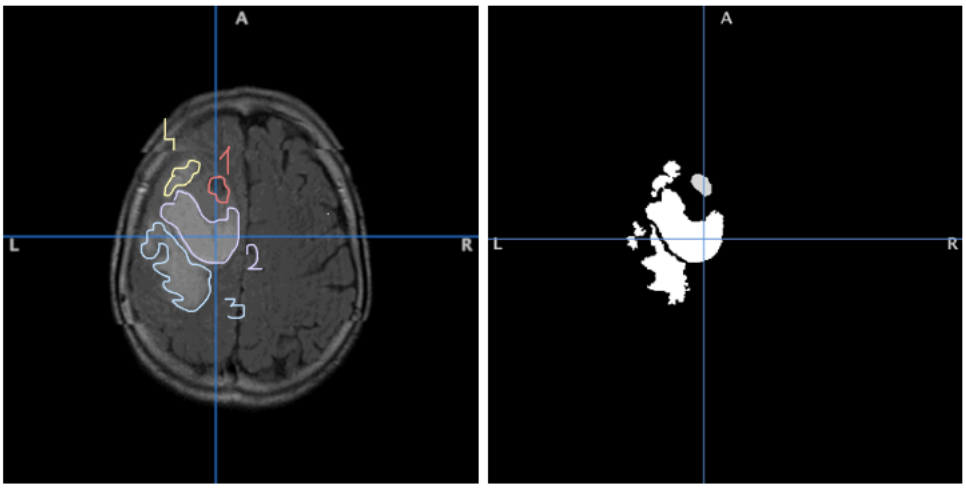
TÉCNICO LISBOA



Evaluation of the impact of deep-learning based Apollo in improving neuroradiological workflows

Carlota de Macedo Santos

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Dr. Akshay Pai
Prof. Ana Catarina Fidalgo Barata

Examination Committee

Chairperson: Prof. Patrícia Margarida Piedade Figueiredo
Supervisor: Dr. Akshay Pai
Member of the Committee: Prof. Rita Homem de Gouveia Costanzo Nunes

July 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was performed at *Cerebriu* (Copenhagen, Denmark), in partnership with Instituto Superior Tecnico (IST) (Lisbon, Portugal) during the period September 2020 - June 2021, under the supervision of Prof. Ana Catarina Fidalgo Barata (IST) and Dr. Akshay Pai (*Cerebriu*), within the frame of the Erasmus + traineeship program.

Acknowledgments

” - How would you characterize *life* in one word ? ”
” - hum...”.
(Silence)
I would say *fusion*.”

Biologically, life starts with a fusion between two haploid cells.
In its essence, life is defined by the choices one makes. Each choice is the product of merging external context and ethical concerns with physical sensations and psychological status.
Knowledge is the fusion of different inspirations and aspirations.
Science is the fusion of observations and theoretical frameworks.
The present document results from all the described fusions, surrounded by the smell of sourdough bread from Mirabelle’s (Copenhagen) and illuminated by the reflection of a bright beam of sunlight on the azulejos (Lisbon).
To Catarina, Akshay, and Becky
To Sofia, Kathrin, Louis, and Artur,
To Aurora and my parents,
To Ottolenghi and Simão,
and To you that is about to read my 8 months journey.

Abstract

Deep Learning (DL) methods for pathology segmentation and classification have gained undeniable relevance in the radiology department. Their promising potential must be balanced with the risks of misclassifications in unseen data. Evaluating robustness with an adequate set of metrics is a crucial step that is usually done suboptimal in the current practices.

In this thesis, we propose a comprehensive evaluation framework that specifically addresses these limitations and jointly assesses the performance of DL models at an image (classification) and lesion (segmentation) levels. Besides analyzing network behaviours across tasks, our method gives a measure of robustness by 1) evaluating the impact of acquisition parameters on performance and 2) applying the framework to an external dataset. The experimental analysis is conducted for two DL solutions, *Apollo* and *nnU-Net*, trained on the same data.

Results show that algorithms are heavily hampered by unintended data bias. In particular, we obtain lower performances for poorly represented pathologies in the training set and verify that the algorithms struggle to predict from out-of-distribution data, *i.e.* acquired with a different sequence or in a different direction. Conversely, more discriminative features are learnt for predominant classes and on prevalent sequence types or orientations. Our experiments also suggest that robustness can be improved by identifying key design decisions in the algorithm pipeline formulation.

By raising awareness on the importance of external validations and by providing alternatives to the current evaluation frameworks, we give a further step towards the seamless integration of DL technologies in medical settings.

Keywords

Deep Learning; Robustness; Unintended Data Bias ; Distributional Shifts; Magnetic Resonance Imaging.

Resumo

Os métodos de Aprendizagem Profunda (DL) têm ganho uma inegável relevância em radiologia, pelo papel preponderante que desempenham na segmentação e classificação de patologias. O seu potencial deve ser contrabalançado pelos riscos de classificações erradas em dados não vistos. Avaliar a robustez de um algoritmo com um conjunto adequado de métricas é um passo crucial que é tendencialmente ignorado na maior parte dos trabalhos.

Nesta tese, propomos um método de avaliação abrangente que aborda estas limitações e avalia conjuntamente o desempenho de modelos DL ao nível da imagem (classificação) e da lesão (segmentação). Para além de analizar o comportamento para uma dada tarefa, o método inclui uma medida de robustez ao 1) avaliar o impacto de parâmetros de aquisição no desempenho e 2) avaliar num conjunto de dados externo. A análise experimental é realizada para *Apollo* e *nnU-Net* treinadas no mesmo conjunto de dados.

Os resultados mostram que algoritmos são fortemente prejudicados pela existência de um enviesamento involuntário de dados. Obtemos desempenhos inferiores para patologias sub-representadas no conjunto de treino e verificamos que os algoritmos têm dificuldade em funcionar com dados adquiridos com uma sequência ou orientação diferente. Inversamente, são aprendidas características mais discriminatórias para classes e tipos de sequência ou orientações prevalecentes. A análise experimental também sugere que a robustez pode ser melhorada através da identificação de decisões chave quanto à formulação do algoritmo.

Ao sensibilizar para a importância de validações externas e fornecer alternativas aos métodos de avaliação actuais, pretendemos agilizar a integração de tecnologias DL em ambientes hospitalares.

Palavras Chave

Aprendizagem Profunda (DL); Robustez; Enviesamento Involuntário dos Dados ; Alterações na Distribuição dos Dados; Imagem por Ressonância Magnética.

Contents

List of Figures

List of Tables

Acronyms

ADC	apparent diffusion coefficient
AI	artificial intelligence
BBB	brain blood barrier
BCa	bias corrected and accelerated
CE	Conformité Européenne
CI	confidence intervals
CNN	convolutional neural network
CPU	central processing unit
CSF	cerebrospinal fluid
CT	computed tomography
DICOM	digital imaging and communications in medicine
DL	deep learning
DWI	diffusion weigthed imaging
FCN	fully convolutional neural network
FDA	Food and Drugs Administration
FID	free induction decay
FLAIR	fluid attenuated inversion recovery
FN	false negative
FP	false positive
GBCAs	gadolinium-based contrast agents
GPU	graphics processing unit
L0	label 0
L1	label 1

L2	label 2
L3	label 3
ML	machine learning
MLP	multi layer perceptron
MRI	magnetic resonance imaging
NIFTI	neuroimaging informatics technology initiative
IST	Instituto Superior Tecnico
PAC	picture archiving and communication
PWI	perfusion weighted imaging
ReLU	rectified linear unit
RF	radio frequency
TR	repetition time
TE	echo time
TN	true negative
TP	true positive
T2 * GRE	T2 * gradient echo
SAFE	Stroke Alliance for Europe
SAH	subarachnoid hemorrhage
SVM	support vector machine
SWAN	susceptibility weighted angiography
SWI	susceptibility weigthed imaging
UI	user interface
WHO	World Health Organization
2D	two dimensional
3D	three dimensional

1

Introduction

Contents

Biomedical engineering is an interdisciplinary field that arose from the fusion between biology and technology. A clear manifestation of that fusion can be found in deep learning (DL), particularly in convolutional neural network (CNN). CNN are one of the greatest applications of transferring biological concepts to automated networks. They have been derived from cat vision system [1] and were firstly integrated in the *Neocognitron* [2] ([3]). With the increase in computational efficiency, the availability of large amount of data, and the fact that societies are moving towards process automation , DL has been evolving from the *Neocognitron* to more advanced and complex neural network architectures. This increase in complexity has extended the landscape of possible DL applications. In Radiology, DL has gained undeniable relevance [4]: from image processing to segmentation, DL is expected to play an essential role in the digital health revolution [5]. In that context, *Cerebriu* [1] developed *Apollo*, a DL software that aims at providing faster diagnosis and improving the efficiency of hospital workflows. *Cerebriu* is a med-tech start-up based in Copenhagen with worldwide partners in Denmark, Israel, the United States, France, Germany, and Japan. Their in-house software *Apollo* is at the core of this master thesis and is currently under clinical validation.

Chapter 1 gives a foretaste of the clinical and scientific relevance of this work developed at *Cerebriu* in collaboration with Instituto Superior Tecnico (IST). This chapter starts by motivating the problem addressed in this thesis from a general (Section 1.1) to a more specific (Section 1.2) *Apollo*-based perspective. The remaining sections are dedicated to the approach description and the thesis organization.

1.1 Motivation

According to European Union, between 2013 and 2018, the use of medical equipment, specially computed tomography (CT) and magnetic resonance imaging (MRI), has been rising (Figure 1.1) [6]. Among the countries that have reported larger increases, Denmark saw its numbers of CT and MRI scans being multiplied by approximately 1.3 and 1.44 respectively. In turn, Portugal registered an increment of 50% in the number of CT scans and of 96% in the number of MRI scans. Higher number of scans imposes an increased of the radiologists workload [5] and may compromise workflows, quality of care and disease management [7] due to the fatigue and shortage of medical staff. As an example, in 2010, it was shown that, in a eight hour shift, a radiologist was asked to interpret one scan (MRI or CT) every 3-4 seconds [7]. This is all the most unfortunate knowing that 60% of the acquired scans are unnecessary [1]. With the development of technology and the increased computational power, DL appears as a promising solution to overcome the aforementioned issues.

[1] David Hubel and Torsten Wiesel contributions, 1964
[2] Kunihiko Fukushima, 1980 [2]

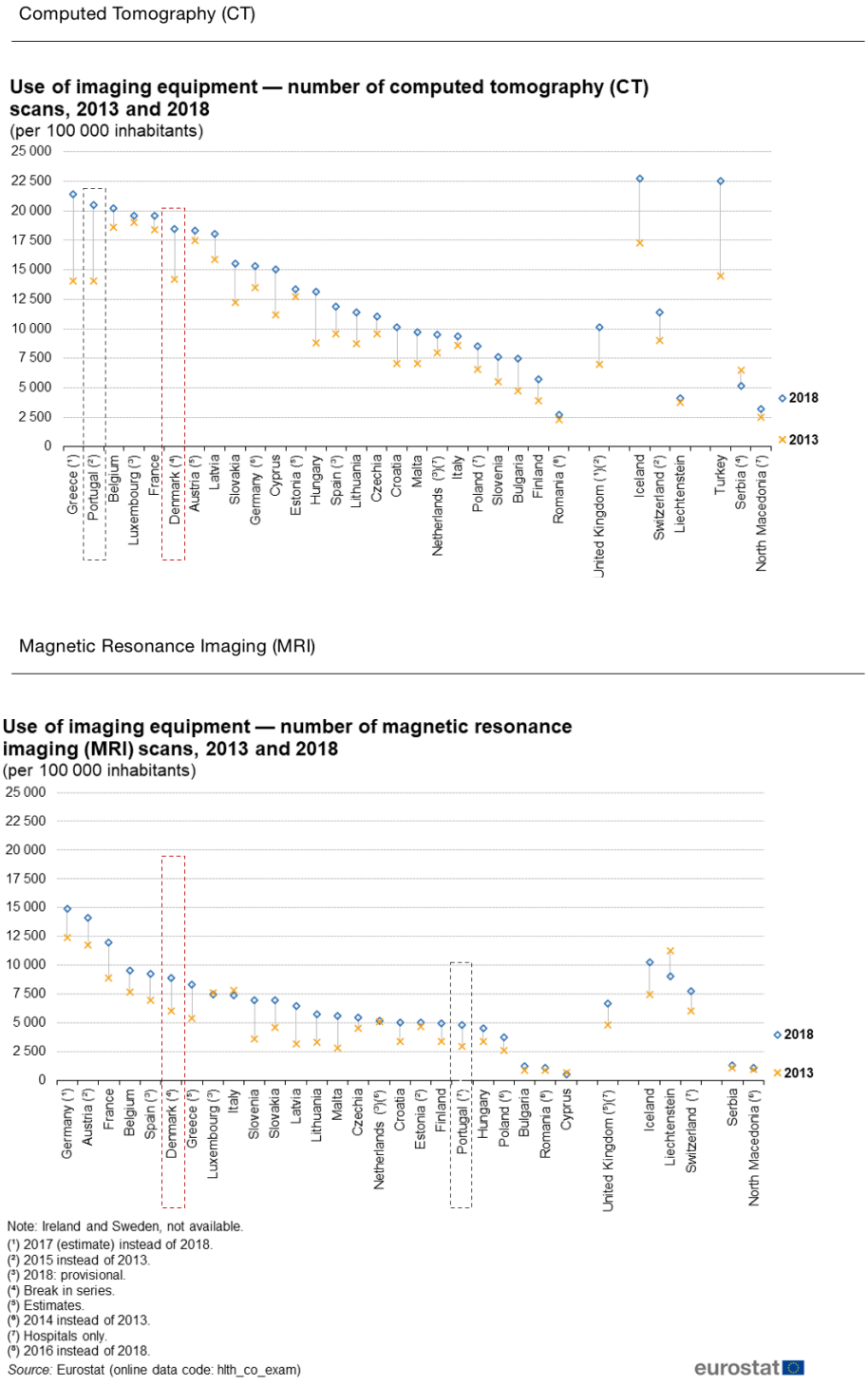


Figure 1.1: Number of (**Top.**) CT and (**Bottom.**) MRI scans per 100 000 inhabitants between 2013 (orange) and 2018 (blue) across European countries.
Denmark and Portugal, referred in the discussion, are highlighted in red and grey, respectively.
This figure was extracted from [6].

Its underlying potential does not only arise from its capacity to handle massive amounts of data and alleviate the radiologist workload, but also from its ability to discover relationships between scan features and patho-physiological attributes that may not be included in the radiologists lexicon [5, 8].

From a patient perspective, the added value of integrating DL is particularly evident in strokes, hemorrhages and tumors managements, as discussed below. Strokes affect 15 million patients worldwide each year [9]. They are currently responsible for more than 4.5 millions deaths and, by 2030, this number is expected to reach 7.8 million [9]. Additionally, more than 30% of the patients suffer from permanent disabilities and the subsequent rehabilitation contributes to the increase of healthcare costs [9]. Therefore, by 2030, with the predicted increase of the disease incidence, strokes will become a cumbersome burden for healthcare systems, from a human and economical aspects [9]. Innovative approaches and guidelines should prioritize, at first, ischemic stroke management as 87% of strokes have an ischemic nature [3]. Ischemic strokes have a time-dependent nature and treatment selection (thrombolysis or mechanical thrombectomy) depends on a correct estimation of the onset and detection of potential hemorrhages at the infarction site [9]. Current delivery of stroke care, selection of reperfusion treatment, and triage for resource-intensive stroke units have failed to provide a satisfactory patient outcome and efficient use of the hospital resources [9]. New procedures should seek to provide faster estimates of the infarction onset and more accurate evidences of hemorrhages findings at the infarction site to enable a more rapid and trustworthy treatment delineation [11]. Therefore, DL algorithms that simultaneously automate the detection of strokes and hemorrhages could be a great support for fast radiology decision and relieve the burden of infarctions in clinical settings.

Similarly to ischemic strokes, managing hemorrhages also requires early and accurate diagnosis. Within the first three hours of onset, the hematoma is growing at a fast rate and the damaged area is expanding [12]. However, diagnosis of intra-cerebral hemorrhages is not straightforward. Depending on its age and location, hemorrhages may appear differently on scans [13]. Hence, DL could contribute to reduce misdetections of specific sub-types of hemorrhages. This is specially relevant for subarachnoid hemorrhage (SAH) for which the high variability in their MRI intensities, the blooming effect produced by adjacent bones, or the dilution of blood with cerebrospinal fluid (CSF) jeopardize its diagnosis and increase the misdetection rate [14–16].

Regarding tumors, current MRI protocols rely on gadolinium-based contrast agents (GBCAs) for diagnosis and monitoring of brain cancer. Vascular network growth is key for tumors proliferation allowing for adequate oxygen and nutrients supply. The generated vasculature is structurally and functionally abnormal, build on leaky and immature blood vessels [17]. Hence, the injected venous contrast accumulates at the pathological site and, based on GBCAs paramagnetic properties, tumors will appear hypo-intense in T_2 or hyper-intense in T_1 scans [18]. However, some concerns have been raised on the potential

[3] Data drawn from the American Stroke Association https://www.stroke.org/en/about-stroke/types-of-stroke [10]

harm of GBCAs related with its deposition in brain tissues and with the risks of nephrogenic systemic fibrosis in patients with renal failure [19] [4]. Therefore, DL could be a viable answer for contrast reduction, enabling the diagnosis without requiring contrast administration to accurately identify the pathology [20].

Even if DL shows clear advantages in supporting healthcare providers, some ethical and medico-legal concerns have been raised [4]. Automating a human-based process is far from being trivial. When it is difficult to discriminate between benign and malignant scenario, doctors are known to over-diagnose malignancy for patient safety. However, this behaviour decreases accuracy and is not always replicated in DL systems [4]. In addition, for healthcare providers, learning is an iterative process, while for DL algorithms parameters are only updated during training. As a result, while doctors are able to cautiously adapt to out-of-distribution data (rare pathological conditions, data acquired with a different scanner, or demographic shift), DL models are hampered by their training and validation data. As a result, variability in disease patterns and in their translation into MRI features may not be understood by the models that become unable to generalize to other scanners, acquisition parameters, populations, or pathology characteristics [4]. Under these assumptions, what would happen if the network fails in identifying a lesion and is responsible for the patient death?

Hence, it is important to assess DL clinical value, safety, and benefit quantification before promoting digitization and automation of healthcare processes [4]. This is a necessary step to bring the technology into clinic and take advantages from its benefits. Fortunately, there is a growing awareness in the scientific community of such topics. Recently, *Radiology* published guidelines to critically appraise current medical DL research from a quality and safety perspective [21]. Similarly, Challen *et al* [4] compiled quality assessment questions to support research and development in DL for clinical applications [4]. Some key considerations are listed below:

- Are the results of the algorithm compared with radiology experts?
- When there are high impact negative outcomes, how does the algorithm adjust its behaviour?
- Are the evaluation metrics comprehensive of the algorithm function?
- Is an external test set used for final statistical reporting?
- Have multi-vendor images or different acquisition protocols been used to evaluate the algorithm?
- Has the system been tested in diverse locations, disease progressions, and populations?
- How is the system going to be monitored and maintained over time to adjust for distributional drift?

Trying to answer some of these questions with the creation of an adequate evaluation framework and finding ways to demonstrate clinical value is the core objective of this thesis. From our test-bed algorithm *Apollo*, we hope to warn researchers of the intrinsic training set dependencies of their models and to inspire more comprehensive evaluation behaviours.

[4] Mechanisms, relevance and potential harm of gadolinium deposition in brain tissues is a fertile field of research. Clinical evidences are needed to balance GBCAs use in clinical settings [19]

1.2 Problem Formulation

Transferring DL concepts to the radiology department for **infarcts**, **tumors** or **hemorrhages** segmentation and classification shows undeniable benefits. *Cerebriu* understood the vast potential of DL in radiology from both hospital and patient perspectives and created *Apollo* to provide decision support at key stages of the diagnostic process [1]. In this section, information on the imaging techniques and the pathologies selected by *Cerebriu* is given, along with the main challenges encountered by the company in that specific context.

1.2.1 MRI Niche Market

While *Cerebriu* decided to focus on the MRI market, CT remains the current gold standard for the diagnosis of most brain pathologies [22]. CT popularity is evidenced in Figure 1.2 with the availability of CT versus MRI scans in European healthcare centers. Except for Germany, discrepancies in the number of CT and MRI scans are striking and Denmark is no exception in that regard. The country shows a number of MRI scans that is about half the number of CT scans.

To understand the different standpoints on the adoption of CT or MRI, Figure 1.3 identifies the key advantages of each modality. On one hand, CT is associated with lower scanning costs and faster acquisition times. It is sufficient to exclude many neurological disorder. On the other hand, MRI is a more versatile technique. By selecting an adequate sequence, higher contrast can be created to differentiate specific brain tissues. Moreover, while CT measures the attenuation of X rays by surrounding tissues [23] [5], MRI signal arises on the interaction of hydrogen with three types of magnetic fields: B_0, B_1, and linear gradient fields G [24] [6]. Figure 1.4 and the following paragraphs discuss the role played by each magnetic field type.

- **External magnetic field** B_0: Hydrogen spins are oriented randomly, resulting in a null net macroscopic magnetic moment M [24]. When an external field B_0 is applied in the longitudinal direction z, spins start precessing at a Larmor frequency w_L (with $w_L = \gamma \times B_0$) and the potential energy E of a magnetic moment $\overrightarrow{\mu}$ in the presence of $\overrightarrow{B} = B_0 \overrightarrow{z}$ is expressed in (1.1):

$$E = \overrightarrow{\mu}.\overrightarrow{B} = -\mu_z B_0 = -\gamma \hbar I_z B_0, \tag{1.1}$$

  where $I_z = \pm\frac{1}{2}$. Two energy states arise with $\Delta E = \gamma \hbar B_0$ [24]. Via Zeeman interaction, the spin magnetic moment vector $\overrightarrow{\mu}$ acquires two possible configurations: a parallel direction (with

[5] Attenuation is defined by the atomic number and physical density of the tissue [13].

[6] MRI phenomenon requires an odd number of protons or odd number of neutrons to present a spin angular momentum S and a subsequent magnetic dipole moment $\overrightarrow{\mu}$ whom magnitude is given by $\mu = \gamma.S$ (where γ is the gyromagnetic ratio). Signal arises from a macroscopic magnetic moment $M = \sum_N \overrightarrow{\mu_N}$, manipulated with static, time-dependent or, spatially-dependent magnetic fields. Hydrogen 1H, with a single proton, by its abundance in the human body under the form of water, is the most sensitive atom to be studied [24].

Medical Equipment Use (2018)

	NUMBER OF SCANS		NUMBER OF SCANS per 100 000 inhabitants	
	CT	MRI	CT	MRI
Belgium	2 307 172	1 090 132	20 190	9 540
Bulgaria	524 078	87 724	7 460	1 249
Czechia	1 178 430	581 713	11 086	5 472
Denmark	1 069 713	516 188	18 464	8 910
Germany (1)	12 662 959	12 334 374	15 320	14 922
Estonia	176 498	66 548	13 351	5 034
Ireland	:	:	:	:
Greece	*2 295 689*	*894 935*	*21 389*	*8 338*
Spain	*5 558 113*	*4 323 169*	*11 877*	*9 238*
France	13 106 629	8 005 980	19 572	11 955
Croatia	415 270	207 233	10 151	5 066
Italy	5 655 395	4 453 226	9 360	7 370
Cyprus	130 763	4 960	15 029	570
Latvia	348 355	124 517	18 076	6 461
Lithuania	319 865	161 071	11 417	5 749
Luxembourg	*119 261*	*45 314*	*19 617*	*7 454*
Hungary	1 287 715	443 741	13 173	4 539
Malta	47 131	27 264	9 725	5 626
Netherlands	*1 633 883*	*898 653*	*9 482*	*5 215*
Austria	1 622 759	1 249 968	18 356	14 139
Poland	3 243 404	1 414 169	8 541	3 724
Portugal (2)	2 107 363	495 337	20 492	4 817
Romania (3)	539 358	213 408	2 738	1 083
Slovenia	157 661	144 631	7 602	6 974
Slovakia	845 085	378 650	15 515	6 952
Finland	317 115	273 257	5 750	4 954
Sweden	:	:	:	:
United Kingdom (2)	*6 724 514*	*4 443 498*	*10 118*	*6 686*

Figure 1.2: As a note, these figures are not exclusive for brain imaging.
Denmark, as being the current *Cerebriu* market, is highlighted in red and Portugal in grey.
This figure is extracted from [25].

	CT	MRI
Brain Imaging	• Detailed evaluation of cortical bone • Sufficient contrast to exclude many neurological disorders	• Higher range of soft tissue contrasts • Higher sensitivity and specificity to detect abnormalities within the brain itself • Better evaluation of structures that may be obscured by artifacts from bone in CT images
Logistics	• Lower sensitivity to patient motion during the examination	• Ability to be performed in any imaging plane without having to physically move the patient
Patient Safety and Outcome	• Suitable for all patients : no underlying risks for patients with implantable medical devices, such as cardiac pacemakers, ferromagnetic vascular clips, and nerve stimulators) • Faster acquisition: golden standard for trauma and acute neurological emergencies	• No use of ionizing radiation: preferred in patients requiring multiple imaging examinations • Smaller risk of causing potentially lethal allergic reaction associated with MRI contrast agents
Availability and costs	• Higher availability • Lower costs per scan	

Figure 1.3: (**Left.**) CT versus (**Right.**) MRI advantages regarding the following attributes: brain imaging quality, logistics, patient safety and, costs and availability.(**Left.**) CT versus (**Right.**) MRI advantages regarding the following attributes: brain imaging quality, logistics, patient safety and, costs and availability.
This figure encompasses information drawn from [22, 24, 26].

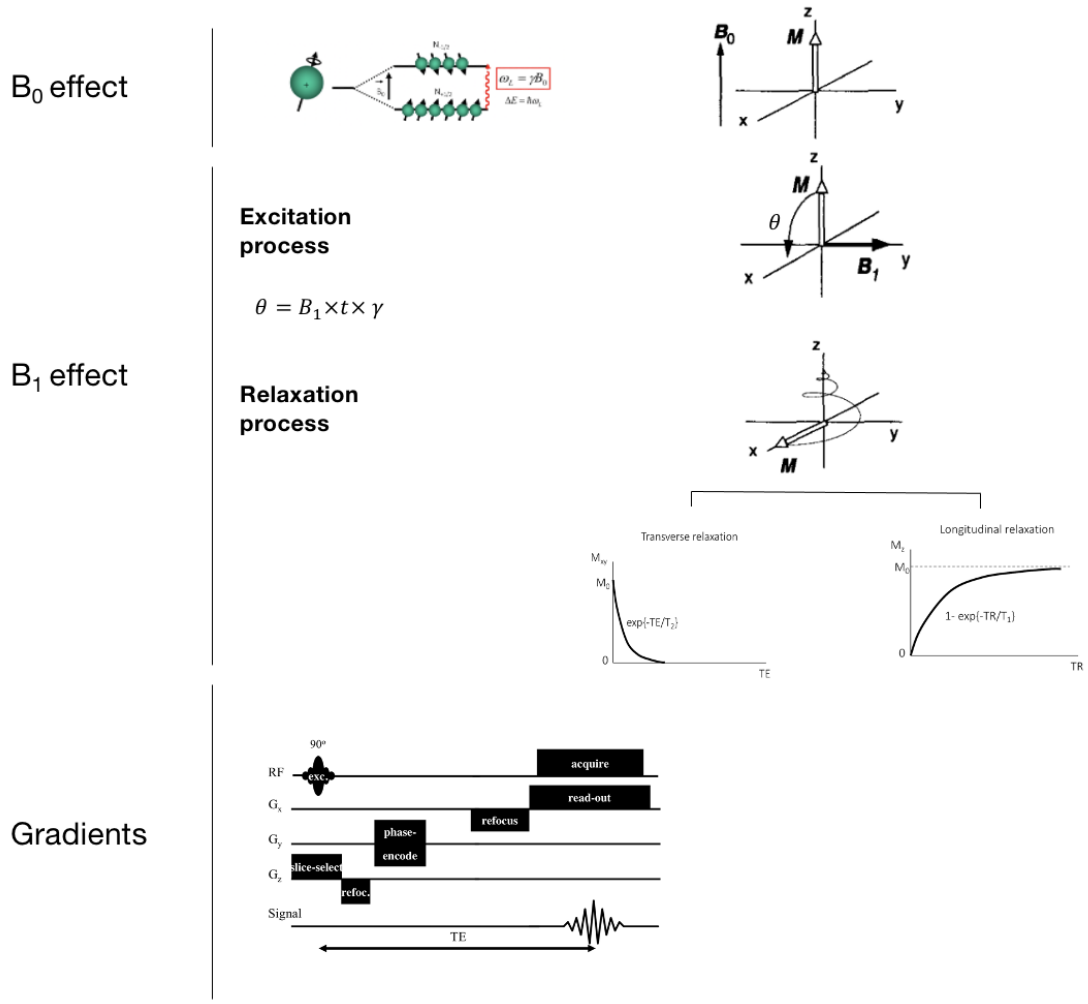


Figure 1.4: MRI basics. Interaction of hydrogen with the three magnetic fields:
(**Top.**) B_0 spin polarization effect, resulting in a positive equilibrium nuclear magnetization M_0 along z direction.
(**Middle.**) B_1 spin excitation and relaxation processes (along the longitudinal and transverse directions). θ refers to the flip angle, *i.e*, the amount of rotation of M_0 during application of B_1 .
(**Bottom.**) linear gradient fields G for spatial encoding in the x, y, and z directions.
Figure adapted from [27] and [28].

$I_z = \frac{1}{2}$) or anti-parallel direction (with $I_z = -\frac{1}{2}$). The lower energy state (with $I_z = \frac{1}{2}$) is more populated but thermal energy is sufficient to allow some migrations to the higher energy level. As a consequence, a positive equilibrium nuclear magnetization M_0 arises in the z direction (Figure 1.4 - Top) [24].

- **radio frequency (RF) pulse** B_1: A radio-frequency field, tuned to the Larmor frequency w_L, is applied in the transverse xy plane to initiate the resonance and excite the spins out of equilibrium. It make the lower energy spins move to higher energy states. B_1 rotates M_0 by an angle θ (flip angle) to the transverse plane with $\theta = B_1.t.\gamma$, with t the duration and B_1 the magnitude of the RF pulse application (*cf* Figure 1.4 - Middle) [24].

When B_1 is switch off, relaxation to equilibrium occurs as a result of two distinct mechanisms. On one hand, the longitudinal relaxation arises from fluctuations at Larmor frequency and molecular vibrations, releasing energy and causing the return to equilibrium along the z-axis at a rate T_1 [24]. On the other hand, transverse relaxation results from fluctuations at Larmor frequency and dephasing (rotations and fluctuations at random frequencies), causing the return to equilibrium along the xy-axis at a rate T_2. Additional mechanisms for loss of spin coherence also occur due to B_0 inhomogeneities. B_0 imperfections, externally applied gradient fields, and intrinsic sample susceptibility difference. These mechanisms increase phase dispersion and fasten the return to equilibrium at a rate T_{2*} with $T_{2*} < T_2$. A refocusing pulse of 180 °at half of the time window between excitation and signal acquisition can be applied to cancel these contributions [24].

From Faraday's law of induction, changes of magnetization in the transverse plane can be perceived by an RF receiver coil. The generated time signal (free induction decay (FID)) is recorded and processed to reconstruct an MRI image [24]. Excitation and readout are defined by two crucial time variables: repetition time (TR) and echo time (TE). While TR accounts for the time between two consecutive excitations (RF pulses), TE is the time between the excitation and the read-out of the FID signal [24, 27].

- **Gradient Fields** G: Gradient Fields G are essential to achieve spatial localization of MRI signals and recreate the two dimensional (2D) or three dimensional (3D) images. When a spatially dependent magnetic field G is applied, the total magnetic field in the $\overrightarrow{r}$ direction is given by: $B(\overrightarrow{r}) = B_0 + \overrightarrow{G}.\overrightarrow{r}$. The frequency of the spins is, then $\omega(\overrightarrow{r}) = \gamma B_{\overrightarrow{r}} = \omega_0 + \gamma\overrightarrow{G}.\overrightarrow{r}$ and the phase dispersion is obtained with $\Delta\phi(\overrightarrow{r}, t) = \gamma\overrightarrow{G}.\overrightarrow{r}t$ [27].

  As illustrated in Figure 1.4 (Bottom), three different linear gradient are applied for slice selection (z axis), frequency encoding (x axis), and phase encoding (y axis) [24, 27]. Designing the MRI sequence and selecting an adequate strength and duration for the gradient are key factors that influence image resolution [27].

Regarding the versatility of the technique, by selecting adequate pulse sequence parameters (flip angle θ, TR, and TE) and taking into account the sample physical characteristics (proton density, T_1, and T_2), particular contrasts can be obtained, enhancing different brain tissues and textures [24, 27]. To make its predictions, *Apollo* combines the information of three common MRI sequences: diffusion weigthed imaging (DWI), fluid attenuated inversion recovery (FLAIR), and gradient echo sequences:

- **Diffusion Weighted Imaging DWI**: DWI is a T_2-weighted image, where the T_2 signal is attenuated based on the Brownian motion of the protons in the brain. The true diffusion coefficient of each brain volume is approximated by an apparent diffusion coefficient (ADC), computed considering free water motion in the brain.

  Taking the direction of measurement x as an example, the basic concept of this sequence is to

apply spatially varying gradient G_x along x and to reverse this gradient $-G_x$ at $\frac{TE}{2}$. If a refocusing pulse is applied, then the gradient is applied in the same direction G_x (Figure 1.5). A "no motion/-coherent motion" scenario translates a null phase shift $\Delta\phi = 0$ as the second gradient undoes the effect of the first spatially varying gradient. A "incoherent motion" scenario translates a non null phase shift $\Delta\phi \neq 0$, as the second gradient is only be able to refocus the phase dispersion caused by the first gradient, but not the phase dispersion caused by motion [29].

Therefore, voxels with higher diffusion (*i.e* accentuated protons motion) present a larger dephasing $\Delta\phi$ and a faster transverse magnetization relaxation rate, appearing hypointense in the image [29]. Conversely, voxels with restricted diffusion appear hyperintense.

- **Fluid Attenuation Inversion Recovery FLAIR**: FLAIR is a T_2-weighted image, where signal cancelling is applied to a specific tissue. The basic concept of this sequence is to apply an inversion pulse of 180°($M_z \rightarrow -M_z$) and excite when the longitudinal magnetization $M_{z_i} = 0$ for a certain tissue i (Figure 1.6 - null point) during relaxation. Therefore, it cancels the contribution of that tissue in the image. Usually, the selected tissue is the cerebrospinal fluid (CSF), allowing for an increased lesion-to-background contrast and enhanced visualization of brain parenchyma abnormalities. Lesions appear hyperintense compared to regular T_2-weighted sequences [24].
- **Gradient echo sequences**: T2 * gradient echo (T2 * GRE), susceptibility weigthed imaging (SWI), and susceptibility weighted angiography (SWAN) are T_2*-weighted images, where the T_2* signal is attenuated based on hemoglobin and its degradation products magnetic susceptibility. The aforementioned sequences are mostly used to characterize brain hemorrhage based on the paramagnetic properties of deoxyhemoglobin, methemoglobin, and hemosiderin versus diamagnetic properties of oxyhemoglobin. Due to their magnetic susceptibility, deoxyhemoglobin and methemoglobin (paramagnetic) and hemosiderin (superparamagnetic) generate a susceptibility difference between blood vessels and surrounding tissues, accentuating the aforementioned T_2* effect. As a result, transverse relaxation is achieved faster and tissues are identifiable by their hypointensity in the image [30].

  The major difference between these gradient echo sequences is that T2 * GRE is a 2D technique while the remaining two are 3D (Figure 1.7). Patented by different MRI vendors, the aforementioned 3D techniques present some distinctive characteristics in the sequence design and the post-processing of the acquired image [7].

1.2.2 Pathological Context

The pathologies targeted by *Cerebriu* are infarcts, tumors, and hemorrhages. *Cerebriu* adopted terminology for ischemic strokes is infarcts and will be kept for all the remaining chapters of the thesis.

[7] SWI relies on the acquisition of one echo at a specific TE and use additional phase information in the post-processing step. SWAN is a multi-echo sequence and exploit only the magnitude information [14].

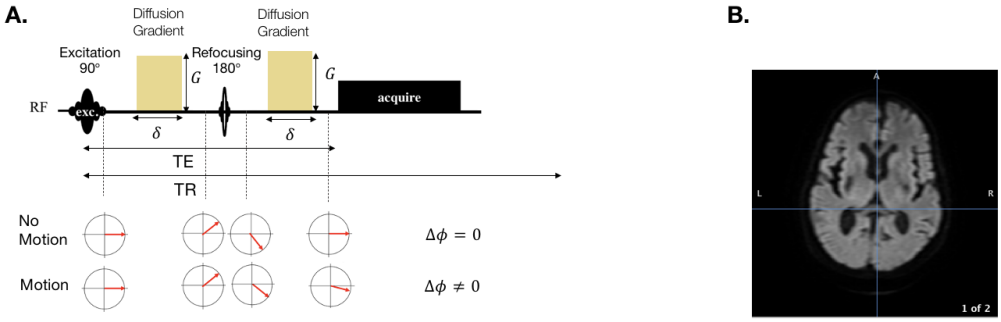


Figure 1.5: DWI - (**A.**) sequence design (adapted from [29] and [27]) and (**B.**) DWI scan example. Diffusion gradients are applied to capture diffusion. A ”coherent motion” versus ”incoherent motion” scenarios with their respective phase shift ϕ are depicted in the figure.

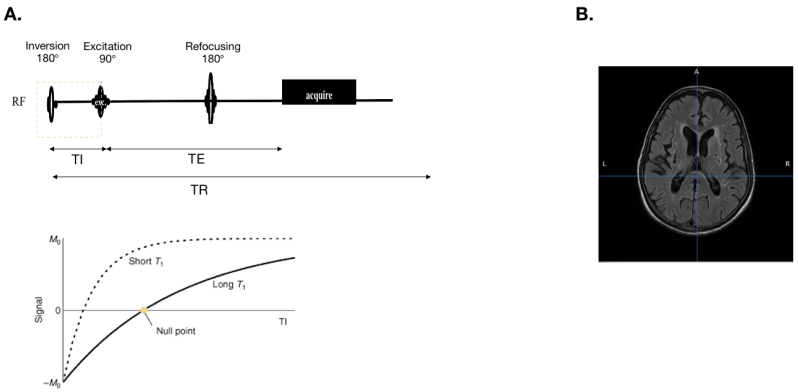


Figure 1.6: FLAIR- (**A.**) FLAIR sequence design (adapted from [27]) and (**B.**) FLAIR scan example. A clear cancellation of CSF signal can be appraised in the figure.

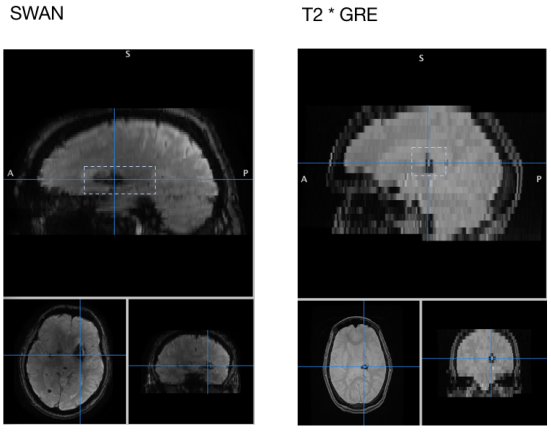


Figure 1.7: Comparison between (**Left.**) SWAN (3D) and (**Right.**) T2 * GRE (2D). The latter technique shows faster acquisition times against worse image resolution. Hemorrhages are highlighted with a blue square.

Ischemic strokes account for 87 % of all strokes diagnosis [10]. The remaining 13% are hemorrhagic strokes, where the blood supply is partially suspended by the rupture of the vessels due to uncontrolled hypertension or underlying blood-vessel abnormalities (cerebral aneurysms, arterio-venous malformations) [9]. Hemorrhagic strokes are considered as hemorrhage in *Cerebriu* pathology formulation. Differences in these pathological mechanisms are illustrated in Figure 1.8.

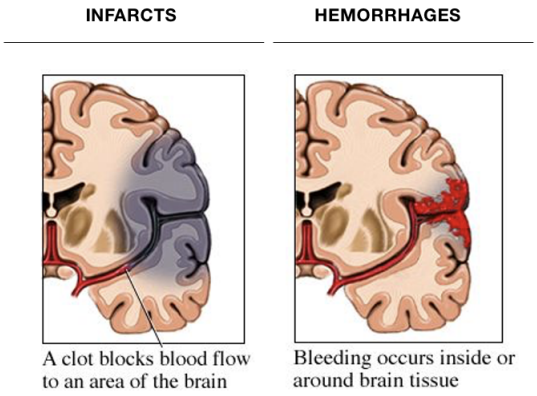


Figure 1.8: Comparison between (**Left.**) ischemic stroke (considered as infarcts in *Cerebriu* nomenclature) and (**Right.**) hemorrhagic stroke (considered as hemorrhages in *Cerebriu* nomenclature). While infarcts occur by a thrombotic or embolic occlusion of the cerebral artery, reducing the flow of oxygen and nutrients to brain tissues, hemorrhages are originated by a rupture of blood vessels in the brain. Figure adapted from [31].

- **Infarcts - Ischemic Strokes**: An ischemic stroke is induced by a thrombotic or embolic occlusion of a cerebral artery. The resulting interruption of blood supply to the brain reduces the flow of oxygen and nutrients to the tissues at the infarction site. Ischemia gives rise to a hypoxia scenario and necrotic tissue appears at the core of the lesion. In the surroundings of this non salvageable area, there is a hypo-perfused region (*i.e* penumbra) that is supplied during infarction by a collateral blood flow. The penumbra is the target of stroke treatment and the portion of potentially salvageable ischemic tissue is estimated under two parameters: the patient collateral blood flow and the infarction onset [11].

  Blood supply can be restored by thrombolytic therapy or by mechanical thrombectomy to remove the clot of the artery. Treatment selection relies on the estimation of the infarct onset and on the detection of hemorrhages in the surrounding of the lesion. For an onset between zero to six hours and in the absence of hemorrhages, thrombolytic therapy is prescribed [9]. Therefore, an efficient estimation of the age of the infarct is essential for a correct disease management. The discrimination of the temporal evolution of ischemic strokes encompasses hyper-acute, acute, sub-acute, and chronic stages but their time delineation varies in the literature [32].

  Figure 1.9 shows the current modalities used in infarct diagnosis. CT is the gold standard for infarct detection. However, MRI has been shown to outperform CT in detecting micro-bleeds and subtle

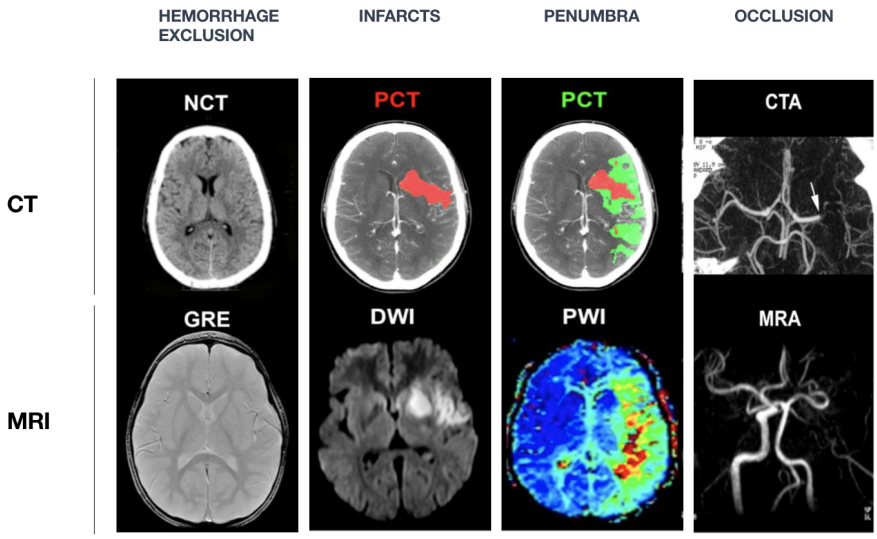


Figure 1.9: CT versus MRI for stroke, penumbra, and occlusion diagnosis. Figure adapted from [35].

hemorrhages and in reducing the ionization dose. Additionally, by allowing a better delineation of the infarct core and ischemic penumbra, MRI seems to be a more powerful technique for stroke triaging [33] [8]. Among the possible sequences acquired in a MRI scanner, DWI and the subsequent ADC map are the most sensitive methods for detecting ischemia at early stages [11]. Changes in the energy metabolisms engenders a loss of ionic gradients and a net transfer of water from the extra to the intra-cellular compartment. The excessive accumulation of water molecules in intra-cellular compartment and the consequent reduced extracellular volume are depicted by DWI as a reduction in water diffusion. New techniques combining DWI with perfusion weighted imaging (PWI) are being developed [34], built on the hypothesis that DWI reflects the non salvageable infarct area while PWI reflects the overall hypo-perfused lesion (infarct + penumbra) [11].

MRI techniques other than DWI are also attracting interest to support the decision-making process and onset estimation. FLAIR can help identifying infarcts within the first three hours of symptom onset. If not hyperintense on FLAIR images, infarct has more than 90% probability of being imaged within the first 3 hours of symptom onset while hyper intensity translates a three to eight hours scenario after onset [11]. Hemorrhages at infarction site can also be detected by SWI, SWAN, or T2 * GRE based on its sensitivity to blood magnetic properties [11].

- **Tumors**: A tumor consists of a mass of abnormal tissue that arises from preexisting body cells. It is characterized by an independent and uncontrollable growth with no associated function. The hyperplasia can be accompanied by anaplasia [36]. Tumors can be divided between benign and malignant. While cells of a benign tumor are normal in shape, size and structure, cells of a malig-

[8] New protocols are emerging to bridge the existing gaps of CT imaging for stroke detection. It includes non-contrast CT for hemorrhages exclusion, perfusion CT for penumbra estimation, and CT-angiography for intracranial thrombus and vascular narrowing identification. These new protocols will not be discussed in the present work, as most of the hospitals do not incorporate them in the routine procedures [9].

nant tumors are usually different from their surrounding tissue. They loose their initial function by reaching a less differentiate state and show higher growth and spread rate. Besides its degree of malignity, tumors can be classified according to their composition and the relative proportion of its solid, cystic, necrotic, hemorrhagic, edema, or protein-rich components [37]. An example of tumor classification according to its constituents can be appraised in Figure 1.10.

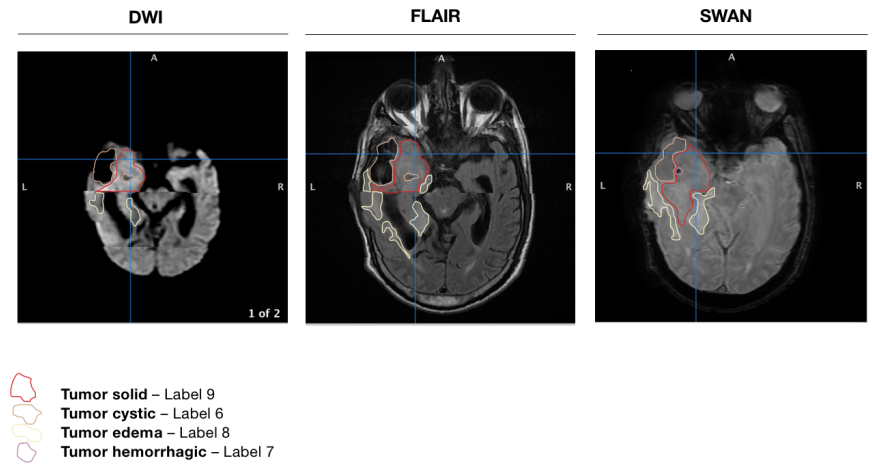


Figure 1.10: Tumors classification according to their composition. Solid tumor (red), cystic tumor (orange), edema tumor (yellow) and hemorrhagic tumor (purple) are identified in (**Left.**) DWI, (**Middle**) FLAIR, and (**Right.**) SWAN scans. Hemorrhagic tumor is only visible in SWAN modality. Other tumor components (necrotic and protein-rich) are not represented in the Figure.
The slices across modalities are the same and the patient is extracted from MedAll dataset [1].

The current recommendations for a standardized brain tumor imaging protocol lean on MRI sequences and contrast administration [38]. The minimum recommended sequences includes pre and post-contrast of T_1-weighted images, pre-contrast T_2 FLAIR and DWI, and post-contrast T_2-weighted spin-echo [38].

- **Intracerebral Hemorrhages**: Intracerebral hemorrhages can have a traumatic or non-traumatic origin. Non traumatic intracerebral hemorrhages are caused by a rupture of blood vessels in the brain [39]. In the absence of vascular malformation, hemorrhagic conversion of an ischemic stroke, intracranial tumor, or coagulopathy, the hemorrhage is referred as primary intracerebral hemorrhage. Primary hemorrhages are predominant, mostly induced by hypertensive arteriosclerosis and cerebral amyloid angiopathy [40]. The release of blood in the extra-vascular space induces a mechanical disruption of the neurons and glia, followed by mechanical deformation causing oligoemia, neurotransmitter release, mitochondrial dysfunction, and membrane depolarisation. Under the new pathological condition, coagulation and hemoglobin breakdown products activate the microglia that induces a disruption of the brain blood barrier (BBB) and the apoptosis of neurons and glia [39].

Brain hemorrhages can be subdivided according to its location in intra-axial (intraparenchymal and intraventricular) and extra-axial (subdural, epidural, and subarachnoid), as shown in Figure 1.11 [41, 42] . Epidural and subdural hemorrhages are distinguished by their morphology and topography.

Apart from the spatial sub-division, hemorrhages can be characterized as a function of time. Hyper-acute, acute, early sub-acute, late sub-acute, and chronic stages are usually identified in the literature [41]. The extra-axial sub-types share similar characteristics on MRI and CT scans with intraparenchymal hemorrhages with slower progression across stages [13, 16]. An extra discrimination between venous and arterial hemorrhages should also be made due to the different oxygen content that may influence the transition between hemoglobin states and, consequently, the diagnosis of the pathology [13].

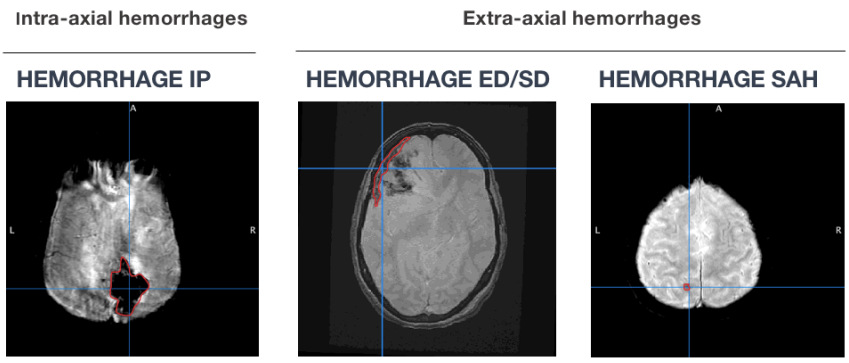


Figure 1.11: SWAN of different hemorrhage sub-types: (**Left.**) Intra parenchymal hemorrhages, (**Middle.**) Extra/-sub dural hemorrhages, and (**Right.**) SAH. Intraventricular hemorrhages are not discriminated in *Cerebriu* annotation protocol

Diagnosis of intracranial hemorrhages is based on CT [39] and MRI [33]. The appearance of hemorrhages on both imaging techniques parallels the temporal evolution of the disease [41]. While CT depicts hemorrhages as a high-attenuation mass within the brain tissues, MRI has proven to be more accurate when it comes to estimate the stage of the hemorrhage and to detect early hemorrhagic conditions [13, 33]. Although CT attenuation varies linearly with protein content (mainly hemoglobin) and hematocrit measurement , some artifacts surrounding the skull can mimic hemorrhage and lead to misclassification and inadequate patient management [13]. MRI, via the introduction of blood-sensitive gradient echo sequences, is able to trace the sequential evolution of an hemorrhage by revealing relevant features of hemoglobin transformation and subsequent changes in its magnetic properties within the hematoma [33]. Given the age of an hematoma, the hemoglobin undergoes a transformation from intra-cellular oxygenated hemoglobin to deoxyhemoglobin and hemosiderin, associated with different oxidation state of its constitutive iron atoms.

While oxyhemoglobin is diamagnetic, deoxyhemoglobin and methemoglobin are weakly paramagnetic and hemosiderin is superparamagnetic [43], creating changes in signal intensity. By being sensitive to hemosiderin, MRI has the ability to not only detect acute and chronic hematomas but also old and clinically silent cerebral microbleeds [42]. In addition to changes in magnetic properties of hemoglobin, the compartmentalization observed in hemorrhages is a necessary condition to create a local field inhomogeneity and trigger spin dephasing and signal loss captured by MRI images [43].

Table 1.1 summarizes the three main disease classes and their corresponding sub-types, as presented in the previous paragraphs.

Table 1.1: (**Left.**) Labels - pathologies correspondence used in Annotation Protocol.
(**Right.**) Labels - pathologies correspondence used in *Apollo Cerebriu* training and inferences.

Annotations

PATHOLOGY		LABEL
Infarct	Hyper acute	2
	Acute	1
	Sub acute	11
	Hemorrhagic Transformation of Infarct	21
Chronic infarct	Hypointense	12
	Hyperintense	20
Tumor	Solid	9
	Solid – isointense in FLAIR	24
	Cystic	6
	Protein Rich Cyst	22
	Edema	8
	Necrotic	10
	Hemorrhagic	7
Intraparenchymal Hemorrhage (IP)	Hyper Acute	26
	Acute	28
	Early Sub Acute	30
	Late Sub Acute	32
	Chronic	34
Epidural/Subdural Hemorrhage (ED/SD)	Hyper Acute	27
	Acute	29
	Early Sub Acute	31
	Late Sub Acute	33
	Chronic	35
Sub Arachnoid Hemorrhage (SAH)	Acute	25
	Chronic	36
Edema		18
Gliosis		19

Training and Inferences

PATHOLOGY	LABEL
Infarcts	1
Tumors	2
Hemorrhages	3
Chronic Infarcts	4

1.2.3 Scientific Challenges

From the general issues addressed in Section 1.1, some need particular attention when it comes to deal with the aforementioned pathologies and imaging techniques.

Challenges of automating a human-based process may arise from the broad landscape of patho-physiological attributes of a determined disease [13, 37]. Based on *Cerebriu* experience, tumor is the most demanding pathology to annotate, as it may be characterized by various heterogeneous histological sub-regions and by broad imaging phenotypes [37]. As shown in Figure 1.10, shapes, textures, and intensities vary across tumor types and sub-regions, depending on the underlying biological properties (solid, edema, cystic, necrotic...) [37]. In-house annotators highlighted that the elaborating and up-dating a tumor annotation protocol has been a cumbersome process. Therefore, transferring the aforementioned non trivial human knowledge into neural network features is expected to be challenging. Large amount of data distributed across classes is needed to guarantee that pathology attributes are being learnt correctly by the network during the training phase.

Unfortunately, the design of a high-performance algorithm is not just a question of automating a correct delineation of the relevant pathology characteristics. DL solutions for MRI segmentation and classification also have to be robust to distributional shifts of the radiomic attributes [4, 5, 44, 45]. This is all the most relevant facing the diversity of radiology platforms, the heterogeneity of processes, formats, and protocols, the variability of intra and inter-site scanner manufacturers, models and versions [46]. One pathology may appear differently depending on the acquisition scanner, protocol, and modality. As presented in Figure 1.7, SWAN and T2 * GRE have different signal intensities and image resolutions. Figure 1.12 supports this idea by highlighting the disparity observed in histograms of images acquired by different scanners and modalities (T_1 versus T_2 weighted images). A direct assessment on two datasets used at *Cerebriu* shows the relative intensities between pathologies and background differences, originated by different scanners (Figure 1.13). Having an out-of-distribution image may lead to an erroneous output and have undesired consequences for the patient outcome or hospital workflows [4]. Again, large amount of data is needed to guarantee the ability of an algorithm to generalize across institutions and scanning attributes.

Taking the aforementioned issues into account, the elaboration of an algorithm for segmentation and detection of brain diseases based on MRI highly relies on the amount, quality, and diversity of the data used during training and validation. Data availability, curation, and distribution across classes are essential steps that are intrinsically associated with the algorithm performance [4, 47, 48]. Nevertheless, in medical settings, the access to medical data is limited and DL-based companies are calling for new procedures that facilitate data sharing and discourage the exclusivity of partnership between developers and institutions [48].

1.3 Objectives and Contributions

The present works aims at designing a comprehensive framework for assessing the performance of DL medical segmentation and classification algorithms.

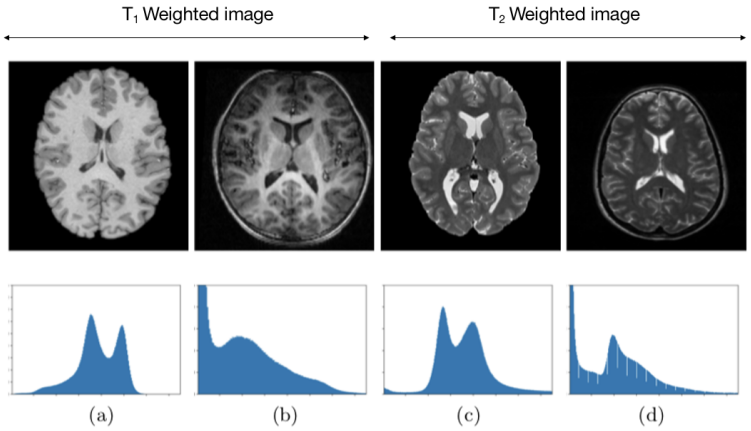


Figure 1.12: Image slices ((**Top**) and their respective histograms of intensities (**Bottom**). Scans are acquired with different scanners ((**Left**) T_1 and (**Right**) T_2 weighted images). The differences observed in the histograms can be easily identified, reinforcing the idea of high variability between images with different acquisition parameters.
Figure drawn from [45].

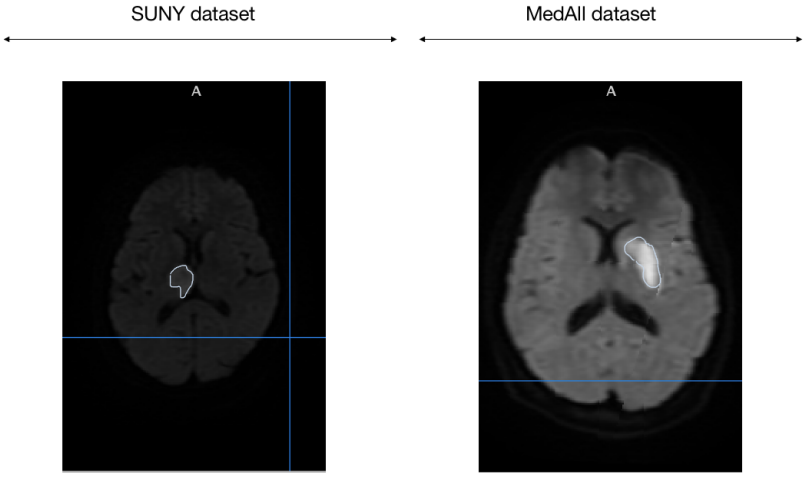


Figure 1.13: Example of infarct from ((**Left**) SUNY dataset acquired with a 1.5 T Siemens scanner and from (**Right**) MedAll dataset acquire with a 1.5 T GE Healthcare scanner. The relative intensity difference between background and foreground classes depends on the clinical sites, scanners and acquisition parameters.

The proposed method is expected to evaluate the robustness of the DL models response to distributional shifts. It is meant to measure the level of agnosticism to the training data and to cover the generalisation ability of an algorithm across hospitals and different MRI hardware parameters. This is a necessary step to bring DL technology from the lab to the clinical practice, tackling the already mentioned concerns of clinicians with respect to its reliability and liability.

Performance will be addressed under three perspectives and statistical validation is provided accordingly:

- **Performance across pathologies** - Gives a deeper understanding of the pathological contexts for which more discriminative features were learnt by DL algorithms. This evaluation could help their integration in meaningful clinical workflows and protocols.
- **Performance across acquisition parameters** - Gives a deeper understanding of the MRI acquisition parameters that allow a better performance of DL algorithms at the clinical sites. It can also be interpreted as a measure of robustness.
- **Performance across datasets** - Gives a deeper understanding of DL algorithms generalization ability. Generalization is defined as the ability to perform well on previously unobserved inputs [3]. It is a preliminary step to ensure their safety at new clinical sites before its implementation.

The conceptual objective of this work is to sensitize DL researchers on conducting an accurate and comprehensive evaluation of their networks and to warn them on the impact of data bias and distributional drifts. The evaluation framework is validated experimentally in two algorithms: *Apollo* [1] (the test-bed algorithm that motivated this project) and *nnU-Net* [49] (the current state-of-the-art DL approach for medical image segmentation).

1.4 Thesis Outline

The remaining chapters are organized as follows. Chapter 2 analyzes how DL can be integrated into clinical pratice and contribute to the modernization of the radiology department. It highlights its potential to optimize hospital resources, implement more efficient workflows, and achieve better patient outcomes. It also identifies the major concerns that hamper the adoption of DL solutions in real clinical settings. Chapter 3 works as an introduction to the two DL algorithms under analysis and describes the step-by-step procedure undertaken for their evaluation. Chapter 4 provides a detailed analysis of the performance of *Apollo* and a comparison with *nnU-Net* is made is parallel. Chapter 5 concludes the thesis, highlighting the importance of assessing the response to distributional shifts in the evaluation guidelines and pointing out directions for future improvements on evaluation methods. A final note on the expected outcome of *Cerebriu* in danish hospitals and its extrapolation to portuguese clinical settings is also provided.

2

Deep Learning: the True Cornerstone of Radiology Revolution

Contents

The goal of **Chapter 2** is to discuss how DL can integrate the routine clinical pratice and contribute to the modernization of the radiology department. Specifically, Chapter 2 provides an overview of DL models applied to disease segmentation and classification of MRI data. Since DL is a family of machine learning (ML) methods, the chapter starts with a brief introduction to ML theoretical concepts. Then, a state of the art analysis is conducted, covering the current research trends, the market situation, and the major barriers for a seamless hospital integration.

2.1 Deep Learning Basics for Clinical Applications

The present section is dedicated to general ML concepts that are progressively refined to DL concepts. It aims at giving the necessary tools to understand the research and market orientations discussed in Section 2.2.
A comprehensive definition of ML was given by Mitchell [3] :

> "A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E."

This definition will serve as a scaffold to introduce and develop the important concepts highlighted in bold. The following subsections address each concept sequentially and succinctly.

2.1.1 The Learning Experience E

The learning experience of an algorithm can follow different learning paradigms: supervised learning, unsupervised learning, or reinforcement learning. Between supervised and unsupervised learning lay different semi-supervised learning experiences, depending on their relative proportion of supervised *versus* unsupervised learning. For medical application, current algorithms mainly rely on **supervised learning** [50], where the algorithm has access to a dataset containing labelled examples during the training and validation steps. In that case, based on observations made on examples of a random vector x and its corresponding label y, the algorithm tries to estimate $p(y|x)$, learning to predict y from x [3].

In order to achieve state-of-the-art performances and face data scarcity, the learning step is usually paired with data augmentation [51] [1]. Augmentation is a key concept when it comes to train a segmentation network in biomedical applications since 1) it is an efficient and pragmatic method to simulate structural and textural changes of anatomical architectures; 2) invariance and robustness to tissue deformation can be learnt by the network; and 3) satisfactory training results are achieved without relying on large training corpora [52].

[1] Data augmentation is a strategy that consists of creating additional data from the already acquired set. Different techniques can be applied from noise addition, flips, rotations, intensity variations to more complex non-rigid deformations

2.1.2 The Type of Task T

The growing role of ML in different domains and a better understanding of principles that underlie intelligence brought forward a diversity of tasks that could be executed by algorithms. Automating processes such as pathologies segmentations and classifications have not only become an attractive field of research but is also starting to be part of the radiologists routine [50]. They are seen as key tasks for pathology diagnosis and evaluation (volume, disease sub-types, location).

A **Classification** task is based on specifying the output category $y \in 1, ..., k$ of the corresponding input x by estimating the mapping function $f : \mathbb{R}_n \rightarrow 1, ..., k$ and computing $y = f(x)$ [3]. A **Segmentation** task is a pixel-wise or voxel-wise classification task. Segmentation clusters image regions into groups that share the same class, assigning a label to each pixel or voxel x of the input image X [49]. Therefore, the algorithm output share the same dimensions as the input and the assigned categories are tightly correlated in space due to image continuity [3]. Accounting for 70% of international medical image analysis competitions, this type of task is an essential ingredient when it comes to clinical applications [49]. Segmentation task at a lesion level is usually combined with classification at an image level. Hence, higher emphasis will be given to segmentation, as it is a pixel-wise classification that determines the final image classification in *Apollo* pipeline.

Each type of task requires a specific model. Traditionally, ML-based image segmentation approaches are designed to perform a segmentation task, using hand-crafted features that were previously extracted from raw data (Figure 2.1 - Left). Hence, the selected features and class labels work both as an input for the classifier to determine the function f such that $y = f(x)$. Defining the number and the type of features or selecting the optimal type of classifier is a challenging and time-consuming process, since it is mainly performed by trial and error [50]. Therefore, ML solutions (support vector machine (SVM), decision trees, or Bayes classifiers) are less suitable when it comes to automate the segmentation of complex diseases that show a broad spectrum of features.

DL-models [53] emerged as a solution by merging feature extraction, selection, and classification into one problem that is optimized during training (Figure 2.1 - Right). It has been proving its superiority with higher quality segmentation and classification accuracy [50]. Different DL building blocks have been proposed to compose more complex and powerful algorithms. From multi layer perceptron (MLP) to the U-Net, an overview across DL solutions landscape for medical segmentation is provided in the following paragraphs and schematic architectures of MLP, CNN and fully convolutional neural network (FCN) are presented in Figure 2.2.

- **MLP**: The MLP architecture (Figure 2.2 - Top) is built upon multiple units, organized in layers, where the number of layers gives the depth of the network. Each unit i is connected to a unit j of

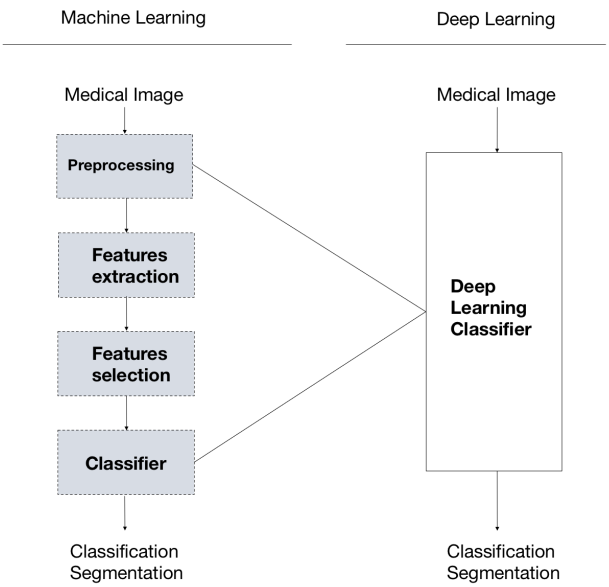


Figure 2.1: (**Left.**) ML framework for image classification and segmentation and (**Right.**) DL framework for image classification and segmentation.
With the high dimensions of medical images (3D), ML cannot compete against DL when it comes to pixel-wise classification. Its trial-and-error process for feature extraction and selection is extremely time-consuming and requires medical expertise.
This figure was adapted from [50].

the next layer through a weight w_{ij} and the output of unit j is given by z_j according to (2.1):

$$z_j = g(w_{0j} + \sum_{i \in previous layers} w_{ij} z_i), \tag{2.1}$$

where g is an activation function continuous and differentiable and z_i the output of unit i [54]. Applying g enables the MLP to extract more complex information from the data and create nonlinear mappings between input and output [55]. During the training, weights ω are optimized to produce the best approximation of the mapping function f given by $y = f(x)$ [3].

The model is described by a direct acyclic graph [3]. The term feed-forward network is drawn from the fact that information is propagated through one single direction, from the input to the output. The term neural network arises from the fact that two MLP units from consecutive layers simulate mathematically a connection between two neurons of the brain.

The reason why MLP are not suitable for image segmentation is that they are computationally expensive to train. Having connections between all units i of layer I and all units j of layer J, each one defined by its own weight w_{ij} make them extremely inefficient for large dimensions inputs and outputs [20].

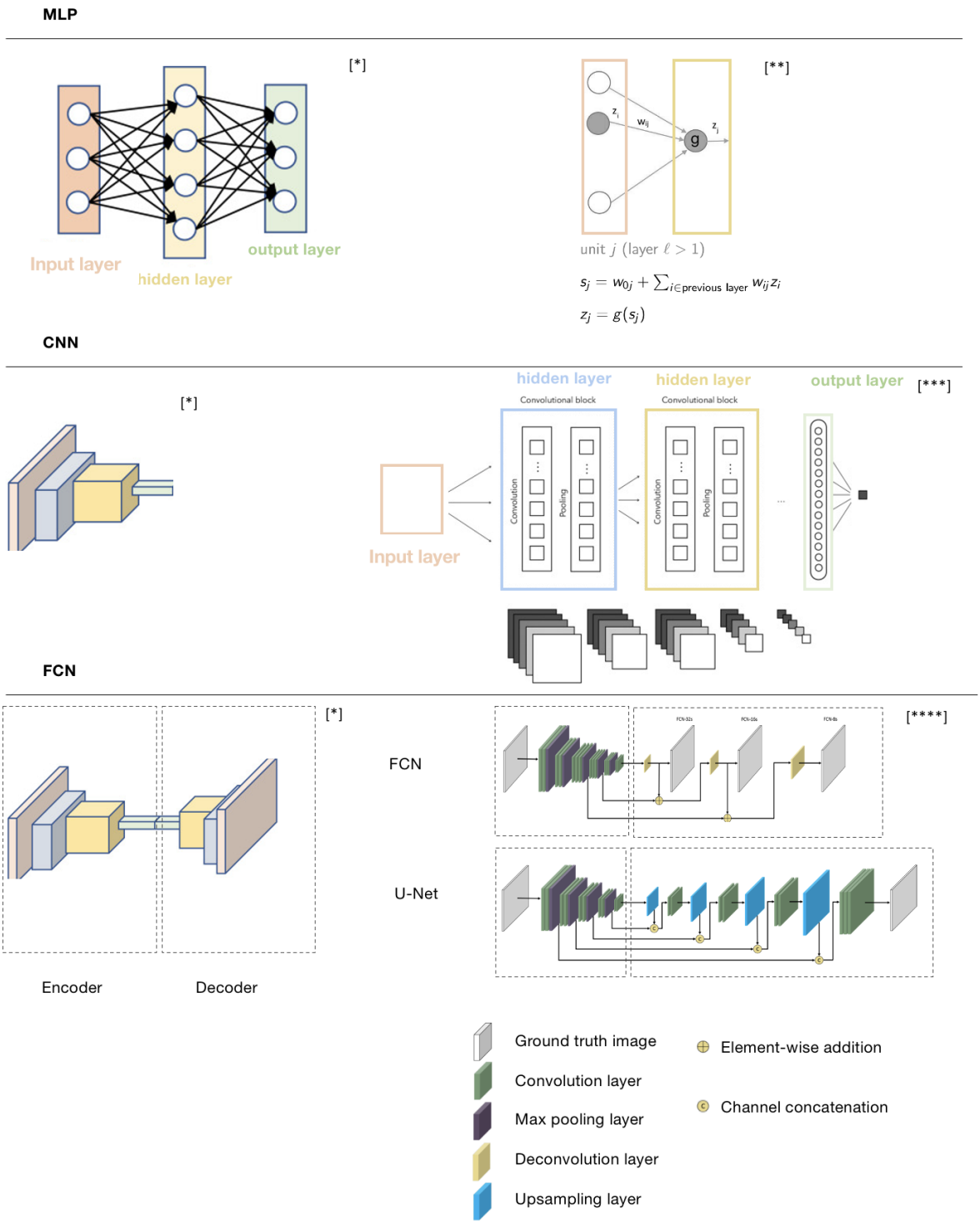


Figure 2.2: (**Top.**) MLP, (**Middle.**) CNN, and (**Bottom.**) FCN schematic architectures. This Figure was adapted from [*] [50], [**] [54], [***] [56] and [****] [57].

- **CNN**: The CNN [53] was introduced by LeCun, Bengio and Hinton (Figure 2.2 - Middle). CNN draw their strengths from the convolution operation from which arise useful properties that enable CNN to achieve a highly efficient representation of the input data without suffering from MLP restrictions and drawbacks [3]. The aforementioned properties (sparse connections, parameter sharing, and equivariant representation to translation) are presented in Figure 2.3 along with the differences between MLP and CNN.

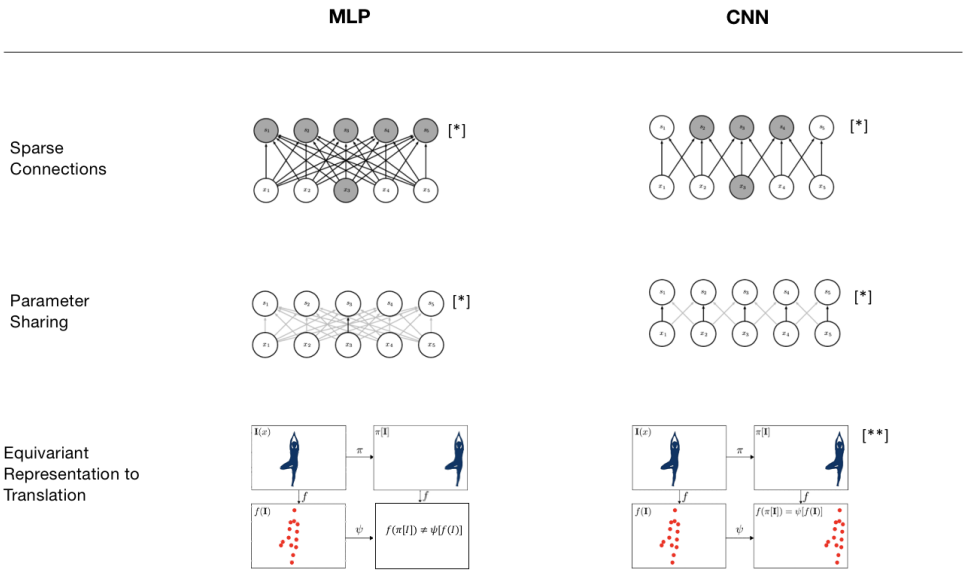


Figure 2.3: (**Left.**) MLP and (**Right.**) CNN approaches with respect to sparse connections, parameter sharing and equivariant representation to translation. The present figure exposes the advantages of CNN framework: it reduces the number of parameters to be learnt during the training, the computational burden and presents an efficient approach for image classification. This figure was adapted from [*] [3] and [**] [58].

Sparse connections and parameter sharing are essential ingredients when it comes to decrease the number of parameters w to be learnt and stored while achieving highly efficient representation of the input data. While in a MLP, all units i of layer I are connected all units j of layer J, sparse connections create a local receptive field connectivity framework, reducing the connections between units of consecutive layers. As the correlation between pixels decreases when the distance between them increases, sparse connections do not deteriorate the quality of segmentation [59]. While in a MLP, each weight w_{ij} is used exactly once in the output layer computation, parameter sharing authorizes the network to share the weights through all the units of the same layer. Sparse connection and parameter sharing are then combined with equivariant representation to translation that allows the network to be insensitive to translations. Therefore, CNN are viable alternative to MLP to process large input dimensions.

CNN architectural design comprises a convolutional layers, non-linear activation layers, pooling layers, and a final fully connected layer:

- *Convolutional layers* use convolution of the input image with a determined filter to extract feature maps.

- *Activation layers* answer the need of introducing non-linearity into neural networks to allow for nonlinear mappings representation between input and output [55]. Among the non-linear activation layers, rectified linear unit (ReLU) and Leaky ReLU [2] are known to increase training speed by reducing second order effects and producing large and consistent gradients.

- *Pooling layers* enhance the computational efficiency by downsampling the obtained feature map and reducing the size of the input fed into the next layer. By reporting a summary statistic for clusters of neighbour pixels, pooling layers also prevent overfitting.

- *The final fully connected layer* and its respective activation function computes the output.

The loss of spatial information with the increase in the network depth may decrease CNN performance for segmentation tasks. Hence, to perform segmentation with a CNN, a patch-wise classification is usually adopted, where the CNN classifies the center pixel of the patch. The image is fully covered by sliding the patch throughout all the pixels. Therefore, while CNN are the gold standard for classification tasks, they become less efficient when it comes to image segmentation [50].

- **FCN**: The FCN [60] was proposed by Long and Shelhamer to overcome the loss of spatial information noticed in CNN. The architecture of a FCN is portrayed as the combination of an encoder, that ensure feature extraction and dimension reduction, and a decoder, that performs semantic segmentation [50]. This architecture enables the network to learn relevant features (size- , shape- , texture- related) and to combine the aforementioned contextual information with a precise spatial location. It consists of replacing the final fully connected layer by a succession of deconvolutional layers. Via up-sampling of the low-resolution feature maps and bi-linear interpolation, the original dimension is restored and a segmentation map of the input image is created. Nevertheless, this up-sampling operation limits the resolution of the segmentation map and produces unsatisfactory and coarse results. To refine spatial location, Long and Shelhamer introduced skip connections that combine coarse information of higher layers with fine information of lower layers (Figure 2.2 - Bottom). It allows to recover fine-grained spatial information that is potentially lost in the pooling and downsampling layers.

  Among FCN architectures, emphasis is given to U-Net, one of the most commonly adopted architectures for medical image segmentation [51]. Its name arises from its symmetric U-shaped architecture, as shown in Figure 2.4. All U-Nets operate with a very common configuration of two convolutional blocks per layer. Each block consists of a 3x3 unpadded convolution, followed by batch normalization and a ReLU non-linearity [52]. Combining context (encoder) with location (de-

[2] ReLU uses the activation function $g(z) = max\{-0, z\}$ based on the principle that not all the neurons/units are activated simultaneously. On examples where the activation is zero, the training via gradient based methods is compromised. Therefore, Leaky ReLU was introduced, with a negative slope of 0.01 for $z \in [-\infty, 0[$ to answer the aforementioned issue [3].

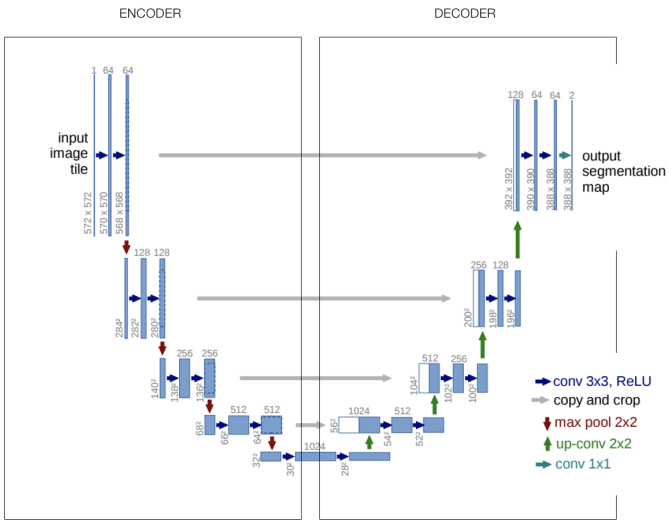


Figure 2.4: General U-Net architecture. The U shape is visible: encoder (contextual information) and decoder (spatial location) are identified.
Blue boxes - multi-channels feature maps. The number of channels (z) is mentioned on the top of the box and in-plane dimensions (x-y) on the lower left edge of the box.
White boxes - copied feature maps.
Figure adapted from [52].

coder) is enabled by incorporating the aforementioned blocks with downsampling and upsampling operations, respectively. In the contracting path, each block is followed by a downsampling step, computed with 2x2 max pooling operation. In the expansive path, each block is preceded by an upsampling step, performed by a 2x2 up-convolution and a concatenation with the correspondingly cropped feature map from the contracting path [52]. This last step takes up the idea of skip connections implemented by Long and Shelhamer, apart from the fact that information of the encoder is added via a concatenation operator and not via element-wise addition. This guarantees that better segmentations are produced and the learning process is refined.

U-net is the building block upon which *Apollo* and *nnU-Net* models are built. Their respective specificities will be detailed in Chapter 3.

2.1.3 The Performance Measure P

Assessing the performance can be conducted during the validation step, to monitor hyper-parameter choices and select the optimal solution, and during the testing step, to judge the generalization ability of the model when evaluated on an unseen dataset [61]. In this work, performance analysis falls into the latter category and is defined as being an objective, indirect and empirical evaluation of an algorithm [43]. Unsupervised evaluations, for which no ground truth is provided, are not reported nor discussed.

Over the past years, there has been an evolution in metrics, adopted in the evaluation frame-

works [43]. Selecting the right metric requires a deep understanding of the algorithm specific applications and the context for which it is designed. Regarding segmentation, the evaluation of an algorithm can be conducted at different levels (pixel-wise, lesion-wise, image-wise) and the evaluation framework needs to be designed accordingly. Based on its application, one could privilege the number of segments, its area or volume , its alignment with ground truth, or its density [62]. Additionally, for segmentation tasks, an equilibrium has to be found between over and under-estimating lesions. Generally, for medical applications, it is common to maximize recall, guaranteeing no lesions is being missed during segmentation, on the cost of precision [62].

It is also important to take into consideration class unbalance or outliers. As a matter of fact, medical data suffers from class unbalance at an image and lesion levels. Most of the available cases correspond to healthy patients, and within the unhealthy patients, the pathological segments are small compared with the background. Higher accuracy for the dominating class will overshadow the lower accuracy associated with the other class, thus providing biased results. Moreover, medical data is mainly noisy (partial volume effect), of low resolution, and presents artifacts. Therefore, metrics sensitive to outliers should be avoided [62].

Figure 2.5 summarizes the adequate evaluation metrics for pathology segmentation with respect to the previously mentioned features. The metrics were reduced to area and alignment measures and are computed based on the confusion matrix (Figure 2.6), defined at pixel, lesion, or image level.

2.2 State of the Art

This section provides an overview of the current trends in supervised DL models applied to medical image segmentation, and discusses the main challenges of transferring these approaches to the clinical setting. The main purpose of this section is to draw a bridge between the theoretical and the empirical settings. Understanding how the theoretical knowledge is applied in practice and what hampers the transition from research to clinic is essential to justify the experimental path described in Chapter 3.

2.2.1 DL at the Core of Healthcare (R)evolution

DL applications in clinical settings have been flourishing over the past years and state-of-the-arts for medical segmentation are being updated regularly [20]. A review conducted by [64] shows that innovations in model architecture accounts for 36 % of the contributions of the recent medical image segmentation papers. The remaining contributions do not exceed 18 % and include optimization of loss function, weak supervision, multi-task models, and data augmentation methods. Regarding the current trends in model architecture, the encoder-decoder architecture with skip connections (*i.e* U-Net based networks) has been given a position of honour with respect to medical segmentation. The top 15

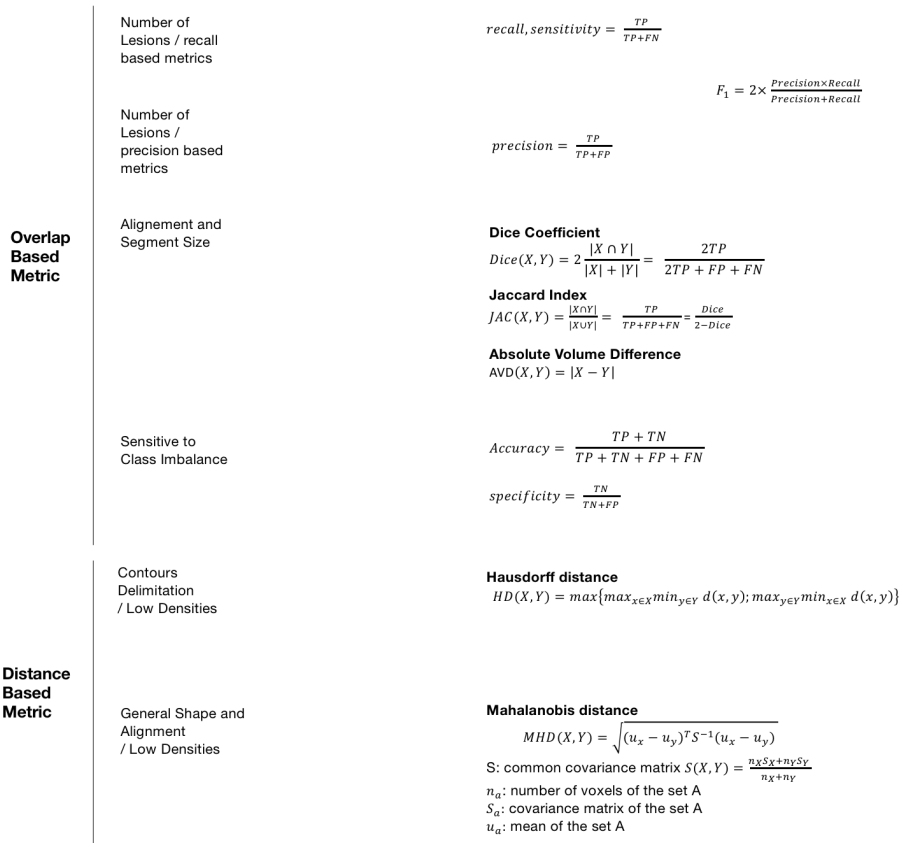


Figure 2.5: (**Top.**) spatial overlap and (**Bottom.**) spatial distance based metrics for medical segmentation evaluation. Metrics were selected according to the specific application and the relevance of the following features: recall, precision, alignment and segment size, contours delimitation, and general shape and alignment. Metrics sensitive to class imbalance are also mentioned, as most of the medical dataset suffers from unequal class repartition.
X is the predicted lesion segmented and *Y* is its correspondent ground truth. *TP* (True Positive), *TN* (True Negative), *FN* (False Negative), and *FP* (False Positive) are drawn from the confusion matrix (Figure 2.6).
This figure was designed based on [62] and [63].

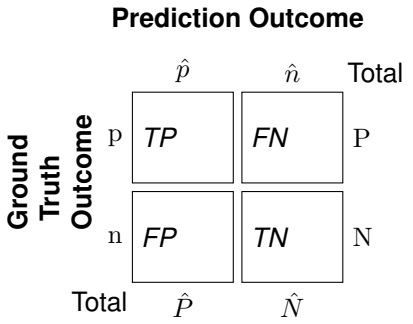


Figure 2.6: Confusion matrix for a binary classification problem where p stands for a positive and n for negative case. The matrix compares the ground truth y with its respective prediction $\hat{y}$. For a positive ground truth, the model can either predict a *TP* (True Positive) or *FP* (False Positive) case. In turn, a negative ground truth can result in a *TN* (True Negative) or *FN* (False Negative) prediction.

DL-based segmentation methods of 2019 KiTs Challenge [65] [3] are a clear evidence of U-Net based models' superiority. Current research encompasses optimizing the amount of data being transferred through skip connections, modifying the upsampling operation from the lower resolution input feature maps, or adapting the U-Net like encoder-decoder skeleton [4] [64]. The two models used in this work fall in the latter category, as detailedly explained in Sections 3.1 and 3.2, respectively.

An outstanding observation when addressing architectural modifications is that performance is not always correlated with innovative architectural designs and complex architectural modifications [49]. Achieving state-of-the-art mostly relies on design choices made during network configuration, where the pipeline fingerprint is designed and parameters are selected. Therefore, identical architectures seem to cover the entire range of evaluation scores (KiTs Challenge) [65]. Nevertheless, network configuration and architectural extensions are not universal across domains and modalities [49]. Method configuration requires dedicate adaptation to each dataset. This supplementary step is even more important than innovative architectural modifications. The aforementioned observation constitute the core and foundation of Isensee *et al.* framework [49], the current status quo in medical segmentation tasks. A simple U-Net is tuned for each dataset, based on heuristic rules applied to extracted data fingerprints [49].

Another key aspect regarding the network architecture is the number of channels. Current algorithms tend to rely on multi-modalities. This scenario allows the network to aggregate more contextual information from different image modalities of the same subject. Hence, it yields a higher quality segmentation and predictive accuracy [66]. Unfortunately, accessing different modalities is sometimes cumbersome as it increases acquisition time, costs, and may affect the hospital workflow [67]. Additional image registration is usually required to provide accurate location of the combined information [66]. Modalities used to feed an algorithm developed for the diagnosis of a pathology have to be in line with the protocols used in hospitals. Besides guaranteeing that the modality is suitable for detecting that pathology, it confirms that the algorithm is answering a valid clinical problem. Moreover, it ensure that data are more likely to be available for training. As highlighted in Chapter 1, CT can be used in a broad range of pathology diagnosis, making it the preferred modality for DL applications. However, due to its versatility, MRI has been recently rising in popularity [50]. In a single MRI session, it is possible to acquire several modalities, through an adequate selection of the scanner parameters within the same "scanning session". Numerous challenges for pathological segmentation with multi-modal MRI scans have been flourishing, underlying the potential of this technique [5].

[3] KiTs stands for *Kidney and Kidney Tumor Segmentation*. The Challenge is hosted by the Medical Image Computing and Computer Assisted Intervention (MICCAI) society.

[4] This can be performed by adding of residual connection or attention blocks, cascading , ensembling , or combining adversarial networks with U-Net based networks [64].

[5] Ischemic Stroke Lesion Segmentation (ISLES) for sub-acute ischemic stroke lesions segmentation and acute stroke clinical outcome; BraTS for glioma segmentation, prediction of patients survival rates, and uncertainty evaluation of the prediction maps; the RSNA intracranial hemorrhage detection for diagnosis and classification of intracranial hemorrhages sub-types.

2.2.2 The Challenges of the Research-to-Clinic Transition

While automated medical segmentation is an active field of research, few solutions are commercialized or in a pilot phase at clinical sites. Several Conformité Européenne (CE)- or Food and Drugs Administration (FDA)-approved artificial intelligence (AI) solutions have been put forward in the last few years for brain pathologies segmentation. Most of them use CT, usually targeting strokes and hemorrhages. A brief selection of companies leading the DL Radiology Revolution for infarcts, tumors and hemorrhages detection using MRI and CT was conducted and presented in Table 2.1. Integrating ML algorithms in clinical workflow is hampered by the lack of: 1) **standardization** in hospital ecosystems, 2) **explainability**, 3) **generalization** assessment, and 4) **evaluation guidelines**.

The success of DL models does not solely rely on its accuracy. It also depends on the performance and impact of the DL solution in real clinical settings. The diversity of radiology platforms, the heterogeneity of processes, formats, and protocols, the variability of intra and inter-site scanner manufacturers make the integration of DL challenging in hospital workflows [46]. Prospective studies evaluating the AI solutions in real clinical settings are slowly being adopted by research groups and companies [68]. They are designed to give insights on how the algorithm responds *in situ* and multiply the metrics used to assess clinical effectiveness. Nevertheless, a standardization of the frameworks and the implementation of incentives for culture change in routine clinical pratice are still missing for a seamless integration of AI in medical ecosystems [48].

Moreover, most of DL solutions are black-box algorithms, capable of drawing hidden relations and faster conclusions than radiologists. From their intrinsic nature, DL solutions have low degree of explainability and the system predictions can only be judged as correct based on the final outcome [4]. The decision-making process or prediction confidence is not usually monitored. Therefore, the reliability of their predictions has been questioned [50] and the liability aspect potentially harnesses the integration of DL into existing clinical workflow [47].

Concerns have also been raised on the ability of the algorithm to generalize to different clinical sites and to be widespread in clinical settings (Section 1.2.3). In current practice, and due to the scarsity of medical data, most of the evaluation processes are not sufficiently broad. A study on performance evaluation of AI algorithms for medical classification shows that only 6% of the 516 reviewed solutions performed external validation, and so far, there is limited research demonstrating the generalizability of these algorithms to widespread clinical practice [48]. Moreover, when performed, the external evaluation is unsuitable for biases detection as the selected dataset present significant overlaps with the ones used for training and validating the models [65]. However, the inherent dependence of the network to the training set and the unintended data bias during the optimization step should be taken more seriously. Training data may not accurately represent the entire population introducing a selection bias based on demographic or acquisition parameters [4]. In [69], a systematic evaluation of the effect of scanner pa-

Table 2.1: Landscape of DL-based companies towards radiology modernization. Pathological context, modalities, ML framework, main clinical outcomes, clearance information and penetration stage in the market are highlighted. Clinical outcomes are similar and comprise faster and more accurate clinical solutions.

Company	Pathology	Modality	Company creation	Technology Name	Machine Learning framework and tasks	Clinical applications	Clearance	Implementation In Hospitals
Cerebriu	Stroke Tumor Hemorrhage	MRI (DWI, SWI, FLAIR, T2 * GRE)	2018 Denmark	Apollo Brain	Deep Learning U-Net Disease segmentation and classification	Real-time decision support during examination Triage Attention-based diagnostics support.	CE-marked FDA current application	Not commercialized Current validation in 14 hospitals
Cercare Medical	Stroke	MRI (PWI, DWI)	Denmark 2013	Cercare Stroke	Deep Learning Estimating the mismatch volume and ratio based on MRI for Stroke lesions quantifications Fully automated identification of core and hypoperfusion lesions	Attention-based diagnostics support and quantitative measurements Decision support in acute ischemic stroke management Easier assessment of treatment eligibility Predicted outcome with stent or thrombolysis	CE marked	Commercialized in Europe Integrated with Siemens Healthineers, Spectra, Wellbeing Software , etc
Biomind	Hemorrhages Tumors Cerebral Small Vessel disease (CSVD)	MRI CT (only for hemorrhage)	China 2017	Biomind	Deep learning Hemorrhages Fast and early detection, location and severity of hemorrhages, and accurate prognosis Tumors Diagnosis of a wide range of brain tumours. CSVD Identification of the presence of CSVD, estimation of the cause and prediction of the risk of stroke.	Attention-based diagnostics support and and quantitative measurements	CE marked China FDA (NMPA) Singapore HAS	Commercialized in Europe and Asia: Hôpitaux Robert Schuman, Beijing Tiantan Hospitals, National Heart Centre
Aidoc	Intracranial Hemorrhage Large Vessel Occlusions	CT	Israel 2016	Aidoc	Deep Learning CNN Disease segmentation and classification	Triage Attention-based diagnostics support	CE marked FDA clearance	Commercialized: +400 global install bases Integrated with GE Healthcare, Philips, FUJIFILM Medical Systems Agfa HealthCare supports, etc
Brainomix	Stroke	CT CTA MRI	Oxford 2010	E-Stroke	Deep Learning CT Estimation of the volume of ischemia Computation of the clinically validated ASPECTS score, to help optimize patient selection for stroke treatments. CTA Status of collaterals estimation Computation of the CTA Collateral Score (CTA-CS). MRI or CTP automation and standardization of the assessment of mismatch and volume Estimation of the volume of ischemia and penumbra Computation of the mismatch ratio to help optimize patient selection for stroke treatments. Detection and measurement of hyperdense regions and vessels, supporting clinicians to find evidence of vessel occlusions and bleeding on simple imaging.	Attention-based diagnostics support and quantitative measurements Decision support in acute ischemic stroke management Easier assessment of treatment eligibility Reduction of contrast and radiation exposure	CE marked FDA pre market notification	Commercialized in more than 25 countries Intergrated with GE Healthcare,Stryker, etc
Zebra medical vision	Intracranial Hemorrhages	CT	Israel 2014	Neuro Solution	Deep Learning CNN (pixelwise prediction - BloodNet) Disease segmentation and classification.	Triage Attention-based diagnostics support	CE marked FDA clearance	Commercialized: Assitance Publique Hôpitaux de Paris, Apollo Hospitals,etc Integrated with Philip Carestream, Jonhson &Jonhson, Wellbeing software, Canon Medical, etc

rameters (manufacturer, magnetic field strength, and slice thickness) on ten extracted features for breast cancer detection demonstrated that 58.22% of the features were affected by the scanner manufacturer, 50.47% by the slice thickness and 38.94% by the field strength. Other types of bias include curation bias (selection of optimal training data, high-quality images) and negative dataset bias (over-representation of positive cases) [4]. A striking example of this phenomenon was the study led by Google Health, assessing the impact of a DL algorithm to detect diabetic retinopathy in real clinical setting [68]. Eleven clinics in Thailand were chosen for the experiment, observing the performance on-site. Consequences of data curation were drastically assessed: trained with high-quality scans, the algorithm responded poorly to the pictures taken by the nurses, rejecting more than one fifth of the images. Instead of achieving the desired outcome, the solution increased frustration and worsened the workflow [68]. All these source of bias contribute for the distributional shift, responsible for a out-of-distribution prediction. Therefore, external evaluation is recommended.

Finally, it is important to adopt adequate evaluation frameworks with appropriate and diversified metrics, intended to cover all requirements needed to be integrated in clinical routine. Nevertheless, the literature review conducted in the present work and presented in Figure 2.7 shows that limited metrics are usually used to evaluate the impact of the algorithm in clinical settings. In medical segmentation tasks, Sørensen Dice coefficient (defined in Figure 2.5) accounts for 40 % of the metrics computed to assess the algorithm performance. Another phenomenon observed during the literature review is that clinically relevant metrics are tendentiously neglected. As an example, Badea *et al* [70] only use accuracy ($accuracy = \frac{TN+TP}{FN+FP+TP+TN}$) to evaluate a model that performs a pixel-wise classification of the degree of dermal burns (normal vs burn) on color and infrared images. However, computing accuracy of normal versus burn segments can be misleading. Normal segments (classified as true negative (TN) or false positive (FP)) are larger when compared with burn segments (classified as true positive (TP) or false negative (FN)). Hence, a high accuracy value is not necessarily correlated with good segmentation of burn areas: high TN can overshadow low TP. Dice coefficient or F_1 measure could be more appropriated.

Therefore, a more complete evaluation framework with a deeper understanding of the performance across metadata properties or pathologies is required to convince stakeholders to adopt and trust DL approaches. Targeting challenges 3) and 4), our key objective is to bring awareness on those issues, to inspire better practices, and provide alternatives to the current evaluation framework of DL models.

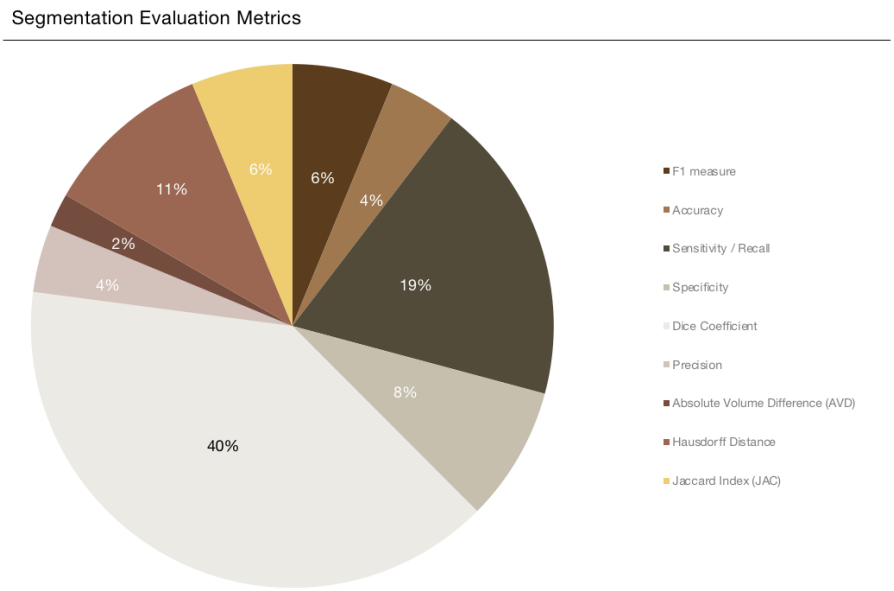


Figure 2.7: Commonly used evaluation metrics for segmentation of tissue types or pathologies in medical images. Dice accounts for 40% of the metrics currently used in medical image analysis.
Data drawn from: [50], [67], [71], [72], [73], [74], [51], [75], [76].

3

Materials and Methods

Contents

Chapter 3 describes our approach that jointly assesses the performance of DL algorithms for classification and segmentation tasks. Through accurate and adequate metrics, the analysis encompasses performance across pathologies, performance as a function of MRI acquisition parameters, and performance on unseen data. Given the context of this internship, performance evaluation is intended to give a deeper understanding of *Apollo* [1] across pathologies, acquisition parameters, and sites. However, to show that the framework is not limited to *Cerebriu*'s test-bed algorithm, the evaluation pathway is extended to the state-of-the-art in biomedical segmentation *nnU-Net* [49].

This chapter starts by explaining the in-house *Apollo* software architecture and its clinical application. Then, a brief description of *nnU-Net*, regarding its architecture and training specificities, is given in Section 3.2. Finally, relevant information about the dataset, adopted post-processing methodologies, and metrics chosen for evaluation are addressed in the remaining sections.

3.1 Apollo

The present section describes the two components of our test-bed algorithm *Apollo* [1]. The DL component, that performs segmentation and classification on MRI scans, and the software component (user interface (UI)) are covered in Section 3.1.1 and Section 3.1.2, respectively. Information about *Apollo* training scheme can be found in Appendix A.1.

3.1.1 *Apollo* Architecture

Apollo is built upon a standard U-Net architecture. In the present case, *Cerebriu* has opted for a symmetric network topology with a four block depth architecture [1]. *Apollo* is composed by three U-Nets, each of them fed with the three 3D MRI sequences described in Section 1.2.1: DWI, FLAIR and T2 * GRE, SWAN, or SWI images. Each network computes a multi label semantic classification and predictions are made on the entire image. Therefore, the expected output is one segmentation map per label/class per network. Four classes are considered: class 1 - infarcts (label 1 (L1)), class 2 - tumors (label 2 (L2)), class 3 - hemorrhages (label 3 (L3)), class 0 being considered as background. Post-processing techniques are then applied to project each segmentation map back to its native modality space [2]. As a result, the final prediction of Apollo, for a specific label is a majority voting between the three predictions of the disease, aligned in its native space, originated by the three different networks. In other words, for label k, the prediction $\hat{y}$ follows equation (3.1):

$$\hat{y}^k = mode\left\{\hat{y}^1_k, \hat{y}^2_k, \hat{y}^3_k\right\} \tag{3.1}$$

[1] Block composition is described in Section 2.1. The only difference with respect to block composition is that instance normalization is performed instead of batch normalization.

[2] Each pathology has a respective space of prediction, mirroring the clinical pratice for diagnosis: DWI, FLAIR and SWI/SWAN/T2 * GRE spaces are used to predicts infarcts, tumors and hemorrhages, respectively. Therefore, the native modality space of an infarct would be the DWI space.

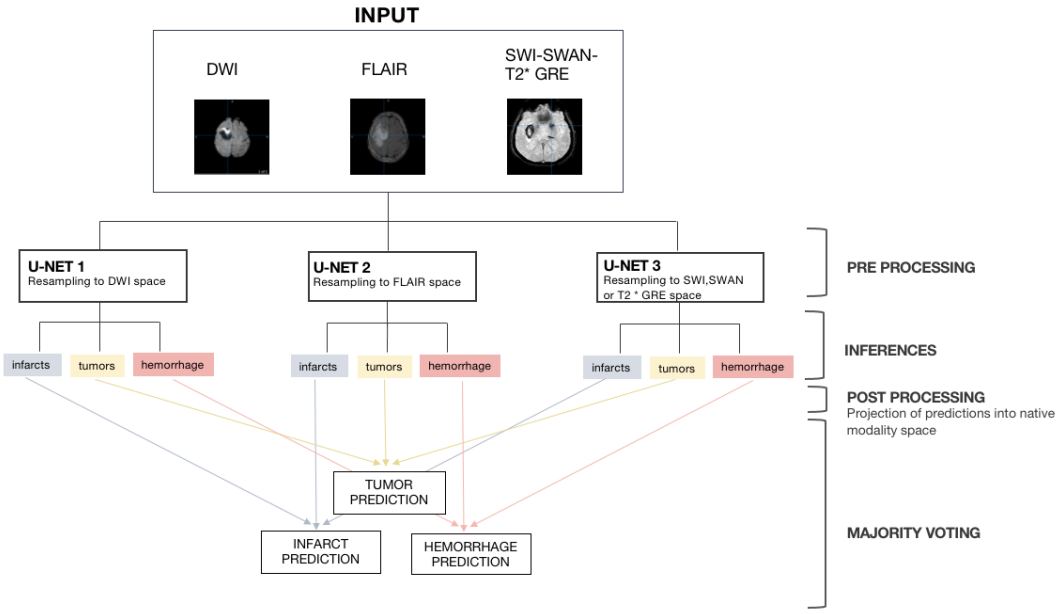


Figure 3.1: *Apollo* inferences process.

where y_k^j with $j = 1, 2, 3$ is the prediction originated by a network j for label k. A visual explanation of the aforementioned inference procedure can be appreciate in Figure 3.1.

In clinical settings, the three networks run in parallel and the final predicted maps (after ensembling) are expected two minutes after the end of the acquisition step, depending on the scanner and the acquisition protocol. Based on these maps , if at least one lesion of a specific class has been segmented, the patient is automatically classified with the same label.

3.1.2 *Apollo*'s Software: User Interface and Key Features

Apollo application in radiology department can be narrowed down to three relevant features, shown in Figure 3.2:

- **Smart Protocol**: Providing protocol decision support during image acquisition, the main objective of this feature is to reduce MRI scan time and unnecessary acquisition. This translates into 1) an improved use of resources (adequate selection of MRI sequences to acquire/ reduction of reexaminations) and 2) a higher quality patient care (reduction of scanning time / prevention of contrast administration). The suggested sequences are based on clinical findings and can be manually configured, according to the hospital protocol.
- **Triage Advisory**: Granting triage decision support during image acquisition, the main objective of this feature is to improve quality care by automatically selecting patients that require urgent review by the radiologist . This results in 1) optimized patient flow, 2) improved use of clinical resources, and 3) avoidance of unnecessary admissions . Triage relies on clinical findings and

the prioritization is set by the hospital according to its respective classification criteria. Triage is updated real time while patients are in the scanning room.

- **Image Reporter**: Allowing the visualization of regions of interest during scan and later in picture archiving and communication (PAC), the main objective of this feature is to verify pathology and accelerate reporting. This results in 1) decreased reliance on specialist radiology services and 2) an improved support to junior medical staff in the diagnostic process.

3.2 *nnU-Net*

nnU-Net is an open-source network developed by Fabian Isensee *et.al.* in 2019 [77]. The proposed framework was able to outperform the most specialized DL pipelines in 19 public international competitions, setting the new state of the art in 33 out of the 53 tasks [49].

3.2.1 *nn-UNet* Architecture

nnU-Net stands for "no new net". The network does not draw its strengths from an improved architecture, a more efficient training scheme or a more appropriate loss function. Its novelty, highlighted in Figure 3.3, resides on its ability to handle a wide disparity of structures and image properties, proposing a tailor-made network without any user intervention. Pre-processing, architecture design, training scheme and post-processing are automatically configured. It by-passes the traditional iterative trial and error and reduces the need of expertise in the ML field when it comes to network design. In addition, *nnU-Net* data efficiency is all the more appreciable for medical applications. Having extracted its encoding design choices from a large and diverse data pool, *nnU-Net*'s performance is not deteriorated by data scarcity.

The network design is interpreted in terms of a data fingerprint and a pipeline fingerprint. Data fingerprints compile the properties of a specific dataset. Pipeline fingerprints summarize the design choices of a segmentation network and are divided into three groups: blueprint (data-independent, already predefined), inferred (data dependent), and empirical parameters (optimized during training). Network optimization and adequacy to the respective dataset rely on heuristic rules, operating on the data fingerprints, to compute the inferred pipeline fingerprints. These are key decisions required to transfer a basic architecture to the actual dataset and segmentation tasks. Blueprint parameters are then combined to identify a high quality pipeline fingerprint for the studied dataset. Three configurations can be chosen for the design of the neural network: a 2D U-Net, a 3D U-Net (full resolution), and a 3D U-Net cascade that creates a refinement on a first low resolution network. Optimal configuration and post-processing choices are assessed post-training, setting the configuration of the empirical parameters (Figure 3.3). Detailed information about fingerprints can be found in Appendix A.2.

Figure 3.2: *Apollo* key features [1].
(**Top.**) **Smart Protocol**. In the figure, protocol A is advised based on clinical findings inferred on DWI and FLAIR images.
(**Middle.**) **Triage Advisor**. In the figure, a color code (low - green, middle - orange, high - red) can be seen on the patient list, according to the prioritizing code set by the Hospital.
(**Bottom.**) **Image reporter**. In the figure, an infarct can be appreciated in the DWI image.

Following the already described U-Net architecture, the depth is configured by determining the number of downsampling operations along each axis depending on the patch size and voxel spacing. In the present case, the depth of the network is set to six blocks. Block composition is slightly different from the standard U-Net: ReLU is replaced by Leaky ReLU, batch normalization by instance normalization, maxpooling by stride convolution, and the upsampling with a 2x2 convolution by transposed convolution [49].

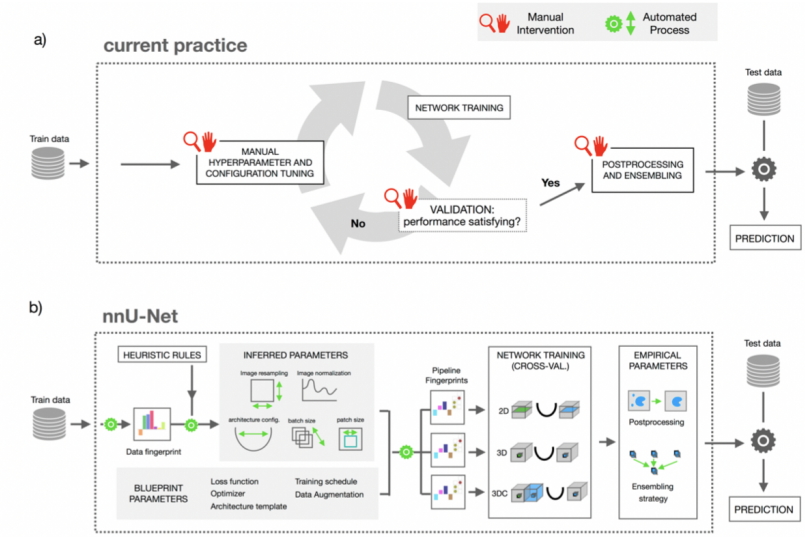


Figure 3.3: Manual and proposed automated configurations of DL method.
a) Current practice of configuring a DL method for biomedical segmentation: iterative trial and error process. In training and validation steps, hyper-parameters, network architectures and topologies are manually set. This procedure is, often, time consuming, and requires accute expertise in the ML field. The optimal architecture usually needs a re-optimization for each new dataset (new modalities, new pathologies...).
b) Proposed automated configuration by *nnU-Net*: Automated DL solution adapted to the dataset. A tailor-made network, fitting the present dataset, specially designed for its image properties, labels and segmentation tasks, is obtained, without manual intervention, based on the identification of robust design decisions and explicitly models key interdependencies.
Image and descriptions are adapted from [49].

Inferences are achieved with a sliding window with the same patch size as used during training [3]. Adjacent predictions overlap by half the size of a patch; voxels located at central window positions have higher weights in the softmax aggregation.

Table 3.1 summarizes *Apollo* and *nnU-Net* pipelines and aforementioned design choices.

3.2.2 *nnU-Net* Training Specificities

Since the network has to be fed with exactly the same dataset used in *Apollo* training, we only trained the 3D *nnU-Net* configuration for fairer and unbiased comparisons [4]. As an additional support to this decision, 3D U-Net was also found to be the best performing method in [49].

To mimic *Apollo*, three 3D U-Net networks were trained. As in *Apollo*, each network has to perform five segmentation tasks: infarcts, tumors, hemorrhages and background. The multi label semantic segmentations originated by each network are then combined by majority voting to obtain the final segmentation maps, exactly as shown for *Apollo* in Figure 3.1.

[3] Similarly to *Apollo*, *nnU-Net* prioritizes large patch sizes, under a given GPU memory constraint, over the batch size.
[4] 3D U-Net cascade was not considered as not suitable for the training dataset. Cascade is only triggered for datasets where the patch size of the 3D full resolution U-Net covers less than 12.5% of the median image shape, reducing the possibility to aggregate sufficient contextual information for optimal training [49].

Table 3.1: Comparative analysis of *Apollo* and *nnU-Net* pipelines

	Apollo	**nnU-Net**
Modification to the U-Net architecture		
Normalization	Instance Normalization	Instance Normalization
Activation Function	ReLU	Leaky ReLU
Downsampling	Max pooling	Stride Convolution
Upsampling	Upsampling	Transposed Convolution
Depth	4 blocks	6 blocks
Training Schedule		
Epochs	250 (iterating over 500 minibatches)	1000 (1 epoch iterating over 250 minibatches)
Patch Size	[160,176,160]	[128 128 128]
Batch Size	2	2
Back propagation		
Algorithm	Adam Optimization	Stochastic Gradient Descent Nesterov Momentum (μ = 0.99)
Learning Rate	η = 0.0001	η = 0.01
Early Stopping	Yes	-
Loss Function	Dice Loss	Cross-entropy and Dice Loss (excluding the background)
Deep supervision	-	Yes - Loss Function is optimized for the 4 last blocks of the decoder
Image pre-processing	Z-scoring Intensity Normalization Oversampling Foreground Regions	Z-scoring Intensity Normalization Oversampling Foreground Regions
Data Augmentation	Rotation ; Translations; Intensity Variation; Random Cropping	Rotation ; Translations; Intensity Variation (Gaussian Blur and Gaussian Noise) Gamma correction; Mirroring ; Scaling; Low Resolution Simulation;
Predictions	On the entire image	With a Gaussian sliding window overlap: half of the patch size

3.3 Data

The present section is devoted to the data used in the project. It is intended to describe the datasets used in the experiments and to give further information on their MRI acquisition parameters, extracted from the image header.

Infarcts, tumors, and hemorrhages are referred as L1, L2, and L3.

3.3.1 Dataset Information

The data on which are based our experiments and analysis come from three hospitals: OUH (Odense University Hospital - Odense, Denmark), MedAll (MedAll Diagnostics - Chennai, India), and SUNY (Suny Upstate University Hospital - New York, United States). Data was curated and annotated before use: one patient can have a multiple labels. Training and validation steps were conducted in 853 and 214 patients, respectively, withdrawn from OUH and MedAll datasets [5]. No information regarding the scanner manufacturer was recovered from the training data. In turn, performance evaluation is conducted in two datasets:

[5] Training data: 94.26 % MedAll and 5.74 % OUH; Validation data: 94.39 % MedAll and 5.61 % OUH.

- an **in-house** dataset - 195 MedAll patients, selected among the validation set (*i.e* 91.12% of the validation set), with 36 % of normal cases. 58.97 % of the scans were acquired with a 1.5 T GE Healthcare scanner [6].

- an **external** dataset - 62 SUNY patients, with no healthy patients. Data was acquired with 1.5 T scanners from different providers (50.00 % Philips, 46.77 % Siemens, and 3.23 % GE Healthcare).

The aforementioned datasets are characterized in Table 3.2 and Figure 3.4. Table 3.2 shows the repartition of patients between classes and describe some lesion attributes, its number, and size. Figure 3.4 presents the spectrum of disease sub-types across datasets (including a comparison with the training set).

Table 3.2: Group description: distribution of infarcts (L1), tumors (L2) and hemorrhages (L3) lesions across in-house and external datasets. $\overline{N_{lesions}}$ and $\overline{A_{lesions}}$ refer to the mean number of lesion and the mean lesion size in voxels.

	L1		L2		L3	
Training set % of unhealthy patients	25,67		11,37		9,96	
Testing set % of unhealthy patients	**In-house**	**external**	**In-house**	**external**	**In-house**	**external**
	25.64	58.06	8.72	19.35	12.31	24.19
$\overline{N_{lesions}}$	180	195	120	52	253	352
$\overline{N_{lesions}}$ / $patient$	3.60	5.42	7.06	5.20	10.54	23.46
$\overline{\mathcal{A}_{lesions}}$ (voxels)	398.83	461.24	6776.38	2114.96	1926.63	1995.69
$\overline{\mathcal{A}_{lesions}}$ / $patient$ (voxels)	793.41	813.60	11301.25	1503.15	3956.58	2016.98

As previously mentioned, the in-house dataset comes from the validation set and shares similar characteristics with the training set (same clinical sites). In contrast, the external dataset is composed of unseen data, acquired with different scanning parameters. Therefore, performance is expected to be optimistic for the in-house evaluation while a slight drop is predicted for the external evaluation.

3.3.2 MRI Acquisition Parameters

Data can be found in a DICOM or neuroimaging informatics technology initiative (NIFTI) format [7]. Following the header description, relevant MRI acquisition parameters are identified in the in-house dataset, as listed below:

- **Scanner type:** The scanner vendor.

[6] When digital imaging and communications in medicine (DICOM) format is not available, no information on scanner manufacturer and on B_0 can be recovered.

[7] While DICOM format allows access to a more complete set of MRI acquisition parameters, NIFTI format lacks information about TR, B_0, scanner vendor, and flip angle.

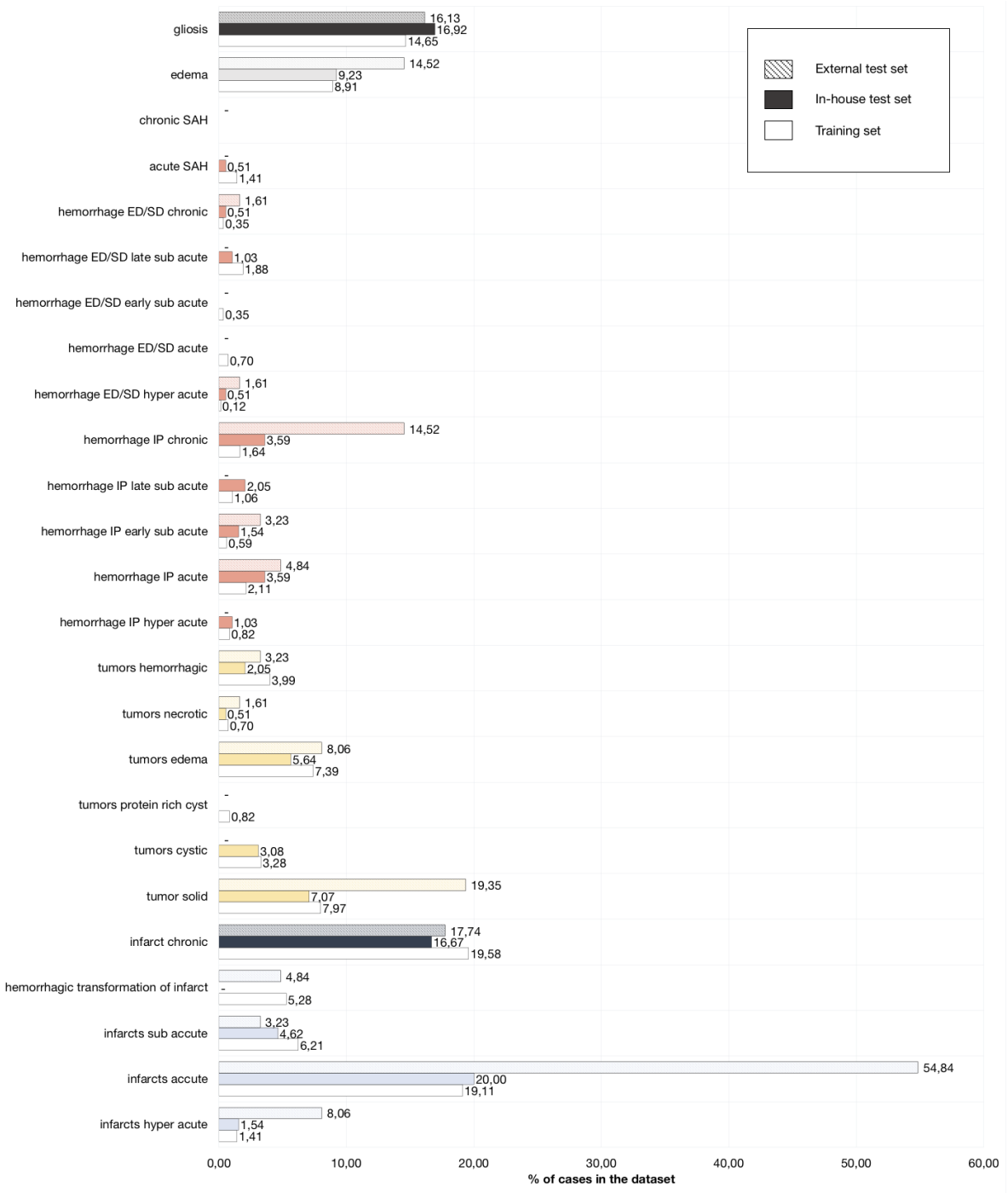


Figure 3.4: Percentage of labels sub-types in the training set (white), in-house testing set (coloured), and external testing set (coloured with dashed line). The percentage of disease k in the dataset j are given by $P_{k,j} = \frac{Np_{k,j}}{Np_j}$, where $Np_{k,j}$ is the number of patient presenting the disease k in the dataset j and Np_j is the total number of patients in the dataset j.

- **The Magnetic field strength** B_0 (in *Tesla*).
- **Acquisition parameters:** TR and TE in *seconds*; flip angle θ in (°).

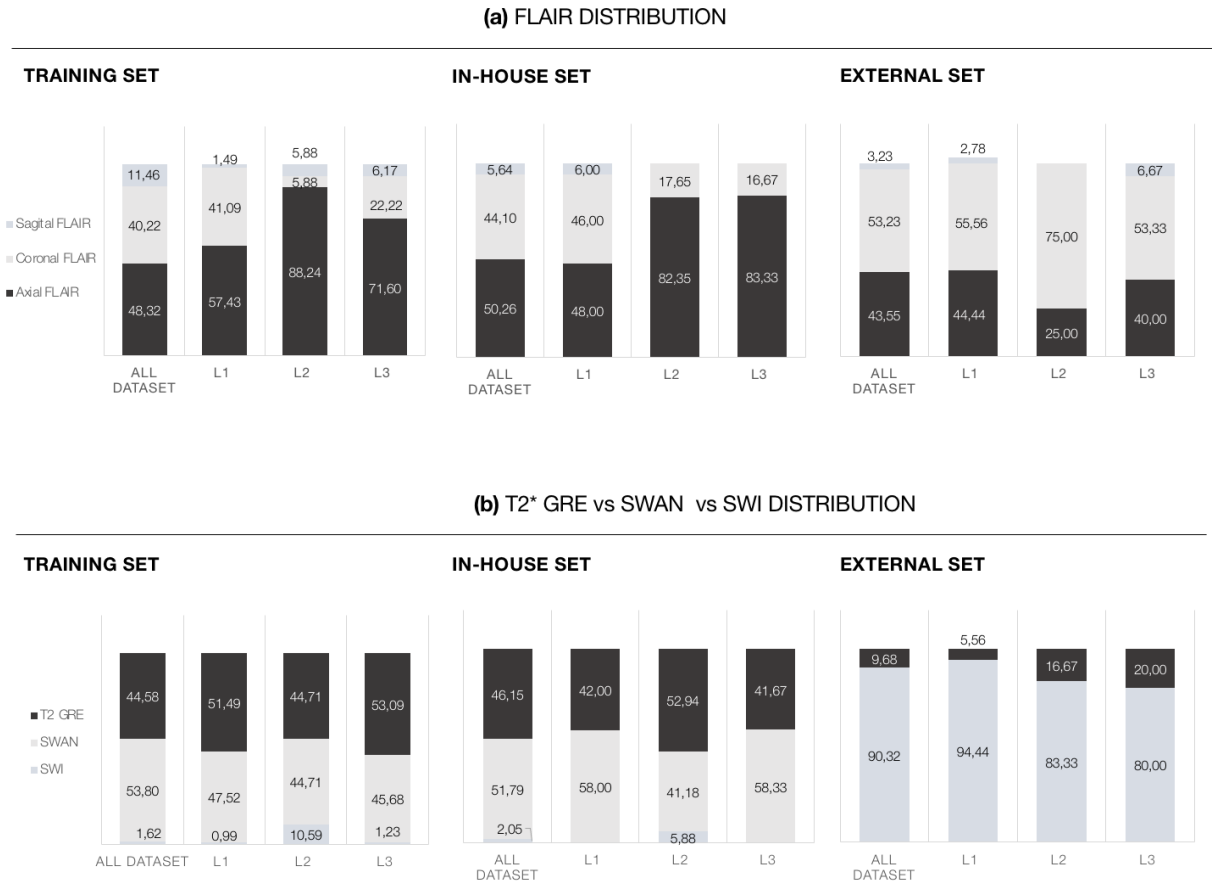


Figure 3.5: Distribution of the acquisition parameters in (**Left.**) the training set, (**Middle.**) the in-house set, and (**Right.**) the external set.(**a.**)FLAIR distribution: Axial, Coronal, and Sagittal ;(**b.**)T2 * GRE, SWI, andSWAN distribution.

- **Sequence type and orientation:** The type of sequences acquired and its orientation in space (axial, coronal or sagittal).
- **Voxel dimension:** Voxel size in the 3D space: x, y and z (in *mm*). It gives an estimation of the image resolution and the impact of partial volume effect in the analysis.

As explained in section 3.1, *Apollo* requires a DWI, FLAIR and SWI, SWAN, or T2 * GRE to make inferences on the data provided. Therefore, for each patient belonging to the datasets only the acquisition parameters of the aforementioned sequences are considered.

The evaluation contemplates how *Apollo* and *nnU-Net* perform in detecting and localizing tumors, infarcts, and hemorrhages and how this performance is affected by MRI acquisition parameters. For a matter of conciseness, this analysis is exclusive to the in-house dataset. The selected clinical parameters are the **type of sequences and orientation** (T2 * GRE, SWI, or SWAN and axial, coronal, or sagittal), as described in Appendix B. Figure 3.5 shows the distribution of Axial, Sagittal, and Coronal

FLAIR (**a**) and the distribution of T2 * GRE, SWI, and SWAN (**b**), in the training, in-house, and external datasets. All the DWI were acquired with Axial orientation.

Regarding sequence orientation, Sagittal FLAIR is not considered in the analysis: along with the fact that it is a 3D acquisition while the other two FLAIR acquisitions are 2D, it is poorly represented in the training and testing sets. In **Axial FLAIR** versus **Coronal FLAIR**, due to the testing set imbalance, performance comparisons are only performed for infarct predictions. Infarcts patients are better distributed between sequence orientations in the in-house and training dataset. The same cannot be said for hemorrhages and tumor patients: Coronal FLAIR only includes 17.65 and 16.67 % of the tumor and hemorrhages patients , *i.e* three and four patients, respectively.

Regarding the type of sequences, SWI and SWAN are encompassed in the same group [8]. In **T2 * GRE** and **SWAN/SWI**, performance comparisons only considers hemorrhage predictions. It is believed that these sequences have a higher contribution, compared with DWI or FLAIR, when it comes to hemorrhage predictions [41, 78] [9]. Training and testing sets are also balanced across sequence types for L3.

The compared groups are characterized in Table 3.3 and Figure 3.6.

Table 3.3: Group Descriptions: (**a.**)FLAIR distribution: Axial *versus* Coronal ;(**b.**)T2 * GRE *versus* SWI /SWAN. $\overline{N_{lesions}}$ and $\overline{A_{lesions}}$ refer to the mean number of lesion and the mean lesion size in voxels.

(a) Group AXIAL / CORONAL FLAIR

		AXIAL FLAIR	CORONAL FLAIR
Total number of patients (in-house set)		98	86
% of patients (in-house/ Training)		50.26 / 48.32	44.10 / 40.22
L1 patients	% of patients (in-house/ Training)	24.49 / 46.15	26.74 / 53.85
	$\overline{N_{lesions}}$	51	74
	$\overline{N_{lesions}}$ / $patient$	2.13	3.22
	$\overline{\mathcal{A}_{lesions}}$ (voxels)	386.27	696.84
	$\overline{\mathcal{A}_{lesions}}$ / $patient$ (voxels)	352	1897,94

(b) Group T2* GRE / SWI-SWAN

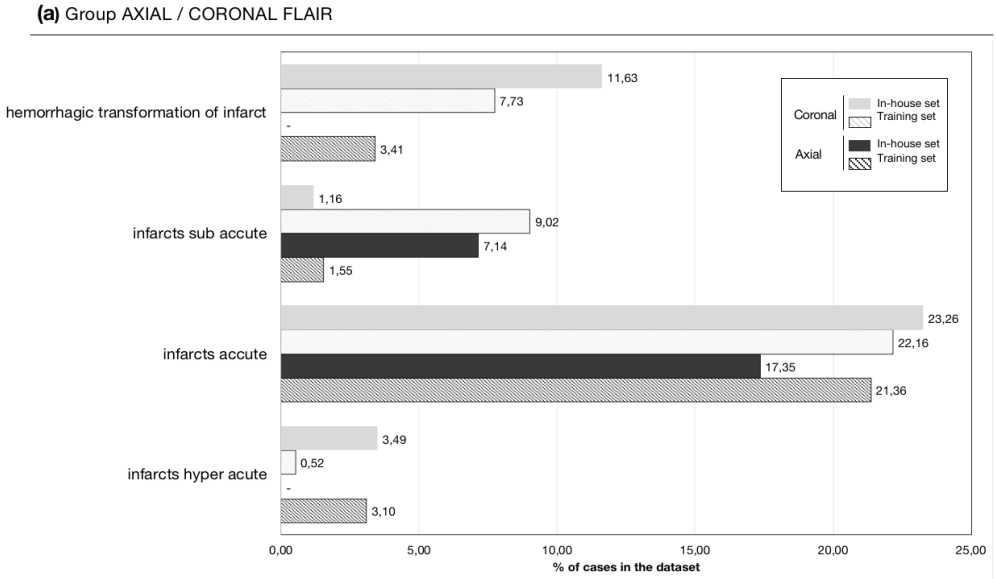
		T2 * GRE	SWI / SWAN
Total number of patients (in-house set)		90	105
% of patients (in-house/ Training)		46.15 / 44.58	53.85 / 55.42
L3 patients	% of patients (in-house/ Training)	11.11 / 29.05	13.33 / 22.02
	$\overline{N_{lesions}}$	40	93
	$\overline{N_{lesions}}$ / $patient$	4.0	6,64
	$\overline{\mathcal{A}_{lesions}}$ (voxels)	1088,675	4768,8
	$\overline{\mathcal{A}_{lesions}}$ / $patient$ (voxels)	1797,03	10623,7

3.4 Post-processing of Predictions and Ground Truth Binary Masks

Post-processing on prediction and ground truth binary masks has to be performed after *Apollo* and *nnU-Net* inferences. Post processing methods are based on **morphological dilation** with a determined coefficient $N_{dilation}$ and **filtering** of lesion areas below a certain threshold A_{HP}. Dilation is performed after filtering. $N_{dilation}$ and A_{HP} are defined for *Apollo* using the in-house dataset and are kept throughout all the analysis. Dilation and filtering are expected to help increase the fairness of the evaluation

[8] Only four patients show a SWI in the in-house dataset.

[9] Infarcts and tumor diagnosis rely only on SWI, SWAN or T2 * GRE to gather additional information about surrounding hemorrhages according to *Cerebriu* annotators.



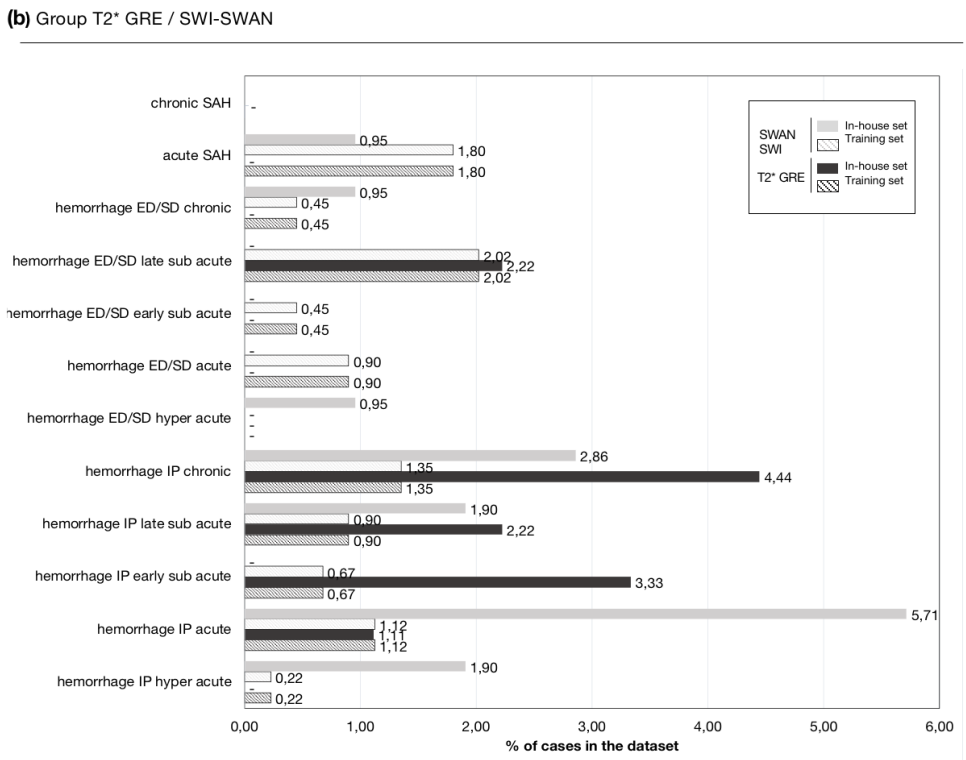


Figure 3.6: Sub-labels distribution. The percentage of disease k in the subset j are given by $P_{k,j} = \frac{Np_{k,j}}{Np_j}$, where $Np_{k,j}$ is the number of patient presenting the disease k in the subset j and Np_j is the total number of patients in the subset j.

and to reflect the real performance of *Apollo*. Dilation allows overlapping of spatially close lesions on both prediction and ground truth whereas filtering removes noisy voxels on prediction and ground truth. Motivation for dilation and filtering can be found in Figures 3.8 and 3.7, respectively.

It is important to bear in mind that post-processing methods have to be undertaken carefully: parameters have to be calibrated rigorously to avoid the production of misleading or distorted results. A visual inference of the dilation and filtering iterative experiment can be seen in Figure 3.9.

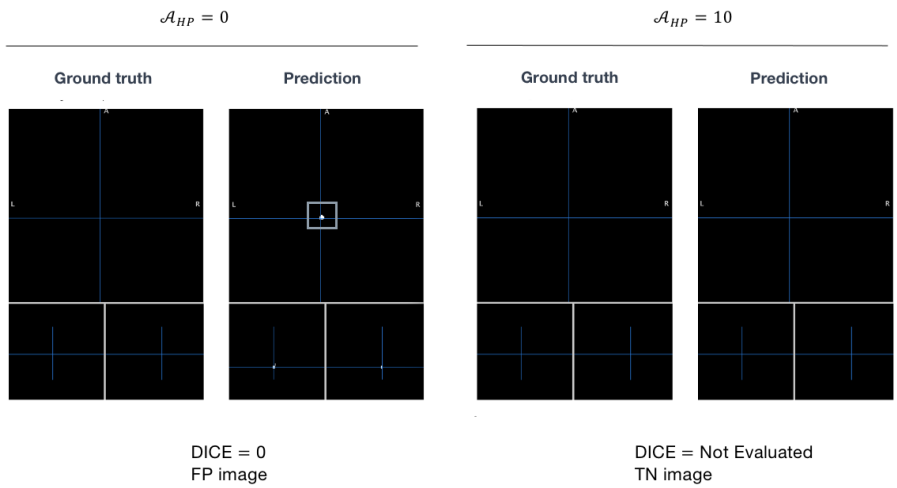


Figure 3.7: Example of hemorrhage inferences with (**Left.**) no filtering and (**Right.**) $A_{HP} = 10$. Ground truth was annotated in the SWI space.
The left prediction mask accounts as a FP image in the confusion matrix due to the noisy voxels of size inferior to 10 voxels; the right prediction mask accounts as a TN image. Small predictions in masks should be removed to reflect a fairer performance evaluation.

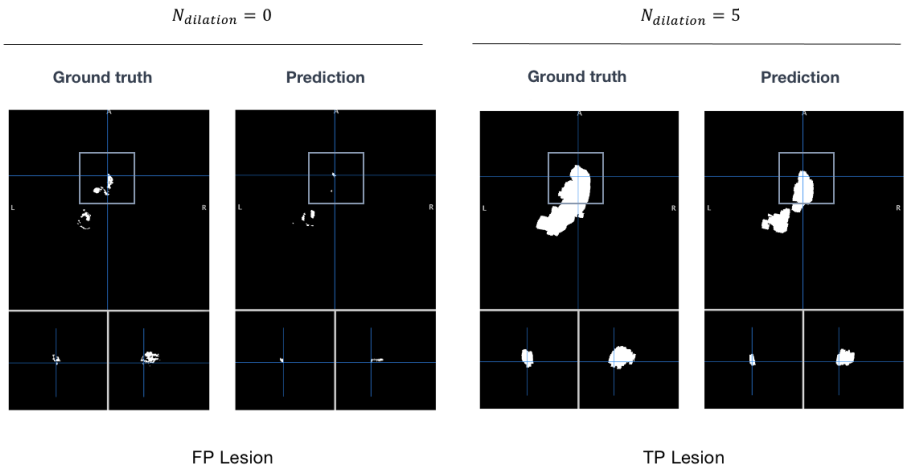


Figure 3.8: Example of hemorrhage inferences with (**Left.**) no dilation and (**Right.**) $N_{dilation} = 5$. Ground truth was annotated in the SWI space.
Prediction and ground truth do not overlap, even though a lesion was predicted by the software. Dilation enables lesions to be considered TP if only a slight deviation from ground truth location is noticed. Performing a mathematical dilation of binary masks could reflect a fairer performance evaluation.

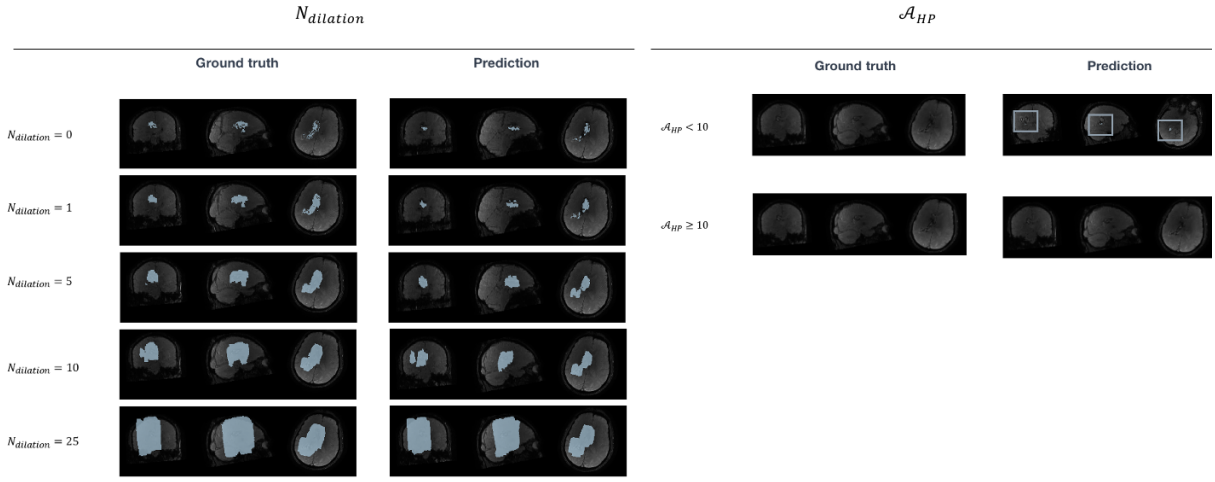


Figure 3.9: 3D hemorrhage ground truth and prediction masks on the original SWI image. (**Left.**) Masks are dilated 1, 5, 10 and 25 times, represented by the variable $N_{dilation}$ and (**Right.**) lesions with a size inferior to H_{HP} are filtered.

3.4.1 Filtering of Small Lesion Areas

Filtering of small lesion areas in ground truth and prediction binary masks is also undertaken. Areas under a defined threshold, referred as A_{HP} are set to zero. Different A_{HP} of 5, 10 , 25 , 50 , 100 voxels are tested alternatively. Ground truth and predictions are filtered simultaneously: small areas are found in both segmentations (Table 3.4).

The optimal thresholding size is set taking into consideration the general performance along with its credibility. Moreover, as True Positive (TP) lesions show higher mean and median size than False Positive (FP) and False Negative (FN) lesions, filtering procedure is anticipated to be beneficial (Table. 3.5). As FP sizes are greater, across labels, than FN sizes, a more significant decrease in FN lesions is expected. Box plots of lesion size distribution across labels are depicted in Appendix C.1 (Figure C.1).

It should be mentioned that filtering is not part of product design pipeline nor considered in a clinical setting. Ideally, *Apollo* should detected lesions under any threshold area. However, in the in-house dataset, annotations were conducted in a way that pathologies are sometimes represented by isolated one-voxel lesions. Therefore, filtering appears as a good compromise to counterbalance this issue.

Table 3.4: Ground truth versus predicted lesions sizes in voxels. Mean and median per label are given. As noticed, median size are relatively small, indicating the presence of small areas annotated and predicted.

	Lesion sizes (voxels)			
	mean		**median**	
	ground truth	prediction	ground truth	prediction
L1	613.47	457.91	26.00	22.50
L2	6776.38	2223.66	4.00	21.00
L3	1926.63	1636.82	13.00	30.50

Table 3.5: TP versus FP and FN lesion sizes in voxels. Mean and median per label are given. As noticed, mean and median of TP lesions is higher than FP and FN. As a note, TP lesions are drawn from predicted segmentations.

	Lesions size (voxels)					
	mean			**median**		
	TP	FP	FN	TP	FP	FN
L1	619	231	15	25	15	9
L2	15570	783	109	166	19	4
L3	2584	745	136	33	24	9

3.4.2 Dilation of Lesion Areas and Bounding Box Experiment

An iterative multi-dimensional binary dilation, defined by a coefficient $N_{dilation}$ is performed 1, 5, 10 and 25 times for ground truths and predictions . A 3D cubic structural element is generated to perform dilation uniformly in 3D space and applied to the ground truth and predicted segmentation maps. It allows for close lesions to be grouped into a single one.

The optimal number of dilation iteration is set taking into consideration the general performance along with its credibility. In that context, evaluation is expected to be more flexible, regarding spatial location of corresponding lesions in predicted and ground truth maps.

Comparison between dilation with optimal coefficient and bounding boxes is also made. Bounding boxes are drawn around each lesion, followed by a 5 times *morphological dilation* and *closing*. The experiment is conducted as a response to the high number of predicted lesions compared to ground truth (for L1 and L2). This discrepancy could be explained by an abnormal behavior of the network that tends to split a ground truth lesion in several prediction lesions (Appendix C.3, Figure C.4). In this context, bounding boxes are expected to revert this situation in a more efficient way than a simple dilation. A visual inference of the aforementioned comparison can be found in Appendix C.3, Figure C.5.

3.5 Implementation

The present section highlights the metrics selected for evaluation (Section 3.5.1) and the steps followed in the statistical validation (Section 3.5.2). Hardware and software information are specified at the end of this section.

Performance is assessed for three classes (L1,L2, and L3) and the evaluation framework is validated on *Apollo* and *nnU-Net*.

3.5.1 Evaluation Metrics

Defining the correct metrics for performance evaluation implies considering the intentions and objectives of the software in the radiology department, as detailed in Section. 3.1. Metrics referred in the following section are mathematically described in Figure 2.5.

Under the designed evaluation framework, the multi-class problem is transformed into a binary problem where each label is analysed independently. A joint analysis, at an image (classification) and lesion (segmentation) levels, is conducted. At an image level, **confusion matrix** (Figure 2.6), **sensitivity** in % , and **specificity** in % are selected. At a lesion level, **confusion matrix** (without TN lesions, as considered non-existent), **recall** in %, and **precision** in % are preferred [10]

In this work, the confusion matrix at a lesion level is calculated with Dice coefficient. Results at an image level are built upon lesion level outcomes.

Means of **Sørensen Dice coefficient** and **Hausdorff distance** per image in *mm* are also computed to have a spatial score between predictions and ground truth. The Hausdorff distance is implemented in 3D across all predicted lesions and correspondent ground truths for a specific label. The Hausdorff distances obtained across individual lesions are then averaged. Hausdorff distance is deeply affected by outliers. However, the filtering post-processing of predictions and ground truths is expected to minimize this effect.

Across metrics, higher scores are preferred with the exception of Hausdorff distance for which a smaller distance indicates a higher quality segmentation [63].

An additional experiment is also undertaken in the in-house dataset to understand the common errors of the model for a given ground truth lesion. It consists of a multi-class analysis that assesses if the network is making confusions between labels. The experiment is based on the assumption that tumors, hemorrhage and infarcts cannot be physiologically present at the same location of the brain. Conducted on the existing ground truth images, presenting at least one lesion, results are displayed under the form of a confusion matrix. A similar experiment is also made at the image level.

3.5.2 Statistical Validation

The statistical validation of the evaluation procedure encompasses the computation of **confidence intervals (CI)** for the selected metrics and **hypothesis testing**.

- **Confidence Intervals**

In statistic inference, parameters of interest of a population θ are estimated via its computed value (*i.e* statistics $\widehat{\theta}$) on the observed sample. The most commonly employed method of estimation is via CI. In probabilistic terms, the computed CI refers to the interval in which the true value of the parameter θ is expected to fall with a probability of α.

In this work, CI are computed for all metrics with the bias corrected and accelerated (BCa) bootstrapping method [79] [11] and $\alpha = 0.95$. A total of $B = 10000$ resampling with replacement are made from

[10] Recall and sensitivity correspond to the same metric. The two different nomenclature is adopted to differentiate between lesion and image levels of analysis.

[11] According to *Efron* and *Tibshirani*, BCa were specially designed to achieve a reasonable performance across a broader range of statistics and distributions, drawing smaller interval length and guaranteeing a higher algorithm numerical stability [80].

the empirical probability distribution of the data under analysis, as suggested in [80]. For convergence purposes, the sample size (*i.e* the number of patients) is set to the original sample size of the dataset. As the original sample size is larger than 30, bootstrapping results are expected to be reliable [80].

- **Hypothesis testing**

Results are also validated statistically for all metrics to ensure non-randomness via hypothesis testing. The alternative hypothesis is non-directional and tests are two-tailed. Significance level is set to 0.05: the null hypothesis is rejected when $\rho < 0.05$. For each analysis, a minimum of ten patients is required and parametric and its correspondent nonparametric tests are computed simultaneously [12]. Statistical tests are conducted on two groups at the time, as discussed below. A group represents a sample (*i.e* a set of patients) drawn from a population. The number of groups and its characteristics (dependent/independent) determine the correct parametric and non-parametric tests to select.

Regarding *comparisons between nnU-Net and Apollo*, the selected groups are dependent: tests were paired via specific patient scans, allowing for an implicit normalization for all the non-paired factors. As a consequence, a pairwise t-test and its homologous Wilcoxon signed-rank test are selected for statistical validation [25] (Test 17 and 18) [13]. Regarding *comparisons between MRI acquisition parameters* and *comparisons between datasets*, groups are independent. A t-test for two independent samples and its peer nonparametric Wilcoxon rank-sum test are considered to be appropriated [25] (Test 11 and 12).

3.5.3 Hardware and Software Specifications

Experiments were conducted in a Linux virtual machine, Ubuntu SMP, with a x86_64 processor. The required central processing unit (CPU) is an Intel(R) Xeon(R) CPU E5-2690 v4, with a capacity of 3.5 GHz, a width of 64 $bits$, a clock of 100 MHz, and a size of 2.6 GHz. The graphics processing unit (GPU) is a NVIDIA Tesla P40 with a memory of 22919 MiB.

nnU-Net is trained using one GPU (along with a strong CPU) and Pytorch (version 1.6). *nnU-Net* and *Apollo* evaluation and post-processing scripts are run on Python, version 3 [82]. Specific libraries are enumerated below:

- **Nibabel** (used to read and work with NIFTI files)

[12] The distinction depends on the assumptions made by the tests. Parametric tests assumes normal distribution of the data and variance homogeneity between groups, while the homologous nonparametric tests do not require normal distributions. When these assumptions are met, parametric tests provide more powerful results, being associated with lower Type II error rate (false null hypothesis retention). However, on small number of samples, the assumptions of a parametric tests are more likely to be violated and parametric tests may lead to erroneous results, inflating the likelihood of committing a Type I error (true null hypothesis rejection). Therefore, it is prudent to compute both parametric and nonparametric, even if, in this work, parametric and nonparametric tests usually lead to convergent conclusions. In the few noticed cases, in agreement with [25], the parametric test was rejecting the null hypothesis, while the nonparametric test was found to be not significant.

[13] According to [81], when comparing two machine learning algorithms with a paired t-test, the minimum number of data sets to guarantee normal distribution should be, at least, 30. Therefore, Wilcoxon signed-rank test should be privileged, as the only assumption made on the distribution is on the symmetry of the difference scores distribution about the median of the population of difference scores.

- **PyDICOM** (used to read and work with DICOM files)
- **Nilearn** (used for plotting purposes)
- **Medpy** (used for Hausdorff distance implementation - *medpy.metric.binary.hd*)
- **Skimage** (used for computations)
- **Scipy** (used for dilation experiments)
- **Pingouin**, **bootstrap_stat** and **Scikit_posthocs** (used for statistical assessment)
- **Pandas** (used to store the results and organize the patients database)

Visual Inferences are made using software Mango [83].

4

Experimental Analysis

Contents

Chapter 4 presents the experimental results. Even though the analysis is centered in *Apollo*, results are compared with *nnU-Net*, the current status quo for biomedical segmentation. Performance is analyzed from three perspectives: performance across pathologies (Section 4.2), performance as a function of clinical parameter (Section 4.3), and ability to generalize to an independent dataset (Section 4.4). As mentioned in Chapter 3, the latter is computed on the external dataset, while the remaining are assessed on the in-house dataset.

This chapter starts by evaluating the impact of the post-processing operations, described in Section 4.1. The optimal parameters, extracted from post-processing experiments, are kept for the remainder of the chapter.

Discussion of the results is made in parallel with their exposition. Pathologies under analysis are infarcts, tumors and hemorrhages, referred as L1, L2, and L3, respectively.

4.1 *Apollo* Post-processing Results

The present section details the selection criteria for the post-processing parameters: the number of times the dilation operation is performed $N_{dilation}$ and the cut-off area under which lesions are high-pass filtered A_{HP}. Dilation and filtering were undertaken independently and their optimal parameters were selected based on the network performance at an image and lesion levels, counterbalanced with the plausibility and realism of the results. Over-performance due to an unreasonable high $N_{dilation}$ or A_{HP} was taken into account.

Regarding **filtering post-processing experiments**: five experiments were conducted with $A_{HP} < 5, 10, 25, 50, 100$. As ground truth and predictions are filtered simultaneously, consequences on the confusion matrix at a lesion level and Dice coefficient are hard to predict and may vary across labels. It depends on each individual lesion size and, for TP lesions, it also relies on the relative sizes between predictions and ground truths. Therefore, behaviours of recall and precision (and consequently of sensitivity and specificity) cannot be extracted as directly as dilation results (Appendix C.1). The only meaningful observation at a lesion level is the overall decrease of the FN lesions (Figure 4.1) which is in line with Section 3.4.

At an image level, the evolution of sensitivity and specificity scores can be seen in Figure 4.2. Starting with L1, higher sensitivity is achieved for A_{HP} under or equal to 10, while specificity rises when A_{HP} increments. For L2, sensitivity remains constant, while higher specificities are obtained as the filtering cut-off area expands. For L3, sensitivity shows a decrease from $A_{HP} > 25$, while specificity slightly increases over the filtering experiment. Larger percentual changes between filtering thresholds in specificity are seen for L1, L2 and L3 at an A_{HP} of 50, 100 and 10, respectively. Again, putting the filtering experiments under the light of reality and credibility, thresholds over 25 were discarded (*i.e* $A_{HP} > 25$),

as being too indulgent for the algorithm evaluation. Therefore, the aforementioned analysis highlights two candidates for optimal filtering: $A_{HP} = 10$ and $A_{HP} = 25$. However, since sensitivity is preferred over specificity and sensitivity for L1 decreases in the interval $A_{HP} \in]10; 25]$, optimal filtering parameter A_{HP} was set to 10.

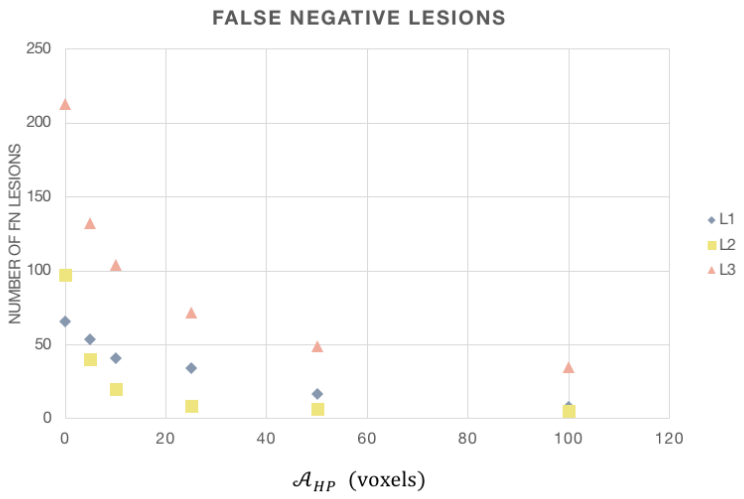


Figure 4.1: *Apollo* performance at a lesion level: FN lesions across labels as a function of the cut-off area under which lesions are high-pass filtered A_{HP} .

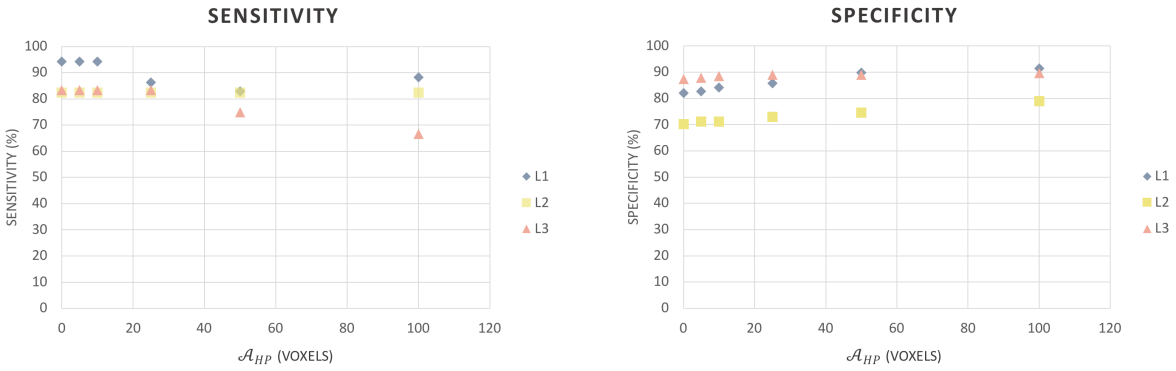


Figure 4.2: *Apollo* performance at an image level: (**Left.**) sensitivity and (**Right.**) specificity in % as a function of the cut-off area under which lesions are high-pass filtered A_{HP}.

Regarding **dilation post-processing experiments**, five different experiments were performed with $N_{dilation} = 0, 1, 5, 10, 25$. Dilation was applied to both predicted and ground truth lesions. Results are reported in Figures 4.3, 4.4, and 4.5. Since ground truth and predicted lesions become larger, expanding their overlapping region, Dice coefficient increases with the number of dilation iterations (Figure 4.3). With the increase of Dice coefficient, the number of TP lesions goes up, while the number of FN and FP lesions goes down (Appendix C.2). The aforementioned results have a direct impact on recall and

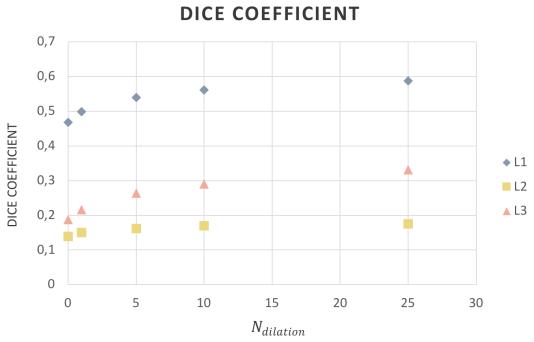


Figure 4.3: Dice Coefficient as a function of the number of dilation iterations $N_{dilation}$. L1 (infarct), L2 (tumors) and L3 (hemorrhages) are represented in blue, yellow and pink, respectively. Dice increases with dilation for all labels.

precision, given in Figure. 4.4. Both metrics increase across dilation iterations and labels with recall showing higher percentual changes than precision. L1, L2 and L3 recall and L1, L2 precision shows higher percentual changes when 5 times dilation are computed.

Nevertheless, changes at a lesion level do not reflect any changes at an image level: sensitivity and specificity remain constant across dilation iteration for L1 and L2. L3 shows an increase in sensitivity from 10 dilation iterations (Figure. 4.5). As a consequence, one could select 10 as the optimal dilation iteration number. A visual inspection deemed the results above 5 dilation iterations to be unrealistic. Therefore, 5 was considered as good compromise between the increase in performance and the visual truthworthiness of the results.

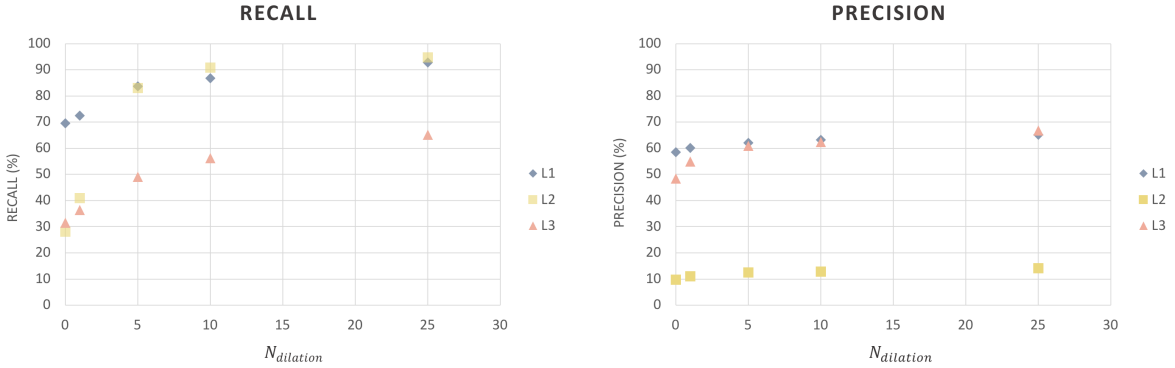


Figure 4.4: *Apollo* performance at a lesion level: (**Left.**) recall and (**Right.**) precision in % as a function of the number of dilation iterations $N_{dilation}$.

Regarding the use of **Bounding Boxes**, poorer results are achieved when compared to the 5 times dilation (Appendix C.3) . The high number of predictions compared to ground truth annotations was not related with the network splitting one ground truth in several predictions, but with the high number of FP lesions. Therefore, it justifies why bounding boxes do not show any improvement in performance.

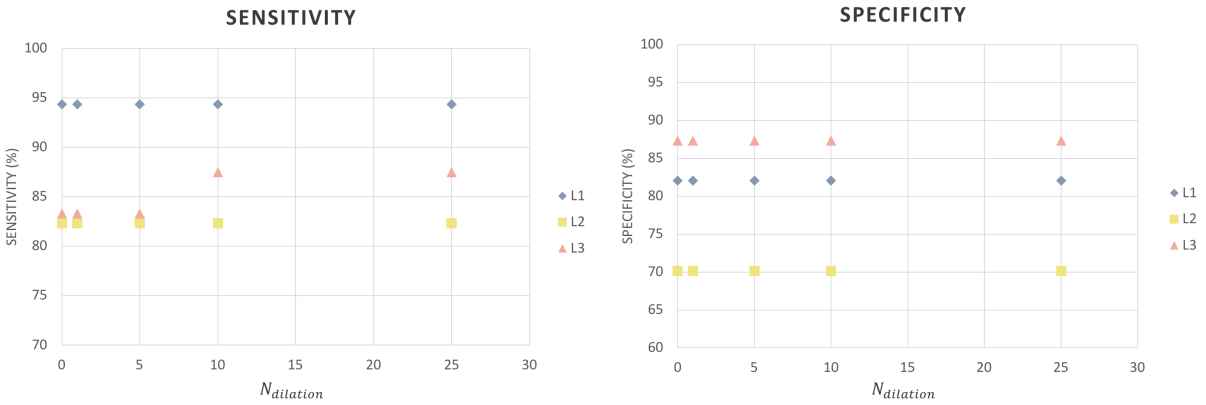


Figure 4.5: *Apollo* performance at an image level: (**Left.**) sensitivity and (**Right.**) specificity in % as a function of the number of dilation iterations $N_{dilation}$.

Figure 4.6 and Table 4.1 summarize the scores of the post-processing. The success of post-processing is clearly observed at lesion level (L2 recall), even though the impact at an image level is reduced. Only a slight improvement in specificity is noticed, as a direct results of filtering. Sensitivity remains constant. Taking the aforementioned into account, post-processing with $N_{dilation} = 5$ and $A_{HP} = 10$ is also performed in all the experiments of Chapter 4, and except for *nnU-Net* L1 sensitivity, an improvement in performance is confirmed across experiments. In addition, by also post-processing the ground truth images, the description of the datasets made in Table 3.2 becomes slightly inaccurate for L2 and L3. For the latter classes, a decrease in the number of lesions and number of lesions per patient is noticeable in both in-house and external datasets. A drop of 69 % in L2 in-house, of 58 % in L2 external, and of 75 % in L3 external is reported in the number of ground truth lesions.

Table 4.1: Post-processing results *versus* no post-processing.

	L1		L2		L3	
	Raw	Post processed	Raw	Post processed	Raw	Post processed
sensitivity (%)	94.34	94.34	82.35	82.35	83.33	83.33
specificity (%)	82.07	84.14	70.17	71.27	87.36	88.51
mean Dice Coefficient	0.468	0.560	0.139	0.167	0.188	0.272
mean Hausdorff Distance (mm)	8.13	27.09	38.58	39.14	33.53	52.23
recall (%)	69.59	82.71	28.15	87.18	31.51	50.64
precision (%)	58.53	63.58	9.74	13.18	48.51	60.31

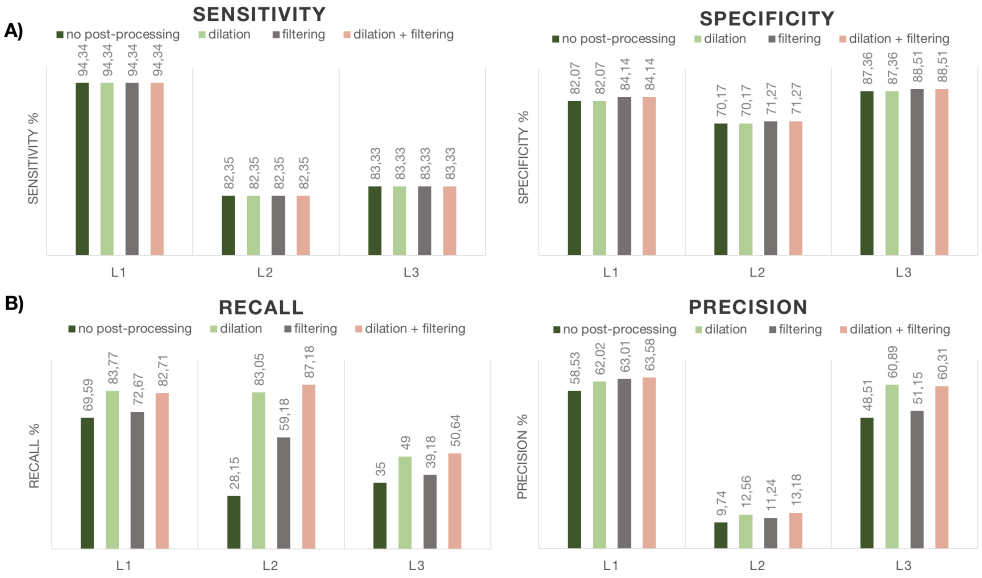


Figure 4.6: Post-Processing results: no post-processing (dark green); dilation (light green); filtering (grey) and dilation and filtering combination (pink).
A) Metrics at an image level: **(Left.)** sensitivity and **(Right.)** specificity in %.
B) Metrics at a lesion level: **(Left.)** recall and **(Right.)** precision in %.

4.2 Performance across Pathologies

The core of this section is the evaluation of *Apollo* performance across pathologies. This perspective of performance is all the more relevant when analyzing a med-tech software with a multi-class segmentation purpose. The aforementioned analysis enables hospitals to select the pathologies for which *Apollo* should be part of the clinical routine.

4.2.1 *Apollo*

Apollo performance is summarized Table 4.2, while Figure 4.7 displays the confusion matrix at an image and lesion levels, providing complementary information to the metrics presented in Table 4.2.

Table 4.2: *Apollo* evaluation for the in-house dataset across pathologies - at an image (sensitivity, specificity, Dice coefficient and Hausdorff distance) and lesion (recall and precision) levels.CI, computed via bootstrap, are presented $[\alpha; \beta]$. Results are obtained after post-processing. Values may slightly differ from Table 4.1 as three patients were removed from the analysis after post-processing. Higher values of performance are highlighted in bold.

	L1	L2	L3
sensitivity (%)	**94.00** [83.33 ; 98.21]	82.35 [54.55 ; 95.24]	83.33 [60.87 ; 95.00]
specificity (%)	84.14 [77.34:89.40]	72.47 [65.32 ; 78.49]	**90.06** [84.77 ; 93.85]
mean Dice Coefficient	**0.544** [0.442 ; 0.638]	0.175 [0.102 ; 0.272]	0.291 [0.193 ; 0.404]
mean Hausdorff Distance (mm)	**25.67** [18.74 ; 34.75]	39.14 [22.61 ; 60.63]	52.23 [37.29 ; 70.33]
recall (%)	80.67 [65.85 ; 90.84]	**87.18** [79.45 ; 97.22]	50.64 [27.33 ; 77.62]
precision (%)	60.38 [44.63 ; 73.61]	14.23 [6.34 ; 29.02]	**64.23** [45.79 ; 80.37]

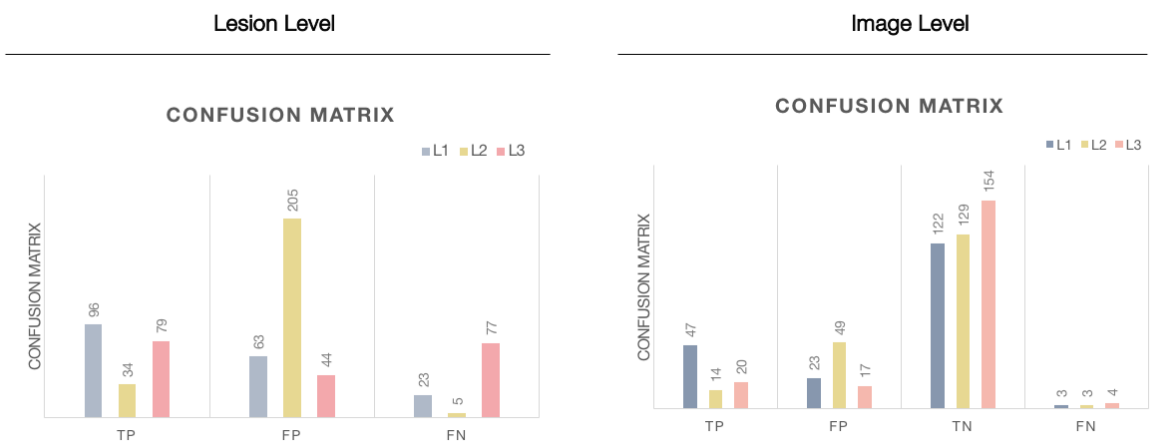


Figure 4.7: Confusion matrix for *Apollo* in the in-house dataset: (**Left.**) lesion level and (**Right.**) image level.

Apollo achieves better performances for L1 than for L2 and L3 and its confusion matrix shows a higher number of TP lesions than FP and FN lesions. It allows L1 to have a more homogeneous behaviour towards recall and precision. While L3 presents a low recall score related with the high number of FN lesions, L2 is characterized by a low precision score explained by the extremely high number of FP lesions. At an image level, discrepancies between L1 and L3 are less pronounced. The high number of FN lesions seems to have a smaller impact in the classification at an image level and sensitivity for L3 is satisfactory. The same cannot be said about L2, for which the high number of FP lesions seems to deeply affect its specificity. Framing this into a clinical context, an infarct patient is more likely to be detected while healthy patients are more likely to be reported as having a tumor.

In terms of the segmentations, L1 presents higher Dice coefficient and lower Hausdorff distance. Infarct segmentations and contours seem to be more similar with their respective ground-truth. Low Dice coefficients are observed for L2 and L3 due to the high number of FP and FN lesions, respectively. L3 is also characterized by a wider Hausdorff distance, 1.93 and 1.33 times larger than L1 and L2 respective distances.

Confidence Intervals (CI) were also computed for each parameter of interest θ. From Table. 4.2, CI appear to be unexpectedly broad and the network scores are highly inconsistent. The high variability of the data can be explained by the small sample size (195 patients). As CI are computed via resampling with replacement, depending on the patients selected in each bootstrap, results can vary drastically. The lesion level is more affected than the image level, and among the metrics, sensitivity for L1 and specificity across labels show higher consistency along bootstraps.

A large disparity is observed in performance across labels. Therefore, an interesting question that arises from these findings is : why does *Apollo* perform better for L1 than for L2 and L3?

A trivial assumption is to explain divergence across labels in terms of their pathophysiological differences. As an example, the high number of FN for hemorrhages could be explained by its intrinsic nature. In general, hemorrhages are characterized by a relatively high amount of spread lesions (shown in Figure 4.8). Therefore, it becomes difficult for the network to detect and segment each of these lesions individually. Among hemorrhages, some sub-types are also more difficult to pick up by the network from their own characteristics. This is the case of SAH: 100% of the acute SAH are missed by *Apollo*.

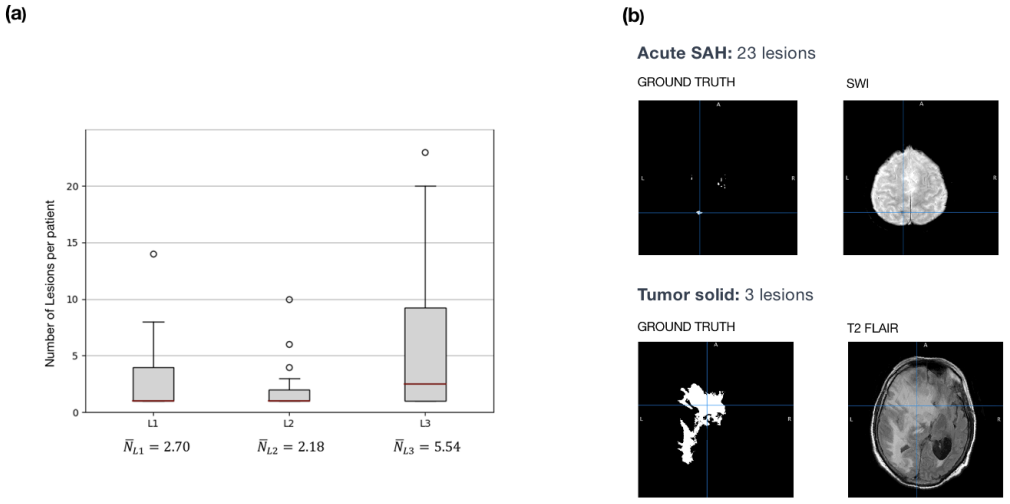


Figure 4.8: Number and size of lesions across labels. **(a)** Distribution of the number of annotated lesions per image across labels. Mean is displayed under the respective boxplot. Mean and median number of ground truth lesions is higher for L3.**(b)** Example of a segmentation mask for a (**Top.**) L3 and (**Bottom.**) L2 patient. Hemorrhages are characterized by higher number of spread lesions across the brain volume.

However, the aforementioned factor cannot explain by itself all the differences observed across pathologies. The higher scores obtained for L1 seem to also stem from the training scheme, most specifically, from the data distribution and class formulation. From Table 3.2, it is possible to observe that 25.67% of the pathologies present in the training set are infarcts, *versus* 11.37% tumors and 9.96% hemorrhages. Therefore, hemorrhage features, for instance, had less opportunity to be learnt and extracted correctly by the network, being a possible explanation to the high number of FN lesions for this label. Diving with more details into the FN lesions corroborates this hypothesis. Poorly represented sub-types in the training set are correlated with higher percentage of missed detections. One example is epidural/subdural chronic hemorrhages: accounting for 0.35% of the training set (from Figure 3.4), 80 % of the lesions present in the in-house dataset are missed [1].

Besides unbalanced class distributions, divergence in performance can also arise from grouping sub-types into three major classes. While infarcts can be divided in sub-acute, acute, and hyper-acute, tumors and hemorrhages show a wider sub-type range (Figure 1.1). Each sub-class is described by

[1] The aforementioned findings are not exclusive to the L3 class. Among the other classes, cystic tumors or hyper-acute infarcts also follow the same pattern.

its own texture, size, and shape, originating a broader landscape of features that needs to be learnt by the algorithm. This makes the classification task more difficult based on this extended diversity. As mentioned in Section 1.2.3, L2 is the most problematic class: its large MRI features lexicon, highlighted in Figure 1.10, seems to have direct consequences in the network performance. As an example, edema tumors seem to share more FLAIR and DWI features with vasogenic edemas than with other tumor sub-types (Figure 4.9). Therefore, it is not surprising that 17.15 % lesions predicted as tumor were annotated as edemas. Similar behaviours are noticed between hemorrhagic tumors and hemorrhages. The network usually detects the hemorrhagic part of the tumor [2] and classifies it as L3 instead of L2.

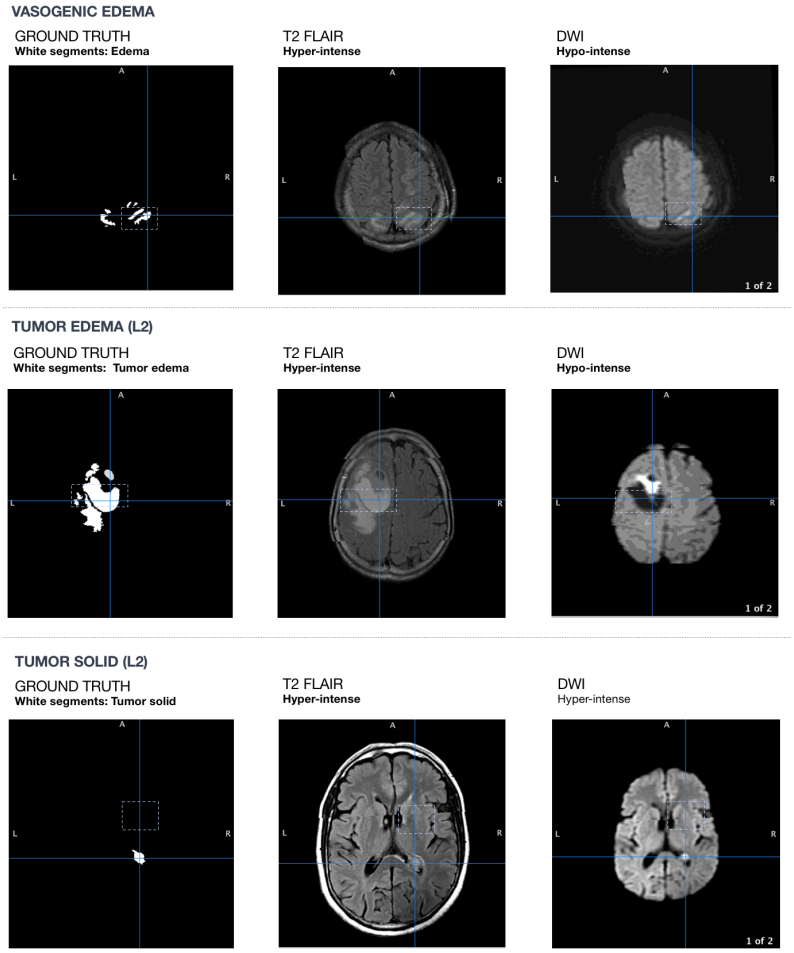


Figure 4.9: Comparison between edema tumors, vasogenic edemas, and solid tumors on (**Middle.**) FLAIR and (**Right.**) images. (**Left.**) Ground truth segmentation for spatial location of pathological areas. Tumor and vasogenic edemas have the same intensity characteristics in FLAIR and DWI without belonging to the same class while solid tumors and edema tumor are both annotated as tumors and do not present the same DWI signal intensity. The high variability in MRI features for L2 may explain the high number of FP observed for this class.

[2] Hemorrhagic characteristics mainly rely on T2 * GRE or SWAN/SWI hypo-intensity. It is thought to be the main reference for *Apollo* to detect hemorrhage.

Another relevant aspects was to determine if the network was confusing labels, *i.e* if it assigns a wrong label to an annotated lesion. A multi-label analysis was conducted, at a lesion level, and its output presented in Table 4.3 (Left) as a confusion matrix. As expected, most lesions are correctly identified (except for L3 that suffers from FN predictions). Confusion between labels is not the main source of FN detections: when missing a lesion, the prediction is usually background. Again, L3 is the label that *Apollo* struggles to associate to the right class. It is clear that the network confounds the various labels, mainly L3. However, do these lesion-wise misclassifications have consequences on the image output? Taking the aforementioned experiments to the image level, a multi-label analysis was conducted and its results presented in Table 4.3 (Right) as a confusion matrix. While for L1 the confusions made with L2 and L3 at a lesion level have no direct impact on the image classification, this is not the case for L2 and L3. However, putting into perspective, even if the percentage of confusions seems high, the actual number of images is low.

Table 4.3: Multi-class analysis - (**Left.**) lesion level and (**Right.**) image level. Confusion matrix in percentage (%). For each ground truth (GT) lesion/image of label i , prediction is assessed and, according to its label type j, lesion is accounted in the matrix $M(i, j)$. In green, the percentage of TP lesions/patients are highlighted.

		PREDICTIONS			
		L0	L1	L2	L3
GT	L1	15.91	82.58	0.76	0.76
	L2	8.11	5.41	86.49	0
	L3	47.41	2.96	8.15	41.48

		PREDICTIONS			
		L0	L1	L2	L3
GT	L1	6.00	94.00	0.00	0.00
	L2	5.88	11.76	82.35	0.00
	L3	8.33	4.17	8.33	83.33

4.2.2 Comparison of *Apollo* and *nnU-Net*

Results of the test-bed algorithm are now analyzed from the perspective of *nnU-Net* performance. In that sense, metrics were averaged across labels for the two networks to provide a higher level analysis. Results are presented in Table 4.4 and Figure 4.10. Supplementary material on the behaviour of *nnU-Net* across pathologies is provided in Appendix D.2. Reference to Appendix will be made during the analysis to support our findings and draw our conclusions.

Table 4.4: *Apollo* (blue) and *nnU-Net* (brown) evaluation across pathologies for the in-house dataset. Results are obtained after post-processing. Metrics are averaged across labels. Higher values of performance are highlighted in bold.

	Apollo	*nnU-Net*
sensitivity (%)	86.40	**91.16**
specificity (%)	81.89	**91.40**
mean Dice Coefficient	0.303	**0.440**
mean Hausdorff Distance (mm)	**37.44**	38.84
recall (%)	**70.88**	66.14
precision (%)	38.07	**60.77**

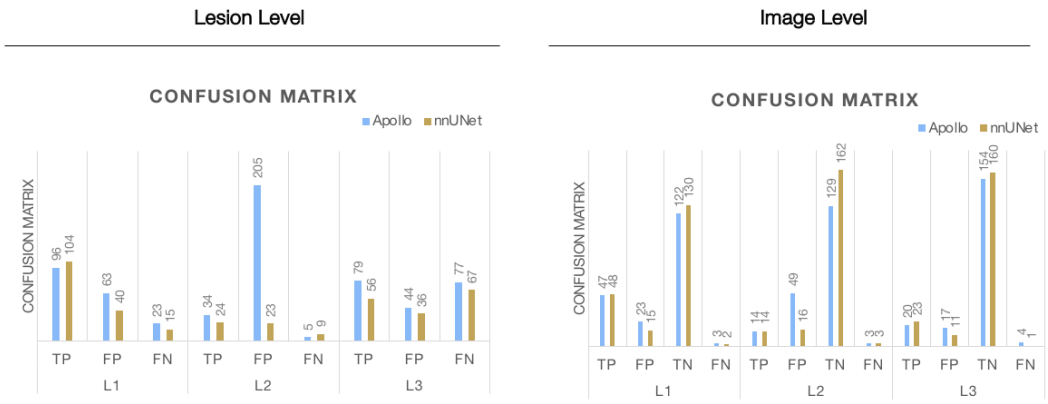


Figure 4.10: Confusion matrix in the in-house dataset for *Apollo* (blue) and *nnU-Net* (brown) at (**Left.**) lesion level and (**Right.**) image level.

On the in-house dataset, *nnU-Net* reaches surprisingly high scores, when compared with *Apollo*. Apart from recall, *nnU-Net* outperforms *Apollo* in the remaining metrics, specially when it comes to precision. Nevertheless, statistical significance was only reached for specificity, L1 Hausdorff distance, and L1 precision.

Additionally, the average performance does not reflect behaviours across labels. What seems to be a general improvement is far from being homogeneous along classes, specially at a lesion level. At a lesion level, while L1 shows a decrease in the number of FP and FN and increase of TP, resulting in an improve recall and precision, the same cannot be said for L2 and L3. For L2, *nnU-Net* tends to under-predict lesions while *Apollo* tends to over-predict (see Figure 4.11). For L3, the decrease in FN and FP lesions does not compensate the decrease of TP lesions, leading to lower recall and precision scores. In that sense, *nnU-Net* managed to solve the FP issue of L2 but struggled in answering the FN issue of L3.

At an image level, discrepancies in performance between networks mostly arise from L2 and L3 classes, while similar behaviours are noticed for L1. It is interesting to report that the lower performance noticed in L3 has no direct consequences in its sensitivity, which shows an increase of 15 %. In turn, the decrease in the number of L2 FP lesions enabled nnU-Net to increase its specificity score by 25 %. These results are supported by the confusion matrix (Figure 4.10 - Right) that shows a general increase of TP and TN and a decrease of FP and FN. It seems that *nnU-Net* managed to learn more discriminative features across classes, independently of their low representation on the training data and broad landscapes of features. By condensing domain knowledge into a set of heuristic rules, *nnU-Net* by-passes the aforementioned L3 FN and L2 FP issues (Section 4.2.1).

In terms of segmentation, the quality of the predictions is especially improved for L3 (higher Dice coefficient combined with a lower Hausdorff distance). Overall, Dice scores are higher for *nnU-Net* across labels, while Hausdorff distance is only lower for hemorrhages.

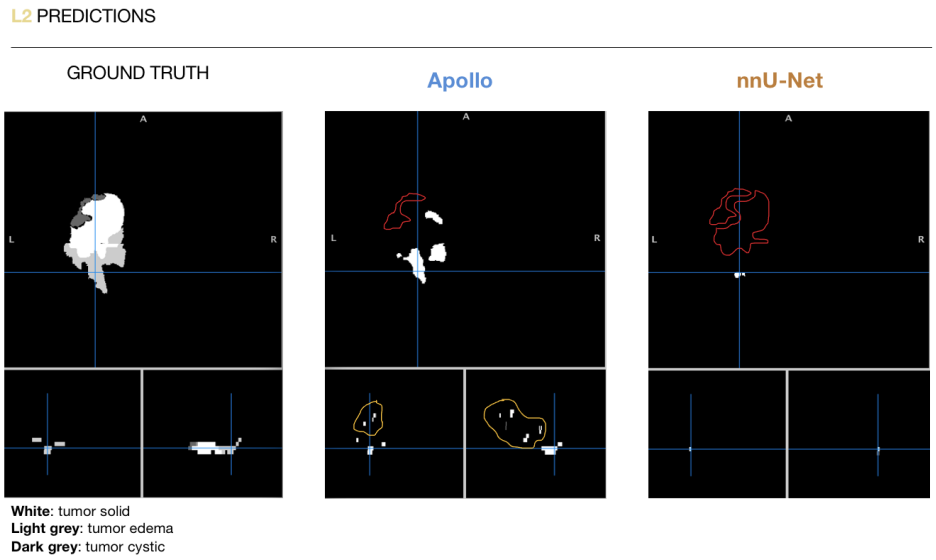


Figure 4.11: Predictions for L2 (tumors) ground truth (shown on the (**Left**) for *Apollo* (**Middle**) and *nnU-Net* (**Right**) on the in-house dataset. FP lesions are reported in yellow and FN lesions in red. *nnU-Net* tends to under-predict while *Apollo* presents non-sense L2 predictions.

Another interesting observation comes from analyzing the ground-truth labels of FP and FN lesions. Regarding FP lesions, most predictions arise from background for L3 in both networks. However, when this is not the case, similar behaviours where FP are associated with tumor hemorrhagic are reported. Differences are found for L2: while, for *nnU-Net*, FP are associated with edemas, gliosis, and chronic infarct, in *Apollo* they are mainly random [3].

Regarding FN lesions, similar patterns are seen for both algorithms. Misclassifications are mainly reported in poorly represented sub-types in the training set. *nnU-Net* and *Apollo* struggle with hyper acute infarcts, cystic tumor, epidural/subdural late sub-acute hemorrhages, and epidural/subdural chronic hemorrhages. *nnU-Net* handles better SAH, epidural/subdural hyper-acute hemorrhages, intraparenchymal acute and chronic hemorrhages, and hyper-acute infarcts. In turn, *nnU-Net* has a higher FN rate regarding solid tumor.

nnU-Net also seems to be less prone to confusing labels at both image and lesion levels.

4.3 Performance as a Function of Sequence Type and Orientation

This section is dedicated to the analysis of the performance of *Apollo* and *nnU-Net* as a function of acquisition parameters, *i.e* sequence types and orientations. The experiments allow for a better understanding of the optimal use of the networks in clinical settings. However, they can be taken to a higher level and express a potential measure of generalization ability to the selected acquisition parameters.

As mentioned in Section 3.5.2, for **Axial** *versus* **Coronal** FLAIR (experiment 1), comparisons were only

[3] When not random, FP lesions share the same ground truth class detailed for *nnU-Net*.

performed for L1 patients. In turn, for **T2 * GRE** and **SWAN/SWI** (experiment 2), comparisons were performed for L3 patients. Since the sample sizes observed in both groups are relatively small to achieve a meaningful statistical analysis, no conclusions were drawn from statistical tests [4] nor from confidence intervals (CI). For the purposes of consistency and uniformity, CI are kept when presenting the results. Extremely broad CI are obtained, especially for sensitivity, recall, and precision scores. Performance is far from being uniform, and highly depends on the patients selected from the distribution.

Confusion matrices for the two experiments conducted in the two neural networks are shown in Figure 4.12.

4.3.1 *Apollo*

The results for **Axial** vs **Coronal** FLAIR, supporting the confusion matrices, are shown in Table 4.5. As in Section 4.2, confidence intervals (CI) were also computed for each parameter of interest θ. However, in the present case, respective datasets are sub-divisions of the entire dataset.

Table 4.5: *Apollo* evaluation for infarct prediction in the in-house dataset as a function of sequence orientation: Axial FLAIR *versus* Coronal FLAIR. Evaluation is performed at an image (sensitivity, specificity, Dice coefficient and Hausdorff distance) and lesion (recall and precision) levels. Dice score is given for the entire dataset. CI, computed via bootstrap, are presented $[\alpha; \beta]$. Higher values of performance are highlighted in bold.

	AXIAL FLAIR	C0R0NAL FLAIR
sensitivity (%)	**100.00** [- ; -]	91.30 [69.23 ; 100.00]
specificity (%)	75.68 [64.93 ; 84.62]	**92.06** [78.05 ; 92.77]
mean Dice Coefficient	0.482 [0.348 ; 0.612]	**0.651** [0.487 ; 0.786]
mean Hausdorff Distance (mm)	**19.34** [11.95 ; 29.21]	33.21 [20.69 ; 49.53]
recall (%)	**91.84** [69.05 ; 100.00]	77.78 [54.90 ; 91.94]
precision (%)	47.37 [28.57 ; 66.67]	**79.03** [55.56 ; 91.84]

From Figure 4.12, higher number of FP and FN lesions are seen for Axial and Coronal FLAIR, respectively. It seems that Axial FLAIR tends to over-predict lesions while Coronal FLAIR tends to under-predict them. Direct consequences can be observed in recall and precision scores and in sensitivity and specificity scores. In term of quality of segmentations, higher mean Dice is achieved by Coronal FLAIR, while lower Hausdorff distance is reported for Axial FLAIR. However, this is explained by the high number of FP lesions, noticed in Axial FLAIR [5]. When a patient has an infarct, Axial orientation seems to lead to higher quality segmentations, combining higher Dice scores with lower Hausdorff distances.

By the fact that higher sensitivity and recall score are achieved with Axial FLAIR, *Apollo* robustness

[4] With the exception of specificity, none of the metrics in *Apollo* FLAIR experiment were statistically significant (ρ *value* > 0.05) for both independent T-test and Mann–Whitney U test and large ρ *value* were obtained for the sensitivity score (0.153 for nonparametric and 0.162 for parametric) and recall (0.104 for non parametric and 0.114 for parametric). In the *Apollo* SWAN/SWI/T2 * GRE experiment, for some metrics, not enough values (less than 10) were reported and the evaluation could not be performed. Significance was only reached for Hausdorff distance. Similar results were achieved with *nnU-Net* (with one additional significant metric, *i.e* precision, in the FLAIR experiment). Meaningful results are expected with a higher pool of patients.

[5] When removing the healthy patients from the dataset (and consequently, removing a high proportion of the FP lesions), higher mean Dice is seen for Axial FLAIR

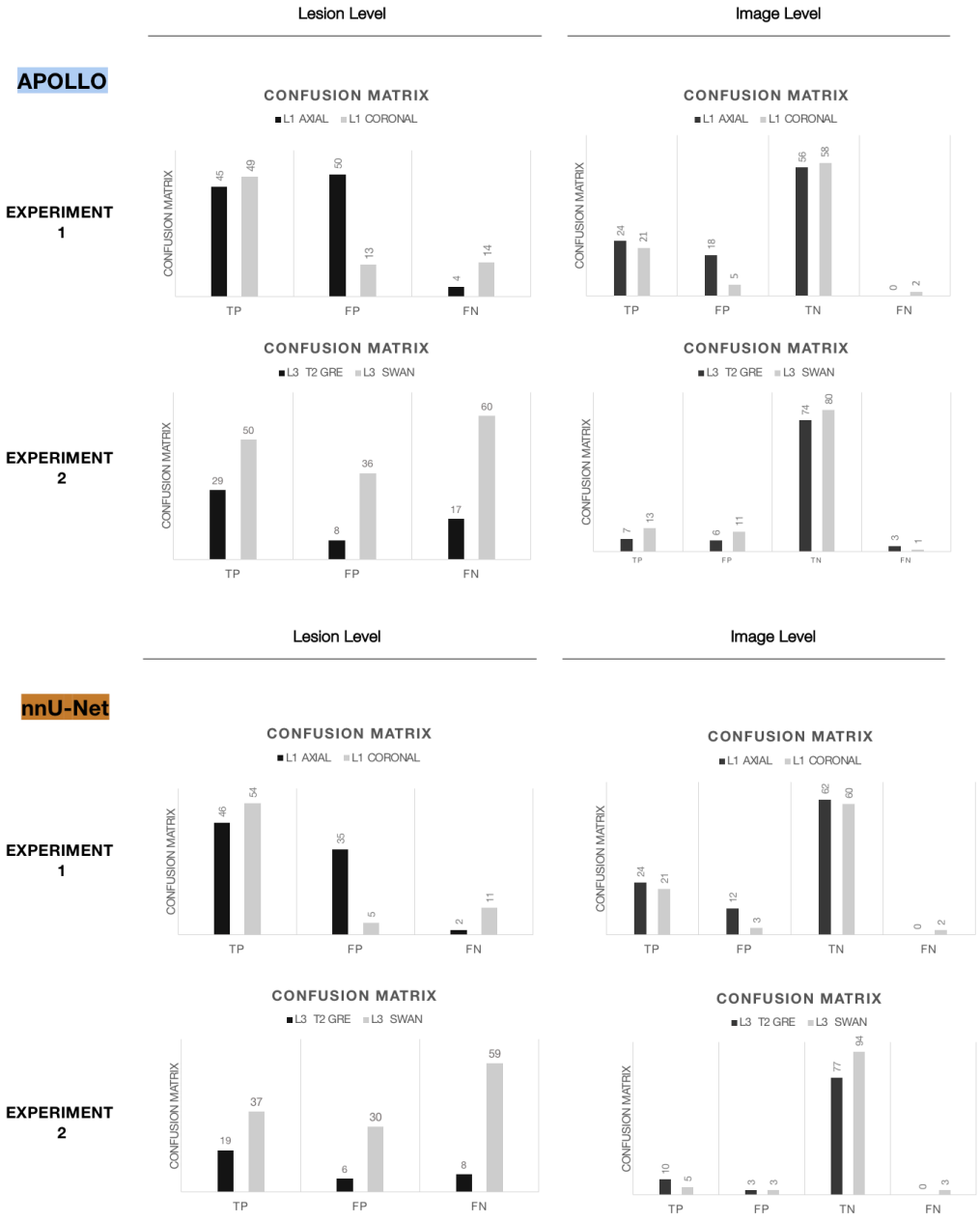


Figure 4.12: Confusion matrices for experiments 1 (Axial FLAIR in dark and Coronal FLAIR in grey) and 2 (T2 * GRE in dark and SWAN/SWI in grey) for (**Top**) *Apollo* and (**Bottom**) *nnU-Net*.
In Axial FLAIR tends to over-predict while Coronal FLAIR tends to under-predict lesions. Direct consequences can be seen at the image level.
Better lesion level performance is achieved by T2 * GRE. In *nnU-Net*, better image level performance is also reported, while in *Apollo*, higher sensitivity score is reached for SWAN/SWI.

in terms of sequence orientation should be questioned. *Apollo* does not seem agnostic to the sequence orientation for infarct predictions. We hypothesize that this difference in performance lies in the training set composition and in intrinsic imaging properties. On one hand, the training dataset includes 1.4 times more infarct patients with Axial orientation. On the other hand, acquiring DWI and FLAIR in the same direction should improve the aggregation of contextual information coming from different inputs.

In medical applications and in this particular case, the trade-off between sensitivity and specificity should always lean towards sensitivity by reducing the rate of undetected pathologies. Therefore, *Apollo* should be used with Axial FLAIR for infarct detections, since its higher sensitivity would translated a higher degree of quality care given to the patients. However, clinicians must be aware that in these conditions, *Apollo* would also induce an increment in hospital costs due to increased admissions rate (from its lower specificity). A trade-off must be achieved to counterbalance these two sides.

In the second experiment, we have decided to focus on the disparities in performance between **T2*GRE** and **SWAN/SWI** for hemorrhages detection and the results can be found in Table 4.6.

Table 4.6: *Apollo* evaluation for hemorrhages prediction in the in-house dataset as a function of sequence type: T2 * GRE *versus* SWAN + SWI.CI, computed via bootstrap, are presented $[\alpha; \beta]$. VHigher values of performance are highlighted in bold.

	T2 GRE	SWI + SWAN
sensitivity (%)	70.00 [28.57 ; 92.31]	**92.86** [57.14 ; 100.00]
specificity (%)	**92.50** [84.15 ; 96.47]	87.91 [79.06;93.41]
mean Dice Coefficient	**0.307** [0.140 ; 0.515]	0.281 [0.163 ; 0.421]
mean Hausdorff Distance (mm)	**28.16** [16.73 ; 47.50]	65.20 [45.33 ; 87.40]
recall (%)	**63.04** [31.58 ; 89.19]	45.45 [18.75 ; 74.03]
precision (%)	**78.36** [45.00 ; 95.00]	58.14 [36.27 ; 79.12]

From Figure 4.12, higher numbers of FP and FN lesions are seen for SWAN/SWI, translating into lower precision and recall. Interestingly, in this case, worse performance at a lesion level is not correlated with worse performance at an image level, and SWAN/SWI reaches higher sensitivity for a similar specificity when compared to T2 * GRE. Analyzing the quality of segmentations, similar mean Dice are seen for both sequence types. However, in terms of Hausdorff distance, lower mean is achieved by T2 * GRE. Therefore, in terms of segmentation, higher quality tends to be associated with T2 * GRE. By outperforming SWAN/SWI in terms of quality of segmentation and lesion level metrics, T2 * GRE seems to produce more discriminative features that facilitate *Apollo* predictions of hemorrhages [6]. Once more, considering the discrepancies between SWAN/SWI and T2 * GRE, *Apollo* does not seem agnostic to the type of sequence.

In clinical settings, SWAN/SWI acquisitions seem to be more beneficial for hemorrhage detections. As explained in Section 3.1, the detection of one lesion in the segmentation maps is enough to cor-

[6] This could also arise from the training set composition in which 53 % of the hemorrhage patients are encompassed in the T2 * GRE group.

rectly classify the patient and alert the radiologists. Despite a lower performance for the segmentation tasks, SWAN/SWI reaches a higher sensitivity score. Therefore, it avoids wrong prioritization and negative patient outcomes. The lower sensitivity noticed for T2 * GRE is not exclusive to automated processes. Radiologists (*Cerebriu* annotators) refer SWAN/SWI as being more sensitive in detecting micro-hemorrhages due to the fact that it is a 3D acquisition technique with a higher out-of-plane resolution.

4.3.2 Comparison of *Apollo* and *nnU-Net*

Again, a comparison between *Apollo* and *nnU-Net* in terms of performance as a function of sequence type and acquisition is performed. Results are reported in Figures 4.12 (Bottom). The response of *nnU-Net* facing different acquisition parameters can easily be perceived from the confusion matrices. As a result, metrics will not be presented.

Regarding **Axial** *versus* **Coronal** FLAIR, similar behaviours are noticed across networks in both confusion matrices and evaluation metrics. As in *Apollo*, *nnU-Net* tends to over-predict with Axial FLAIR and under-predict with Coronal FLAIR.

Regarding **T2 * GRE** and **SWAN/SWI**, feeding *nnU-Net* with T2 * GRE for hemorrhages prediction leads to better results across all metrics. Again, T2 * GRE seems to be more beneficial for feature extraction and lesion segmentation. However, while similar behaviours are reported at a lesion level, the networks diverge at an image level. In *nnU-Net*, T2 * GRE tends to over-predict and SWAN/SWI under-predict patients with hemorrhages. Surprisingly, and against clinicians preferences, T2 * GRE classifies a higher number of TP patients with hemorrhages.

As a final observation, by achieving satisfactory results across groups at an image level and in terms of segmentation quality, *nnU-Net* seems more robust facing differences in sequence types and orientations. While in *Apollo*, T2 * GRE sensitivity for hemorrhage prediction or Axial FLAIR specificity for infarct prediction are below 80 %, *nnU-Net* shows decent scores, independently of the MRI settings.

4.4 Generalization Ability

The present section assesses the robustness of *Apollo* and *nnU-Net* by evaluating them in an external dataset with unseen MRI acquisitions parameters. Results for *Apollo* and *nnU-Net* are described in Table 4.7. Confusions matrices in Figure 4.13 and 4.14 are also exposed in the same order to bring additional insights. Metrics are discriminated per label in Appendix D.1 and D.2.

4.4.1 *Apollo*

Table 4.7: (**Left.**) *Apollo* and (**Right.**) *nnU-Net* evaluation in the in-house (grey) and external (purple) datasets. Results are reported after post-processing that lead to a general increase in performance for both networks. Metrics are averaged across labels. Higher values of performance are highlighted in bold.

	in-house	external
sensitivity (%)	**86.40**	72.25
specificity (%)	**81.89**	13.44
Dice Score	**0.303**	0.141
Hausdorff (mm)	**37.44**	53.90
recall (%)	70.88	**83.36**
precision (%)	**38.07**	18.28

	in-house	external
sensitivity (%)	**91.16**	61.95
specificity (%)	**91.40**	67.60
Dice Score	**0.440**	0.234
Hausdorff (mm)	**38.84**	53.01
recall (%)	66.14	**68.98**
precision (%)	**60.77**	30.62

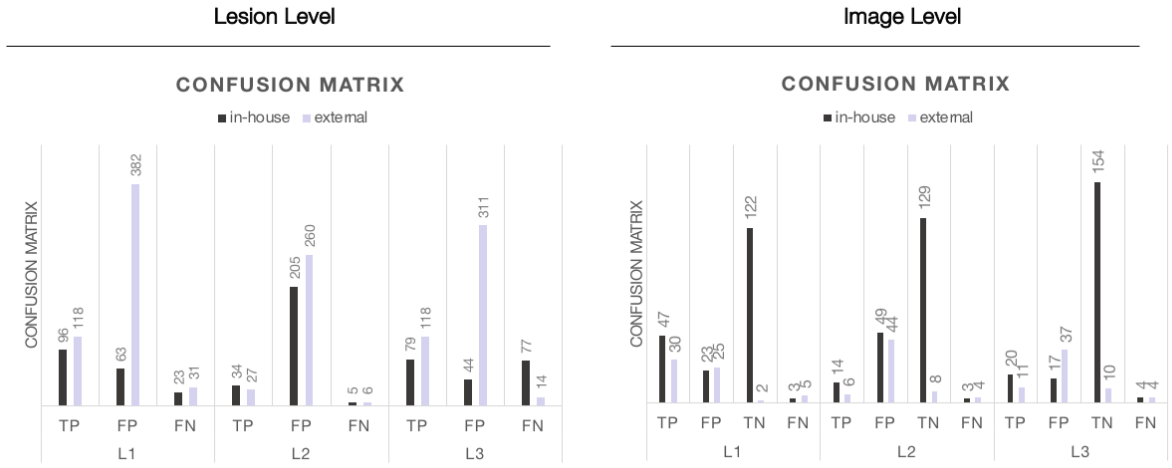


Figure 4.13: Confusion matrix for *Apollo* in the in-house (grey) and external (purple) datasets. Results are shown for the (**Left.**) lesion level and (**Right.**) image level.

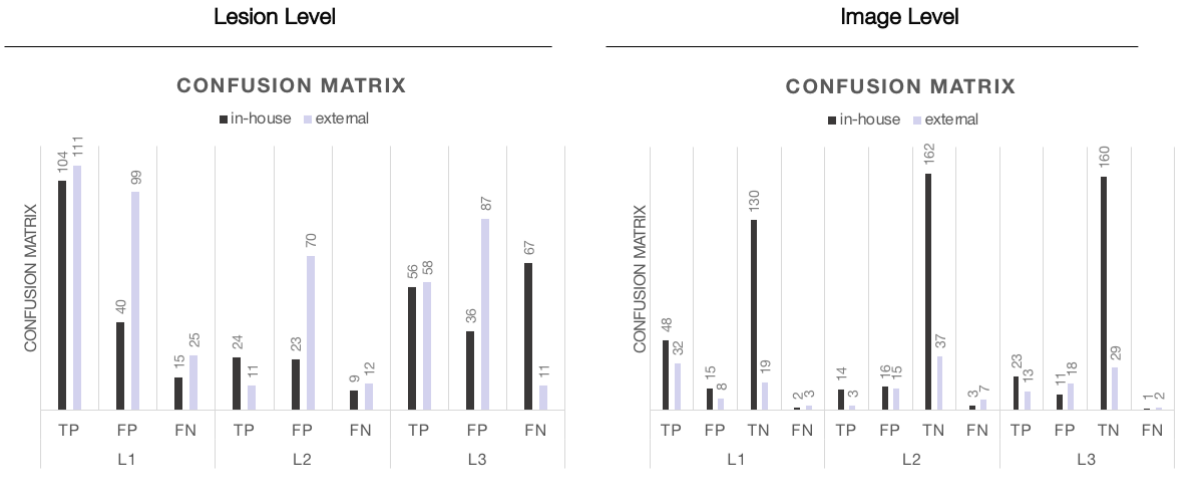


Figure 4.14: Confusion matrix for *nnU-Net* in the in-house (grey) and external (purple) datasets. Results are shown for the (**Left.**) lesion level and (**Right.**) image level.

The experimental findings corroborate the hypothesis defined in Section 3.3 on the expected degradation in performance for unseen data. *Apollo* makes noisy predictions and has troubles in distinguishing pathology *versus* background. FP are predominant in the confusion matrix at a lesion level. While in pre-

vious experiments L2 was the only label exhibiting this issue, FP lesions have now spread to all classes in the new dataset. It affects the precision score, especially for L1 and L3 . By over-predicting lesions on the external dataset, L3 transferred its FN issue to a FP problem and increased its recall score. At an image level, FP lesions have a major impact in the confusion matrix. When taking the proportion of FP classifications among all the patients of the datasets, it reaches 70% and 60% for the L2 and L3 classes. Similarly, the proportion of TN patients has also suffered a decrease: it has been divided by 19 for L1 and 5 for L2 and L3. This has a cumbersome impact on the specificity and the average score has dropped 84 % between datasets. L1 is the most affected class.

In terms of segmentation quality, poorer performance is also noticed in unseen data: the average Dice score is halved and Hausdorff distance is increased when translating from in-house to external set. In particular, L2 Dice score has reached an extremely low value.

Statistical significance was reached for specificity, L1 and L3 precision and Dice score, and L1 Hausdorff distance. Therefore, results highlight the dependency of *Apollo* to the characteristics of the training set, and this, across labels.

4.4.2 Comparison of *Apollo* and *nnU-Net*

A poorer performance is also seen for *nnU-Net* when facing unseen data. However, unlike *Apollo*, the effect is less widespread across labels and L2 suffers a higher drop in its scores. These findings are supported by the confusion matrix at a lesion level where the increase in FP (across labels) is paired with a severe decrease in TP (L2). This affects both precision and recall that show a decrease of 73 % and 34 %, respectively. Like in *Apollo*, L3 recall increases in the external dataset, but it is not reflected in the average due to L2 score. This behaviour towards L2 is also reported at an image level. *nnU-Net* seems to struggle more than *Apollo* in maintaining its average sensitivity score. However, the higher loss in sensitivity for *nnU-Net* is easily perceived as being again a L2 issue, and not a general tendency across class.

In terms of the segmentation quality across datasets, *nnU-Net* and *Apollo* react similarly to unseen data. In particular, L2 Dice coefficient and L3 Hausdorff distance are the more affected scores.

Ignoring L2, *nnU-Net* managed to better regulate the relative numbers of FP and TP lesions, maintaining precision to decent levels. In a similar fashion, *nnU-Net* handled smoothly the decrease of TN and the increase of FP patients. These findings justify how *nnU-Net* is able to preserve a satisfactory specificity. Instead of a 84 % decrease, a slight drop of 26 % is registered.

The statistical analysis corroborates the aforementioned findings and validates the poorer performance for L2 across metrics for *nnU-Net*, with the exception of the Hausdorff distance. L1 and L3 specificity and L3 Dice also present a $\rho < 0.05$.

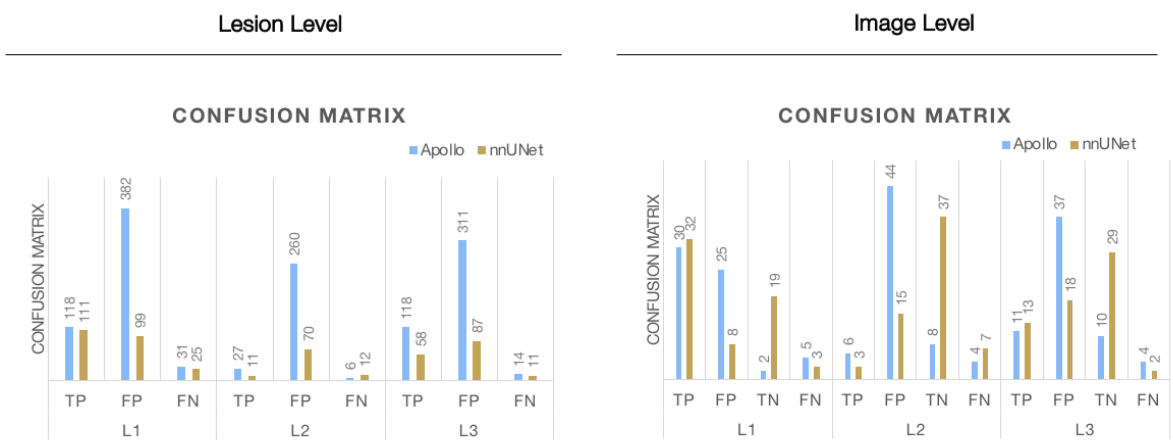


Figure 4.15: Confusion matrix in the external dataset for *Apollo* (blue) and *nnU-Net* (brown) at (**Left.**) lesion level and (**Right.**) image level.

It is also relevant to perform a comparison between the two networks in the external dataset. Confusion matrices are presented in Figure 4.15. An overlap is noticed between the FP and FN patients across labels of the two networks: except for L2 FP, all the misclassified patients by *nnU-Net* were also misclassified by *Apollo*. In general, *nnU-Net* seems to be more resilient when facing unseen data. Despite performance deterioration and ignoring L2 sensitivity, recall, and precision, the remaining scores obtained in the external dataset are still acceptable and outperform *Apollo*. *nnU-Net* is able to by-pass the poor scores in specificity and precision, observed in *Apollo*, and preserve sensitivity and Dice coefficients for L1 and L3 to relatively high values. The differences observed are statistically validated for specificity across labels and Dice, Hausdorff, and precision for L1 and L3 [7]. It is also interesting to note that *nnU-Net* has the tendency to under-predict lesions and *Apollo* to over-predict them. An example of L2 predictions is presented in Figure 4.16 to support these findings.

The last topic that still needs to be addressed concerns the possible reasons behind the decrease in performance observed for both networks in the external set. From Figures 3.4 and 3.5, we hypothesize that the reasons behind the observed discrepancies across datasets lie on the fact that DL solutions are predicting on out-of-distribution data:

- Shifts in class distribution - training/external:

Shifts in class distribution between the external and the training sets is one possible assumption to justify the drop in performance noticed for unseen data. Labels that had less opportunity to be learnt and extracted correctly by the network are present in higher proportions in the external set, impacting its performance in a more pronounced manner. As an example, the analysis conducted in the in-house

[7] L2 discrepancies across networks could not be statistically confirmed, as the number of paired patients the analysis did not reach 10.

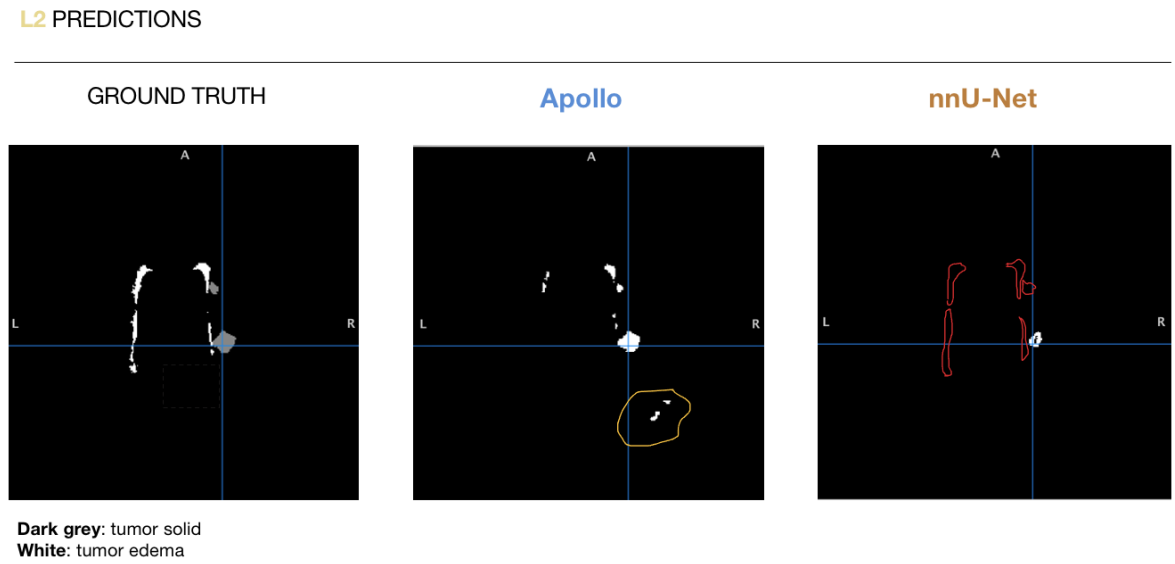


Figure 4.16: Predictions of L2 (tumors) for *Apollo* (**Middle**) and *nnU-Net* (**Right**) on the external dataset. Ground truth is shown on the (**Left**). FP lesions are reported in yellow and FN lesions in red. L2 predictions are embedded with FP for *Apollo*, while *nnU-Net* struggles in detecting lesions.

dataset revealed that *Apollo* and *nnU-Net* struggled in detecting hyper acute infarcts and this was justified by their low occurrence in the training set. By including six times more hyper acute infarcts than the in-house dataset, the external dataset is more susceptible to suffer from FN classifications. The fact that the highest percentage of FN lesions in L1 falls in the hyper acute class corroborates our assumptions. This effect can also be appraised in the L3 predictions. *Apollo* struggled in predicting intraparenchymal acute, intraparenchymal chronic hemorrhages, and epidural/subdural chronic hemorrhages that are represented in higher proportions in the external set.

- Shifts in class distribution - in-house/external:

Shifts in class distribution between the external and the in-house sets can also exacerbate the impact of some labels in the overall performance. If a poorly represented label in the in-house set is misdetected, its effect is overshadowed by the performance of more prevalent labels. However, if in the external dataset, this label is present in a higher proportion, its impact will have a higher weight on the overall performance. This effect is prevalent for *nnU-Net* tumor predictions and may explain the behaviour of the network towards that class. In the in-house experiments, our findings reported that *nnU-Net* was missing a high percentage of solid tumors that are more prevalent in the external dataset. In addition, among the tumors predicted by *nnU-Net*, FP lesions mostly arise from chronic infarcts and edemas that are in higher proportions in the external dataset. Similar behaviours can be noticed for *nnU-Net* and L3 predictions: in the external dataset, FP lesions are related with hemorrhagic transformation of infarct that are not represented in the in-house dataset.

- Shifts in MRI parameters:

Out-of-distribution data can also emerge from differences in the scanners and acquisition parameters. This type of distributional shift seems to deeply affect the L3 class. While 80 % of the hemorrhage patients in the external set present a SWI scan, they only represent 1.23% of the training set. Transferring the discriminative features learnt on T2 * GRE (Section 4.3) to SWI images is far from being trivial and may explain the behaviours of DL solutions when predicting on previously unseen data. Similar shifts are also noticed, in a lower measure, for L2 predictions on FLAIR acquisitions. Algorithms trained with most of the tumors annotated in the Axial orientation have to predict on the Coronal orientation in the external dataset. Therefore, it is understandable why higher percentages of misclassification (FP and FN) are found for Coronal FLAIR scans.

5

Conclusions

Contents

Chapter 5 outlines the main findings of the conducted experiments and suggests possible extensions and future research directions.

5.1 Summary of Findings

This thesis introduces a complete evaluation framework for DL algorithms performing lesion segmentation and image classification on MRI images. Our heuristic approach is built upon a deep understanding of its clinical application. By presenting a broad panorama of metrics and steps, along with a strong statistical analysis, experiments have given a comprehensive perception of the strengths and weaknesses of the models across segmentation and classification tasks. In particular, using evaluation metrics such as Hausdorff distance has lead to a better assessment of segmentation quality and border delineation. Including CI computations in the analysis enables to sense how consistent the results are. Nevertheless, the major contribution of our framework is the assessment of distributional shifts in the algorithm performance. By evaluating performance across acquisition parameters and, at a bigger scale, across different datasets, our method gives insights on how networks are hampered by their training data and its respective MRI parameters.

By extensively addressing the generalization ability, our analysis evidences the impact of unintended data bias on the performance of DL models. In both datasets and across experiments, lower performances were reported for sub-label types and MRI acquisition parameters poorly represented in the training data. This is even more evident on the external and unseen test set. The reason behind this phenomenon is that models are opportunists. Algorithms tend to learn over-represented pathologies that better solve the optimization problem of the learning step and struggle to predict on data acquired with different acquisition parameters. Findings reinforce the importance of conducting an external evaluation to assess robustness across clinical sites. Crucially, by-passing this step may result in sub-optimal performances and misclassifications, preventing a seamless integration of DL solutions in hospitals. To address these limitations, recent guidelines published by *Radiology* [21] and by Challen *et al.* [4] help radiologists in gauging models from a quality and safety perspectives.

Another key aspect in the design of our evaluation framework is the comparison of a DL algorithm with the status quo in biomedical segmentation, *nnU-Net*. From a DL perspective, it is impressive how *nnU-Net* is able to condense domain knowledge and identify robust design decision to adapt to *Cerebriu* data and tasks. Trained with one-fold cross-validation instead of the recommended five, *nnU-Net* outperforms *Apollo* that was specially designed towards that data and tasks. *nnU-Net* has shown higher homogeneity in performance across tasks and seems more robust towards different sequence types and orientations. Except for tumors, satisfactory results were obtained on unseen data, highlighting a higher generalization ability. We hypothesize that the reason behind *nnU-Net* better performance on unseen data lies in its pipeline formulation and key design choices. This observation is in line with Isensee *et al.* [49] that

justifies the state-of-the-art performance of *nnU-Net* as a direct result of ”the distillation of knowledge from a large data pool into a set of robust design choices ” made to automate its configuration and promote its agnosticism to tasks or datasets. We expect divergence in performance to arise from the following design choices (Table 3.1):

- *nnU-Net* was optimized on a more comprehensive loss function. While *Apollo* only considers Dice (with background), *nnU-Net* combines Dice (without background) and Cross entropy. Removing the background in the loss computation may be beneficial as it penalizes errors on smaller areas such as lesions, reaching a better segmentation output. Cross entropy, in turn, allows for a faster convergence of the gradient in early steps of training.
- By making the network more robust to variations in the training data, higher segmentation quality can also arise from the extensive and complex data augmentation procedures undertaken by *nnU-Net*.
- *nnU-Net* was trained with deep supervision: the total loss function is computed with all but the two lowest resolution in the decoder. It may help decrease the percentage of FP lesions in tumors.
- *Apollo* was trained with early stopping while it was not the case for *nnU-Net*. A longer training allows the algorithm to extract better discriminative features and enables a finer tuning of lower prevalence class. Removing early stopping in *Apollo* may reduce the percentage of FN lesions in hemorrhages.

However, this superiority in performance comes with the cost of being more demanding in terms of computational memory and requiring extremely long training times. While *Apollo* trains one fold in three days, *nnU-Net* takes more than one week and a half.

5.2 Limitations and Future Work

This section comprises the limitations encountered in this work and explores potential directions of future work.

While our framework has shown its potentials in evaluating DL algorithms , validating in a larger pool of patients would further improve the quality and convergence of our findings. From finding hospitals willing to share clinical data to generating ground truth segmentation maps, having data ready for training and evaluation was a cumbersome and time-consuming process. Data scarcity currently hampers the selection of additional MRI acquisition parameters and jeopardizes the statistical validation of the results.

Besides data availability, the information of the headers was far from being homogeneous across datasets and data formats. Compared to DICOM, NIFTI format lacked information on several MRI acquisition parameters. In addition, when the information was available, a large discrepancy in how headers were filled was noticed. This is a direct result of the high DICOM flexibility, also noticed in [46].

By allowing radiologists to manually enter the *Series Description* without any standardized guidelines, storing the type of sequence and its orientation involved a case-to-case confirmation by visual inferences to obtained an uniform database.

Finally, another minor limitation was the availability of only one GPU, making the training task of *nnU-Net* a timely process with a ratio of one week and a half per fold.

A simple task that deserves attention in the immediate future is to perform an external evaluation in a larger pool of patients in order to achieve meaningful statistical validation and assess the impact of additional MRI acquisition parameters in the network performance. In particular, understanding how sequence type and orientation influences the performance in the external dataset would complement the analysis conducted in the in-house dataset.

Another interesting line of thought would be to include this framework in *Cerebriu* evaluation routine and to add additional human-centered and clinically relevant metrics. It would enable a more accurate monitoring of the network performance across the different software up-dates or clinical sites during pilot studies and would guarantee a comprehensive overview of its impact in real settings.

Moreover, even if our approach can serve as an accurate baseline to evaluate DL solutions, it was specially designed for *Apollo*. We believe that our framework can serve as a good starting point to further extend to other types of imaging techniques and tasks.

5.3 Final Considerations

This section is meant to give insights on how *Apollo* is expected to create value in danish and portuguese hospitals.

5.3.1 *Apollo* in Danish Hospitals

In early 2021, *Cerebriu* decided to refine its scope and shifted from infarcts, tumors, and hemorrhages predictions to an infarct-specific scenario. Under this new clinical setting, *Apollo* main application is to automate infarcts detection (rule-in and rule-out) in MRI scans. It is meant to guarantee faster treatment decision when infarct is detected or to provide a more confident rule-out.

Apollo is still to be implemented in danish clinical settings, therefore data to evaluate the socio-economic outcomes for the radiology department is scarce. Some institutions, as Herlev Hospital, are incorporating the software in their scanning protocol and initiating retrospective trials. In collaboration with the latter hospital, a study case was conducted within *Cerebriu*, under specific assumptions. Currently, Herlev Hospital only includes MRI in its stroke management protocol from 08h00 to 20h00, based on workforce requirements. Using *Apollo* in the scanning room could extend the schedule from 08h00

to 00h00 and allow for extra 1 070 potential stroke patients to undergo an MRI scan per year and benefit from a better disease management. From the Herlev's representative feedback, patients arriving in these extra four hours undergo a CT and are kept in the hospital until 08h00 the next day to perform a MRI. Moreover, it is assessed that approximately 60% of the hospitalizations arise from preventive measures. Therefore, under these assumptions, 1 070 CT scans and 60% of the admissions could be avoided during the extended hours in one year. Following the reimbursement policies of Denmark, these would contribute in a yearly cost reduction of 2 148 025 DKK and 5 326 717 DKK, respectively. Therefore, the total benefits of using *Apollo* would result in more than 7 000 000 DKK per year [1]. Financial return is also expected within the first year.

5.3.2 *Cerebriu* Penetration in the Portuguese Market

According to the Stroke Alliance for Europe (SAFE) report [84] (2015), 15 208 strokes and 5 297 strokes are reported yearly, in Portugal and Denmark. Incidences are estimated at 75.4 in Portugal *versus* 56.5 in Denmark per 100 000 inhabitants. Mortality rates are also higher in Portugal where strokes are responsible of 67.9 deaths *versus* 44.9 in Denmark per 100 000 inhabitants. The aforementioned figures justifies the implementation of faster and more efficient stroke management protocol in Portugal.

Unfortunately, *Apollo* implementation does not solely rely on stroke incidence or mortality rates. It also requires hospitals to include MRI in their stroke management protocol. Criteria for market selection and penetration are based on MRI scanners availability and integration in stroke management protocol. As the internship is a partnership between IST and *Cerebriu*, it is relevant to evaluate the viability of implementing *Apollo* in Portugal and stipulate a potential integration scenario.

In Portugal, the difference in medical equipment availability is abysmal between MRI and CT scans. Per 100 000 inhabitants, the country holds four times more CT scans than MRI units [6] [2]. Moreover, the current stroke management follows the "Via Verde AVC" protocol [85]. The protocol guidelines for imaging techniques corroborate with the recommendation of World Health Organization (WHO) [86] and mention a CT and angio-CT. MRI is only prescribed in case of inconclusive CT diagnosis. In view of the low MRI availability and the CT-based stroke protocol, the transfer of *Apollo* technology to portuguese clinics is limited.

[1] This figure does not take into account costs related with implementation, utilization and staff training.
[2] According to [6], Portugal possesses 272 CT scans versus 94 MRI scanners.

Bibliography

[1] Cerebriu A/S . (2020) Apollo for brain - cerebriu solutions. Accessed 14-December-2020. [Online]. Available: https://www.cerebriu.com/apollo-detailed-workflow/

[2] K. Fukushima, "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position Kunihiko," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[4] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-atanasova, "Artificial intelligence , bias and clinical safety," *BMJ Quality & Safety*, vol. 28, pp. 231–237, 2019. [Online]. Available: https://qualitysafety.bmj.com/content/28/3/231

[5] D. L. Rubin and M. P. Lungren, "Preparing Medical Imaging Data for Machine Learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020. [Online]. Available: http://dx.doi.org/10.1148/radiol.2020192224.Preparing

[6] Eurostat, "Healthcare resource statistics - technical resources and medical technology," *Statistics Explained*, pp. 1–18, 2020. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php/Healthcare_resource_statistics_-_technical_resources_and_medical_technology#Availability_of_technical_resources_in_hospitals

[7] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nat. Rev. Cancer.*, vol. 18, no. 8, pp. 500–510, 2018. [Online]. Available: http/dx.doi.org/10.1038/s41568-018-0016-5.Artificial

[8] B. Kocak, E. [U+FFFD] Durmaz, E. Ateş, and Ö. Kılıçkesmez, "Radiomics with artificial intelligence: a practical guide for beginners," *Diagnostic and interventional radiology*, vol. 25, no. 6, pp. 485–495, 2019. [Online]. Available: https://doi.org/10.5152/dir.2019.19321

[9] D. Harpaz, E. Eltzov, R. C. S. Seet, R. S. Marks, and A. I. Y. Tok, “Point-of-care-testing in acute stroke management : An unmet need ripe for technological harvest,” *Biosensors*, vol. 7, no. 30, pp. 1–39, 2017. [Online]. Available: http://dx.doi.org/10.3390/bios7030030

[10] American Stroke Association. Ischemic Stroke. Accessed 12-January-2021. [Online]. Available: https://www.stroke.org/en/about-stroke/types-of-stroke/ischemic-stroke-clots

[11] M. P. Lin and D. S. Liebeskind, “Imaging of ischemic stroke,” *Continuum*, vol. 22, no. 5, pp. 1399–1423, 2016. [Online]. Available: http://dx.doi.org/10.1212/CON.0000000000000376

[12] S. J. An, T. J. Kim, and B.-w. Yoon, “Epidemiology , Risk Factors , and Clinical Features of Intracerebral Hemorrhage : An Update,” *Journal of Stroke*, vol. 19, no. 1, pp. 3–10, 2017.

[13] P. M. Parizel, S. Makkat, E. Van Miert, J. W. Van Goethem, L. van den Hauwe, and A. M. De Schepper, “Intracranial hemorrhage : principles of CT and MRI interpretation,” *European radiology*, vol. 11, no. 9, pp. 1770–1783, 2001. [Online]. Available: http://dx.doi.org/10.1007/s003300000800

[14] J. Hodel, M. Rodallec, S. Gerber, R. Blanc, A. Maraval, S. Caron, L. Tyvaert, M. Zuber, and M. Zins, “Séquences IRM ≪ SWAN , SWI et VenoBOLD ≫ exploitant le phénomène de susceptibilité magnétique : principes techniques et applications cliniques,” *J Neuroradiol*, vol. 39, pp. 71–86, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.neurad.2011.11.006

[15] J. J. Heit, M. Iv, and M. Wintermark, “Imaging of intracranial hemorrhage,” *J. Stroke*, vol. 19, no. 1, pp. 11–27, 2017. [Online]. Available: http://dx.doi.org/10.5853/jos.2016.00563

[16] J. S. Whang, M. Kolber, D. K. Powell, and E. Libfeld, “Diffusion-weighted signal patterns of intracranial haemorrhage,” *Clin Radiol*, vol. 70, no. 8, pp. 909–916, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.crad.2015.04.006

[17] N. Nishida, H. Yano, T. Nishida, T. Kamura, and M. Kojiro, “Angiogenesis in cancer,” *Vascular health and risk management*, vol. 2, no. 3, pp. 213–219, 2006. [Online]. Available: https://doi.org/10.2147/vhrm.2006.2.3.213

[18] Z. Zhou and Z.-R. Lu, “Gadolinium-based contrast agents for magnetic resonance cancer imaging,” *WIRES NANOMED NANOBI*, vol. 5, pp. 1–18, 2013. [Online]. Available: http://dx.doi.org/10.1002/wnan.1198

[19] V. Gulani, F. Calamante, F. G. Shellock, E. Kanal, and S. B. Reeder, “Gadolinium deposition in the brain : summary of evidence and recommendations,” *The Lancet Neurol.*, vol. 16, no. 7, pp. 564–570, 2017. [Online]. Available: http://dx.doi.org/10.1016/S1474-4422(17)30158-8

[20] A. Selvikv Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on MRI,” *Zeitschrift fur medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2018. [Online]. Available: https://doi.org/10.1016/j.zemedi.2018.11.002

[21] D. A. Bluemke, L. Moy, M. A. Bredella, B. B. Ertl-Wagner, K. J. Fowler, V. J. Goh, E. F. Halpern, C. P. Hess, M. L. Schiebler, and C. R. Weiss, “Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the Radiology Editorial Board,” *Radiology*, vol. 294, no. 3, pp. 487–489, 2020. [Online]. Available: https://doi.org/10.1148/radiol.2019192515

[22] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, “Deep learning for brain MRI segmentation : State of the art and future directions,” *J. Digit. Imaging*, vol. 30, pp. 449–459, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5537095/

[23] P. Figueiredo, “Medical imaging computed tomography - CT,” Instituto Superior Tecnico, Lisboa, Tech. Rep., 2019.

[24] D. G. Nishimura, *Principles of Magnetic Resonance Imaging*, 1st ed. Stanford University, 2010.

[25] S. David, *Parametric and Nonparametric Statistical Procedure*, 2nd ed. Chapman & Hall/CRC, 2000.

[26] C. Hess. Exploring the brain: Is CT or MRI better for brain imaging? Accessed 09-April-2020. [Online]. Available: https://radiology.ucsf.edu/blog/neuroradiology/exploring-the-brain-is-ct-or-mri-better-for-brain-imaging

[27] P. Figueiredo, “Medical imaging - magnetic resonance imaging,” Instituto Superior Técnico, Lisbon, Tech. Rep., 2019.

[28] ETH (Eidgenössische Technische Hochschule) Zurich. MW Pulses and Spin Dynamics. Accessed 20-April-2021. [Online]. Available: https://epr.ethz.ch/education/basic-concepts-of-epr/mw-pulses---spin-dyn-.html

[29] R. G. Nunes, “Diffusion Weighted Magnetic Resonance Imaging,” Instituto Superior Técnico, Lisboa, Tech. Rep., 2020.

[30] P. Figueiredo, “Neuroimaging - functional MRI I,” Instituto Superior Técnico, Lisbon, Tech. Rep., 2020.

[31] Interventional Neuroradiology - UCLA Health. Acute stroke. Accessed 12-January-2021. [Online]. Available: https://www.uclahealth.org/radiology/interventional-neuroradiology/acute-stroke

[32] Deng,Francis and Gaillard, Frank. Ischaemic stroke. Accessed 10-January-2021. [Online]. Available: https://radiopaedia.org/articles/ischaemic-stroke

[33] R. von Kummer, “MRI: The new gold standard for detecting brain hemorrhage?” *Stroke*, vol. 33, pp. 1748–1749, 2002. [Online]. Available: https://doi.org/10.1161/01.STR.0000019882.06696.D6

[34] K. Nagenthiraja, P. Brian, M. B. Hansen, L. Østergaard, and K. Mouridsen, “Automated decision-support system for prediction of treatment responders in acute ischemic stroke,” *Front. Neurol.*, vol. 4, pp. 1–8, 2013. [Online]. Available: http://dx.doi.org/10.3389/fneur.2013.00140

[35] M. Shams, S. Shams, and M. Wintermark, “What is new in imaging of acute stroke ?” *Intensive Care Med.*, vol. 46, pp. 1453–1456, 2020. [Online]. Available: https://doi.org/10.1007/s00134-020-06070-x

[36] Encyclopaedia Britannica. Tumour. Accessed 15-April-2021. [Online]. Available: https://www.britannica.com/science/tumor

[37] Center for Biomedical Image Computing Analytics. Multimodal Brain Tumor Segmentation Challenge 2020: Scope. Accessed 24-March-2021. [Online]. Available: https://www.med.upenn.edu/cbica/brats2020/

[38] B. M. Ellingson, M. Bendszus, J. Boxerman, D. Barboriak, B. J. Erickson, M. Smits, S. J. Nelson, E. Gerstner, and B. Alexander, “Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials,” *Neuro-Oncology*, vol. 17, no. 9, pp. 1188–1198, 2015. [Online]. Available: http://dx.doi.org/10.1093/neuonc/nov095

[39] A. I. Qureshi, D. A. Mendelow, and D. F. Hanley, “Intracerebral haemorrhage,” *Lancet*, vol. 373, no. 9675, pp. 1632–1644, 2011. [Online]. Available: http://dx.doi.org/10.1016/S0140-6736(09)60371-8

[40] R. Sahni and J. Weinberger, “Management of intracerebral hemorrhage,” *Vasc Health Risk Manag*, vol. 3, no. 5, pp. 701–709, 2007.

[41] C. S. Kidwell and M. Wintermark, “Imaging of intracranial haemorrhage,” *Lancet Neurol*, vol. 7, no. 3, pp. 256–267, 2008. [Online]. Available: http://dx.doi.org/10.1016/S1474-4422(08)70041-3

[42] F. M. Siddiqui, S. V. Bekker, and A. I. Qureshi, “Neuroimaging of hemorrhage and vascular defects,” *NeuroRx*, vol. 8, no. 1, pp. 28–38, 2011. [Online]. Available: http://dx.doi.org/10.1007/s13311-010-0009-x

[43] Z. Wang, E. Wang, and Z. Ying, “Image segmentation evaluation : a survey of methods,” *Artificial Intelligence Review*, vol. 53, pp. 5637–5674, 2020. [Online]. Available: https://doi.org/10.1007/s10462-020-09830-9

[44] R. Cattell, S. Chen, and C. Huang, “Robustness of radiomic features in magnetic resonance imaging : review and a phantom study,” *Visual Computing for Industry, Biomedicine, and Art*, vol. 2, no. 19, 2019. [Online]. Available: http://dx.doi.org/10.1186/s42492-019-0025-6

[45] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, “A lifelong learning approach to brain MR segmentation across scanners and protocols,” 2018, arXiv:1805.10170 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1805.10170

[46] R. Gauriau, C. Bridge, L. Chen, F. Kitamura, N. A. Tenenholtz, J. E. Kirsch, K. P. Andriole, M. H. Michalski, B. C. Bizzo, and C. Bridge, “Using DICOM Metadata for Radiological Image Series Categorization : a Feasibility Study on Large Clinical Brain MRI Datasets,” *J. Digit. Imaging*, vol. 33, no. 3, pp. 747–762, 2020. [Online]. Available: https://doi.org/10.1007/s10278-019-00308-xUsing

[47] A. P. Brady, , and E. Neri, “Artificial intelligence in radiology — ethical considerations,” *Diagnostics*, vol. 10, p. 231, 2020. [Online]. Available: http://dx.doi.org/10.3390/diagnostics10040231

[48] B. Allen Jr, S. E. Seltzer, C. P. Langlotz, and K. P. Dreyer, “A Road Map for Translational Research on Artificial Intelligence in Medical Imaging : From the 2018 National Institutes of Health / RSNA / ACR / The Academy Workshop Concept to Market,” *JACR*, vol. 16, no. 9, pp. 1179–1189, 2019. [Online]. Available: https://doi.org/10.1016/j.jacr.2019.04.014

[49] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, pp. 203–211, 2021. [Online]. Available: http://dx.doi.org/10.1038/s41592-020-01008-z

[50] I. R. I. Haque and J. Neubert, “Deep learning approaches to biomedical image segmentation,” *IMU*, vol. 18, 2020. [Online]. Available: https://doi.org/10.1016/j.imu.2020.100297

[51] J. Cho, K.-s. Park, M. Karki, E. Lee, S. Ko, J. K. Kim, D. Lee, J. Choe, J. Son, M. Kim, S. Lee, J. Lee, C. Yoon, and S. Park, “Improving sensitivity on identification and delineation of intracranial hemorrhage lesion using cascaded deep learning models,” *J. Digit. Imaging*, vol. 32, pp. 450–461, 2019.

[52] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *LNCS*, vol. 9351, pp. 234–241, 2015.

[53] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14539

[54] J. S. Marques, “Neural Networks,” Instituto Superior Tecnico, Lisbon, Tech. Rep., 2017.

[55] S. Sharma, S. Sharma, and A. Athaiya, “Activation functions in neural networks,” *IJETMAS*, vol. 4, no. 12, pp. 310–316, 2020.

[56] R. Ñanculef, P. Radeva, and S. Balocco, “Training Convolutional Nets to Detect Calcified Plaque in IVUS Sequences,” in *Intravascular Ultrasound*, S. Balocco, Ed. Elsevier, 2020, ch. 9, pp. 141–158. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128188330000096

[57] S. Li. Simple introduction about hourglass-like model. Accessed 15-April-2021. [Online]. Available: https://medium.com/@sunnerli/simple-introduction-about-hourglass-like-model-11ee7c30138

[58] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," 2017, arXiv:1612.04642 [cs.CV]. [Online]. Available: http://arxiv.org/abs/1612.04642

[59] R. E. Turner, "Representational learning in sensory cortices : connecting receptive fields to natural scene statistics," Gatsby Computational Neuroscience Unit, London, Tech. Rep., 2013. [Online]. Available: http://www.gatsby.ucl.ac.uk/~turner/teaching/4g3/2013/stats-recep-field.pdf

[60] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *PAMI*, vol. 39, no. 4, pp. 640–651, 2017. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2016.2572683

[61] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and Reproducibility in Deep Learning for Medical Image Segmentation," *Sci. Rep*, vol. 10, pp. 1–16, 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-69920-0

[62] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation : analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 29, 2015. [Online]. Available: http://dx.doi.org/10.1186/s12880-015-0068-x

[63] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. Lee, N. Aung, E. Lukaschuk, M. Sanghvi, F. Zemrak, K. Fung, J. Paiva, and V. Carapella, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *JCMR*, vol. 20, no. 65, 2018. [Online]. Available: https://doi.org/10.1186/s12968-018-0471-x

[64] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images : a review," *Artif. Intell. Rev.*, vol. 54, pp. 137–178, 2020. [Online]. Available: https://doi.org/10.1007/s10462-020-09854-1

[65] N. Heller, F. Isensee, K. H. Maier-hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han, G. Yao, Y. Gao, Y. Zhang, Y. Wang, and F. Hou, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the kits19 challenge," 2020, arXiv:1912.01054 [eess.IV]. [Online]. Available: https://arxiv.org/abs/1912.01054

[66] Y. Kabir, M. Dojat, B. Scherrer, F. Forbes, and C. Garbay, "Multimodal MRI segmentation of ischemic stroke lesions," in *29th Conf Proc IEEE Eng Med Biol Soc*. IEEEXplore, 2007, pp. 1595–1598. [Online]. Available: http://dx.doi.org/10.1109/IEMBS.2007.4352610

[67] L. Chen, P. Bentley, and D. Rueckert, “Fully automatic acute ischemic lesion segmentation in dwi using convolutional neural networks,” *NeuroImage Clin*, vol. 15, pp. 633–643, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.nicl.2017.06.016

[68] E. Beede, A. Iurchenko, L. Wilcox, and L. M. Vardoulakis, “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376718

[69] A. Saha, X. Yu, D. Sahoo, and M. A. Maciej, “Effects of MRI scanner parameters on breast cancer radiomics,” *Expert Syst. Appl*, vol. 87, pp. 384–391, 2018. [Online]. Available: https://doi.org/10.1016/j.eswa.2017.06.029

[70] M.-S. Badea, I.-I. Felea, L. M. Florea, and C. Vertan, “The Use of Deep Learning in Image Segmentation , Classification and Detection,” 2016, arXiv:1605.09612 [cs.CV]. [Online]. Available: http://arxiv.org/abs/1605.09612

[71] D. Pustina, H. B. Coslett, P. E. Turkeltaub, N. Tustison, M. F. Schwartz, and B. Avants, “Automated segmentation of chronic stroke lesions using linda: Lesion identification with neighborhood data analysis.” *Hum. Brain Mapp.*, vol. 37, no. 4, pp. 1405–1421, 2016. [Online]. Available: http://dx.doi.org/10.1002/hbm.23110

[72] A. Myronenko, “3D MRI brain tumor segmentation using autoencoder regularization,” 2018, arXiv:1810.11654 [cs.CV]. [Online]. Available: http://arxiv.org/abs/1810.11654

[73] M. Hssayeni, M. Al-Janabi, A. Salman, H. Al-khafaji, Z. Yahya, and B. Ghoraani, “Intracranial hemorrhage segmentation using a deep convolutional model,” *Data*, vol. 5, p. 14, 2020. [Online]. Available: http://dx.doi.org/10.3390/data5010014

[74] P. D. Chang, E. Kuoy, J. Grinband, B. D. Weinberg, M. Thompson, R. Homo, J. Chen, H. Abcede, M. Shafie, L. Sugrue, C. G. Filippi, M.-Y. Su, W. Yu, C. Hess, and D. Chow, “Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT,” *AJNR*, vol. 39, no. 9, pp. 1609–1616, 2018. [Online]. Available: http://dx.doi.org/10.3174/ajnr.A5742

[75] Z. Jiang, C. Ding, M. Liu, and D. Tao, “Two-stage cascaded u-net : 1st place solution to brats challenge 2019 segmentation task,” in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2019*. Springer, 2020, pp. 231–241.

[76] K. Kamnitsas, W. Bai, E. Ferrante, S. Mcdonagh, and M. Sinclair, “Ensembles of multiple models and architectures for robust brain tumour segmentation,” 2017, arXiv:1711.01468 [cs.CV]. [Online]. Available: http://arxiv.org/abs/1711.01468

[77] F. Isensee. (2020) nnU-Net. Accessed 28-December-2020. [Online]. Available: https://github.com/MIC-DKFZ/nnUNet

[78] J. S. Whang, M. Kolber, D. K. Powell, and E. Libfeld, “Diffusion-weighted signal patterns of intracranial haemorrhage,” *Clinical Radiology*, vol. 70, no. 8, pp. 909–916, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.crad.2015.04.006

[79] B. Efron and R. Tibshirani, “Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy,” *Statistical Science*, vol. 1, no. 1, pp. 54–75, 1986. [Online]. Available: https://doi.org/10.1214/ss/1177013815

[80] M.-t. Puth and M. Neuh, “On the variety of methods for calculating confidence intervals by bootstrapping,” *J Animal Ecology*, vol. 84, pp. 892–897, 2015. [Online]. Available: http://dx.doi.org/10.1111/1365-2656.12382

[81] J. Demsar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[82] Python Software Foundation, “Python.” [Online]. Available: https://www.python.org/download/releases/3.0/

[83] J. L. Lancaster and M. J. Martinez, “Mango.” [Online]. Available: http://rii.uthscsa.edu/mango/

[84] Stroke Alliance for Europe. Data Comparison - The Burden of Stroke in Europe Report. Accessed 08-May-2021. [Online]. Available: https://strokeeurope.eu/data-comparison/

[85] Direção Geral da Saúde. Norma nº 015/2017 - Via Verde do Acidente Vascular Cerebral no Adulto. Accessed 05-May-2021. [Online]. Available: https://normas.dgs.min-saude.pt/2017/07/13/via-verde-do-acidente-vascular-cerebral-no-adulto/

[86] World Health Organization. Guidelines for Management of Stroke. Accessed 08-May-2021. [Online]. Available: https://extranet.who.int/ncdccs/Data/MNG_D1_1.%20Clinical%20guideline%20of%20Acute%20Stroke%20.pdf

A

ML Methods / Hardware and Software Information

Appendix A brings some additional information on training specificities of *Apollo* (Section A.1) and *nnU-Net* (Section A.2).

Apollo and *nnU-Net* were trained with the same data, 853 patients for training and 214 for validation. For each patient, three types of anisotropic images were required: DWI, FLAIR, SWI, SWAN, or T2 * GRE. A linear resampling of the MRI and annotated images was undertaken prior to the training to a 1x1x1 mm^3 space. Four classes are specified: label 0 (L0) (background), L1 (infarcts), L2 (tumors) and L3 (hemorrhages). Chronic infarcts (class 4) were included in the training process but excluded from the testing process.

A.1 *Apollo*

Apollo was trained and weights were obtained prior to the start of the internship. Each network took a total of three days to train. Relevant training material is reported in the current section, for curiosity, but choices regarding the training schedule do not fall into this thesis framework.

Apollo is an ensemble of three U-Nets, working in different reference modality spaces. The entire 3D image was fed into the each U-Net, after an intensity z-scoring normalization, allowing for more contextual information and higher segmentation performance. Patch size was set to [160,176,160]. Large patch sizes implies decreasing the batch size to 2 which results in noisier gradients during back propagation [49]. Each network was trained for 250 epochs (with early stopping), each epoch iterating over 500 minibatches. Weights were initialized randomly. Adam Optimization was implemented with a dice loss and an initial learning rate of 0.0001. Data augmentation was also undertaken [1].

A.2 *nn-UNet*

Nn-UNet [49] pipeline configuration is illustrated in Figure A.1. Like *Apollo*, our *nnU-Net* is the ensemble of three U-Nets, working in different reference modality spaces.

Regarding the **data fingerprints**, each network is trained with the exactly same datasets and same labels used in *Apollo* (Section 3.3).

Regarding the **pipeline fingerprints**, some training parameters are predefined (blueprints in blue) and do not depend on the training set. Independently of the dataset, each U-Net is trained for 1000 epochs; one epoch iterating over 250 minibatches. Stochastic gradient descent with Nesterov momentum ($\mu = 0.99$) and an initial learning rate of 0.01 is used for learning network weights. *nnU-Net* is trained with deep supervision: for the last four blocks of the decoder, a downsampled segmentation map is created and used for loss computation. The training objective is to minimize the weighted sum of all losses L_x, resulting in (A.1):

$$L = \omega_1 L_1 + \omega_2 L_2 + \omega_3 L_3 + \omega_4 L_4. \tag{A.1}$$

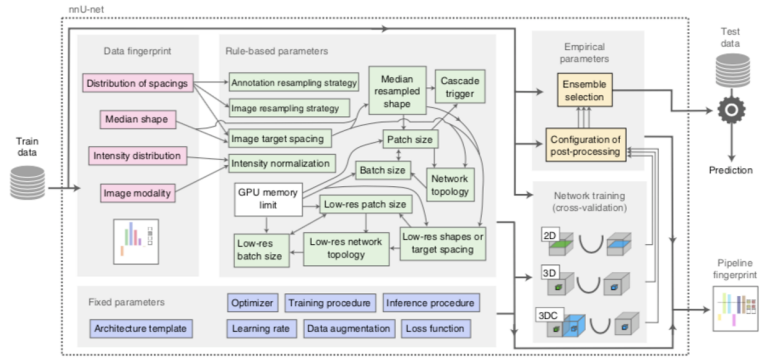
Higher weights are given to the losses computed with the higher resolution maps. Each L_x is the sum of cross-entropy (including the background) and Dice loss (excluding the background). Empirically, combining the dice loss with a cross-entropy loss improves training stability and segmentation accuracy. Data augmentation is also completed stochastically during training [2]. Furthermore, to avoid ignoring rare classes, oversampling foreground regions is also applied.

The empirical parameters (in yellow) are set after the training of each U-Net and require multiple con-

[1] Linear deformations (rotation and translations), non-linear deformations, intensity variation or random cropping to the available training images and random switching of diffusive channels, were implemented.

[2] Rotation, translation, scaling, low resolution simulation, Gaussian blur, Gaussian noise, intensity variation (brightness, contrast), simulation of low resolution, gamma correction, mirroring were undertaken.

figurations to be trained on a five-fold cross-validation. In our case, to mimic *Apollo*, only the 3D full resolution with one-fold cross-validation was trained, and empirical parameters could not be evaluated. Hence, only the inferred parameters (in green) are specific to our data and need to be specified. In the present case, patch size was set to [128 128 128] and batch size to 2. Moreover, as in *Apollo* pipeline, a pre-processing step is usually performed to decrease the dependencies on voxel spacing and image contrast. As the data fed into the networks had already been resampled to a 1x1x1 *mm* space, no re-sampling was undertaken by *nnU-Net*, according to the heuristic rules on resampling strategies. Hence, only intensity z-scoring normalization and one-hot encoding conversion of the annotated images were conducted.



Design choice	Required input	Automated (fixed, rule-based or empirical) configuration derived by distilling expert knowledge (more details in online methods)
Learning rate	–	Poly learning rate schedule (initial, 0.01)
Loss function	–	Dice and cross-entropy
Architecture template	–	Encoder–decoder with skip-connection ('U-Net-like') and instance normalization, leaky ReLU, deep supervision (topology-adapted in inferred parameters)
Optimizer	–	SGD with Nesterov momentum ($\mu = 0.99$)
Data augmentation	–	Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring
Training procedure	–	1,000 epochs × 250 minibatches, foreground oversampling
Inference procedure	–	Sliding window with half-patch size overlap, Gaussian patch center weighting
Intensity normalization	Modality, intensity distribution	If CT, global dataset percentile clipping & *z* score with global foreground mean and s.d. Otherwise, *z* score with per image mean and s.d.
Image resampling strategy	Distribution of spacings	If anisotropic, in-plane with third-order spline, out-of-plane with nearest neighbor Otherwise, third-order spline
Annotation resampling strategy	Distribution of spacings	Convert to one-hot encoding → If anisotropic, in-plane with linear interpolation, out-of-plane with nearest neighbor Otherwise, linear interpolation

Image target spacing	Distribution of spacings	If anisotropic, lowest resolution axis tenth percentile, other axes median. Otherwise, median spacing for each axis. (computed based on spacings found in training cases)
Network topology, patch size, batch size	Median resampled shape, target spacing, GPU memory limit	Initialize the patch size to median image shape and iteratively reduce it while adapting the network topology accordingly until the network can be trained with a batch size of at least 2 given GPU memory constraints. for details see online methods.
Trigger of 3D U-Net cascade	Median resampled image size, patch size	Yes, if patch size of the 3D full resolution U-Net covers less than 12.5% of the median resampled image shape
Configuration of low-resolution 3D U-Net	Low-res target spacing or image shapes, GPU memory limit	Iteratively increase target spacing while reconfiguring patch size, network topology and batch size (as described above) until the configured patch size covers 25% of the median image shape. For details, see online methods.
Configuration of post-processing	Full set of training data and annotations	Treating all foreground classes as one; does all-but-largest-component-suppression increase cross-validation performance? Yes, apply; reiterate for individual classes No, do not apply; reiterate for individual foreground classes
Ensemble selection	Full set of training data and annotations	From 2D U-Net, 3D U-Net or 3D cascade, choose the best model (or combination of two) according to cross-validation performance

Figure A.1: Proposed automated method configuration for deep learning-based biomedical image segmentation. Given a new segmentation task, dataset properties are extracted in the form of a dataset fingerprint. A set of heuristic rules models parameters interdependencies and operates on this fingerprints to infer the data-dependent rule based parameters (or inferred fingerprints) (**green**) of the pipeline. These are complemented by the fixed parameters (or blueprint fingerprints) (**blue**) which are predefined and do not required adaptation. Up to three configurations are trained in a 5-fold cross-validation. Finally, nnU-Net automatically performs empirical selection of the optimal ensemble of these models and determines whether post-processing is required (empirical parameters) (**yellow**).
The image and descriptions are excerpted from [49].

B

Acquisition Parameters

Appendix B is devoted to the selection of the acquisition parameters for whom impact on performance will be assessed. Selection is performed under two criteria that maximize the number of patients per groups [1] and enable a representative analysis :

- Selected parameters should be described in both DICOM (58.97 % of the in-house dataset) and NIFTI (41.03 % of the in-house dataset) headers.
- Selected parameters should be distributed evenly across groups.

[1] A group is one sub-division of the selected acquisition parameter pool of values. As an example, taking B_0 as selected parameter, groups could be drawn as follow: group 1 (1.5 T) and group 2 (3.0 T).

Groups are presented in Figure B.1 for each acquisition parameter and the selection process is described below:

- B_0 and **scanner manufacturer** do not comply with any of the two criteria. One unique group can be drawn from these parameter and no information is given for the NIFTI format.
- **TR**, **TE**, and **flip angle** do not respect the first criterion. In addition, their dispersion depends on the sequence type, and within each sequence type, the range of values is not broad enough to draw meaningful groups. From Figure B.1, TE presents a wider dispersion. However, this dispersion is not evenly distributed: 62.2% of the values are equal to 7.7 *ms* and 33.33% to 19 *ms*.
- An interesting parameter is the **voxel dimension**. For each acquisition technique, the higher resolution does not show relevant disparities (Figure B.1). In turn, the lower resolution plane presents a larger deviation. Nevertheless, intervals are hard to draw, as not homogeneously distributed, and prevent to continue with voxel dimension in the analysis. For instance, SWI condenses 62.3% of the values in $z_{dim} = 4$ *mm* and T2 * GRE presents 60.5 % of the values in $z_{dim} = 6.5$ *mm*.
- The unique suitable parameter that satisfies our criteria of selection is **Sequence type and orientation**. Its distribution in the in-house dataset can be find in Figure 3.5 and groups are described in Section 3.3.2.

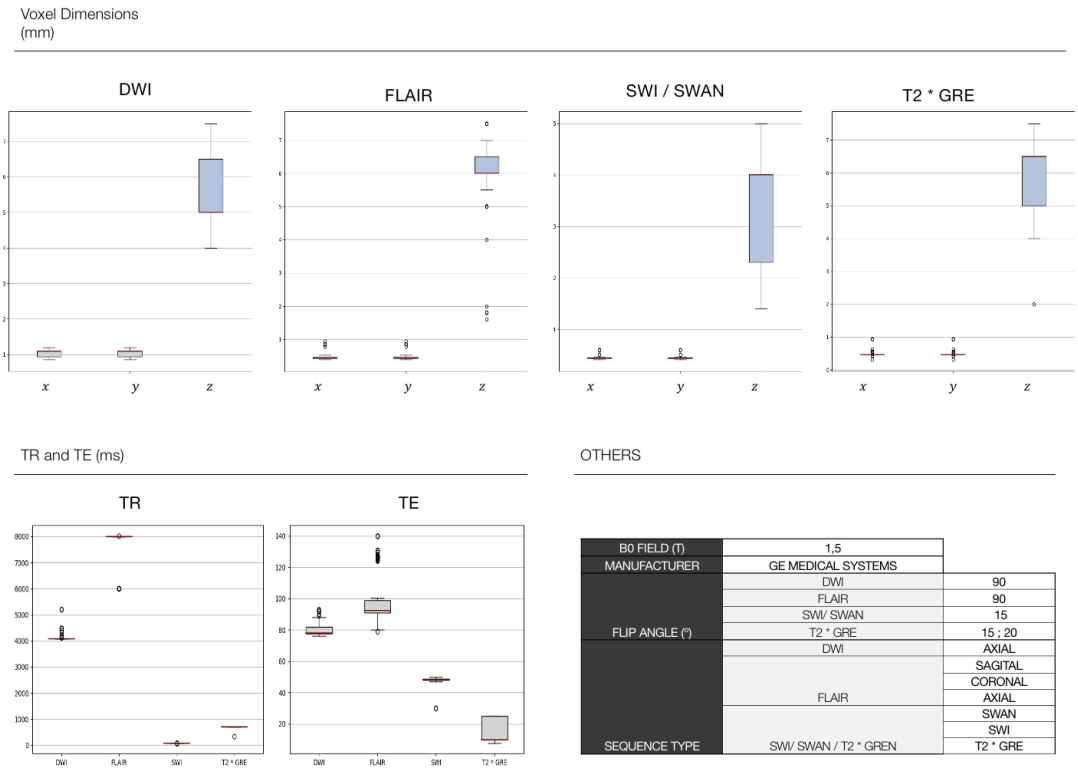


Figure B.1: Dispersion of MRI acquisition parameters in the in-house dataset.

C

Post-processing Results

Appendix C focuses on the post-processing experiments: filtering (Section C.1), dilation (Section C.2), and bounding boxes (Section C.3). The main purpose of this section is to support Sections 3.4 and 4.1, giving additional motivation and presenting supplementary results, not shown in the main corpus.

C.1 Filtering

Considering filtering as an additional step of our evaluation pipeline is motivated in Figure C.1. From the filtering experiment described in Section 3.4.1, lesion level metrics (recall and precision) are shown in Figure C.2 across the filtering thresholds A_{HP}.

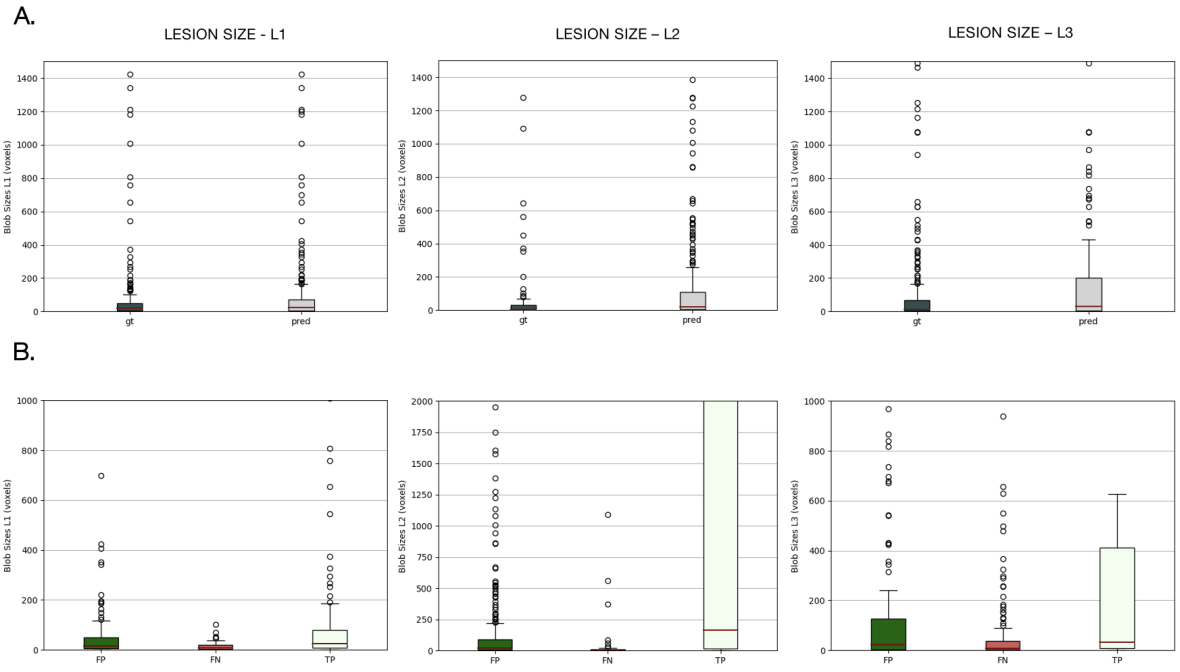


Figure C.1: Boxplots of lesion size distributions across labels.
A) ground truth versus predicted lesion sizes. Filtering is needed in both annotations and predictions: the median size (red line) of annotated and predicted lesions is low.
B) TP (light green) versus FP (dark green) and FN (dark pink) lesion sizes. Filtering is expected to affect more significantly FP and FN lesions. Mean and median lesion sizes for these categories is lower than TP mean and median lesion sizes.

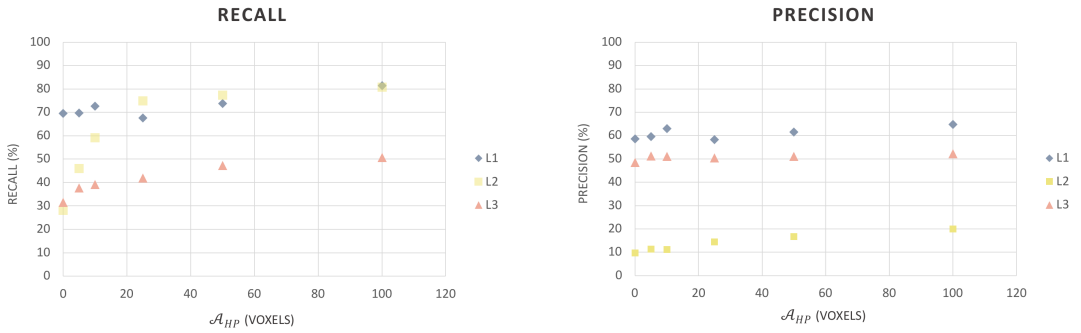


Figure C.2: *Apollo* (**Left.**) recall and (**Right**) precision in % as a function A_{HP} in the in-house dataset.

C.2 Dilation

The confusion matrix at a lesion level for the dilation experiment (described in Section 3.4.2) is shown in Figure C.3.

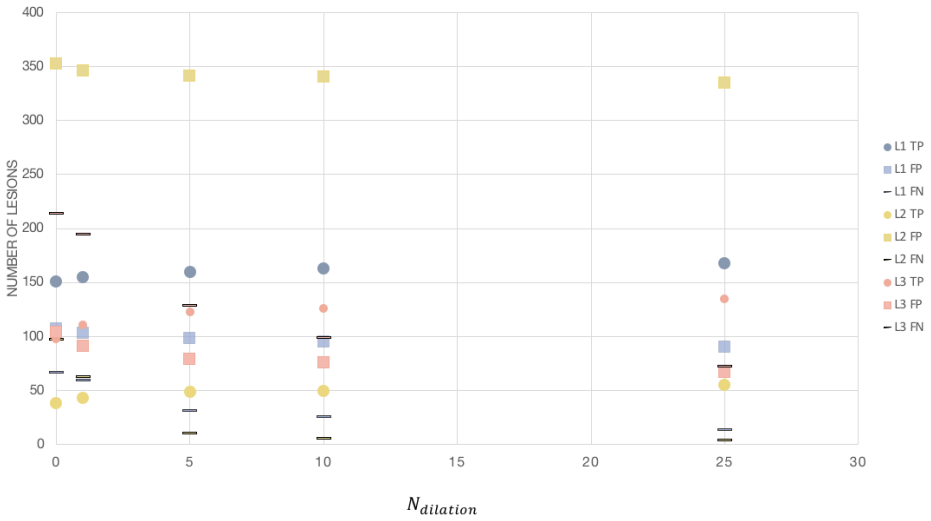


Figure C.3: Confusion matrix at a lesion level across $N_{dilation}$.

C.3 Bounding Boxes

Bounding boxes experiments arose from the discrepancy between the number of predicted and ground truth lesions (Figure C.4). Higher predicted lesions could have been explained by the network splitting one ground truth lesion in several predictions. This hypothesis was latter rejected: for each annotation, only one lesion was segmented by the network, and this is verified across labels. The observed discrepancy is due to the high number of FP lesions, predominant for L2 (as referred in Section 4.2). Bounding boxes were created as described in Section 3.4 and the results are compared to the raw data and 5 times dilation performances. A visual inference of lesions across the three different analysis is presented in Figure C.5 and results are shown in Figure C.6.

At an image level, sensitivity and specificity are not influenced by dilation nor by the use of bounding boxes (Section 4.1). However, at a lesion level, recall and precision show an improvement across labels when using dilation and bounding boxes. Nevertheless, bounding boxes are associated with lower recall (except for L3) and lower precision across pathologies. Therefore, the use of bounding boxes in our evaluation framework was rejected.

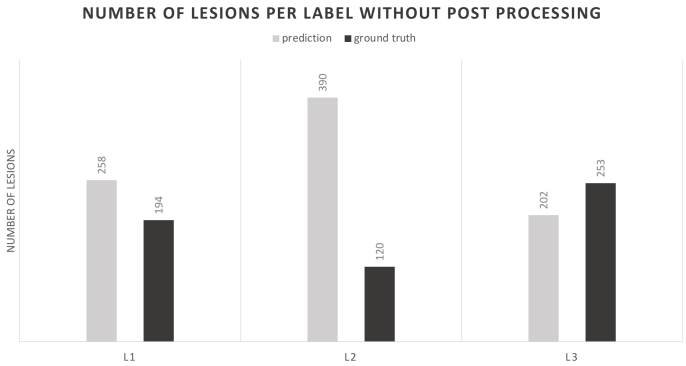


Figure C.4: Number of lesions per label: predicted Lesions (light grey) versus ground truth lesions (dark). Except for L3, higher number of lesions are predicted compared to ground truth.

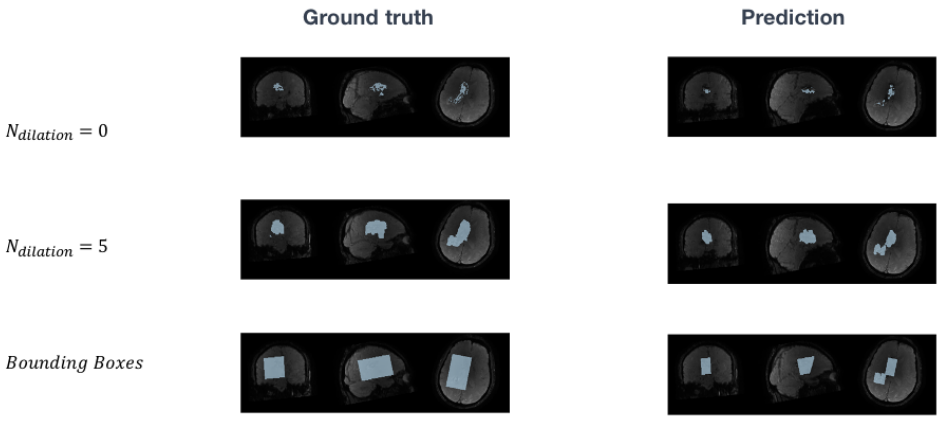


Figure C.5: Comparison of post-processing *versus* no post-processing for hemorrhage predictions.
(**Top.**) without any post-processing; (**Middle.**) with 5 times dilation; (**Bottom.**) with bounding boxes creation, 5 times dilation and morphological closing.

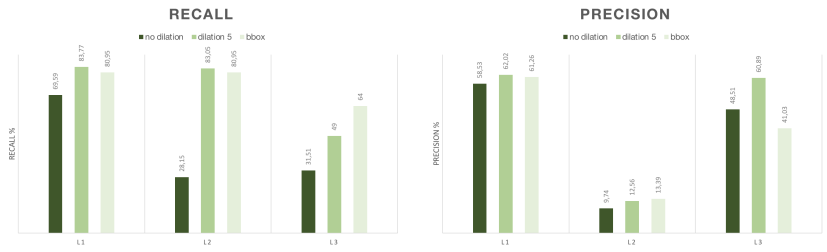


Figure C.6: Bounding boxes performance at a lesion level.
(**Left.**) Recall and (**Right.**) precision in % across labels for no dilation experiment (grey), dilation 5 times experiment (dark green), and bounding boxes experiment (light green). Bounding Boxes are relevant only for L3 in terms of recall, as it shows a sharper decrease of FN lesions, compared to the other two methods. L3 is characterized by its smaller and dispersed ground truth annotations.

D

Apollo-nnU-Net: Additional findings

Appendix D gives further information on *Apollo* (Section D.1) and *nnU-Net* (Section D.1) respective performances. Results are shown after post-processing and CI (computed via bootstrap) are referred as $[\alpha; \beta]$.

D.1 *Apollo*

Apollo evaluation in the external dataset is reported in Table D.1. Poor performance is seen across labels and metrics. In particular, specificity and precision are extremely low, and *Apollo* seems to over-predict lesions for all classes. Large CI highlight once again the inconsistency of the results.

Table D.1: Apollo evaluation across pathologies for the external datasets.

	L1	L2	L3
sensitivity (%)	85.71 [70.27 ; 94.44]	60.00 [22.22 ; 87.5]	73.33 [42.86 ; 92.31]
specificity (%)	7.41 [0.00 ; 23.81]	15.38 [7.27 ; 26.92]	21.28 [10.87 ; 34.69]
mean Dice Coefficient	0.281 [0.193 ; 0.380]	0.076 [0.030 ; 0.158]	0.132 [0.067 ; 0.226]
mean Hausdorff Distance (mm)	40.11 [30.36 ; 52.97]	42.34 [33.69 ; 47.83]	92.20 [48.90 ; 186.29]
recall (%)	79.19 [67.28; 88.16]	81.82 [42.86 ; 96.55]	89.39 [70.59 ; 98.09]
precision (%)	23.6 [14.14 ; 37.21]	9.41 [3.53 ; 18.97]	27.51 [10.87 ; 53.73]

D.2 *nnUNet*

nnU-Net evaluation in the in-house and external datasets are reported in Table D.2 - *Top* and Table D.2 - *Bottom*, respectively. In the **in-house** dataset, *nnU-Net* shows a more homogeneous performance across labels, and results are improved when compared with *Apollo*. Nevertheless, similar standard deviations are recorded in the metric distributions in both networks, showing that performance is highly dependent on the data pool selected for bootstrap. In the **external** dataset, *nnU-Net* manage to keep acceptable scores for sensitivity, specificity, recall, and precision in L1 and L3. It mainly struggles with L2 and seems to miss the ground-truth lesions. By including a smaller number of patients, the external dataset presents even broader CI than the in-house dataset.

Table D.2: *nnU-Net* evaluation across pathologies for the (**Top.**) In-house and (**Bottom.**) external datasets.

	L1	L2	L3
sensitivity (%)	96.00 [86.36 ; 100.00]	82.35 [56.25 ; 95.23]	95.83 [76.47 ; 100.00]
specificity (%)	89.66 [83.67;93.89]	91.01 [85.96 ; 94.48]	93.57 [88.82 ; 96.53]
mean Dice Coefficient	0.626 [0.553 ; 0.718]	0.273 [0.154 ; 0.423]	0.500 [0.361 ; 0.632]
mean Hausdorff Distance (mm)	27.39 [20.31; 36.30]	48.43 [29.03 ; 67.70]	44.16 [30.10 ; 61.29]
recall (%)	87.39 [77.48 ; 93.33]	72.73 [56.25 ; 87.50]	45.53 [22.97 ; 68.29]
precision (%)	72.22 [49.19 ; 83.33]	51.06 [29.78 ; 73.58]	60.87 [42.50 ; 75.36]
	L1	L2	L3
sensitivity (%)	91.43 [76.67 ; 97.37]	30.00 [8.33 ; 63.64]	86.67 [57.14 ; 100.00]
specificity (%)	70.37 [50.00 ; 85.71]	71,15 [57.41 ; 82.14]	61.10 [46.67 ; 74.51]
mean Dice Coefficient	0.496 [0.371 ; 0.615]	0.098 [0.026 ; 0.253]	0.263 [0.144 ; 0.410]
mean Hausdorff Distance (mm)	39.58 [29.52; 51.97]	46.59 [44.33 ; 45.27]	80.78 [40.23 ; 165.75]
recall (%)	81.62 [71.53 ; 88.43]	47.83 [0.00 ; 84.21]	84.06 [60.61 ; 96.05]
precision (%)	52.86 [35.19 ; 69.68]	13.58 [2.27 ; 38.95]	40.00 [17.73 ; 63.93]