular neurology, MI is providing a series of precise and minimal invasive tools that are suitable to investigate the brain anatomy and functions, both in healthy and pathological conditions: Magnetic Resonance Imaging (MRI) and electroencephalography (EEG). MRI, thanks to the richness of imaging parameters and details, is capable to furnish invaluable contributions to better understand brain and brain functions, when functional MRI (fMRI) is used. Besides, EEG represents a valid functional support to MRI when fMRI is not suitable, too invasive or expensive to be used. Moreover, EEG offers its excellent temporal resolution which could greatly help in functional studies, when integrated with MRI and fMRI in hybrid systems for simultaneous acquisitions. The advantage of hybrid architectures is the exploitation of optimal spatial resolution of MRI and the excellent temporal resolution of EEG [5], though the last just usable to pick up cortical signals. The mini-invasive nature of MRI and EEG has allowed their massive usage that, in the downing era of data-drive health sciences, is responsible of producing huge amount of data.

The obtained data must be deeply analysed, often through comparisons with previous examinations (temporal follow-up), to establish functional relationships that often reside in the finest details. For this reason, objective quantification is a challenge that can only be performed through objective, automatic strategies of measurement and calculation. Though, in principle, it could be also performed by radiologists, it is tedious, time-consuming, source of errors and, hence, impractical for clinical routine.

Artificial Intelligence (AI), which to date has largely focused on non medical applications, such as computer vision, provides to be an instrumental toolkit that will help unleash the potential of MI. In fact, the significant variability in anatomy across individuals, the lack of specificity of the imaging techniques, the unpredictability of the diseases, the weakness of the biological signals, the presence of noise and artifacts and the complexities of the underlying biology often make it impossible to derive deterministic algorithmic solutions for the problems encountered in neurology. Besides, medical signal/image analysis (MIA) often concerns:

- quantification of specific geometric features of the objects of interest;
- assessment of changes over time;
- detection and characterization of morphological variations between subjects;
- analysis of shape and shape variability in objects and features;
- quantification of local or regional contrast or contrast differences.

The peculiarities of MI and the requirements of MIA are also challenges for AI that, being often applied for the detection/recognition of an object in an image whereby the precise geometry of the objects is irrelevant or may be known a-priori, may encounter difficulties which the scarcity of labeled data can further worsen.

AI is designed to mimic the layers of neurons in the human brain to process and extract information, allowing computers to learn by data, without being explicitly programmed.

Therefore learning from data to build models about multivariate and dynamic relationships among variables and utilizing these models to make inferences is an indispensable procedure in tackling the challenges of medical image analysis for neurology. AI provides effective solutions.

Aim of this thesis is to develop AI models capable of carrying out quantitative, objective, accurate and reliable analyzes of multi-modal medical signals used in neurology. Beyond the development of AI models, we focus the attention on the quality of data which can be lowered by the "uncertainty" produced by the issues cited above. In this thesis the uncertainty affecting data will be highlighted, discussed and addressed. An increasing size approach will be followed: first one-dimensional signals (EEG), then multidimensional (3D) and multi-modal images (MRI) will be treated. Following this concept, in the first Section:

- Chapter 1 presents an AI-based framework for automatic artifact removal from EEG signals;
- Chapter 2 uses the framework presented in Chapter 1 to interpret with weak and disturbed EEG signals of infants affected by 3M syndrome.

in the second Section:

- Chapter 3 describes the materials and methods used for MRI processing: the used MRI data set, the way to deal with inter-raters variability in MRI and the criteria used to evaluate the proposed AI-models (metrics and scores);
- Chapter 4 introduces several guidelines for the effective training of an AI model for automatic segmentation of MS lesions by MR images;

- Chapter 5 provides an optimization of the pre-processing pipeline of MRI, through a local contrast normalization algorithm;
- Chapter 6 presents an AI framework for MS lesion segmentation by single modal MRI;
- Chapter 7 proposes a novel AI architecture that uses multiple MRI modalities and evaluates the contribution of each modality to the diagnosis.

A final Section presents the conclusions and the future work.

It is important to clarify that the thesis reports methods and findings published on international journals and conference proceedings, co-authored by the candidate, appropriately referenced.

Part II

Artificial Intelligence For EEG Analysis

Chapter 1
A CNN-based Architecture for Automated Artifact Recognition from EEG Signals Represented in Scalp Topographies of Independent Components

Electroencephalography (EEG) measures the electrical brain activity in real-time by using sensors placed on the scalp. Artifacts due to eye movements and blinking, muscular/cardiac activity and generic electrical disturbances, have to be recognized and eliminated to allow a correct interpretation of the Useful Brain Signals (UBS). Independent Component Analysis (ICA) is effective to split the signal into Independent Components (IC) whose re-projection on 2D topographies of the scalp (images also called Topoplots) allows to recognize/separate artifacts and UBS. Topoplot analysis, a gold standard for EEG, is usually carried out offline either visually by human experts or through automated strategies, both unenforceable when a fast response is required as in online Brain-Computer Interfaces (BCI). This chapter presents a fully automatic, effective, fast, scalable framework for artifacts recognition from EEG signals represented in IC Topoplots. The framework, composed by three 2D Convolutional Neural Networks (CNN), divides Topoplots in 4 classes: 3 types of artifacts and UBS.

The content of this chapter appeared in [6].

1.1 Introduction

EEG measures the neuronal activity through electrodes placed on the scalp with an excellent temporal resolution. Optimal temporal resolution and low invasiveness make EEG particularly suitable for real-time usage [7, 8, 9]. Extraneous signals produced by eye movements and blinking, muscular spasms, cardiac activity and generic interferences [10, 11] can obscure UBS since skull and scalp (including muscles) are between brain and sensors.

Blinking and eye movements produce electrooculography artifacts (EOG) mainly recorded by frontal sensors and which propagate across the scalp [12]. Three categories of EOG exist: eye blinking (BEOG), vertical (VEOG) and horizontal (HEOG)

EOG. EOG has often much higher amplitude than UBS and frequencies in the range 10-40 Hz where also UBS are present.

Cardiac activity produces electrocardiography artifacts (ECG) [13]. ECG effects can be reduced by subtracting the signal of a peripheral sensor from those located on the scalp [11]. Residual ECG effects are lower than brain signals but are still present.

Cranial muscles produce electromyogram artifacts (EMG). The main feature of EMG is the wide spectral distribution with maximum power in the range of 15-30 Hz where also UBS insist [14].

Finally, generic discontinuities are generated by impedance fluctuations or electric/electronic interferences (IF) affecting single sensors with large fluctuations in amplitude [15].

Artifacts could be much intense than UBS and propagate to large regions [12, 15, 16]. UBS could be completely obscured and signal wrongly interpreted if a selective preprocessing strategy was not employed. As stated above, artifacts and UBS share some bands of frequencies and preprocessing alternatives to frequency analysis are required to separate their respective contribution. Fortunately, EEG signals can be viewed as a mixture of independent linear source components, some mainly due to artifacts and others to UBS [17]. In addition to components also their reprojection on the scalp is important because shape and localization are a fundamental information to classify sources [18]. An effective method to retrieve source components from EEG signals is Independent Component Analysis (ICA) [16] which, from an n channel EEG measurement, allows a calculation of at most n independent components on a given temporal window of the signal. A component is defined by an array of weights representing the contributions of the sensors to it. Weights can be interpolated in 3D by using the spatial map of the sensors on the scalp and reprojected on a 2D Topoplot [19] to allow shape and spatial localization analysis (Fig 1.1).

Visual inspection of Topoplots by a human expert allows to produce source classification [19]. In fact, although additional information such as power spectrum density (PSD) and autocorrelation can improve classification accuracy, it is rarely used due to the enormous increase in preprocessing time (visual inspection takes about 5 -7 sec per Topoplot: additional parsing information could more than double the time). Visual inspection is effective, widely used for off-line removal of artifacts and therefore supported by most EEG signal processing tools [20, 21].

Automated strategies, in particular CNN based approaches, have been already employed with success for EEG signals classification and artifacts removal [22, 23, 24, 25]. However, these strategies are neither optimized for interactive fast responses nor scalable. Moreover, they are sensitive to dynamic signal modifications

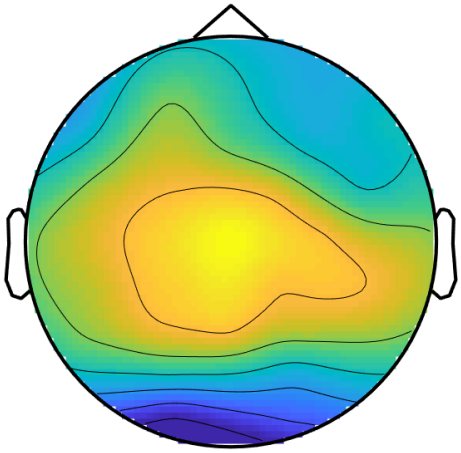

Figure 1.1: An example of an Independent Component represented in form of Topoplot. Colours are indicative of the underlying brain activations: blue represents low activation, yellow corresponds to high activation.

not solvable without re-training. These are the main objectives of the proposed chapter.

In fact, we propose an effective, fast, scalable, easily trainable, and robust automatic framework for EEG artifacts recognition by scalp topographies of Independent Components with a CNN-based approach that emulates the human visual interpretation process to be used in interactive EEG. Rapidity, scalability and robustness are necessary conditions for interactivity. In particular, the system response must be fast enough to justify interactivity with BCI, compatible with the time required for stimulus/response. Scalability, both with respect to the number of sensors and recognizable artifacts, is important for BCI in disabled people: it is very common that EEG configurations have to be specially designed and new artifacts arise due to specific muscle spasms and uncontrolled ocular or head movements. Furthermore, not ideal operating conditions could dynamically change the signal quality which the system should be robust. In addition to the framework, we also define a public dataset of labelled Topoplots on data collected by DEAP, a web-based public dataset of EEG signals [26].

1.2 Related Works

Several techniques have been designed and used for EEG artifacts detection [27, 28, 10] which can be grouped into the following categories: Regression Methods; Filtering Algorithms; Wavelet Transform; Empirical Mode Decomposition; Blind Source Separation.

Regression Methods assume that artifacts are measured through dedicated channels [29, 28]. Measurements are necessary to estimate the propagation coefficients

that need to be subtracted by brain signals. Though these strategies have a good computational performance, they have two major drawbacks: one or more reference channels are needed, being this a severe limitation for BCI [30]; the reduction of artifacts also implies the removal of relevant UBS [15].

Filtering includes several approaches, the most widely used being the adaptive filter [31]. This method assumes that UBS are unrelated to artifacts and requires a dedicated channel to measure the artifacts to be subtracted from the signal. Filtering strategies suffer from the same limitations as well as the Regression Methods [32].

Unlike Regression and Filtering, Wavelet Transform (WT) does not require reference signals. WT transforms the signals from time domain to time-frequency domain and low amplitude WT coefficients are zeroed before being inversely transformed. The main limitation is that artifacts which overlap in frequency with UBS, or which are too specific, are not completely removed [28, 10, 33].

Empirical Mode Decomposition (EMD) [34] and Multivariate EMD (MEMD) [35] are tools for signal decomposition into amplitude and frequency modulated basis functions (Intrinsic Mode Functions, IMFs). EMD is specific for single channel data and MEMD is the extension of EMD to multichannel data. These methods are robust to noise and suitable for muscle artifacts removing [36] but, being slow, they are not suitable for BCI [10].

Blind Source Separation (BSS) does not require prior information and/or reference signals. Over time, two main algorithms have been used: Principal Component Analysis (PCA) and ICA. The first use of PCA was in 1991 [37] but in 1997 it was shown that PCA is unable to completely separate artifacts from UBS [38]. In fact PCA transforms the time-domain observations of correlated variables into a set of linearly uncorrelated variables using orthogonal transformations: when UBS and artifacts are not orthogonal, PCA fails to separate the corresponding components [28]. The ICA defined above overcomes PCA limitations [16, 39].

More recently, other forms of BSS strategies have been proposed. One of these, Canonical Correlation Analysis (CCA), is more effective than ICA in removing muscle artifacts by EEG signals [40]. CCA assumes a relatively low autocorrelation of muscle artifacts with respect to brain activity considered to be maximally autocorrelated. For this reason, CCA compares the current EEG signals to a delayed version of the past signals and preserves sources that maximize autocorrelation between the two datasets. This strategy has been shown to work well for the removal of muscle artifacts but must be combined with other strategies to cope with other sources of artifacts or with a few-channel EEG [41].

Another recent method is Independent Vector Analysis (IVA) [42, 43]. It includes in a single strategy both the need of assuming that some artifacts have different autocorrelation properties than UBS (muscular activity) and that other artifacts

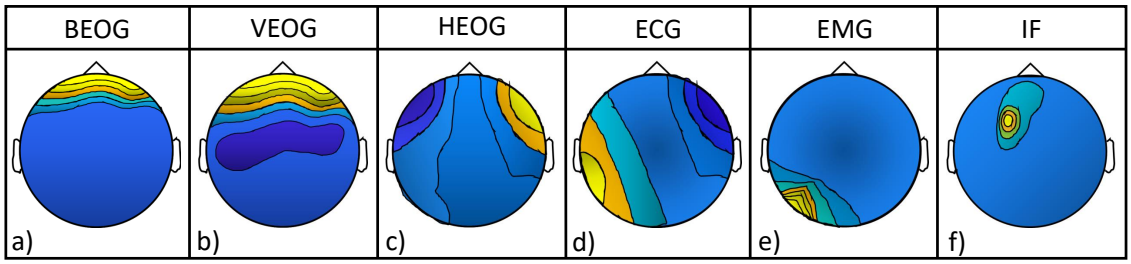

Figure 1.2: ICA components related to artifacts. BOEG (a) and VEOG (b) have similar shapes. The same occurs between HEOG (c) and ECG (d) and between EMG (e) and IF (f). EOG and ECG have locations well-defined on the head; EMG and IF are represented by isolated peaks.

are independent of UBS (ocular movements). IVA outperforms ICA and CCA in isolating muscle and ocular artifacts, especially in low-quality signals, but it does not handle all types of artifacts.

Despite recent proposals for increasingly effective EEG signal preprocessing strategies, which can also be usable in freely-available tools [44], ICA is still considered the most general one, thanks to its ability to treat all types of known artifacts [45, 46, 10], although for some of them showing a sub-optimal performance. In fact, the possibility of representing Independent Components (IC) calculated by using ICA in 2D scalp topographies allows them to be recognized and classified visually. For this reason, the framework presented is based on the analysis of the Topoplots of the IC calculated by using ICA to detect artifacts. Fig 1.2 shows characteristic topographies of the artifacts presented above [15, 19, 47].

As it can be seen, BEOG is concentrated on the frontal region of the head (Fig 1.2.a) as well as VEOG (Fig 1.2.b), though VEOG spreads through the head more than BEOG. For their similarity both in shape and meaning, they can be grouped into a single class.

In the case of HEOG, two peaks of opposite sign are positioned around the nose (Fig 1.2.c). Similarly to HEOG, ECG (Fig 1.2.d) is composed of two peaks of opposite sign localized on the edges of the head, around the ears (ECG differs from HEOG only for a different orientation). The similarity between HEOG and ECG suggests their inclusion in the same class (the recognition between them, outside the scope of this manuscript, could be based on the orientation of the peaks).

EMG and IF are isolated peaks, the former usually found on the border of the head near the neck and face where muscular activity is pronounced (Fig 1.2.e), while the latter is often located in the middle of the head (Fig 1.2.f). Due to their similar shape, EMG and IF are included in the same class, although their nature is very different (EMG are due to muscle activation while IF are due to electrical

disturbances). The distinction between a channel failure and a muscle artifact is difficult even for a human expert: position, power and frequency of occurrence increase the probability of one over the other but, for our purposes, they must be both discarded.

As specified above, the role of IC topography is fundamental in the identification of artifacts by a human expert. However, EEG measurements often consist of several time windows (trials) also divided into many sub-windows (sub-trials) and the amount of IC Topoplots could easily reach several thousands. These numbers make visual inspection and recognition impossible, especially when a quick response is required.

In the paper [27], out of 46 works reviewed, only 3 use IC [48, 49, 50], all of these designed to treat specific artifacts and not for removing all of the above types. Recently, CNN have revolutionized computer vision, particularly for what concerns automated object recognition [51, 52, 18]. CNN have been successfully used in several EEG classification studies [8, 53, 54, 25, 55]. A recently proposed automatic method relies on CNN to classify artifacts by low resolution IC Topoplots (32x32), PSD and autocorrelation [25]. In addition to Topoplots, further data is necessary to support the information loss that occurs when using low resolution IC Topoplots (low resolution is necessary to gain efficiency). Though effective, the resulting classification strategy is particularly difficult to train due to the difficulty of obtaining labelled data from human experts. In fact, a lot of time is required for manual classification and for the definition of threshold parameters both for PSD and for autocorrelation. Furthermore, the strategy is poorly adaptable and generalizable to time-variable signal-to-noise ratio scenarios.

Croce et al. [23] proposed a CNN-based approach for the automatic classification of IC from EEG and Magnetoencephalography (MEG) signals. While effective, similarly to [25] it uses low-resolution Topoplots and PSD in an off-line mode. In fact, the architecture has not been designed and optimized for fast-response applications, although it is particularly suitable for signal interpretation of two electrophysiological modalities and it uses multimodality to improve performance.

To the best of our knowledge, none of the state of the art automatic classification strategies allows to: recognize all types of actually known artifacts; achieve high accuracy; use only information coming from IC Topoplots; be easy and easily trainable; be robust to dynamic changes in the signal quality; be independent of the number of channels; be scalable with respect to newly discovered sources of artifacts; be fast enough for BCI. Aim of this chapter is to satisfy all the above requirements.

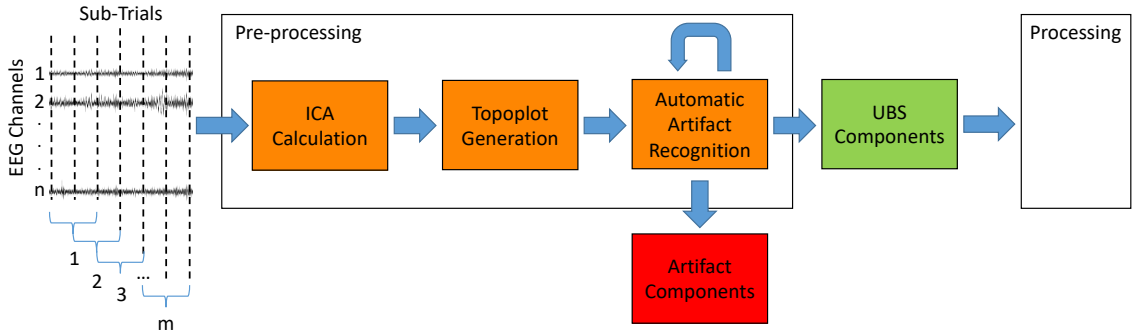

Figure 1.3: Pre-processing pipeline: ICA calculation of partially overlapping EEG temporal sub-trials enclosed by curly brackets (left); generation of the resulting Topoplots; automatic recognition of artifacts from Topoplot images. Artifacts are discarded and UBS are passed to the processing stage.

1.3 The Overall Architecture

In what follows, we propose to replace the role of a human expert and his visual inspection of IC Topoplots with a fully automatic CNN-based analysis of the Topoplots of EEG signals separated in partially overlapping time windows to account for the transient nature of the artifacts, as reported in Fig 1.3.

Our objective is to allow the separation of artifacts from UBS and to classify artifacts into the three classes defined above. This last choice is not justified by a simple recognition/elimination process: our aim is to separate artifacts into classes in order to make quick decisions about them. For instance, if an artifact of the class EMG/IF occurs frequently, it can be argued that this is caused by a sensor failure (IF) rather than due to muscle spasms (EMG): in which case, definitive exclusion of the sensor could be more convenient and efficient than continuously discovering/eliminating the artifacts it generates. This type of decision could be of paramount importance in BCI. Other objectives are to: push on generality and scalability by providing for the treatment of new Topoplots of future artifacts without redefining the overall structure and, more importantly, without re-training the entire system; reduce the training dataset and labelled data for training.

To pursue these objectives, the architecture in Fig 1.4 is proposed, consisting of 3 parallel CNN of the type described in [56, 57]. In fact, due to their common patterns, BEOG and VEOG are grouped into a first CNN (B_V CNN), HEOG and ECG (H_E CNN) into a second CNN and EMG and IF (E_I CNN) in a third CNN. The reason of grouping different artifacts in a single CNN is threefold: 1) artifacts grouped in the same CNN are often indistinguishable from each other even for a trained human expert; 2) grouped artifacts share the same treatment;

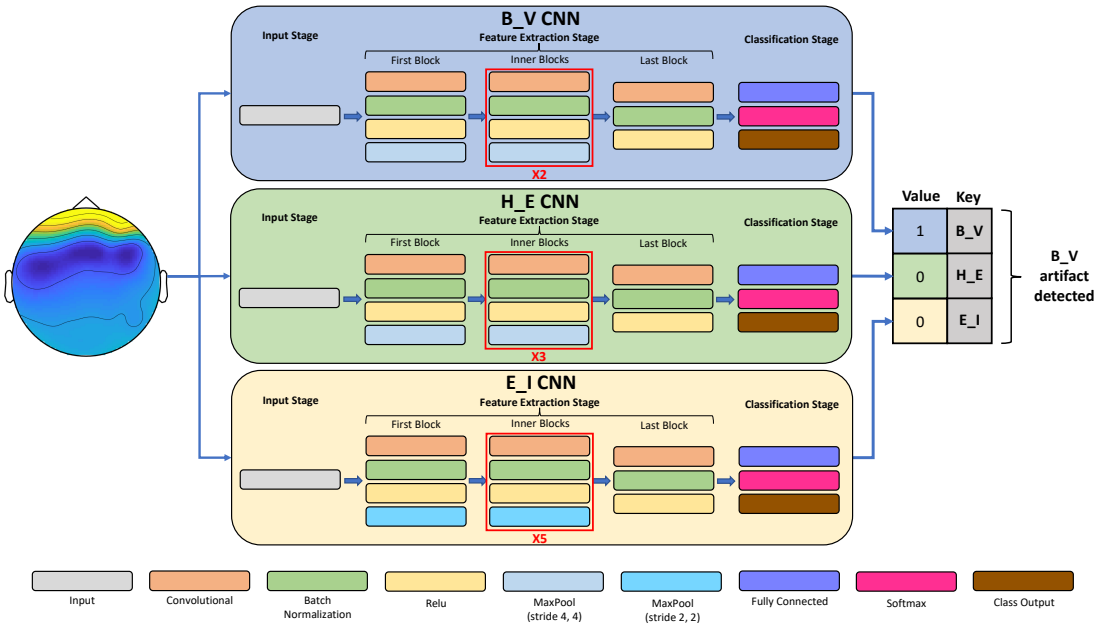

Figure 1.4: Framework architecture. The current Topoplot is passed to three CNN separately (the number of CNN is the same of the artifact classes). As example, a BEOG is passed to the framework: the resulting output indicates its recognition by the first CNN (1) and its refutation by the other two (0). The Input and the Classification Stages have the same design in all CNN. Regarding the Feature Extraction stage, B_V CNN contains 2 Inner Blocks, H_E CNN 3 Inner Blocks and E_I 5 Inner Blocks..

3) computational efficiency improves. Besides the advantages of clustering, the proposed framework has also some advantages over extreme clustering, that is the use of a single CNN. In fact, it allows to: accelerate training, reduce datasets for training, accelerate convergence and increase generality. These advantages can be explained mainly by the fact that each CNN can be trained separately.

1.3.1 Framework Design

CNN are based on feed-forward Artificial Neural Networks (ANN). A CNN consists of input and output layers as well as of multiple hidden layers for feature extraction which include convolutional layers. The main advantages of CNN over classic ANN is that the neurons in one layer do not connect to all the neurons in the next layer, but only to a small subset. The three CNN used therein are structured as in Fig 1.4.

Each CNN is organized in 3 stages: an Input Stage, a Classification Stage, interspersed with a Feature Extraction Stage. The Input Stage consists of only one Input Layer. The Classification Stage consists of a Fully Connected Layer (of dimension 2), a Softmax Layer and a Classification Layer. Input and Classification Stages are the same for all CNN. The Feature Extraction Stage extracts different features for each class of artifacts (geometrical properties, position and orientation within the Topoplot, intensity, etc), it is specific for each CNN and is organized in "Blocks". Each Block contains a Convolutional Layer, a BatchNorm Layer, a Relu Layer and a MaxPool Layer, except for the last Block where the MaxPool Layer is absent.

In the B_V CNN, the Feature Extraction Stage consists of a First Block, 2 Inner Blocks and a Last Block. For each block, the Convolutional Layers use respectively 8, 16, 32 and 64 filters (kernel size 3x3) and the MaxPool Layers have size 2x2, stride [4,4] and padding 0. H_E CNN contains a First Block, 3 Inner Blocks and a Last Block. In all blocks, the Convolutional Layers use respectively 8, 16, 32, 64 and 128 filters (kernel size 3x3) and the MaxPool Layers have the same size as B_V CNN. Finally, in E_I CNN, the Feature Extraction Stage consists of a First Block, 5 Inner Blocks and a Last Block. In all blocks, the Convolutional Layers use 8, 16, 32, 64, 128, 256 and 256 filters. In the latter case, the Max-Pool Layers are still 2x2 in size but stride [2, 2] (this because Impedance and EMG artifacts are composed of a small cluster of pixels). In all CNN, the First Block has the dimensions of an Inner Block but it has been separated from the Inner Blocks to indicate that the number of the latter may vary due to the complexity of the class to be recognized. The Feature Extraction Stage is specific for each class because:

- B_V patterns are well defined and localized (low complexity);

- H_E patterns are less defined than B_V ones (medium complexity);
- E_I patterns are neither well localized nor well defined (high complexity).

The number of Inner Blocks, of filters in Convolutional Layers and the stride in MaxPool Layers are optimized for each CNN. The process starts from a minimal architecture for all classes (the B_V CNN) by repeating training and Inner Block increment until the maximum accuracy is achieved. The highest accuracy is achieved first by B_V CNN, then by H_E CNN and finally by E_I CNN. In our architecture the number of parameters, $1.1x10^6$, is distributed in CNN as follows: $2.5x10^4$ in B_V, $9.5x10^4$ in H_E and $9.5x10^5$ in E_I. In terms of number of parameters, using a single CNN to classify the 4 classes would probably be more demanding than using three different CNN, as confirmed by the huge increase of the network parameters when the complexity of the class to be recognized increases. In terms of number of parameter, our framework is lighter than SqueezeNet [58], one of the most competitive CNN architectures.

1.4 Experimental evaluation

The framework was implemented in Matlab (The MathWorks Inc., https://mathworks.com/) on a PC with Intel Core I7-6700, 32GB of RAM and Nvidia GeForce GTX 1080.

1.4.1 Experimental Dataset

The experimental dataset consists of EEG signals collected from the DEAP dataset [26], a public multicentre database containing a collection of EEG signals of negative and positive emotional states recorded from 32 participants (16 men and 16 women, aged between 19 and 37, average: 26.9) while watching 40 music videos (1 minute each) on different topics. Participants rated each video in terms of arousal, valence, like/dislike, dominance and familiarity. The EEG signals, sampled at 512 Hz, were recorded on the following 32 positions of the international 10-20 positioning system [59]: Fp1, AF3, F3, F7, FC5, FC1, C3, T7, CP5, CP1, P3, P7, PO3, O1, Oz, Pz, Fp2, AF4, Fz, F4, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, P4, P8, PO4, and O2. For more information on DEAP please refer to [26]. For our purposes, the raw data from DEAP where used after filtering with a notch filter [60] at 50Hz and 60Hz to suppress power-line interferences. Data were divided into temporal sub-trials, windows of 8 seconds overlapping each other for 4 seconds (4 seconds of "present" signal joint to 4 seconds of "past" signal) and used to generate IC and the corresponding Topoplots. The past signal was used to support the current signal to satisfy both the following seemingly conflicting requirements:

1. obtain a signal long enough to ensure ICA convergence;

2. shorten the "present" window of the signal for quick EEG applications.

This choice represents also a good compromise to reduce transient artifacts while preserving UBS (longer time sequences would mean artifacts and UBS).

ICA was calculated on each sub-trial by obtaining a maximum of 32 components, at most one per channel. The Topoplot corresponding to an IC was generated and managed as a 134x136 RGB image having a fixed position and orientation. This resolution represents a good compromise between high spatial precision and reduced execution time . Another constraint is that Topoplots were represented in the 64 colors Parula palette. This palette is commonly chosen for problems solved with the use of CNN ([61],[62],[63]). Since our framework has been trained on Topoplots represented in Parula palette, it only treats images in Parula: transformations are necessary if other palettes are used. The number of obtained Topoplots was 992800. From this huge dataset, images were extracted for training/validation/test of CNN.

Data augmentation (changes of orientation, scaling, translation and brightness) was not used because:

1. the orientation of each Topoplot is fixed and rotations would change its meaning and, therefore, would be unjustified;

2. scaling and translation would create redundancy because the interpretation of a Topoplot always refers to the external silhouette of the head;

3. modifying brightness would be wrong because Topoplots are generated with fixed color maps and fixed brightness scale, as discussed above;

4. a huge number of labelled datasets are available.

1.4.2 Training

DEAP data of subjects 1-8 were manually classified and labelled independently by 5 human experts into 4 different categories: BEOG ∪ VEOG (B_V), HEOG ∪ ECG (H_E), EMG ∪ IF (E_I) and UBS. A consensus dataset was obtained considering each Topoplot belonging to the most voted class. In case of tie result (the case in which the most voted classes were 2 with 2 votes each), the resulting Topoplot was discarded. The consensus dataset agreed 95.7% with all human experts (agreement between consensus and each of the experts was between 97.2% and 99.1%).

All artifacts were extracted from the labelled consensus and just one subset of UBS was randomly selected. The composition of the training sets is illustrated in Table 1.4.2.

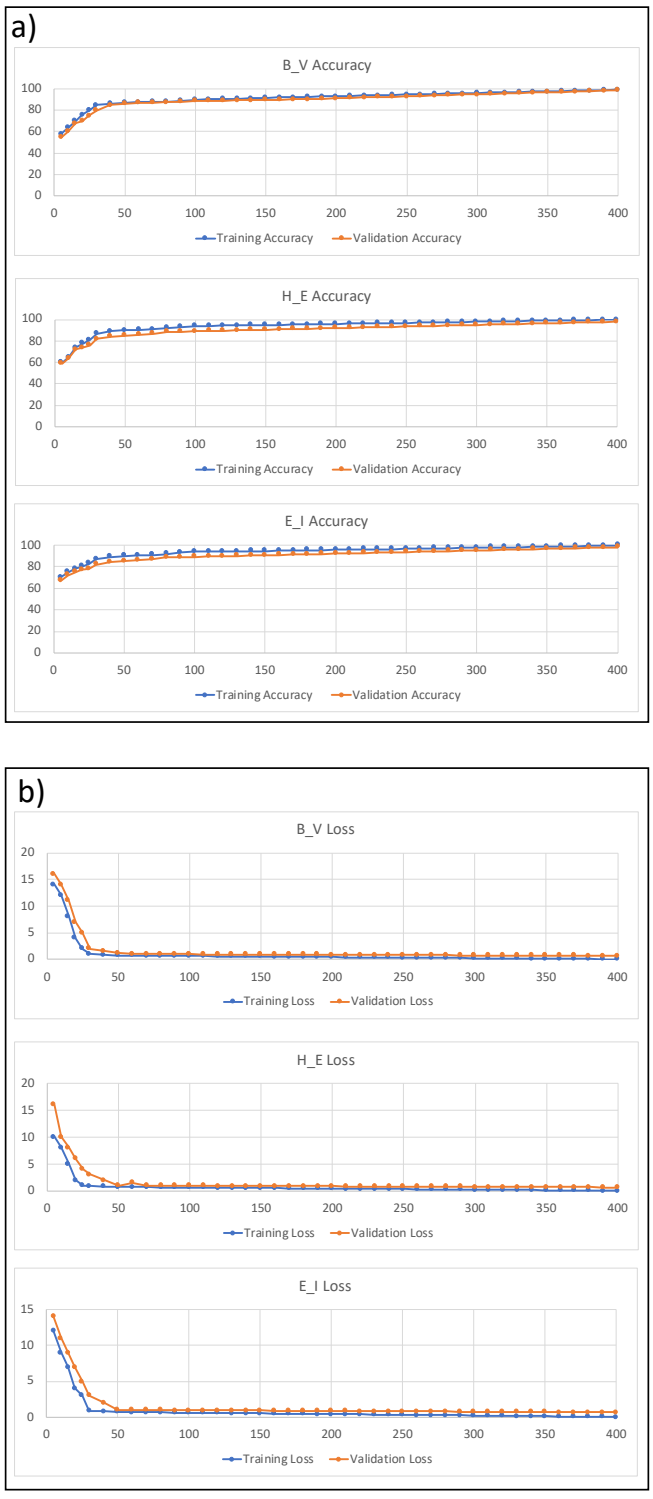

Figure 1.5: Accuracy % (a) and Loss values (b), with respect to the epochs, for the three CNN. Blue is used for training and red for validation.

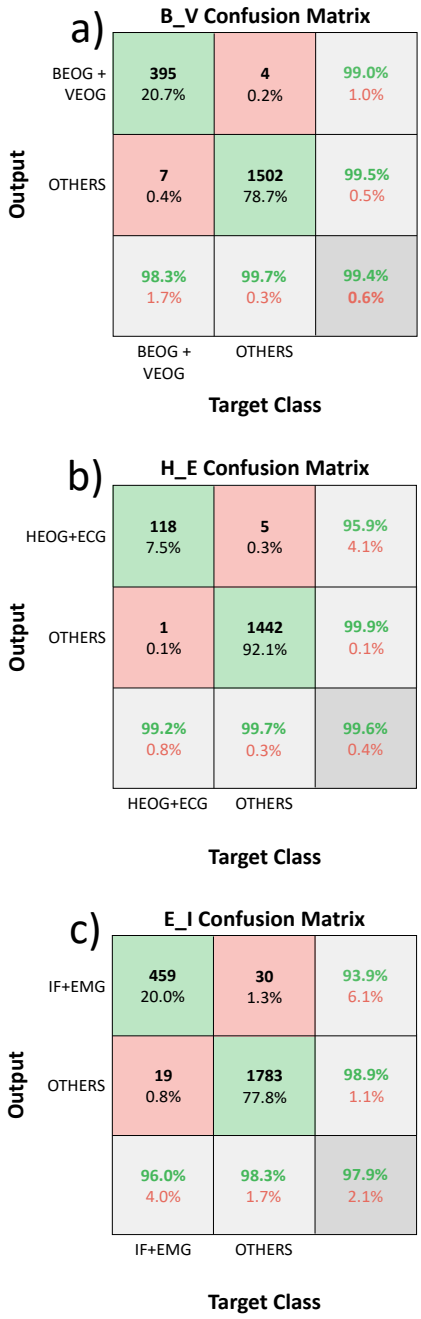

Figure 1.6: Confusion matrices for the fifth iteration of the validation process of: B_V (a); H_E (b); E_I (c). The upper left 2x2 sub-matrix contains true positives (1,1), false positives (1,2), false negatives (2,1) and true negatives (2,2). Each cell contains the absolute value (bold) and the corresponding % of the total number of elements (plain text) . The remaining cells always contain two numbers %. The third column consists of: positive predictive value (green) and false discovery rate (red), in position (1,3); negative predictive value (green) and false omission rate (red), in position (2,3). The third row consists of: sensitivity or true positive rate (green) and false negative rate (red), in position (3,1); specificity or true negative rate (green) and false positive rate (red), in position (3,2). Cell (3,3) contains accuracy (green) and misclassification rate (red).

B_V CNN	H_E CNN	E_I CNN
1341 (B_V)	398 (H_V)	1592 (E_I)
5020 (H_E + E_I + UBS)	4823 (B_V + E_I + UBS)	6044 (B_V + H_E + UBS)

Table 1.1: Configuration of the datasets used to train the framework.

		Classification Results			
		Others	**Artifacts**	**Double Detections**	
Framework	**B_V**	310720	30120	**H_E**	**E_I**
				340	4480
	H_E	332429	8411	**B_V**	**E_I**
				340	320
	E_I	296669	44171	**B_V**	**H_E**
				4480	320

Table 1.2: Classification of 340890 Topoplots. The "Others" column contains other artifacts + UBS. The "Artifacts" column contains the number of artifacts recognized as its own by the CNN shown on the left. "Double detections" are artifacts considered to be their own by two CNN simultaneously and are reported twice. For instance, the Topoplots belonging to B_V and, at the same time, to H_E (340) are the same ones that belong to H_E and, at the same time, to B_V. "Triple detections" never occurred and not reported.

In particular, for each CNN the training set was organized by separating the Topoplots into two classes: the first class containing the artifacts to be recognized by CNN and the second class containing all the others (other artifacts + UBS). In this way, each CNN was trained to recognize its own artifacts and separate them from the others (the cardinality of the three classes roughly reflects the real frequency of occurrence).

The dataset for each CNN was split 70% for training and 30% for validation. The optimization algorithm used was the Stocastic Gradient Descendent with momentum=0.9, with L_2-norm. The max number of epochs was fixed to 400, although all the CNN converged well below 100 epochs (Fig 1.5). The training process, taking about 40 min, was repeated 10 times and resulted in the following average accuracy: 99.4±0.4 % for B_V, 99.6±0.3% for H_E and 97.9±0.6% for E_I. The confusion matrices of one of the validation processes are shown in Fig 1.6. They confirmed that the class E_I was the most difficult to recognize.

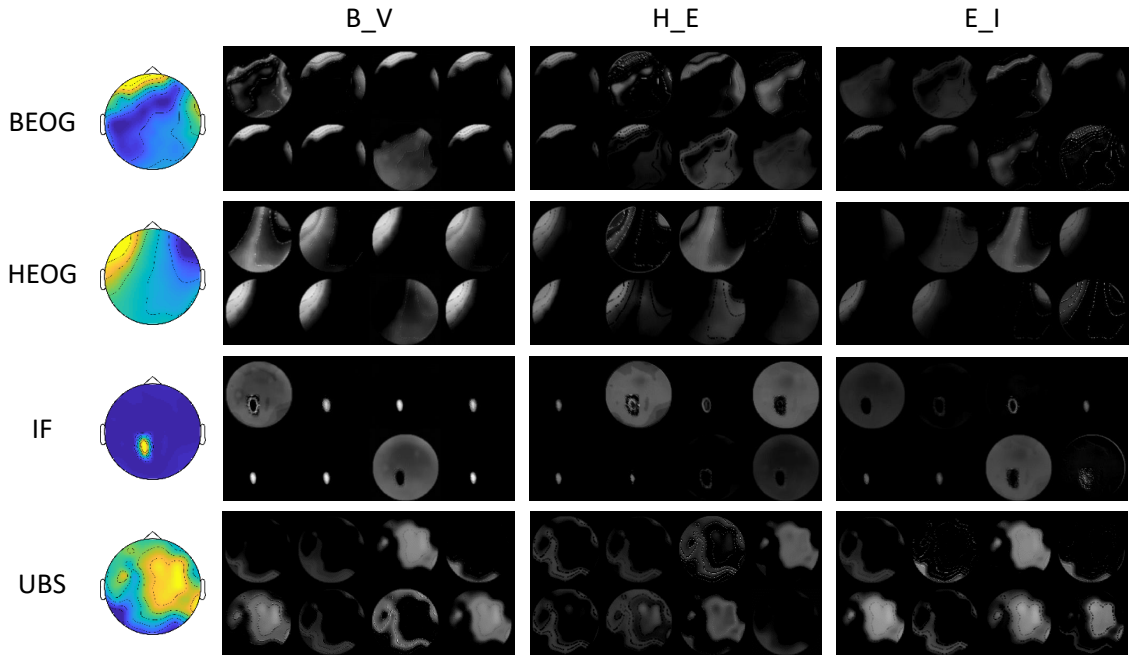

Figure 1.7: First feature extraction step when each CNN acts on Topoplots belonging to the three classes of artifacts or to UBS (columns) or when different CNN act on the same Topoplot (rows).

1.4.3 Results

The proposed framework was tested on data from subjects 20-32 of the DEAP dataset, corresponding to 340890 images. The results of classification, summarized in Table 1.2, show that, among the Topoplots considered artifacts, a very small percentage had double membership ("Double Detections"). Upon thorough examination with human experts, most of them were actually found to be ambiguous. It is worth noting that the numbers are repeated twice in "Double Detections" columns: Topoplots shared between classes j and k have been counted by both. No "Triple Detections" occurred, that is no Topoplot was classified as a member of the 3 classes at the same time and, for this reason, a "Triple Detections" column is not present in Table 1.2.

The behaviour of each CNN with respect to the others, when acting on different types of Topoplots, is shown in Fig. 1.7. In particular, the output of the first feature extraction step is presented, both when acting on the same Topoplot (rows) and on different Topoplots (columns). As it can be seen, CNN reacted very differently to the same Topoplot, thus demonstrating good fixation activity [64]. The Topoplots shown in Fig. 1.7 are characteristic instances of all classes.

We also performed Gradient-weighted Class Activation Mapping (Grad-Cam) to produce a coarse localization map showing which regions in the image were used for

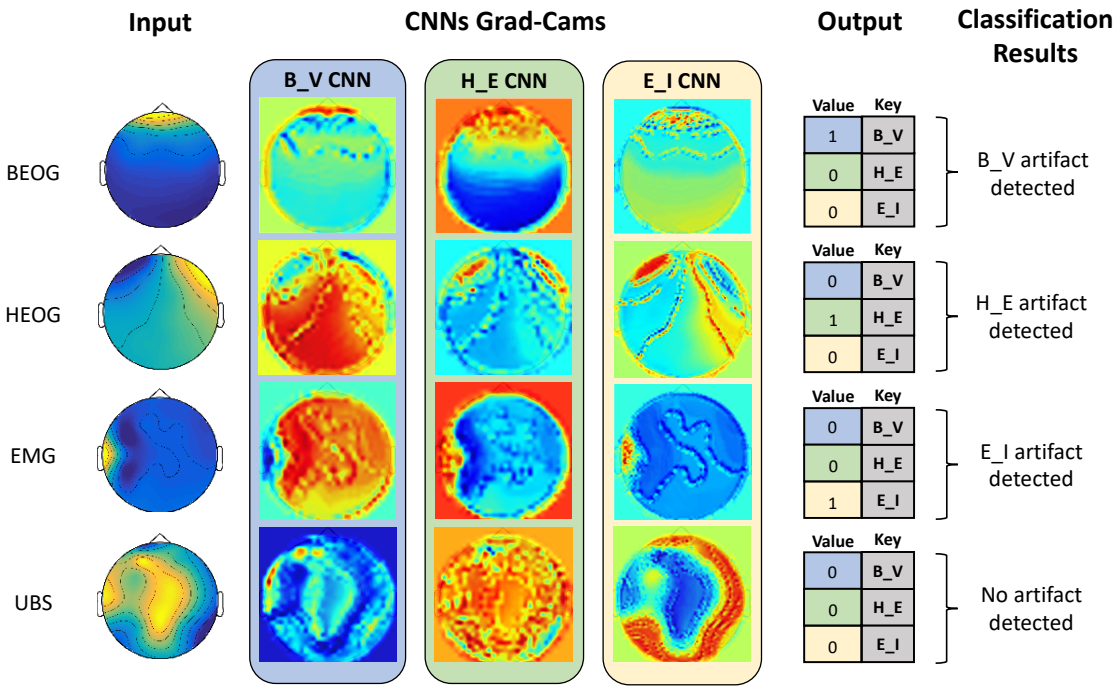

Figure 1.8: Grad-Cams: the Topoplots belonging to the three classes of artifacts or to UBS are inputs of each CNN (columns); the same Topoplot is the input of different CNN (rows). The last two columns on the right show "Output" and "Classification results", respectively.

prediction [65]. Fig. 1.8 shows examples of grad-cams in a table: rows report the 4 different classes of Topoplots (artifacts + UBS) and columns represent the respective grad-cams of each CNN. In the first row, a B_V artifact is the CNN input. The B_V CNN grad-cam shows that the activations are correctly localized on the frontal region of the head. In the second row, a H_E artifact is the CNN input and the H_E CNN grad-cam shows that the activations are correctly located on the lateral edges of the head: this confirms that the H_E CNN is not biased by the strongest positive activation (yellow) in the Topoplot. In the third row, an E_I artifact is the CNN input and the E_I CNN grad-cam shows that the positive values are well located on the artifact while the rest of the map shows negative values. Finally, in the last row an UBS Topoplot is the CNN input: none CNN recognizes it as its own.

To check in depth the behaviour of the framework, another Topoplot dataset was generated from subjects 9-19 of the DEAP dataset and 1/10 of them (a total of 29500), selected at random, was submitted to the 5 human experts for visual classification and labelling. The consensus contained 22795 UBS, 2190 B_V, 526 H_E and 3871 E_I, respectively. A set of 117 Topoplots was discarded due to tie. Then, automatic recognition was performed using the proposed framework: the

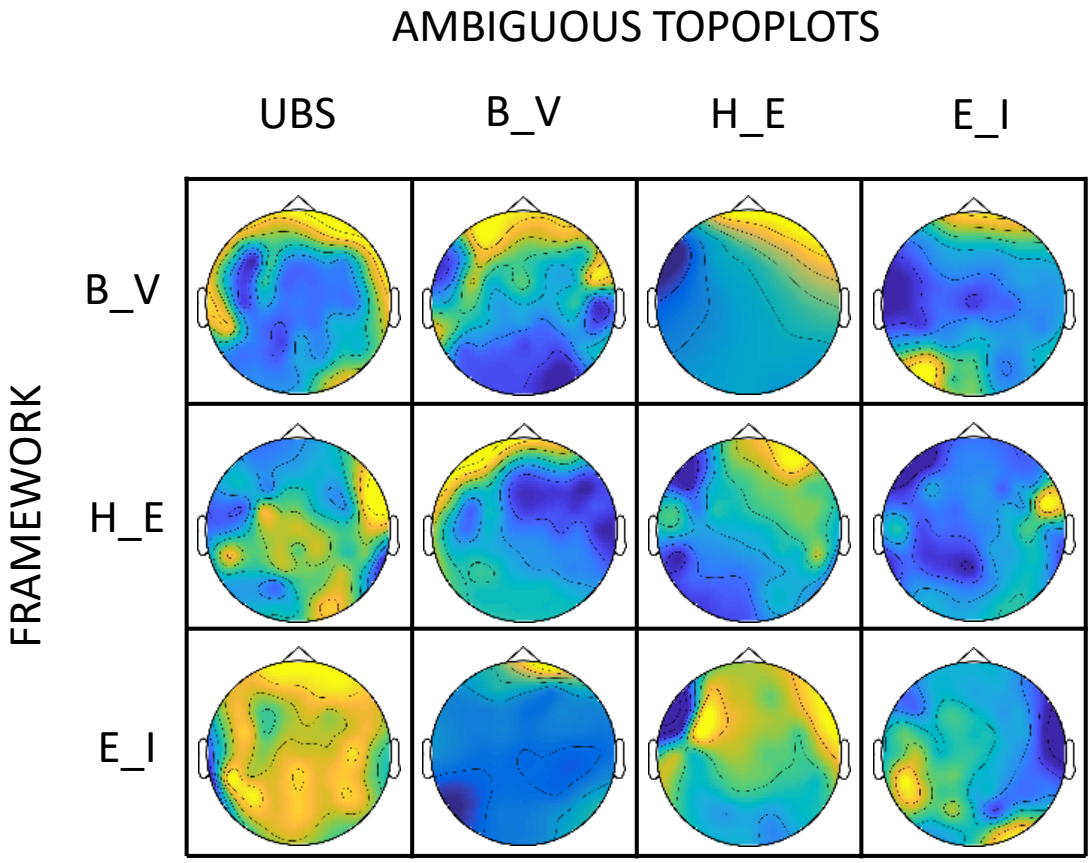

Figure 1.9: Examples of ambiguous Topoplots. The meaning is the same as for Table 1.3 with Topoplots instead of numbers and the exclusion of the columns "TOTAL" and "PERFORMANCE".

resulting errors, compared to manual consensus, and the performance of each CNN are reported in Table 1.3.

In particular, for each classifier (rows) the number of Topoplots wrongly classified as proper but belonging to the other classes, the false positives, are reported (columns). When the class matches the classifier, the corresponding cell contains the number of false negatives (FN). For each classifier, the right part of Table 1.3 respectively reports accuracy, sensitivity and specificity.

Performance data show that our framework has very good accuracy, sensitivity and specificity (>98%). Moreover, we verified that humans experts disagreed on most Topoplots misclassified by the framework, thus confirming that most of them exhibited ambiguous patterns. Examples of ambiguous Topoplots are reported in Fig 1.9.

Although the performance of the proposed method did not differ from that of a human expert (see above), we also decided to make an indirect comparison between our framework and the classification strategy proposed by Pion-Tonachini et al. [25]. The comparison was indirect because we did not implement the Pion-Tonachini's strategy but we used the online tools of the web-site (https://labeling.ucsd.edu/tutorial) that the Authors decided to share. To this end, we asked to our 5 human experts (the same ones who helped us build the labelled dataset) to self-instruct on how to use PSD and autocorrelation to classify IC in addition to Topoplots.

Then, for a dataset of 1000 artifacts we generated Topoplot, PSD and autocorrelation and we asked the experts to perform a manual classification both using only Topoplots and the three types of information with the method learned on the web-site (two different sortings of dataset were used). In this way, we had the opportunity to compare the proposed framework in terms of performance both with respect to the strategy on which it was trained and to a classification strategy that involves multiple sources of information. Moreover, we were able to ascertain the added value of multiple information.

Two consensus were generated from manual classifications: one among Topoplot-only classifications (MCT) and the other among multimodal classifications (MCTPA). The two consensus were compared: they agreed at 95.4%, although the time spent on each classification of MCT was about 1/5 of that required for one of MCTPA. The strong consensus agreement shows that multiple information could only contribute a small percentage towards improved performance. Indeed, the differences between the two consensus are in line with disagreement among experts (see above) and do not change significantly when multiple information are used: the huge increase in classification time is unwarranted, especially where speed is a requirement. This is confirmed by the performance results summarized in Table 1.4. As can be noticed, the proposed framework confirms the performance compared to human

experts on MCT (almost the same as Table 1.3). However, due to differences in human decisions when including further information, the performance of the proposed framework decreases compared to MCTPA. Although reduced, performance remains above 96%, thus confirming a behaviour of the framework similar to that of a human expert even when compared with MCTPA. Therefore, the use of the proposed framework on fast-response EEG is completely justified because the analysis of the Topoplots alone is sufficiently specific to define the nature of IC, as confirmed also in [45, 46, 10].

		CLASSIFICATION ERRORS (#)						PERF. (%)		
		UBS	**B_V**	**H_E**	**E_I**	**TOTAL**		**Acc.**	**Sens.**	**Spec.**
Framework	**B_V**	346	20 (FN)	17	235	618		98.7	99.0	98.7
	H_E	114	16	9 (FN)	22	161		99.5	98.2	99.6
	E_I	162	134	138	73 (FN)	507		99.1	98.1	99.2

Table 1.3: Recognition errors for each CNN (rows) when acting on a labelled dataset of Topoplots generated from subjects 9-19 of DEAP dataset. False positives (FP) are those classified as proper by a CNN although belonging to another class (columns). False negatives (explicitly indicated with FN close to the number to distinguish them from FP) are reported in the cells where the classifier corresponds to the class. The "TOTAL" column contains the sum of the values on the left (FP+FN). The last three columns show, for each classifier, accuracy, sensitivity and specificity, respectively.

		MCT PERF. (%)				MCTPA PERF. (%)		
		ACC.	**SENS.**	**SPEC.**		**ACC.**	**SENS.**	**SPEC.**
Framework	**B_V**	98.5	98.9	98.8		96.8	97.0	96.4
	H_E	99.6	98.4	99.6		97.7	97.2	98.1
	E_I	98.8	98.3	99.1		96.4	97.2	96.8

Table 1.4: Performance of the proposed framework compared with MCT and MCTPA, respectively.

In terms of computational performance, the proposed framework took 1.4 sec for 32 Topoplots, of which: 0.3 sec to calculate the IC with fast-ICA [66] ; 0.9 sec to generate Topoplots; 0.21 sec for classification. As it can be seen, the bottleneck is represented by the Topoplot generation (necessary both for manual analysis and for automated strategies). Fortunately, the time for Topoplot generation does not increase linearly with the number of images because some calculations, such as those required for channel positioning on the scalp model, are executed only once. Compared to the strategy in [25], which, in addition to Topoplots, has to analyze further information, the proposed framework is simpler and probably more efficient. Furthermore, although no direct indication of efficiency was given in [25], the number of network parameters used in [25] is equal to $2.8x10^6$, about 2.5 times higher than our architecture. The computational results show that our framework is sufficiently fast to be compatible with fast EEG-applications, being the pre-processing faster than the time it takes to measure the signal to be analyzed (one sub-trial in Fig 1.3 took 8 sec, 4 of them are of "present" signal), thus making artifacts recognition effective and timely for interactive BCI [67, 68, 69].

1.5 Discussion

Artifact recognition/classification from IC Topoplots is fundamental and is considered the gold standard for this purpose. To this end, we demonstrated that a fully automated framework, based on CNN operating only on IC Topoplots, is effective, fast enough to be used in interactive EEG, easily scalable and robust, without any additional information. The proposed framework has high accuracy (∼99%), specificity (∼98%) and sensitivity (∼99%), is very close to the performance of human experts and its results are in line with the most advanced methods [25]. Our framework, thanks to its scalable structure, fits perfectly with the present and future requirements of artifacts recognition because it can be easily adapted and trained to manage future artifact patterns without re-training the existing architecture (specific CNN must be added to the framework and trained separately).

The proposed framework is capable of operating in 1.4 sec for 32 Topoplots (including the time necessary to generate Topoplots), fast if compared to the time necessary to collect EEG signals for an EEG-based BCI (usually between 2.5-5 sec).The execution time can be further reduced if fewer sensors are used (as usual in BCI) without any framework modification.

Chapter 2
Analysis and Interpretation of Weak EEG Signals of 3M Syndrome Infants

3M syndrome is a rare disorder that involves the gene cullin-7 (CUL7). CUL7 modulates odour detection, conditions the olfactory response (OR) and plays a role in the development of the olfactory system. Despite this, there are no direct studies on olfactory functional effects in 3M syndrome. This is because obtaining high-quality EEG signals from infant recordings, compared to adults, is very difficult. In fact, EEG signals from newborns are very much noisy and greatly affected by artifacts mainly caused by uncontrollable movements. Artifact removal is crucial and preliminary to any form of possible interpretation. In this chapter, a framework for the interpretation of EEG signals of two twins infants affected by 3M syndrome and one healthy infant, recorded during olfactory stimuli, is described. The purpose was to analyse the cortical OR through chemosensory event-related potentials (CSERPs) and power spectra calculated EEG signals. The resulting analysis has been possible thanks to the preliminary usage of the artifact removal framework presented in the Chapter 1.

The content of this chapter appeared in [70].

2.1 Introduction

3M syndrome is a "rare autosomal recessive dwarf syndrome" [71]. The distinctive features of this little-known syndrome are limited prenatal growth, facial dysmorphism, absence of microcephaly and cognitive impairment. Since 3M syndrome is autosomal recessive, both inherited copies of the gene have mutations. Mutually exclusive genetic mutations in cullin-7 (CUL7), obscurin-like 1 (OBSL1) and coiled-coil domain-containing protein 8 (CCDC8) cause the pathology, as confirmed by a study conducted by Dan Hanson and collaborators [72]. They noted that, in terms of the clinical and biochemical 3M syndrome phenotype, children with CUL7 mutations were significantly shorter than those with OBSL1 or CCDC8 mutations. However, the aetiological mechanisms that lead to the observed growth disability

in 3M syndrome remain unclear, but they are probably related to abnormalities in basic cell growth and changes in cellular responses to growth factor stimulation.

Although 3M syndrome is considered a relatively rare disease, it is probably an under-recognised condition; its main characteristics, including impaired pre- and post-natal growth, are shared with all gestational age children with growth failure. This population includes many children who do not yet have a clear mechanism of growth impairment [73]. It is likely that 3M syndrome is often misdiagnosed or unrecognised due to normal mental development, mild dysmorphic facial features and good patient health.

Residual clinical features (triangular face, pointed chin, mouth and prominent lips, fleshy nose with anteverted nostrils, short stature, large skull and prominent forehead) and clinical history (low birth weight) are typical of 3M syndrome [71]. Epidemiological data about 3M syndrome are not known. Today, approximately 200 cases have been reported worldwide [74].

2.1.1 3M Syndrome and Potential Olfactory Involvement

As mentioned above, genetically confirmed patients with 3M syndrome carry mutations in CCDC8 (5%), OBSL1 (25%) or CUL7 (70%) [75]. CUL7 interacts with other cellular proteins and contributes to the formation of an E3 ubiquitin ligase complex that ubiquitinates specific targets. CUL7 mutations may disrupt insulin-like growth factor 1 (IGF-1) and growth hormone (GH) signalling pathways and contribute to growth alteration [74]. Insulin receptor substrate 1 (IRS1) is a target of the CUL7-SCF ubiquitin ligase. IRS1 is a signalling molecule that is a member of a family of adaptor molecules downstream of GH, IGF-1 and insulin receptors [75, 76]. Insulin receptors are expressed in olfactory receptor neurones of rat olfactory mucosa, a fact that suggest insulin plays a role in odour detection modulation at the olfactory mucosa level [77, 78]. CUL7-FBXW8 is a component of an E3 ubiquitin ligase that localises to the Golgi apparatus in neurones and is required for dendrite growth and organisation. Inhibition of this ligase in neurones alters Golgi morphology, impairs vesicle trafficking and disrupts dendrite morphogenesis and arborisation [79]. The ubiquitin ligase activity is linked to axon guidance during pathfinding in the development of the olfactory system.

2.1.2 Olfactory Perception and Chemosensory Event-Related Potentials (CSERPs) in Infants

Olfactory perception is highly developed in newborns and infants. Recent research indicates that olfactory system activity is already present in 1-day-old newborns

[79, 80]. Furthermore, smells can modulate nociception [81], by inducing greater stability during painful procedures and lower severity of central apnoea. Moreover, unpleasant or irritating odours promote disadvantageous evolutionary responses, such as decreased respiratory rate (up to apnoea) [82]. The sense of smell is also compromised in children with cerebral malformations, genetic diseases (e.g., trisomy 13 or 18, Kallmann syndrome or Riley-Day syndrome), endocrine disorders (such as hypothyroidism and gonadal dysgeneses) and in infants borne to diabetic mothers [83]. A recent work showed that it is possible to record olfactory event-related potentials (OERPs) in infants [84]. OERPs and CSERPs are electrophysiological components that allow researchers to evaluate chemosensory and chemoperceptual responses to olfactory stimuli [85]. The main difference between OERP and CSERPs is that the former is elicited by purely olfactory stimulation, while the latter is elicited by chemical stimulation, which may also include trigeminal activation [86]. Schriever and colleagues research, however, highlights the difficulty in observing OERPs in infants. This phenomenon is likely because there are more recording artefacts. OERP components in infants are the same as in adults: early components N1 and P2 [16] and late positive components (LPCs) [87]. N1 and P2 are the early sensorial components and are modulated by stimulus concentration and typology. LPCs include P3a and P3b and are modulated by the cognitive aspects of the stimulus (e.g., presentation frequency or stimulus salience) [87]. Moreover, time-frequency analysis highlights increased low frequencies (4–7 Hz) in a temporal range that corresponds to LPC [84] Even though there are no previous ERP (and specifically CSERPs) studies in 3M infants, one could hypothesise that the olfactory system could be dysfunctional in infants with CUL7 mutations [88]. Based on the integrated CSERPs approach, the aim of this study was to investigate whether there are implications at the level of olfactory perception both in OERP components and with regard to the main rhythms associated with rhinoencephalon [89] and entorhinal cortex [90] activity in 3M syndrome. Since no study has evaluated the use of CSERPs or OERPs to investigate olfactory functional responses in 3M infants, olfactory function in this rare syndrome is poorly characterised. Moreover, to the best of our knowledge, no study has been conducted using electroencephalogram (EEG) signals from the subjects with 3M syndrome using signal processing and analysis strategies. There are multiple potential benefits from this study. If the 3M syndrome subjects differ from the controls with respect to the olfactory response, early OERP screening, which represents an economic and non-invasive tool compared to genetic screening, could then lead to a possible subsequent genetic investigation (if it is positive). Furthermore, this research could allow us to deduce functional CUL7 involvement in the human chemosensory/olfactory response, a prospect that has not yet been studied.

2.2 Materials and Methods

The research was conducted at the Neurology Unit of the Vito Fazzi Hospital in Lecce with subjects recruited at the Neonatal Intensive Care Unit (UTIN). Data collection was performed in compliance with the Code of Ethics of the World Medical Association (Declaration of Helsinki) and authorized by the ASL-Lecce Ethics Committee (Approval record N°7, Date 19 July 2017). Written informed consent was obtained from the parents.

2.2.1 Subjects

Three subjects (males) with a diagnosis of 3M syndrome were recruited for the study. The subjects were siblings, two 5-month-old twins (3M-N) and their 18-month-old brother (3M-O). CUL7 genetic analysis (exons 14-23/24) highlighted the presence of pathogenic variants c2781delC (p.Ser928Leufs*5) and c.4391 A>C (p.His1464Pro) in the state of compound heterozygosity. The diagnostic conclusion for all three siblings was 3M syndrome due to familiar mutations. The laboratory data is compatible with the segregation of the family pathology in the foetus. The 3M syndrome group presented the following medical history: prematurity, low birth weight (LBW), small size for gestational age, syndromic facies, triangular face, prominent frontal drafts, bulbous nose, flat angiomas of the median line, short neck and thorax, hypospadias and suspected bow curvature, fleshy and prominent heels, prenatal 3M diagnosis based on amniocentesis karyotype, glandular hypospadias, transient hypocalcaemia and transient oliguria. Moreover, the subjects with 3M syndrome showed a larger cranial circumference (75–90°). Our sample size represents about 1.5% of the approximately 200 cases described worldwide [74]. The control group was recruited with the criterion of having the same gestational and post-conceptional ages as the 3M subjects, but no apparent clinical abnormalities from anamnestic data. The controls consisted of two healthy 12-month-old male twins (HS-O) and two 4-month-old male twins (HS-N).

The subjects were compared, separately, according the post conceptional age; the HS-O twin pair served as controls for the 3M-O and the two HS-N served as controls for the 3M-N. Independent comparisons were performed, due to different characteristics from neonatal EEGs and developmental brain behaviour [91]. Additionally, all infants were examined by the hospital paediatrician to rule out nasal congestion or other temporary respiratory diseases. Both healthy and subjects with 3M syndrome were subjected to neonatal auditory screening and the auditory brainstem response (ABR) and did not show any significant clinical abnormalities. No other behavioural olfactory or chemosensory assessment was performed.

2.2.2 OERP Assessment

Subjects performed a CSERPs task that involved the eucalyptus scent (natural eucalyptol oil, 1,3,3-Trimethyl-2-oxabicyclo [2.2.2] octane; Sigma-Aldrich, CAS Number 470-82-6). The experimental eucalyptus concentration was 20 μL in 10 mL Vaseline oil. The odorous solutions were prepared in 20 mL transparent glass vials and kept sealed with plastic film in a darkened cabinet. The scent was administered via an olfactometer [92]. The OERP presentation paradigm consisted of sequences of olfactory stimulations; each stimulation lasted 340 ms, with an inter-stimulus interval (ISI) of 20 s. In total, the subjects were exposed to 20 stimulations (a sufficient number of stimulations, because the minimum number to elicit OERP is 8 [84]). In accordance with recommendations based on previous research, the ISI was greater than 10 s to avoid habituation [88]. The device used to record the presentation of odorous stimuli allowed us to measure, in a controlled and automated way, the CSERPs evoked by olfactory stimuli synchronized to the acquisition of the EEG signal. The administration of the odorant, which took place through the olfactometer, was presented through a plexiglass tube that was positioned in the centre of the two nostrils. The odorant was delivered as binarinal stimuli in front of the nose. During the electroencephalographic recording, the children were seated in the arms of the mother, who in turn was sitting in a comfortable armchair placed inside the EEG recording room. The children were in a relaxed condition, were in a post-prandial state (i.e., they had eaten about an hour before the EEG recording) [93] and were in a waking state [94]. The choice of the eucalyptol odorant, which has a mixed component both olfactory and trigeminal, allowed us to keep the children in arousal during the CSERPs recordings [95, 96, 97, 98].

2.2.3 EEG Recording

The EEG signals were recorded using a Micromed 19-channel amplifier (Fp1; Fp2; F7; F3; Fz; F4; F8; T3; C3; Cz; C4; T4; T5; P3; Pz; P4; T6; O1; and O2). The scalp electrodes were applied according to the International 10-20 system. The EEG signal processing was performed using a Brain Vision Analyzer (Brain Products GmbH). The impedance was maintained below 8 kΩ, and the sampling rate was 256 Hz.

2.2.4 OERP Pre-Processing

The electrodes were online referenced to FCz, and offline, they were postponed with a common offline reference [99]. The signal was filtered offline (0.01–50 Hz, 24 dB), and the artefact rejection threshold was set to > 125 [32]. ERP epochs included a 100-ms pre-stimulus reference period and a 500-ms post-stimulus segment. The

peaks were automatically detected for all channels. The OERP components were labelled as N1 and LPC according to Pause et al. [87]. The latency windows were set to 100–400 ms for N1 and 350–600 ms for LPC [88, 100]. Main regions of interest (ROIs) were extracted though the linear derivation process: central left (C3-A1-T3), central right (C4-A2-T4), temporo-parietal left (P3-T5-O1), temporo-parietal right, (P4-T6-O2), frontal left (Fp1-F3-F7), frontal right (Fp2-F4-F8), central (Cz), parietal (Pz) and frontal (Fz). This process was defined a priori to reduce the number of electrode/channel comparisons, according to the definition of the two hemispheres and the lobes [101]. The linear derivation process allows one to synthetize new channels from linear combinations of recording existing electrodes/channels [102].

2.2.5 EEG Signal Pre-Processing

We further analysed the original EEG signals with signal processing strategies, since the sample was necessarily small and the study could be reduced exclusively to a single case. Thus, we investigated EEG rhythms on pieces (trials) of signal collected after each olfactory stimulation by searching for the presence of recurrent common trends in the subjects with 3M syndrome with respect to the controls.

We divided the signals in trials of 10 seconds (to ensure the convergence of the ICA algorithm)

First, to correct different amplification effects, each trial was normalised with respect to its baseline level, obtained by calculating the mean value of the power spectrum in a frequency band (65–75 Hz), which is usually only occupied by noise. Thus, all the resulting trials showed the same amplification. Then, the signal was subjected to a band-pass filter (0.01–49 Hz, 24 dB) in the frequency domain in order to eliminate noise and offset.

Finally, each trial was elaborated for eliminating artifacts. To this end, ICA was performed and the components were calculated, trasformed in topoplot and classified with the method presented in Chapter 1.

The residual components were projected back to the signal space to obtain a filtered version of the signal that composed each trial.

Examples of artifacts removal, for Trial 5, 20 and 26, are reported in Figures 2.1, 2.2 and 2.3.

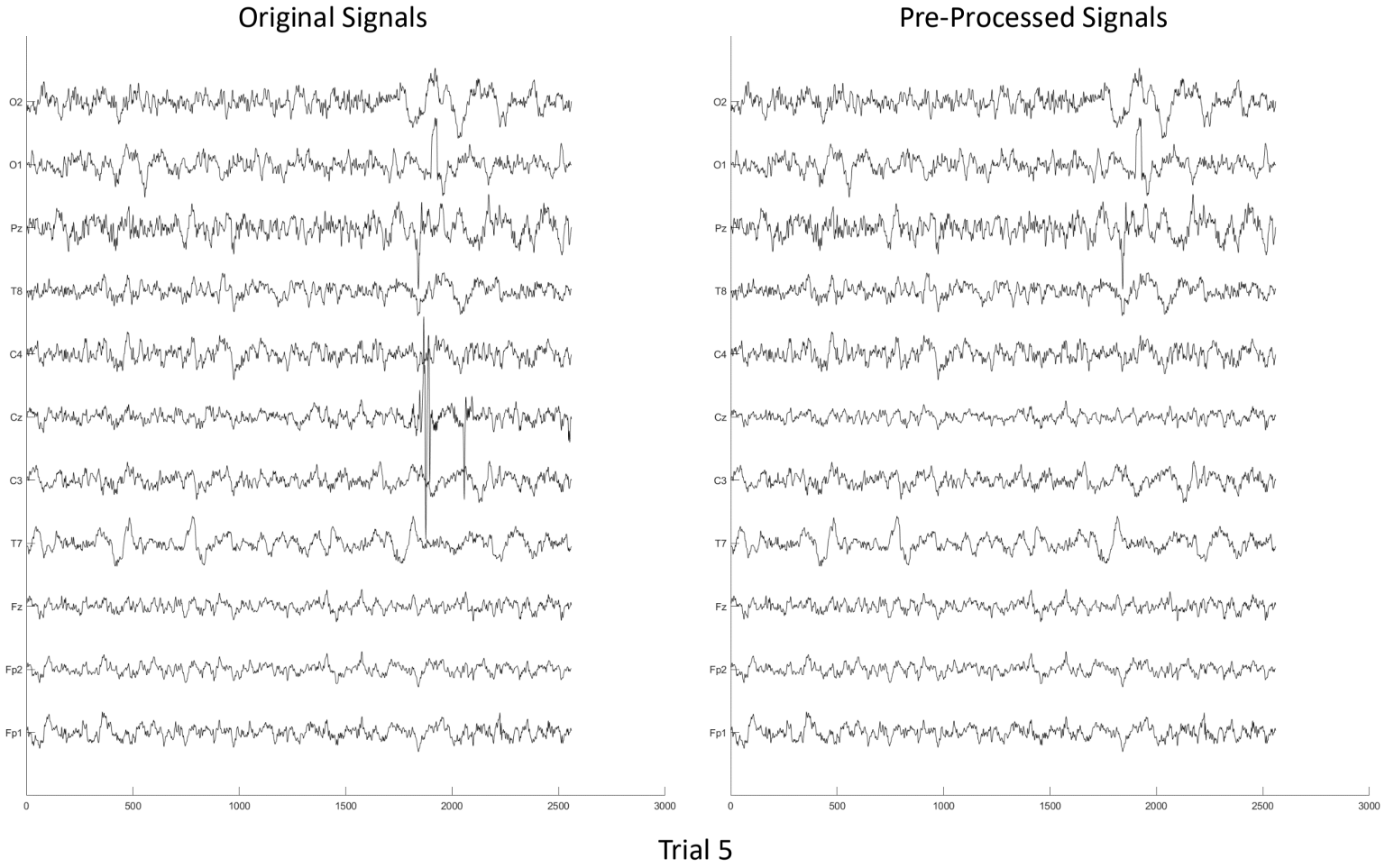

Figure 2.1: Respect to the original signals, spikes present in channels 02, 01, C4, Cz and C3 were removed.

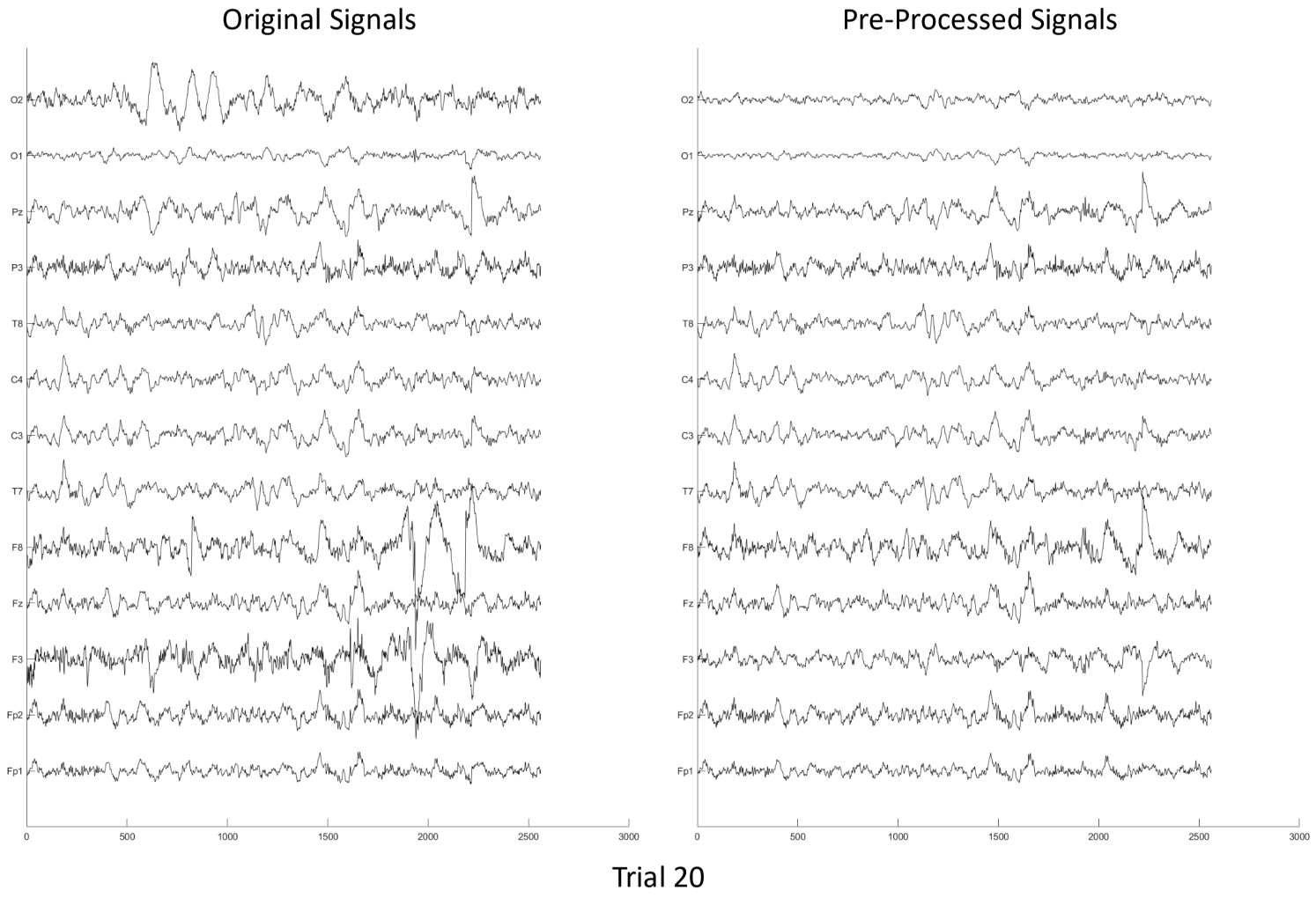

Figure 2.2: Respect to the original signals, spikes present in channels F8 and F3 were removed.

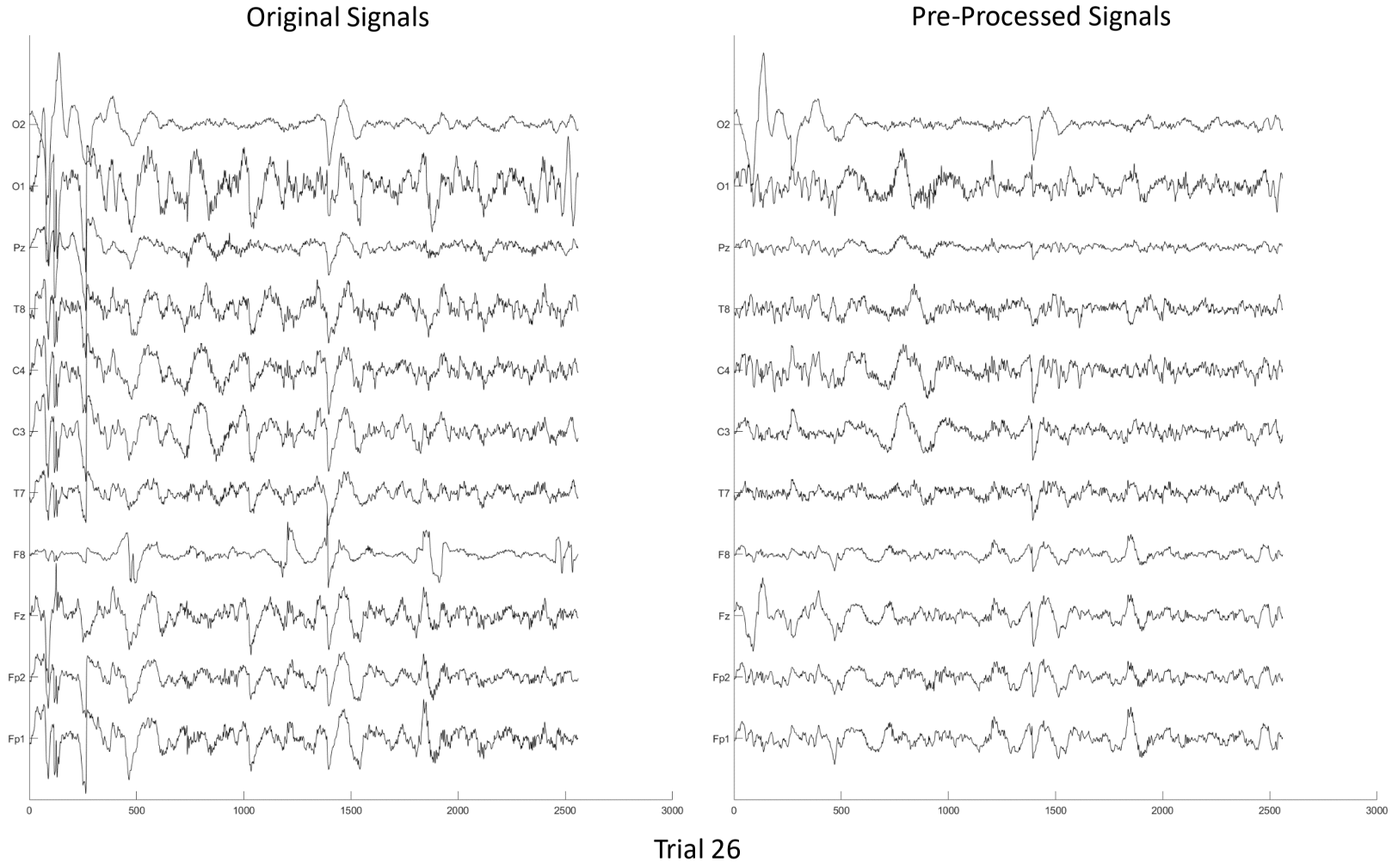

Figure 2.3: Respect to the original signals, spikes present in almost all channels were removed.

2.3 Results

2.3.1 OERP Data Analysis

Due to the sample size, we performed initial explorative and descriptive analyses to investigate N1 and LPC OERP components (Tables 2.1 and 2.2). The OERP results revealed that 3M-O showed greater N1 amplitudes and faster latencies on frontal left (3M-O 17.34 μV vs HS-O 1.42 μV), frontal right (3M-O -25.37 μV vs HS-O -12.55 μV), central left (3M-O -11.17 μV vs HS-O -4.1 μV), central right (3M-O -7.97 μV vs HS-O -2.7 μV) and temporal left (3M-O -6.8 μV vs HS-O -2.2 μV). Cz showed faster latency (3M-O 109 ms vs 129 ms) and Pz showed greater amplitude (3M-O -2.80 μV vs HS-O -1.34 μV) in 3M-O. LPC data followed the same pattern as N1, except for central right (3M-O 9.75 μV vs HS-O 12.23 μV), Fz (3M-O 5.78 μV vs HS-O 9.58 μV) and Pz (3M-O 4.26 μV vs HS-O 22.06 μV), where 3M-O had a decreased amplitude.

	Group	N1 Amplitude	N1 Latency	LPC Amplitude	LPC Latency
Frontal Left	3M-O	-17.34	66	33.32	191
	HS-O	-1.42	137	10.05	328
Frontal Right	3M-O	-25.37	125	13.40	230
	HS-O	-12.55	234	3.85	188
Central Left	3M-O	-11.17	141	24.02	180
	HS-O	-4.10	160	12.9	402
Central Right	3M-O	-7.97	164	9.75	282
	HS-O	-2.70	102	12.23	203
Temporal Left	3M-O	-6.80	140	24.4	188
	HS-O	-2.20	105	18.28	227
Temporal Right	3M-O	-4.52	148	14.83	2,93
	HS-O	-14.5	113	8.83	2,54
Cz	3M-O	-4.28	109	1.51	328
	HS-O	-11.80	129	1.20	262
Fz	3M-O	-6.97	164	2.57	254
	HS-O	-28.24	133	9.38	191
Pz	3M-O	-2.80	160	4.26	246
	HS-O	-1.34	113	22.06	156

Table 2.1: Results of descriptive analysis of amplitude (μV) and latency (ms) of N1 and Late Positive Component (LPC) in 18-month-old 3M-O and HS-O subjects.

	Group	N1 Amplitude	N1 Latency	LPC Amplitude	LPC Latency
Frontal Left	3M-N	-3.23	121	6.56	267
	HS-N	-8.94	203	4.89	297
Frontal Right	3M-N	-9.32	188	–	–
	HS-N	-2.38	184	–	–
Central Left	3M-N	-7.68	121	5.88	262
	HS-N	-5.76	168	4.19	297
Central Right	3M-N	-1.47	195	1.30	215
	HS-N	-6.49	148	–	–
Temporal Left	3M-N	-5.85	105	0.311	297
	HS-N	-9.26	172	5.08	316
Temporal Right	3M-N	-0.988	137	6.07	207
	HS-N	-2.15	164	7.85	270
Cz	3M-N	-16.02	180	–	–
	HS-N	-11.80	129	1.20	262
Fz	3M-N	-8.95	109	5.78	160
	HS-N	-6.67	223	9.58	250
Pz	3M-O	-6.12	0.27	14.27	164
	HS-O	-9.96	105	5.58	188

Table 2.2: Results of the descriptive analysis of the averaged amplitude (µV) and latency (ms) for N1 and LPC in 3M-N and HS-N. Two dashed lines indicate a lack of signal

2.3.2 EEG Spectral Analysis

After pre-processing, each trial was analysed 0–1000 ms after onset (since the brain response signal is zero 1 s after olfactory stimulation). The resulting signals, each sampled for 1 s at 256 points, were analysed with Fourier transform (FT) in 4 Hz windows (0.01–4, 4–8, 8–12, 12–16, etc., until 48 Hz), and the power spectrum was calculated. The analysis was performed for each trial and each channel separately, and the results were analysed in the form of a power spectrum represented graphically and as topoplot images. The obtained results demonstrated that the frequencies generated by olfactory stimulations mostly occurred in the (0.01,8] Hz interval in both the subjects with 3M syndrome and healthy subjects. However, the examined subjects with 3M syndrome had low frequencies ($\leq$ 4 Hz) elicited by olfactory stimulation, while higher frequencies ($>$ 4 Hz) were mostly activated for healthy subjects. Figure 2.4 highlights these aspects by reporting, for all subjects and for each channel, the percentage of trials for which 60% of the EEG power spectrum area was in the (0.01, 4] Hz interval (green) or $>$ 4 Hz (blue). Vertical red lines indicate the mean percentage (averaged for all ROIs) of 60% of the power that occurred before 4 Hz. This presentation clearly shows a right displacement for patients with 3M syndrome with respect to the corresponding controls. These data confirm, for subjects with 3M syndrome, the increment of trials for which the power spectrum concentrated in the (0.01, 4) Hz interval.

This effect was most noticeable between 3M-O (row #1, column #1) and HS-O (row #1, columns #2 and #3) with respect to 3M-N (rows #2 and #3) and HS-N (row #2, columns #2 and #3). Moreover, some ROIs were more involved than others in this process, as shown in Figure 2.5, which reports the percentage difference of the green area in Figure 2.4 between patients with 3M syndrome and corresponding controls (and separated by ROIs). Table 2.3 describes a ROI analytic evaluation showing percentage difference between the green regions (area of the power spectrum $\leq$ 4 Hz) in Figure 2.4 for subjects with 3M syndrome and control subjects; Table 2.3 confirms that, for 3M-O, the power spectrum was concentrated in the (0.01, 4] Hz interval (positive values, highlighted in "orange") for all ROIs with respect to both controls. However, for 3M-N, this behaviour was confirmed just for some specific ROIs (regions with discordant signs were not considered).

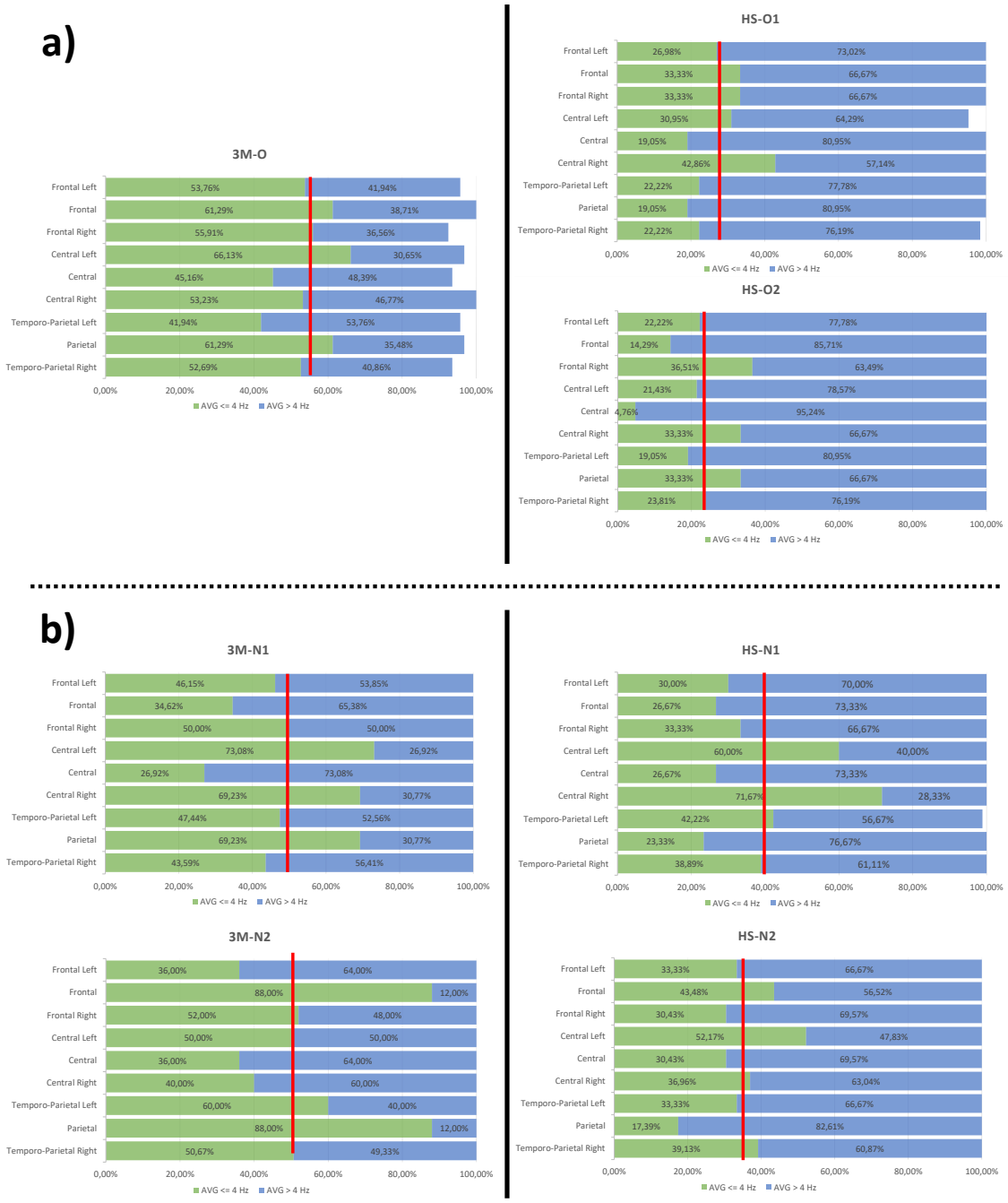

Figure 2.4: Representation of the percentage of trials (horizontal axis) divided by ROIs (vertical axis), for which 60% of the power spectrum area was ≤ 4 Hz (green) or > 4 Hz (blue) for each of the treated infants. A sum less than 100% indicates that some trials were too corrupted to be treated and, hence, discarded; this phenomenon mainly occurred for subject 3M-O. Data regarding subject 3M-O and the corresponding controls HS-O1/HS-O2 are reported in a) and 3M-N1/3M-N2 and the corresponding shared controls are reported in b). Vertical bars indicate the average threshold; differences are apparent between patients and control

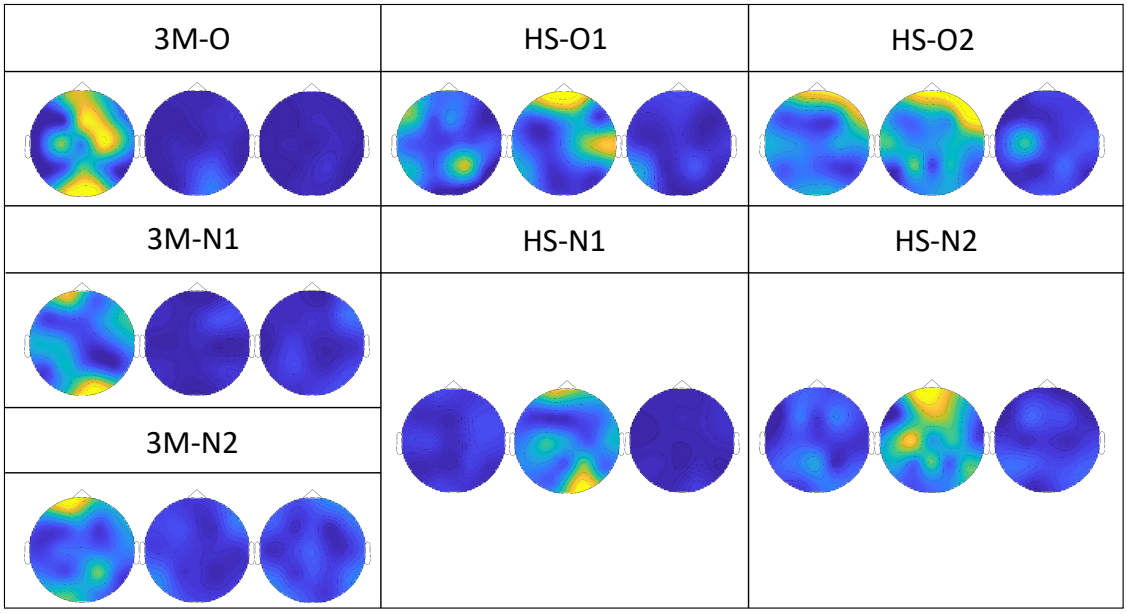

Figure 2.5: Topoplot images that report the power spectrum distribution of one of the typical trials for each infant. Each of the three topoplots refers to an analysed bandwidth: 0.01–4 Hz (left), 4–8 Hz (middle) and 8–12 Hz (right). The scale was normalized between 0 and 1 (0 = intense blue, 1 = intense yellow) for all subjects and is not shown for convenience. For patients with 3M syndrome (left column), the left topoplot (0.01–4 Hz) carried most of the power; for healthy subjects (middle and right columns), most of the power was concentrated in the middle topoplot (4–8 Hz). For all subjects, the right topoplot (8–12 Hz) contained negligible power with respect to the lower frequency windows.

Area	3M-O / HS-O1	3M-O / HS-O2	3M-N1 / HS-N1	3M-N1 / HS-N2	3M-N2 / HS-N1	3M-N2 / HS-N2
Temporo-Parietal Right	57,82%	54,81%	10,78%	10,23%	23,25%	22,77%
Parietal	68,92%	45,61%	66,30%	74,88%	73,48%	80,24%
Temporo-Parietal Left	47,01%	54,58%	10,99%	29,73%	29,63%	44,44%
Central Right	19,48%	37,37%	-3,52%	46,62%	-79,17%	7,61%
Central	57,82%	89,46%	0,95%	-13,04%	25,93%	15,46%
Central Left	53,19%	67,60%	17,89%	28,60%	-20,00%	-4,35%
Frontal Right	40,38%	34,71%	33,33%	39,13%	35,90%	41,47%
Frontal	45,61%	76,69%	22,96%	-25,60%	69,70%	50,59%
Frontal Left	49,81%	58,67%	35,00%	27,78%	16,67%	7,41%

Table 2.3: Percentage difference, by ROIs, between the green regions (area of the power spectrum ≤ 4 Hz) in Figure 2.4 for subjects with 3M syndrome and control subjects. Data that exhibit concordant positive values are highlighted in orange.

We observed larger differences in EEG spectral power displacement in the subjects with 3M syndrome, although this effect was different between 3M-O and 3M-N. In particular, 3M-O showed more activation in all the ROIs, but 3M-N showed more activation in temporo-parietal (right and left), parietal, frontal right and frontal left, where both clinical samples exhibited intersection. These results highlight that similar EEG patterns are present in the same clinical categorisation (e.g., 3M). The signal behaviour between 3M-O and its corresponding control is differentially distributed with respect to 3M-N and HS-N, although the general effect was the same. 3M-N had very similar responses; the different responses between patients of different ages were more pronounced than those between controls of different ages, where similar patterns, although at different scales, were maintained. Figure 2.5 shows the screenshots of typical power spectrum topoplots (with normalised scales) in the first three frequency windows (0.01–4 Hz, 4–8 Hz and 8–12 Hz, respectively) for one trial for each of the analysed subjects. Notably, Figure 2.5 confirms that patients with 3M syndrome responded more in the 0.01–4 Hz bandwidth, while healthy subjects were mostly active in the 4–8 Hz bandwidth. Power was negligible in the 8–12 Hz bandwidth for all subjects.

2.4 Discussion

3M syndrome is extremely rare and difficult to diagnose. Its peculiarity lies in bone alterations and genetic variations, which, among various aspects that these changes modulate, also affects the olfactory response. Indeed, CUL7, a gene involved in the 3M syndrome, can modulate odour detection and condition the OR and plays a role in the development of the olfactory system [78, 79]. Despite this involvement, there are no direct studies on the functional effects of this syndrome. This paucity of data is due to the fact that the syndrome is one of the rarest genetic disorders and evaluation of cortical responses to olfactory stimuli in infants and newborns is one of the less frequent investigations within psychophysiology and cognitive neuroscience [84].

The purpose of the present chapter was to analyse the cortical olfactory response, recorded through CSERPs, in infants with 3M syndrome. We first evaluated the CSERPs responses with a direct descriptive comparison (since this study was comparable to a single case study due to the small sample size) on the trends of the olfactory stimulus sensory and perceptive components. In particular, we analysed the sensory components N1 and LPC elicited by the stimulation paradigm [87]. The CSERPs results demonstrated that the 3M-O infant exhibited increased N1 amplitudes and faster latencies. Furthermore, we found a faster latency in Cz, which is positioned on the precentral gyrus, and greater amplitude in Pz, which is located in

the middle parietal lobe. The precentral gyrus and parietal cortex are considered sites for olfactory working memory. We interpret these findings as an indication of greater allocation of attentional resources, enhanced olfactory working memory and olfactory perception, which is visible in the N1 component, which is involved in the sensorial detection of olfactory stimulation [87]. We suppose that this enhancement could be related to the CUL7 alteration. LPC data followed the same results as N1, except for central right, where we observed a decreased amplitude. The wider LPC is a consequence of the processing of olfactory information visible through the N1 component. The 3M-N twins also showed increased amplitude in the precentral gyrus and faster N1 and LPC latencies, although the results in younger infants were apparently less defined and exhibited less LPC typing than for the 3M-O and control groups. This minor typing is evident with the difficulty of identifying the LPC in frontal right, central right and Cz ROIs [84].

In 3M syndrome, olfactory processing appears to be clearly diversified. Specifically, comparison of the N1 and LPC indicates substantial differences in 3M syndrome that may be a consequence of a modified olfactory processing pattern. Moreover, the subjects with 3M syndrome showed different arousal localisations from olfactory stimulation, data that implicate much larger areas that range from the left hemisphere to the midline sites (i.e., Fz, Cz and Pz). These differences were more distributed and evident in the infant rather than younger twins, but in general, they seemed to be constant with respect to the CSERPs trend. As a further signal control, we performed a new analysis based on the assumption that the slow and high CSERPs frequencies are related. Indeed, we considered the rhythms within the signal and considered the greater cortical response, which in our case coincided, at a temporal level, with the CSERPs-elicited response. These results demonstrated that the frequencies generated by olfactory stimulations were mostly present in the (0.01,8] Hz interval in subjects with 3M syndrome and healthy subjects. However, the behaviour observed in the examined subjects was that low frequencies, in particular δ (≤ 4 Hz) were elicited by olfactory stimulation in subjects with 3M syndrome, while higher frequencies (> 4 Hz) were mostly activated for healthy subjects. Moreover, we argue that some ROIs are more involved than others in this process. In particular, 3M-O showed involvement in all ROIs, although parietal, central left, central, frontal and frontal left exhibited greater activation; 3M-N showed elevated activation in temporo-parietal (left and right), parietal, frontal left and frontal right. Overall, similar EEG patterns were present for the same clinical categorisation (e.g., 3M-O and 3M-N). The δ EEG rhythm appears to be more structured in 3M-O, although the general effect in 3M-N was the same, but less strong. The different age-related responses in 3M infants were more pronounced than those between controls, where similar patterns were maintained in CSERPs and EEG spectral analysis.

Moreover, the presence of δ rhythms in patients with 3M syndrome clearly implicates olfactory response involvement, since this rhythm is closely connected to olfactory perception [103, 104].

Although we were unable to perform robust statistical analysis due to the limited number of subjects, our results are the first assessing, in a preliminary way, CSERPs in 3M subjects, and this could be of interest for basic research and clinicians. For basic research, these results highlight, for the first time in human infants, a functional aspect of the cortical olfactory response linked to the CUL7 gene. From the clinical point of view, these results suggest that a diagnostic evaluation of the cortical olfactory response at an early age may provide indications for subsequent genetic screening, which is more complex and expensive than a CSERPs assessment. The first limitation of this preliminary study comes from the sampling of 3M subjects. These subjects, in fact, belong to the same family, therefore they could show similar electrophysiological characteristics due to their familiarity and not due, exclusively, to the olfactory system, despite the peculiarity of these subjects is precisely having a variation of the CUL7 gene, closely connected with olfaction. The other limitations concern the chemical nature of stimulation and the sample size. In fact, regarding the first one, we did not use a purely olfactory stimulus (e.g., phenethyl alcohol) to prevent the child from relaxing and falling asleep during the EEG recording [98]. The administration of eucalyptus, in fact, on the one hand allowed us to keep the children in a state of mood increased vigilance, but on the other hand has ensured that the elicited component is of a mixed type (both olfactory and trigeminal) [86, 98] The sample size, even if it is a limitation, also partly represents a strength. The small number of subjects actually represents a larger percentage of subjects with 3M syndrome than the percentage that would usually be represented by afflicted individuals in clinical studies. We can conclude, albeit in a preliminary way, that the chemosensory investigation of this syndrome, could open new connections between purely clinical aspects, such as the identification of a potential biomarker, and basic research aspects, to understand how and at what time a genetic alteration can modify a sensory and subsequently perceptive and / or cognitive response.

Part III

Artificial Intelligence For MRI Analysis

Chapter 3
Materials And Methods

Preliminary to the design and development of any supervised AI model, at least 3 elements need to be considered:

- Data Set: it should contain the sufficient amount of data to represent all the potential patterns (and pattern variations) which the AI model must be able to cope with. Beyond quantity, also the quality of the data is fundamental.
- Data Labels: data annotation is the process of defining labels to the training data set. In MI, data annotation involves tagging specific biological images with tags identifying lesions. Data annotation requires a lot of work and is often done manually by a team of specialists, each one repeating the process independently from the others on the same data to cope with inter rater uncertainty. It provides that initial setup for training AI models.
- Scores and metrics: used to quantify the performance of the model. Many are necessary in order to evaluate the model under a broad range of characteristic parameters.

In this Chapter the Data Set used, the proposed Data Labels and the evaluations criteria are presented. The problem to be solved is the automatic recognition and segmentation of the lesions produced by multiple sclerosis from MRI.

3.1 Data Set

The Medical Image Computing and Computer Assisted Society (MICCAI) is a non-profit corporation founded on the 29 July 2004. The society mission is promote the research and the expertise in the field of medical image computing and computer assisted medical interventions including biomedical imaging. In order to do that, among all the activities, the MICCAI organization has created and maintained some of well know useful data set, such as the one used in this thesis for experimentation: the Multiple Sclerosis SEGmentation data set (MSSEG) [1].

Center	Scanner	Modality	Matrix	Slices	Voxel resolution (mm)
1	Siemens Verio 3T	Sagittal 3D FLAIR	512x512	144	0.5×0.5×1.1
		Sagittal 3D T1	256x256	176	1×1×1
		Axial 2D PD-T2	240x320	44	0.69×0.69×3
3	General Electrics Discovery 3T	Sagittal 3D FLAIR	512×512	224	0.47×0.47×0.9
		Sagittal 3D T1	512×512	248	0.47×0.47×0.6
		Axial 2D DP-T2	512×512	From 28 to 44	0.43×0.43×3 Gap: 0.5
7	Siemens Aera 1.5T	Sagittal 3D FLAIR	256×224	128	1.03×1.03×1.25
		Sagittal 3D T1	256×256	176	1.08×1.08×0.9
		Axial 2D PD-T2	320×320	25	0.72×0.72×4 Gap: 1.2
8	Philips Ingenia 3T	Sagittal 3D FLAIR	336×336	261	0.47×0.47×0.9
		Sagittal 3D T1	336×336	200	0.47×0.47×0.6
		Axial 2D PD-T2	512×512	46	0.43×0.43×3 Gap: 0.5

Table 3.1: Acquisition details for center. Table is from [1].

The MSSEG is composed by MR images collected from the following centers: University Hospital of Rennes (Center 1), University Hospital of Bordeaux (Center 3) and University Hospital of Lyon with 2 different scanners (Center 7 and Center 8). All the centers provided 4 acquisition modality: 3D fluid-attenuated inversion recovery (Flair), 3D T1 weighted pre and post-Gadolinium injection and axial dual proton density (PD) and T2 weighted.

The equipment used for each center is summarized in Table 3.1

Data were furnished both in unpreprocessed and in preprocessed form. Preprocessing refers to a series of mathematical adjustments to MR images before segmentation [105] for reducing the effects of noise and imaging artifacts, equalizing space, eliminating outliers and stabilizing the contrast. As previously stated, the segmentation from MRI is difficult due to the variability of imaging parameters, overlapping intensities, noise, gradients, motion, blurred edges, anatomical variations and susceptibility artifacts [106, 107]. For this reason, images undergo pre-processing to make classification robust with respect to imaging and scanners.

For the MSSEG data set, the preprocessed data consisted in performing the following steps:

- Denoising of each modality;
- Rigid registration of each modality on the FLAIR image;
- Brain extraction (skull stripping) from T1-w image and applied to other modalities;
- Bias field correction of each modality.

For each patient, The MSSEG data set includes 7 different manual delimitation (segmentation) of the lesions (each voxel was identified as lesion/not lesion) made

by 7 different expert radiologist. All the segmentation are merged into a consensus through a statistical fusion (Lop-STAPLE) [108, 109].

The Lop-STAPLE is an iterative algorithm, that fuse the input segmentations using the Expectation-Maximization approach. The peculiarities of Lop-Staple is that during the merging, it penalize the individual deviations from agreement between manual experts segmentations. Moreover, the algorithm is robust to differences between manual expert segmentations, and it allows the computation of agreement scores with respect to the consensus segmentation considered then as ground truth [1].

The MSSEG collects 53 MS patients MRI examinations divided in two groups: those from 15 patients were furnished to the research community for research purpose and the remaining 38 patients were maintained secret and used by the society to evaluate the performance of the methods participating to the challenge organized by them.

MS is a degenerative disease of the brain and spinal cord which can vary greatly between patients in severity and symptoms [110]. The majority of patients transit into a progressive phase consisting in an unremitting and progressive accumulation of disability. MS origins are not well understood but characteristic signs of tissue damages are recognizable, such as white matter lesions and brain atrophy or shrinkage due to degeneration. These signs can be observed by MRI which is a special tool to follow-up MS patients with reduced invasiveness due to the usage of specific contrast agents. In fact, focal lesions in the brain and spinal cord are primarily visible in the white matter on structural MRI observable as hyperintensities on T2-weighted, PD, and Flair images and as hypointensities, or "black holes", on T1-wheighted images [111].

Multiple sclerosis (MS) is a degenerative disease of the brain and spinal cord which can vary greatly between patients in severity and symptoms [110]. The majority of patients transit into a progressive phase consisting in an unremitting and progressive accumulation of disability. Actually there is no cure for MS and existing therapies focus on symptomatic management and prevention of further damage, with variable effectiveness, though recent advancements are promising. MS origins are not well understood but characteristic signs of tissue damages are recognizable, such as white matter lesions and brain atrophy or shrinkage due to degeneration. These signs can be observed by MRI which is a special tool to follow-up MS patients with reduced invasiveness due to the usage of specific contrast agents. In fact, focal lesions in the brain and spinal cord are primarily visible in the white matter on structural MRI observable as hyperintensities on T2-weighted images, PD or FLAIR images and as hypointensities, or "black holes", on T1-wheighted images [111].

Physicians often use FLAIR images for WM lesion detection and other modalities

mostly to ascertain the lesion stage. Complementary information is collected to visualize cortical lesions by means of MPRAGE and MP2RAGE imaging sequences [112, 113, 114].

An examination consists in thousands of images mostly collected pre and post contrast agent administration. MRI is used routinely in clinical practice but it is unspecific for MS and not well correlated to the clinical disability progression (physical and cognitive), to the neuro-plasticity and to the effects of demyelinization of nerves, the last being a critical effect which is invisible to MRI. Indeed, WM could appear normal though it has reduced myelin: for a MS patient, the "healthy" brain tissue is usually referred as "apparently healthy" [115]. Healthy anatomical structures similar to lesions and close to lesions could contribute to create further ambiguity

The MSSEG data set perfectly fits with the scope of this thesis because it contains images from different scanner, different modalities for each scanner and the original manual delimitation of each expert radiologist. For this reasons, the MSSEG data contains all the data necessary to investigate all the aspects discussed in the Introduction section.

3.2 Ternary ground-truth

In [1], the following sentence perfectly summarize the difficulties related to the development of the consensus from multiple individual manual segmentations: "*MS lesions segmentation is known to be expert and center-dependent, which can lead to relatively large discrepancies between individual manual segmentations*".

In fact, in medical imaging it is often made the simplifying assumption that there is a single, unknown, true segmentation map of the underlying anatomy, and each human rater produces an approximation with variations reflecting individual experience. The concept of a single-truth assumption may be correct when assuming that there exists only one (true) boundary of the physical objects captured in an image and the ambiguities in interpretation are due to human mistakes and disagreements. In the opposite case, it can be assumed that the variable annotations from experts are all realistic and acceptable instances of the true segmentation.

As it often occurs, the truth is in the middle: some ambiguities are indeed specific to human subjectivity or imperfections (extrinsic), while some others are due to the problem itself (intrinsic). Actually, both are important, but intrinsic ambiguities have the highest role, being due both to MS presentation and to MRI non specificity: lesions are not well separated from healthy tissue in MS (PVE) and MRI is neither sufficiently specific for MS nor sufficiently precise. Regarding human subjectivity, this produces differences that are due to a mix of prior assumptions, like

experience in the field, greater or lesser exploitation of additional meta-information (such as anatomical/radiological/clinical knowledge), mistakes or oversights which often are concentrated on small and/or low intensity lesions and lesion borders.

When raters are forced to provide a binary segmentation, as in MSSEG, they cannot express any doubt, whatsoever is the cause. The binary segmentation does not allow the representation of the intrinsic uncertainty and, furthermore, induces a human rater to assume polarized decisions which, from one side, could not correspond to what the rater really believes in and, from the other side, could be confusing and misleading for an automated strategy. In fact, ambiguous decisions might have been assumed by the rater in similar situations (an uncertain region could be considered healthy tissue in one case and lesion in another) which could influence the automated strategy [116].

For this reason, in order to train an automated method to recognize the intrinsic uncertainty of the problem, it is necessary to integrate the binary ground-truth with human uncertainty (doubts), making it robust to out-of-training-set examples and adversarial examples [117, 118].

3.2.1 Ternary ground-truth with Staple

Though we have used an implementation which is very similar to the original [108] we have redefined it and introduced little but significant variations to completely fit our scopes: in fact, we do not use the binary maps of the segmentation as input of STAPLE but the probability mass functions of the identifications. In fact, in each pixel there is not the final decision (1 for a lesion and 0 for a healthy tissue) but the probability that the pixel could allow to a lesion: this number could assume any number between 0 and 1 and, obviously, when the extremes occur it means that a net decision has been assumed. The output are the joint mass probability function derived from those of the two raters and the performance measure of both raters. In order to show that STAPLE can deal with the case we propose, we give an essential redefinition of the method by using notation necessary to the dimension of our problem, that is the presence of R raters (for specific details on STAPLE, please, refer to [108]). Consider an image of N voxels, and the task of classifying a structure in that image by indicating the probability of presence or absence of the structure at each voxel (that we also call fuzzy presence, where 0 indicates surely absence, 1 indicates surely presence and any other intermediate value indicate uncertain presence with that probability). In this way we could also model the human uncertainty with respect to some lesion candidates (doubtful region d). Let θ be an 3x3xR matrix where θ_j indicate, for the rater j, a 3x3 matrix (one entry for each class, 1, 0 and d) where each entry $\theta_{(j,s's)}$ indicates the probability that

rater j will decide the label s' when the true label is s. While the true label s is indicated by a single value (1 for lesion, 0 for no lesion and d for doubt), s' is not represented by a single value but by one of the following three continuous intervals:, $[1-\delta, 1]$ for lesion, $[0, \delta]$ for no lesion and $(\delta, 1-\delta)$ for doubt. The perfect rater would be characterized by having 1 in the diagonal and 0 elsewhere, that is an identity matrix. Let $\mathbf{D}$ be an NxR matrix describing probabilistic decisions (fuzzy) made at each voxel of the image by each rater. Let $\mathbf{T}$ be an indicator vector of elements representing the hidden binary true segmentation, where for each voxel the lesion is recorded as present (1) or absent (0) or doubt (d). Let the complete data be ($\mathbf{D}$,$\mathbf{T}$) and let the probability mass function of the complete data be $f(\mathbf{D}, \mathbf{T}|\theta)$. Our goal is to estimate the performance level parameters of the R raters characterized by θ which maximize the complete data log likelihood function

$$\theta = \arg\max_{\theta} \ln f(\mathbf{D}, \mathbf{T}|\theta) \tag{3.1}$$

Being Di, j the value of the measured voxel i by the rater j, then

$$\theta_{j,11} = p_j = Pr(D_{i,j} \geq 1 - \delta | T_i = 1) \tag{3.2}$$

represents the "lesion fuzzy fraction" (relative frequency of lesion outcome by rater j when the true outcome is lesion),

$$\theta_{j,22} = q_j = Pr(D_{i,j} \leq 0 + \delta | T_i = 0) \tag{3.3}$$

represents the "negative fuzzy fraction" (relative frequency of not-lesion outcome by rater j when the true outcome is not-lesion) and

$$\theta_{j,33} = r_j = Pr(0 + \delta < D_{i,j} < 1 - \delta | T_i = d) \tag{3.4}$$

represents the doubtful fuzzy fraction (relative frequency of doubt outcome by rater j when the true outcome is doubt). These three parameters define the principal diagonal of θ_j.

Obviously the parameters of $\theta_j \in [0, 1]$, are characteristics of the raters and are different for each rater and can be easily calculated when T is known, which is not the real situation. A reasonable assumption is that the segmentations performed by the R raters are conditionally independent each other, given the true segmentation $\mathbf{T}$ and the performance θ that is

$$(D_{i,j}|T_i, \vartheta_j) \perp (D_{i,j_1}|T_i, \vartheta_{j_1}), \ \forall \ j_1 \neq j \tag{3.5}$$

This assumption is justified by the fact that the R raters derive the segmentation of the same image independently each another and that the quality of segmentation

is captured by θ. Morevoer, raters are trained in a similar way and the decision about segmentation may differ mainly due to systematic differences between raters, the first having more to do with experience on lesions and the second most with information gained by the surrounding tissues. A probabilistic estimate of the true segmentation can be derived as a constructive combination of the probabilistic decisions assumed by the two raters. The expectation-maximization problem in Eq.1 can be solved iteratively by considering that some maximum likelihood problems would be simplified if some missing data are available, as in the case we are dealing with. The observable data, the segmentation decisions at each voxel, are incomplete because true segmentation is unknown and are regarded as an observable function of the complete data where also true segmentation is known. Here, the complete data is the segmentation probabilities D augmented with the true segmentation of each voxel T. T is called the missing or hidden data, and is unobservable. θ are the unknown parameters characterizing the performance of the two raters. According to the above definition of probability mass function of complete data, we write the complete data log likelihood function as

$$lnL_c\{\vartheta\} = \ln f(\mathbf{D}, \mathbf{T}|\vartheta) \tag{3.6}$$

The process to identify the expert quality parameters and ground truth consists of iterating between 1) estimation of the hidden ground truth given a previous estimate of the expert quality parameters, and 2) estimation of the expert quality parameters based on how they performed given the new estimate of the ground truth. This algorithm can be recognized as an expectation maximization (EM) algorithm, in which the parameters that maximize the log likelihood function are estimated based upon the expected value of the hidden ground truth. The EM algorithm approaches the problem of maximizing the incomplete data log likelihood equation

$$lnL_c\{\vartheta\} = \ln f(\mathbf{D}|\vartheta) \tag{3.7}$$

Since the complete data log likelihood function is not observable, it is replaced by its conditional expectation given the observable data D using the current estimate of θ. Computing the conditional expectation of the complete data log likelihood function is referred to as the E-step, and identifying the parameters that maximize this function is referred to as the M-step. The algorithm initialization can start by assuming that the experts are each equally good and have high values for p, q and r, though not infallible. This is equivalent to initializing the algorithm by estimating an initial ground truth as an equal weight combination of each of the expert segmentations. Another assumption is the voxelwise independence (the classification of each voxel is independent of the classification of close voxels). The requirements necessary to carry out the EM algorithm are to have a specification of the complete

data and to have the conditional probability density of the complete data given the observed data. At the iteration k, the problem in Eq. 3.1 can be rewritten as:

$$\begin{aligned}
\vartheta^{(k)} &= \arg\max_{\vartheta}\ E\left[\ln f\left(\mathbf{D},\mathbf{T}|\vartheta\right)|\mathbf{D},\vartheta^{(k-1)}\right] \\
&= \arg\max_{\vartheta}\ E\left[\ln \frac{f\left(\mathbf{D},\mathbf{T},\vartheta\right)}{f\left(\vartheta\right)}|\mathbf{D},\vartheta^{(k-1)}\right] \\
&= \arg\max_{\vartheta}\ E\left[\ln \frac{f\left(\mathbf{D},\mathbf{T},\vartheta\right) f\left(\mathbf{T},\vartheta\right)}{f\left(\mathbf{T},\vartheta\right) f\left(\vartheta\right)}|\mathbf{D},\vartheta^{(k-1)}\right] \\
&= \arg\max_{\vartheta}\ E\left[\ln \frac{f\left(\mathbf{D}|\mathbf{T},\vartheta\right) f\left(\mathbf{T}\right) f\left(\vartheta\right)}{f\left(\mathbf{T},\vartheta\right) f\left(\vartheta\right)}|\mathbf{D},\vartheta^{(k-1)}\right] \\
&= \arg\max_{\vartheta}\ E\left[\ln \frac{f\left(\mathbf{D}|\mathbf{T},\vartheta\right) f\left(\mathbf{T}\right)}{f\left(\mathbf{T},\vartheta\right)}|\mathbf{D},\vartheta^{(k-1)}\right]
\end{aligned} \tag{3.8}$$

with the assumption that T is independent of the performance parameters, that is $f\left(\mathbf{T},\vartheta\right) = f\left(\mathbf{T}\right) f\left(\vartheta\right)$.

The estimator of the unobserved true segmentation is derived by first deriving an expression for the conditional probability density function of the true segmentation given the expert decision and the previous estimate of the performance parameters:

$$\begin{aligned}
f\left(\mathbf{T}|\mathbf{D},\vartheta^{(k)}\right) &= \frac{f\left(\mathbf{D}|\mathbf{T},\vartheta^{(k-1)}\right)\mathbf{f}\left(\mathbf{T}\right)}{\sum_{T\prime} f\left(\mathbf{D}|\mathbf{T}\prime,\vartheta^{(k-1)}\right)\mathbf{f}\left(\mathbf{T}\prime\right)} \\
&= \frac{\prod_i \left[\prod_j f\left(D_{i,j}|T_i,{\vartheta_j}^{(k-1)}\right) f\left(T_i\right)\right]}{\sum_{T_1'}\cdots\sum_{T_N'}\prod_i \left[\prod_j f\left(D_{i,j}|T_i',{\vartheta_j}^{(k-1)}\right) f\left(T_i'\right)\right]}
\end{aligned} \tag{3.9}$$

and for each voxel i, we have

$$f\left(T_i|D_i,\vartheta^{(k)}\right) = \frac{\prod_j f\left(D_{i,j}|T_i,{\vartheta_j}^{(k-1)}\right) f\left(T_i\right)}{\sum_{T_i'}\prod_j f\left(D_{i,j}|T_i',{\vartheta_j}^{(k-1)}\right) f\left(T_i'\right)} \tag{3.10}$$

where $f\left(T_i\right)$ is the prior probability of T_i and the conditional independence of classifications allows to write the joint probability as a product of rater-specific probabilities. Again, a voxel-wise independence is used. The previous equation can be made explicit for $T_i = 0$, $T_i = 1$ and $T_i = d$. By using the definition of p_j, q_j and r_j and by considering that the sum of each column of ϑ_j, $1 - p_j$ is the probability of $\neq 1$ outcome by rater j when the true outcome is 1, that $1 - q_j$ is the probability of $\neq 0$ outcome by rater j when the true outcome is 0 and that $1 - r_j$ is the probability of $\neq d$ outcome by rater j when the true outcome is d, we can write

$$\begin{aligned}
a_i^{(k)} &\equiv f\left(T_i = 1\right)\prod_j f\left(D_{i,j}|T_i = 1,{\vartheta_j}^{(k)}\right) \\
&= f\left(T_i = 1\right)\prod_{j:D_{i,j}\geq 1-\delta} p_j^{(k)} \prod_{j:D_{i,j}<1-\delta}\left(1 - p_j^{(k)}\right)
\end{aligned} \tag{3.11}$$

$$
\begin{aligned}
b_i^{(k)} &\equiv f\left(T_i = 0\right) \prod_j f\left(D_{i,j} | T_i = 0, {\vartheta_j}^{(k)}\right) \\
&= f\left(T_i = 0\right) \prod_{j: D_{i,j} \leq \delta} q_j^{(k)} \prod_{j: D_{i,j} > \delta} \left(1 - q_j^{(k)}\right)
\end{aligned}
\tag{3.12}
$$

and

$$
\begin{aligned}
c_i^{(k)} &\equiv f\left(T_i = d\right) \prod_j f\left(D_{i,j} | T_i = d, {\vartheta_j}^{(k)}\right) \\
&= f\left(T_i = d\right) \prod_{j: \delta < D_{i,j} < 1-\delta} r_j^{(k)} \prod_{j: D_{i,j} \leq \delta \vee D_{i,j} \geq 1-\delta} \left(1 - r_j^{(k)}\right)
\end{aligned}
\tag{3.13}
$$

where the notation $j : D_{i,j} \geq 1 - \delta$ denotes the set of indices j (raters) that at voxel i gave a decision value $D_{i,j} \geq 1 - \delta$ (the presence of a lesion). With these expressions, we can define a compact description for the conditional probability of the true segmentation at each voxel:

$$
W_{i,1}^{(k)} \equiv f\left(T_i = 1 | \mathbf{D}_i, \vartheta^{(k)}\right) = \frac{a_i^{(k)}}{a_i^{(k)} + b_i^{(k)} + c_i^{(k)}}
\tag{3.14}
$$

for the class $s = 1$ and with an equivalent for $s = 0$ and $s = d$ (in those cases, on the numerator we substitute $a_i^{(k)}$ with $b_i^{(k)}$ and $c_i^{(k)}$, respectively). In general:

$$
\begin{aligned}
W_{i,\,s}^{(k)} &\equiv f\left(T_i = s | \mathbf{D}_i, \vartheta^{(k)}\right) \\
&= \frac{f\left(T_i = s\right) \prod_j f\left(D_{i,j} | T_i = s, {\vartheta_j}^{(k)}\right)}{\sum_{s'} f\left(T_i = s'\right) \prod_j f\left(D_{i,j} | T_i = s', {\vartheta_j}^{(k)}\right)}
\end{aligned}
\tag{3.15}
$$

The weight $W_{i,\,s}^{(k)}$ indicates the conditional probability that the true label at voxel i is s, given the set of segmentations and the estimate of the performance values. With these values, we can now calculate the values of the rater performance that maximize the conditional expectation of the complete data log likelihood:

$$
\begin{aligned}
\vartheta^{(k)} &= \arg\max_{\vartheta} E\left[\ln f\left(\mathbf{D}, \mathbf{T} | \vartheta\right) | \mathbf{D}, \vartheta^{(k-1)}\right] \\
&= \arg\max_{\vartheta} \sum_j \sum_i E\left[\ln f\left(D_{i,j} | T_i, \vartheta_j\right) | \mathbf{D}, \vartheta^{(k-1)}\right]
\end{aligned}
\tag{3.16}
$$

and for each rater j

$$
\begin{aligned}
\vartheta_j^{(k)} &= \arg\max_{\vartheta_j} \sum_i E\left[\ln f\left(D_{i,j} | T_i, \vartheta_j\right) | \mathbf{D}, \vartheta^{(k-1)}\right] \\
&= \arg\max_{\vartheta_j} \sum_i \sum_s \left[W_{i,\,s}^{(k)} \ln f\left(D_{i,j} | T_i = s, \vartheta_j\right)\right] \\
&= \arg\max_{\vartheta_j} \sum_{s'} \sum_{i:\ D_{i,j} \in s'} \sum_s \left[W_{i,\,s}^{(k)} \ln f\left(D_{i,j} \in s' | T_i = s, \vartheta_j\right)\right] \\
&= \arg\max_{\vartheta_j} \sum_{s'} \sum_{i:\ D_{i,j} \in s'} \sum_s W_{i,\,s}^{(k)} \ln \vartheta_{j,s's}.
\end{aligned}
\tag{3.17}
$$

The set of parameters that maximizes the above expression can be found by solving the following constrained optimization problem [rif. a STAPLE 2004]:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \vartheta_{j,n'n}} \left[\sum_{s'} \sum_{i:\ D_{i,j} \in s'} \sum_{s} W_{i,\ s}^{(k)} \ln \vartheta_{j,s's} + \lambda \sum_{s'} \vartheta_{j,s's} \right] \\ &= \sum_{i:\ D_{i,j} \in n'} W_{i,\ n}^{(k)} \frac{1}{\vartheta_{j,n'n}} + \lambda \end{aligned} \tag{3.18}$$

that is

$$\vartheta_{j,s's} = \frac{\sum_{i:\ D_{i,j} \in s'} W_{i,\ s}^{(k)}}{-\lambda} \tag{3.19}$$

and, by using the fact that $\sum_{s'} \vartheta_{j,s's} = 1$, we have

$$\vartheta_{j,s's} = \frac{\sum_{i:\ D_{i,j} \in s'} W_{i,\ s}^{(k)}}{\sum_{i} W_{i,\ s}^{(k)}} \tag{3.20}$$

As can be seen, the class s' is represented as a set (an interval). The previous equation also includes the case in which no doubt region is used (binary case).

It is worth noting that when the input images are absolutely polarized (binary maps are used), the proposed variation falls in the method proposed in [rif staple 2004]. We use the ternary consensus as a ground truth for our framework to obtain ternary classification from each of the CNN proposed.

3.2.2 "Safe" Ternary ground-truth

At the same time, to maintain the possibility of comparing different strategies on the same ground-truth, it is not recommended to completely redefine it [119, 120, 121], but just to consider as uncertainty what at least two of the seven human raters of MSSEG have considered as lesions, while the binary consensus has not. In this way, the original Lesion of the binary consensus is not altered but the space for the Uncertainty is gnawed from the Background.

This method is quite different from other strategies used to define the Uncertainty [119, 121] and it has the following motivations:

1. to maintain the original structure of the lesion ground-truth calculated in MSSEG with STAPLE and its derivations [109, 119, 120];

2. to account for the Uncertainty affecting both the problem and the raters;

3. to avoid the new class (Uncertainty) could change the original MSSEG consensus which could prevent a direct comparison with other methods;

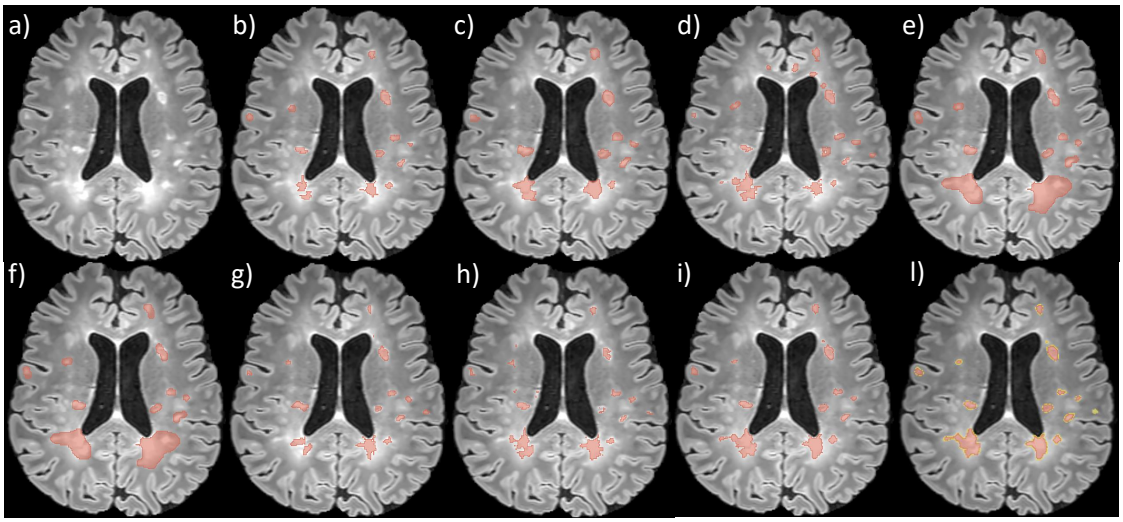

Figure 3.1: A sample FLAIR image from the 2016 MICCAI data set (a), the binary classifications from the 7 human raters (b-h), the binary consensus (i) and the ternary consensus (l). The Lesion is annotated in red. In the ternary consensus, the Uncertainty is indicated in yellow.

4. to quantify the gain the proposed framework could effectively get when the Uncertainty is introduced with respect to its absence;

5. to allow the learning strategy to consider as uncertain not only lesion borders, as other Authors do [119], but also whole regions not necessarily connected to lesions. In fact, the Uncertainty could regard both the lesion borders, where damaged tissues could coexist with healthy tissues (PVE), and whole structures, where doubts are due to MRI unspecific nature for MS.

Fig.3.1 reports a FLAIR image example with the seven human binary classifications, the binary consensus and the proposed ternary consensus.

The Uncertainty, in yellow in the ternary consensus (Figure 3.1 l), indicates doubtful regions where discordant decisions are assumed by raters but on which at least two raters agree.

3.3 Evaluation Criteria

As far as we need an exhaustive comparison between all the raters involved therein (artificial, single humans and ground truth), and being a unique performance parameter unavailable, we define and calculate all the mostly known metrics. In what follows, we define all the used metrics by separating those oscillating in the interval $[0, 1]$, whose ideal value is 1, from those oscillating in the interval $[0, \infty)$, whose

best value is 0. The two groups are distinguished for graphical purposes. For more details about the reported metrics, please refer to [122, 123, 124, 2, 125].

3.3.1 Scores

Sensitivity (also called recall or true positive rate) is defined as:

$$SENS = \frac{TP}{TP + FN} \tag{3.21}$$

$SENS$ measures the portion of positive voxels that are correctly identified, that is the capability of a method to correctly classify the voxels, without underestimation. In fact, sensitivity ranges between 0 ($TP = 0$) and 1 (when $FN = 0$). We also distinguish an object sensitivity, $OSENS$, defined as:

$$OSENS = \frac{TP_o}{TP_o + FN_o} \tag{3.22}$$

in which the prefix O and the subscript o indicate we are referring to whole objects and not to single voxels. An object is considered as TP if the intersection with the corresponding object in the ground-truth is not empty.

Specificity ($SPEC$) is defined as:

$$SPEC = \frac{TN}{TN + FP} \tag{3.23}$$

$SPEC$ represents the portion of negative voxels N that have been correctly identified. For the treated case, since classes are strongly unbalanced, $SPEC$ is biased by the fact that most of the image surface is covered by background: for this reason the high specificity does not guarantee a good performance (we have reported it for completeness).

Accuracy (ACC) is defined as:

$$ACC = \frac{TP + TN}{P + N} \tag{3.24}$$

but, due to unbalancing, we use the following normalized ($ACCN$) definition:

$$ACCN = \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) /2 \tag{3.25}$$

to make it more representative.

Positive Predicted Value (PPV), also called Precision, is defined as:

$$PPV = \frac{TP}{TP + FP} \tag{3.26}$$

PPV represents the portion of voxels identified as positives which are really positives (TP). PPV measures how the method correctly classifies voxels in the

correct class without overestimating the class itself. In fact, PPV ranges between 0 (TP=0) and 1 (FP=0).

As for $OSENS$, we have defined an object-based PPV, $OPPV$, as follows:

$$OPPV = \frac{TP_o}{TP_o + FP_o} \tag{3.27}$$

in which the prefix O and the subscript o indicate we are referring to whole objects, as above. $OPPV$ represents the portion of objects identified as positives which are really positives (TP_o). $OPPV$ has the same meaning of PPV but for whole objects, not for single voxels.

Dice score, also called Sorensen–Dice coefficient, is defined as:

$$Dice = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.28}$$

Dice score measures the similarity between two data sets. This index is widely used in AI for the validation of image segmentation algorithms. We refer to Dice score as Global Dice score to distinguish it from Image Dice score.

Image Dice score uses the same equation of Global Dice score but, while Global Dice score is calculated on the whole data set, Image Dice score is applied on each single image and finally averaged on the number of images. Image Dice score allows to the so called per-image metrics [123]. Per-image metrics are important because they tend to highlight the local behaviour.

A score similar to *Dice* is the Intersection Over Union (IoU):

$$IoU = \frac{TP}{TP + FP + FN} \tag{3.29}$$

where the difference is in the weight of TP.

The $F1$ Score (calculated for whole objects and not for single voxels) is defined as:

$$F1 = 2 * \frac{OSENS * OPPV}{OSENS + OPPV} \tag{3.30}$$

where $OSENS$ and $OPPV$ are defined above.

BF score is a per-image version of $F1$ score.

Pearson Correlation Coefficient (PCC), between two data sets A and B, is defined as:

$$PCC(A, B) = \frac{cov(A, B)}{\sigma_A * \sigma_B} \tag{3.31}$$

where $cov(A, B)$ is the covariance of A and B and σ_A and σ_B are the standard deviation of A and B, respectively. PCC ranges in the interval $[-1, 1]$ and a negative value of PCC indicates a similarity of the object A with the negative version of the object B.

3.3.2 Metrics

The following metrics are those used in the present manuscript in which the ideal value is 0.

Extra Fraction (EF), is defined as:

$$EF = \frac{FP}{TP + FN} \tag{3.32}$$

Detection error rate (DER) is defined as:

$$DER = \frac{DE}{MTA} \tag{3.33}$$

where DE is the detection error calculated as the sum of the voxels of a connected region marked as positive by the rater and the mean total area (MTA) is defined as the average between the number of positive voxels from the rater and the ground-truth. DER measures the disagreement in detecting the same regions between the rater under evaluation and the ground-truth.

Outline Error Rate (OER) is defined as

$$OER = \frac{OE}{MTA} \tag{3.34}$$

where OE is the outline error calculated as the difference between the number of voxels of the union and that of the intersection between the positive connected regions of both the rater and the ground-truth. OER measures the disagreement in outlining the same object between the rater under evaluation and the ground truth.

False Detection Ratio (FDE) is defined as:

$$FDE = \frac{FP}{P} \tag{3.35}$$

Relative Area Error (RAE) is defined as:

$$RAE = \frac{TP + FP - P}{P} \tag{3.36}$$

Hausdorff Distance (HD) between two objects A and B is defined as:

$$HD(A, B) = max(h(A, B), h(B, A)) \tag{3.37}$$

where $h(A, B)$ is:

$$h(A, B) = max_{a \in A} min_{b \in B} \parallel a - b \parallel \tag{3.38}$$

HD measures how far two subsets are from each other. In other words, two sets are close with respect to HD if every point of one set is close to a certain point of the other set.

Euclidean Distance (ED) between two objects A and B is defined as:

$$ED(A, B) = max(d(A, B), d(B, A)) \tag{3.39}$$

where $d(A, B)$ is defined as:

$$d(A, B) = \frac{1}{N} \sum_{a \in A} min_{b \in B} \parallel a - b \parallel \tag{3.40}$$

Surface Distance (SD) is defined as:

$$SD = \frac{\sum_{i \in A_S} d(x_i, G_S) + \sum_{j \in G_S} d(x_j, A_S)}{N_A + N_G} \tag{3.41}$$

where A_S and G_S are two segmentations (one is the rater segmentation and the other is the ground truth), d denotes the minimal ED between voxels on both surfaces, while N_A and N_G denote the number of points of each surface.

Chapter 4
Automatic Multiple Sclerosis Lesion Segmentation from MRI: Guidelines for Effective Outcomes.

The richness of parameters of MRI makes it possible, for radiologists, to identify some characteristic signs of tissue damages, such as white matter lesions and brain atrophy or shrinkage. Besides the undoubted advantages, the MRI variability MRI makes the design of an efficient AI model a real challenge because images and, hence, their corresponding features, continuously change with imaging parameters. Moreover, scanners from different manufacturers produce images with slightly different contrast. A trained human eye could adapt to MRI variability with difficulty: the task is greatly challenging for AI models.

In this chapter, the general constraints for automatic identification/segmentation related to the MS lesions by MRI are discussed and guidelines are presented for the effective training of a AI model. A convolutional neural network (CNN) based model is trained and preliminary results, demonstrating the improvements, are reported.

The content of this chapter appeared in [126].

4.1 Introduction

MS origins are not well understood but characteristic signs of tissue damages are recognizable, such as white matter lesions and brain atrophy or shrinkage due to degeneration. These signs can be observed by MRI which is a special tool to follow-up MS patients with reduced invasiveness due to the usage of specific contrast agents. In fact, focal lesions in the brain and spinal cord are primarily visible in the white matter on structural MRI observable as hyperintensities [111].

These imaging procedures are all performed in a single MRI examination and the corresponding slices (hundreds) are all used for MS monitoring and follow-up (also comparisons with previous examinations are necessary). Identification of the lesions affecting the white matter and their count and volume calculation by MRI have become well established protocols for assessing the disease progression and pharmacological efficacy. For this reason, MRI is currently used routinely in clinical

practice: imaging markers are capable to capture volumetric changes but need to be assisted by an expert, either human or automatic. However, the richness of MRI parameters/imaging modalities if, by one side, constitutes an advantage for gathering fundamental information about MS lesions, by the other it makes the design of efficient automatic experts a real challenge because images and, hence, the corresponding features, change with magnetic field strength, imaging parameters, sequences and scanners from different manufacturers (Siemens, Philips, GE, etc.). To these modifications, a trained human eye suddenly adapts but an automatic expert has to be deeply trained before its adaptation. But, is this really necessary?

In what follows we describe some guidelines for automatic segmentation of MS lesions identification/segmentation by MRI and discuss how to allow an automatic system to perform at best. Moreover, we present a strategy to improve lesion identification and segmentation. To the best of our knowledge, the proposal of preliminary conditions for correct MS lesion identification/segmentation by MRI is new and necessary to obtain better performance from automatic methods.

4.2 Related Work

Several attempts have been proposed for automatic segmentation of MS lesions by MRI, though the variability of MS lesions in size, shape, intensity and localization make automatic and accurate identification and segmentation really challenging [127, 128, 129]. Even if, classical techniques such as for example based on shapes [130, 131] could be effective, deep neural networks seem to be more promising also because require low manual intervention with respect to other approaches. In fact, the great advantage of deep learning is that the relevant feature set is no longer defined by the user but learned directly by the system from the training images. This is a crucial aspect because it is not trivial for people to characterize features that best serve to separate healthy tissue from MS lesions. From the perspective of deep learning application, the high dimensionality of the MR images, the difficulty of obtaining reliable ground truth and the high accuracy required for clinical practice, all contribute to make MS lesion identification/segmentation a worthy test application. CNN have demonstrated breaking performance also in brain imaging segmentation [132, 133, 134]. In particular in [132] is presented one of the first attempt of an automated learning approach for MS lesion segmentation. Besides the architecture of the used system, the method analyzes 3D patches of the MRI volume instead of the entire volume or single slices. In 2015, [133] proposed a method that used 3D CNNs to learn features by different datasets of the same patient: T1-w, T2-w, PD and FLAIR MRIs. The method proposed in [134] has proven to use efficiently the information carried on by different MRI imaging modalities by reducing the number of

parameters (and hence the training set) through the usage of two CNNs in cascade, trained separately. To date, the method presented in [134] represents for MS lesion segmentation one of the benchmark architectures. In fact, in the comparative study of algorithms for MS lesion segmentation for MICCAI 2016 international challenge, presented in [129], demonstrated that the method in [134] was established as one of the most effective for MS lesion segmentation, though the best method was that obtained by creating a consensus between the results of all the compared methods.

However, though advanced computer vision techniques have been compared in [129], the results were not brilliant with respect to other field of applications. For this reason, In what follows we discuss the reasons of poor results and suggest guidelines to allow better efficacy for automatic strategies.

4.3 General considerations and guidelines definition

Though MRI is considered a gold standard, the correct interpretation of MS lesions through MRI is a not trivial operation and it is a subject of debate [111] due to the fact that MS lesions can be easily misdiagnosed or erroneously interpreted (confused with other diseases and/or artifacts and/or tissue modifications with age) also by expert, trained radiologists and guidelines for radiologists are continuously updated to overcome misdiagnosis [111, 135]. Moreover, in [111] it is also affirmed that misdiagnosis also depends on the used MRI scanner. As a consequence, expert radiologists often disagree when performing independent diagnosis of the same data, both due to the ambiguity between MS lesions and other diseases and because they could have gathered their experience on different scanners.

Besides the considerations in [129], some important aspects have to be underlined [136]:

1. MS lesion identification/segmentation depends mostly on imaging scanners due to differences in imaging parameters, temporization, features, magnetic field values and homogeneity, etc.. These differences could bad influence automatic methods more than on human experts. In fact, humans use also other implicit information, such as clinical or anatomical concepts, to evaluate the image content. a huge increment of data for training should be necessary to include differences between scanners into an automatic system;

2. MS lesion identification/segmentation depends on the used data pre-processing strategy which should be part of the method itself: the indistinct free usage of data (preprocessed or unpreprocessed) could greatly affect the convergence of the method and the training dataset dimension;

3. An MS lesion identification/segmentation strategy depends on the imaging modalities it uses (FLAIR and T2-w images are more informative than PD or T1-w [111]: the indistinct usage of all the modalities to train an automatic strategy probably results in a decrement of convergence speed and has to imply an increment of the dataset used for training.

The previous considerations found their confirmation in the contrasting results reported in [129]: the methods performance decreased when used on data from a previously unseen scanner; methods which used preprocessed data were not all better than those using unpreprocessed data; methods using all the imaging modalities were not always better than those using just some imaging modalities.

To better explain these apparently strange behaviours, please consider data presented in Figures 4.1 and 4.2, where some images, from the MICCAI2016 dataset, collected by different scanners are reported for all the imaging modalities, both before (Figure 4.1) and after preprocessing (Figure 4.2). For the same images, an horizontal line of data (red line) is also plotted below (Figure 1b and Figure 2b). As can be noted, unpreprocessed data show relevant differences between scanners (though data allowed to different patients, it is clearly visible the ratio between the amplitude of different tissues in the same image are different for the two scanners, as it is also confirmed by comparing the image corresponding to the same imaging sequences): these differences, which distinguish MRI from CT (where images from different scanners are scalable in amplitude and easily compared), are due to different imaging parameters optimization by different manufacturers, though using the same imaging sequences.

In Figure 4.2, the situation after preprocessing, an amplitude normalization between different images has occurred. In fact, the images of different scanners are more similar than those before preprocessing. However, from Figure 2b it can be observed that the preprocessing step produced a variation on the baseline of some of the images (the signal outside the brain, which should be zero, has a level well above zero). Moreover, each image was normalized independently from the other: this implied a modification which has been different from one image to the other, thus introducing substantial differences also on data from the same scanner. Finally, the amplitude ratio between different tissues in the same image has not been rightly corrected and, in some cases, differences between data coming from different scanners were increased. This is probably the reason why some automatic strategies, though using preproessed data, performed worse than those using original, unpreprocessed, data. Finally, from both Figure 4.1 and 4.2, it can be noted that the information carried on by different imaging modalities regarding MS lesions is completely different: iperintense regions on FLAIR images which are also iperintense on

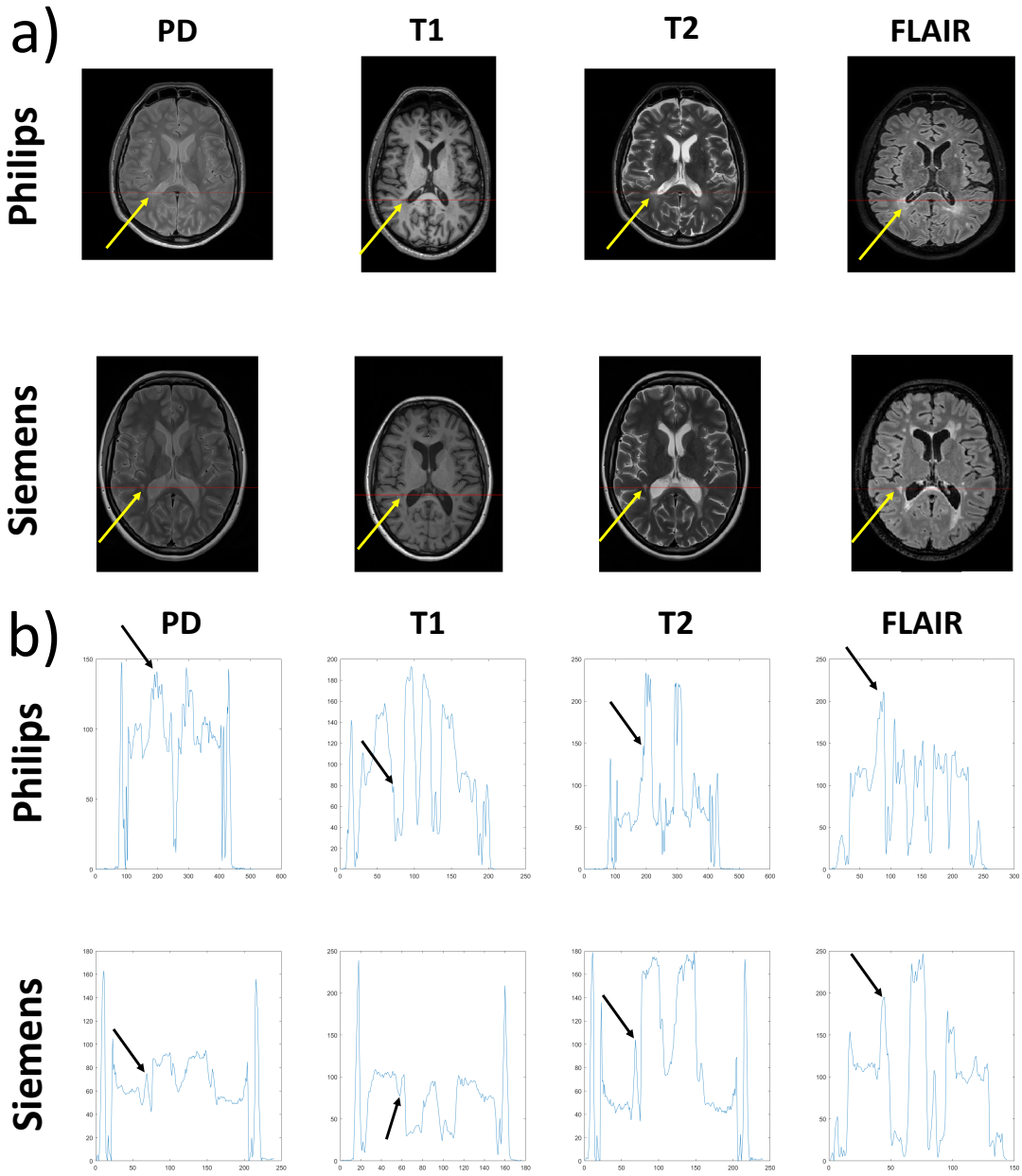

Figure 4.1: Raw, unprocessed, data from different scanners (rows) and from different imaging modalities (columns). Images are reported in (a) and plots of a single row of the images (along the red line) are shown in (b). The position of a lesion along the red line is indicated by an arrow. The shrinkage of the FLAIR image from Siemens scanner is due to a different (greater) dimension of the voxel in the horizontal direction.

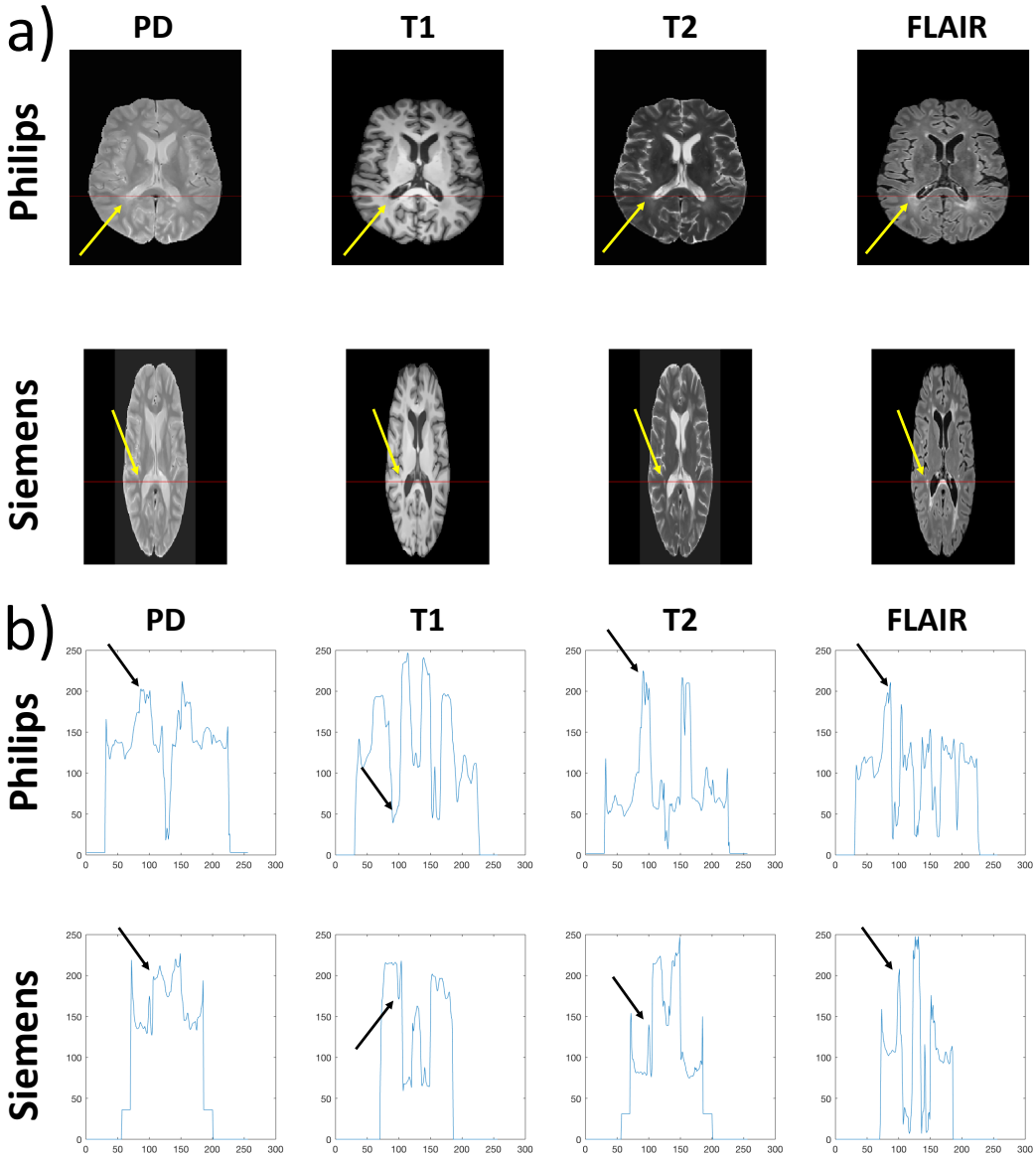

Figure 4.2: Data of Figure 1 after preprocessing. Images are reported in (a) and plots of a single row of the images (along the red line) are shown in (b). The position of a lesion along the red line is indicated by an arrow. Images have been also reshaped after their co-registration.

the corresponding T2-w images surely indicate MS lesions [111]. The other imaging modalities (T1-w and PD) do not add anything more and, often, their content is confusing and not clearly interpretable (as in the MS lesions indicated by the green arrows, both in Figure 4.1 and Figure 4.2).

Form the above considerations, the following guidelines could be derived:

1. The training of the method should be done on data from a single scanner (also humans adapt to the scanner they normally use): when data from different scanners need to be interpreted and, may be, compared, the system has to be trained separately to each scanner (in this way, the training set can be reduced, the procedure shortened and the performance increased);

2. A preprocessing strategy, consisting in the rigid registration of each modality on the FLAIR image, is necessary to obtain images of different modalities which are spatially correspondent. Other forms of preprocessing, especially those consisting in amplitude corrections, have to be performed on the whole volume and not differently on each single slice. Moreover, preprocessing has to become part of the automatic segmentation method;

3. The image modalities to be used in the identification/segmentation process have to be chosen in advance to avoid useless/confusing information, unjustified increment of the training dataset, convergence deceleration and performance reduction (FLAIR and T2-w images are sufficient).

In what follows, we show how, by applying the previously defined guidelines, it is possible to improve the performance of a lesion segmentation method.

4.4 MS lesion identification/segmentation

Being a benchmark method, we have used the supervised CNN-based paradigm presented in [134] that has also been used, in a modified version, in [137]. In particular, by following the previously defined guidelines, we operated the following choices:

1. the dataset used for training, validation and test was the MICCAI2016 dataset but just using data from a single 3T scanner (Philips manufacturer);

2. raw, unpreprocessed, data were preprocessed by performing rigid registration of each modality on the FLAIR image followed by brain extraction (skull stripping) from T1-w image and applied to other modalities;

3. only FLAIR and T2-w imaging modalities were used for identification/segmentation. In this way, we provided a simpler task to the system, thus reducing

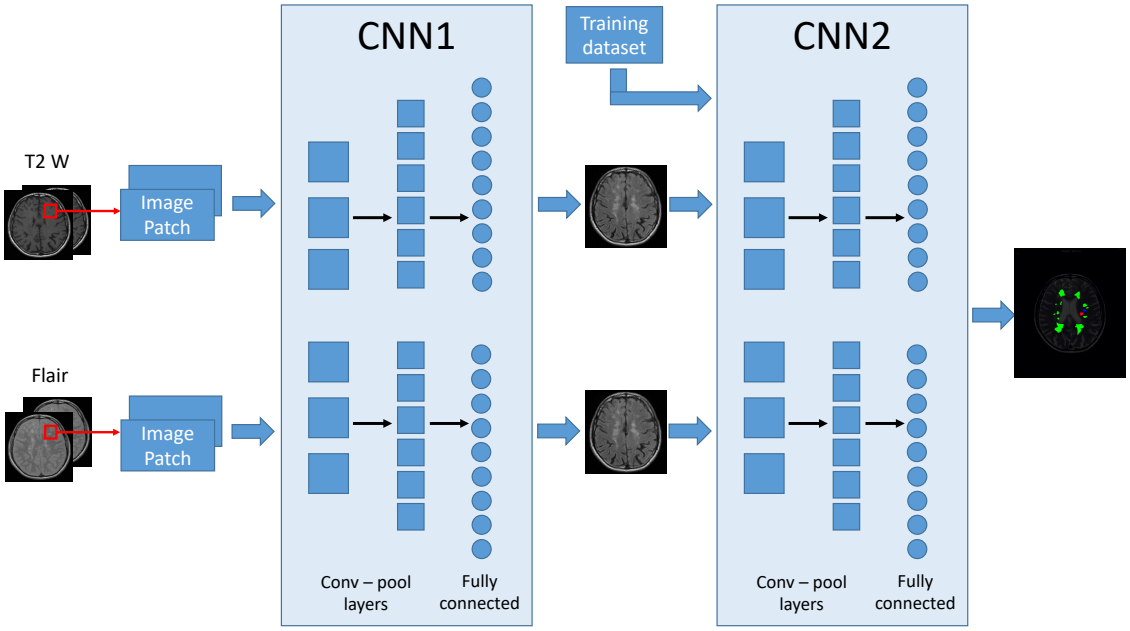

Figure 4.3: Two stage CNNs architecture used for identification/segmentation of MS lesions. Input of the system are the registered volumes by FLAIR and T2-w images. Training of CNN2 is made with a separated dataset.

the dimension of the training, labelled, dataset. The images selected from the dataset were distributed in three subsets: 800 for training, 200 for validation and 100 for test. A scheme of the assembly used for MS lesion identification/segmentation is reported in Figure 4.3.

The method is based on a cascade of two CNNs. The low variation in contrast of MRI images, the use of images from just one scanner and the reduction of imaging modalities, allow simple network architectures and a reduction of the training set dimension. The system consists of a 7-layers architecture for each of the two CNNs. Each network is composed by two stacks of convolution and max-pooling layers with 32 and 64 filters, respectively. Convolutional layers are followed by a fully-connected layer of size 256 and a soft-max fully connected layer of size 2 whose output is the probability of each voxel to belong to a lesion. For a complete settlement of the used parameters, please refer to [134]. MS lesions are calculated using 3D neighboring patch features. The used 3D patches are cubic, 11x11x11 voxels. The splitting in two different CNNs allows to separate the training procedure in two and this allows a reduction of the number of parameters without reducing accuracy. To reorder data balance for training, that is to equilibrate the number of "positive" patches (containing lesions) with "negative" patches (containing no lesions, much greater than the other), the dataset used for training consists of the whole dataset of positive patches and of an equal number of randomly selected negative, healthy patches. In this way,

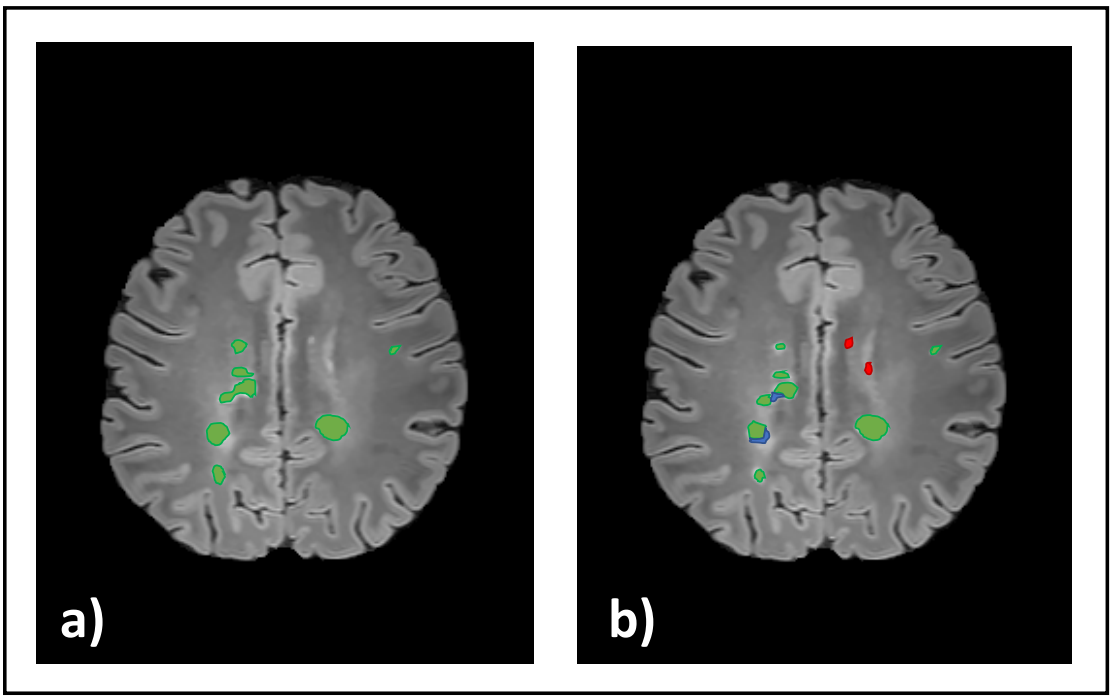

Figure 4.4: MS lesion identification/segmentation on one of the images (FLAIR) by MICCAI2016 used for test. In (a), the ground-truth identification/segmentation is reported in green; in (b), the same image is reported with indicated, in colors, the voxels identified/segmented by the method: the voxels rightly identified/segmented are indicated in green; in red are those wrongly identified as lesions (false positive); in blue those are those wrongly recognized as healthy tissue (false negative).

the first network (CNN1) is trained by using the resulting balanced dataset and then tested on the whole dataset, thus obtaining a list of probabilities for each voxel of each patch to be "positive" (part of a lesion). After that, a balanced dataset is created by using the previous test results and by considering as positive all patches containing voxels whose probability is greater than 0.5. As for the previous balanced training dataset, negative patches (those in which all voxels had probability <0.5), are randomly selected to be the same number of "positive" patches. The second network (CNN2) is trained from scratch with the dataset resulting from CNN1. Once the whole pipeline is trained, new unseen MRI volumes can be processed using the same, two stage, architecture. The dataset is first decomposed in patches and, then, all volume patches are evaluated using CNN1. CNN1 discards all voxels with low probability (< 0.5). The rest of the voxels, included into corresponding patches, are re-evaluated by CNN2 to obtain the final probabilistic lesion mask. Resulting binary masks (ones where lesion are present, zeros elsewhere) are computed by thresholding the probability lesion masks (prob > 0.5 are considered lesions). Finally, an additional false positive reduction is performed by discarding binary connected regions with very low number of positive voxels (this number is calculated with respect to the minimal volume of the lesions used for testing). The method had an average F1 score of 0.68 and an average Dice score of 0.71 (about 25% better than the original method [134] and 15% better than the modified method in [137] without using any artificial strategy for increasing the training dataset of patches. The improvement with respect to [137], relevant if we consider that it has been obtained with half of the imaging modalities, is mainly due to the fact that it has been obtained by training the method on data from a single scanner and just from the most significant imaging modalities, which simplifies the identification/segmentation process. Moreover, these results are significant because they allow to overcome the score of the automatic "Team fusion" and also of the worst human expert [129], thus making automatic identification/segmentation acceptable for MS diagnosis/analysis. In order to show the results on the images, Figure 4.4 reports the worst-case automatic identification/segmentation: the method allows a discrete identification of the lesions (false positives are in red) and a good segmentation (false negatives are in blue).

4.5 Discussion

We have discussed some limitations that occur when using automatic identification/segmentation of MS lesions by MRI data: the richness of imaging parameters and internal variability of MRI scanners make the problem ambiguous and difficult. By considering these limitations we have extracted a set of basic guidelines that

the training dataset should have in order to avoid confusion when training a supervised automatic identification/segmentation strategy. Finally, we have applied these guidelines and used them while performed training of a CNN-based strategy used as a benchmark. The results are better than those obtained without using the constraints on the training dataset, thus making the automatic method similar, in performance, to a human expert. Moreover, we have obtained a faster convergence of the method with respect to use it with data from multiple scanners and/or when using data from indistinct imaging modalities.

Chapter 5
MRI Stabilization through Local Contrast Normalization

As discussed in the previous chapter, MRI images can have different contrast though collected with the same imaging sequence. Furthermore, images contrast varies also due to the internal properties of the scanned body. This makes MRI very different from other imaging techniques, such as Computed Tomography (CT) where images can be easily normalized for different tissues, both in healthy and pathological conditions, by using Hounsfield's units. Since MR images do not have standardized amplitudes, an AI model trained by using images from one scanner could completely fail to analyse images made with another scanner. For this reason, MRI images should be pre-processed in order to normalize them and thus reduce their contrast variability. This chapter presents a local contrast normalization algorithm for a specific MRI imaging sequence, the Flair, because this sequence is used to study inflammatory processes of the brain. The application of the proposed strategy on the images from different MRI scanners are reported and compared. Results are reported and discussed.

The content of this chapter appeared in [138].

5.1 Introduction

The richness of parameters makes MRI so special but, at the same time, it results in a huge variability in image contrast, also collected with the same instrument and with the same imaging sequence, due to the fact that variability is also implicit on the imaged sample, on how it interacts with the MRI system and its chemical intrinsic properties. Contrast variability is greatly enhanced in images from scanners of different manufacturers, magnetic field strength, electromagnetic field homogeneity, etc. [139]. This make MRI very different from other imaging techniques, such as CT where images, depending just on one parameter (X-ray attenuation), can be easily normalized for different tissues and organs, both in healthy and pathological conditions, in Hounsfield units. Contrast instability in MRI makes identifica-

tion/segmentation task very difficult also for expert radiologists but, in particular, for automated strategies. Indeed, automated strategies which are trained to cope very well with data from one scanner, could completely fail with data from another. To make automated segmentation strategies robust to contrast variability, a lot of images from different scanners have to be used for training thus making this process very long, and the segmentation results are suboptimal. For this reason, images are preprocessed before they are passed to automatic segmentation. Though several strategies have been proposed to standardize MRI similarly to CT [140, 141, 142], results are not enough accurate because a general MRI normalization, feasible both for each imaging sequence and for each anatomical district, was attempted while preprocessing, due to huge MRI variability, has to be specific for each imaging sequence and for each imaged anatomical region. In what follows we present a local contrast normalization strategy for MRI of the brain, related to a specific MRI imaging sequence, the FLAIR, one of the imaging sequences used for inflammatory diseases.

5.2 MRI Preprocessing

Preprocessing refers to a series of mathematical adjustments to MR images before segmentation [105] for reducing the effects of noise and imaging artifacts, equalizing space, eliminating outliers and stabilizing contrast. Though it is well known [124] how important is to match image contrast before segmentation, due to variable sequences, overlapping intensities, noise, field inhomogeneity, partial volume, gradients, motion, echoes, blurred edges, anatomical variations and susceptibility artifacts [106][107]. Some of this variability can be reduced with specific hardware [143][144][145][146][147] but most has to be correctly with appropriate software [148][149][150][151][106][152][107]. For this reason, MRI has to undergo preprocessing to stabilize and make effective segmentation.

Preprocessing for MRI (Figure 5.1) consists of: registration and alignment of images, noise reduction, skull stripping, bias field correction and contrast normalization.

Step 6 in Figure 5.1 represents the proposed technique for local contrast normalization.

Rigid registration and alignment are necessary because images obtained by different imaging modalities are not registered and might have different spacing and thickness. Robust methods are used for this scope [153, 154] to obtain images in axial orientation, the orientation used therein, with the same resolution along the three spatial axes and we used MRITOTAL [155] whose source code is available at https://github.com/bic-mni.

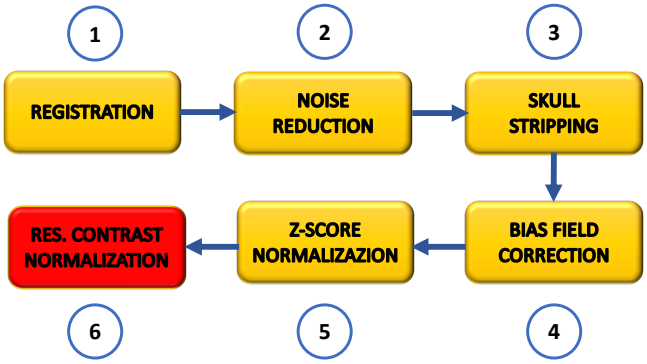

Figure 5.1: MRI preprocessing pipeline: the first 5 steps are usually applied to MRI; the last step (n.6) is the contrast normalization that we improve with a local contrast matching to reduce residual contrast mismatch.

Experimental noise affects MRI [156, 107], whose power is inversely proportional to the magnetic field strength: noise reduction would serve to make images more robust to MRI equipment at different magnetic field strength and to reduce segmentation outliers. Anisotropic diffusion noise reduction filtering to preserve edges is routinely applied in MRI and we merged the strategies in [157, 151, 158] to keep advantage from each one. Noise reduction is the first step to prepare data at best for the following steps.

Skull stripping is another important preprocessing step since fat, skull, skin and other non-brain tissues may cause misclassifications and problems to amplitude normalization. Skull stripping is performed via the FMRIB Software Library (FSL) [159].

Bias field correction is necessary to reduce the non-uniform intensity effects in MRI due to magnetic fields inhomogeneity or applied radio-frequency fields within the scanner: it is very important to reduce them to make segmentation robust sample positioning inside the scanner and to homogenize amplitude for the same grain tissue. In-homogeneity effects are corrected therein by using the N4 algorithm [160].

Finally, MR images undergo contrast normalization and all the intensity levels of the various scans are rearranged and normalized in the same interval through the z-score [142], that is by subtracting the mean value of the image to the image itself and by dividing the result for its standard deviation.

Due to the extreme variability in parameters settings, MR images collected with the same sequence could have variable contrast in equipment from different manufacturers or when the same system is used for different patients (reciprocal amplitudes

are not standardized in MRI) and residual contrast difference often remain after z-score. Residual amplitude variations between different brain tissues could negatively influence the following image analysis. To this aim, Figure 5.2 shows a comparison between FLAIR images collected by different systems on different patients before (A) and after their preprocessing with z-score (B). The other preprocessing steps have yet been applied to all the proposed images. In particular, the first column shows the images and the second column shows the corresponding histograms. As can be noticed, contrast differences are high in original images. Moreover, though attenuated, these differences remain after z-score calculation, as confirmed by histograms. In fact, z-score eliminates scaling between images but internal reciprocal contrast differences between soft brain tissues are almost unaltered. This is explainable because z-score normalization acts on the whole image by displacing and stretching (or enlarging) its histogram, but it leaves unchanged reciprocal amplitude displacement between different soft brain tissues, as can be noticed from the reciprocal peaks positions on the histograms both before and after z-score. However, a whole image histogram equalization would not be feasible since it would tend to eliminate the contrast gained by the imaging sequence.

5.3 The Proposed Strategy

We introduce a specific reinforcement for the standard preprocessing (the first 5 steps in Figure 5.1), summarized by Step 6 of Figure 5.1: a local histogram matching strategy for FLAIR images which differentiates between white matter (WM), grey matter (GM) and cerebrospinal fluid (CSF). Images (sub-images) corresponding to WM, GM and CSF segmentation, respectively, are treated separately (due to MRI variability, in one the amplitude shift could be in one direction and in another could follow the opposite direction) and then recombined to obtain the whole contrast normalized image.

The technique we propose aims at comparing the grey levels of WM, GM and CSF of an axial FLAIR image of the current patient examination, central in the brain where all the three classes are well recognizable, with the corresponding classes of a reference central image r, whose values, once calculated, are fixed, stored in memory and used as a ground truth, being the reference image selected between those of an examination collected with a MRI system used for reference (it is important to note that the choice of the reference scanner is irrelevant). To calculate the reference values, first the z-score is applied to r, then the soft brain tissues are segmented and separated in three complementary classes, corresponding to three images (WM, GM, and CSF), and the peak positions in the histograms are calculated of the three resulting sub-images and stored in memory. The Flow-chart of the method is

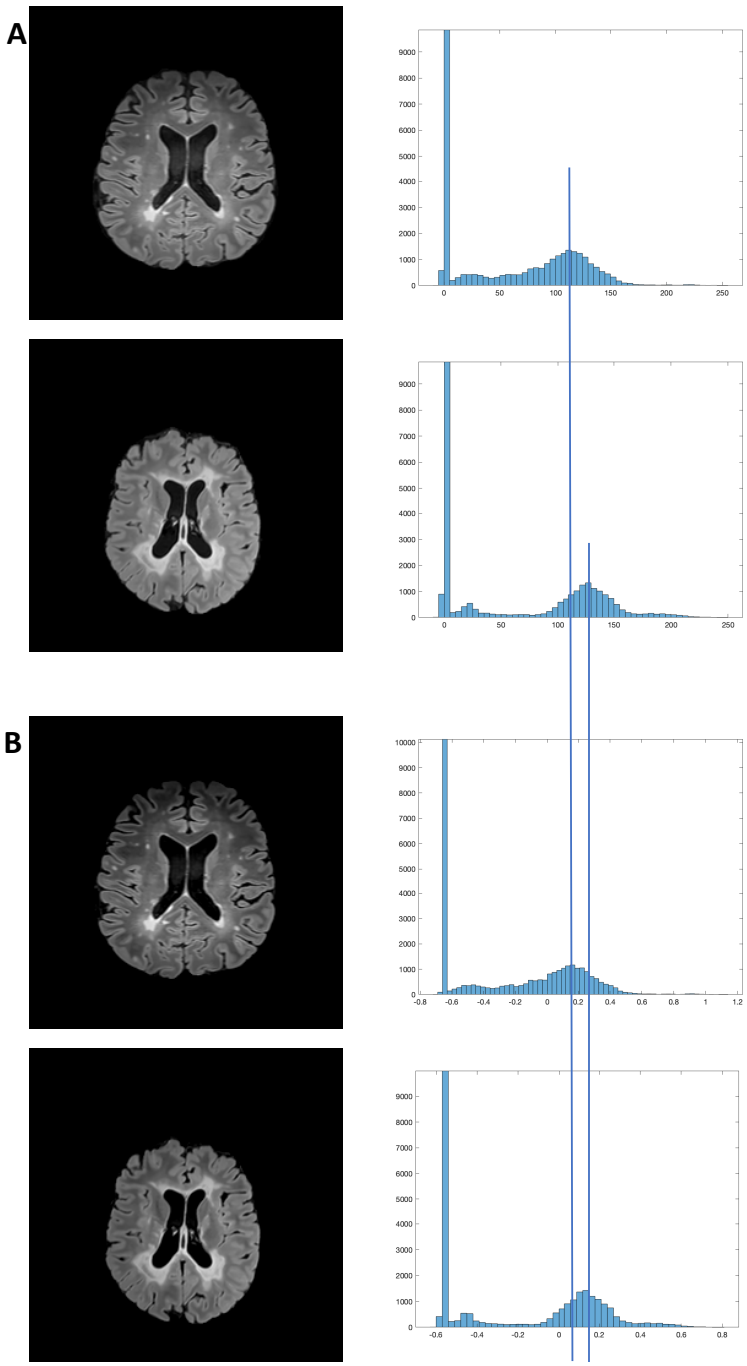

Figure 5.2: (A) Brain FLAIR images from different MRI scanners (Philips 3T and Siemens 3T, respectively) and related histograms. (B) z-score results of the images in (A) and related histograms. Relevant intensity mismatch remains after z-score. In the histograms, vertical axis was cut at 10.000 to better highlight lower details.

reported in Figure 5.3, where the operations performed on r are not shown because it undergoes steps 1 and 2 of the same Flow-chart. The method collects the central image k of the FLAIR data set to be treated and does the same as in the reference image r (steps 1 and 2). Then, for each class $H \in \{WM, GM, CSF\}$, the shift of the maximum in k with respect to r, DH, and the gray-scale band, BH, are calculated (steps 3 and 4). The band BH is necessary to collect the gray scale range allowing to each class H without repeating classification. DH and BH are calculated just once, for K and then applied to all the image of the current examination. In fact, for each image i, the sub-image allowing to each histogram range BH is realigned in amplitude for compensating the shift (step 5). Finally (step 6), the realigned sub-images are summed together to recreate the final image. Note that the segmentation is necessary just for image k and not for the other images it because the selection of the three classes is made by choosing the respective ranges of amplitudes, BH, which are complementary each other. Some points are important: 1) We need to use segmentation to separate different brain tissues from the histogram of the whole image both to find the position of local maximum amplitude and to calculate the range of amplitudes allowing to each class. The segmentation strategy we used therein is that proposed in [161] which has been proven to be one of the best. 2) For each image of the current examination, we divide the three sub images by using BH and apply DH to each of them (the correction could be different for each range) and then we recombine the three corrected images into the whole image (images in which one of the three classes is absent would not be corrected, having it no pixels for the corresponding amplitude range). In this way, all the images of the current examination undergo to the same process. 3) The calculation of the maximum in the histograms is performed on a filtered version of the histogram plot (3 points CAR filter [162] is used) to force stability and reducing unjustified over or under corrections. 4) The ground truth image and its respective histogram ranges and amplitude peaks for the three classes are selected from a central image of an examination collected with the reference scanner and saved. 5) Data from all patients, including those collected with the reference scanner, are processed by the proposed histogram correction algorithm to stabilize amplitude also coming from the reference equipment: this could help to reduce patient and scanner dependencies. The fact that contrast normalization is applied for last is to avoid that noise, bias field, skull presence, differences in dynamic range and other disturbing effects could negatively influence normalization.

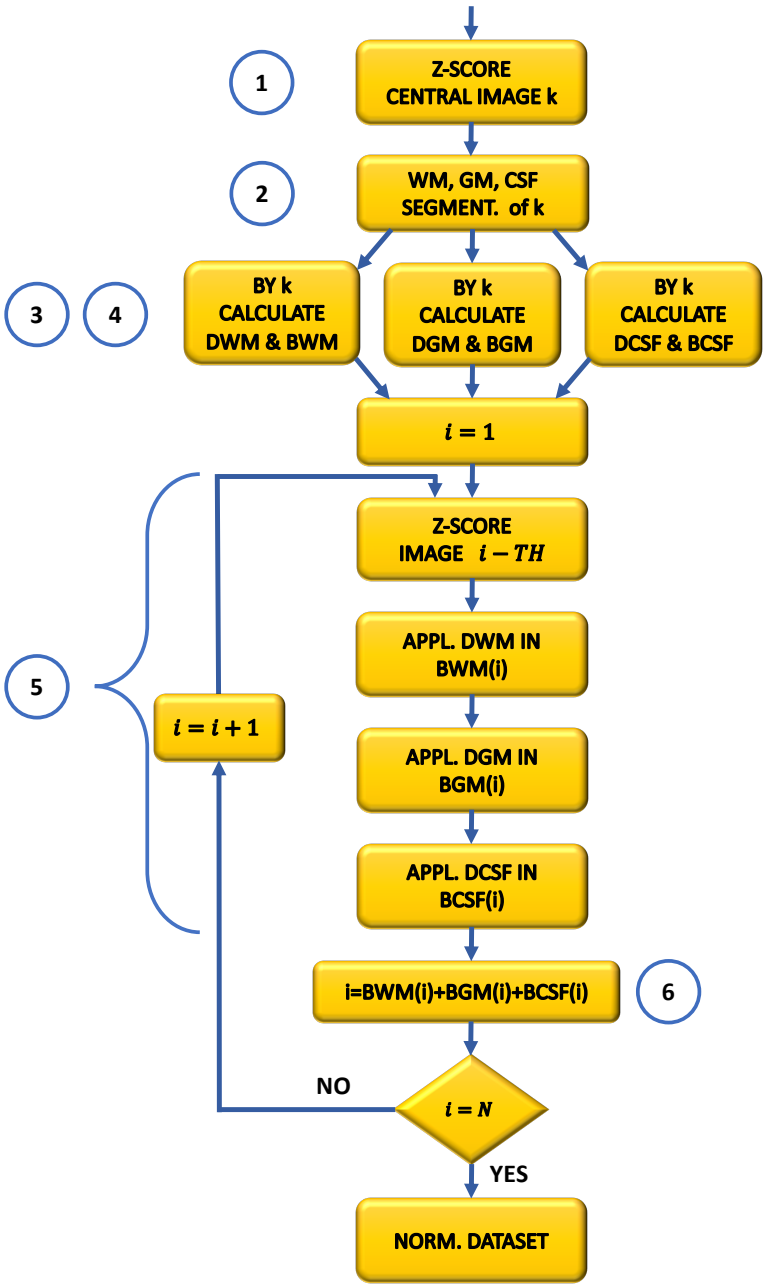

Figure 5.3: Amplitude realignment procedure. DH, for $H \in \{WM, GM, CSF\}$, are the amplitude displacements calculated on the sub-images WM, GM and CSF of the image k, respectively. BH, for $H \in \{WM, GM, CSF\}$, are the histogram band sub-images WM, GM and CSF of the image k, respectively. BWM(i), BGM(i) and BCSF(i) are the histogram ranges of sub-images WM, GM and CSF of the image i, respectively.

5.4 Experimental Evaluation

To test the effectiveness of the proposed strategy in the context of preprocessing pipeline, we applied both the traditional MRI preprocessing pipeline (steps 1-5 in Figure 5.1) and that integrated with the local contrast normalization (steps 1-6 in Figure 5.1) on 1500 FLAIR brain images from a public dataset of data coming from patients affected by MS. In what follows, the used dataset is described and the results presented and discussed.

5.4.1 Results and Discussion

The 1500 images from MSSEG dataset all undergo standard preprocessing. Of the original images, 256 grey-level images, histograms were calculated in single amplitude values, but graphically represented in 64 bins (4 intensity levels for each bin) to improve readability. The data of a patient, selected randomly among the patients whose data were collected with the Philips 3T scanner, were used for reference. To this aim, from a central image r of the chosen examination, points 1 and 2 of the proposed algorithm were applied to calculate the histogram peak position of WM, GM and CSF and stored to be used for the correction of data from the other patients and scanners. At the end of the process all images must have the same internal contrast between the three segmented tissues.

The final results, showing the average peak shift and the corresponding standard deviation, in amplitude units, affecting original data (before preprocessing), after standard preprocessing (steps 1-5 of Figure 5.1) and after final preprocessing (steps 1-6 of Figure 5.1), are reported in Table 5.1. Amplitude displacements could be, for each class, positive, negative or null. What is reported in Table 5.1 is the average of the histogram peak position of the final image, after it has been recombined from the sub-images of the three classes, once they have been separately realigned. This is the reason why residual histogram mismatch also occur after correction. As it can be observed, also the reference scanner Philips 3T expressed a non zero displacement due to patient/scanner dependencies. Note that we are interested to the histogram stabilization, the local contrast equalization, not to the amount of the shift (a scanner is not better than another when the shift is lower: shift just depends on the scanner used for reference and it is not a quality parameter).

Results in Table 5.1 demonstrates that the amplitude shift has been almost completely corrected (residual final displacement remains inside a single bin, consisting of 4 intensity values, for all the considered scanners): obviously, the amplitude shift starting differences were so huge, also after the application of the z-score, that images appeared too different (see Figure 5.2) and needed to be corrected.

Scanner Type	No Prep.	Prep.	Prep. + C. Norm.
Siemens 1.5T	15 ± 3	12 ± 2	2 ± 1
Siemens 3T	18 ± 4	14 ± 3	2 ± 1
Philips 3T	4 ± 2	3 ± 1	1 ± 1

Table 5.1: Average displacement and standard deviation, in intensity units, without preprocessing (second column), after standard preprocessing (third column) and after final preprocessing (fourth column), separately for each scanner (rows).

As an example, Figure 5.4 shows the images of Figure 5.2, with the application of the method to the image of second row in order to make it normalized to that in the first row. The histograms of the resulting images are also reported for comparison (right column). As can be noticed, images have very similar contrast and the original differences, clearly visible in Figure 5.2 also after the application of z-score, is greatly reduced as confirmed by the histogram. Note that the proposed strategy has been applied to translate a higher contrast image to a lower contrast one: the opposite could have been done without any limitation.

The effect of preprocessing on segmentation is important because it stabilizes data and, pushing toward generalization, simplifies the role of the following segmentation/interpretation process. The proposed local contrast normalization method, as an integration to the general preprocessing strategy, greatly contribute to homogenize FLAIR images from different scanners. In this way, a relevant gain is furnished to radiologists, who are not forced to retrain themselves when using data from different scanners, but, more important, to automatic segmentation/interpretation strategies which could be trained by using lower data sets (being data yet equalized to different scanners, the automatic strategies would not need to be trained with data from several scanners) and, hence, the training process could be faster. Finally, as a really good consequence, the automatic strategies could result more general (data coming from a scanner which is completely new to the automatic systems would be effectively treated) and final accuracy would improve.

It is important to note that, with the exception of FLAIR in which the three above classes show well separated intensities, as shown in Figure 5.2, the proposed strategy has not been tested on other imaging modalities and it does not necessarily performs well in all of them, for example when the contrast among classes in the original image is very low, though a similar strategy could be also attempted to translate images from one imaging modality to the others. This could be very helpful in reducing acquisition time while maintaining the advantage of using information by different modalities (some modality could be calculated from others and not directly

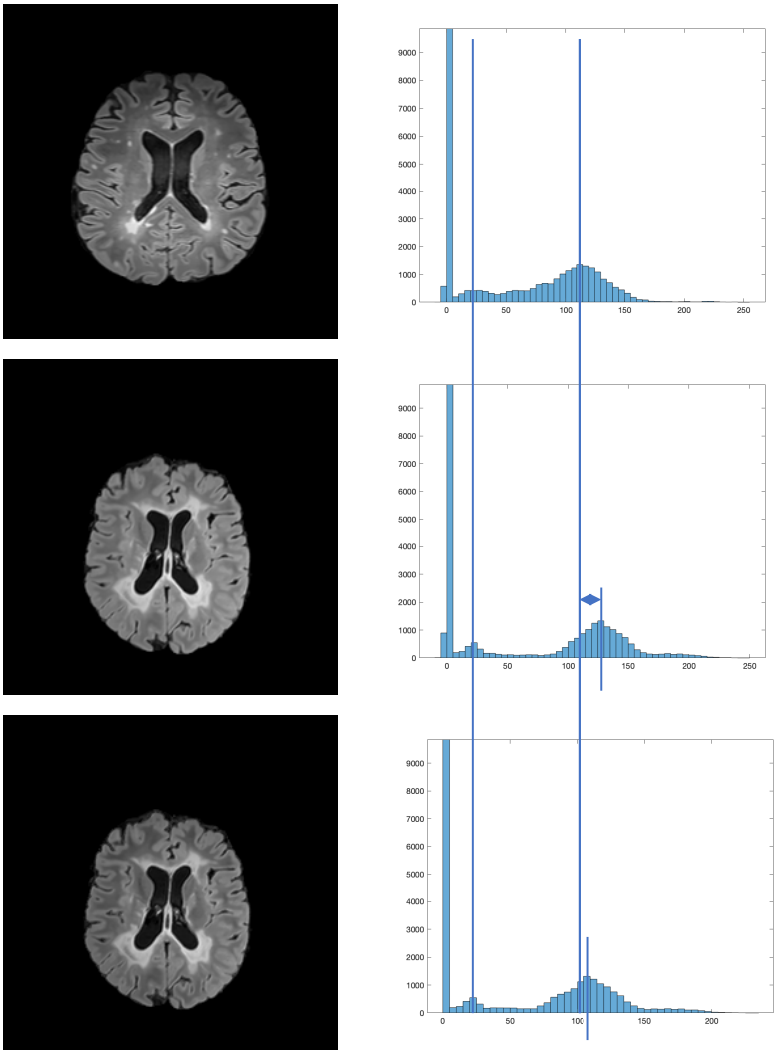

Figure 5.4: Local contrast normalization. The image in the third row represents the local amplitude realignment of the image in the second row to the image in the first row. Lines connecting histograms (right column) serve to evaluate the respective positions of the peak, before and after correction, with respect to the reference image. Vertical axis of the histogram was cut to zoom low values.

acquired).

5.5 Discussion

A new local contrast normalization strategy has been proposed for MRI FLAIR images of the brain. It is based on the preliminary segmentation of a reference image in WM, GM and CSF and on the calculation of the peak positions of the histograms of these segmented images. These values are used as reference amplitude positions. When using a new examination, its central image undergoes the same procedure as the reference image and the displacements from the reference images are calculated for the classes WM, GM and CSF. Finally, for each image, corrections are applied to the histogram of each class separately and then recombined in a single image. The proposed histogram normalization strategy has been experimentally tested on 1500 images from 3 different MRI scanners: its good performance has been numerically demonstrated and graphically illustrated.

The proposed strategy could be very effective to generalize and improve the training process and the final accuracy of automated segmentation/identification strategies due to its good property of reducing scanner and patient specific mismatches.

However, the proposed method is sequence (FLAIR) and organ (brain) specific: its application to other imaging sequences has not been attempted and, as it is, we think it could be easily extended to those imaging sequences whose original contrast between WM,GM and CSF is well defined; ts application to different body districts is out the scope of the research, being it included in a project for the study of the inflammatory processes of the brain [163].

Chapter 6
Multiple Sclerosis Lesions Identification/Segmentation in MRI using an Ensemble of CNN

Beyond the different images contrast produced by different MRI scanners (presented in the previous Chapter), inappropriate image interpretation and wrong application of MRI diagnostic criteria contribute to misdiagnosis. For example, it is difficult to distinguish pathologies such as neuromyelitis optica spectrum disorders, Susac syndrome and MS using MR images since some MRI diagnostic criteria are common. Even for the same disease, distinguishing healthy tissue from diseased tissue is not trivial. Regarding MS, white matter of the brain could appear normal despite the lack of myelin in it. Therefore, the "healthy" brain tissue is usually referred as "apparently healthy" [115]. Healthy anatomical structures similar to lesions and close to them could create further ambiguity. Moreover, partial volume effect (PVE) occurs when lesions and healthy tissue are present in the same place. These ambiguities originated by the unspecific nature of MRI with respect to a particular disease often create ambiguities and disagreement among radiologists (inter-raters variability) as well as uncertainty in a same radiologist (intra-rater variability) mainly in defining the borders of the lesions, but also in labeling whole regions. This could make the manual segmentation inaccurate and usually to avoid errors multiple manual segmentations are taken into consideration. Moreover, physicians partially manage the uncertainty generated by ambiguity relying on their personal radiological/clinical/anatomical background and experience but this information are not available during the training phase of AI Models. This chapter presents an automated framework based on three pivotal concepts to better emulate human reasoning:

1. the handling of the uncertainty (defined in Chapter 3) class;
2. the proposal of two, separately trained, CNN, one optimized with respect to lesions themselves and the other to the environment surrounding lesions, respectively repeated for axial, coronal and sagittal directions;
3. the definition of an ensemble classifier to merge the information collected by all CNN.

Compared to the framework proposed in Chapter 4, that analyzes patches of the images, this one analyze the whole image for training the method also regarding the position of the lesions inside the image (white matter) and to reduce outliers.

The comparison, made with the consensus (the ground-truth) between 7 human raters and with each of the 7 human raters, proves that there is no significant difference between the automated and the human raters. The results of our framework concerning the uncertainty are also reported, even if a comparison with the raters is impossible because they don't recognize this class.

The content of this chapter appeared in [3].

6.1 Introduction

Several automatic frameworks have been recently proposed for MS lesions segmentation [2, 124, 164, 165, 166, 167, 121] and also for evaluating MS temporal progression [168, 169, 170].

However, to date the results are still far from those of human experts, despite the efforts have been huge. Actually, this has led to an increase in the model complexity not corresponding to the expected improvement. Indeed, often state of the art methods have failed when tested on data from a different data set [171]. This mainly occurs because automated strategies are not robust to MRI variability, not even sufficiently able to model medical knowledge, human operational capacity and flexibility.

Regarding implicit medical knowledge and experience, they mostly remain unexpressed and are not reported on the labelled data sets used to train the automatic strategies.

The same regards the reasoning methodology used by radiologists during 3D data analysis: data are mostly analyzed in 2D axial slices with a continuous view of coronal and sagittal slices to confirm an hypothesis, to give spatial continuity to a lesion or to check the environment in which the hypothetical lesion is localized [172, 115]. A recent paper [173] highlights the usage of 3D CNN in the pipelines for MS lesion segmentation strategies, but this is quite different by another recent strategy [174] in which 2D U-net ensemble models are preferred for automated strategies for WM hyper-intensities evaluation in a way which is similar to the human methodology.

Further, the uncertainty affecting expert radiologists when classifying some regions is not reported in the public data sets used for training: a binary choice is often insufficient to represent the evaluation of an expert. If represented, the uncertainty could greatly help an automatic strategy to better segment also undoubted lesions. This pushed several scientists to investigate on uncertainty in medical data [175, 176] and to the effect that the rater style could transfer in terms of uncertainty

to an automated strategy [116].

Implicit information in automated methods is difficult to be modelled and more, if introduced with external supports (dictionary, anatomical atlases, etc.) [177], is insufficient to fill the gap with human experts.

We aim at filling this gap, both in performance and in reasoning, by proposing a framework which includes:

1. the classification of uncertainty as an intermediate class between the background and lesions;
2. the optimization of two CNN (2D U-net models), one for the class lesion and one for the class background to contextualize lesions with respect to the surrounding anatomical structures for the three spatial directions (axial, coronal and sagittal);
3. the definition of an ensemble classifier to merge the information collected by all CNN.

just on FLAIR images.
In this chapter is presented:

1. the usage of uncertainty to emulate uncertain reasoning for improving lesion identification/segmentation;
2. the contemporary exploitation of lesions and lesions in the context of the surrounding environment, for all the spatial directions;
3. the definition of the ensemble of CNN-based automated raters approaching the problem from different points of view; the demonstration that just a single MRI modality, FLAIR, is sufficient to classify/segment MS lesions in WM;
4. the demonstration that an automatic strategy behaves and performs like a human expert.

6.2 Related work

Medical image analysis is greatly performed with automated methods, mostly involving deep learning [178]. Automated MS lesion identification/segmentation is still an active field of research and several methods have been provided in the last decade and well reviewed along time [179, 180, 181, 2, 124, 164, 165] and the role of AI-based methods is emerging [182]. Automated strategies can be classified into

three main groups: methods using pre-selected features modelling (PSFM), methods using a-priori information modelling (APIM) and methods using deep learning modelling (DLM).

PSFM calculate pre-selected features and learn from previously segmented training images to separate lesions from healthy tissue [183]. Some PSFM use a large set of features and select the more discriminant ones through labelled training. One of them is an atlas-based technique, employing topological and statistical atlases for WM lesion segmentation [184]. Another includes the usage of Decision Random Forests [185]. Similarly, a framework for segmentation of contrast-agent enhanced lesions using conditional random fields is defined in [186]. [187] propose a set of features, including contextual features, registered atlas probability maps and an outlier map, to automatically segment MS lesions through a voxel by voxel approach. A rotation-invariant multi-contrast non-local means segmentation is proposed in [188] for the identification and segmentation of lesions from 3D MRI images. Supervised learning by PSFM has been widely employed in tasks where the training database and the pre-selected feature set cover all possible cases [189]. Nevertheless, when the heterogeneity of the disease and the potential variability of imaging are large, as it occurs for MS and MRI, the dimension of the training database and, mostly, the choice of the pre-selected features are critical.

APIM does not require labelled training data to perform segmentation, but usually exploit some a-priori information, such as the intensity clustering, to model tissue distribution [190]. In [191], a likelihood estimator to model the distribution of intensities in healthy brain MR images is presented. Other methods use threshold with post processing refinement [192, 193] or are based on probabilistic models [194, 195]. A big challenge for APIM is that the outliers are not specific for lesions because they could be due to artifacts, intensity inhomogeneity and small anatomical structures like blood vessels: this often produces false positives [196]. Moreover, APIM is strongly based on the information extracted and simplified by the knowledge of specific experts.

Though the dimension of the training database is also crucial in DLM, this has no concern regarding the pre-selection of features as in PSFM or regarding a-priori information modelling as in APIM.

In fact, during the last years DLM has gained popularity in medical imaging especially with CNN [197] and, in particular, with U-nets ad their variants [198, 199, 200, 201]. CNN, compared to machine learning approaches, has achieved remarkable success in biomedical image analysis [196, 202, 203]. DLM trains and learns to design features directly from data [204] and provides best results in MS lesion identification/segmentation [164, 205, 206, 207, 121, 167]. This has also been confirmed in recent reviews [2, 124, 164, 165].

CNN applied to MS often use 2D spatial convolutional layers [208, 204], others use 3D convolutional layers to incorporate 3D spatial information simultaneously [209, 210, 211, 212] or merge spatial with temporal information [168, 170]. All these methods perform segmentation with a minimum lesion volume threshold to avoid the inclusion of small outliers.

However, CNN performance is still far from that obtained by human experts or its performance dramatically drops with other data sets [171]. In what follows, we present a robust framework based on CNN that can reach human performance if some human methodological insights are modelled in it.

6.3 The proposed framework

The framework we propose, sketched in Figure 6.1, consists of the following steps:

1. deep learning automatic classification, of the images (2D) composing the MRI model, in three classes: Background, Uncertainty and Lesion (capital letter to imply the concept 'class'), optimized for Lesion (lesions from inside) and for Background (lesions seen in the context of the surrounding environment), separately for axial, coronal and sagittal directions (resulting in 6 classifiers);

2. class fusion (separately for Lesion and Uncertainty, starting from Lesion) by performing the Union of the 2 axial segmentation (step 2a in Figure 6.1), followed by a majority vote taken from the remaining segmentations and used for confirmation of the class (if the class is not confirmed, this is downgraded (step 2b in Figure 6.1);

3. final output.

For the framework we propose, the following three hypotheses hold:

1. the MSSEG pre-processed data from just one single MRI modality, FLAIR, are the input of the framework;

2. the binary labelled ground-truth is revised to contain, besides Lesion and Background, also the Uncertainty class which is created from part of the original Background, and leaving Lesion unchanged (see Chapter 3);

3. the three considered classes are supposed to be ordered, Background $<$ Uncertainty $<$ Lesion: as far as just Lesion and Uncertainty are the subjects of fusion, their downgrading consists in the passage from Lesion to Uncertainty and from Uncertainty to Background, respectively: the process starts from Lesion to allow Uncertainty fusion on the upgraded data set.

Step 2a of Figure 6.1 serves to include, besides common information, also complementary information coming from the specificity of each of the two axial CNN and, at the same time, to model the reasoning of radiologists who use axial orientation to make the first hypotheses. Step 2b is used to vote for each object resulting from the axial processing (Lesion or Uncertainty), its permanence in the assigned class, or its downgrading. Objects are confirmed when at least two of the other four raters (two coronal and two sagittal) agree with the axial classification. This ensures that false positives are greatly reduced and that 3D contextualization with the environment is maintained. In this way, the model agrees with the radiologist's reasoning regarding the usage of coronal and sagittal orientations, so to have a confirmation of the hypothesis and to better define 3D object continuity.

The choices regarding the usage of one single imaging modality, the classification in three classes and the use of an ensemble framework are clarified below.

Being supervised, each classifier needs training, validation and test carried on by using data from a public data set. In what follows we first describe the used data set, the ternary ground truth, the CNN architecture, the used loss function, the hyper-parameters optimization and the ensemble, final, classification of Figure 6.1.

6.3.1 CNN architecture

The task we are facing with is the classification of a FLAIR volume, separated into slices, in one of the three classes: Background, Uncertainty and Lesion. Since U-nets [198] are specifically designed for these tasks, we use the U-Net 2D architecture depicted in Figure 6.2 to classify the images composing a volume.

The U-Net is a fully convolutional neural network composed by 2 main sections, Contraction and Expansion, connected by a Bottleneck section. The corresponding Contraction and Expansion modules are also connected through skip connections.

Compared to the traditional U-net architecture, we insert a batch normalization layer in each block to mitigate the effects of the gradient amplification [213] in the regions surrounding the lesions, though this with a relevant increase of computational costs (about 30%).

An important parameter for a U-net is the number of blocks in the Contraction and Expansion sections. If the number of blocks is too low, the network could not have enough features for learning complex structures. On the other hand, if the number of the blocks is too high, the network memorizes complex structures (overfitting).

To optimize the number of blocks, n, we have performed preliminary training, with $n \in \{3, 4, 5\}$. We have not gone outside this set because for $n = 5$ the U-

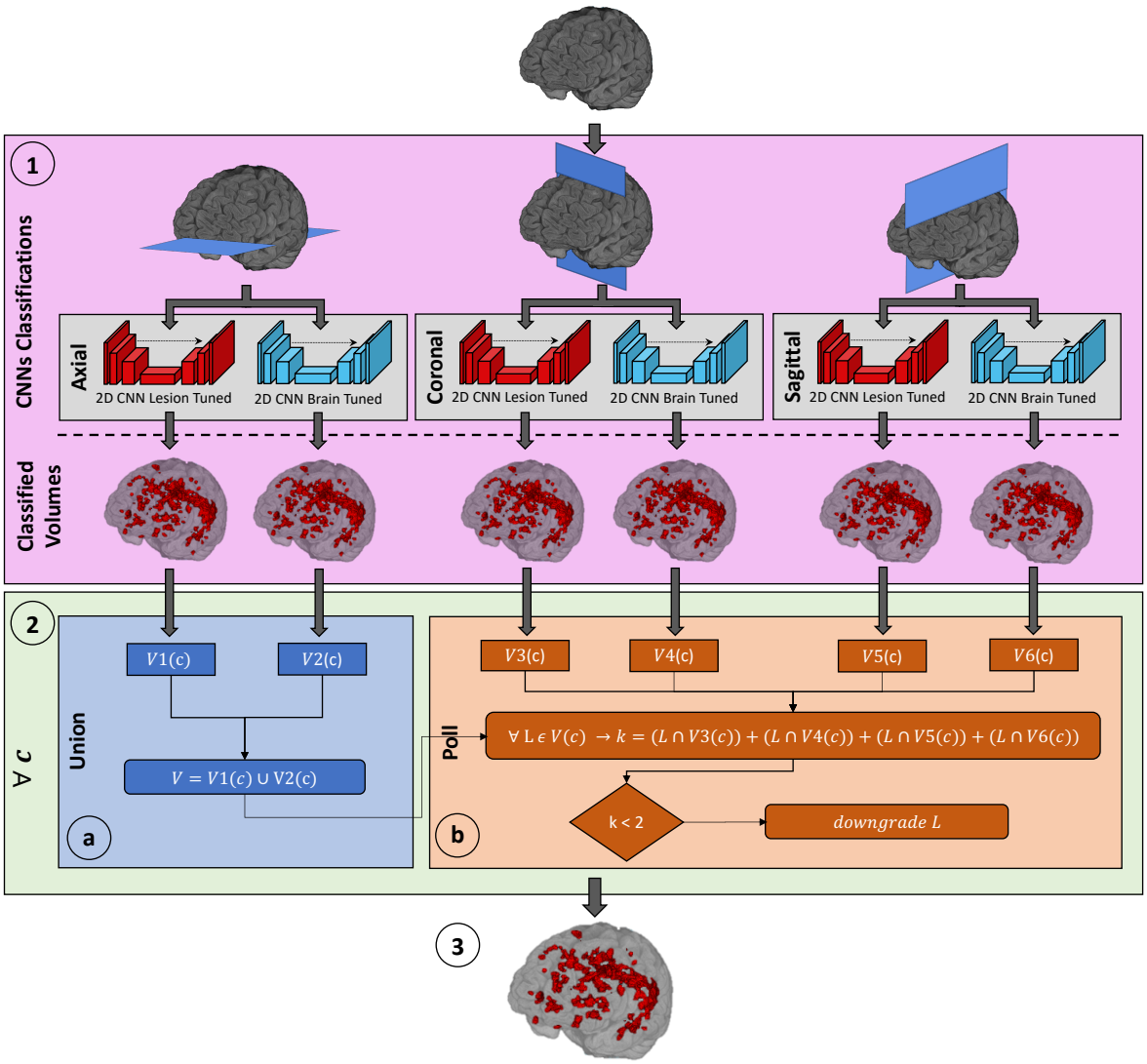

Figure 6.1: The proposed identification/segmentation pipeline which divides the brain tissue in three classes: healthy tissue (Background), tissue that has uncertain nature (Uncertainty), and MS lesions (Lesion). The strategy operates independently on axial, coronal and sagittal images, each processed by two separately trained U-nets, one optimized for Lesion, to directly focus on lesions, and the other optimized for Background, for contextualizing lesions with respect to the environment. After that, it recombines the results by using the Union of axial volumes followed by a majority vote strategy on the coronal and sagittal volumes, for confirmation. Voxels whose classification is not confirmed are downgraded (Lesion becomes Uncertainty and Uncertainty becomes Background). The framework operates separately for Lesion and Uncertainty, starting from Lesion. In step 2b, the procedure is applied voxel by voxel: L is referred to each single voxel of the class $c \in \{Lesion, Uncertainty\}$.

Net started to overfit, even when using high values of L_2-Regularization, and for $n = 2$ a dramatic drop of performance occurred. With $n = 4$, the problems related

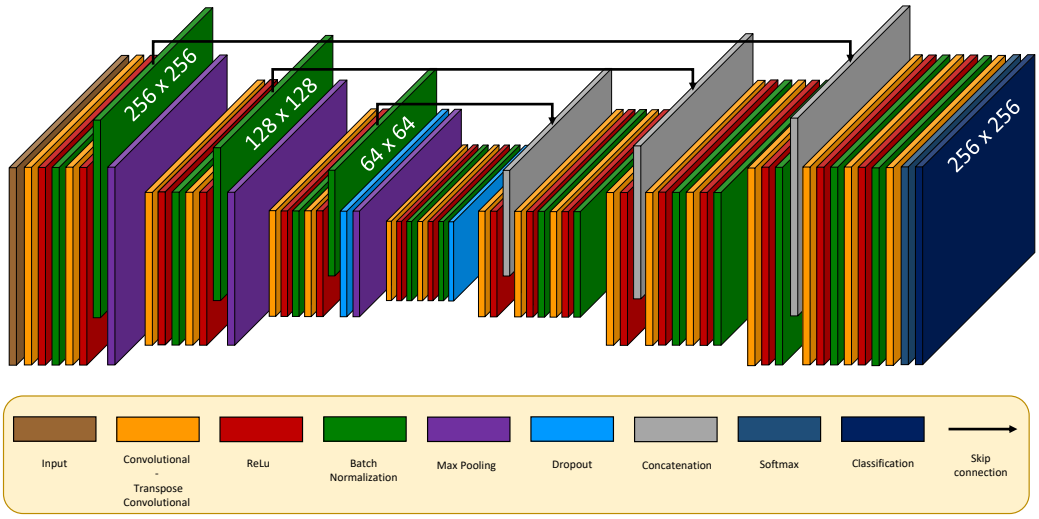

Figure 6.2: The used U-net "D architecture. The architecture is the same for 6 classifiers, though they have been trained separately.

to overfitting have disappeared and the performance was good. However, we have noticed from the feature maps that some redundancy is present. For this reason, we have trained the CNN with $n = 3$ and verified that redundancy is greatly reduced and training converges faster than for $n = 4$: hence, $n = 3$ is the number of blocks used thereinafter.

6.3.2 Loss function and process optimization

The architecture we use has to solve a three class automatic annotation, for which a Multi-label Cross Entropy Loss Function is necessary, defined as follows:

$$loss = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K}(T_{ni}log(Y_{ni}) + (1 - T_{ni})log(1 - Y_{ni})) \tag{6.1}$$

where N and K are the numbers of observations and classes, respectively.

The use of three classes, besides the problem stabilization (the presence of the Uncertainty class gives better confidence in defining both lesions and background), allows also to consider another important aspect. Indeed, we can optimize two CNN, sharing the same architecture and the same loss function (Eq.6.1), but with a different learning process deriving from different focuses: one optimized on the Lesion and the other on the Background (the environment in which lesions are immersed). The Uncertainty, is used as a sort of "*buffer class*". In the case of a binary classification problem this would not have been possible: the optimization of

one would automatically lead to the optimization of the other (what is not Lesion is Background and vice-versa). The usage of the Uncertainty gave to both CNN a new choice to break that constraint.

The training process of a neural network can be controlled through hyperparameters. Different hyperparameters lead to a different learning path and, finally, to a different performance of the neural network. In literature, it is well known that the hyperparameter optimization serves to achieve faster training and better performance [214, 215, 216]. In this study, the hyperparameter optimization is also used to train the two CNN separately, which leads to different paths discovered by the Gradient Descent.

The hyperparameter setting is driven by automatic optimization through a Bayesian approach [217]. Besides, the hyperparameters to be optimized are:

1. **Starting Learning Rate**: it is related to the data set and to the type of neural network.

2. **L2-Regularization**: it prevents overfitting.

3. **Class balancing**: it optimizes the amplification factor for the represented classes and improves training. Here we have three classes, hence two weights are sufficient (Lesion Weight and Background Weight.

Of the above, the first two are standard for CNN, while Class balancing is specific for our CNN because it helps to differentiate the path of optimization between the CNN optimized with respect to Lesion and that optimized with respect to Background.

The resulting optimization problem is the following:

$$x^* = argmin_{x \in X} f(x) \tag{6.2}$$

where X is the domain of x, $f(x)$ represents an objective function to be minimized and x^* is the hyperparameter setting that yields the optimal value of $f(x)$.

In this study $f(x)$ is defined as

$$f(x) = 1 - IoU(x) \tag{6.3}$$

where IoU is the Intersection over Union score [2] defined in Chapter 3.

Regarding the two CNN used therein, for that optimized for the Lesion, the IoU is calculated with respect to the Lesion class and, for that optimized for the Background, the IoU is calculated with respect to the Background class.

Table 6.1 reports the hyperparameter setting for both the optimized CNN ('In' and 'Out' indicate Lesion optimization and Background optimization, respectively)

CNN	Learning Rate	L2-Reg.	Lesion Weight	Background Weight
Axial In	5,32E-04	3,66E-10	7,98E-02	8,91E-01
Axial Out	4,54E-04	1,14E-10	2,99E-02	9,00E-01
Cor. In	6,50E-04	3,66E-10	7,98E-02	8,71E-01
Cor. Out	3,18E-04	7,92E-09	6,59E-02	8,58E-01
Sag. In	1,01E-04	5,53E-09	7,01E-02	8,74E-01
Sag. Out	1,08E-04	3,41E-09	6,02E-02	8,50E-01

Table 6.1: The hyperparameter values for the CNN, each trained with the corresponding oriented images. The suffixes In and Out are used to indicate whether Lesion or Background is optimized, respectively.

in each direction (axial, coronal and sagittal). As it can bee observed, the overall change of the hyperparameter setting justifies different training paths for the CNN and different points of convergence for each of them. Figure 6.3 shows the different behaviour of the two CNN highlighted in the segmentation grad-cams of a sample image, for the axial direction. The CNN optimized for Lesion, tends to enlarge Lesion and Uncertainty with respect to the CNN optimized for Background, which surrounds lesions from outside.

6.3.3 Ensemble Classification

It is well known that ensemble classifiers often perform better and more robustly than their single components [218, 219, 220, 174]. For the classification of lesions we use 2D slices (axial, coronal and sagittal) of the whole volume, with specific CNN trained separately with axial, radial and sagittal slices, respectively. In this way, we can avoid that a particular orientation could be favourable to lesions (the classifier is deceived) or to the classifier (the good classification of some lesions could be a lucky outcome). Further, it serves to ensure 3D continuity to the classification. Moreover, as explained above, we look at lesions both as they are and with respect to the surrounding environment.

We obtain a set of 6 classifiers, two for each of the three orientations, axial, coronal and sagittal, whose classification has to be merged to produce a single output resembling the reasoning of the radiologists: though 3D FLAIR data are collected following sagittal planes to account for clinical/physical issues, radiologists often use axial images for data interpretation and use the other orientations for confirmation [172, 115]. Accordingly, we prefer axial classifications and use coronal and sagittal output for confirmation.

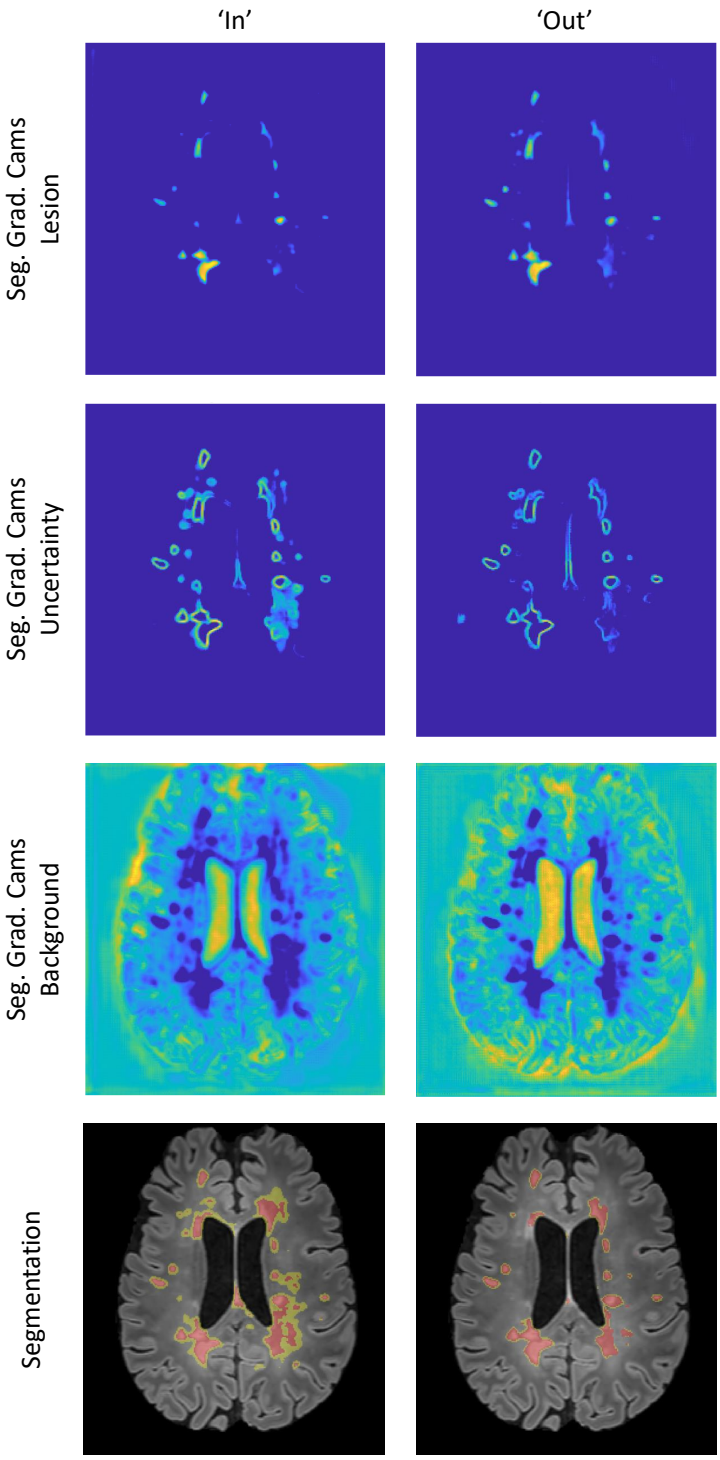

Figure 6.3: Grad-cam representation for an axial sample image for both CNN, Lesions from inside (left) and from outside (right), respectively. Grad-cams are shown for the three classes, Lesion (first row), Uncertainty (second row) and Background (third row). The fourth row shows the classification of each of the CNN.

Regarding axial classifications, since each of the two CNN referring to the same direction operates in the same scenario but with different approaches, they collect

specific information in those regions where reasoning is specific. Since both specific contributions are important, besides common findings, a Union operation between the two classifications is required. This is in accordance with the procedure which expert radiologists would ideally follow [221]. Other forms of fusion, for example statistical fusion through STAPLE [108, 119, 120], are inappropriate because in our case we are combining classifications obtained by different approaches and not classifications from similarly reasoning raters (where it is supposed the use of almost the same approach). Regarding this last point, it is important to notice that also human experts with different experience could assume different decisions [175].

However, being the classified volume a three values data set, the Union does not correspond to the classic binary union operation. In our case, the Lesion is privileged, then comes the Uncertainty and, finally, the Background. In fact, a voxel is classified in a lesion if at least one of the two classifications considers it as a lesion; elsewhere, if at least one of the two classifications considers it as uncertain, it is classified as Uncertainty, elsewhere, it is considered as Background (both classifications affirm the voxel is Background).

After the Union application, false positives are more present than in each single classifier: their number is reduced by using the majority vote between the other 4 classifications (two coronal and 2 sagittal, being the comparison performed along axial planes). In fact, for each voxel the class is maintained if at least two of the other classifiers confirm it, elsewhere it is downgraded by one (a potential Lesion becomes Uncertainty, a potential Uncertainty becomes Background): a double step is not allowed. There is always at most 1-step downgrading simultaneously. This means that first a decision is made on the Lesion and, then, on the Uncertainty by using the data set resulting from the application of the process to the Lesion.

The ensemble of different classifiers is justified both by the fact that the Union has to join common information, as well as specific information from each of the two axial classifiers, and because each potentially positive voxel needs confirmation from the coronal and sagittal classifiers (in this case, also 3D spatial information is considered). Though we don't have measured the diversity degree between classifiers [218] we have preferred to resemble the usual procedure used by radiologists and to rely on the usual benefits of using an ensemble classifier [219].

In the proposed automatic pipeline, we have copied the human behavior by privileging axial sections with respect to the others, but we have also performed trials regarding the preference of the other orientations in the fusion process and the results, not shown, confirm that the axial preference gives the best results, closely followed by the coronal and, at a great distance, by the sagittal, though this is the direction used for 3D FLAIR data collection. The thing could be explained, at least partially, by the fact that axial and coronal slices show highly symmetrical shape

both regarding brain anatomy and lesion shapes, also across different subjects, thus making the learning process easier than for sagittal slices. For sagittal directions, in fact, symmetry is absent and a huge variation of the image content could correspond to a little rotation of the head.

6.4 Results

The proposed framework has been trained, validated and tested on the ternary consensus defined above. As far as the ternary consensus maintains unaltered all the lesion voxels of the original MSSEG binary consensus, we can guarantee a direct comparison with human raters and, at the same time, with already existing automated methods tested on the same data set. Regarding the segmented Uncertainty, a comparison is possible just with respect to the ternary ground-truth, since for the human raters Uncertainty is unavailable. In principle, we could define the Uncertainty for each rater by considering as uncertain the voxels that the rater has considered Lesion while the binary ground-truth has not. Though we have made some experiments in this direction, we believe this comparison should deserve a specific and deep discussion, being it based on approximated hypotheses (the intention of each rater would be guessed, not real), which is out the scope of this manuscript.

The evaluation of all the raters, of the proposed framework and of the human radiologists, with respect both to the ground-truth and to each-other, is performed by applying the cross-validation approach defined in Subsection 3.1. Average and standard deviation values are calculated for the metrics defined in Chapter 3 and divided in two groups: those whose ideal value is 1 and those whose ideal value is 0.

The first results, reported in Fig 6.4, are those between the raters and the ground-truth performed on the Lesion class. This is also an indirect comparison, through the ground-truth, between the proposed framework and the human raters. For a better overview, the mean values are also shown in Fig. 6.5 by using a radar visualization: they confirm that the behaviour of the proposed method is inside the inter-rater variability.

As it can be observed, our framework is never the best or the worst, for at least one of the metrics, as instead occurs for human raters. This can be explained, at least partially, by the fact that it has been trained with the consensus that, for its nature, tends to average pros and cons of the raters from which it has been derived.

A Wilcoxon signed-rank test of the vectors of metric values confirms, with a significance level of 0.01, that there is no significant difference between the behaviour of our method and that of the 7 human raters, with respect to the ground-truth, on the Lesion class. This means that, if data are shown without labels, it would be impossible to recognize the automated rater from humans.

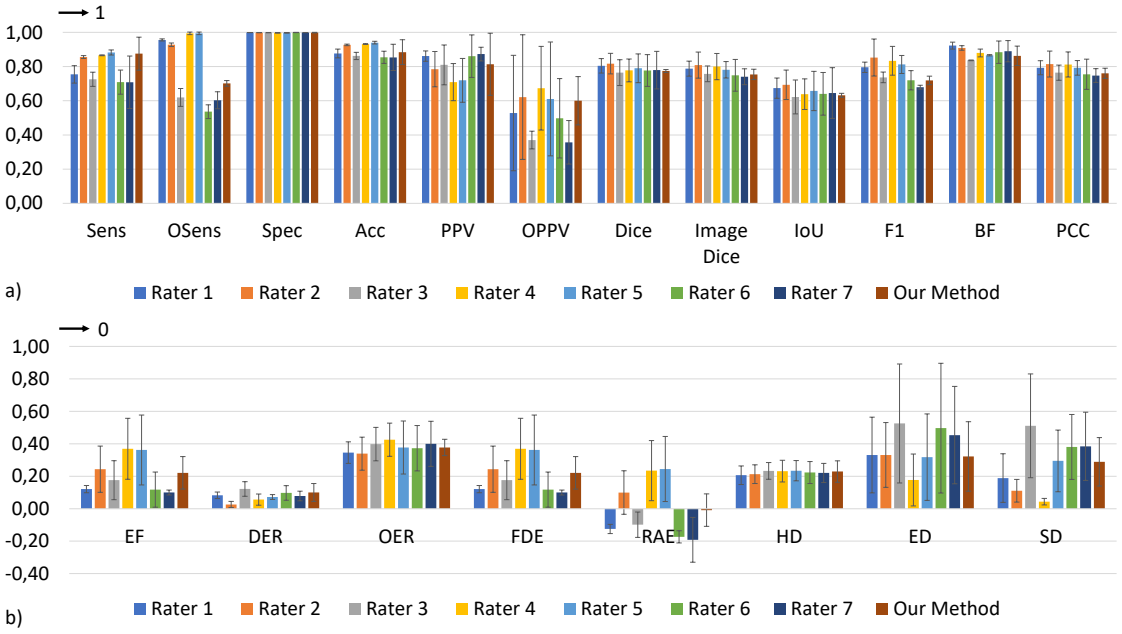

Figure 6.4: Comparison between the raters and the ground truth, performed on the Lesion class. The reported metrics are separated in those whose ideal value is 1 (a) and those whose ideal value is 0 (b). Average and standard deviation are reported. Euclidean, Hausdorff and Surface distances are shown in cm units.

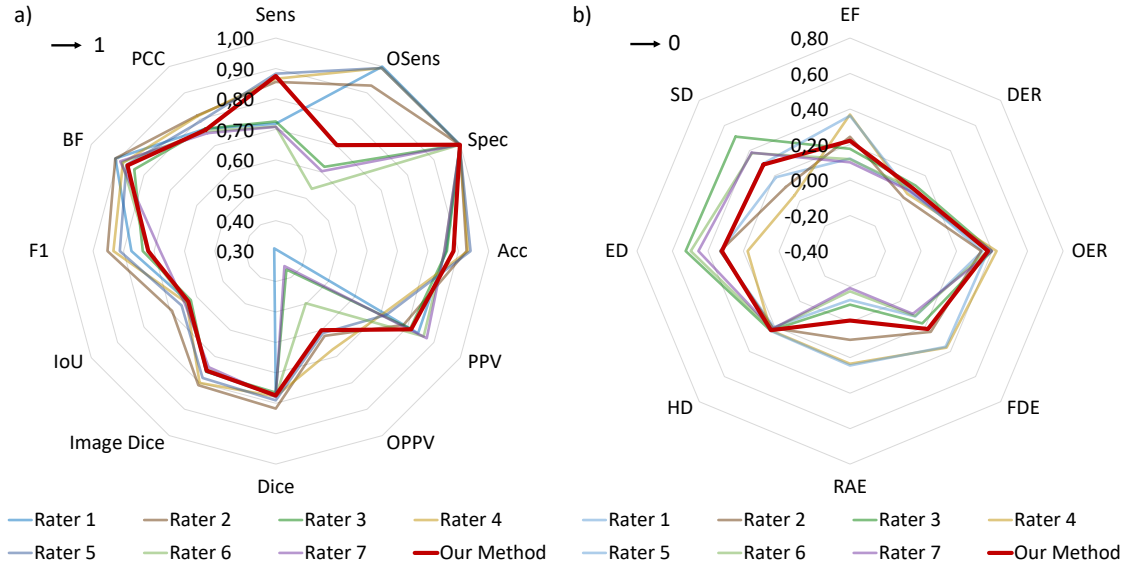

Figure 6.5: Comparison between raters in the same conditions of Fig. 6.4. For graphic purposes, only the average values are reported and the line of the proposed framework (red) is highlighted with respect to the others.

As far as the lesion size could greatly affect the performance of the classification [124] and the previously reported results are averaged with respect to the lesion volume, we repeat the comparison by changing the lesion volume. To this aim, we

consider all the lesions separately and the calculations are performed lesion by lesion, by maintaining separated also the lesions of the same volume. In this way we can: 1) visualize potential outliers; 2) represent lesion density; 3) avoid local averaging that could mask specific contributions to the metrics. The results, reported just for the most commonly used metrics [124], are shown in Fig. 6.6.

Results again confirm the analogy of behavior between the proposed framework and the 7 human raters, though for some volumes our framework has exhibited results close to the borders of oscillation of the 7 human raters. Indeed, a greater dispersion can be observed for F1-score and a relatively high average value is observable for SD, though in line with that of some human raters.

The above good results are not sufficient alone to affirm that our framework behaves like human raters because the comparison is mediated through the consensus. In other words, our framework could be at the same 'distance' as the human raters are from the ground-truth, but from opposite sides. For this reason, a direct comparison is necessary to finally confirm the similarity between the proposed framework and the human raters. To this aim, we perform the experiment of comparing all the raters to each other by considering ground-truth all of them, including our framework, in rotation. Results, for the most frequently used metrics, are reported in Fig. 6.7. In this representation, placing together metrics converging to 1 and to 0, the angular position indicates the metric value: clock wise versus for metrics converging to 1, anti-clock wise versus for metrics converging to 0. Radial information is only used to separate the current ground-truth raters. Ground-truth raters have colored bullets placed on the vertical line.

These results unequivocally confirm that the proposed framework behaviour does not differ from that of the other human raters and that it is not polarized toward a specific rater or toward the consensus. Moreover, as other Authors have highlighted [121], results show similarities among some human raters (R4 with R5 and R6 with R7). Fortunately, results also confirm that the MSSEG consensus is not biased by the similarity between some couples of raters, and that it maintains a "human" behaviour, being it very close to the raters R1 and R2. This is a fundamental aspect because it means that all the people who are attempting to train automated systems with respect to the MSSEG consensus, including ourselves, are not following a "chimera" to which, paradoxically, the closer we get, the more we move away from the proper objective.

After discussing the behaviour of the proposed framework regarding the Lesion, we also have to look at the classification regarding the Uncertainty class. To this aim, Fig. 6.8 reports average and standard deviation values for the metrics calculated with respect to the corresponding class of the ternary ground truth. If compared with the values obtained for Lesion, results are very poor, especially for metrics

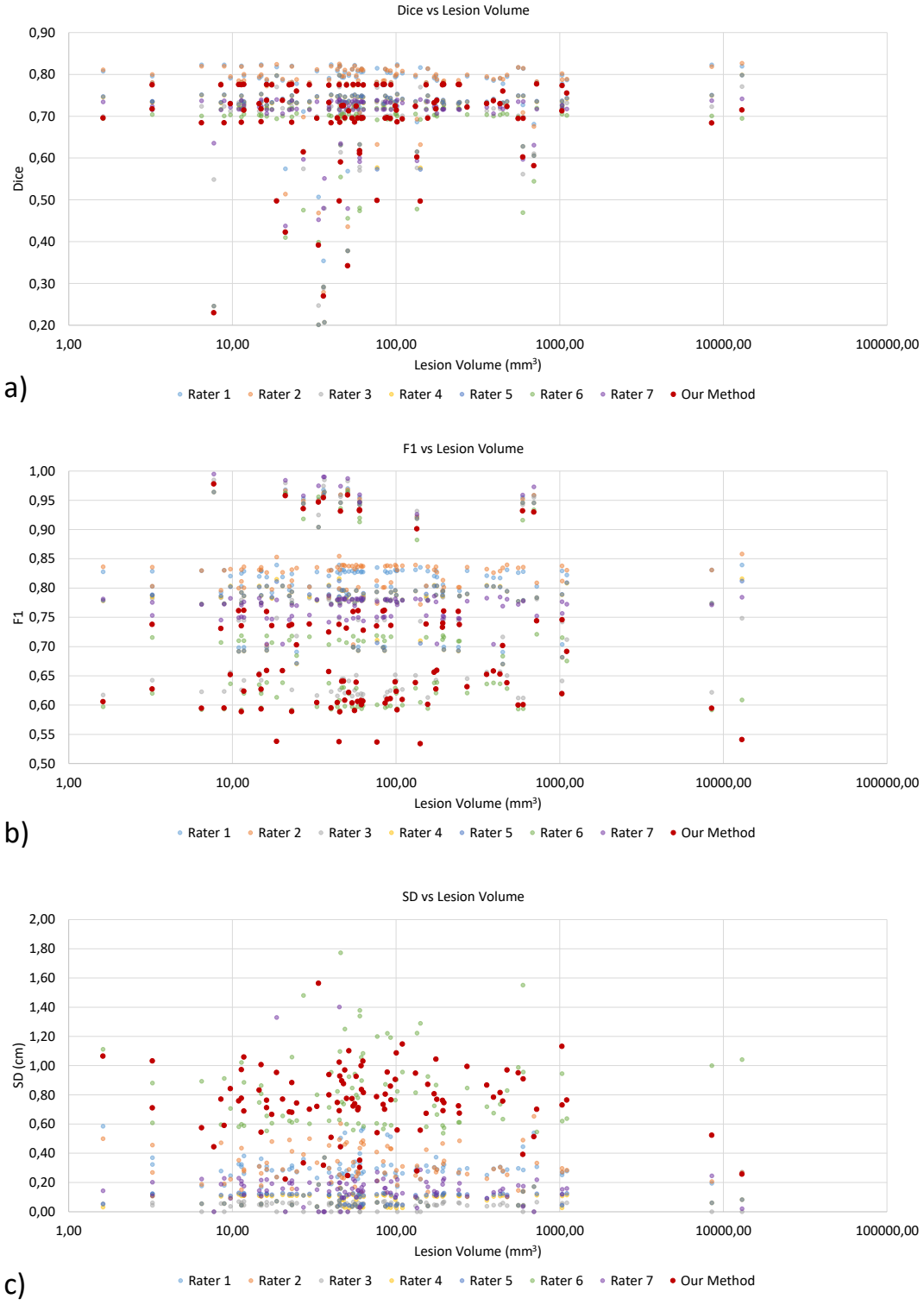

Figure 6.6: Dice score (a), F1-score (b) and Surface Distance (c), calculated for each lesion and shown with respect to the lesion volume for the human raters and the proposed framework. To improve readability, the logarithmic scale is used for the lesion volume and framework's values (red) are highlighted.

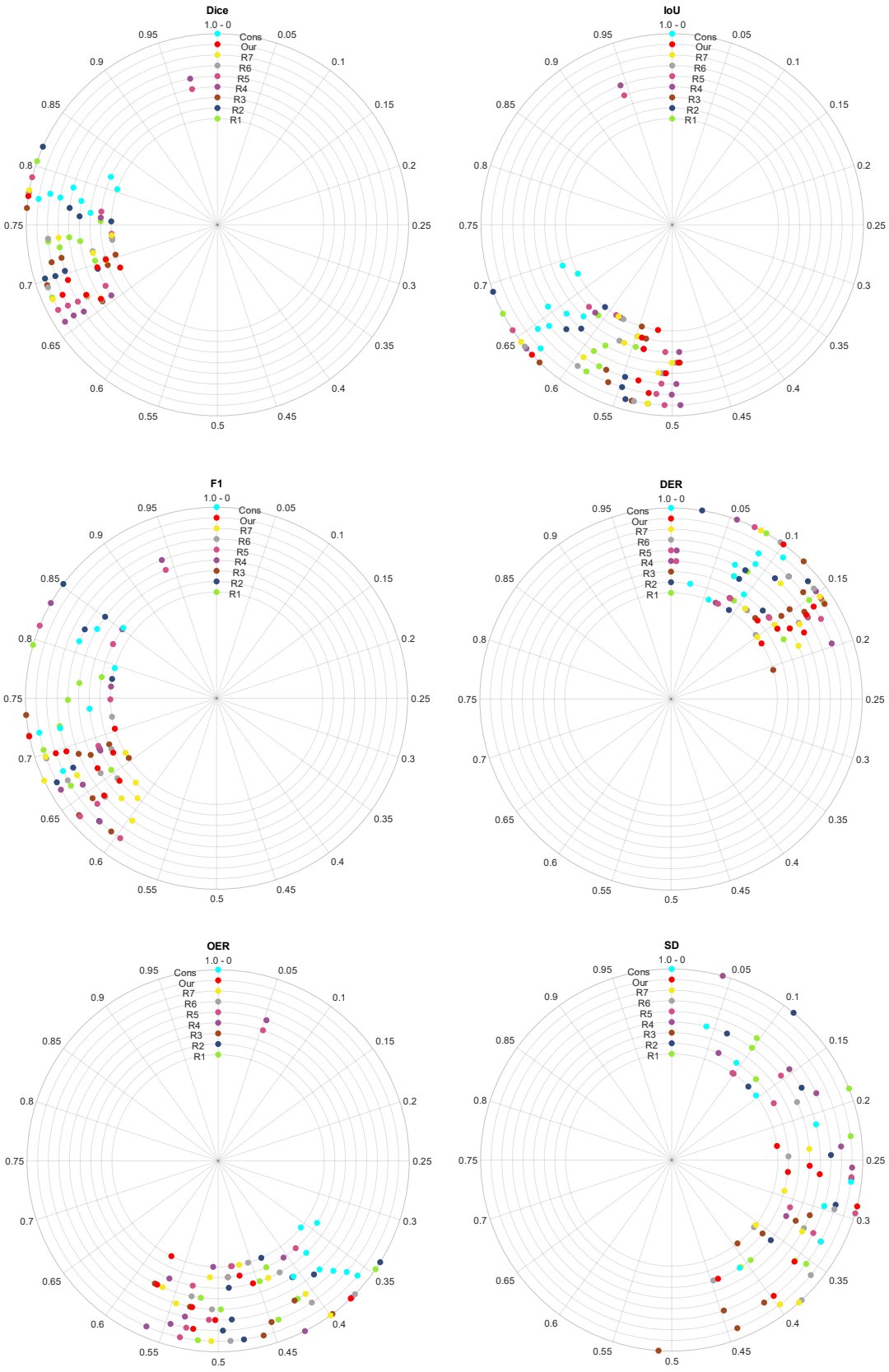

Figure 6.7: Comparison between all the raters with respect to each other, including our framework and consensus, each alternately considered as the ground-truth.

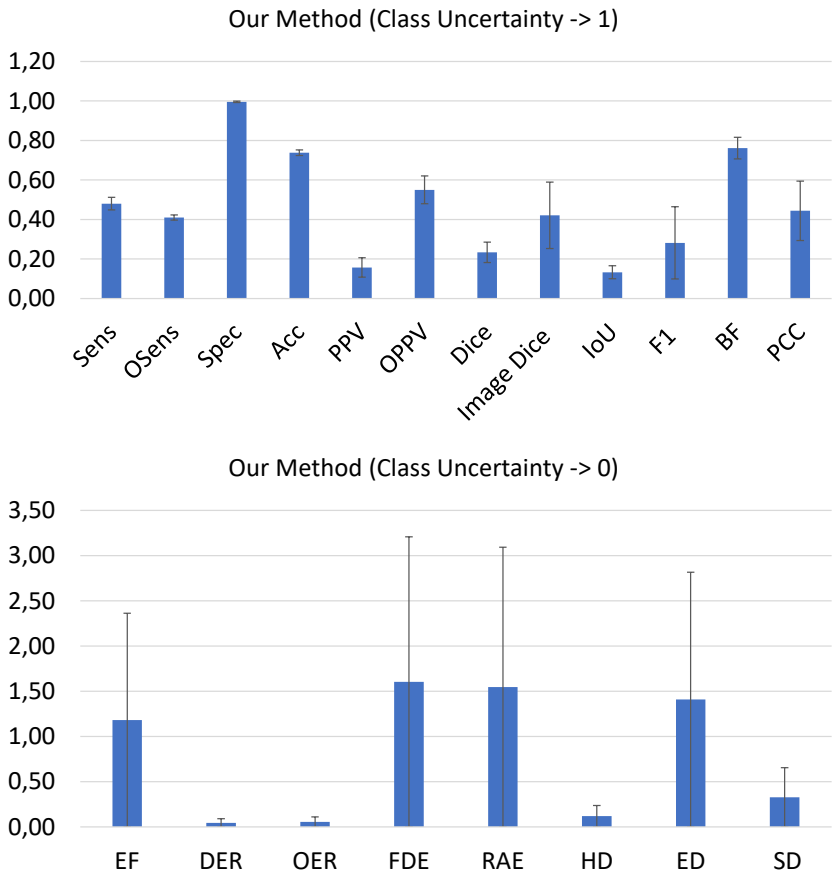

Figure 6.8: Average and standard deviation values of the metrics calculated for Uncertainty with respect to the same class in the ternary ground-truth. Values are represented as those converging to 1 (a) and those converging to 0 (b).

whose ideal value is 1. However, the 3D annular shape exhibited by Uncertainty around lesions, which includes two borders (external and internal) often discontinuous, greatly contributes to lower the results. Moreover, we don't have human references for Uncertainty and, for this reason, a comparison is impossible. Hence the results for Uncertainty are just reported for completeness and future comparison.

A visual overview of the behaviour of the proposed framework in the whole process of identification/segmentation, both for Lesion and for Uncertainty, with respect to the ternary ground truth, is shown in Fig. 6.9. The ternary ground-truth is reported on the left side, the corresponding segmentation obtained with the proposed framework is presented on the right side, for the same subject and slices. Both for Lesion and Uncertainty, the proposed framework selects more than necessary (FP are evident). Interestingly, FN are almost absent from the segmented volume. The other interesting property shown by the proposed framework is the good spatial continuity of the lesion structures in the 3D model of Lesion (upper right panel, where Lesion is red colored while Uncertainty is yellow colored).

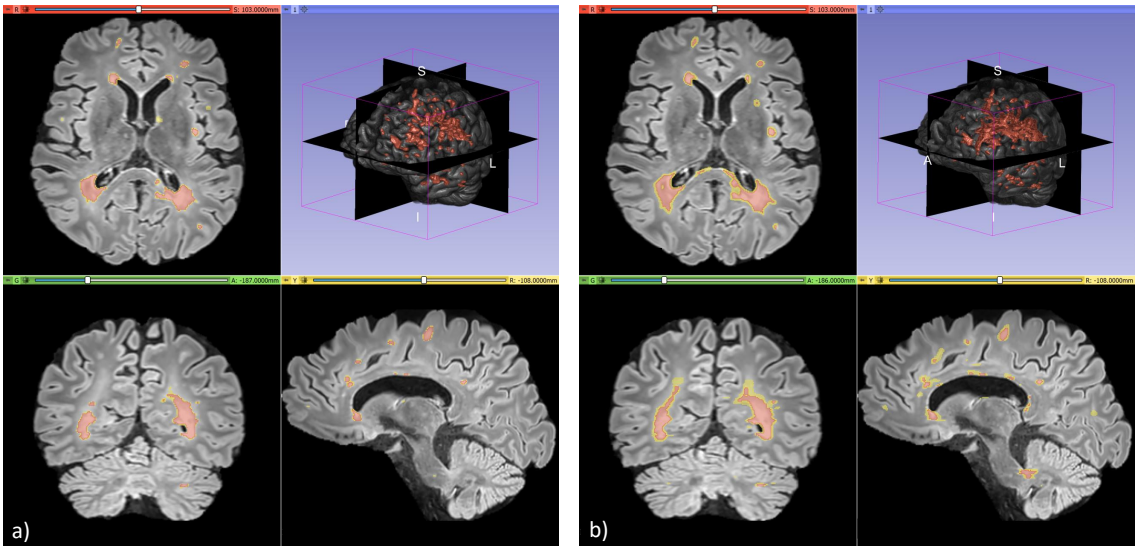

Figure 6.9: Comparison between the ternary ground truth (left) and the proposed automated framework (right). Lesion is red and Uncertainty is yellow. For readability purposes, the upper right panel of each side shows just the healthy brain and Lesion in 3D.

To complete our discussion, it remains to compare the proposed framework with recently proposed automated strategies. To this aim, Table 6.2 contains this indirect comparison on the metrics calculated for at least one of the other methods. The necessary condition for a method to be considered in Table 6.2 is to have been trained, validated and tested on the 2016 MSSEG data set. In this way, we can ensure that the comparison is homogeneous and performed on the same conditions of that obtained with respect to the 7 human raters.

Though a global ranking is difficult, data reported in Table 6.2 are clear: the proposed framework is the most stable with respect to different metrics and it generally outperforms the other methods, including those methods which use multiple imaging modality. This could have an interesting implication: as the automated framework performs like human raters just using FLAIR, it means that FLAIR would contain sufficient information, not only the one necessary, to identify and segment all MS lesions occurring in the WM, independently of their stage. Potentially positive consequences are: a) due to the huge variability of MRI and of each single modality, described above, the usage of a single modality could increase the performance above the use of multiple modalities because it could greatly contribute to stabilize automatic identification/classification; b) the acquisition time and stress for the patient can be reduced.

An important aspect that has determined the outstanding performance of the proposed framework is the use of the ternary ground-truth. Indeed, Fig. 6.10 shows the average performance results when the proposed framework is trained without the

Table 6.2: Comparison between the proposed framework and the state of the art methods. Average data are reported for each metric and the symbol '-' is used when data are unavailable. The reported metrics are those on which at least one method different from the proposed framework has been evaluated.

Method	MRI mod.	Sens	OSens	TPR	Acc	PPV	OPPV	Dice	F1	SD
Team Fusion in [124]	FLAIR, PD, T2 T1, G-E T1	0.71	0,60	**0.99**	-	0.65	0.53	0.64	0.50	0.91
[166]	FLAIR, PD, T2 T1, G-E T1	0.65	-	0.86	**0.97**	-	-	0.76	-	-
[212]	FLAIR, T1	0.55	-	-	-	-	0.79	0.63	-	-
[167]	FLAIR, PD T1, G-E T1	0.76	-	-	-	-	-	**0.82**	-	-
[171]	FLAIR, T1, T2	-	-	-	-	-	-	0.76	0.59	-
Our Framework	FLAIR	**0.88**	**0.77**	0.98	0.88	**0.81**	**0.81**	0.77	**0.72**	**0.27**

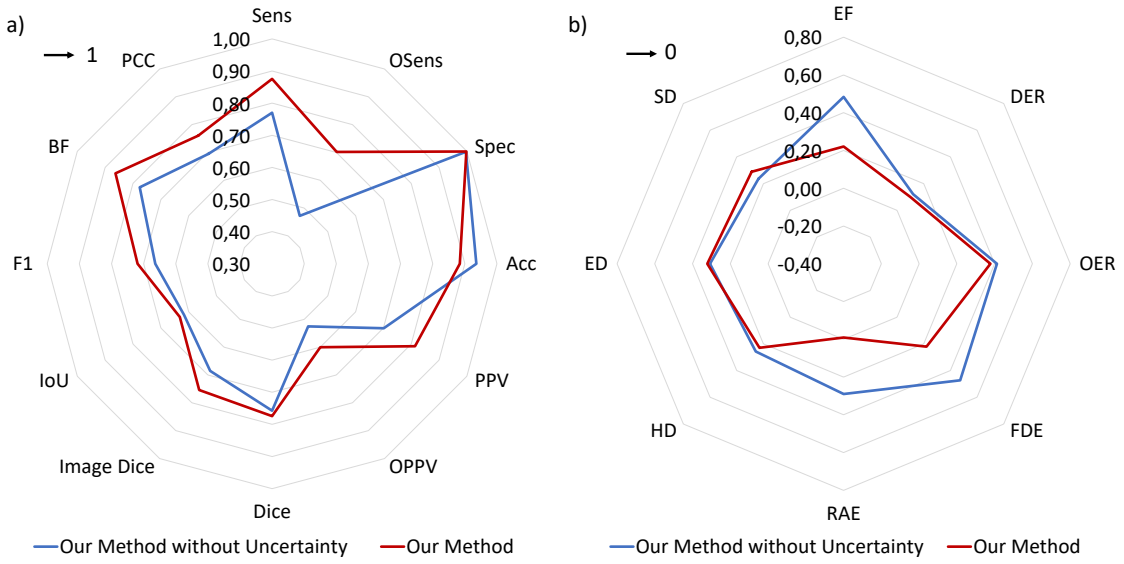

Figure 6.10: The proposed method when trained without and with Uncertainty, compared both on metrics whose ideal value is 1 (a) and for metrics whose ideal value is 0 (b).

Uncertainty (on the binary consensus) as compared to those obtained when trained on the ternary consensus. The ensemble method trained without the Uncertainty outperforms similar automated strategies (team fusion in [124]), though it is still far from humans: the step which places the proposed framework among humans is the inclusion in the pipeline of the class Uncertainty.

This is in line with what reported in [119, 120, 121]: the framework learns better what is surely Lesion, what is surely Background and uses Uncertainty for doubts, as a buffer class. Indeed, the polarized and ambiguous classification of doubtful voxels, sometimes as Lesion and others as Background, disorients any automated

strategy and deviates it from the correct reasoning.

6.4.1 Performances gain through Uncertainty

In Chapter 3 we explained how we constructed the class Uncertainty, in Section 7.3.2 how we used it for the hyper-parameters tuning and how we use it during the training phase. In this section we want to discuss the impact of the Uncertainty on the performances of each CNN and how it influence the different path followed by the Gradient Descent Algorithm.

Figure 6.11 and figure 6.12 show the comparison between each CNN without Uncertainty and the respective CNN with Uncertainty using radar plots in order to better visualize the differences between the metrics values. The figures are organized as follow: 3 rows in which each one shows the plot from the same point of view (Axial, Coronal and Saggital, respectively) and 2 columns that divide the analysis between Lesion and Background, respectively.

The most affected CNN from the introduction of the Uncertainty is the CNN Lesion of the Axial point of view (figure 6.11.a and 6.12.a): Sens, PPV, OPPV, Image Dice, F1, BF, PCC, SD and ED are greatly improved respect to the same CNN without Uncertainty. On the contrary, Acc is worsen respect to the CNN without Uncertainty. The remain metrics are almost the same. From the same Axial view, the CNN Background (figure 6.11.b and figure 6.12.b) gains less advantages from the Uncertainty Class except for the PPV and Sens (that improves). Notable the gain respect to the HD (figure 6.12.b). From a qualitative analysis, both the CNNs with Uncertainty find more true lesions and, as a consequence, also more false lesions. Regarding the latter, the major vote of the method will almost remove all of them.

The same situation occurs in the Coronal View but this time is the CNN Background that gains more advantages from the Uncertainty. Also in this case Sens, OSens, PPV, Image Dice, F1 and BF improve respect to the CNN without the Uncertainty (figure 6.11.d). Also the ED is much better (figure 6.12.d)

The Pixel Sensitivity, the Accuracy and this time also the Lesion Sensitivity show some performance deterioration 6.11.d) (same phenomenon occurs in the Axial view) but in this case the difference are quite small (as shown in figure 6.11.c).

Regarding the Sagittal point of view, the Uncertainty class is not introducing relevant benefits. In fact, the CNN Lesion with Uncertain have little improvements regarding F1 score, BF score and Image Dice score (figure 6.11.e). The others metrics are almost the same, with the exception of the Euclidean Distance that is worse (figure 6.12.e).

To summarize the discussion we can assert that the Axial Lesion CNN and the

Coronal Brain CNN have gained more benefits from the introduction of the class Uncertainty compared to the others CNNs in which the gain is less accentuate. The Sagittal Background CNN is the only CNN in which the performance gain is negligible.

Further analysis, show another interesting phenomena that is closely related to the performances gain just discussed: from the hyper-parameters values of each CNN it is possible to understand which CNN gained more performances from the introduction of the Uncertainty. Table 6.1 shows all the hyper-parameters of each CNN and for the CNNs with Uncertainty also the percentage difference of each hyper-parameter compared with the respective CNN without Uncertainty. It is remarkable that both the CNNs with the best performances gain (Axial Lesion CNN and Coronal Brain CNN both with Uncertainty) have also a gain of Learning Rate of +415,62% and +114,67%, respectively. And it is also remarkable that for all the CNNs with Uncertainty the learning rate increase except for the CNNs for the Saggital view which are the ones with less performances gain. Moreover, the Saggital Brain CNN, that has the worst performance gain, has also the major Learning Rate decrease (-14,92%). This results show that Uncertainty is very effective because increase the learning rate especially for the Axial and Coronal view.

Also the L2 Regularization decreases for all the CNNs with Uncertainty and that means that the model is better capable to generalize and less overfitting.

Finally we analyzed the Lesion Weight and the Background Weight for balancing the cross entropy loss. The Uncertainty class decreases the Lesion Weight and leaves the Background Weight as it is (the differences are quite small).

From these results we can assert that the introduction of the Uncertainty is very effective because in almost all point of view the performances gain are evident.

6.5 Discussion

An automated framework for the identification/segmentation of MS lesions from FLAIR MRI images has been presented. We have demonstrated that traditional CNN architectures, if placed in a context emulating the procedures of human specialists, could effectively behave like a human expert. The strength points of the proposed framework are the following:

1. to train the system both to recognize the lesions as they are and with respect to the environment they are immersed in, thus allowing to incorporate also a sort of meta-information regarding the environment where MS lesions mostly occur;

2. to resemble radiologists in consulting axial slices to discover potential lesions

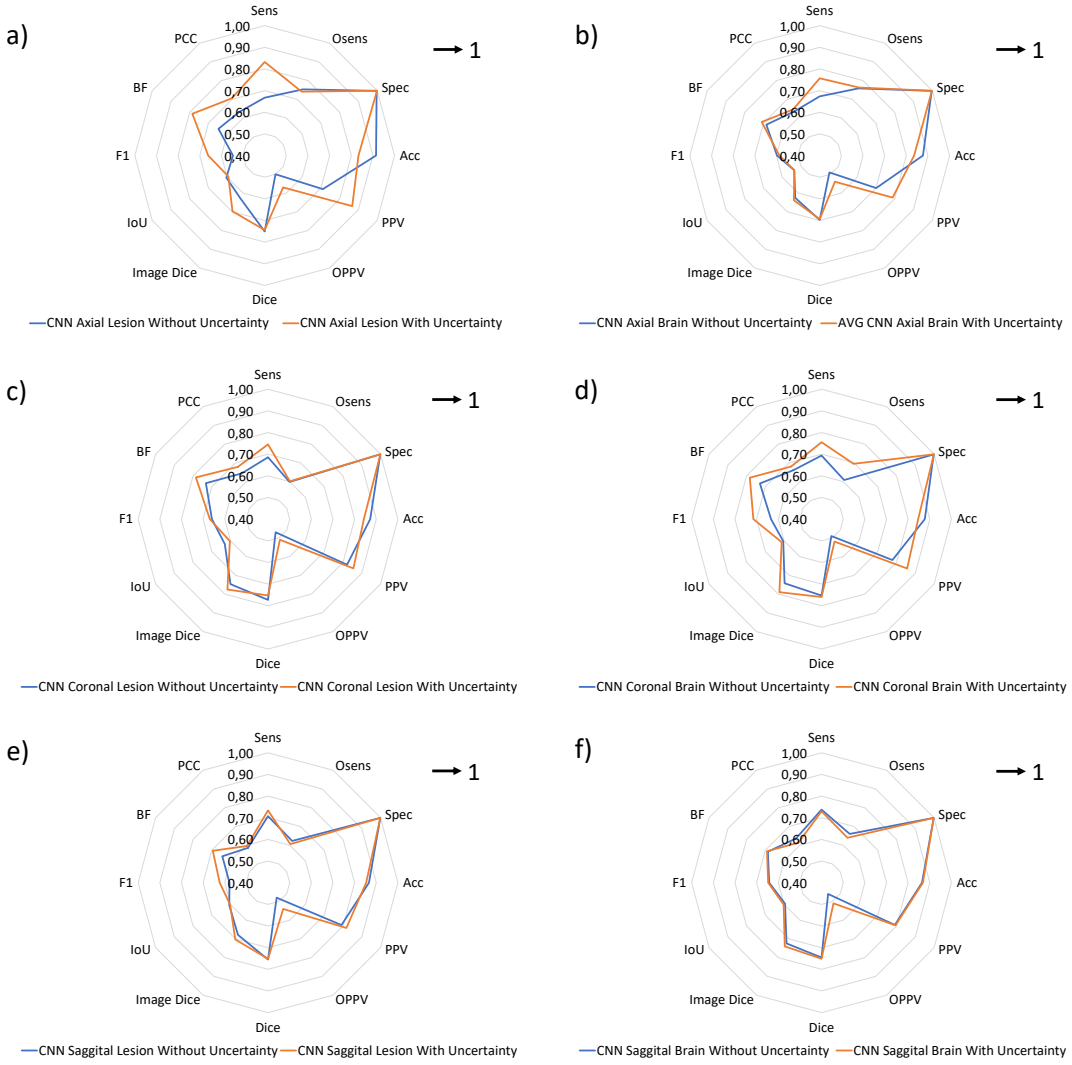

Figure 6.11: Comparison between each CNN with and without the class Uncertainty (blue and orange respectively) for each point of view (Axial, Coronal and Saggital). Higher values of the scores represent best performances.

and to check radial and sagittal slices for confirmation, as well as to maintain 3D continuity to their findings;

3. to use an ensemble classification that usually performs better than its components

4. to use an artificially generated Uncertainty class to improve the performance of an automated strategy and to make it more similar to the human reasoning;

5. to operate just on FLAIR images.

Results have shown that the proposed framework resembles human raters both in behaviour and in performance, when compared with the MSSEG consensus on

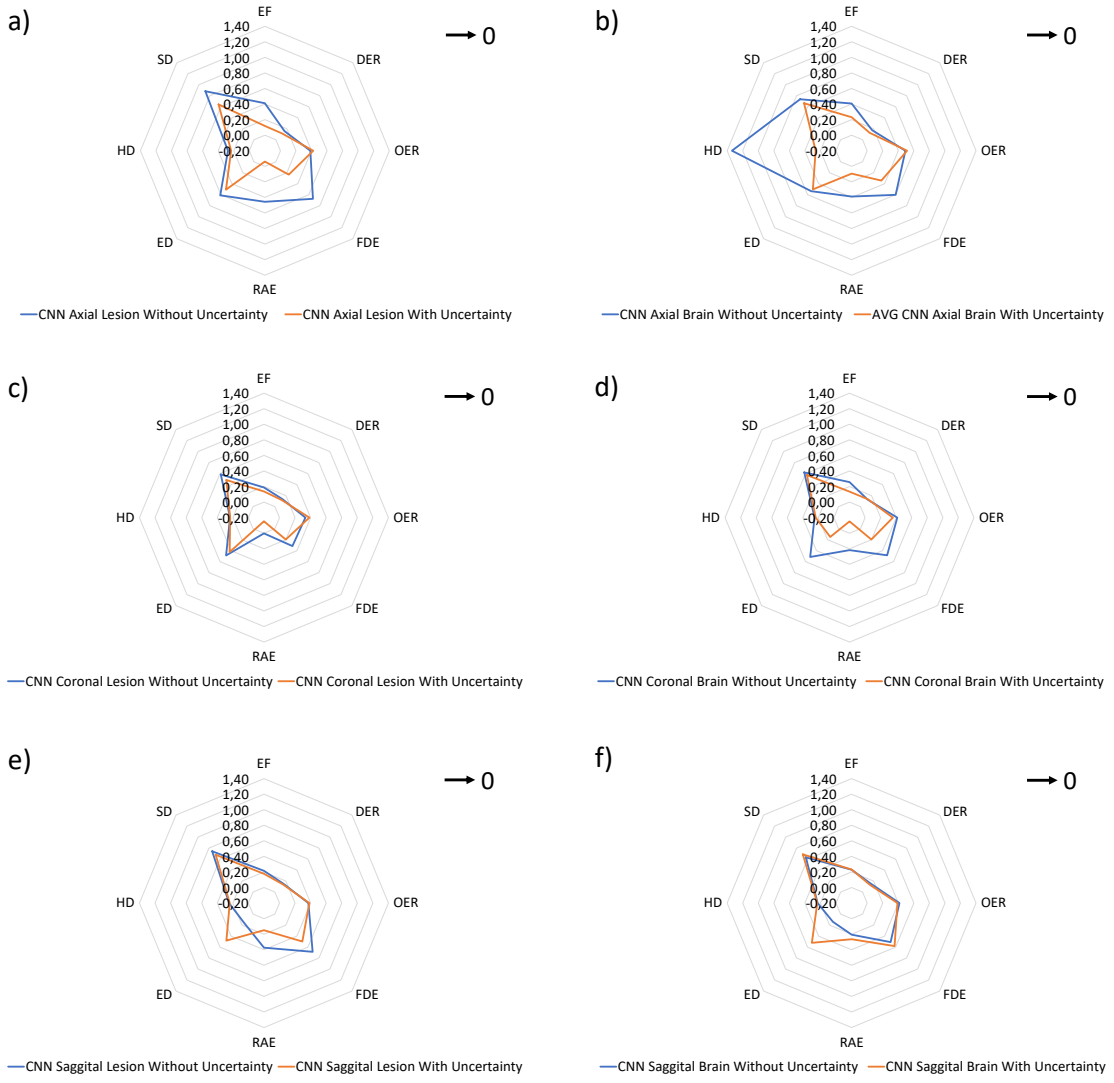

Figure 6.12: Comparison between each CNN with and without the class Uncertainty (blue and orange respectively) for each point of view (Axial, Coronal and Saggital). Lower values of the scores represent best performances

Lesion. Indeed, Wilcoxon statistical test has assessed the framework ability to exhibit a behaviour that is equivalent to, or indistinguishable from, that of a human rater.

Results have also confirmed that the proposed framework outperforms the state of the art strategies which have been trained, validated and tested on the MSSEG data set. Regarding the Uncertainty class, a comparison has been impossible because the human segmentation of Uncertainty is unavailable. However, results have demonstrated that the usage of the Uncertainty during training greatly helps to improve the performance of the framework with respect to not using it. In a recent report [222], the JASON Advisory Group has identified several key recommendations

for advancing computation technology into routine clinical practice. One of them is that new technologies should address a significant clinical need, be practical in use and reduce medical system costs. The demonstration that a better performance is possible by including some concepts (Uncertainty and Ensemble) to enrich traditional CNN architectures, more than continuing to search for even more complex single CNN architectures, goes in the direction of the previous Jason's recommendation.

Chapter 7
Star-Net: a Multi-Branch Convolutional Network for Multiple Source Imaging

In the previous chapter, a framework of CNNs for MS lesions segmentation on a single MRI modality has been presented. Nevertheless, MRI allows to acquire several image modalities in order to extract and combine relevant details from different sources and gather more information. However, the combination of multiple sources of information could not imply an effective gain in the interpretation process either because some sources could contain redundant information (no relevant information is added by some sources) or, worse, could negatively influence the others. In other words, choose which modalities feed into an AI model is not trivial. This chapter presents Star-Net, a multi-branch convolutional network architecture.

It evaluates the contribution of multiple imaging sources in the corresponding layers of different networks, one for each source, weighted according to their contribution. The weights are different in each layer of the network, are case-specific, and are dynamically calculated for each layer.

With this architecture, we reward the active sources while penalizing the inactive ones. This prevents that the irrelevance of the last could dilute the contribution of the former. Star-Net considers the non-linear behaviour of image interpretation, for which the active role of one source in a layer can be reduced/absent in another and can grow-up again in a following layer. When used in the field of multi-modal MRI segmentation, we found that Star-Net can reduce the training convergence with respect to traditional CNN architectures and, more important, it allows to perform case-specific analysis of network activation, to evaluate the effectiveness of each imaging source in the whole interpretation process and, finally, it increases network transparency.

7.1 Introduction

Multiple source imaging is fundamental in several computer vision applications of medical imaging (multimodal imaging and sequences), etc.. The goal of multiple

source imaging is to extract and combine relevant details from different sources to gather more information than from each single source.

However, this is not always trival. There are, in fact, cases in which the combination of multiple sources of information does not imply an effective gain in the interpretation process either because some sources could contain redundant information (no relevant information is added by some sources) or, worse, could negatively influence the others [223].

CNN allow to collect and combine information from multiple source images at any level, at the beginning (early fusion), inside (intermediate fusion) or at the end (late fusion) and the strategy used for fusion is strongly task dependent [224].

However, as it has recently been shown that a CNN ensemble works better than just a single one [225, 137], it seems reasonable to assume that late fusion, being implementable as an ensemble of CNN (one per source), is more effective than early or intermediate fusion, where just one CNN remains after fusion.

Moreover, early fusion can only learn complementary information between modalities leaving out highly non-linear relationships that can be learned only at the higher-level layers [226, 227]. Finally, early fusion is neither optimized with respect to single source contributions nor useful when the contribution of single sources has to be ascertain.

Indeed, many applications would benefit from source contribution separation to:

1. understand what components better contribute to the process and to discover eventual irrelevant sources (simplify the problem);

2. drive the operation process of the CNN and to understand better its internal behaviour (improve transparency);

3. improve simplification by reducing/eliminating irrelevant features;

4. reduce the training set dimension and speed up convergence;

5. improve performance.

In this chapter presents a multi-branch CNN architecture (Fig.7.1), composed by several CNN, one for each imaging source, connected each other through a central weighting normalizer unit whose role is to calculate and redistribute relative activation among the branches. This architecture is designed for explicitly reward the active sources while penalizing the inactive ones.

While the various CNN, that we also call "satellite" for their position, have their feature maps normalized according to reciprocal activation, it is allowed that each progresses separately until the end, when the output are fused before the final output.

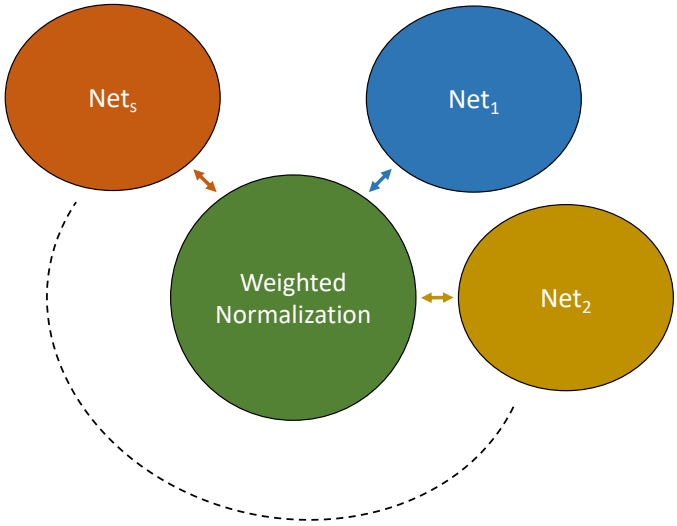

Figure 7.1: Star-Net paradigm: S satellite networks, one for each imaging source, are connected through a central normalizing unit which calculates the activation contribution of each network to the imaging process and applies it to perform a re-balance among networks.

7.2 Related Work

Network architecture has been part of neural network research since their discovery and the recent popularity of neural networks has also revived this research domain. The increasing number of layers in modern architectures amplifies the differences between networks and motivates the exploration of different connectivity patterns.

The new concepts about dense nets [228, 229, 227] open the way to new deep "longitudinal" CNN architectures. A different approach for making networks denser, is to increase the network width. In [230, 231] an inception module is used to concatenate feature-maps produced by filters of different sizes. In [232], a variant of ResNets with wide generalized residual blocks is proposed to improve performance.

However, though the trend to make networks even more connected is strong, strategies for reducing redundancy and to optimize the number of parameters are also very actual, being them finalized at reducing network complexity, power consumption and increasing transparency. Highway networks [233], for example, employ gating mechanisms to regulate shortcut connections. Moreover, dropping modalities [234] and dropping layers [235] have demonstrated the possibility of reducing layers during training while improving also performance. This demonstrates that not all layers (and relative feature maps) may be necessary, that redundancy is frequently present and that, sometimes, redundancy could also be an obstacle for performance.

A lot of work has been conducted on feature reduction through fusion [236]. For example, in [237] similar imaging sources are fused using a heuristic method. In

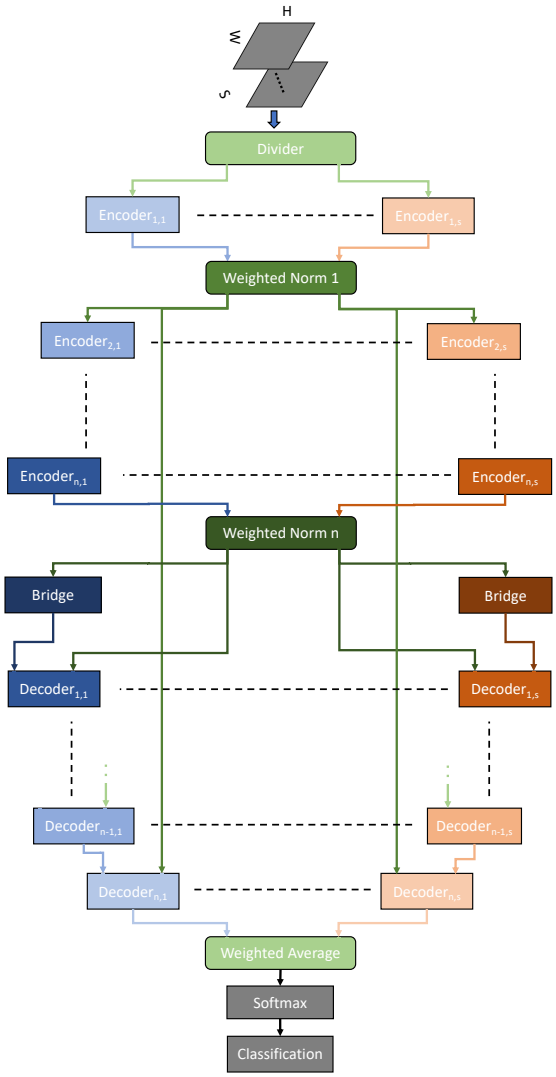

Figure 7.2: Star-Net architecture: in a central unit, the S CNN, are connected through "weighted normalizers" (W_N) at the end of each layers. In each of these W_N modules, the contribution of each modality is upgraded before the process continues in the following layer. The final "Weighted Average" calculate the weights as in W_N but it also merges them in a single feature map, weighted average of the feature maps from the final decoders of the S CNN.

[238], fusion utilizes correlations across representations to combine multiple layers over all modalities in a single one. In [239, 240, 241], multilayer cross connection are proposed aimed at sharing information between modalities of different dimensions. With these approaches, imaging sources are all projected into the same multimodal space, e.g. using concatenation, element-wise products, etc, where a joint learning representation is obtained. In [242] joint representations are opposed to coordinated representations where some constraints between the modalities force the representations to be more complementary. These constraints aim at maximizing the correlation between the multimodal representations, as in [243] and [244].

In an interesting paper, [224], joint representations and coordinated representations are held together by presenting a network architecture, the CentralNet, consisting of a series of satellite deep CNN, to process each modality independently, connected by an additional central network dedicated to the projection of the features coming from different modalities into the same common space. While each satellite network proceeds independently each other, for each layer, the feature maps of the satellite networks are weighted and used as input for the corresponding layer of the central CNN. The central CNN automatically identifies the best levels for fusing information from satellite networks and how these levels should be combined. The output of the central network is used as the output of the process. The weights used in each layer are optimized during trained and fixed.

Though the idea of fusing the contribution of single modalities, layer by layer, is of great inspiration, the previous architecture has some limitations: 1) the weighting parameters, being trainable, are not representative of case-specific details; 2) as a consequence of point 1, the contribution of each modality in each layer remains fixed; 3) the specificity of the output of satellite CNN are excluded from the final decision; 4) as a consequence of point 3, satellite CNN are not influenced by the contribution of the others; 5) it is impossible to evaluate, in a case-specific way, the contribution of each modality to the whole decision process. If solved, this last point would be particularly interesting for two reasons: a) to improve network transparency; b) to ascertain the real contribution of a given modality to the process. Regarding the last point b), in fact, it is important to remark that in many multiple imaging source problems, for example medical imaging [245], the potential contribution of each imaging source to the interpretation process is not always well understood and it is itself a subject of study [115]. Recently proposed methods [246, 247] have been presented to balance inter-modal fusion and intra-modal processing either by dynamically exchanging channels between sub-networks of different modalities [246] or by introducing asymmetric multi-layer fusion. These methods have contributed to improve performance, but the above limitations remain.

Our paper is inspired by the, apparently contrasting, situations of maintaining dense the network and, in the same time, tend to reduce the effects of redundancy, if not necessary. In fact, if from one side, it is necessary to push toward dense structures for transferring information from surface layers to deep ones, from the other side, a smart way is required to reward "active" feature maps and, proportionally, to attenuate the "inactive" ones.

We propose a network architecture whose aim is to enlarge the number of features from multiple-source imaging by stratifying copies of the network (branches), one for each source, in the orthogonal direction with respect to the network depth. In the same time, we create bridges across branches, in the corresponding layers, to re-

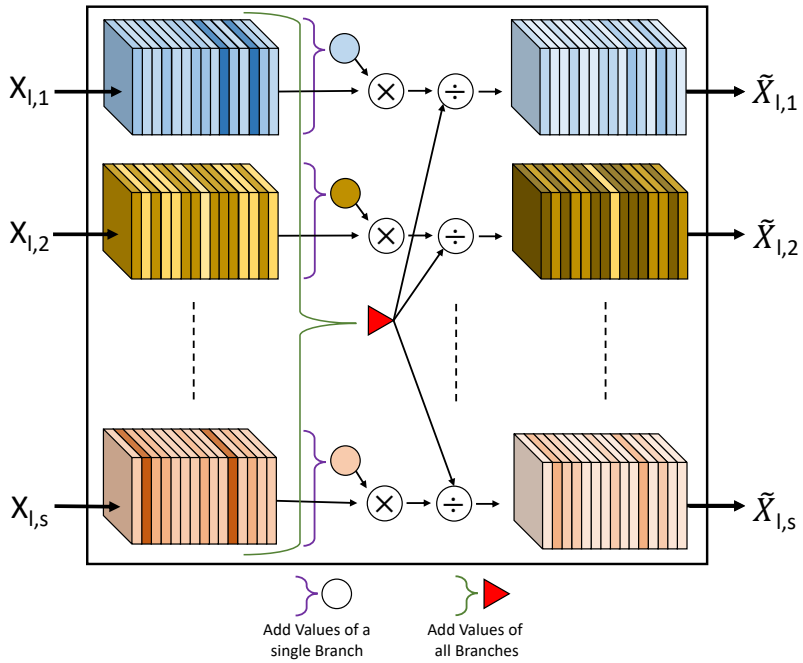

Figure 7.3: Sketch of a W_N calculation/application to the S feature maps of a given layer l of Star-Net. The feature map values of a given branch k are first summed together (circle), then the resulting value is divided by the sum of the sum of the feature map values of all the S feature maps (red triangle) and, finally, the resulting value is used as a multiplier for the current feature map k.

distribute the weights of the feature maps, by increasing those of active feature maps while reducing those of inactive ones, in run time. In this way, dynamic weights are obtained which are case and source specific.

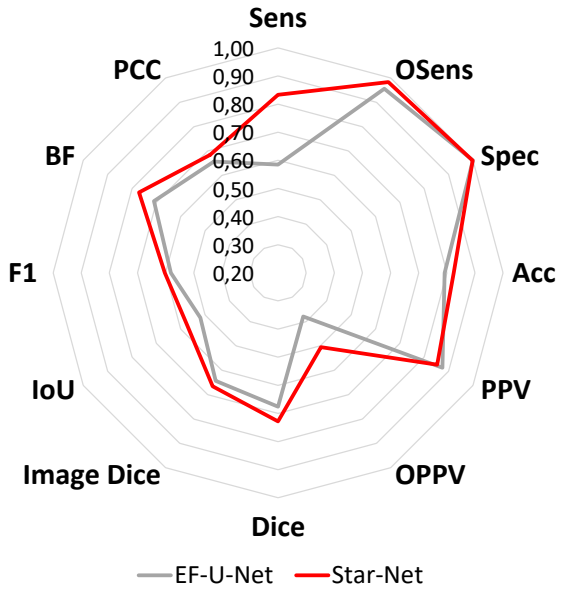

Figure 7.4: Star-Net scores (red line) compared an early fusion U-Net, EF-U-Net, for the same scores (gray line). Score labels indicate (in a clock-wise order from the top): Sensitivity (Sens), Objective Sensitivity (OSens), True Positive Rate (TPR), Accuracy (ACC), Positive Predicted Value (PPV), Objective Positive Predicted Value (OPPV), Correct Detection Ratio (CD), global Dice (Dice), Image-specific Dice (Image Dice), Intersection Over Union (IoU), F1, Boundary F1 (BF) and Pearson Correlation Coefficient (PCC). The scores are defined elsewhere [1, 2, 3]

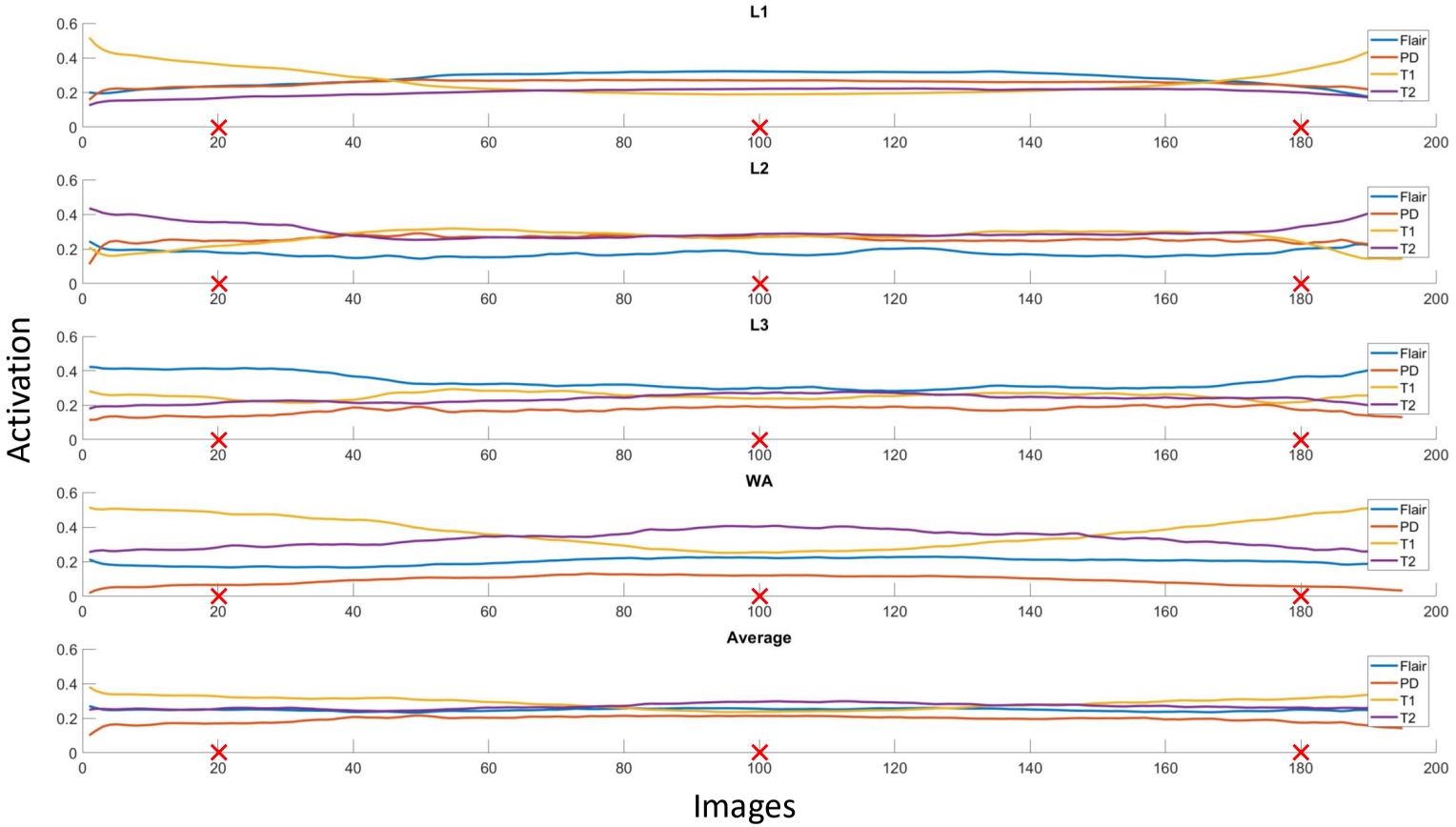

Figure 7.5: Relative activation of each imaging modality for a whole data set of 195 axial MRI images of the brain from the neck (low numbers) to the top of the head (high numbers). The activation values are distinct for images, layers (L1, L2, L3), Weighted Average (WA), and imaging modalities (FLAIR, PD, T1-w and T2-w. The bottom panel indicates the average of corresponding values in the first three panels. Red crosses indicate the positions of the images reported in Fig. 7.6.

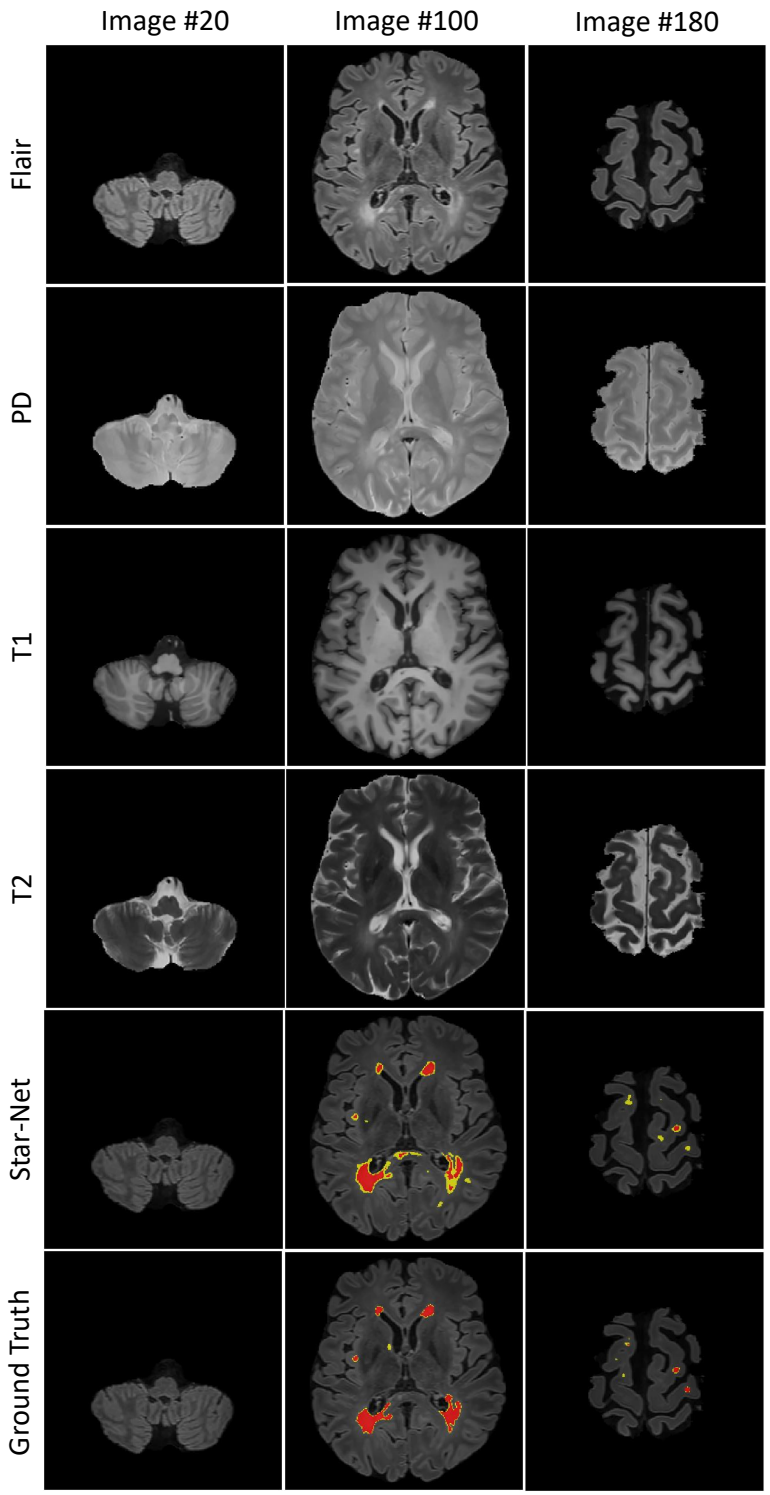

Figure 7.6: Images corresponding to the points indicated with red crosses in Fig.7.5 (in columns). Rows indicate the imaging modalities (rows 1-4), Star-Net segmentation (row 5) and the Ground truth (row 6). In the last two rows, red patches corresponds to MS lesions, yellow patches corresponds to 'uncertain' lesions and the background has no color associated.

7.3 Proposed Architecture

Star-Net, whose longitudinal architecture is shown in Fig.7.2, uses a link of several parallel CNN (branches or paths), one for each source, in which the connecting part is constituted by weighting normalizers (W_N), located at the end of each layer, in which balancing weights are calculated and used to reward the active branches and to penalize, proportionally, inactive ones (the global weight to be distributed is 1). The normalized output of a layer is passed to the following layer and connected to the corresponding decoder through a skip connection. The process of feature maps re-balance, at a given layer l, is sketched in Fig.7.3. In details, let $\mathbf{x}_l$ be the output of the l^{th} layer. Typically, in CNN $\mathbf{x}_l$ is obtained from the output of the previous layer $\mathbf{x}_{l-1}$ through a mapping operator H_l consisting in a convolution followed by regularization and non-linear activation:

$$\mathbf{x}_l = H_l(\mathbf{x}_{l-1}). \tag{7.1}$$

In Star-Net, the feature map of the branch k^{th} of the l^{th} layer, $\mathbf{x}_{k,l}$, is multiplied by a normalizing weight ($W_{N_{k,l}}$), specific for the branch k in the layer l, before it is passed to the $(l+1)^{th}$ layer, as follows:

$$\tilde{\mathbf{x}}_{k,l} = W_{N_{k,l}} \cdot \mathbf{x}_{k,l} \tag{7.2}$$

where

$$W_{N_{k,l}} = \frac{\sum_i x_{k,l,i}}{\sum_{k=1}^{S} \sum_i x_{k,l,i}} \tag{7.3}$$

The i^{th} index extends to all the features composing the current feature map $\mathbf{x}_{k,l}$ (Fig.7.3). The final step in Star-Net is used to fuse the final feature maps of all satellite networks though the following additional weighing average operator:

$$\mathbf{x}_{fin} = \sum_{k=1}^{S} (W_{N_{k,fin}} \cdot \mathbf{x}_{k,fin}) \tag{7.4}$$

where $W_{N_{k,fin}}$ are defined as in Eq.7.3 for the final layer fin and $\mathbf{x}_{fin}$ represents the weighted average of the feature maps allowing to all the S branches and, for this reason, is no more dependent on a specific branch k.

The coefficients $W_{N_{k,l}}$ and $W_{N_{k,fin}}$ are dynamically defined and are not subject to training. For our purposes, this is desirable because, by performing a case-specific weight re-balance, it allows a case-specific optimization of the contributions of each source. In Fig.7.2, U-Net [248] are used as satellite networks, the same necessary for the segmentation experiments described in Section 7.4.

Layer	Flair Branch	PD Branch	T1 Branch	T2 Branch
L1	28% ± 4%	25% ± 2%	27% ± 8%	20% ± 2%
L2	18% ± 2%	26% ± 2%	26% ± 5%	30% ± 5%
L3	34% ± 4%	17% ± 2%	25% ± 2%	24% ± 2%
WA	20% ± 2%	9% ± 3%	38% ± 9%	33% ± 5%
Average	25%	19%	29%	27%

Table 7.1: Relative activation presented in Fig.7.5, averaged along the horizontal directions (images) and represented in percentage (mean and standard deviation are shown). The last row represents the average of the columns (standard deviation is not considered).

7.4 Experiments

Since we were interested in solving a segmentation problem, the CNN used for the branches of Star-Net were 4 U-Net (one for each imaging modality). The Encoder/Decoder depth of the U-Net was 3. The Encoders were composed by a series of Convolutional Layers (filters size 3*3, stride [1, 1]), Batch-Normalization Layers and ReLU Layers (the term H_l in Eq.7.1). The Decoders consisted of a series of Convolutional Layers (filters size 3*3, stride [1, 1]), Transposed Convolutional Layers (filters size 2*2, stride [2, 2]) Batch-Normalization Layers and ReLU Layers. The size of the data structure representing a single instance was 256*256*4 (256*256 were the dimensions of each image and 4 was the number of modalities). Since each satellite network had 7.7 millions of parameters, Star-Net resulted in 30.8 millions of parameters. Adam Optimizer and mini batch size of 4 were used for training, with the Cross Entropy as loss function.

The experiments were performed with the Deep Learning Toolbox of Matlab 2021a on a computer with the following characteristics: 1 CPU AMD Ryzen 5 3600x, 2 GPUs Nvidia 2080 Super, RAM 32 GB, 1 TB of SSD and 2 TB of HDD.

A Bayesian approach was used to optimize the hyper-parameter setting (Starting Learning Rate, L2-Regularization and Class Balancing) [217]. Before final training, 30 small training attempts (10 epochs each) were executed using the hyper-parameters suggested by the Bayesian function. Then, the best hyper-parameter configuration was chosen and used for the final training (50 epochs).

With the proposed architecture, our principal focus is not to obtain the best performance but to demonstrate its feasibility to exploit the case-dependent relationships between sources in different layers and to gather information regarding how and how much each source contributes to the process. As far as the previous characteristics are really innovative and peculiar of Star-Net, the comparison with

other existing architectures was impossible on these. However, to give an idea of Star-Net performance and to demonstrate it is trainable and suitable for the proposed segmentation task, we show a comparison, both in terms of training speed and performance, with an U-Net in which early fusion was adopted between the four imaging modalities (we call it EF-U-Net to distinguish it from the U-Net used in Star-Net but it has the same structure except the input).

Regarding training, Star-Net converged faster than EF-U-Net. In fact, the Bayesian method found the best hyper-parameter configuration after 15^{th} attempts while EF-U-Net found its best after 27^{th} attempts. Further, in the final training, Star-Net reached the maximum accuracy after 34 epochs while EF-U-Net after 42 epochs. This confirms the hypothesis that Star-Net, bein structured to maintain the specificity of the modalities while re-defining their weight according to a global vision, allows to converge faster than EF-U-Net.

After training, Star-Net and EF-U-net were tested on the final data set and results were compared with the ground-truth through a set of metric scores [3] whose values, averaged for the whole test data set, are summarized in Fig. 7.4. Though the essence of these results was to demonstrate that Star-Net is trainable and coherent with the assigned task, they also demonstrate that Star-Net is effective.

In order to evaluate the image-specific contribution of each modality to the task, we calculated the activation contribution of each modality in each layer of the satellite networks and also in the weighted averaging layer: the values for 195 images of a complete 3D MRI examination are shown in Fig.7.5. Results are very interesting: the contribution of each modality is strongly image-dependent and layer-dependent. In particular: for the extreme images (close to the neck and to the top of the head), T1-w is stronger than the others in layer 1 and in the weighted average layer, T2-w outperforms the others in Layer 2, and FLAIR is the best in layer 3; for the central images (the middle of the brain), FLAIR is the best in layers 1 and 3, in layer 2 the main contribution is shared among T1-w, T2-w and PD, and T2-w overcomes the others in the weighted average layer. This confirms that the run-time calculation of the weighting parameters is important, being the role of different modalities greatly dependent on the images. By using the proposed run time strategy, case-specific variations can be captured and exploited and, most important, it allows to "read" inside the network and to better understand what is occurring in its stages (network transparency).

To give a visual idea of Star-Net behavior, Fig.7.6 reports three sample images of the data set in Fig.7.5: the 20^{th}, the 100^{th} and the 180^{th} of the series. The Star-Net segmentation is in good agreement with the ground truth, except some false positive in the yellow class ("uncertainty").

Data of Fig. 7.5 have been also averaged along the image direction and reported,

in percentage, in Table 7.1. These results confirm that the effects of each modality is layer-dependent and that a given modality can be very active in one layer, turned off in another and active again in a following. A further interesting aspect is that some modalities (FLAIR, T1-w and T2-w) are in general more active than PD. This information is important to ascertain the effectiveness of the contribution of a modality with respect to the others. In particular for medical imaging, the possibility of eliminating some modality could greatly contribute to reduce the duration of the examination, to improve patient comfort and to optimize the usage of the imaging equipment [115].

7.5 Discussion

Among the proposed CNN architectures for dealing the problem of multiple-source computer vision, Star-Net represents an innovative evolution because it allows to modulate the single contributions of a series of satellite networks, each associated to a specific source, through a central unit composed by layer-dependent weighted normalizers. The contributions of the satellite CNN are weighted averaged, in a late fusion modality, before an unique and final response is collected.

Star-Net peculiarities are that: its central unit is not a CNN, it has not trainable parameters and it is light (the number of weights used inside it are $S \cdot L$, where S is the number of imaging sources and L is the number of layers composing each branch); the parameters are calculated run-time and they are case-specific; it allows to evaluate the contribution of each source in each layer and to study how and how much the values of the central weights change as a function of the analyzed images. This last aspect, in particular, could be used both to evaluate the contribution of each single source to the process and to improve network readability and transparency.

The reported experiments confirm that the performance are in line with classical architectures used for early fusion multiple-source imaging, or slightly better, with a good improvement in training speed. This improvement is mostly due to the fact that Star-Net maintains separated the specific contributions of all modalities, while re-modulating their weights according to the case-specific inter-modality interaction.

Part IV

Conclusions and Future Developments

Nowadays, AI is being demonstrated as one of the most promising solutions to many real-life problems, including some important regarding medicine, and for this reason many efforts are made to use it as much as possible.

Nonetheless, the trend is to use AI as a "*black box*" because the word intelligence mislead the developers making them thinking that with the minimum effort AI can be capable to do everything.

We can put a lot of efforts to develop even more complex AI models, containing more and more parameters, but this does not ensure a gain and, more important, it does not prevent against failures.

To achieve the objective of being able to use AI extensively and safely for routine activities, including clinical ones, it is necessary to take into account several aspects: the quality and the numerosity of the data sets used for training the models, the procedures for data labelling, the creation of consistent ground truths, noise removal, data normalization, the treatment of uncertainty and the procedures and metrics to evaluate the quality of the results.

In this thesis, we strongly highlighted these concepts and we discussed them regarding to neurological MI, one of the most important fields in medicine. In particular, we focused on the development of AI-based computer-aided analysis and interpretation of EEG signals and MRI since we were interested to investigate the brain functioning and diseases.

Our approach consisted to firstly look at the data and then think about AI model design. Moreover, we studied the workflow of physicians and we tried to mimic it through AI. Where it was needed, we extended this workflow to make it compatible with AI models.

Regarding EEG signal, the first step before the analysis was artifact removal. The proposed AI method presented in Chapter 1 is based on what physicians do: convert the EEG signal in a more representative form (in this case the Topoplots) and analyze it. The obtained results with the proposed AI-based model reached an accuracy of about 98%. This made it possible to apply the method in a challenging scenario: the analysis of weak and noisy EEG signals of infants affected by 3M syndrome, presented in Chapter 2. Physicians who supervised the AI-based analysis discovered, from the artificial model, new ways to read the obtained results and they were capable to transfer these findings to clinical practice.

Regarding MR images, the development of AI models was even more challenging. We focused on MR images of patients affected by multiple sclerosis and this allowed us to deeply investigate the behavior of radiologists and consequentially to understand several barriers that limit the performance of AI models. Firstly, in Chapter 3 we discussed the challenge issue regarding the ground truth generation, being radiologists often at odds on how to segment the data (inter-raters variabil-

ity). To mitigate this problem, we proposed a method for modeling the uncertainty. Afterwards, in Chapter 4 we studied the effects of MRI variability on the training process of AI models. Physicians, even with a lot of effort and by recurring to their long experience, can adapt themselves to this variability but for AI models it could be impractical. For this reason, we proposed some guidelines and we applied them for the training of a benchmark framework of AI models for the segmentation of MS lesions. The results showed a consistent performance improvement of 15%.

To further mitigate the variability of the amplitude of MRI, in Chapter 5 we proposed a Local Contrast Normalization algorithm. Results showed that the amplitude shift was almost completely corrected. In this way we reached to make AI models independent of MRI scanners and imaging variability.

In Chapter 6 we designed our own framework for MS lesions segmentation. Compared to the one used in Chapter 4, which analyzed patches of the images, our framework analyzed the whole image in order to include into the model the information regarding the position of the lesions with respect to the other brain structures and, hence, to reduce outliers. To this aim, we also introduced the information regarding the three views of the 3D MRI model: axial, coronal and sagittal views. In this way, our model was capable to mimic the radiologist behaviour. Moreover, we made the proposed AI model capable to deal with uncertainty. The obtained results showed that our method overcame several state of art methods and, to a deep comparison of the proposed framework with the MRI labelled by 7 radiologists, they proved that our framework was indistinguishable from human raters (it passed the Touring test). The other important implicit result we obtained was the demonstration that just a single imaging modality was sufficient to perform identification/segmentation of MS lesions from MRI. In order to check this results, in Chapter 7 we proposed a new CNN architecture capable to contain the whole set of imaging modalities and to study the effects that each one has on the MS lesion segmentation. The results conformed that the combination of multiple MRI modalities could not imply an effective gain in the interpretation process either because some modalities could contain redundant information or, worse, could negatively influence the others. The information we gathered from the proposed AI-based model could be extremely useful to radiologists for redefining the MRI acquisition protocols used for MS and, more important, to reduce the stress for the patients, by lowering the acquisition time.

Another aspect that we want to underline is that all the proposed architectures had one common design strategy: the use of an ensemble of AI models.

This approach has demonstrated several technical advantages:

- it makes easy to deal with small models

- it makes the proposed frameworks easy to be scaled.

In conclusion, the results demonstrated that it is possible to develop high performances AI frameworks for supporting doctors and researchers with augmented information. This thesis underlined the importance of understanding the context of the problem first, in order to simplify the design of robust AI models that fit it well.

Future developments will regard the generalization of the proposed strategies to cope with different diseases or with different applications of MI. Particular attention will be paid to the optimization of the models and to understand the processes underlying their behaviour. To this aim, specific strategies for checking the deep structures of the proposed architectures will be studied. In this way, besides model optimization, it would be possible to get the functional relationships among the features generating from the model and use them to improve human knowledge (a sort of inverse transfer learning).

Part V

Bibliography

Bibliography

[1] *MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, (Athènes, Greece), 2016.

[2] A. Danelakis, T. Theoharis, and D. A. Verganelakis, "Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging," *Computerized Medical Imaging and Graphics*, vol. 70, pp. 83 – 100, 2018.

[3] G. Placidi, L. Cinque, F. Mignosi, and M. Polsinelli, "Multiple sclerosis lesions identification/segmentation in magnetic resonance imaging using ensemble cnn and uncertainty classification," *arXiv preprint arXiv:2108.11791*, 2021.

[4] M. Laal, "Innovation process in medical imaging," *Procedia-Social and Behavioral Sciences*, vol. 81, pp. 60–64, 2013.

[5] G. Mele, C. Cavaliere, V. Alfano, M. Orsini, M. Salvatore, and M. Aiello, "Simultaneous eeg-fmri for functional neurological assessment," *Frontiers in Neurology*, p. 848, 2019.

[6] G. Placidi, L. Cinque, and M. Polsinelli, "A fast and scalable framework for automated artifact recognition from eeg signals represented in scalp topographies of independent components," *Computers in Biology and Medicine*, vol. 132, p. 104347, 2021.

[7] H.-I. Suk and S.-W. Lee, "A novel bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 286–299, 2012.

[8] H. Cecotti and A. Graser, "Convolutional neural networks for p300 detection with application to brain-computer interfaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 433–445, 2010.

[9] G. Placidi, D. Avola, A. Petracca, F. Sgallari, and M. Spezialetti, "Basis for the implementation of an EEG-based single-trial binary brain computer interface

through the disgust produced by remembering unpleasant odors," *Neurocomputing*, vol. 160, pp. 308–318, jul 2015.

[10] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal—state-of-the-art and guidelines," *Journal of Neural Engineering*, vol. 12, p. 031001, jun 2015.

[11] B. Noureddin, P. D. Lawrence, and G. E. Birch, "Online removal of eye movement and blink eeg artifacts using a high-speed eye tracker," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2103–2110, 2011.

[12] C. A. Joyce, I. F. Gorodnitsky, and M. Kutas, "Automatic removal of eye movement and blink artifacts from EEG data using blind component separation," *Psychophysiology*, vol. 41, pp. 313–325, mar 2004.

[13] P.-F. Lin, M.-T. Lo, J. Tsao, Y.-C. Chang, C. Lin, and Y.-L. Ho, "Correlations between the signal complexity of cerebral and cardiac electrical activity: A multiscale entropy analysis," *PLoS ONE*, vol. 9, no. 2, 2014.

[14] H. Shibasaki and J. Rothwell, "Emg-eeg correlation. the international federation of clinical neurophysiology.," *Electroencephalography and clinical neurophysiology. Supplement*, vol. 52, pp. 269–274, 1999.

[15] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, pp. 229–240, feb 2011.

[16] A. Delorme, T. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis," *NeuroImage*, vol. 34, pp. 1443–1449, feb 2007.

[17] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. Mckeown, V. Iragui, and T. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.

[18] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for eeg-based human intention recognition," *IEEE transactions on cybernetics*, 2019.

[19] T. Radüntz, J. Scouten, O. Hochmuth, and B. Meffert, "EEG artifact elimination by extraction of ICA-component features using image processing algorithms," *Journal of Neuroscience Methods*, vol. 243, pp. 84–93, mar 2015.

[20] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, pp. 9–21, mar 2004.

[21] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data," *Computational Intelligence and Neuroscience*, vol. 2011, pp. 1–9, 2011.

[22] P. Garg, E. Davenport, G. Murugesan, B. Wagner, C. Whitlow, J. Maldjian, and A. Montillo, "Using convolutional neural networks to automatically detect eye-blink artifacts in magnetoencephalography without resorting to electrooculography," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 374–381, Springer, 2017.

[23] P. Croce, F. Zappasodi, L. Marzetti, A. Merla, V. Pizzella, and A. M. Chiarelli, "Deep convolutional neural networks for feature-less automatic classification of independent components in multi-channel electrophysiological brain recordings," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 8, pp. 2372–2380, 2018.

[24] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018.

[25] L. Pion-Tonachini, K. Kreutz-Delgado, and S. Makeig, "Iclabel: An automated electroencephalographic independent component classifier, dataset, and website," *NeuroImage*, vol. 198, pp. 181 – 197, 2019.

[26] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[27] M. K. Islam, A. Rastegarnia, and Z. Yang, "Methods for artifact detection and removal from scalp EEG: A review," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 46, pp. 287–305, nov 2016.

[28] M. M. N. Mannan, M. A. Kamran, and M. Y. Jeong, "Identification and Removal of Physiological Artifacts From Electroencephalogram Signals: A Review," *IEEE Access*, vol. 6, pp. 30630–30652, 2018.

[29] M. M. C. van den Berg-Lenssen, J. A. M. van Gisbergen, and B. W. Jervis, "Comparison of two methods for correcting ocular artefacts in EEGs," *Medical and Biological Engineering and Computing*, vol. 32, pp. 501–511, sep 1994.

[30] J. Minguillon, M. A. Lopez-Gordo, and F. Pelayo, "Trends in EEG-BCI for daily-life: Requirements for artifact removal," *Biomedical Signal Processing and Control*, vol. 31, pp. 407–418, jan 2017.

[31] S. Romero, M. A. Mañanas, and M. J. Barbanoj, "A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: A simulation case," *Computers in Biology and Medicine*, vol. 38, pp. 348–360, mar 2008.

[32] K. T. Sweeney, T. E. Ward, and S. F. McLoone, "Artifact Removal in Physiological Signals—Practices and Possibilities," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, pp. 488–500, may 2012.

[33] A. Acharyya, P. N. Jadhav, V. Bono, K. Maharatna, and G. R. Naik, "Low-complexity hardware design methodology for reliable and automated removal of ocular and muscular artifact from eeg," *Computer methods and programs in biomedicine*, vol. 158, pp. 123–133, 2018.

[34] N. E. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, C. Tung, and H. Liu, "he empirical mode decomposition and hilbert spectrum for nonlinear and nonstationary time series analysis," *Proceedings of the Royal Society A*, vol. 545, no. 1971, pp. 903–995, 1998.

[35] N. Rehman and D. P. Mandic, "Multivariate empirical mode decomposition," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 466, no. 2117, pp. 1291–1302, 2010.

[36] X. Chen, A. Liu, J. Chiang, Z. J. Wang, M. J. McKeown, and R. K. Ward, "Removing muscle artifacts from eeg data: Multichannel or single-channel techniques?," *IEEE Sensors Journal*, vol. 16, no. 7, pp. 1986–1997, 2015.

[37] P. Berg and M. Scherg, "Dipole modelling of eye activity and its application to the removal of eye artefacts from the EEG and MEG," *Clinical Physics and Physiological Measurement*, vol. 12, pp. 49–54, jan 1991.

[38] T. D. Lagerlund, F. W. Sharbrough, and N. E. Busacker, "Spatial Filtering of Multichannel Electroencephalographic Recordings Through Principal Component Analysis by Singular Value Decomposition," *Journal of Clinical Neurophysiology*, vol. 14, pp. 73–82, jan 1997.

[39] Q. Zhao, B. Hu, Y. Shi, Y. Li, P. Moore, M. Sun, and H. Peng, "Automatic identification and removal of ocular artifacts in eeg - improved adaptive predictor filtering for portable applications," *IEEE Transactions on Nanobioscience*, vol. 13, no. 2, pp. 109–117, 2014.

[40] W. De Clercq, A. Vergult, B. Vanrumste, W. Van Paesschen, and S. Van Huffel, "Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram," *IEEE transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2583–2587, 2006.

[41] X. Chen, X. Xu, A. Liu, M. J. McKeown, and Z. J. Wang, "The use of multivariate emd and cca for denoising muscle artifacts from few-channel eeg recordings," *IEEE transactions on instrumentation and measurement*, vol. 67, no. 2, pp. 359–370, 2017.

[42] X. Chen, H. Peng, F. Yu, and K. Wang, "Independent vector analysis applied to remove muscle artifacts in eeg data," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 7, pp. 1770–1779, 2017.

[43] X. Chen, A. Liu, Q. Chen, Y. Liu, L. Zou, and M. J. McKeown, "Simultaneous ocular and muscle artifact removal from eeg data by exploiting diverse statistics," *Computers in biology and medicine*, vol. 88, pp. 1–10, 2017.

[44] X. Chen, Q. Liu, W. Tao, L. Li, S. Lee, A. Liu, Q. Chen, J. Cheng, M. J. McKeown, and Z. J. Wang, "Remae: User-friendly toolbox for removing muscle artifacts from eeg," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 2105–2119, 2019.

[45] E. Urrestarazu, J. Iriarte, M. Alegre, M. Valencia, C. Viteri, and J. Artieda, "Independent Component Analysis Removing Artifacts in Ictal Recordings," *Epilepsia*, vol. 45, pp. 1071–1078, sep 2004.

[46] R. Vigario, J. Sarela, V. Jousmiki, M. Hamalainen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE Transactions on Biomedical Engineering*, vol. 47, pp. 589–593, may 2000.

[47] T. Radüntz, J. Scouten, O. Hochmuth, and B. Meffert, "Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features," *Journal of Neural Engineering*, vol. 14, p. 046004, aug 2017.

[48] C. James and O. Gibson, "Temporally constrained ica: an application to artifact rejection in electromagnetic brain signal analysis," *IEEE Transactions on Biomedical Engineering*, vol. 50, pp. 1108–1116, sep 2003.

[49] I. Winkler, S. Brandl, F. Horn, E. Waldburger, C. Allefeld, and M. Tangermann, "Robust artifactual independent component classification for BCI practitioners," *Journal of Neural Engineering*, vol. 11, p. 035013, jun 2014.

[50] I. Winkler, S. Brandl, F. Horn, E. Waldburger, C. Allefeld, and M. Tangermann, "Robust artifactual independent component classification for BCI practitioners," *Journal of Neural Engineering*, vol. 11, p. 035013, jun 2014.

[51] A. A. M. Al-Saffar, H. Tao, and M. A. Talab, "Review of deep convolution neural network in image classification," in *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pp. 26–31, IEEE, 2017.

[52] L. Song, J. Liu, B. Qian, M. Sun, K. Yang, M. Sun, and S. Abbas, "A deep multimodal cnn for multi-instance multi-label image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 6025–6038, 2018.

[53] E. Askari, S. K. Setarehdan, A. Sheikhani, M. R. Mohammadi, and M. Teshnehlab, "Modeling the connections of brain regions in children with autism using cellular neural networks and electroencephalography analysis," *Artificial intelligence in medicine*, vol. 89, pp. 40–50, 2018.

[54] X. Tang, T. Wang, Y. Du, and Y. Dai, "Motor imagery eeg recognition with knn-based smooth auto-encoder," *Artificial Intelligence in Medicine*, vol. 101, p. 101747, 2019.

[55] X. Gao, X. Yan, P. Gao, X. Gao, and S. Zhang, "Automatic detection of epileptic seizure based on approximate entropy, recurrence quantification analysis and convolutional neural networks," *Artificial Intelligence in Medicine*, vol. 102, p. 101711, 2020.

[56] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[57] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[58] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[59] M. R. Nuwer, G. Comi, R. Emerson, A. Fuglsang-Frederiksen, J.-M. Guérit, H. Hinrichs, A. Ikeda, F. J. C. Luccas, and P. Rappelsburger, "Ifcn standards for digital recording of clinical eeg," *Electroencephalography and clinical Neurophysiology*, vol. 106, no. 3, pp. 259–261, 1998.

[60] S. Leske and S. S. Dalal, "Reducing power line noise in eeg and meg data via spectrum interpolation," *NeuroImage*, vol. 189, pp. 763 – 776, 2019.

[61] Z. Liu, L. Li, H. Xu, and H. Li, "A method for recognition and classification for hybrid signals based on deep convolutional neural network," in *2018 International Conference on Electronics Technology (ICET)*, pp. 325–330, IEEE, 2018.

[62] C. Ito, X. Cao, M. Shuzo, and E. Maeda, "Application of cnn for human activity recognition with fft spectrogram of acceleration and gyro sensors," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pp. 1503–1510, ACM, 2018.

[63] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Transactions on Affective Computing*, 2019.

[64] K. R. Mopuri, U. Garg, and R. V. Babu, "Cnn fixations: an unraveling approach to visualize the discriminative image regions," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2116–2125, 2018.

[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[66] S.-H. Hsu, T. R. Mullen, T.-P. Jung, and G. Cauwenberghs, "Real-Time Adaptive EEG Source Separation Using Online Recursive Independent Component Analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, pp. 309–319, mar 2016.

[67] S. Jalilpour, S. H. Sardouie, and A. Mijani, "A novel hybrid bci speller based on rsvp and ssvep paradigm," *Computer Methods and Programs in Biomedicine*, vol. 187, p. 105326, 2020.

[68] X. Chai, Z. Zhang, K. Guan, T. Zhang, J. Xu, and H. Niu, "Effects of fatigue on steady state motion visual evoked potentials: Optimised stimulus parameters for a zoom motion-based brain-computer interface," *Computer Methods and Programs in Biomedicine*, p. 105650, 2020.

[69] S. K. Khare and V. Bajaj, "A facile and flexible motor imagery classification using electroencephalogram signals," *Computer Methods and Programs in Biomedicine*, p. 105722, 2020.

[70] S. Invitto, A. Grasso, D. D. Lofrumento, V. Ciccarese, A. Paladini, P. Paladini, R. Marulli, V. De Pascalis, M. Polsinelli, and G. Placidi, "Chemosensory event-related potentials and power spectrum could be a possible biomarker in 3m syndrome infants?," *Brain sciences*, vol. 10, no. 4, p. 201, 2020.

[71] I. Marik, O. Marikova, M. Kuklik, D. Zemkova, and K. Kozlowski, "3-m syndrome in two sisters," *Journal of paediatrics and child health*, vol. 38, no. 4, pp. 419–422, 2002.

[72] D. Hanson, A. Stevens, P. G. Murray, G. C. Black, and P. E. Clayton, "Identifying biological pathways that underlie primordial short stature using network analysis," *Journal of molecular endocrinology*, vol. 52, no. 3, pp. 333–344, 2014.

[73] P. G. Murray, D. Hanson, T. Coulson, A. Stevens, A. Whatmore, R. L. Poole, D. J. Mackay, G. Black, and P. E. Clayton, "3-m syndrome: a growth disorder associated with igf2 silencing," *Endocrine connections*, vol. 2, no. 4, pp. 225–235, 2013.

[74] P. E. Clayton, D. Hanson, L. Magee, P. G. Murray, E. Saunders, S. N. Abu-Amero, G. E. Moore, and G. C. Black, "Exploring the spectrum of 3-m syndrome, a primordial short stature disorder of disrupted ubiquitination," *Clinical endocrinology*, vol. 77, no. 3, pp. 335–342, 2012.

[75] D. Hanson, P. Murray, T. Coulson, A. Sud, A. Omokanye, E. Stratta, F. Sakhinia, C. Bonshek, L. Wilson, E. Wakeling, *et al.*, "Mutations in cul7, obsl1 and ccdc8 in 3-m syndrome lead to disordered growth factor signalling," *Journal of molecular endocrinology*, vol. 49, no. 3, p. 267, 2012.

[76] A. Sarikas, X. Xu, L. J. Field, and Z.-Q. Pan, "The cullin7 e3 ubiquitin ligase: a novel player in growth control," *Cell cycle*, vol. 7, no. 20, pp. 3154–3161, 2008.

[77] F. Derakhshan and C. Toth, "Insulin and the brain," *Current diabetes reviews*, vol. 9, no. 2, pp. 102–116, 2013.

[78] M.-C. Lacroix, K. Badonnel, N. Meunier, F. Tan, C. S.-L. Poupon, D. Durieux, R. Monnerie, C. Baly, P. Congar, R. Salesse, *et al.*, "Expression of insulin system in the olfactory epithelium: first approaches to its role and regulation," *Journal of neuroendocrinology*, vol. 20, no. 10, pp. 1176–1190, 2008.

[79] N. Litterman, Y. Ikeuchi, G. Gallardo, B. C. O'Connell, M. E. Sowa, S. P. Gygi, J. W. Harper, and A. Bonni, "An obsl1-cul7fbxw8 ubiquitin ligase signaling mechanism regulates golgi morphology and dendrite patterning," *PLoS biology*, vol. 9, no. 5, p. e1001060, 2011.

[80] H. Varendi and R. Porter, "Breast odour as the only maternal stimulus elicits crawling towards the odour source," *Acta Paediatrica*, vol. 90, no. 4, pp. 372–375, 2001.

[81] M. Delaunay-El Allam, R. Soussignan, B. Patris, L. Marlier, and B. Schaal, "Long-lasting memory for an odor acquired at the mother's breast," *Developmental science*, vol. 13, no. 6, pp. 849–863, 2010.

[82] P. Kuhn, D. Astruc, J. Messer, and L. Marlier, "Exploring the olfactory environment of premature newborns: a french survey of health care and cleaning products used in neonatal units," *Acta Paediatrica*, vol. 100, no. 3, pp. 334–339, 2011.

[83] G. Gauthaman, L. Jayachandran, and K. Prabhakar, "Olfactory reflexes in newborn infants," *The Indian Journal of Pediatrics*, vol. 51, no. 4, pp. 397–399, 1984.

[84] V. A. Schriever, M. Góis-Eanes, B. Schuster, C. Huart, and T. Hummel, "Olfactory event-related potentials in infants," *The Journal of Pediatrics*, vol. 165, no. 2, pp. 372–375, 2014.

[85] G. Kobal and T. Hummel, "Olfactory (chemosensory) event-related potentials," *Toxicology and Industrial Health*, vol. 10, no. 4-5, pp. 587–596, 1994.

[86] S. Invitto and A. Grasso, "Chemosensory perception: a review on electrophysiological methods in "cognitive neuro-olfactometry"," *Chemosensors*, vol. 7, no. 3, p. 45, 2019.

[87] B. M. Pause, B. Sojka, K. Krauel, and R. Ferstl, "The nature of the late positive complex within the olfactory event-related potential (oerp)," *Psychophysiology*, vol. 33, no. 4, pp. 376–384, 1996.

[88] B. M. Pause and K. Krauel, "Chemosensory event-related potentials (cserp) as a key to the psychology of odors," *International Journal of Psychophysiology*, vol. 36, no. 2, pp. 105–122, 2000.

[89] A. Brodal, "The hypocampus and the sense of smell; a review.," *Brain: a journal of neurology*, 1947.

[90] F. Mormann, H. Osterhage, R. G. Andrzejak, B. Weber, G. Fernández, J. Fell, C. E. Elger, and K. Lehnertz, "Independent delta/theta rhythms in the human hippocampus and entorhinal cortex," *Frontiers in Human Neuroscience*, vol. 2, p. 3, 2008.

[91] M. L. Nunes, R. L. Khan, I. Gomes Filho, L. Booij, and J. C. da Costa, "Maturational changes of neonatal electroencephalogram: a comparison between intra uterine and extra uterine development," *Clinical Neurophysiology*, vol. 125, no. 6, pp. 1121–1128, 2014.

[92] S. Invitto, G. Montagna, S. Capone, and P. Siciliano, "Method and system for measuring physiological parameters of a subject undergoing an olfactory stmulation," May 11 2017. US Patent App. 15/412,265.

[93] R. Soussignan, B. Schaal, and L. Marlier, "Olfactory alliesthesia in human neonates: prandial state and stimulus familiarity modulate facial and autonomic responses to milk odors," *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, vol. 35, no. 1, pp. 3–14, 1999.

[94] R. Soussignan, B. Schaal, L. Marlier, and T. Jiang, "Facial and autonomic responses to biological and artificial olfactory stimuli in human neonates: re-examining early hedonic discrimination of odors," *Physiology & Behavior*, vol. 62, no. 4, pp. 745–758, 1997.

[95] M. Sirous, N. Sinning, T. R. Schneider, U. Friese, J. Lorenz, and A. K. Engel, "Chemosensory event-related potentials in response to nasal propylene glycol stimulation," *Frontiers in human neuroscience*, vol. 13, p. 99, 2019.

[96] B. Stuck, T. Moutsis, U. Bingel, and J. Sommer, "Chemosensory stimulation during sleep–arousal responses to gustatory stimulation," *Neuroscience*, vol. 322, pp. 326–332, 2016.

[97] C. Heiser, J. Baja, F. Lenz, J. Sommer, K. Hörmann, R. Herr, and B. Stuck, "Effects of an artificial smoke on arousals during human sleep," *Chemosensory Perception*, vol. 5, no. 3, pp. 274–279, 2012.

[98] G. Badre, M. Wloszczynski, and I. Croy, "Neural activation of putative sleep-wake affecting and relaxing promoting odors," *Sleep Medicine*, vol. 64, p. S19, 2019.

[99] S. J. Luck, *An introduction to the event-related potential technique*. MIT press, 2014.

[100] B. M. Pause, B. Sojka, K. Krauel, G. Fehm-Wolfsdorf, and R. Ferstl, "Olfactory information processing during the course of the menstrual cycle," *Biological psychology*, vol. 44, no. 1, pp. 31–54, 1996.

[101] S. Invitto, G. Piraino, A. Mignozzi, S. Capone, G. Montagna, P. A. Siciliano, A. Mazzatenta, G. Rocco, I. D. Feudis, G. F. Trotta, *et al.*, "Smell and meaning: an oerp study," in *Multidisciplinary Approaches to Neural Computing*, pp. 289–300, Springer, 2018.

[102] G. Goelman and R. Dan, "Multiple-region directed functional connectivity based on phase delays," *Human brain mapping*, vol. 38, no. 3, pp. 1374–1386, 2017.

[103] A. Piarulli, A. Zaccaro, M. Laurino, D. Menicucci, A. De Vito, L. Bruschini, S. Berrettini, M. Bergamasco, S. Laureys, and A. Gemignani, "Ultra-slow mechanical stimulation of olfactory epithelium modulates consciousness by slowing cerebral rhythms in humans," *Scientific reports*, vol. 8, no. 1, pp. 1–17, 2018.

[104] K. Sowndhararajan and S. Kim, "Influence of fragrances on human psychophysiological activity: With special reference to human electroencephalographic response," *Scientia pharmaceutica*, vol. 84, no. 4, pp. 724–751, 2016.

[105] Ž. Špiclin, F. Pernuš, B. Likar, T. Jerman, and D. Ravnik, "Dataset variability leverages white-matter lesion segmentation performance with convolutional neural network," in *Medical Imaging 2018: Image Processing* (E. D. Angelini and B. A. Landman, eds.), SPIE, mar 2018.

[106] M. Bosc, F. Heitz, J.-P. Armspach, I. Namer, D. Gounot, and L. Rumbach, "Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution," *NeuroImage*, vol. 20, pp. 643–656, oct 2003.

[107] G. Placidi, "Mri: Essentials for innovative technologies (1st ed.)." `https://doi.org/10.1201/b11868`, 2012.

[108] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[109] A. Akhondi-Asl, L. Hoyte, M. E. Lockhart, and S. K. Warfield, "A logarithmic opinion pool based STAPLE algorithm for the fusion of segmentations with associated reliability weights," *IEEE Transactions on Medical Imaging*, vol. 33, pp. 1997–2009, Oct. 2014.

[110] L. Steinman, "Multiple sclerosis: a coordinated immunological attack against myelin in the central nervous system," *Cell*, vol. 85, no. 3, pp. 299–302, 1996.

[111] M. Filippi, P. Preziosa, B. L. Banwell, F. Barkhof, O. Ciccarelli, N. De Stefano, J. J. Geurts, F. Paul, D. S. Reich, A. T. Toosy, *et al.*, "Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines," *Brain*, vol. 142, no. 7, pp. 1858–1875, 2019.

[112] M. Brant-Zawadzki, G. D. Gillan, and W. R. Nitz, "MP RAGE: a three-dimensional, t1-weighted, gradient-echo sequence–initial experience in the brain.," *Radiology*, vol. 182, pp. 769–775, mar 1992.

[113] F. Nelson, A. Poonawalla, P. Hou, J. Wolinsky, and P. Narayana, "3d MPRAGE improves classification of cortical lesions in multiple sclerosis," *Multiple Sclerosis Journal*, vol. 14, pp. 1214–1219, oct 2008.

[114] T. Kober, C. Granziera, D. Ribes, P. Browaeys, M. Schluep, R. Meuli, R. Frackowiak, R. Gruetter, and G. Krueger, "MP2rage multiple sclerosis magnetic resonance imaging at 3 t," *Investigative Radiology*, vol. 47, pp. 346–352, jun 2012.

[115] M. Filippi, P. Preziosa, B. L. Banwell, F. Barkhof, O. Ciccarelli, N. De Stefano, J. J. G. Geurts, F. Paul, D. S. Reich, A. T. Toosy, A. Traboulsee, M. P. Wattjes, T. A. Yousry, A. Gass, C. Lubetzki, B. G. Weinshenker, and M. A. Rocca, "Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines," *Brain*, vol. 142, pp. 1858–1875, 06 2019.

[116] O. Vincent, C. Gros, and J. Cohen-Adad, "Impact of individual rater style on deep learning uncertainty in medical imaging segmentation," *ArXiv*, vol. abs/2105.02197, 2021.

[117] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, OpenReview.net, 2017.

[118] J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019.

[119] E. Kats, J. Goldberger, and H. Greenspan, "Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, apr 2019.

[120] E. Kats, J. Goldberger, and H. Greenspan, "A soft STAPLE algorithm combined with anatomical knowledge," in *Lecture Notes in Computer Science*, pp. 510–517, Springer International Publishing, 2019.

[121] C. Gros, A. Lemay, and J. Cohen-Adad, "Softseg: Advantages of soft versus binary training for image segmentation," *Medical Image Analysis*, vol. 71, p. 102038, 2021.

[122] D. S. Wack, M. G. Dwyer, N. Bergsland, C. D. Perri, L. Ranza, S. Hussein, D. Ramasamy, G. Poloni, and R. Zivadinov, "Improved assessment of multiple sclerosis lesion segmentation agreement via detection and outline error estimates," *BMC Medical Imaging*, vol. 12, July 2012.

[123] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?.," in *Bmvc*, vol. 27, pp. 10–5244, 2013.

[124] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Améli, J.-C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi, A. Mahbod, C. Wang, R. McKinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Lladó, M. M. Santos, W. P. Santos, A. G. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F. J. Vera-Olmos, N. Malpica, C. Guttmann, S. Vukusic, G. Edan, M. Dojat, M. Styner, S. K. Warfield, F. Cotton, and C. Barillot, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific Reports*, vol. 8, sep 2018.

[125] A. Carass, S. Roy, A. Gherman, J. C. Reinhold, A. Jesson, T. Arbel, O. Maier, H. Handels, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, D. L. Pham, C. M. Crainiceanu, P. A. Calabresi, J. L. Prince, W. R. G. Roncal, R. T. Shinohara, and I. Oguz, "Evaluating white matter lesion segmentations with refined sørensen-dice analysis," *Scientific Reports*, vol. 10, May 2020.

[126] G. Placidi, L. Cinque, and M. Polsinelli, "Guidelines for effective automatic multiple sclerosis lesion segmentation by magnetic resonance imaging.," in *ICPRAM*, pp. 570–577, 2020.

[127] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Medical image analysis*, vol. 17, no. 1, pp. 1–18, 2013.

[128] A. Danelakis, T. Theoharis, and D. A. Verganelakis, "Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging," *Computerized Medical Imaging and Graphics*, vol. 70, pp. 83–100, 2018.

[129] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Ameli, J.-C. Ferré, *et al.*, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific reports*, vol. 8, no. 1, p. 13650, 2018.

[130] D. Franchi, P. Gallo, L. Marsili, and G. Placidi, "A shape-based segmentation algorithm for x-ray digital subtraction angiography images," *Computer methods and programs in biomedicine*, vol. 94, no. 3, pp. 267–278, 2009.

[131] A. Maurizi, D. Franchi, and G. Placidi, "An optimized java based software package for biomedical images and volumes processing," pp. 219–222, 2009.

[132] Y. Yoo, T. Brosch, A. Traboulsee, D. K. Li, and R. Tam, "Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation," in *International Workshop on Machine Learning in Medical Imaging*, pp. 117–124, Springer, 2014.

[133] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi, "Longitudinal multiple sclerosis lesion segmentation using 3d convolutional neural networks," *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pp. 1–2, 2015.

[134] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[135] A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman, *et al.*, "Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria," *The Lancet Neurology*, vol. 17, no. 2, pp. 162–173, 2018.

[136] S. Roy, J. A. Butman, D. S. Reich, P. A. Calabresi, and D. L. Pham, "Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks," *arXiv preprint arXiv:1803.09172*, 2018.

[137] G. Placidi, L. Cinque, M. Polsinelli, A. Splendiani, and E. Tommasino, "Automatic framework for multiple sclerosis follow-up by magnetic resonance imaging for reducing contrast agents," in *International Conference on Image Analysis and Processing*, pp. 367–378, Springer, 2019.

[138] G. Placidi and M. Polsinelli, "Local contrast normalization to improve preprocessing in mri of the brain," in *International Conference on Bioengineering and Biomedical Signal and Image Processing*, pp. 255–266, Springer, 2021.

[139] G. Placidi, *MRI: essentials for innovative technologies.* CRC Press, 2012.

[140] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, C. M. Crainiceanu, *et al.*, "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, 2014.

[141] J. Ford, N. Dogan, L. Young, and F. Yang, "Quantitative radiomics: impact of pulse sequence parameter selection on mri-based textural features of the brain," *Contrast media & molecular imaging*, vol. 2018, 2018.

[142] A. Carré, G. Klausner, M. Edjlali, M. Lerousseau, J. Briend-Diop, R. Sun, S. Ammari, S. Reuzé, E. A. Andres, T. Estienne, S. Niyoteka, E. Battistella, M. Vakalopoulou, F. Dhermain, N. Paragios, E. Deutsch, C. Oppenheim, J. Pallud, and C. Robert, "Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics," *Scientific Reports*, vol. 10, jul 2020.

[143] M. Alecci, J. Brivati, G. Placidi, L. Testa, D. Lurie, and A. Sotgiu, "A submicrosecond resonator and receiver system for pulsed magnetic resonance with large samples," 1998.

[144] S. Di Giuseppe, G. Placidi, J. Brivati, M. Alecci, and A. Sotgiu, "Pulsed epr imaging: image reconstruction using selective acquisition sequences," *Physics in Medicine & Biology*, vol. 44, no. 6, p. N137, 1999.

[145] S. Di Giuseppe, G. Placidi, and A. Sotgiu, "New experimental apparatus for multimodal resonance imaging: initial epri and nmri experimental results," *Physics in Medicine & Biology*, vol. 46, no. 4, p. 1003, 2001.

[146] G. Placidi, M. Alecci, and A. Sotgiu, "First imaging results obtained with a multimodal apparatus combining low-field (35.7 mt) mri and pulsed epri," *Physics in Medicine & Biology*, vol. 47, no. 10, p. N127, 2002.

[147] M. Alfonsetti, V. Clementi, S. Iotti, G. Placidi, R. Lodi, B. Barbiroli, A. Sotgiu, and M. Alecci, "Versatile coil design and positioning of transverse-field rf surface coils for clinical 1.5-t mri applications," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 18, no. 2, pp. 69–75, 2005.

[148] G. Placidi, M. Alecci, and A. Sotgiu, "Angular space-domain interpolation for filtered back projection applied to regular and adaptively measured projections," *Journal of Magnetic Resonance, Series B*, vol. 110, no. 1, pp. 75–79, 1996.

[149] G. Placidi, M. Alecci, S. Colacicchi, and A. Sotgiu, "Fourier reconstruction as a valid alternative to filtered back projection in iterative applications: implementation of fourier spectral spatial epr imaging," *Journal of Magnetic Resonance*, vol. 134, no. 2, pp. 280–286, 1998.

[150] G. Placidi, M. Alecci, and A. Sotgiu, "ω-space adaptive acquisition technique for magnetic resonance imaging from projections," *Journal of Magnetic Resonance*, vol. 143, no. 1, pp. 197–207, 2000.

[151] G. Placidi, M. Alecci, and A. Sotgiu, "Post-processing noise removal algorithm for magnetic resonance imaging based on edge detection and wavelet analysis," *Physics in Medicine and Biology*, vol. 48, pp. 1987–1995, jun 2003.

[152] G. Placidi and A. Sotgiu, "A novel algorithm for the reduction of undersampling artefacts in magnetic resonance images," *Magnetic resonance imaging*, vol. 22, no. 9, pp. 1279–1287, 2004.

[153] M. Dadar, V. S. Fonov, and D. L. Collins, "A comparison of publicly available linear MRI stereotaxic registration techniques," *NeuroImage*, vol. 174, pp. 191–200, July 2018.

[154] X. Zhang, Y. Feng, W. Chen, X. Li, A. V. Faria, Q. Feng, and S. Mori, "Linear registration of brain MRI using knowledge-based multiple intermediator libraries," *Frontiers in Neuroscience*, vol. 13, Sept. 2019.

[155] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3d intersubject registration of mr volumetric data in standardized talairach space," *Journal of Computer Assisted Tomography*, vol. 18, no. 2, 1994.

[156] O. Dietrich, J. G. Raya, and M. F. Reiser, "Magnetic resonance noise measurements and signal-quantization effects at very low noise levels," *Magnetic Resonance in Medicine*, vol. 60, pp. 1477–1487, nov 2008.

[157] P. Bao and L. Zhang, "Noise reduction for magnetic resonance images via adaptive multiscale products thresholding," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 1089–1099, sep 2003.

[158] P. Coupe, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot, "An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 27, pp. 425–441, apr 2008.

[159] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "FSL," *NeuroImage*, vol. 62, pp. 782–790, aug 2012.

[160] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 1310–1320, jun 2010.

[161] C. Li, J. C. Gore, and C. Davatzikos, "Multiplicative intrinsic component optimization (MICO) for MRI bias field estimation and tissue segmentation," *Magnetic Resonance Imaging*, vol. 32, pp. 913–923, sep 2014.

[162] S. A. Villar, S. Torcida, and G. G. Acosta, "Median filtering: A new insight," *Journal of Mathematical Imaging and Vision*, vol. 58, pp. 130–146, dec 2016.

[163] G. Placidi, L. Cinque, M. Polsinelli, A. Splendiani, and E. Tommasino, "Automatic framework for multiple sclerosis follow-up by magnetic resonance imaging for reducing contrast agents," in *International Conference on Image Analysis and Processing*, pp. 367–378, Springer, 2019.

[164] A. Kaur, L. Kaur, and A. Singh, "State-of-the-art segmentation techniques and future directions for multiple sclerosis brain lesions," *Archives of Computational Methods in Engineering*, feb 2020.

[165] H. Zhang and I. Oguz, "Multiple sclerosis lesion segmentation - a survey of supervised cnn-based methods," *ArXiv*, vol. abs/2012.08317, 2020.

[166] P. Ghosal, P. K. C. Prasad, and D. Nandi, "A light weighted deep learning framework for multiple sclerosis lesion segmentation," in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pp. 526–531, IEEE, 2019.

[167] A. Alijamaat, A. NikravanShalmani, and P. Bayat, "Multiple sclerosis lesion segmentation from brain mri using u-net based on wavelet pooling," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2021.

[168] M. Salem, S. Valverde, M. Cabezas, D. Pareto, A. Oliver, J. Salvi, À. Rovira, and X. Lladó, "A fully convolutional neural network for new t2-w lesion detection in multiple sclerosis," *NeuroImage: Clinical*, vol. 25, p. 102149, 2020.

[169] G. Placidi, L. Cinque, M. Polsinelli, A. Splendiani, and E. Tommasino, "Automatic framework for multiple sclerosis follow-up by magnetic resonance imaging for reducing contrast agents," in *Image Analysis and Processing – ICIAP 2019* (E. Ricci, S. Rota Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, eds.), (Cham), pp. 367–378, Springer International Publishing, 2019.

[170] N. Gessert, J. Krüger, R. Opfer, A.-C. Ostwaldt, P. Manogaran, H. H. Kitzler, S. Schippling, and A. Schlaefer, "Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs," *Computerized Medical Imaging and Graphics*, vol. 84, p. 101772, Sept. 2020.

[171] R. McKinley, R. Wepfer, F. Aschwanden, L. Grunder, R. Muri, C. Rummel, R. Verma, C. Weisstanner, M. Reyes, A. Salmen, *et al.*, "Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural networks," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[172] A. Traboulsee, J. Simon, L. Stone, E. Fisher, D. Jones, A. Malhotra, S. Newsome, J. Oh, D. Reich, N. Richert, *et al.*, "Revised recommendations of the consortium of ms centers task force for a standardized mri protocol and clinical guidelines for the diagnosis and follow-up of multiple sclerosis," *American Journal of Neuroradiology*, vol. 37, no. 3, pp. 394–401, 2016.

[173] U. Macar, E. N. Karthik, C. Gros, A. Lemay, and J. Cohen-Adad, "Team neuropoly: Description of the pipelines for the miccai 2021 ms new lesions segmentation challenge," *ArXiv*, vol. abs/2109.05409, 2021.

[174] V. Sundaresan, G. Zamboni, P. M. Rothwell, M. Jenkinson, and L. Griffanti, "Triplanar ensemble u-net model for white matter hyperintensities segmentation on MR images," vol. 73, p. 102184, Oct. 2021.

[175] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical Image Analysis*, vol. 59, p. 101557, 2020.

[176] R. Alizadehsani, M. Roshanzamir, S. Hussain, A. Khosravi, A. Koohestani, M. H. Zangooei, M. Abdar, A. Beykikhoshk, A. Shoeibi, A. Zare, M. Panahiazar, S. Nahavandi, D. Srinivasan, A. F. Atiya, and U. R. Acharya, "Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991–2020)," *Annals of Operations Research*, Mar. 2021.

[177] C. Liu, X. Zeng, K. Liang, Y. Yu, and C. Ye, "Improved brain lesion segmentation with anatomical priors from healthy subjects," pp. 186–195, Springer International Publishing, 2021.

[178] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, dec 2017.

[179] D. Mortazavi, A. Z. Kouzani, and H. Soltanian-Zadeh, "Segmentation of multiple sclerosis lesions in MR images: a review," *Neuroradiology*, vol. 54, pp. 299–320, may 2011.

[180] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and À. Rovira, "Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches," *Information Sciences*, vol. 186, pp. 164–185, mar 2012.

[181] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Medical Image Analysis*, vol. 17, pp. 1–18, Jan. 2013.

[182] H. M. R. Afzal, S. Luo, S. Ramadan, and J. Lechner-Scott, "The emerging role of artificial intelligence in multiple sclerosis imaging," *Multiple Sclerosis Journal*, p. 135245852096629, Oct. 2020.

[183] M. Zurita, C. Montalba, T. Labbé, J. P. Cruz, J. D. da Rocha, C. Tejos, E. Ciampi, C. Cárcamo, R. Sitaram, and S. Uribe, "Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data," *NeuroImage: Clinical*, vol. 20, pp. 724–730, 2018.

[184] N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, and D. L. Pham, "A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions," *NeuroImage*, vol. 49, pp. 1524–1535, jan 2010.

[185] E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images," *NeuroImage*, vol. 57, pp. 378–390, July 2011.

[186] Z. Karimaghaloo, M. Shah, S. J. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain mri using conditional random fields," *IEEE Transactions on Medical Imaging*, vol. 31, no. 6, pp. 1181–1194, 2012.

[187] M. Cabezas, A. Oliver, S. Valverde, B. Beltran, J. Freixenet, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó, "BOOST: A supervised approach for multiple sclerosis lesion segmentation," *Journal of Neuroscience Methods*, vol. 237, pp. 108–117, nov 2014.

[188] N. Guizard, P. Coupé, V. S. Fonov, J. V. Manjón, D. L. Arnold, and D. L. Collins, "Rotation-invariant multi-contrast non-local means for MS lesion segmentation," *NeuroImage: Clinical*, vol. 8, pp. 376–389, 2015.

[189] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, M. J. Cardoso, N. Cawley, O. Ciccarelli, C. A. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. O. Iheme, D. Unay, S. Jain, D. M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P. A. Calabresi, C. M. Crainiceanu, L. M. Ellingsen, D. S. Reich, J. L. Prince, and D. L. Pham, "Longitudinal multiple sclerosis lesion segmentation: Resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, mar 2017.

[190] Y. Yoo, T. Brosch, A. Traboulsee, D. Li, and R. Tam, "Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation," in *MLMI*, 2014.

[191] L. S. Aït-Ali, S. Prima, P. Hellier, B. Carsin, G. Edan, and C. Barillot, "STREM: A robust multidimensional parametric method to segment MS lesions in MRI," in *Lecture Notes in Computer Science*, pp. 409–416, Springer Berlin Heidelberg, 2005.

[192] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, B. Hemmer, and M. Mühlau, "An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis," *NeuroImage*, vol. 59, pp. 3774–3783, feb 2012.

[193] E. Roura, A. Oliver, M. Cabezas, S. Valverde, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, and X. Lladó, "A toolbox for multiple sclerosis lesion segmentation," *Neuroradiology*, vol. 57, pp. 1031–1043, jul 2015.

[194] R. Harmouche, L. Collins, D. Arnold, S. Francis, and T. Arbel, "Bayesian ms lesion classification modeling regional and local spatial information," in *18th International Conference on Pattern Recognition (ICPR06)*, IEEE, 2006.

[195] M. Strumia, F. R. Schmidt, C. Anastasopoulos, C. Granziera, G. Krueger, and T. Brox, "White matter ms-lesion segmentation using a geometric brain model," *IEEE Transactions on Medical Imaging*, vol. 35, no. 7, pp. 1636–1646, 2016.

[196] T. Brosch, L. Y. W. Tang, Y. Yoo, D. K. B. Li, A. Traboulsee, and R. Tam, "Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.

[197] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[198] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[199] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 424–432, Springer International Publishing, 2016.

[200] F. Milletari, N. Navab, and S. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[201] Y. Chen, B. Shi, Z. Wang, P. Zhang, C. D. Smith, and J. Liu, "Hippocampus segmentation through multi-view ensemble ConvNets," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, IEEE, apr 2017.

[202] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi, "Longitudinal multiple sclerosis lesion segmentation using 3d convolutional neural networks," in *In Proc. International symposium on biomedical imaging, New York.*, 2015.

[203] S. Roy, A. Carass, J. L. Prince, and D. L. Pham, "Subject specific sparse dictionary learning for atlas based brain mri segmentation," in *Machine Learning in Medical Imaging* (G. Wu, D. Zhang, and L. Zhou, eds.), (Cham), pp. 248–255, Springer International Publishing, 2014.

[204] S. Aslani, M. Dayan, L. Storelli, M. Filippi, V. Murino, M. Rocca, and D. Sona, "Multi-branch convolutional neural network for multiple sclerosis lesion segmentation," *NeuroImage*, vol. 196, pp. 1–15, 2019.

[205] G. Kang, B. Hou, Y. Ma, F. Labeau, Z. Su, *et al.*, "Acu-net: A 3d attention context u-net for multiple sclerosis lesion segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1384–1388, IEEE, 2020.

[206] Y. S. Vang, Y. Cao, P. D. Chang, D. S. Chow, A. U. Brandt, F. Paul, M. Scheel, and X. Xie, "Synergynet: a fusion framework for multiple sclerosis brain mri segmentation with local refinement," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 131–135, IEEE, 2020.

[207] H. Zhang, J. Zhang, R. Wang, Q. Zhang, S. A. Gauthier, P. Spincemaille, T. D. Nguyen, and Y. Wang, "Geometric loss for deep multiple sclerosis lesion segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 24–28, IEEE, 2021.

[208] S. Roy, J. A. Butman, D. S. Reich, P. A. Calabresi, and D. L. Pham, "Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks," 2018.

[209] S. Valverde, M. Cabezas, E. Roura, S. González-Villà, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, A. Oliver, and X. Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, jul 2017.

[210] S. R. Hashemi, S. S. Mohseni Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection," *IEEE Access*, vol. 7, pp. 1721–1735, 2019.

[211] F. L. Rosa, M. J. Fartaria, T. Kober, J. Richiardi, C. Granziera, J.-P. Thiran, and M. B. Cuadra, "Shallow vs deep learning architectures for white matter lesion segmentation in the early stages of multiple sclerosis," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 142–151, Springer International Publishing, 2019.

[212] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, and X. Lladó, "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks," *NeuroImage: Clinical*, vol. 21, p. 101638, 2019.

[213] S. Santurkar, D. Tsipras, A. Ilyas, and A. Mądry, "How does batch normalization help optimization?," in *Proceedings of the 32nd International Conference*

on Neural Information Processing Systems, NIPS'18, (Red Hook, NY, USA), p. 2488–2498, Curran Associates Inc., 2018.

[214] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. 10, pp. 281–305, 2012.

[215] E. Hazan, A. Klivans, and Y. Yuan, "Hyperparameter optimization: a spectral approach," in *International Conference on Learning Representations*, 2018.

[216] X. Zhang, X. Chen, L. Yao, C. Ge, and M. Dong, "Deep neural network hyperparameter optimization with orthogonal array tuning," in *Neural Information Processing* (T. Gedeon, K. W. Wong, and M. Lee, eds.), (Cham), pp. 287–295, Springer International Publishing, 2019.

[217] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, (Red Hook, NY, USA), p. 2951–2959, Curran Associates Inc., 2012.

[218] L. I. Kuncheva and C. J. Whitaker *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[219] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: a survey and categorisation," *Information Fusion*, vol. 6, pp. 5–20, Mar. 2005.

[220] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional neural network committees for handwritten character classification," in *2011 International Conference on Document Analysis and Recognition*, IEEE, Sept. 2011.

[221] H. Geijer and M. Geijer, "Added value of double reading in diagnostic radiology, a systematic review," *Insights into Imaging*, vol. 9, pp. 287–301, Mar. 2018.

[222] JASON, "Artificial intelligence for health and heath care.," *JSR-17-Task-002*, 2017.

[223] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[224] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: a multilayer approach for multimodal fusion," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

[225] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ACM, Nov. 2017.

[226] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[227] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1116–1126, 2018.

[228] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 2261–2269, IEEE Computer Society, jul 2017.

[229] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," in *ICLR*, 2017.

[230] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[231] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

[232] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," 2016.

[233] S. Yang and D. Ramanan, "Multi-scale recognition with dag-cnns," in *2015 IEEE International Conference on Computer Vision (ICCV)*, (Los Alamitos, CA, USA), pp. 1215–1223, IEEE Computer Society, dec 2015.

[234] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: Adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 38, pp. 1692–1706, aug 2016.

[235] M. Zhang and Y. He, "Accelerating training of transformer-based language models with progressive layer dropping," in *Advances in Neural Information*

Processing Systems (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 14011–14023, Curran Associates, Inc., 2020.

[236] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, p. 100004, 2019.

[237] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Computer Vision - ECCV 2014 Workshops* (L. Agapito, M. M. Bronstein, and C. Rother, eds.), (Cham), pp. 474–490, Springer International Publishing, 2015.

[238] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *Proceedings of the 24th ACM international conference on Multimedia*, ACM, Oct. 2016.

[239] C. Cangea, P. Velickovic, and P. Lio', "Xflow: 1d-2d cross-modal deep neural networks for audiovisual classification," *ArXiv*, vol. abs/1709.00572, 2017.

[240] Z. Gu, B. Lang, T. Yue, and L. Huang, "Learning joint multimodal representation based on multi-fusion deep neural networks," in *ICONIP*, 2017.

[241] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for sar ship detection," *Remote Sensing*, vol. 9, no. 8, 2017.

[242] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 423–443, Feb. 2019.

[243] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 1247–1255, PMLR, 17–19 Jun 2013.

[244] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, "Correlational Neural Networks," *Neural Computation*, vol. 28, pp. 257–285, 02 2016.

[245] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 162–169, 2019.

[246] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[247] Y. Wang, F. Sun, M. Lu, and A. Yao, "Learning deep multimodal feature representation with asymmetric multi-layer fusion," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3902–3910, 2020.

[248] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

A mia madre e a Yara. A mio padre e a mio fratello.