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

Carlota de Macedo Santos
carlota.m.santos@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

July 2021

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 sub-optimal in the current practices.

Here, we propose a comprehensive 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 data curation. 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. 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.

1. Introduction

According to the European Union, the use of Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) has been rising [1], and radiologists are asked to interpret one scan every 3-4 seconds [2]. Accurate diagnosis and fast disease management are compromised by the high inter- and intra-clinicians variability and fatigue of medical staff [3, 2]. This is all the most unfortunate knowing that 60% of the acquired sequences are unnecessary [4]. By addressing these issues, deep learning (DL) has been garnering special attention in Radiology and is expected to play an essential role in the digital health revolution [5, 6]. Its underlying potential does not only arise from its capacity to handle massive amounts of data and alleviate the workload of radiologists, but also from its ability to discover relationships between scan features and patho-physiological attributes that may not be included in the radiologists lexicon [6, 7]. Promising outcomes encompass a better resources allocation (adequate triaging, no unnecessary admissions and reexaminations), an optimized hospital workflow, and higher quality of care (faster decisions, more confident diagnosis, and less invasive approaches) [4]. This is specially relevant for ischemic stroke that have a time-dependent nature and treatment selection (thrombolysis or mechanical thrombectomy) based on a correct estimation of the onset and detection of potential hemorrhages at the infarction site [8].

However, even if DL shows clear advantages in supporting healthcare providers, automating a human-based process is far from being trivial. Predicting diseases characterized by a broad landscape of patho-physiological attributes can be challenging. Figure 1 illustrates this idea by segmenting the different sub-types of tumors described in [9]. In that context and to avoid wrong predictions, it is important to guarantee that each

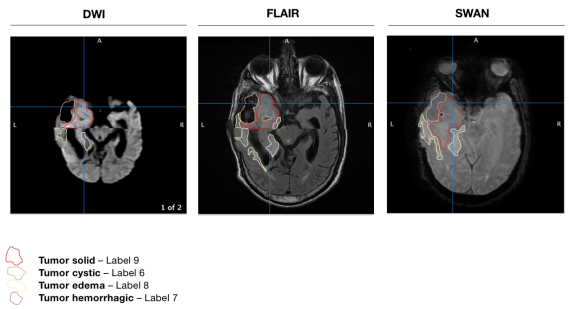

Figure 1: Tumors classification according to their composition. Solid tumor (red), cystic tumor (orange), tumor edema (yellow) and hemorrhagic tumor (purple) are identified in (**Left.**) DWI, (**Middle**) FLAIR, and (**Right.**) SWAN scans. Hemorrhagic tumor is only visible in SWAN modality. Shapes, textures, and intensities vary across tumor sub-types, depending on the underlying biological properties

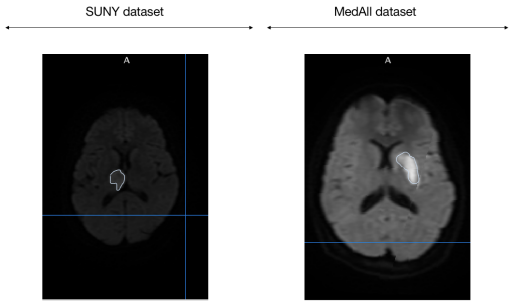

Figure 2: Example of an infarct acquired with ((**Left**) a 1.5 T Siemens scanner and (**Right**) a 1.5 T GE Healthcare scanner. The relative intensity difference between background and foreground classes depends on the scanner and its acquisition parameters.

pathology attributes is being learnt correctly by the network during the training. Challenges also arise from making predictions for out-of-distribution data (rare pathological conditions, data acquired with a different scanner, or demographic shift). Pathologies may appear differently according to the scanner acquisition parameters (Figure 2) and these differences may not be understood by the algorithms. As a result, they may not be able to generalize to other scanners, acquisition parameters, populations, or pathology characteristics [5, 10, 11, 6]. Other types of unintended data bias can arise from data curation (selection of optimal training data, high-quality images) and from the over-representation of positive cases [5]. Addressing these limitations is crucial to bring the technology into clinic and 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 [12, 10, 13]. DL algorithms need to be prepared to predict on previously unseen samples. However, since learning only happens during training, the quality of the predictions and generalization ability highly depend on the data attributes provided during training and validation phases [5].

Under these assumptions, some ethical and medico-legal concerns have been raised in case of mispredictions. Hence, it is important to assess DL clinical value, safety, and benefits before promoting the digitization and automation of healthcare processes [5]. In current practice, and due to the scarsity of medical data, most of the evaluation processes are not sufficiently broad. This is reported in [14] where only 6% of the 516 reviewed AI-based solutions performed external validation. In addition, evaluation frameworks are often not supported by a diversified and comprehensive set of metrics. Unfortunately, selected metrics do not always cover all the requirements for the integration of an algorithm in clinics [15].

Objectives In this work, we outline a new path of evaluation by designing a comprehensive framework for assessing the performance of DL medical segmentation and classification algorithms. Performance will be addressed under three perspectives:

- *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 MRI acquisition parameters* - Gives a deeper understanding of the MRI parameters that allow a better performance of DL algorithms at clinical sites. It can also be interpreted as a measure of robustness. In this work, MRI parameters include the type of sequence and the type of orientation.
- *Performance across datasets* - Gives a deeper understanding of DL algorithms generalization ability. It is a preliminary step to ensure their safety at new clinical sites before their implementation.

2. DL Solutions for Medical Segmentation and Classification

Among DL architectures, U-Net [16] is one of the most commonly adopted for medical image segmentation [17, 18]. Its name arises from its symmetric U-shaped architecture, as shown in Figure 3. It consists of a combination of an encoder, that ensures feature extraction and dimension reduction, and a decoder, that performs semantic segmentation [19]. 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. Skip connections (in grey in the Figure)

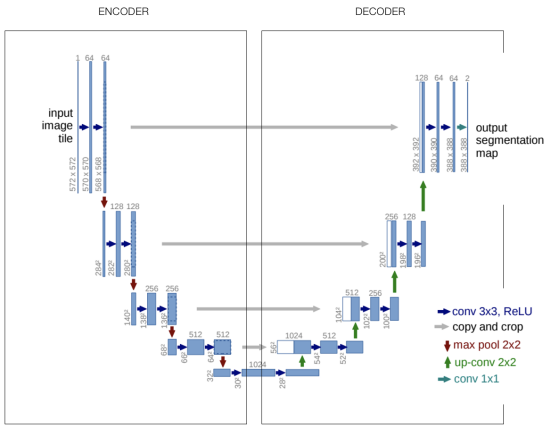

Figure 3: General U-Net architecture. The U shape is visible: encoder (contextual information) and decoder (spatial location) are identified. *Blue boxes* represent the multi-channels feature maps; *White boxes* represent the copied feature maps. Figure is adapted from [16].

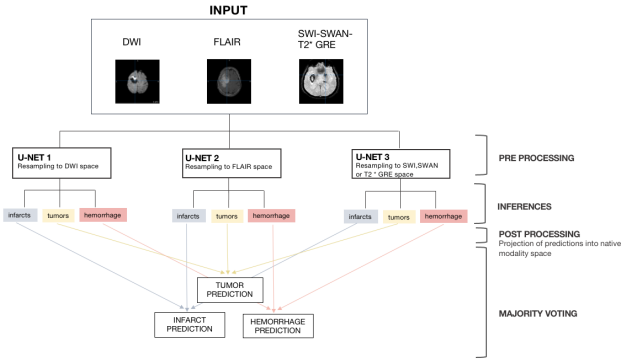

Figure 4: *Apollo* inferences process for infarcts, tumors, and hemorrhages segmentation. The segmentation maps predicted by the 3 U-Nets, working in different reference modality spaces (DWI, FLAIR, and T2*GRE, SWAN, or SWI) are combined by majority voting to create the final output.

guarantee that better segmentations are produced by recovering fine-grained spatial information that is potentially lost in the pooling and downsampling layers.

In order to face data scarcity and achieve state-of-the-art performance, U-Nets and DL networks in general are usually paired with data augmentation [17]. It is all the most appreciable in medical 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 extensive training corpus [16].

More complex variations can be obtained, built upon the standard U-Net architecture, by adapting the U-Net like encoder-decoder skeleton and its convolutional building blocks, modifying the loss function, or following more sophisticated data augmentation techniques [20]. Two examples are described in Section 2.1 and Section 2.2 and compared in Table 1. They will be used as experimental frameworks for validating our approach.

2.1. Apollo

Apollo [4] is a DL solution developed by *Cerebriu*, a danish med-tech company based in Copenhagen. It is currently under clinical validation in hospitals across Denmark. Its architecture consists in ensembling three U-Nets, working in different reference modality spaces. Each network computes a multi label semantic classification and one segmentation map per task is predicted. These maps are combined by majority voting, after being projected in a pre-defined native space.

Apollo is currently designed to segment infarct, tumors, and hemorrhages. Its implementation at clinical sites requires the following sequences : a diffusion weighted imaging (DWI), a fluid attenuated inversion recovery (FLAIR) and a T2 * gradient echo (T2* GRE), a susceptibility-weighted angiography (SWAN), or a susceptibility-weighted imaging (SWI)). Its main objective is to grant triage decision support to radiologists. From the segmentation maps obtained as described in Figure 4), if at least one lesion of a specific class has been segmented, the patient is automatically classified with the same label. Based on the prioritization protocol set by the hospital, *Apollo* automatically selects the patients that require urgent review by the radiologist.

2.2. nnU-Net

An outstanding observation when addressing architectural modifications reveals that performance is not always correlated with innovative architectural designs and sophisticated modifications [21]. Achieving state-of-the-art mostly relies on choices made during network configuration, where the pipeline fingerprint is designed and parameters are selected.

From these findings, the ”no new net”, *i.e nnU-Net* [22], was created. The network does not draw its strengths from an improved architecture, a more efficient training scheme or a more appropriate loss function. Its novelty resides on its ability to handle a wide disparity of structures and image properties, proposing a tailor-made network without any user intervention. The design choices of a segmentation network are divided into three groups: blueprint (data-independent, already predefined, identified as robust common configurations), inferred (data dependent, selected based on heuristics rules), and empirical parameters (optimized during training). Having extracted its encoding design choices from large and diverse

Table 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

data pool, performance is not deteriorated by data scarcity, making *nnU-Net* data efficient and a solid baseline for biomedical segmentations.

3. Materials and Methods

We introduce a complete evaluation framework for DL solutions (Section 3.1), built upon a deep understanding of their clinical applications. By jointly assessing the performance at the image and lesion levels, it is intended to cover segmentation and classification outcomes simultaneously. In order to validate our approach experimentally, experiments are conducted on the two previously described algorithms, *Apollo* and *nnU-Net* and exposed in Section 3.2.

3.1. Evaluation Framework

The following metrics are selected considering the intentions and objectives of the two networks in the radiology department. They are computed based on the calculation of a confusion matrix. Components are abbreviated as followed: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). At an image level, **confusion matrix**, **sensitivity** (1) in % , and **specificity** (2) in % are computed, with higher relevance granted to sensitivity:

$$sensitivity = \frac{TP}{TP + FN} \times 100, \tag{1}$$

$$specificity = \frac{TN}{TN + FP} \times 100. \tag{2}$$

At a lesion level, **confusion matrix** (without TN lesions, as considered non-existent), **recall** (*i.e* sensitivity [1]) (1) in %, and **precision** (3) in % are preferred.

$$precision = \frac{TP}{TP + FP} \times 100 \tag{3}$$

Mean of **Sørensen Dice coefficient** (4) and **Hausdorff distance** (5) in *mm* are also estimated to address the quality of the segmentation in terms of overlap and contours delineation [23]:

$$DSC = 2 \times \frac{X \cap Y}{X \cup Y}, \tag{4}$$

$$d_H(X, Y) = max\left\{A, \overline{A}\right\}, \tag{5}$$

where X and Y are the ground truth and segmentation maps, $A = max_{x \in X} min_{y \in Y} d(x, y)$, and $\overline{A} = max_{y \in Y} min_{x \in X} d(x, y)$

[1] Different nomenclatures are used to distinguish between image and lesion levels.

The confusion matrix at a lesion level is estimated through the Dice score. Results at an image level are built upon lesion level outcomes.

The aforementioned metrics are supported by the estimation of their confidence intervals (CI), computed with the bias-corrected and accelerated bootstrapping method [24, 25] and $\alpha = 0.95$. A total of $B = 10000$ resampling with replacement of the size of the original sample are made from the empirical probability distribution of the data under analysis [25]. As the original sample size is larger than 30, bootstrapping results are expected to be reliable [25].

3.2. Implementation

The evaluation framework is validated experimentally in two supervised learning algorithms, *Apollo* and *nnU-Net*. The evaluation encompasses three pathologies (infarcts, tumors, and hemorrhages), predicted on the MRI sequences required by *Apollo*. The evaluation framework is applied to each class individually, transforming a multi-task problem into a binary problem. The impact of the sequence type (SWI, SWAN, or T2 * GRE) is addressed for hemorrhage predictions and of the sequence orientation (Axial or Coronal FLAIR) for infarct predictions. Robustness is assessed by comparing performance in an in-house and external datasets.

Experiments are run on Python, version 3 [26] in a Linux virtual Machine, Ubuntu SMP, with a x86_64 processor.

Apollo was trained prior to the experiments. To mimic *Apollo*, three 3D *nnU-Net* [22] are trained on the same dataset with one fold cross-validation using one GPU (NVIDIA Tesla P40 - memory 22919 MiB) along with a strong CPU (Intel(R) Xeon(R) - size 2.6 GHz and capacity 3.5 GHz) and Pytorch (version 1.6). Predicted maps are obtained as detailed in Figure 4.

Datasets Data was curated and annotated for the use of *Cerebriu*. Networks were trained on MedAll (MedAll Diagnostics - Chennai, India) and OUH (Odense University Hospital - Odense, Denmark) data. The training set is composed by 25.67% infarct patients, 11.37% tumor patients, and 9.95% hemorrhage patients. The remainders are healthy patients. Performance evaluation is conducted on two datasets:

- an **in-house** dataset - 195 MedAll patients, selected among the validation set (*i.e* 91.12% of the validation set). The in-house set is composed by 25.64% infarct patients, 8.72% tumor patients, and 12.31% hemorrhage patients.
- an **external** dataset - 62 SUNY (Suny Upstate University Hospital - New York, United States) patients. The external set is composed by 58.06% infarct patients, 19.35% tumor patients, and 24.19% hemorrhage patients.

Figure 5 presents the spectrum of classes sub-types, while Figure 6 shows the distribution of the relevant MRI parameters across datasets .

Post Processing Post-processing on prediction and ground truth binary masks is 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}. $N_{dilation} = 5$ and $A_{HP} = 10$ are defined for *Apollo* using the in-house dataset and are kept throughout all the analysis. Calibration is rigorously undertaken to avoid the production of misleading or distorted results. Our analysis reveals that dilation and filtering helped increasing the fairness of the evaluation by allowing overlapping of spatially close lesions and removing noisy voxels on prediction and ground truth, respectively. Dilation is performed after filtering.

4. Results

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

Performance across Pathologies

Apollo performs differently across pathologies: Table 2 shows how *Apollo* addresses each task. When applying the model for L1 detection and segmentation, better performances are achieved and a higher trade-off between sensitivity/specificity and recall/precision is reached. In terms of quality of segmentation, infarcts area and contours are better captured by the network. The network seems to suffer from L3 FN and L2 FP lesions in its predictions. In particular, the noisy segmentation maps observed for L2 deeply affects its specificity and healthy patients are more likely to be reported as having a tumor than an infarct or a hemorrhage.

Differences in performance arise from unintended data bias: A trivial assumption is to explain divergence across labels in terms of their pathophysiological differences. By being characterized by a relatively high amount of spread lesions, L3 lesions are more difficult to detect and segment individually. Additionally, some hemorrhages sub-types are hard to segment manually and this has repercussions when automating the process. A clear example is subarachnoid hemorrhages (SAH) that are not detected by *Apollo* and for which the high variability in MRI intensities, the blooming effect produced by adjacent bones, or the dilution of

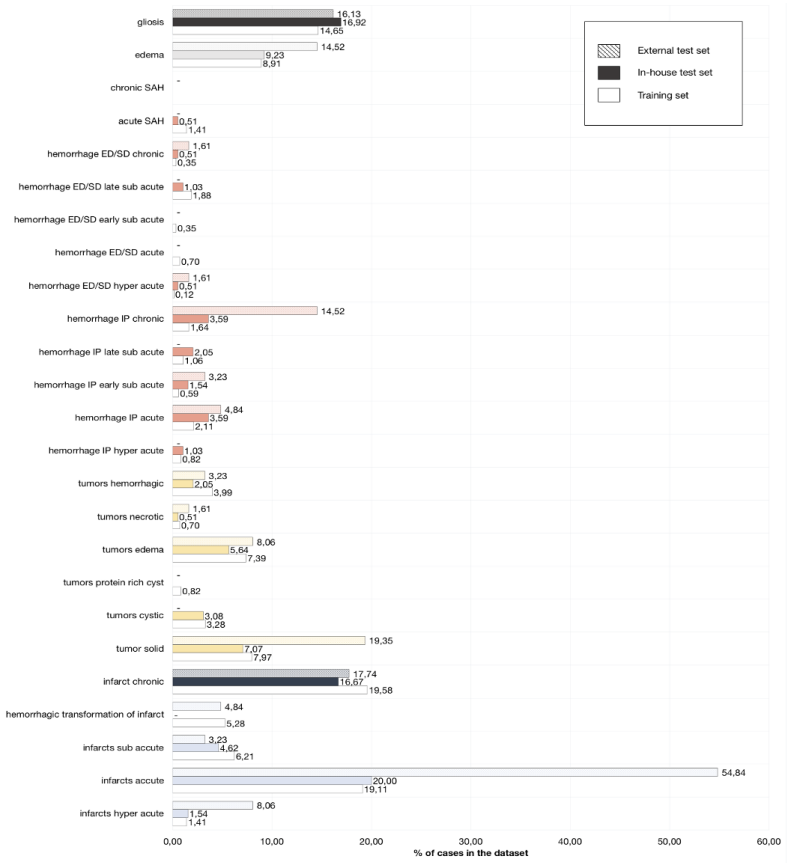

Figure 5: Percentage of classes sub-types in the training set (white), in-house testing set (coloured), and external testing set (coloured with dashed line).The percentage of disease k in the dataset j are given by $P_{k,j} = \frac{Np_{k,j}}{Np_j}$, where $Np_{k,j}$ is the number of patient presenting the disease k in the dataset j and Np_j is the total number of patients in the dataset j. Infarct, tumor, and hemorrhage sub-types appear in light blue, yellow, and pink, respectively.

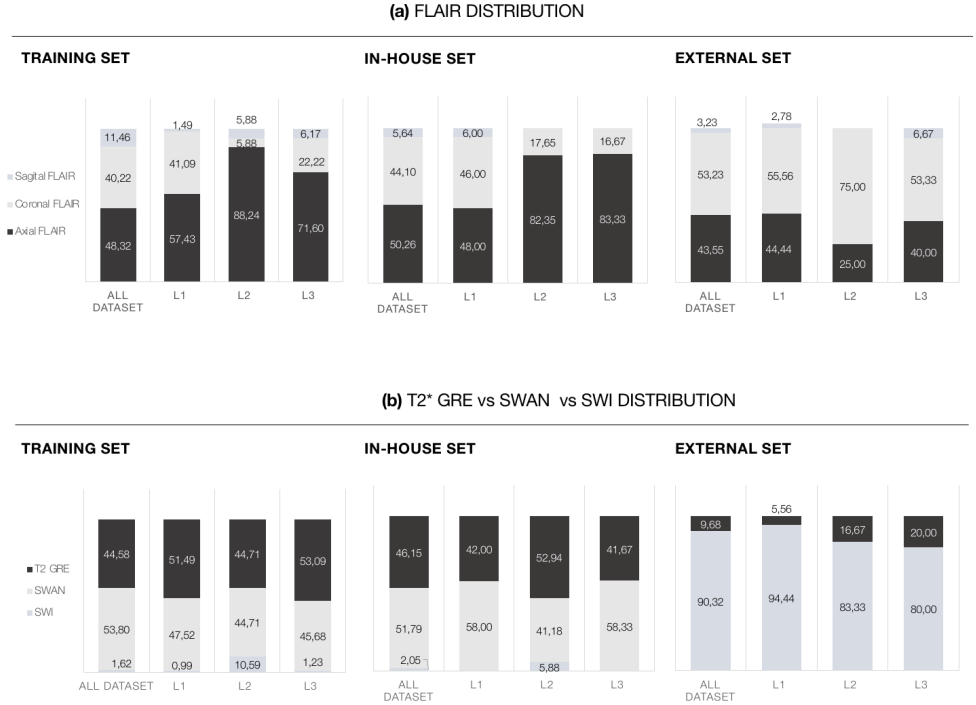

Figure 6: Distribution of the acquisition parameters in (**Left.**) the training set, (**Middle.**) the in-house set, and (**Right.**) the external set. (**a.**)FLAIR distribution: Axial, Coronal, and Sagittal ;(**b.**)T2 * GRE, SWI, and SWAN distribution.

Table 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 $[\alpha; \beta]$.

	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]

Table 3: *nnU-Net* evaluation across pathologies.

	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]

blood with cerebrospinal fluid (CSF) is known to jeopardize its diagnosis [27].
However, the aforementioned factor cannot explain by itself all the differences observed across pathologies. The higher scores obtained for L1 seem to stem from the training data distribution that accounts for 25.67 % infarcts *versus* 11.37 % tumors and 9.95 % hemorrhages. Conversely, poorly represented sub-types in the training set are correlated with higher percentages of missed detections, by having less opportunity to be learnt and extracted correctly by the network. Our analysis reveals that *Apollo* struggles in detecting hyper acute infarcts, cystic tumor, epidural/subdural late subacute hemorrhages, and epidural/subdural chronic hemorrhages.

Besides unbalanced class distributions, divergence in performance could also come 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-types range. Each sub-class is described by its own texture, size, and shape that originates a broader landscape of features, that needs to be learnt by the algorithm (Figure 1). This makes the classification task more difficult based on this extended diversity and could explain the high percentage of FP affecting L2 performance.

nnU-Net has a more homogeneous behaviour across pathologies: From Table 3, *nnU-Net* reaches an outstanding performance when compared with *Apollo*, and this behaviour is seen across pathologies. In particular, an increase of 15 % and 25 % are seen in L3 sensitivity and L2 specificity, while L2 precision is multiplied by 3.59. One one hand, it seems that *nnU-Net* managed to learn more discriminative features across classes. On the other hand, it seems that *nnU-Net* is less dependent on the low representation of pathologies in the training data. While misclassifications are mainly seen for the same sub-classes as *Apollo*, *nnU-Net* handles better SAH, epidural/subdural hyper-acute hemorrhages, intraparenchymal acute and chronic hemorrhages, and hyper-acute infarcts.

Consistency in performance is not verified: Confidence intervals (CI) appear to be unexpectedly broad and both networks score are highly inconsistent. The high variability of the data can be explained by the small sample size and their computation is deeply affected by the patients selected in each bootstrap.

Robustness across different sequence types or orientations

Recall and sensitivity are improved with Axial FLAIR: From Figure 7, by producing a higher recall and sensitivity, *Apollo* and *nnU-Net* seem to better perform when fed Axial FLAIR. Possible explanations could lean on the fact that the training set includes 1.4 more L1 patients with Axial FLAIR and on the hypothesis that better contextual information can be aggregated when inputs share the same low-dimensional direction.

*T2*GRE leads to higher quality L3 segmentations:* *Apollo* and *nnU-Net* converge to better segmentation maps with T2*GRE scans. Higher Dice scores, recall, and precision and lower Hausdorff distances are obtained with T2*GRE. In particular, higher discrepancies are noticed for *nnU-Net* recall that is halved when fed with SWI/SWAN and *Apollo* Hausdorff distance that rises from 28.16 to 65 *mm*. We hypothesize that the higher lesion performance

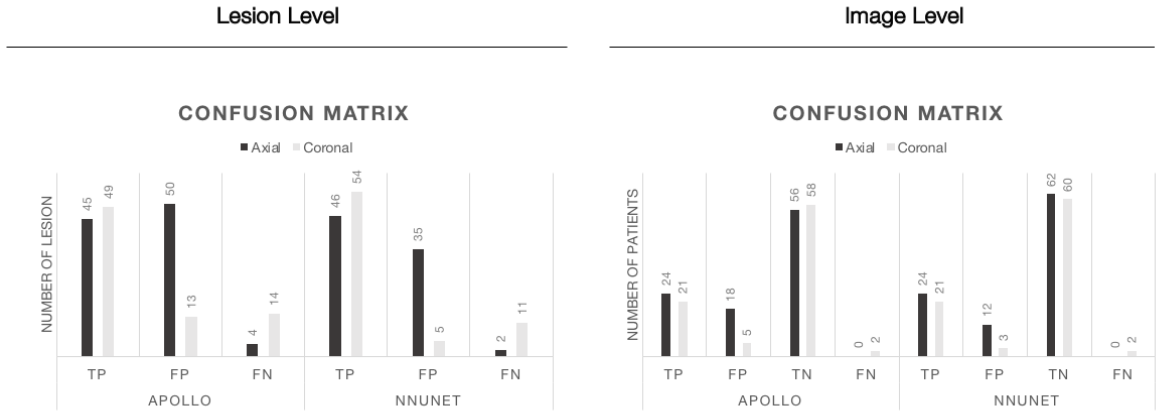

Figure 7: Confusion matrix for L1 prediction in percentage the in-house dataset for *Apollo* and *nnU-Net*. Comparisons between Axial FLAIR (dark) and Coronal FLAIR (light grey) are performed at (Left.) lesion level and(Right.) image level.

Table 4: (**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.

	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

with T2*GRE lies on the image characteristics that seem to produce more discriminative features, facilitating L3 predictions, and on the higher percentage of T2*GRE among hemorrhage patients on the training set.

nnU-Net outperforms Apollo in terms of generalization ability towards the input sequences: 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 L3 prediction or Axial FLAIR specificity for L1 prediction are below 80 %, *nnU-Net* shows decent scores, independently of the MRI parameters considered in the experiments.

Robustness on an external dataset

External validation is required to have reliable insights in models performance: From Table 4, *Apollo* and *nnU-Net* report a drop in performance between the in-house and the external dataset. *Apollo* makes noisy predictions and has troubles in distinguishing pathology *versus* background. This has a cumbersome impact on the specificity with a drop of 84 % between datasets. *nnU-Net* manages to better handle unseen data and reduces the FP predictions. The average results shown in the table do not reflect the fact that L2, by suffering a larger decrease in performance, overshadows L1 and L3 satisfactory scores. This is particularly relevant in the sensitivity: while L2 reports a score of 30%, L1 and L3 scores reach 91.12 % and 86.67%, respectively. Similar behaviours are noticed across all the metric board.

Unintended data bias hampers the generalization ability of DL models: From Figure 5 and Figure 6, 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. On one hand, the external dataset reveals evident shifts in class distribution when compared the training datasets. 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. This is the case of hyper acute infarcts and chronic intraparenchymal hemorrhages. Moreover, behaviours across datasets can also arise from the different compositions between in-house and external data pool. From the analysis conducted in the in-house dataset, among the tumors predicted by *nnU-Net*, FP lesions mostly arise from chronic infarcts and edemas. These labels are in higher proportions in the external dataset and can partially explain the struggle of *nnU-Net* with L2 predictions.

On the other hand, out-of-distribution data can emerge from differences in the scanners and acquisition parameters. While 80 % of the L3 patients in the external set present a SWI scan, they represent 1.23% of the training set. Transferring the discriminative features learnt on SWAN and T2* GRE to SWI images is far from being trivial and may explain the behaviours of DL solution when predicting on previously unseen data. Similar shifts are also noticed, in a lower measure, for L2 predictions on FLAIR acquisitions.

5. Conclusions

We introduce a complete evaluation framework for DL algorithms performing lesion segmentation and image classification on MRI images. Built upon a deep understanding of their clinical application, it conducts a joint analysis at an image (classification) and lesion (segmentation) level, supported by a broad panorama of metrics and steps. The strong performance of our heuristic approach arises by giving a comprehensive perception of the strengths and weaknesses of the models across pathologies and datasets, which was previously neglected in other works.

By extensively addressing the generalization ability, our analysis evidences that DL models are not agnostic to data and their performances are highly affected by unintended data bias present in the training 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 performing an external evaluation to assess robustness across clinical sites. Crucially, this observation may result in sub-optimal performances and misclassifications, hampering the seamless integration of DL solutions at hospitals. To address these limitations, recent guidelines published by *Radiology* [28] and by Challen *et al.* [5] help radiologists in gauging DL models from a quality and safety perspective.

Our comparison between *Apollo* and *nnU-Net* reveals that *nnU-Net* is more robust to unseen data. We hypothesize that the reason behind *nnU-Net*'s better performance on unseen data lies in its pipeline formulation and key design choices. This observation is in line with Isensee *et al.* [21] 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" made to automate its configuration to any task and dataset. From Table 1, we expected divergence in performance to arise from the more sophisticated optimization function, the deep supervision process, complex data augmentation techniques, and the rejection of early stopping. These key choices seem to enable *nnU-Net* to learn more discriminative features, resulting in a more homogeneous performance across pathologies and a more resilient behaviour across datasets.

While our framework has shown its potentials in evaluating DL algorithms , validating in a larger pool of patients would further improve the quality of our findings. Data scarcity currently prevents the selection of additional MRI acquisition parameters and jeopardizes the statistical validation of the results. Practical limitations encompass finding hospitals willing to share clinical data and generating ground truth segmentation maps through manual annotation.

References

[1] Eurostat, "Healthcare resource statistics - technical resources and medical technology," *Statistics Explained*, pp. 1–18, 2020.

[2] A. Hosny et al., "Artificial intelligence in radiology," *Nat. Rev. Cancer.*, vol. 18, no. 8, pp. 500–510, 2018.

[3] Y. Kabir et al. , "Multimodal MRI segmentation of ischemic stroke lesions," in *29th Conf Proc IEEE Eng Med Biol Soc*. IEEEXplore, 2007, pp. 1595–1598.

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

[5] R. Challen et al., "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

[6] D. L. Rubin and M. P. Lungren, "Preparing Medical Imaging Data for Machine Learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[7] B. Kocak et al., "Radiomics with artificial intelligence: a practical guide for beginners," *Diagnostic and interventional radiology*, vol. 25, no. 6, pp. 485–495, 2019.

[8] D. Harpaz et al., "Point-of-care-testing in acute stroke management : An unmet need ripe for technological harvest," *Biosensors*, vol. 7, no. 30, pp. 1–39, 2017.

[9] 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/

[10] R. Cattell et al., “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.

[11] N. Karani et al. , “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

[12] R. Gauriau et al., “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.

[13] E. Beede et al., “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.

[14] B. Allen Jr et al., “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.

[15] M-S. Badea et al., “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

[16] O. Ronneberger et al., “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *LNCS*, vol. 9351, pp. 234–241, 2015.

[17] J. Cho et al., “Improving sensitivity on identification and delineation of intracranial hemorrhage lesion using cascaded deep learning models,” *J. Digit. Imaging*, vol. 32, pp. 450–461, 2019.

[18] N. Heller et al., “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

[19] I. R. I. Haque and J. Neubert, “Deep learning approaches to biomedical image segmentation,” *IMU*, vol. 18, 2020.

[20] S.A. Taghanaki et al., “Deep semantic segmentation of natural and medical images : a review,” *Artif. Intell. Rev.*, vol. 54, pp. 137–178, 2020.

[21] F. Isensee et al., “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, pp. 203–211, 2021.

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

[23] W. Bai et al., “Automated cardiovascular magnetic resonance image analysis with fully convolutional networks,” *JCMR*, vol. 20, no. 65, 2018.

[24] 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.

[25] 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

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

[27] J.S. Whang et al., “Diffusion-weighted signal patterns of intracranial haemorrhage,” *Clin Radiol*, vol. 70, no. 8, pp. 909–916, 2015.

[28] D.A. Bluemke et al., “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.