



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

---

Year : 2023

## Weakly Supervised Deep Learning Models for Anomaly and Change Detection in Radiology

Di Noto Tommaso

Di Noto Tommaso, 2023, Weakly Supervised Deep Learning Models for Anomaly and Change Detection in Radiology

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB\_85AA0FB1009B5

### **Droits d'auteur**

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

### **Copyright**

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



**UNIL** | Université de Lausanne

Faculté de biologie  
et de médecine

**Département de Radiologie,**

**Centre Hospitalier Universitaire Vaudois (CHUV)**

# **Weakly Supervised Deep Learning Models for Anomaly and Change Detection in Radiology**

**Thèse de doctorat ès sciences de la vie (PhD)**

présentée à la

Faculté de biologie et de médecine  
de l'Université de Lausanne

par

**Tommaso Di Noto**

Master de l'Università Campus Bio-Medico di Roma

## **Jury**

Prof. Xavier Crevoisier, Président  
Prof. Patric Hagmann, Directeur de thèse  
Dr. Jonas Richiardi, Co-directeur de thèse  
Dre. Meritxell Bach Cuadra, Co-directrice de thèse  
Prof. Henning Müller, Expert  
Prof. Clarisse Dromain, Experte  
Prof. Georg Langs, Expert

Lausanne  
(2023)



# Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

<b>Président·e</b>	Monsieur	Prof.	Xavier	<b>Crevoisier</b>
<b>Directeur·trice de thèse</b>	Monsieur	Prof.	Patric	<b>Hagmann</b>
<b>Co-directeurs·trices</b>	Monsieur	Dr	Jonas	<b>Richiardi</b>
	Madame	Dre.	Meritxell	<b>Bach Cuadra</b>
<b>Expert·e·s</b>	Monsieur	Prof.	Henning	<b>Müller</b>
	Madame	Prof.	Clarisse	<b>Dromain</b>
	Monsieur	Prof.	Georg	<b>Langs</b>

le Conseil de Faculté autorise l'impression de la thèse de

## **Tommaso Di Noto**

Master - Laurea Magistrale in Ingegneria biomedica, Università Campus Bio-Medico, Italie

intitulée

## **Weakly supervised deep learning models for anomaly and change detection in radiology**

Lausanne, le 3 février 2023

pour le Doyen  
de la Faculté de biologie et de médecine



Prof. Xavier Crevoisier



*“Nothing in life is to be feared,  
it is only to be understood.  
Now is the time to understand more,  
so that we may fear less.”*

*- Marie Curie*



# Contents

<b>Abstract - English</b>	<b>X</b>
<b>Résumé - Français</b>	<b>XI</b>
<b>Acronyms</b>	<b>XII</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Anomaly and change detection in radiology . . . . .	1
1.1.1 Computer-Aided Detection/Diagnosis . . . . .	1
1.1.2 Goal of the PhD thesis . . . . .	4
1.2 The advent of Deep Learning, and its limitations . . . . .	4
1.2.1 Classical Machine Learning . . . . .	5
1.2.2 Deep Learning (DL) . . . . .	7
1.2.3 Weak labels to mitigate the annotation bottleneck . . . . .	9
1.2.4 Transfer Learning to alleviate data scarcity and domain shift . . . . .	11
1.2.5 Prior anatomical knowledge to improve interpretability . . . . .	13
1.3 Automated Detection of Cerebral Aneurysms . . . . .	14
1.3.1 Clinical background . . . . .	14
1.3.2 Aneurysm visual detection . . . . .	15
1.3.3 Automated detection . . . . .	16
1.3.4 Proposed approach . . . . .	16
1.3.5 Clinical evaluation . . . . .	22
1.4 Change Detection in Longitudinal Glioma Imaging . . . . .	23
1.4.1 Clinical background . . . . .	23
1.4.2 MRI-based longitudinal monitoring . . . . .	24



1.4.3	Extracting weak labels from radiology reports with Natural Language Processing	25
1.4.4	Glioma change detection using weak labels and transfer learning . . . . .	27
<b>2</b>	<b>Summary of Results</b>	<b>32</b>
2.1	Automated Aneurysm Detection . . . . .	32
2.2	Glioma Change Detection . . . . .	32
2.2.1	Report classification with Natural Language Processing . . . . .	32
2.2.2	Image change detection with transfer learning and weak labels . . . . .	33
<b>3</b>	<b>Discussion</b>	<b>35</b>
3.1	Main contributions . . . . .	35
3.2	Limitations and future steps . . . . .	39
3.3	Conclusion . . . . .	42
	<b>References</b>	<b>44</b>
	<b>Papers</b>	<b>67</b>

# Acknowledgments

First, I would like to thank my supervisor, Prof. Patric Hagmann, for providing me this thrilling opportunity. Your support, both on a scientific and personal level, has been fantastic, and your clinical advice throughout the PhD has been invaluable to steer the research towards more concrete and applicable outcomes.

Second, I would like to thank my co-supervisors, Dr. Jonas Richiardi and Dr. Meritxell Bach Cuadra. Your help and guidance has been incredible! When I was stuck, you were there to make me reason and to propose amazing ideas. When I was sad cause nothing worked, you were there to teach me it's normal to fail. And when things were going well, you were there to celebrate with me. Our routine discussions, both professional and personal, have been insightful and I deeply thank you for that. I feel privileged to have been supervised by both of you.

I would also like to thank all the members of the Connectomics, MIAL, and TML labs. You guys are great, and I feel grateful to have had the chance to be in all three groups.

A special thank goes to Francesco, Yasser, Emeline, Gian Franco, Adele, Cipo, Hamza, H el ene and Jaume. Looking back at all the moments we shared together makes me realize how fortunate I am to have crossed your lives. You made these years memorable and I'll always be thankful for that.

To my Italian friends Mastro, Bruzzo and Nico. You guys are awesome and every time I go back home you make me feel like nothing ever changed. I know I can always count on you.

To Marie, for always being there for me, in the good and the bad moments. I look forward to sharing many more adventures together and overcoming all the obstacles that life will throw at us. I'm lucky to have you, and couldn't ask for a better person by my side.

Last, I want to thank my family. Despite the distance, you've always been my anchor. To Giachi, for being the captain of our unforgettable cruises, but most importantly the point of reference I can always rely on. To my parents, for giving me the possibility to follow my dreams abroad, and the strength to face all the difficulties along the way. None of this would have been possible without your unconditional love and wholehearted support. I love you.

---

## Abstract - English

**PhD Candidate (department):** Tommaso Di Noto (Department of Radiology, CHUV)

The goal of this PhD thesis was to address some recurrent limitations that are associated with Deep Learning (DL)-based Computer-Aided Detection/Diagnosis (CAD) systems in medical imaging. We focused our analysis on two clinical tasks routinely performed in radiology: the detection of intracranial aneurysms on Magnetic Resonance Angiography (MRA) scans and the longitudinal monitoring of high-grade gliomas in T2-weighted MR scans. The first limitation that we tackled was the *lack of annotated data*; to mitigate this, in both tasks we made use of **weak labels** to drive the learning process. These are coarse labels that are typically imprecise, but fast/cheap to obtain. For the aneurysm detection task, our weak labels corresponded to oversized annotations which resulted to be 4 times faster to create with respect to their voxel-wise counterparts. In the glioma change detection task, the weak labels were automatically extracted from textual radiology reports with a Natural Language Processing (NLP) framework, and allowed us to increase the amount of labeled data more than 3 fold. A further contribution of this thesis to reduce data scarcity in medical DL applications is the **open release of our two in-house datasets**. To date, our cohort for aneurysm detection is the largest in the community (N=284, 127 controls, 157 patients with aneurysms), while our cohort for glioma change detection (N=183 patients, 1693 difference maps) is the first longitudinal dataset ever made open access. The second limitation that we addressed was *domain shift*, which corresponds to a change in data distribution between a model’s training data, and the unseen test data it will be fed with at inference time. To alleviate domain shift, we investigated **Transfer Learning (TL)**, and in particular we automated the choice of TL sub-type treating it as a hyperparameter to optimize which avoids empirical and sub-optimal choices that are frequent in similar works. The last limitation of DL-based tools that we tried to mitigate was the *lack of interpretability*, with DL models often being addressed as incomprehensible “black-box” models: in the aneurysm project, we added **prior anatomical knowledge** constraining the analysis only to areas of the brain that are plausible for aneurysm occurrence. This aimed at simulating the radiologists’ reading of images and avoiding unrealistic model outputs. Although the use of prior anatomical knowledge is not the most frequently used technique to improve interpretability, it has been shown to help increasing model transparency in related works. Overall, we believe that the combination of our open-source contributions and open-access datasets have the potential to make DL-based CAD tools more reproducible, and bring them closer to clinical application.

---

## Résumé - Français

**Doctorant (Service):** Tommaso Di Noto (Service de radiodiagnostic et radiologie interventionnelle, CHUV)

L'objectif de cette thèse est d'aborder certaines limitations récurrentes associées aux systèmes de *diagnostic assisté par ordinateur* (CAD) basés sur le *Deep Learning* (DL). Nous avons concentré notre analyse sur deux tâches couramment réalisées en radiologie : la détection d'anévrismes cérébraux sur des examens d'angiographie par résonance magnétique et le suivi de gliomes de haut grade sur des images par résonance magnétique pondérées T2. La première limitation que nous avons abordée est le *manque de données annotées*. Pour atténuer ce problème, nous utilisons dans les deux tâches des ***weak labels*** : il s'agit d'annotations généralement imprécises, mais rapides et peu coûteuses à obtenir. Pour le projet des anévrismes, nos *weak labels* correspondent à des annotations grossières qui ont été créées quatre fois plus rapidement que leurs équivalents à l'échelle d'un voxel. Dans le projet des gliomes, les *weak labels* ont été extraites automatiquement des rapports radiologiques avec un modèle de *Natural Language Processing*, et nous ont permis de tripler la quantité de données annotées. Une autre contribution majeure de cette thèse pour pallier la pénurie de données repose sur la publication de nos *datasets*. À ce jour, notre *dataset* pour la détection des anévrismes est le plus important de la communauté en terme de nombre de sujets (N=284), tandis que notre *dataset* pour la détection des changements dans les gliomes (N=183) est le premier *dataset* longitudinal jamais publié. La deuxième limitation que nous avons abordée est le *domain shift*, un scénario fréquent en imagerie médicale qui se traduit par un changement dans la distribution des données entre les données d'apprentissage d'un modèle et les données de test non vues qui lui seront fournies au moment de l'inférence. Pour atténuer le *domain shift*, nous avons étudié le ***Transfer Learning*** (TL), et en particulier nous avons automatisé le choix du type de TL en le traitant comme un hyperparamètre à optimiser, évitant ainsi des choix empiriques, fréquents dans des études similaires. La dernière limitation que nous avons abordée est le *manque d'interprétabilité des modèles* : dans le projet des anévrismes, nous avons inclus des **connaissances anatomiques a priori** en restreignant l'analyse aux zones du cerveau susceptibles d'être impactées par la survenue d'un anévrisme. Cette démarche vise à simuler la lecture des images par les radiologues et à éviter des résultats irréalistes. Nous sommes convaincus que nos contributions à la science ouverte ont le potentiel de rendre les systèmes de CAD basés sur le DL plus reproductibles et de les rapprocher de l'application clinique.

# Acronyms

- ADC** Apparent Diffusion Coefficient. 23
- AI** Artificial Intelligence. 4, 13, 42
- BraTS** Brain Tumor Segmentation. 24, 28
- CAD** Computer-Aided Detection/Diagnosis. 2, 3, 5, 42, 43, X
- CADe** Computer-Aided Detection. 1, 2, 4, 16, 22, 25, 39–41
- CADx** Computer-Aided Diagnosis. 1, 2
- CBCT** Cone-Beam Computed Tomography. 10
- CNN** Convolutional Neural Network. 8, 30, 67
- CT** Computed Tomography. 2, 13
- CTA** Computed Tomography Angiography. 14
- CV** Computer Vision. 7, 10, 37
- DL** Deep Learning. 2–5, 7–10, 12, 13, 16, 24, 25, 35–37, 39, 42, 43, X
- DNN** Deep Neural Network. 7, 8, 12, 13
- DSA** Digital Subtraction Angiography. 14
- DSC** Dynamic Susceptibility Contrast. 23
- DTI** Diffusion Tensor Imaging. 23

## Acronyms

---

- DWI** Diffusion-Weighted Imaging. 23
- ESNR** European Society of NeuroRadiology. 23
- FLAIR** FLuid Attenuated Inversion Recovery. 23
- ML** Machine Learning. 2–6, 8–11, 13, 24, 43, 67
- MRI** Magnetic Resonance Imaging. 2, 23
- NLP** Natural Language Processing. 10, 25, 28, 36, X
- NN** Neural Network. 7
- ROI** Region Of Interest. 2
- SWI** Susceptibility Weighted Imaging. 23
- T1w** T1-weighted. 23
- T2w** T2-weighted. 23
- TL** Transfer Learning. 3, 4, 11, 12, 22, 25, 27, 28, 30, 31, 37, 40, 43
- TOF-MRA** Time-Of-Flight Magnetic Resonance Angiography. 4, 14
- UIA** Unruptured Intracranial Aneurysm. 14–17
- WHO** World Health Organization. 23, 42

# Chapter 1

## Introduction

### 1.1 Anomaly and change detection in radiology

Pathologies in medical imaging correspond to rare deviations from a distribution of normal, healthy samples [1]; in other words, they can be considered as outliers (or anomalies) in the standard population data. One primary task performed by radiologists in cross-sectional imaging is the visual detection of these anomalies with the goal of tailoring diagnosis and follow-up treatment. In addition to anomaly detection, a recurrent task carried out in radiology departments is change detection: this corresponds to a longitudinal monitoring of a pathology over time. Regardless of the physicians' expertise, both anomaly and change detection can still be limited by several factors, such as image noise, fatigue, distraction, or work overload [2]. To overcome these limitations, in the early 1980s the medical imaging community began to invest resources in a new line of research dedicated to Computer-Aided Detection (CADe), and Computer-Aided Diagnosis (CADx) systems [3, 4], whose dissemination was facilitated by the rapid advent of digital technologies, the creation of ever larger medical databases, and the concurrent steady improvements in the field of Machine Learning (ML).

#### 1.1.1 Computer-Aided Detection/Diagnosis

CADe and CADx can be defined as complementary, “second opinion” tools for detection/diagnosis that radiologists can use when reading medical images [5]. The ultimate goal of both is to improve the productivity of radiologists by increasing the accuracy and consistency of diagnoses, while ideally reducing image reading time [6]. The difference between the two lies in the aid provided to clinicians: a CADe highlights potential abnormalities on diagnostic exams, whereas a CADx analyzes an image

## 1.1 Anomaly and change detection in radiology

---

finding and helps to discriminate between different disease processes (e.g. benign versus malignant lesion) [3]. CAD is often used as an umbrella term to include both CADe and CADx. From their onset in the '80s, CAD tools found widespread adoption in several clinical tasks, especially in screening settings [7–10].

The first wave of CAD systems was mostly based on standard image processing expedients such as image filtering, enhancement methods, edge detectors, adaptive histogram equalization, or image subtraction after registration [11]. Then starting from the early 2000s, a second wave of CAD systems based on pattern recognition and Machine Learning (ML) started to spread. These systems used “classical” ML architectures which will be briefly described later in section 1.2.1. For the sake of this paragraph, it suffices to understand the general pipeline behind these tools: after some pre-processing and the definition of a Region Of Interest (ROI), the core blocks of a CAD based on classical ML are feature extraction and classification. During the former, we extract relevant image features that describe the ROI, whereas in the latter we feed these extracted features to a statistical computer algorithm that finds patterns in the features and learns how to discern between relevant classes (e.g. normal vs. abnormal brain MRI [12], brain aneurysm vs. infundibulum [13]). The expansion of CAD systems during the first decade of the third millennium has been impressive: first of all, CAD tools were developed for virtually every imaging modality; second, they also expanded to clinical tasks which were not necessarily part of a screening program [14]. To have an overview of all these applications, interested readers are encouraged to delve into seminal review papers such as [15–17] for applications in breast mammography and Magnetic Resonance Imaging (MRI), [18–20] for chest radiography and Computed Tomography (CT), [21, 22] for colonography, [23] for brain tumors, [24] for dementia, or [25, 26] for aneurysm detection, to mention just a few.

A third wave of CAD tools that goes under the name of “radiomics” started around 2010 [27–30]: many traits of radiomics overlap with classical CAD systems since we still extract large amounts of quantitative imaging features from medical images with the intent of uncovering relevant biological biomarkers. The major differences are the standardization of the feature extraction process (with initiatives such as [31, 32]), the high throughput and diversity of features (e.g. shape-based, histogram-based, texture-based), and the increased focus on combination of imaging data with other patient information (demographic, histologic, genomic, or proteomic data).

The most recent wave of CAD systems began around 2015 and is centered around Deep Learning (DL). In section 1.2, we will explore the differences between shallow and deep learning architectures. For now, we can imagine DL models as specific types of artificial neural networks (a type of ML architecture) that bypass the feature extraction block of traditional CADs and rather learn discriminative



## 1.1 Anomaly and change detection in radiology

---

features directly from raw data [33]. Out of all the revolutions in the field of CAD systems, the deep learning one was the most disruptive, with CAD tools reaching performances on par, or superior to trained radiologists for some specific and limited pathologies and modalities [34]. Nevertheless, the DL revolution also posed several hurdles that need to be overcome before DL-based CAD tools can be safely adopted during routine clinical practice [33].

The first limitation is the lack of large (thousands/millions of samples), multicentric and heterogeneous datasets. This hinders predictive performance, because DL algorithms have high capacity and benefit from more data than classical ML algorithms. Also, it hampers generalization ability, because there is no guarantee that an algorithm trained on local data would perform well on data from other centres. Collecting such datasets is a costly process, especially if together with images we also want to **collect manual labels/annotations generated by radiologists**. One possible solution to mitigate this annotation bottleneck is the use of “**weak**” labels [35–40] which is at the core of this PhD thesis, and will be discussed in sections 1.2.3, 1.3.4 and 1.4.3.

A second limitation of DL-based CAD tools is the lack of generalization when applied to unseen real-world data that exhibit a feature distribution that differs from the training distribution, a phenomenon known as **domain shift** [41]. One solution that is commonly adopted to alleviate domain shift is **Transfer Learning (TL)**: the main idea behind TL is to leverage knowledge acquired from a specific task or domain (source) to solve the real-world related task (target). The use of TL to alleviate domain shift and data scarcity is also a recurrent theme of this PhD thesis, and will be discussed in sections 1.2.4, 1.3.5 and 1.4.4.

A third limitation of DL-based tools is the lack of model **interpretability** [42, 43], with deep neural networks (explained in section 1.2.2) often being addressed as incomprehensible “black boxes”. This issue is especially delicate in medical imaging, where decisions made by DL-based CAD tools should be understandable/explainable to clinicians, patients, and ML practitioners. Among the numerous techniques being proposed in the literature to address model interpretability [44, 45], one is the use of **prior anatomical knowledge** [33, 46] which can help to understand whether we have extracted relevant features, can narrow the analysis to anatomically-plausible areas, or can help interpreting the model output. The use of prior anatomical knowledge is also a contribution of this thesis and will be later discussed in sections 1.2.5, and 1.3.4.

## 1.2 The advent of Deep Learning, and its limitations

---

### 1.1.2 Goal of the PhD thesis

In this PhD thesis, we address two clinical tasks that are routinely performed in radiology departments. The first task is the detection of unruptured intracranial aneurysms in Time-Of-Flight Magnetic Resonance Angiography (TOF-MRA) scans, while the second task is the longitudinal change detection for patients with high-grade gliomas. For each task, we develop a specific DL-based CAde tool and, as we will see in the following paragraphs, we address some of the DL limitations that were presented above. The ultimate goal for both projects is then the development of a CAde system: for the aneurysm project, the CAde system would help radiologists to increase their sensitivity and thus avoid risky oversights, while for the glioma project the CAde tool would rather be used to highlight the parts of the scans that have changed over time in order to facilitate and speed up the diagnosis, while ideally providing quantitative indicators to radiologists. For both tasks we address the lack of large annotated dataset by devising time-saving weak labels. In addition, though in slightly different ways, for both projects we leverage TL to increase performances on the ultimate target task. Last, for the aneurysm detection project we make use of prior anatomical knowledge to constrain the analysis only to anatomically-plausible locations, an expedient that increases detection performances and reduces model opaqueness. In Figure 1.1, we report the inherent limitations of medical DL that we aimed to mitigate in this PhD thesis, together with our proposed solutions.

The rest of the thesis is organized as follows: in section 1.2, we describe the advent of DL, starting from classical (shallow) ML models and ending with an overview of its current limitations. In section 1.3, we introduce the first clinical task of brain aneurysm detection. In section 1.4, we present the second clinical task of change detection for longitudinal glioma imaging. Following the format of “thesis with articles” recommended by the University of Lausanne, we then present in Chapter 2 a summary of the results for the 3 most important manuscripts written throughout the PhD (2 accepted [35,47] and 1 submitted [36]). In Chapter 3, we discuss and contextualize the results, then we provide an outlook on future steps, and finally we draw the conclusions of the thesis. Last, the 3 manuscripts are attached.

## 1.2 The advent of Deep Learning, and its limitations

Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) that gives computers the ability to learn through experience and without being explicitly programmed to do so [48]. ML can be broadly divided into two main sub-fields: classical ML and DL [49]. Although the two clinical tasks

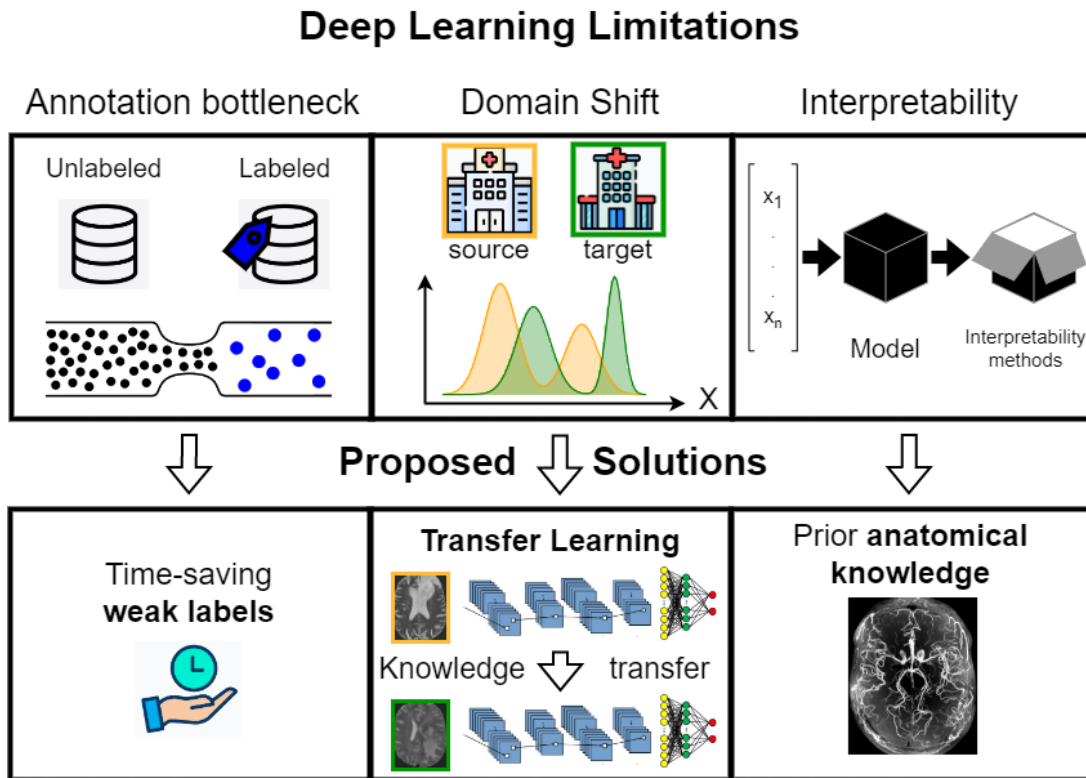


Figure 1.1: **Top row:** recurrent limitations of DL-based CAD tools in medical imaging: the annotation bottleneck corresponds to the difficulty of retrieving large amount of labeled data; domain shift is a change in the feature distribution between an algorithm’s training dataset, and the dataset it encounters when deployed; interpretability is the lack of understanding associated to DL models. **Bottom row:** proposed solutions to mitigate such limitations.

addressed in the thesis are based on DL algorithms, it is worth understanding how the field moved from classical (shallow) ML towards deeper architectures, and what the key differences are.

### 1.2.1 Classical Machine Learning

Classical ML refers to computer algorithms (or models<sup>1</sup>) that are fed with a variety of input data (e.g. tabular, image, text, audio) and learn how to become better at evaluating and acting on that data over time. A more formal definition was given by Tom Mitchell in 1997 and says: “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.” [48].

<sup>1</sup>The terms *algorithm* and *model* will be used interchangeably throughout the thesis.

## 1.2 The advent of Deep Learning, and its limitations

---

A common notation used to summarize this concept defines ML algorithms as models that learn a target function ( $f$ ) that best maps input variables ( $X$ ) to an output variable ( $Y$ ), following:

$$Y = f(X) \tag{1.1}$$

where  $f$  is unknown and needs to be learned from data. In classical ML,  $X$  comprises a set of discriminative **features** that describe the samples in the dataset. If considering a supervised scenario (more details in section 1.2.3), together with features,  $X$  would also include corresponding **labels** which indicate the class each sample belongs to. The combination of features (denoted  $\mathbf{x}$ ) and labels (denoted  $\mathbf{y}$ ) forms what is typically called the training dataset  $\mathcal{D}_{train} = \{X, Y\} = \{\mathbf{x}_i, \mathbf{y}_i | i = 1 : N_{train}\}$ , where  $N_{train}$  corresponds to the number of samples. Once the dataset is collected, the next step in a classical ML pipeline is the choice of one model (often called *classifier*) that will learn a consistent relationship (function  $f$ ) between the features and the labels of  $\mathcal{D}_{train}$ . Many different classifiers have been developed in the past decades [50], and describing their differences is out of the scope of this thesis. What is important to know is that each classifier has a unique learning process (called *training*) during which its internal parameters are updated in the best possible way in order to distinguish the classes in  $\mathcal{D}_{train}$ . Supposing we choose a multi-layer perceptron classifier [51, 52], at the end of the training process our model would have updated its parameters in the best possible configuration to distinguish the classes of the samples in  $\mathcal{D}_{train}$ . The last step of the classical ML pipeline is *inference*: during this phase, we use the trained classifier to generate predictions for new, distinct samples. These unseen samples make up what is called the test dataset  $\mathcal{D}_{test} = \{\mathbf{x}_i | i = 1 : N_{test}\}$ , where  $N_{test}$  is the number of test samples for which we want to generate predictions  $\hat{\mathbf{y}}_i$ . If training was successful, the classifier should be able to assign, with a certain degree of confidence, the new  $N_{test}$  samples to their corresponding classes.

The process of computing/choosing which features to use in the classical ML pipeline is called *feature engineering*: although this approach is straightforward and explainable, it has the main drawback of requiring domain expertise which, depending on the field of application, can be hard to obtain. In addition, even when domain expertise is available, the generation (engineering) of hand-crafted features is not always optimal, since potentially informative descriptors might be neglected. These limitations of manual feature engineering, combined with the explosion of available data and advances in computing power of the last decade (low-cost graphical processing units, and increasing storage capacity), progressively led to the widespread adoption of deep learning [52].

## 1.2 The advent of Deep Learning, and its limitations

---

### 1.2.2 Deep Learning (DL)

The multi-layer perceptron classifier that we saw in the previous paragraph belongs to a broad family of models called Neural Networks (NNs). Their commonality is a layered structure of interconnected nodes that is inspired by the neuronal connections in the human brain. NNs learn from training data how to recognize useful patterns, and then make predictions for future events. During training, the connections between nodes (commonly denoted as *weights*) are adjusted according to a specified learning rule in order to improve predictions. As shown in Figure 1.2, a NN is made of an input layer, one or more hidden layers, and an output layer. Each layer is composed of several nodes and the nodes in each layer use the outputs of all nodes in the previous layer as inputs. By doing so, all neurons interconnect with each other through the different layers. The models that we have

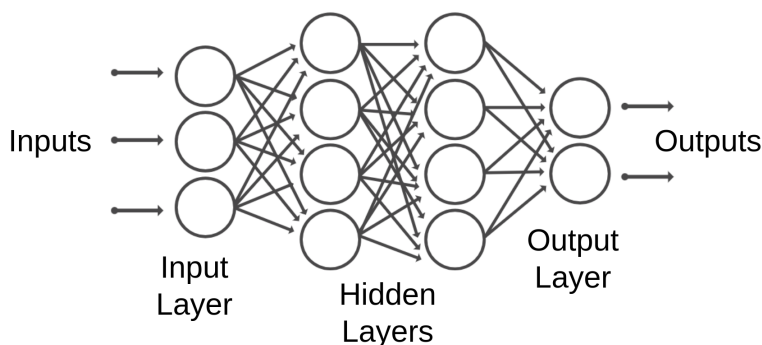


Figure 1.2: Typical architecture of a feed-forward neural network.

discussed so far are usually denoted as *shallow* NNs, and understanding their structure is important to introduce the concept of DL because most DL methods are based on NN architectures. As a matter of fact, DL models are often referred to as Deep Neural Networks (DNNs), where the adjective *Deep* simply indicates that we are dealing with a NN that has more than 3 hidden layers. In general, the higher the number of layers (and thus of parameters), the deeper the NN will be. Since the first perceptron model introduced in 1958 by Frank Rosenblatt [53], increasingly complex NNs have been developed by researchers, a path that led to today's deepest networks that can contain up to 530 billion parameters (e.g. Megatron-Turing model [54]). Overall, NNs are especially suitable for modeling non-linear relationships. In the field of DL, deep neural networks have been extensively adopted for numerous tasks in the fields of Computer Vision (CV) (e.g. image recognition, object detection, etc.), natural language processing (e.g. machine translation, speech recognition, etc.), optimized search engines, and content recommendation, to name just a few. A more detailed list of

## 1.2 The advent of Deep Learning, and its limitations

---

ML applications was provided by [55]. The most important novelty brought by DL with respect to the classical ML approach is that the feature engineering step disappears from the pipeline: instead of creating hand-crafted features which are later passed to a ML model, DNNs can be fed with minimally preprocessed data (e.g. directly with input images) and then automatically learn features from this raw data through multiple nonlinear layers of representation. Because of this concept, DL can also be referred to as *representation learning*. Figure 1.3 illustrates this key difference for a toy image classification problem. Explaining the different sub-types of DNNs is out of the scope of this

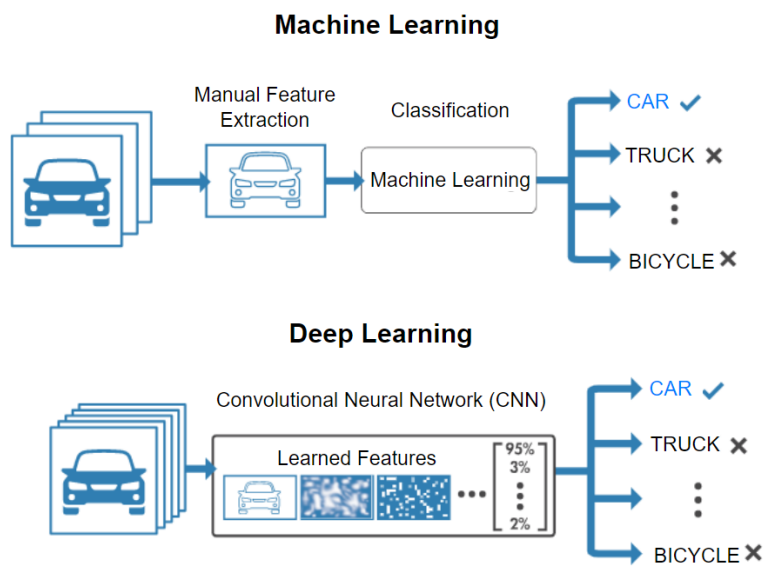


Figure 1.3: Difference between classical ML and DL for image classification. The manual feature extraction (feature engineering) disappears from the DL pipeline where features are extracted directly from the input images.

thesis, though a detailed overview can be found in [56]. Since this work focuses on DL models in the field of radiology, it is sufficient to know that the most common architectures applied in this domain are Convolutional Neural Networks (CNNs), mainly because of their excellent capacity to deal with images as input samples [57]. In the next paragraph, we shed light on the main limitations of DL algorithms, with a special focus on medical imaging.

## 1.2 The advent of Deep Learning, and its limitations

---

### 1.2.3 Weak labels to mitigate the annotation bottleneck

Machine learning can be divided in 3 main categories<sup>2</sup>: unsupervised, semi-supervised and supervised, depending on the amount of labels that are available; supervised learning is the most common, thus only a brief introduction will be provided for the unsupervised and semi-supervised learning paradigms, while more attention will be given for supervised learning where the annotation bottleneck is more prominent.

**Unsupervised Learning** - We refer to unsupervised learning when there are no labels available in the dataset. The goal of the ML algorithm in this scenario is to discover hidden patterns and/or create clusters (i.e. subgroups) in the dataset without the need for human intervention. Unsupervised learning models can be utilized for three main tasks: clustering, association, and dimensionality reduction. Clustering groups unlabeled data based on their similarities or differences. Association rules are techniques for discovering relationships across the samples of the dataset. Last, dimensionality reduction is used when the number of features in the dataset is too high and must be reduced.

**Semi-supervised Learning** - We refer to semi-supervised learning when a problem involves a small number of labeled examples and a (typically larger) number of unlabeled examples [59]. Semi-supervised learning is well-suited for applications where data is relatively easy to obtain, but the subsequent labeling process is challenging or expensive, as in medical imaging [60]. The goal in semi-supervised learning is to use the set of unlabeled data to improve the predictions for the task at hand. Two main goals can be targeted in self-supervised learning: either we try to predict the labels for future data or we try to predict labels for the already available unlabeled data. For a detailed overview of semi-supervised applications in medical imaging, interested readers can refer to [60].

**Supervised Learning** - This corresponds to applications where all available data is labeled. Supervised learning can be divided into two subcategories: classification and regression. The main difference between the two is that in classification we predict/classify discrete values (e.g. email is spam or not spam), while in regression we predict a continuous value (e.g. price of a house based on the city where it is built). As we have seen in sections 1.2.1 and 1.2.2, during training the ML algorithm (be it a classifier or a regressor) searches for patterns in the data that correlate with the desired output labels. Then, during inference, the model is fed with new unseen samples and generates predictions according to the patterns that were learned.

During the last two decades, countless supervised algorithms have been developed for medical imaging tasks, both based on classical ML and on DL [61]. However, it soon became clear that

---

<sup>2</sup>we neglect reinforcement learning here, currently still under-explored in medical imaging [58]

## 1.2 The advent of Deep Learning, and its limitations

---

achieving high discriminative power for complex medical patterns required sufficiently large training cohorts [33], especially for DL-based models. As mentioned in section 1.1.1, the collection of labeled data is extremely costly in medical imaging: unlike CV tasks, where image annotations are relatively simple (and can even be performed via crowdsourcing [62]), the creation of labels in medical imaging requires technical expertise and is extremely time-consuming for radiologists. In this PhD thesis, we explore the use of “*weak*” labels, which represents one potential solution to mitigate the manual annotation bottleneck. Weak labels can be defined as coarse, incomplete, limited, undersized, or oversized annotations that are less precise, but considerably faster to create with respect to standard labels. The ultimate goal of weak labels is to simplify and speed up the annotation process for clinicians who often face work overload and should rather spend their time on more clinically relevant tasks. In line with this trend, numerous works have been published in the medical imaging community. For instance, the authors in [37] developed a DL pipeline to predict the concentration of different stains in multiplex immunohistochemistry images. Instead of labeling all the pixels for each cell, pathologists were only asked to mark a dot at the center of each cell. From each dot, the authors then used SLIC [63] to generate superpixels (meaningful regions). Similarly, the authors in [38] asked their specialists to annotate teeth in Cone-Beam Computed Tomography (CBCT) 3D images only drawing rough bounding boxes on certain axial slices. These bounding boxes are much faster to draw with respect to a voxel-wise labeling of the teeth done in every slice. In another study [39], the experts were asked to only draw a rough region inside each object of interest on Scanning Electron Microscopy (SEM) images. In line with these works, one of the papers of this PhD thesis [35] investigates the use of oversized weak labels for automated aneurysm detection and will be discussed in section 1.3.4.

An alternative approach for the creation of weak labels that has gained increasing interest in the community is the use of Natural Language Processing (NLP) on textual medical reports. NLP is the branch of ML that helps computers understand, interpret, and manipulate human language [64]. In medical imaging, NLP has the goal of extracting clinically relevant information from radiology reports. Medical reports are a valuable source of information since they are always stored together with corresponding images, they contain high-level insights from physicians, and they are less demanding than images from a computational point of view. In fact, reports require less disk space to be stored and training time of report-based ML models is much shorter compared to image-based models. Despite all these advantages, medical reports also come with intrinsic drawbacks; in particular, most reports are stored as unstructured, free-text documents and exhibit a strong degree of ambiguity, uncertainty and lack of conciseness [65]. Nevertheless, advances in NLP have enabled



## 1.2 The advent of Deep Learning, and its limitations

---

the extraction of relevant labels from radiology [66–72] and histopathology [40, 73] reports. One of the manuscripts presented in this thesis [47] follows this line of research and aims to extract weak labels from radiology reports of patients with high-grade gliomas. More details about this work are presented in section 1.4.3.

### 1.2.4 Transfer Learning to alleviate data scarcity and domain shift

Humans have an inherent capacity to transfer knowledge acquired from one task to a related, similar task [74]. Typically, the “closer” the tasks, the easier it is for us to transfer our knowledge and skills. For instance, learning how to drive a motorbike, having previously ridden a bicycle, can benefit from transferring some of the bicycle skills. Contrary to human learning, ML models were historically developed to solve specific, isolated tasks. However, this one-model-one-task paradigm soon revealed its obvious limitation: whenever the feature space of the new input data was changing (phenomenon known as “domain shift” [75]), performances at inference time decreased considerably [76], and the model had to be re-trained from scratch including new observations that resembled the target feature distribution. In medical imaging, domain shift can be induced by several factors such as different image acquisition protocols, different acquisition devices, different image resolution or even more fundamental differences like variations in populations’ features (e.g. source cohort younger than target cohort) [77]. Transfer Learning (TL) is one research field that tries to mitigate domain shift and overcome the isolated learning paradigm by utilizing knowledge acquired from a specific task or domain (source) to solve a downstream, related task (target) [76]. In addition to domain shift, TL techniques can also help in tackling the problem of data scarcity, especially for scenarios in which collecting additional training data is too costly, time-consuming, or even unrealistic [78], such as in medical imaging, where datasets used by most research groups are typically small (hundreds of samples, more rarely thousands) [79]. In any TL scenario, there are three fundamental questions that must be answered: 1) **What to transfer?** Indeed, we must understand which part of the knowledge can be transferred from source to target in order to improve the performance on the target task. For instance, we can try to uncover which portion of knowledge is source-specific and which one is shared between source and target. 2) **When to transfer?** We must ensure that the transfer of knowledge is worth the effort; there can be scenarios in which knowledge transfer is detrimental for performances (phenomenon known as negative transfer [76]). 3) **How to transfer?** Once the other two questions have been addressed, we can start analyzing different techniques to transfer knowledge across domains and tasks. Adopting the notation from [80], we formally define a domain  $D$  and a task

## 1.2 The advent of Deep Learning, and its limitations

---

$T$  as  $D = \{X, P(X)\}$  and  $T = \{Y, f(\cdot)\}$ , where  $X$  is the feature space,  $P(X)$  is the corresponding marginal probability distribution,  $Y$  is the label space, and  $f(\cdot)$  is the objective predictive function. Moreover, we use the notations  $D_s$ ,  $D_t$ ,  $T_s$ , and  $T_t$  to indicate source domain, target domain, source task and target task, respectively. Exploring all TL techniques is out of the scope of this thesis, but interested readers can refer to [81, 82] for an overview of TL and domain adaptation (specific type of TL) in medical imaging applications, both for shallow and deep architectures. Since the models used in this PhD thesis are all DL-based, we report hereafter the main variants of TL that can be adopted with deep neural networks [83]:

1. **Fine-tuning**: the DL model is pre-trained on the source domain  $D_s$  and then all its weights are fine-tuned on the target domain  $D_t$ .
2. **Feature Extraction**: the DL model is pre-trained on  $D_s$  and then only some of its weights (typically the last linear layers) are fine-tuned on  $D_t$ . Instead, the convolutional backbone layers are usually “frozen” (i.e. not trained again).
3. **Mixed Training**<sup>1</sup>: the DL model is trained only once on a mixed dataset composed of  $D_s$  and the training portion of  $D_t$ .

The core idea behind fine-tuning and feature extracting is that DNNs learn different features at different layers, with initial layers that have been shown to represent abstract, generic features, and later ones which capture more specific features related to the task at hand [84]. Most papers dealing with medical TL focused on the choice of the source domain  $D_s$ , trying to understand which is the best  $D_s$  from which we should transfer knowledge. For instance, several works investigated the use of natural images (e.g. the ImageNet dataset [85]) for model pre-training [86–89]. Conversely, more recent works showed that the use of natural images for model pre-training could lead to negligible performance improvements [90], and rather suggested that using a medical domain as source is preferable [91–93]. Similarly to the discussion about the choice of  $D_s$ , there is also a lack of consensus in the medical imaging community regarding which type of TL is the most effective (e.g. is fine-tuning better than feature extraction?), with most of the works trying several combinations empirically [83].

In one of the manuscripts presented in this thesis [36], we explore the use of TL: we aim to understand to what extent it is possible to transfer knowledge from a source domain which has a different label distribution from the target domain, a scenario called “*inductive*” TL ( $T_s \neq T_t$

---

<sup>1</sup>Although strictly speaking there is no transfer of knowledge for this subgroup, we loosely include Mixed Training among the TL types.

## 1.2 The advent of Deep Learning, and its limitations

---

because of distinct label spaces  $Y_s \neq Y_t$ ) [80]. More specifically, we address the task of glioma change detection with  $Y_s$  consisting of the above-mentioned weak labels generated automatically from radiology reports, and with  $Y_t$  consisting of manual labels created by human experts ( $Y_s \neq Y_t$ ), again from radiology reports. Details about this work will be provided in section 1.4.4.

### 1.2.5 Prior anatomical knowledge to improve interpretability

In addition to the annotation bottleneck and the domain shift effect, DL models suffer from a third major limitation: the lack of interpretability, an issue often referred to as the “black box problem” of AI [94]. Over the last decade, research on model transparency and explainability has grown steadily, also because of the pervasive adoption of DL systems across the most diverse fields of applications and domains [95]. According to [96], interpretability can be defined as “the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model”. Simple ML models that are traditionally considered interpretable are linear models (linear regression, logistic regression) or decision trees, especially when fed with a limited number of hand-crafted, clinically derived features (e.g. age, sex, history of smoking, etc.) [97]. Conversely, DNNs are considered the least interpretable models because of their inner feature extraction process, their hierarchical structure and their high number of trainable parameters. Several lines of research have been pursued to improve model interpretability in radiology [97]. Among the most common, we find visualization techniques such as saliency maps [98,99], guided backpropagation [100], and gradient-weighted class activation maps (Grad-CAM) [101]. Although in different ways, the main idea behind these techniques is to highlight areas of an image that drive the prediction of the DL model. Other tools that can help increasing interpretability are model-agnostic techniques, such as the Local interpretable model-agnostic explanations (LIME) [102], or Regression Concept Vectors to assess the importance across features [103].

One (less frequent) expedient that can be adopted to reduce model opacity is the use of prior anatomical knowledge [104] as in [35,46,105,106]. The idea behind these studies is to inject anatomical prior knowledge somewhere along the pipeline by specifying body parts of interest, and then guide the ML/DL model to learn from these regions. For instance, the authors in [105] improved contrast phase classification for dynamic CT images by narrowing the analysis only to relevant landmark points. Similarly, the authors in [46] leveraged prior anatomical knowledge to improve the quality of post-hoc explanations generated with LIME [102] on histopathology images. Along the same lines as these works, in one of the manuscript presented in this thesis [35], we exploit prior anatomical

## 1.3 Automated Detection of Cerebral Aneurysms

---

knowledge to constrain the analysis only to parts of the brain that are plausible for the task of aneurysm detection. Details about this work will be described in section 1.3.4. In the next section, we dive into the first clinical task addressed in this PhD thesis: the detection of cerebral aneurysms.

## 1.3 Automated Detection of Cerebral Aneurysms

### 1.3.1 Clinical background

Unruptured Intracranial Aneurysms (UIAs) are abnormal dilatations in the brain arteries caused by a weakness in the blood vessel wall [107]. UIAs typically appear in the form of a bulge or balloon and their prevalence in the adult population ranges between 1% and 5% [108]. UIAs are typically small structures (average diameter  $\approx 5$  mm) and are the predominant cause of nontraumatic SubArachnoid Hemorrhages (SAH) [109]. The mortality rate of aneurysmal SAH is around 40% and only half of post-SAH patients return to independent life. The majority of UIAs occur primarily in proximal arterial bifurcations in the circle of Willis (see Figure 1.4) and 85% of these lesions are anterior in location. About 20% of patients with UIAs have more than one aneurysm [110]. UIAs are more common in women than in men, with a 3:1 ratio. Also, they are more common in elderly people and rarer in children. Although Digital Subtraction Angiography (DSA) is considered the gold standard for diagnosing cerebral aneurysms, its invasive nature limits routine application [111]. Thus, the two non-invasive alternative techniques routinely applied to detect UIAs are TOF-MRA or Computed Tomography Angiography (CTA). In this thesis, we focus on TOF-MRA which, compared to CTA, has the advantage of avoiding radiation exposure. Screening is typically recommended for first-degree relatives of affected family members when two or more members of the family have UIAs or SAH. Untreatable risk factors include old age, female sex, and genetic factors, while treatable factors include smoking and hypertension. When an UIA is found, several factors must be considered to identify the optimal patient management. Even though there are no randomized clinical trials that define the optimal management of an UIA, clinicians can rely on data from prospective or retrospective studies to take an informed decision. To speed up the decision process, Greving et al. devised a standardized grade called PHASES score that indicates the absolute 5-year risk of rupture of UIAs [112]. Since the publication of the work, the PHASES score has been widely adopted since it is fast to compute and it efficiently summarizes the main risk factors. Each risk factor has several categories with associated points which are then summed to obtain a final risk score. The risk factors can be directly derived from the PHASES acronym (**P**: population, **H**: hypertension, **A**: age, **S**: size of aneurysm, **E**: earlier

### 1.3 Automated Detection of Cerebral Aneurysms

---

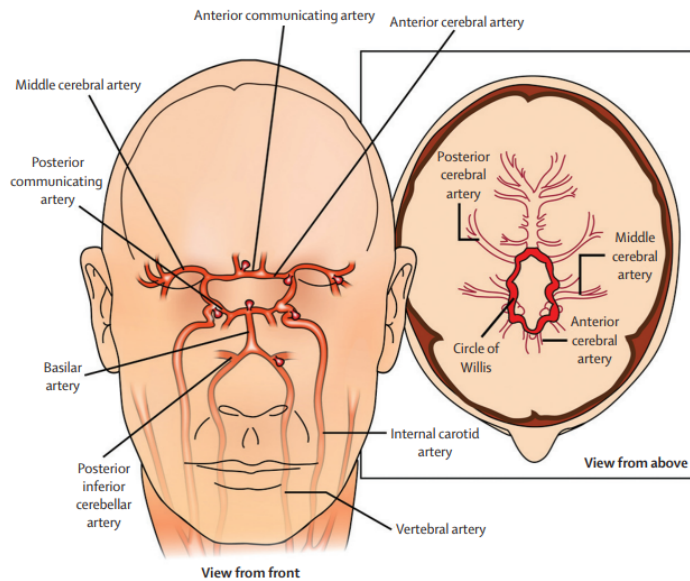


Figure 1.4: Frequent sites of formation of intracranial saccular aneurysms.

SAH, **S**: site of aneurysm). There are two main interventional options for UIAs: surgical clipping and endovascular management, with no large clinical trials comparing the two. For more details, readers are referred to [109].

#### 1.3.2 Aneurysm visual detection

Manually assessing a TOF-MRA scan is a time-consuming process which requires high expertise from experts: radiologists detect aneurysms by selectively scrolling through the TOF-MRA volumes in different planes. For instance, they check in the axial plane the most recurrent locations where aneurysms can occur. Then, the sagittal view permits better views of areas like the basilar trunk; afterwards, the coronal view can be used for areas like the anterior cerebral arteries or the Sylvian segments. In addition, Maximum Intensity Projection (MIP) images can be used to search for stenoses, or to confirm potential aneurysms that were spotted [35]. The visual detection of aneurysms carried out by radiologists presents several limitations: first, sensitivity for small aneurysms ( $<5\text{mm}$ ) can be as low as 35% [113], especially for inexperienced radiologists. More generally, it has also been reported that about 10% of all UIAs are missed during routine clinical practice [114]. Moreover, it may be hard to spot even medium-sized aneurysms on maximum intensity projection (MIP) images because of overlap with adjacent vessels and unusual locations [6]. For these reasons, the development

## 1.3 Automated Detection of Cerebral Aneurysms

---

of a CADe tool able to help clinicians detecting UIAs would be highly beneficial, especially considering that the workload of radiologists is projected to increase in coming years [115].

### 1.3.3 Automated detection

There have been several research groups that proposed automated CADe tools for detecting UIAs throughout the last 20 years. Before the advent of Deep Learning (DL), [25] detected aneurysms with image filtering techniques, [116] proposed a method based on lesion candidate extraction and subsequent false positive reduction, and later [26] used candidate points of interest in the brain arteries to locate aneurysms. Then, starting from 2016, most studies shifted towards the development of DL algorithms, which have permitted to achieve unprecedented performances and have become the de facto standard for UIA detection [117–121]. Despite their success, these DL approaches are still constrained by one of the major bottlenecks presented in sections 1.2.3 and 1.2.4: the lack of large, labeled datasets. This is mainly due to two factors: first, the creation of voxel-wise labels for medical images is tedious and time-consuming; second, none of these TOF-MRA studies to date made their dataset publicly available. This hampers reproducibility and multi-site analyses that are paramount for building robust DL architectures. A significant leap forward for the community was brought by the Aneurysm Detection And segMentation (ADAM) challenge [122]. This allowed for the first time to obtain a fair and unbiased comparison across methods and it revealed the true difficulty of the detection task, considering that none of the top-5 algorithms exceeded a sensitivity of 70% (i.e. 30% of UIAs are missed by all automated methods). In addition, the ADAM training dataset was the first open dataset that could be used in the community for benchmarking. Of all the related studies mentioned above, only [121] evaluated their models on the challenge dataset.

### 1.3.4 Proposed approach

In this section, we describe the methodological contributions of the the first paper included in this PhD thesis [35]. The goal of this manuscript was to develop a fully automated DL network for UIA detection. The work has 4 main contributions. First, to mitigate the data availability bottleneck we explored the use of **weak labels**. Second, to improve model interpretability we leveraged **prior anatomical knowledge** to constrain the analysis only to areas that are anatomically plausible. Third, we **released our in-house dataset** to the community to foster reproducibility and benchmarking. To date (December 2022), this is the largest TOF-MRA labeled dataset available online. Last, we assessed **multi-site generalization** by evaluating our model on the external ADAM

### 1.3 Automated Detection of Cerebral Aneurysms

---

challenge data. Below, we report a summarized version of the Materials & Methods section of the manuscript, starting from an overview of the in-house dataset, and ending with the description of the experiments that we performed. For more details, readers are referred to [35]. Results related to this work will be shown in section 2.1.

**In-house dataset** - We included consecutive patients that underwent TOF-MRA between 2010 and 2015, and for which the corresponding radiological reports were available. Patients with ruptured/treated aneurysms or with other vascular pathologies were excluded. Totally thrombosed aneurysms and infundibula (dilatations of the origin of an artery) were likewise excluded. In total, we retrieved 284 TOF-MRA subjects: 157 had one (or more) UIAs, while 127 did not present any. The dataset was anonymized and organized according to the Brain Imaging Data Structure (BIDS) standard [123]. It is available on OpenNeuro [124] at <https://openneuro.org/datasets/ds003949>.

**Aneurysm annotation, size, location and risk groups for in-house dataset** - Aneurysms were annotated by one radiologist with 2 years of experience in neuroimaging, and double-checked by a senior neuroradiologist with over 15 years of experience. Two annotation schemes were followed:

1. Weak labels: for most subjects (246/284), the radiologist used the Multi-image Analysis GUI (Mango) software to create the aforementioned weak labels. These correspond to spheres that enclose the whole aneurysm, regardless of the shape. The size of the spheres was chosen manually by our radiologist on a case-by-case basis ensuring that the whole aneurysm was entirely enclosed within the sphere. A visual example of one weak label is shown in Figure 1.5.
2. Voxel-wise labels: for the remaining subjects (38/284), the radiologist used ITK-SNAP (v.3.6.0) [125] to create voxel-wise labels drawn slice by slice scrolling in the axial plane.

The overall number of UIAs included in the study was 198 (178 saccular, 20 fusiform). These were grouped according to the PHASES score presented in section 1.3.1. In addition, for post-hoc analyses, we divided the UIAs into two groups based on their risk of rupture: low-risk and medium-risk. Aneurysms in the low-risk group are those that are monitored over time, but do not require any intervention. Instead, aneurysms in the medium-risk group can be considered for treatment. We computed for each aneurysm a partial PHASES score that only considered size, location, and patient's age: if an aneurysm had partial PHASES score  $\leq 4$ , it was assigned to the low-risk group, while if it had a partial score  $> 4$ , it was assigned to the medium-risk group. After removing fusiform UIA (the PHASES score was built for saccular UIA) and extracranial carotid artery UIA (they do not bleed in the subarachnoid space), we ended up with 141 low-risk and 23 medium-risk aneurysms.

### 1.3 Automated Detection of Cerebral Aneurysms

---

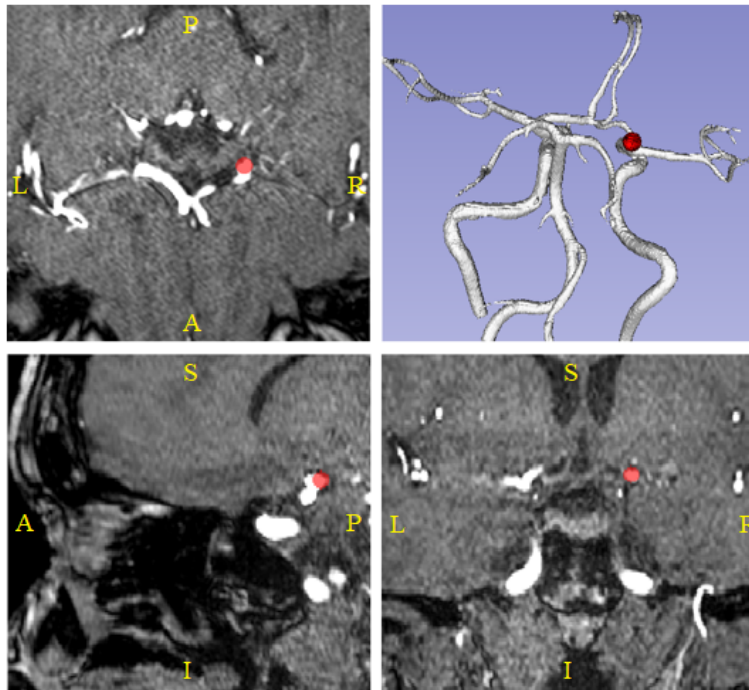


Figure 1.5: TOF-MRA orthogonal views of a 62-year-old female patient. Red areas correspond to our oversized spherical weak labels. Top-left: axial plane; top-right: 3D posterior reconstruction of the cerebral arteries; bottom-left: sagittal plane; bottom-right: coronal plane.

**Data processing** - Several preprocessing steps were carried out for each subject. First, we performed skull-stripping with the FSL Brain Extraction Tool [126]. Second, we applied N4 bias field correction [127]. Third, we resampled all volumes to a median voxel spacing. Last, a probabilistic vessel atlas built from multi-center MRA datasets [128] was co-registered to each patient’s TOF-MRA using ANTS [129]. The atlas was used both during training and inference (see below).

**Network, Cross-Validation, Metrics and Statistics** - The deep learning model used in this study is a custom 3D UNET, inspired by the original work [130]. We used 3D TOF-MRA patches as input to our network. We set the side of the input patches to 64x64x64 voxels to include even the largest aneurysms. All patches were Z-score normalized, as is common practice [131]. Further details about the network can be found in [35].

**Cross-validation** - To evaluate detection performances, we conducted a five fold cross-validation on the 246 subjects with weak labels (details in [35]). In order to make results comparable across experiments, we always used the same cross-validation split. In all experiments on the in-house dataset, we always pre-trained our network on the whole ADAM training dataset and then fine-



### 1.3 Automated Detection of Cerebral Aneurysms

---

tuned it on the in-house training data. Ablation experiments of domain adaptation across the two datasets can be found in Supplementary Materials of the manuscript. The code used for this work is available at [https://github.com/connectomicslab/Aneurysm\\_Detection](https://github.com/connectomicslab/Aneurysm_Detection)

**Metrics and Statistics** - In line with the ADAM challenge, we used sensitivity and false positive (FP) rate as detection metrics. A detection was considered correct if the center-of-mass of the predicted aneurysm was located within the maximum aneurysm size of the ground truth mask. In addition, we computed the Free-response Receiver Operating Characteristic (FROC) curve [132]. To compare different model configurations, we used a two-sided Wilcoxon signed-rank test of the areas under the FROC curves across test subjects, as similarly performed in [133]. To compare the performances of a configuration with respect to aneurysm rupture risk, location and size we performed several Chi-squared tests [134]. For all tests, we set a significance threshold  $\alpha = 0.05$ .

**Experiments** - In this section, we will present the four experiments that we conducted.

**1) Use of Weak Labels** - The goal of this experiment was to answer the following questions: 1) how much faster is the creation of weak labels with respect to the creation of voxel-wise labels? 2) what is the impact of using weak labels in terms of detection performances when comparing to voxel-wise labels? To answer the first question, we selected a subset of 14 patients (mean aneurysm size (s.d.) = 5.2 (1.0) mm), and we assessed the time difference between the two annotation schemes (i.e. all 14 patients were annotated first with weak labels, and then with voxel-wise labels). To answer the second question, we used the 38 subjects with voxel-wise labels and for these patients we artificially generated corresponding weak spherical labels (“weakened” labels, details in Supplementary of the paper). Then, to evaluate the influence of annotation quality (weakened vs. voxel-wise) in terms of detection performances, we conducted 3 experiments in which we used an increasing number of patients with voxel-wise labels: (i) all 38 patients with weakened labels, (ii) 19 patients with weakened labels and 19 with voxel-wise labels, and (iii) all 38 patients with voxel-wise labels.

**2) Use of Anatomical Information** - Because the task of aneurysm detection is extremely spatially constrained, we exploited the prior information that aneurysms a) must occur in vessels, and b) tend to occur in specific locations of the vasculature. To include this domain knowledge, one of our radiologists pinpointed in the probabilistic vessel atlas the location of 20 landmark points where aneurysm occurrence is most frequent (list in Supplementary Materials). These points were chosen according to the literature [109] and were co-registered to the TOF-MRA space of each subject, as illustrated in Figure 1.6. We exploit this domain knowledge both during training and inference:

**Training** - We apply an anatomically-informed selection of training patches to sample both negative (without aneurysms) and positive (with aneurysms) patches. Specifically, 8 positive patches

### 1.3 Automated Detection of Cerebral Aneurysms

---

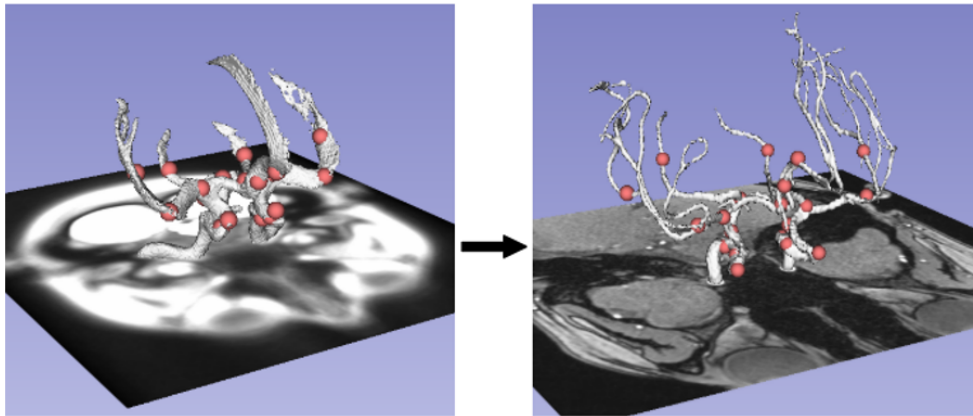


Figure 1.6: left: 20 landmark points (in red) located in specific positions of the cerebral arteries (white segmentation) in MNI space. right: same landmark points co-registered to the TOF-MRA space of a 21-year-old, female subject without aneurysms.

per aneurysm were randomly extracted in a non-centered fashion. Then, we extracted 50 negative patches per TOF-MRA volume. Out of these, 20 were centered in correspondence with the landmark points, 20 were patches containing vessels (details in Supplementary Materials), and 10 were extracted randomly. Overall, this sampling strategy allows us to extract most of the negative patches which are comparable to the positive ones in terms of average intensity. To mitigate class imbalance, we applied data augmentations on positive patches: namely, rotations ( $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ ), flipping (horizontal, vertical), contrast adjustment, gamma correction, and addition of gaussian noise.

**Inference** - The patient-wise evaluation was performed following a sliding window approach. We exploited again the prior anatomical knowledge described above by retaining only the patches which are both within a minimum distance from the landmark points and fulfill specific intensity criteria (details in Supplementary). The rationale behind this choice was to only focus on patches located in the main cerebral arteries, as shown in Figure 1.7. Two post-processing steps were adopted: first, we kept a maximum of 5 candidate aneurysms per patient (only the 5 most probable). Second, we applied test-time augmentation to increase sensitivity.

**Validation** - To validate the effectiveness of our two anatomically-informed expedients, we first evaluated an anatomically-agnostic baseline where none of the two expedients is used and the 38 added subjects have weakened labels. Second, we evaluated the same anatomically-agnostic baseline (none of the two expedients used) but with the 38 subjects having voxel-wise labels. Third, we tested one model where only the anatomically-informed patch sampling is carried out. Last, we computed

### 1.3 Automated Detection of Cerebral Aneurysms

---

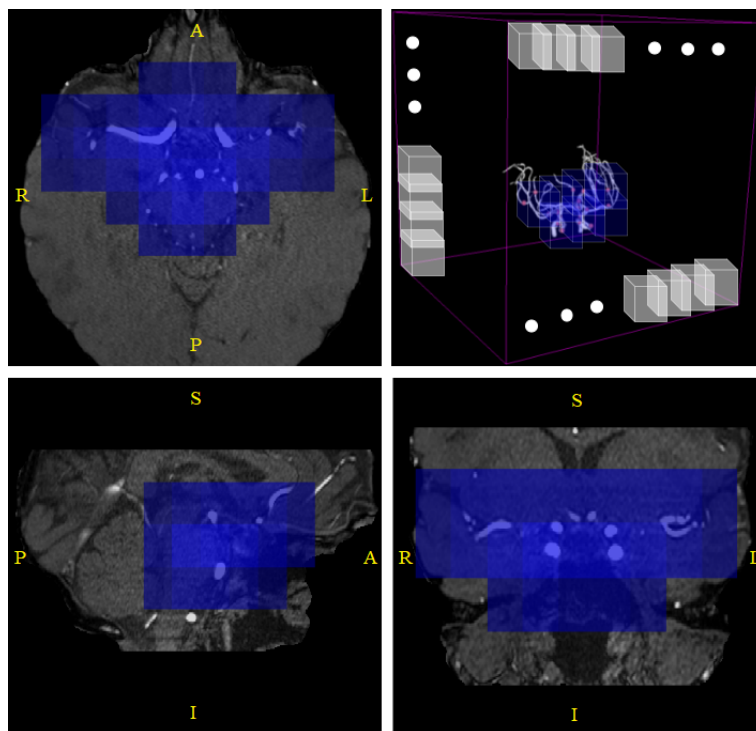


Figure 1.7: TOF-MRA orthogonal views of a 62-year-old female subject: blue patches are the ones which are retained in the anatomically-informed sliding-window approach. (top-right): 3D schematic representation of sliding-window approach; out of all the patches in the volume (white patches), we only retain those located in the proximity of the main brain arteries (blue ones).

performances when only the anatomically-informed sliding window is performed.

**3) Participation to the ADAM Challenge** - To evaluate model performances in data coming from a different institution, we participated to the ADAM challenge. The ADAM training dataset is composed of 113 TOF-MRA (93 patients with 125 UIAs, 20 controls). The voxel-wise annotations were drawn in the axial plane by two radiologists. Instead, the test dataset is made of 141 cases (117 patients, 26 controls) and it is solely used by the organizers to compute patient-wise results.

**4) Performances with Respect to Risk-of-rupture, Location and Size** - We investigated how detection performances would vary with respect to the risk-of-rupture groups described above. In addition, we explored how performances would vary with respect to aneurysm location and size.

### 1.3.5 Clinical evaluation

In this section, we introduce the next step for the aneurysm detection project. This is a work in progress that we are planning to submit soon for consideration in a journal. The goal of this new work will be to assess the clinical validity of the CADe tool developed in [35]. Specifically, we are planning to conduct a retrospective, within-subject reading with two radiologists, including both patients with UIAs and control subjects. The two readers (one senior and one junior) will be asked to visually inspect 140 TOF-MRA scans (70 patients, 70 controls, distinct from the 284 of [35]) under two different settings:

1. **Baseline:** the readers only have access to the original TOF-MRA sequence. The readers can freely explore all three views (axial, sagittal, coronal), as well as the MIP reconstruction.
2. **ML-assisted:** the readers have access both to the original TOF-MRA sequence (3 views + MIP) and to the same sequence that contains potential aneurysm candidates generated by the DL model proposed in [35]. We denote this sequence as *overlay*.

Patient age and sex will be visible to the readers under both settings. We will conduct two reading sessions: on the first session, the 140 subjects will be randomly assigned either to the Baseline or to the ML-assisted setting, while on the second session, the opposite scenario will be presented to the readers. Although the order of Baseline/ML-assisted is random, we will keep the same random order for both readers (junior and senior), since patient order might influence detection performances, for instance due to fatigue. The time interval between the two reading sessions will be of one month. The readers will be blinded to the performance levels of the CADe system, and to the exact aneurysm prevalence in the test set, in order to avoid expectation bias [135]. The outcomes that we are planning to monitor across the two settings (Baseline and ML-assisted) are, for each reader, sensitivity, specificity, reading time, and confidence in the diagnosis. Moreover, we will also measure intra- and inter-rater agreement. The statistical test used to measure aneurysm-wise sensitivity and subject-wise specificity will be the McNemar’s test [136]. To establish the number of patients with aneurysm needed to run the paired reading (N=70), we ran a sample size calculation assuming an increase of 15% in sensitivity for the junior radiologist, a power of 80% and a Type I error  $\alpha = 0.05$ , and the same for establishing the number of controls (N=70). To find the top-performing model that will be used for inference on the 140 TOF-MRA scans, we will leverage once again TL, with the goal of uncovering the most efficient way to transfer knowledge from the ADAM source dataset to our in-house target dataset (e.g. finetuning vs. feature extracting vs. mixed training).

# 1.4 Change Detection in Longitudinal Glioma Imaging

In this section, we introduce the second radiological task addressed in this PhD thesis: change detection in glioma imaging.

### 1.4.1 Clinical background

Gliomas represent 30% of all primary brain tumors, 80% of all malignant ones, and account for most deaths caused by primary brain tumors [137]. Like most other tumors, they are essentially genetic alterations of single cells whose pattern shapes the clinical features of the tumor. This group of tumors is highly heterogeneous and includes astrocytomas, oligodendrogliomas, mixed oligoastrocytic gliomas or ependymomas. The World Health Organization (WHO) has established a grading system for glioma tumors that goes from I to IV, with grade IV (glioblastoma) indicating the most aggressive variation [138]. The grading is based on five histopathology criteria that are related to the degree of anaplasia (lack/loss of differentiation): cellular density, nuclear atypia, mitosis, endothelial proliferation and necrosis. Readers interested in the grading system can refer to [138,139]. The incidence of gliomas increases with age and the only factor that is recognized as a direct causative agent is ionizing radiation (exposure to a therapeutic dose) [140–142]. Gliomas are associated with rare familial syndromes such as Neurofibromatosis type 1, neurofibromatosis type 2, tuberous sclerosis, Li–Fraumeni, and Turcot syndrome. However, these syndromes only account for  $\leq 1\%$  of all gliomas [143]. The most common symptoms that might suggest the presence of glioma are headache, nausea, memory loss, seizure, personality changes, weakness in the arms, face or legs, numbness and problems with speech. The principal treatment for gliomas is surgical resection followed by radiation and/or chemotherapy, with temozolomide (Temodar) being the most frequently used chemotherapy drug. The gold standard imaging technique to detect glioma tumors is MRI, though the definitive diagnosis of gliomas can only be obtained via histology [144,145]. A multitude of MRI sequences is routinely performed to monitor glioma patients [146]. The standard imaging protocol recommended by the European Society of NeuroRadiology (ESNR) includes the following sequences: 3D unenhanced T1-weighted (T1w), 3D T1w enhanced with gadolinium-based contrast agent, axial 2D T2-weighted (T2w), 2D FLuid Attenuated Inversion Recovery (FLAIR), and axial 2D Diffusion-Weighted Imaging (DWI)/Apparent Diffusion Coefficient (ADC). Protocol extensions might also include more advanced sequences such as Diffusion Tensor Imaging (DTI), Susceptibility Weighted Imaging (SWI), Dynamic Susceptibility Contrast (DSC)-perfusion and MR spectroscopy. To date, no clear consensus exists for these advanced sequences. However, it has been shown that multi-parametric MRI (the combined use of

## 1.4 Change Detection in Longitudinal Glioma Imaging

---

standard and advanced sequences) leads to a more accurate tumor characterization [147]. During the last decade, a growing body of research has focused on quantifying tumor structures on multi-parametric MR images and finding correlations between imaging and clinical features (as in [148]). A crucial step to achieve these clinical tasks is the segmentation of tumor sub-compartments (such as enhancing tumor, tumor core, whole tumor, necrotic core), a task that was facilitated by the Brain Tumor Segmentation (BraTS) challenge [149–151]. This was first organized in 2015 and was later re-proposed every year with constant improvements to the dataset and tasks. The open BraTS dataset is one of the largest annotated datasets for segmentation and has become a milestone reference for benchmarking ML and DL algorithms.

### 1.4.2 MRI-based longitudinal monitoring

Neuroradiological monitoring of gliomas has a crucial role for primary diagnosis and post-therapeutic follow-up. For grades I and II (Low-Grade Gliomas, LGGs), imaging is conducted to monitor tumor stability and evaluate possible anaplastic transformations. For grades III and IV (High-Grade Gliomas, HGGs) imaging serves to distinguish, among other things, therapy-induced changes from actual tumor-related changes. Since these two phenomena exhibit overlapping features (e.g. surrounding oedema), discrimination for radiologists can be challenging. To facilitate and standardize diagnostic monitoring of gliomas, several guidelines have been proposed in the literature. Since in this PhD thesis we work with HGGs, here we report the most famous guidelines for HGGs. Readers interested in longitudinal monitoring of LGGs can refer to [152]. For HGGs, one of the first widely used guidelines were the Macdonald criteria [153]. However, with time these showed some limitations [154] which were overcome in 2010 with the Response Assessment in Neuro-Oncology (RANO) criteria [155]. Along with the 4 main types of response already present in the Macdonald criteria (progressive disease (PD), partial response (PR), complete response (CR) and stable disease (SD)), the RANO criteria helped to deal with more subtle scenarios involving non-enhancing tumour areas, radiation-induced pseudoprogression, surgery-induced enhancements, and pseudoresponse after treatment with antiangiogenic therapies. Although more accurate, also the RANO criteria showed some limitations. For instance, it was shown that bidirectional (2D) measurements of contrast enhancing tumor overestimate tumor volume and have high reader discordance [156]. In addition, the thresholds used to define response and progression are relatively arbitrary and are not backed up by sufficient literature. To mitigate these limitations, the modified RANO (mRANO) criteria were proposed in [156]. Details regarding mRANO can be found in [156].

### 1.4.3 Extracting weak labels from radiology reports with Natural Language Processing

As we have seen in previous paragraphs, radiological monitoring of gliomas is extremely complex and heterogeneous. In the following sections, we present 2 of the manuscripts included in this PhD thesis. The goal of these works is to investigate the creation of weak labels from radiology reports, the use of TL between weakly-annotated and human-annotated data, and the impact of model capacity in medical TL. The clinical scenario that we envision is one where the developed CADe tool would highlight relevant, tumor-related changes with respect to the previous exam to facilitate and speed up the diagnosis for radiologists, while ideally providing quantitative indicators. The proposed CADe tool is a DL-based model and it is explained more in detail in section 1.4.4. As we have learned in section 1.2.3, the availability of large annotated dataset is a recurrent bottleneck for DL applications. One solution to overcome this bottleneck that was discussed in section 1.2.3 is the use of NLP on medical reports and that is exactly what we carried out for the glioma change detection project. In the rest of this section, we provide a summarized version of the Materials & Methods section of the work [47], the second manuscript of this PhD thesis.

**Dataset** - We retrospectively included 164 subjects that underwent longitudinal MR glioma follow-up in the university hospital of Lausanne (CHUV) between 2005 and 2019. 71% of the patients in the cohort had Glioblastoma Multiforme (GBM), while the remaining 29% had either an oligoastrocytoma or an oligodendroglioma. At every session, a series of MR scans were performed including structural, perfusion and functional imaging. For the sake of this study, we only focused on the native T1-weighted (T1w) scan, the T2-weighted (T2w) scan and the T1w-gad (post gadolinium injection, a contrast agent). For 25 patients, we collected images and reports across multiple sessions (on average, 9 sessions per subject). For the remaining 139 patients, we only retrieved images and reports from 1 random session. Overall, we ended up with a dataset of 361 radiology reports to use for the NLP pipeline. Every report was written (dictated) in French during routine clinical practice by a junior radiologist after exploring all sequences of interest. Then, a senior radiologist reviewed each case amending the final report when necessary.

**Report Tagging** - In order to build a supervised document classifier, one radiologist (4 years of experience in neuroimaging) tagged the reports with labels of interest. For each report, the annotator was instructed to perform two separate tasks: first, she had to assign 3 classes to the reports; one class that indicated the global conclusion of the report, one class to indicate the evolution of the enhanced part of the lesion (T1w conclusion) and the last one to indicate the evolution of the lesion

## 1.4 Change Detection in Longitudinal Glioma Imaging

---

on T2-weighted sequences (T2w conclusion). For each of these three groups, the annotator could choose between the following labels:

- **Stable**: assigned when the tumor did not change significantly with respect to the previous comparative exam.
- **Progression**: assigned when the tumor worsened with respect to the previous comparative exam. This class included cases where the enhanced part of the tumor increased in size or when the T2 signal anomalies surrounding the tumor increased in extension.
- **Response**: assigned when the tumor responded positively to the treatment.
- **Unknown**: used when the annotator was not able to assign any of the three classes above.

The second task of the annotator was to highlight the most recent comparative date in the reports. Since the reports are not structured, this helped linking the current report being tagged with the most meaningful previous one. For simplicity, in this work we only focused on the global conclusion of the reports, and not on the T1w and T2w conclusions. Also, we removed all cases that were tagged as **unknown** (21 reports) and we merged **progression** and **response** into one unique class which we denote as **unstable**. By doing this, we narrowed the task to a binary classification problem where the model tries to distinguish between **stable** (N=191) and **unstable** (N=149) reports. To facilitate the annotation process, we utilized the open-source software Daturks<sup>3</sup>.

**Text Preprocessing & Embedding** - Several preprocessing steps were carried out to reduce the vocabulary size. First, we removed all proper nouns such as physicians' and patients' names. Second, all the words in the reports were converted to lowercase. Third, we removed punctuation and the most common French stop words. Among these, we ensured to keep the French negation 'pas' (*not*) since it is very frequent in the reports, and reverses the meaning of the sentence. Fourth, all reports were tokenized using the *wordpunct* class of the Natural Language Toolkit framework (version 3.6.1) [157]. As last step, since all the reports contain the three sections '*indications*', '*description*' and '*conclusion*', we removed all content before the '*indication*' section, which is either useless (e.g. department phone number) or sensitive (e.g. patient identifier).

A key step in any NLP pipeline is text embedding. This corresponds to the conversion of tokenized text into numerical vectors. In this work, we compare two widespread embedding techniques, namely TF-IDF [158] and Doc2Vec [159]. The former is a standard term-weighting embedding scheme

---

<sup>3</sup>OpenSource Data Annotation tool - <http://github.com/DataTurks/DataTurks>



## 1.4 Change Detection in Longitudinal Glioma Imaging

---

(classical ML) that preserves the length of the tokenized documents, while the latter is a DL-based technique that creates dense vectors which encode word order and context.

**Experiments** - All experiments were run in a 5-fold, nested, stratified cross validation (CV). The internal CV was used to tune the hyperparameters of the pipeline with a custom Grid Search algorithm. Instead, the external CV was used to compute results on hold-out test samples. For details regarding hyperparameter optimization, readers are referred to [47]. To avoid overoptimistic predictions, we also ensured that the reports from multiple sessions of the same subject were not present some in the train set and some in the test set. Furthermore, to reduce the bias introduced by the random choice of patients at each CV split, the whole nested CV was repeated 10 times, each time performing the splitting anew, and results were averaged. For all experiments, we adopted the Random Forest algorithm [160] to classify the embedded documents. To compare the two pipelines (Doc2Vec vs. TF-IDF embedding), we computed all standard classification metrics, and we plotted the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. The reports indicating tumor stability were considered as negative samples, whereas those indicating a change in the tumor were considered as positives. The classification metrics and the curves were averaged across the 10 runs. To statistically compare the results, a Wilcoxon signed-rank test was performed [161] on the area under the ROC curve (AUC) across the 10 runs, with a significance threshold level  $\alpha = 0.05$ .

We conducted an explainability analysis with the LIME toolkit to identify the most important words that influenced the final prediction. This was run on the TF-IDF pipeline only since it resulted in higher performances (as we will see in section 2.2.1). We set the best hyperparameters obtained across the random runs and we ran LIME over all test reports. All the code developed for this study is available at [https://github.com/connectomicslab/Glioma\\_NLP](https://github.com/connectomicslab/Glioma_NLP).

### 1.4.4 Glioma change detection using weak labels and transfer learning

In this paragraph, the third manuscript [36] included in the PhD thesis is presented. The goal of this work is to tackle image-based glioma change detection. The main contributions of the manuscript are the following: (i) we propose a Transfer Learning (TL) approach that leverages inexpensive and fast-to-create weak labels generated from the report classifier of [47]; (ii) we automate the choice of TL type, treating it as another hyperparameter to optimize, and thus avoiding manual empirical trials; (iii) we assess the impact of model size on the proposed TL pipeline and (iv) we evaluate our pipeline on the longitudinal subjects of the public BraTS 2015 dataset. In the following paragraphs, the Materials & Methods section of the paper is summarized.

## 1.4 Change Detection in Longitudinal Glioma Imaging

---

**Related works** - Previous works addressing glioma change detection, NLP to generate weak labels, the automation of TL, or the influence of model size on TL are described in [36].

**In-house dataset** - We retrieved 2100 MR scans belonging to 183 retrospective patients with high-grade gliomas who were scanned between 2004 and 2019 at the Lausanne University Hospital. At every session, a series of MR scans including structural, perfusion and functional imaging were performed. For simplicity, in this work we only focused on the T2-weighted (T2w) scans. Scans that were too close to surgery (within 4 weeks) were excluded since they contained exaggerated intensity changes and brain deformations around the resection cavity. In addition, we extracted the radiology reports associated with each session. We released an anonymized version of our dataset on Zenodo<sup>4</sup>.

**BraTS dataset** - To assess the generalization of our pipeline to an external dataset, we ran inference on the longitudinal subjects of the BraTS 2015 dataset. We selected the 2015 edition because it is the only one that contains patients with multiple scans. Out of the 20 available longitudinal patients, we discarded 5 because they only contained two scans (before and after resection). For the remaining 15 subjects, we used 59 MR scans (average of 4 scans per subject), again only focusing on T2w scans. From these 59 scans, we generated 51 difference maps (creation process described below) which were tagged by one radiologist with over 18 years of experience, using the labels presented in the next paragraph. We openly released these labels for other researchers.

**Report Tagging** - From the 183 glioma patients of the in-house dataset, we created two sub-datasets: a Human-Annotated Dataset and a Weakly-Annotated Dataset.

**Human Annotated Dataset (HAD)** - For this sub-dataset, three radiologists tagged the MR radiology reports with labels of interest following the same procedure described in section 1.4.3. In total, 381 reports (belonging to 169 distinct patients) were manually annotated by human experts. Out of these 381, 39 reports (belonging to 39 distinct patients) were tagged by a senior radiologist with over 18 years of experience in neuroimaging (P.H), while 342 reports (belonging to 162 patients) were tagged by two radiologists both with 4 years of experience. Cohen’s kappa coefficient between the two readers for the T2w conclusion was  $k = 0.80$  which is considered a “substantial agreement” according to [162]. The 41/342 reports for which the two annotators disagreed were discarded. Also, we discarded 90 reports for which the T2w conclusion was different from the global conclusion. Last, we also excluded the 17 reports for which the T2w conclusion was unknown.

**Weakly Annotated Dataset (WAD)** - For this sub-dataset, reports were tagged with the classifier proposed in our previous work [47] (described in section 1.4.3). We denote the labels generated from the report classifier as *weak* because the classifier will commit errors, and because, differently from

---

<sup>4</sup>DOI: 10.5281/zenodo.7214605

## 1.4 Change Detection in Longitudinal Glioma Imaging

human readers, it cannot abstain when the reports are unclear (i.e. there is no **unknown** label).

Both for **HAD** and **WAD** we merged **progression** and **response** into one class which we denote as **unstable**. This narrowed the task to a binary classification problem where we try to distinguish between **stable** and **unstable** reports. After these modifications, HAD contained 233 reports (159 stable, 74 unstable), whereas WAD contained either 795 (333 stable, 462 unstable) or 361 (165 stable, 196 unstable) reports, depending on the hyperparameter *fraction\_of\_WAD* presented below.

**Image-based change detection** - Every radiology report links two time points, namely the current scan and a previous scan which is used as baseline for comparison. Thus, for each report, we generated a corresponding T2w absolute difference map as illustrated in Figure 1.8. The rationale

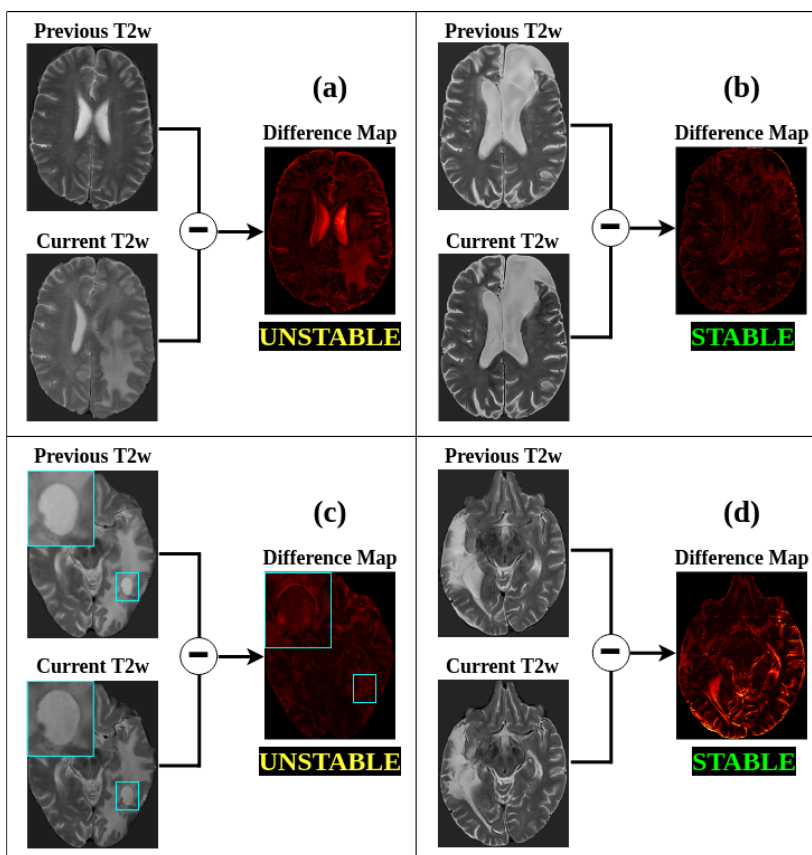


Figure 1.8: Creation of T2w difference maps. After registration and normalization of the previous and current T2w volumes, the maps are computed via voxel-wise absolute difference. (a) 58-year-old male patient with a progressing gliosarcoma. (b) 59-year-old male patient with a stable astrocytoma. (c) 60-year-old male patient with seemingly stable glioblastoma, but with enlarging cystic lesion (zoom in cyan color). (d) 60-year-old male patient with a (less evident) stable oligodendroglioma.

## 1.4 Change Detection in Longitudinal Glioma Imaging

---

behind these difference maps is that parts of the tumor that either progress or respond to treatment (unstable) should appear as hyper-intense; instead, if the tumor is stable across the two time-points, the difference map will likely be hypo-intense overall. To generate the difference maps, we first applied N4 bias field correction with ANTs [163] both to the previous and to the current T2w volumes. Second, we registered the previous volume to the current volume. Third, we skull-stripped both volumes (previous warped and current) with HD-BET [164]. Fourth, we applied z-score normalization on both volumes. Last, we computed the absolute voxel-wise difference of the normalized volumes.

**Classification Networks** - The image-based change detection is treated as a binary classification problem: as for the reports, we try to classify the difference maps into stable and unstable. We used two Convolutional Neural Networks (CNNs) for the classification of the T2w difference maps: a custom 3D-VGG [165] (henceforth called VGG) and a 3D-ResNeXt [166] with Squeeze-and-Excitation [167] (henceforth called SEResNeXt). Details about the networks can be found in [36].

**Experiments & Hyperparameter tuning** - To create the weak labels, we adapted the report classifier [47] and trained it to classify the T2w conclusion (in [47] it was trained to classify the global conclusion). We ran a nested 5-fold cross-validation on the 233 HAD reports, selected the best hyperparameters, and finally performed inference with the best model on all WAD reports to obtain the weak labels later used for the image-based change detection.

**Image-based glioma change detection** - Because of computational constraints, we decided to fix some hyperparameters, and tune others. Among the fixed (i.e. not tuned) hyperparameters we chose a batch size of 4 and 60 training epochs with early stopping. Depending on the experiments detailed below, other hyperparameters were tuned using the Optuna framework [168] (details in [36]). To understand which TL type is the most appropriate to improve classification performances and how model capacity can influence TL results, we performed two experiments (called **Baseline** and **TL**) with the two CNNs models described above (VGG and SEResNeXt): in the **Baseline** experiment, we conducted a 5-fold cross-validation only on HAD, and WAD was intentionally not used (no TL). Evaluation was performed on the test subjects of each cross-validation fold and then results were aggregated. The only two hyperparameters that were tuned for the **Baseline** experiment were *learning\_rate* and *weight\_decay*. The former was chosen from  $\{1 \times 10^{-4}, 1 \times 10^{-5}, 1 \times 10^{-6}\}$ , whereas the latter was chosen from  $\{0, 0.01\}$ . In the **Transfer Learning (TL)** experiment, we still performed a 5-fold cross-validation on HAD, but this time we also exploited the WAD difference maps. In addition to *learning\_rate* and *weight\_decay* (which are tuned identically to the **Baseline**), here we also searched for the best TL configuration. Specifically, we tuned 3 additional hyperparameters: *mixed\_training*, *feature\_extraction* and *fraction\_of\_WAD*.

## 1.4 Change Detection in Longitudinal Glioma Imaging

---

- *mixed\_training* can either be True or False: if True, we use for training a mixed shuffled dataset that is composed of WAD difference maps and the difference maps of the training HAD patients; if instead *mixed\_training* is False, we either perform feature extraction if *feature\_extraction* is True, or fine-tuning if *feature\_extraction* is False.
- *fraction\_of\_WAD* indicates which portion of WAD to use. We added this hyperparameter because not all weakly-labeled data is necessarily useful. In other words, by tuning *fraction\_of\_WAD* we wanted to understand whether some reports (and hence some difference maps) are more informative than others. The tunable values that we chose for *fraction\_of\_WAD* were  $\{WAD > 0.75, WAD > 0.95\}$  where 0.75 and 0.95 are the output probabilities (soft labels) of the report classifier from [47]. For instance, when using  $WAD > 0.95$  we only use the reports for which the report classifier is highly confident (output probability  $> 0.95$ ).

Since running all combinations would have been impractical, we only ran each TL experiment (VGG-TL and SEResNeXt-TL) for 4 days. To summarize, we ran 4 experiments: VGG-Baseline, VGG-TL, SEResNeXt-Baseline, and SEResNeXt-TL. The comparisons between **Baseline** and **TL** aimed to assess the effectiveness of the weak labels in WAD. Instead, comparisons between the two CNNs aimed to understand the influence that model capacity can have on TL strategies for our task.

**Cross-Validation** - For the **Baseline** experiments, we performed a 5-fold cross-validation on HAD. At each cross-validation split, 80% of the subjects were used to train the CNN (either VGG or SEResNeXt), while the remaining 20% of the subjects were used to compute test results. Within each cross-validation fold, we also used 25% of the training subjects as validation set for tuning the hyperparameters. To avoid over-optimistic results, the cross-validation splits were always performed at the subject level. For the **TL** experiments, we performed the same 5-fold cross-validation on HAD, but then adapted the learning strategy according to the hyperparameters chosen during hypertuning. We ensured that the same splits were performed on HAD both for the **Baseline** and **TL** experiments.

**Metrics, Statistics & Code** - The task that we address is binary classification of the T2w difference maps which are labeled either as **stable** or **unstable**. We report in the Results section accuracy, sensitivity, specificity, F1 score, AUC, and Area Under the Precision-Recall curve (AUPR). We consider the class **unstable** as “positive”, and the class **stable** as “negative”. To statistically compare the four different models presented above, we ran permutations tests using the difference in AUCs as test statistic, as similarly performed in [169]. We set a significance threshold  $\alpha = 0.05$  and we ran 10000 permutations for each test. The code used for this paper and the corresponding configuration files are available at [https://github.com/connectomicslab/Glioma\\_Change\\_Detection\\_T2w](https://github.com/connectomicslab/Glioma_Change_Detection_T2w).

# Chapter 2

## Summary of Results

In the following pages, the main results of the papers included in the thesis [35,36,47] are summarized.

### 2.1 Automated Aneurysm Detection

In [35], our weak labels resulted to be four times faster to generate with respect to their voxel-wise counterparts (two-sided Wilcoxon signed-rank test – annotation timings,  $W=0$ ,  $p=0.001$ ). The model that achieved the highest sensitivity (83%, with false positive rate of 0.8) across the test folds was the one for which we applied the anatomically-informed sliding window approach during inference, but not the anatomically-informed patch sampling during training. Figure 2.1 illustrates the FROC curves of 3 models under different configurations. Model 7 is the top-performing. When evaluating this model on the ADAM test dataset, we achieved a sensitivity of 68% (false positive rate of 2.5) and ranked 4th/18 on the open leaderboard. We found no significant difference in sensitivity between aneurysm risk-of-rupture groups ( $p=0.75$ ), locations ( $p=0.72$ ), or sizes ( $p=0.15$ ).

### 2.2 Glioma Change Detection

#### 2.2.1 Report classification with Natural Language Processing

The report classifier developed in [47] reached 89% AUC when distinguishing reports indicating tumor stability and tumor instability, with the TF-IDF embedding performing significantly better than Doc2Vec (Wilcoxon signed rank test comparing AUC distributions,  $P = 0.009$ ). Table 2.1 illustrates classification results for both embedding schemes.

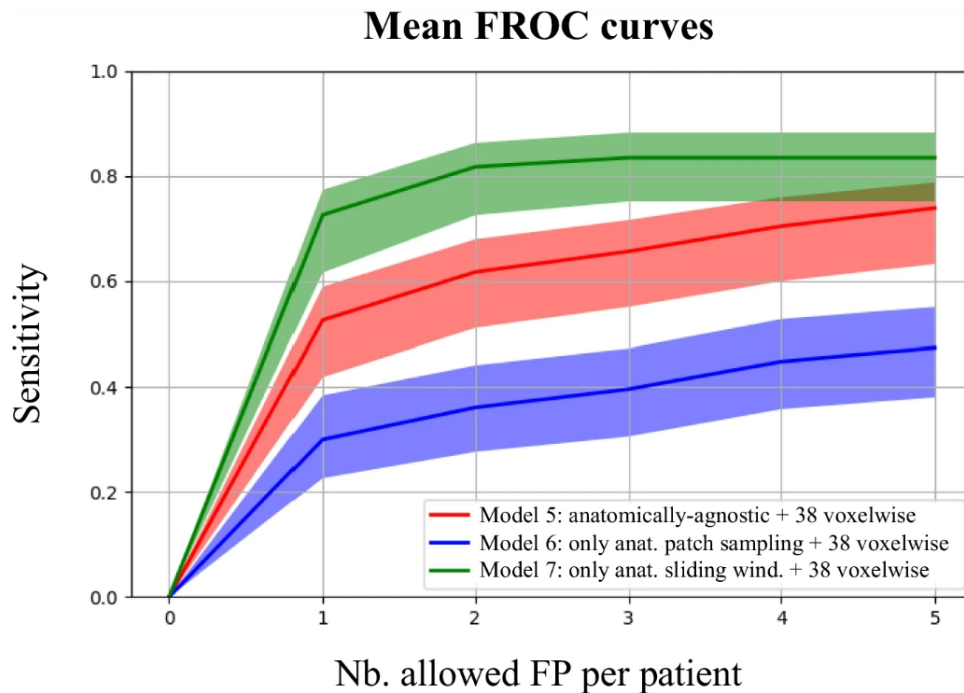


Figure 2.1: Mean Free-response Receiver Operating Characteristic (FROC) curves across the five test folds of the cross-validation. Shaded areas represent the 95% Wilson confidence interval. The three models correspond to Model 5, Model 6, and Model 7. Anatomically-agnostic model: none of the two anatomically-informed expedients are used. Anat: Anatomically-Informed.

The interpretability analysis run with LIME showed that words like ‘progression’, ‘augmentation’ and ‘diminution’ that all indicate some sort of change were recurrent for predicting true positive samples. Similarly, words like ‘no’, ‘stability’ and ‘unchanged’ are predominant when predicting true negative reports. However, the error analysis also highlighted some recurrent mistakes, such as the importance given to the words ‘t2 ’ and ‘axial’ which are not directly linked to the task at hand. Overall, we deemed the high AUC attained with the TF-IDF pipeline to be sufficient for generating report-based weak labels for a subsequent study, which led us to the last manuscript of the thesis [36].

### 2.2.2 Image change detection with transfer learning and weak labels

In [36], the weak labels allowed us to increase the amount of labeled T2-weighted difference maps more than 3-fold. As we can observe in Table 2.2, this increase in dataset size permitted to raise classification performances both for the VGG and the SEResNeXt models on the in-house dataset. In fact, AUC and AUPR of the VGG-TL and SEResNeXt-TL experiments are higher with respect

## 2.2 Glioma Change Detection

Table 2.1: Report classification results across the 10 random runs. Values are presented as mean  $\pm$  standard deviation. Bold values indicate the highest performances. Acc = accuracy; Sens = sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value; F1 = F1-score; AUC = area under the ROC curve; AUPR = area under the PR curve.

Embedding	Acc %	Sens %	Spec %	PPV %	NPV %	F1 %	AUC	AUPR
<b>TF-IDF</b>	<b>88</b> $\pm$ 1	91 $\pm$ 1	<b>75</b> $\pm$ 0	<b>95</b> $\pm$ 0	<b>60</b> $\pm$ 2	<b>93</b> $\pm$ 0	<b>.89</b> $\pm$ .01	<b>.97</b> $\pm$ .00
<b>Doc2Vec</b>	86 $\pm$ 2	<b>94</b> $\pm$ 3	38 $\pm$ 10	89 $\pm$ 1	57 $\pm$ 10	92 $\pm$ 1	.83 $\pm$ .05	.96 $\pm$ .01

to VGG-Baseline and the SEResNeXt-Baseline, respectively. However, this raise was only significant for the VGG model (AUC permutation test,  $p=0.05$ ). In addition, we also found that model capacity is negligible for the task at hand: when comparing the VGG-Baseline with the SEResNeXt-Baseline experiment we found no significant difference (AUC permutation test,  $p=0.17$ ) and similarly we found no difference when comparing the VGG-TL with the SEResNeXt-TL experiment (AUC permutation test,  $p=0.39$ ). Overall, these results suggest that the VGG is preferable for the task at hand because it is simpler and more computationally efficient (having  $\approx 2.5X$  fewer parameters). When checking the most frequent hyperparameters chosen from Optuna, we found a peculiar pattern for the TL experiments: the hyperparameter *mixed\_training* was always True. This means that training from scratch with a mixed dataset (WAD + training HAD) consistently leads to higher performances with respect to either fine-tuning or feature extraction. Regarding classification performances on the external BraTS dataset, the SEResNeXt-Baseline model showed the highest AUC, though it did not significantly outperform the SEResNeXt-TL ( $p=0.46$ ), or the VGG-Baseline ( $p=0.39$ ).

Table 2.2: Classification test results. Upper part: in-house dataset. Lower part: BraTS-2015 dataset. **Bold** values indicate the highest performances. N=# of difference maps; Baseline=pipeline where only HAD data is used. TL=Transfer Learning: pipeline where both HAD and WAD are used. ACC=accuracy; SENS=sensitivity; SPEC=specificity; F1=F1 score; AUC=Area Under ROC Curve; AUPR=Area Under Precision-Recall Curve; PARAMS=# of parameters in the model.

Dataset	N	Model	Acc	Sens	Spec	F1	AUC	AUPR	Params
In-house	233	VGG-Baseline	70	55	77	54	.74	.55	7.5 M
		VGG-TL	<b>79</b>	<b>80</b>	79	<b>71</b>	.82	.72	
		SEResNeXt-Baseline	76	50	<b>88</b>	57	.79	.63	19.4 M
		SEResNeXt-TL	77	78	76	68	<b>.83</b>	<b>.73</b>	
BraTS 2015	51	VGG-Baseline (inference)	75	82	50	83	.66	.90	7.5 M
		VGG-TL (inference)	76	92	25	86	.59	.89	
		SEResNeXt-Baseline (inference)	73	69	<b>83</b>	79	<b>.76</b>	<b>.93</b>	19.4 M
		SEResNeXt-TL (inference)	<b>78</b>	<b>95</b>	25	<b>87</b>	.60	.60	



# Chapter 3

## Discussion

In this last chapter, we will contextualize and discuss the results presented above. Then, we will provide an overview of future steps for both the aneurysm and the glioma project. Last, we will conclude the thesis with some final thoughts and remarks regarding the impact of the work.

### 3.1 Main contributions

The goal of this PhD thesis was to investigate the use of DL models for two routine tasks conducted in radiology departments: aneurysm detection on TOF-MRA scans, and longitudinal monitoring of patients with high-grade gliomas on T2-weighted MR scans. For both projects, we tried to tackle some of the recurrent limitations that are associated to any DL-based pipeline in medical imaging. Specifically, the main contributions of this PhD thesis are the following:

1. The use of weak labels to mitigate the manual annotation bottleneck
2. The use and automation of transfer learning to alleviate domain shift and data scarcity
3. The use of prior anatomical knowledge to reduce model opaqueness
4. The open release of our two in-house datasets and open-source models

In the following paragraphs, we will further discuss these contributions and highlight their impact in the medical imaging community. The first research focus of this PhD thesis was the use of **weak labels**. Although several initiatives have been put in place to facilitate the sharing of data across research groups (e.g. The Cancer Imaging Archive [170], re3data [171], Grand Challenges in

### 3.1 Main contributions

---

Biomedical Image Analysis [172], etc.), the need for large, multicentric and heterogeneous labeled data will likely remain the major challenge for the deployment of robust DL-based systems in coming years [173]. The main reason for this is that small sample sizes and the lack of diverse geographic areas hinder model generalization, and DL models need to be continuously updated to cope with *concept drift* [174]. The weak labeling strategy presented in this thesis aims at mitigating this annotation bottleneck of medical DL and thus at increasing the amount of labeled data. In the aneurysm project, we leveraged oversized weak labels which allowed us to reduce the annotation time fourfold with respect to voxel-wise labels, while still maintaining competitive detection performances (ablation experiments, Table 4 of [35]). If reasoning in terms of larger datasets (e.g., thousands of patients), the proposed annotation process is a scalable and time-saving solution which can significantly alleviate the annotation bottleneck. Similarly, in the glioma project, we generated weak labels from textual radiology reports using an NLP classifier. Unstructured, semi-structured and structured radiology reports represent an underexploited resource for numerous applications in medical informatics [175], such as disease information and classification, diagnostic surveillance, quality compliance, cohort creation, and source of weak labels for downstream imaging tasks [176]. Despite being more general than image labels, weak labels extracted from radiology reports hold great potential for mitigating the annotation bottleneck. The main advantage of NLP-generated weak labels is that report classifiers are normally fast to train as compared to image-based models. (e.g. the NLP classifier presented in [47] takes  $\approx 10$  minutes, while even the small-capacity VGG network in [36] takes  $\approx 6$  hours). Moreover, even before the actual model training, also the creation of report labels is typically faster than the creation of image labels (e.g. identifying “tumor progression” in a report is faster than generating a voxel-wise manual mask of the progressing parts of the tumor). However, neither in [47] nor in [36] we conducted a proper timing for the generation of report labels, plus the difference in generation time might be task dependent. Nonetheless, once we have a trained report classifier, labeling hundreds (or even thousands) of new samples becomes extremely fast and inexpensive. On top of this, radiology reports have the advantage of being less computationally cumbersome in terms of storage and data transfer which can have a drastic impact in large, multicentric studies. In the work [47], we explored the use of NLP for classifying French radiology reports of patients with high-grade gliomas with two embedding strategies. As pointed out in [176], and subsequently shown in other works [177, 178], classical ML embedding techniques can lead to comparable results with respect to DL techniques when properly tuned. Moreover, they are still frequent when the dataset size is limited such as in medical imaging applications. Our work confirms this trend since, given the same classifier (Random Forest), the TF-IDF pipeline statistically outperformed the Doc2Vec one.

### 3.1 Main contributions

---

In [36], the weak annotation pipeline devised in [47] allowed us to obtain a more than 3-fold increase in sample size (233 difference maps for the Human Annotated Dataset vs. 795 for the TL pipeline with Weakly-Annotated Data  $> 0.75$ ) at very little added cost. Results in section 2.2.2 showed that the automatically labeled dataset WAD helped improving classification results both for the VGG and the SEResNeXt model between Baseline and TL experiments, although the difference was only significant for the VGG model. Nevertheless, as similarly reported in [179], we expect performances of both models to increase even further as more weakly-labeled samples are added.

The second contribution of this thesis relates to the **use of Transfer Learning (TL) to mitigate domain shift and data scarcity**. Although TL in medical imaging applications is not novel [90, 91, 180–182], the parallelism between TL in computer vision (CV) tasks and clinical tasks remains underexplored [90]. For instance, it is unclear whether the direct correlation between larger model capacity and increase in classification/segmentation performances usually witnessed in CV holds true for medical DL models that typically operate in lower data regimes (hundreds, rarely thousands of samples). Furthermore, the optimal type of TL (e.g. fine-tuning, feature extracting, mixed training, etc.) also remains to be discovered. In the work [36], we tried to shed light on these unanswered questions by exploring the impact that model capacity can have on medical TL, and by automating the choice of TL type by framing the problem as a hyperparameter optimization task. When studying the impact that model size can have on classification performances, we found no significant difference between VGG and SEResNeXt neither for the Baseline nor for the TL experiment. Therefore, for our application, we conclude that the VGG model is preferable because it is simpler and faster to train. A similar result was found in [183] where a VGG19 model outperformed much deeper networks in a TL pipeline for COVID-19 detection. Also, we found that the TL experiment with weakly annotated data led to significantly higher performances with respect to the Baseline experiment only for the VGG. This result differs from the similar study [90] since in the small data regime we found the smaller network (VGG) to benefit more from TL with respect to the larger SEResNeXt. In general, both our results and the ones in [183] show that the high-capacity networks and TL strategies typically used for CV tasks in the high-data regime are not necessarily optimal for medical imaging tasks, where models often operate in the low-data regime. Given that DL scaling studies typically show log-linear or power laws [184, 185] relating loss to dataset size, including for TL [186], it is possible that the higher-capacity SEResNeXt model in our study would be superior if much more data was available. However, this is not visible with our small dataset as we are far from the performance asymptote. Regarding the automation of the TL pipeline, instead of searching for the best TL type manually (which is the standard approach in similar studies [83]), we

### 3.1 Main contributions

---

framed the TL experiments as a hyperparameter optimization problem. Because the optimal value of other hyperparameters (such as the learning rate) depends on the TL type, our approach avoids the arbitrary choice of a TL type which can be potentially suboptimal. We believe that our pipeline can be adopted by similar works that aim to automate TL for image classification. Surprisingly, we found that mixed training TL led to higher classification performances with respect to fine-tuning or feature extracting. From a computational and environmental point of view, this finding is alarming because it indicates that the longest-running, least resource-efficient TL pipeline could be preferable, at least for longitudinal monitoring of high-grade gliomas. However, further investigations on similar tasks are needed to assess if this trend is isolated or recurrent.

The third contribution of the thesis is the use of **prior anatomical knowledge** to improve detection results and increase model interpretability. In the work [35], our model leverages the underlying anatomy of the brain vasculature (i.e., we “*anatomically-informed*” our network) in order to simulate the radiologists’ exploration of the TOF-MRA scans. First, most of the negative patches (i.e. patches without aneurysms) extracted during training either contained a vessel or were located in correspondence with the aneurysm landmark points. Second, we limited the sliding window approach only to regions of the brain that are plausible for aneurysm occurrence. These constraints aimed at mimicking the radiologists’ behavior in the sense that only regions containing vessels, or at higher risk for aneurysms are scanned, while the rest of the brain is neglected. Regarding model interpretability, narrowing the analysis to these anatomically plausible areas makes the model more easily explainable to clinicians because, for instance, we avoid false positive predictions in areas of the brain that are too peripheral and therefore unrealistic. The results in section 2.1 showed that the anatomically-informed sliding window is an effective expedient since it increases sensitivity, while reducing the average FP rate. Instead, the anatomically-informed patch sampling proved to be negligible when combined with the anatomically-informed sliding-window, or even detrimental when the sliding window was anatomically-agnostic. We hypothesize that applying only the anatomically-informed patch sampling leads to a domain shift issue: specifically, the model is trained using intensity-matched patches, but then is tested with any patch in the brain (because there is no anatomically-informed sliding window). We think this difference between training and test domain is what causes the decrease in performances. Nevertheless, the anatomically-informed sliding window expedient suggests that injecting prior anatomical knowledge in the pipeline can improve detection performances. We believe this general principle is also applicable to other pathologies with sparse spatial extent.

The last noteworthy contribution of this PhD thesis, both for the aneurysm [35] and the glioma project, [36] is the **open release of our in-house datasets**. The state of the art for automated

## 3.2 Limitations and future steps

---

aneurysm detection methods clearly lacks multi-site validation which is paramount if we plan to safely applying these CADe tools during routine clinical practice. Although [117, 118] did publish results obtained from multiple institutions, none of them released their dataset publicly which makes comparisons between algorithms unfeasible. The comparisons between methods are further hindered by the use of non-standardized evaluation metrics (e.g. FROC/lesion-wise sensitivity/subject-wise specificity) or by the fact that not all related studies include both patients (subjects with aneurysms) and controls (subjects without aneurysms). By openly releasing our dataset, we aim to bridge the data availability gap and foster reproducibility in the medical imaging analysis community. The combination of our in-house dataset and the ADAM dataset will allow researchers to assess the realistic robustness of proposed algorithms on heterogeneous data generated from different scanners, acquisition protocols and study population. Moreover, the availability of both datasets (in-house and ADAM) will allow researchers to try different supervised domain adaptation techniques and uncover which is the most effective for the task at hand. For the glioma change detection project, the release of our in-house dataset is arguably even more significant since, to best of our knowledge, ours is the first longitudinal labeled dataset available in the community for monitoring high-grade gliomas. Also for this project, the availability of both our in-house dataset and the labeled BraTS-2015 dataset (that we released) will allow to investigate domain adaptation, even though the amount of labeled BraTS data is limited (N=51 session pairs). We believe both our in-house datasets (aneurysm: [35], glioma: [187]) will foster reproducibility in the community and allow a more rigorous benchmarking for automated DL models.

## 3.2 Limitations and future steps

In this section, we present the major limitations of the developed DL models, and we point to future steps that can be undertaken for overcoming such limitations.

**Aneurysm Detection** - Focusing on [35], even combining our in-house dataset with the ADAM dataset, the number of subjects is still limited when compared to some related TOF-MRA [117, 118] or CTA [188, 189] studies. Also, we acknowledge that the number of patients for whom we compared the different annotations schemes (i.e., weak vs. voxel-wise) is limited (N =38); it is possible that statistically significant performance differences could be found with a larger sample size. Furthermore, we have to further increase detection performances if we plan to deploy our model as a second reader for radiologists, especially to detect tiny aneurysms or aneurysms in rare locations which are more frequently overlooked [114]. Although our top-performing model reached a sensitivity

### 3.2 Limitations and future steps

---

of 83%, this value might not be high enough, especially when assisting senior radiologists. At the same time, there should not be too many false positive predictions per subject (ideally not more than 2 or 3, on average), otherwise reading time might become prohibitively long.

Finally, ablation experiments have shown that pre-training the model on the ADAM dataset did not increase detection performances, thus different transfer learning techniques should be explored.

In the paper under preparation described in section 1.3.5, we are planning to overcome some of these limitations. First, we will retrieve 81 new subjects (65 controls, 16 patients) to re-train the model presented in [35] and further improve detection results. These 81 subjects are distinct from the 140 that will be later used for the within-subject reading. In addition, we will compare different TL techniques (fine-tuning, feature extraction, mixed training) to better exploit knowledge acquired from the ADAM dataset on the target in-house dataset. This supervised TL scenario is similar to the one faced in [36], though this case would correspond to a *transductive* TL scenario [80] where  $D_s$  (ADAM dataset)  $\neq D_t$  (in-house dataset). Last, we will use the top-performing configuration discovered in [35] (i.e. no anatomically informed patch sampling, but anatomically informed sliding window) to run inference on the new 140 subjects. This work will represent the ultimate step of the CAde development and will help us understand the real practical value of our tool in a radiological setting in terms of reading time, added clinical value, and acceptance by the readers.

Beyond the planned work, additional steps that can be explored to improve detection performances in the future might include for instance trying different architectures to segment (e.g. V-Net [190], UNETR [191], Swin UNETR [192], nnUnet [193]) or detect (e.g. nnDetection [121]) aneurysms, ideally combining these models with our anatomically-driven expedients. Also, one might consider using a multi-scale approach with patches of larger (or smaller) size. In addition, it would be useful to conduct further error analyses (as the one shown in Figure 3.1) to identify common patterns for both false positive and false negative predictions.

**Glioma Change Detection** - In this clinical application, the first limitation of our approach [47] is that the report annotations were performed by one single radiologist which is not the optimal scenario for ambiguous NLP tasks. A second major limitation of the glioma project (both in [47] and in [36]) is that we narrowed the classification problem to a binary scenario in which we only distinguished stable vs. unstable tumor, mainly because we did not have enough cases of tumor response in our cohort. This is a simplification because *progression* and *response* are distinct clinical indicators. Another limitation of [36] is that we only focused on T2w MRI volumes, even though a multi-modal assessment of glioma evolution would be more accurate [146]. A fourth limitation is that the reports from the HAD for which the two annotators disagreed were discarded. Additionally,

### 3.2 Limitations and future steps

---

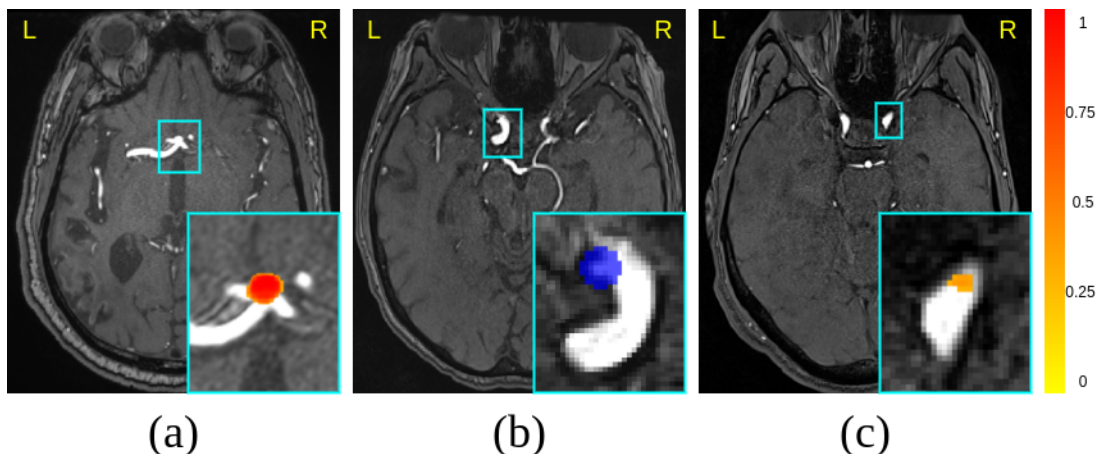


Figure 3.1: Qualitative analysis of predictions and errors. The heatmap generated by the network ranges from 0 (low probability, yellow color) to 1 (high probability, red color) (a) True positive prediction in the anterior communicating artery. (b) False negative in the internal carotid artery. The ground truth label mask is shown in blue. (c) False positive prediction in the internal carotid artery.

we only evaluated one approach for fine-tuning (i.e. all the layers of the networks were re-trained).

Future works should aim at extracting new patients and adapt the classification towards a 3-class (stable, progression, response) or even 5-class (stable, progression, response, pseudoprogression, pseudoresponse) problem. Ideally, the reports linked to these new patients should be annotated by more than one radiologist (the more, the better) in order to have more robust ground truth labels after a consensus has been reached for discordant examples. As shown in section 1.4.2, adding more classes would require the use of additional sequences (e.g. enhanced T1w, FLAIR, etc.) on top of the T2w scans. If moving toward this more granular scenario, researchers should be wary because, as previously shown in [194], results might change significantly. Another approach to increase performances could be to change the layers that are frozen/re-trained during fine-tuning and the number of epochs during which they are re-trained, as in [89]. Eventually, the final goal of the glioma change detection project would also be a clinical assessment of the utility of the CAdE system. Once a multi-modality approach will be put in place, with satisfactory classification performances, a prospective trial should be run to understand whether the changes highlighted in the new scan are relevant for the radiologists and actually facilitate and speed up diagnosis and subsequent report writing.

## 3.3 Conclusion

Estimates of the WHO indicate that the proportion of the world’s population over 60 years of age will be 22% by 2050, nearly double that of 2015 [195]. And “the older population requires more imaging”, said Dr. Harprit Bedi, vice chairman of radiology education at Boston University School of Medicine. This aging trend is worrying because the number of practicing radiologists will likely be insufficient to meet the growing demand for imaging care [196, 197]. Some studies have reported that, in some extreme cases, an average reader should interpret one image every 3–4 seconds in an 8-hour shift to meet workload demands [198]. This mismatch between available radiologists and disproportionate grow of imaging data is projected to have dire consequences for patients who will experience ever longer waiting lists, and for radiologists who will have to cope with increasing backlog of examinations.

One possible solution to face this looming scenario is the integration of CAD systems and machine learning algorithms into the radiological workflow, with the intent of increasing image reading throughput while preserving high diagnostic accuracy. Tremendous progress has been made in the field of CAD systems since their first introduction in the 1980s, with the latest wave, the one based on deep neural networks, showing impressive performances across a rising number of specific, radiological tasks [199]. As a matter of fact, DL-based algorithms excel at automatically recognizing complex patterns in imaging data and providing quantitative, rather than qualitative, information about radiographic characteristics [200]. According to [200], the three radiological tasks in which DL-based CAD systems will likely have a large impact are anomaly detection (as in the case of aneurysm detection presented in this thesis), subsequent characterization of objects of interest via segmentation, diagnosis and staging, and finally the longitudinal monitoring of objects for diagnosis and assessment of treatment response (as for the monitoring of high-grade gliomas described in previous sections).

Despite all the excitement about the added value that AI will bring to radiology, there is still a great debate regarding the speed with which novel DL models will be implemented in clinical practice [201]. As we have seen throughout this PhD thesis, there are still some undeniable limitations linked to DL-based systems that need to be overcome, or at least mitigated, before these tools become pervasive and routinely adopted in the clinics.

The first limitation that we addressed was the lack of large annotated medical dataset. This data scarcity is considered one of the biggest obstacles for reaching human-level clinical performances (or higher), because models trained with too few samples tend to overfit to the training data and lose the



### 3.3 Conclusion

---

ability to generalize [202]. This issue is even more pronounced for rare diseases, where the retrieval of large cohorts is extremely complicated. Our proposed solutions to mitigate this phenomenon were the use of time-saving weak labels to speed up the collection of annotated data and the open release of both our in-house datasets.

The second limitation that we addressed was domain shift, a recurrent phenomenon that occurs when the unseen test samples have a different feature distribution with respect to the training samples. To compensate for domain shift, in the glioma project we investigated the use of transfer learning, and in particular the automation of TL types to avoid empirical choices. To further validate this approach, we are planning to experiment it also in the clinical paper described in section 1.3.5.

A third limitation of DL models is the lack of interpretability. Even though there is a growing trend in the ML community towards open-sourcing data and code, a strong theoretical understanding of deep learning still needs to be established [203]. This lack of understanding complicates failure prediction and makes it hard to isolate the logic behind a specific conclusion drawn by the model. However, although model interpretability is paramount and will need extensive further investigation, it has also been pointed out that numerous safe and effective Food and Drug Administration (FDA)-approved drugs also have unknown mechanisms of action [204, 205], which opens the floor to a more general discussion that tries to answer the question: “up to which level do we want our model to be explainable?”. In this PhD thesis, we tried to alleviate model opaqueness using prior anatomical knowledge. Specifically, in [35] we constrained the analysis only to the areas of the brain that are plausible for aneurysm occurrence. Although this approach is currently not the most widespread in the literature, it can help to simulate the physicians’ exploration of medical images and reduce unexpected behaviors of the model.

A fourth limitation that was not addressed in this work but that is worth mentioning is the inability of DL models to address more than one task (a quality referred to as *narrow intelligence*, or *specific-purpose intelligence*). A comprehensive (or *general-purpose*) DL system capable of performing multiple tasks such as detecting different anomalies within the entire human body is yet to be developed [200].

In summary, this PhD work explored several expedients that aim at mitigating intrinsic limitations of DL-based CAD systems in radiology. Although we narrowed our analysis only to two specific tasks (aneurysm detection and glioma change detection), we believe our contributions can help researchers who are facing similar tasks (anomaly/change detection) to overcome such limitations and bring their models closer to clinical application.

# References

- [1] Maximilian E Tschuchnig and Michael Gadermayr. Anomaly detection in medical imaging-a mini review. *Data Science–Analytics and Applications*, pages 33–38, 2022.
- [2] Maryellen L Giger. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3):512–520, 2018.
- [3] Ronald A Castellino. Computer aided detection (cad): an overview. *Cancer Imaging*, 5(1):17, 2005.
- [4] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.
- [5] Kenji Suzuki. A review of computer-aided diagnosis in thoracic and colonic imaging. *Quantitative imaging in medicine and surgery*, 2(3):163, 2012.
- [6] Kunio Doi. Current status and future potential of computer-aided diagnosis in medical imaging. *The British journal of radiology*, 78(suppl\_1):s3–s19, 2005.
- [7] Heang-Ping Chan, Kunio Doi, Simranjit Galhotra, Carl J Vyborny, Heber MacMahon, and Peter M Jokich. Image feature analysis and computer-aided diagnosis in digital radiography. i. automated detection of microcalcifications in mammography. *Medical physics*, 14(4):538–548, 1987.
- [8] Rangaraj M Rangayyan, Fabio J Ayres, and JE Leo Desautels. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of the Franklin Institute*, 344(3-4):312–348, 2007.
- [9] Ayman El-Baz, Garth M Beache, Georgy Gimel’farb, Kenji Suzuki, Kazunori Okada, Ahmed Elnakib, Ahmed Soliman, and Behnoush Abdollahi. Computer-aided diagnosis systems for

## REFERENCES

---

- lung cancer: challenges and methodologies. *International journal of biomedical imaging*, 2013, 2013.
- [10] Hiroyuki Yoshida, Yoshitaka Masutani, Peter Maceneaney, David T Rubin, and Abraham H Dachman. Computerized detection of colonic polyps at ct colonography on the basis of volumetric features: pilot study. *Radiology*, 222:327–336, 2002.
- [11] Bradley J Erickson and Brian Bartholmai. Computer-aided detection and diagnosis at the start of the third millennium. *Journal of digital imaging*, 15(2):59–68, 2002.
- [12] Yudong Zhang, Zhengchao Dong, Lenan Wu, and Shuihua Wang. A hybrid method for mri brain image classification. *Expert Systems with Applications*, 38(8):10049–10053, 2011.
- [13] Laurence AG Marshman, Peter J Ward, Paul H Walter, and Robert S Dossetor. The progression of an infundibulum to aneurysm formation and rupture: case report and literature review. *Neurosurgery*, 43(6):1445–1448, 1998.
- [14] Jorge Hernández Rodríguez, Francisco Javier Cabrero Fraile, María José Rodríguez Conde, and Pablo Luis Gómez Llorente. Computer aided detection and diagnosis in medical imaging: a review of clinical and educational applications. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, pages 517–524, 2016.
- [15] Monique D Dorrius, Marijke C der Weide, Peter van Ooijen, Ruud M Pijnappel, and Matthijs Oudkerk. Computer-aided detection in breast mri: a systematic review and meta-analysis. *European radiology*, 21(8):1600–1608, 2011.
- [16] Meredith Noble, Wendy Bruening, Stacey Uhl, and Karen Schoelles. Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. *Archives of gynecology and obstetrics*, 279(6):881–890, 2009.
- [17] Jinshan Tang, Rangaraj M Rangayyan, Jun Xu, Issam El Naqa, and Yongyi Yang. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE transactions on information technology in biomedicine*, 13(2):236–251, 2009.
- [18] Bram Van Ginneken, BM Ter Haar Romeny, and Max A Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on medical imaging*, 20(12):1228–1241, 2001.

## REFERENCES

---

- [19] Shigehiko Katsuragawa and Kunio Doi. Computer-aided diagnosis in chest radiography. *Computerized Medical Imaging and Graphics*, 31(4-5):212–223, 2007.
- [20] Ren Yuan, Patrick M Vos, and Peter L Cooperberg. Computer-aided detection in screening ct for pulmonary nodules. *American Journal of Roentgenology*, 186(5):1280–1287, 2006.
- [21] Hiroyuki Yoshida and Abraham H Dachman. Computer-aided diagnosis for ct colonography. In *Seminars in Ultrasound, CT and MRI*, volume 25, pages 419–431. Elsevier, 2004.
- [22] Didier Bielen and Gabriel Kiss. Computer-aided detection for ct colonography: update 2007. *Abdominal Imaging*, 32(5):571–581, 2007.
- [23] A El-Sayed, Heba M Mohsen, Kenneth Revett, and Abdel-Badeeh M Salem. Computer-aided diagnosis of human brain tumor through mri: A survey and a new algorithm. *Expert systems with Applications*, 41(11):5526–5545, 2014.
- [24] Esther E Bron, Marion Smits, Wiesje M Van Der Flier, Hugo Vrenken, Frederik Barkhof, Philip Scheltens, Janne M Papma, Rebecca ME Steketee, Carolina Méndez Orellana, Rozanna Meijboom, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: the caddementia challenge. *NeuroImage*, 111:562–579, 2015.
- [25] Hidetaka Arimura, Qiang Li, Yukunori Korogi, Toshinori Hirai, Hiroyuki Abe, Yasuyuki Yamashita, Shigehiko Katsuragawa, Ryuji Ikeda, and Kunio Doi. Automated computerized scheme for detection of unruptured intracranial aneurysms in three-dimensional magnetic resonance angiography<sup>1</sup>. *Academic radiology*, 11(10):1093–1104, 2004.
- [26] Xiaojiang Yang, Daniel J Blezek, Lionel TE Cheng, William J Ryan, David F Kallmes, and Bradley J Erickson. Computer-aided detection of intracranial aneurysms in mr angiography. *Journal of digital imaging*, 24(1):86–95, 2011.
- [27] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A Eschrich, Matthew B Schabath, Kenneth Forster, Hugo JWL Aerts, Andre Dekker, David Fenstermacher, et al. Radiomics: the process and the challenges. *Magnetic resonance imaging*, 30(9):1234–1248, 2012.
- [28] Marius E Mayerhoefer, Andrzej Materka, Georg Langs, Ida Häggström, Piotr Szczypiński, Peter Gibbs, and Gary Cook. Introduction to radiomics. *Journal of Nuclear Medicine*, 61(4):488–495, 2020.

## REFERENCES

---

- [29] Stephen SF Yip and Hugo JWL Aerts. Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61(13):R150, 2016.
- [30] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563, 2016.
- [31] Alex Zwanenburg, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020.
- [32] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.
- [33] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227, 2020.
- [34] Katie Chockley and Ezekiel Emanuel. The end of radiology? three threats to the future practice of radiology. *Journal of the American College of Radiology*, 13(12):1415–1420, 2016.
- [35] Tommaso Di Noto, Guillaume Marie, Sebastien Tourbier, Yasser Alemán-Gómez, Oscar Esteban, Guillaume Saliou, Meritxell Bach Cuadra, Patric Hagmann, and Jonas Richiardi. Towards automated brain aneurysm detection in tof-mra: open data, weak labels, and anatomical knowledge. *Neuroinformatics*, pages 1–14, 2022.
- [36] Tommaso Di Noto, Meritxell Bach Cuadra, Chirine Atat, Eduardo Gamito Teiga, Monika Hegi, Andreas Hottinger, Patric Hagmann, and Jonas Richiardi. Transfer learning with weak labels from radiology reports: application to glioma change detection. *arXiv preprint arXiv:2210.09698*, 2022.
- [37] Shahira Abousamra, Danielle Fassler, Le Hou, Yuwei Zhang, Rajarsi Gupta, Tahsin Kurc, Luisa F Escobar-Hoyos, Dimitris Samaras, Beatrice Knudson, Kenneth Shroyer, et al. Weakly-supervised deep stain decomposition for multiplex ihc images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 481–485. IEEE, 2020.

## REFERENCES

---

- [38] Matvey Ezhov, Adel Zakirov, and Maxim Gusarev. Coarse-to-fine volumetric segmentation of teeth in cone-beam ct. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 52–56. IEEE, 2019.
- [39] Rihuan Ke, Aurélie Bugeau, Nicolas Papadakis, Peter Schuetz, and Carola-Bibiane Schönlieb. Learning to segment microscopy images with lazy labels. In *European Conference on Computer Vision*, pages 411–428. Springer, 2020.
- [40] Niccolò Marini, Stefano Marchesin, Sebastian Otálora, Marek Wodzinski, Alessandro Caputo, Mart van Rijnthoven, Witali Aswolinskiy, John-Melle Bokhorst, Damian Podareanu, Edyta Petters, et al. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. *NPJ digital medicine*, 5(1):1–18, 2022.
- [41] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.
- [42] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.
- [43] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [44] Daniel T Huff, Amy J Weisman, and Robert Jeraj. Interpretation and visualization techniques for deep learning models in medical imaging. *Physics in Medicine & Biology*, 66(4):04TR01, 2021.
- [45] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.
- [46] Mara Graziani, Iam Palatnik de Sousa, Marley MBR Vellasco, Eduardo Costa da Silva, Henning Müller, and Vincent Andrearczyk. Sharpening local interpretable model-agnostic explanations for histopathology: Improved understandability and reliability. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–549. Springer, 2021.

## REFERENCES

---

- [47] Tommaso Di Noto, Chirine Atat, Eduardo Gamito Teiga, Monika Hegi, Andreas Hottinger, Meritxell Bach Cuadra, Patric Hagmann, and Jonas Richiardi. Diagnostic surveillance of high-grade gliomas: towards automated change detection using radiology report classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 423–436. Springer, 2021.
- [48] Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.
- [49] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [50] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.
- [51] Guoqiang Peter Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):451–462, 2000.
- [52] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.
- [53] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [54] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.
- [55] Pramila P Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (IC3UBEA)*, pages 1–6. IEEE, 2018.
- [56] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74, 2021.

## REFERENCES

---

- [57] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.
- [58] Mingyu Kim, Jihye Yun, Yongwon Cho, Keewon Shin, Ryoungwoo Jang, Hyun-jin Bae, and Namkug Kim. Deep learning in medical imaging. *Neurospine*, 16(4):657, 2019.
- [59] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning. adaptive computation and machine learning series, 2006.
- [60] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.
- [61] Kenji Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017.
- [62] Victor S Sheng and Jing Zhang. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9837–9843, 2019.
- [63] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.
- [64] G Chowdhury Gobinda. Natural language processing. *Annual Review of Information Science and Technology*, 37:51–89, 2003.
- [65] Lawrence H Schwartz, David M Panicek, Alexandra R Berk, Yuelin Li, and Hedvig Hricak. Improving communication of diagnostic radiology findings through structured reporting. *Radiology*, 260(1):174, 2011.
- [66] Po-Hao Chen, Hanna Zafar, Maya Galperin-Aizenberg, and Tessa Cook. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *Journal of digital imaging*, 31(2):178–184, 2018.



## REFERENCES

---

- [67] Kenneth L Kehl, Haitham Elmarakeby, Mizuki Nishino, Eliezer M Van Allen, Eva M Lepisto, Michael J Hassett, Bruce E Johnson, and Deborah Schrag. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA oncology*, 5(10):1421–1429, 2019.
- [68] Saeed Hassanpour, Graham Bay, and Curtis P Langlotz. Characterization of change and significance for clinical findings in radiology reports through natural language processing. *Journal of digital imaging*, 30(3):314–322, 2017.
- [69] Selen Bozkurt, Emel Alkim, Imon Banerjee, and Daniel L Rubin. Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *Journal of digital imaging*, 32(4):544–553, 2019.
- [70] Anne-Dominique Pham, Aurélie Névéol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello, and Anita Burgun. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC bioinformatics*, 15(1):1–10, 2014.
- [71] Imon Banerjee, Yuan Ling, Matthew C Chen, Sadid A Hasan, Curtis P Langlotz, Nathaniel Moradzadeh, Brian Chapman, Timothy Amrhein, David Mong, Daniel L Rubin, et al. Comparative effectiveness of convolutional neural network (cnn) and recurrent neural network (rnn) architectures for radiology text report classification. *Artificial intelligence in medicine*, 97:79–88, 2019.
- [72] Matthew C Chen, Robyn L Ball, Lingyao Yang, Nathaniel Moradzadeh, Brian E Chapman, David B Larson, Curtis P Langlotz, Timothy J Amrhein, and Matthew P Lungren. Deep learning to classify radiology free-text reports. *Radiology*, 286(3):845–852, 2018.
- [73] Carlos R Oliveira, Patrick Niccolai, Anette Michelle Ortiz, Sangini S Sheth, Eugene D Shapiro, Linda M Niccolai, and Cynthia A Brandt. Natural language processing for surveillance of cervical and anal cancer and precancer: algorithm development and split-validation study. *JMIR medical informatics*, 8(11):e20826, 2020.
- [74] Ling Shao, Fan Zhu, and Xuelong Li. Transfer learning for visual categorization: A survey. *IEEE transactions on neural networks and learning systems*, 26(5):1019–1034, 2014.

## REFERENCES

---

- [75] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, 2020.
- [76] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.
- [77] Mehran Javanmardi and Tolga Tasdizen. Domain adaptation for biomedical image segmentation using adversarial training. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 554–558. IEEE, 2018.
- [78] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.
- [79] Martin J Willeminck, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.
- [80] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [81] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.
- [82] Padmavathi Kora, Chui Ping Ooi, Oliver Faust, U Raghavendra, Anjan Gudigar, Wai Yee Chan, K Meenakshi, K Swaraja, Pawel Plawiak, and U Rajendra Acharya. Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*, 2021.
- [83] Hee E Kim, Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):1–13, 2022.
- [84] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

## REFERENCES

---

- [85] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [86] Yiting Xie and David Richmond. Pre-training on grayscale imagenet improves medical image classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.
- [87] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [88] Vladimir Iglovikov and Alexey Shvets. Ternaunet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018.
- [89] Wan Hang Keith Chiu, Varut Vardhanabhuti, Dmytro Poplavskiy, Philip Leung Ho Yu, Richard Du, Alistair Yun Hee Yap, Sailong Zhang, Ambrose Ho-Tung Fong, Thomas Wing-Yan Chin, Jonan Chun Yin Lee, et al. Detection of covid-19 using deep learning algorithms on chest radiographs. *Journal of thoracic imaging*, 35(6):369–376, 2020.
- [90] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.
- [91] Laith Alzubaidi, Mohammed A Fadhel, Omran Al-Shamma, Jinglan Zhang, J Santamaría, Ye Duan, and Sameer R Oleiwi. Towards a better understanding of transfer learning for medical imaging: a case study. *Applied Sciences*, 10(13):4523, 2020.
- [92] Hong-Yu Zhou, Shuang Yu, Cheng Bian, Yifan Hu, Kai Ma, and Yefeng Zheng. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–407. Springer, 2020.
- [93] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image

## REFERENCES

---

- analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 3–13. Springer, 2021.
- [94] Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, 3(4):966–989, 2021.
- [95] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, et al. A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, pages 1–32, 2022.
- [96] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. Explainable software analytics. In *Proceedings of the 40th international conference on software engineering: New ideas and emerging results*, pages 53–56, 2018.
- [97] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020.
- [98] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [99] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.
- [100] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
- [101] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

## REFERENCES

---

- [102] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.
- [103] Mara Graziani, James M Brown, Vincent Andrearczyk, Veysi Yildiz, J Peter Campbell, Deniz Erdogmus, Stratis Ioannidis, Michael F Chiang, Jayashree Kalpathy-Cramer, and Henning Müller. Improved interpretability for computer-aided severity assessment of retinopathy of prematurity. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, pages 450–460. SPIE, 2019.
- [104] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.
- [105] Ke Yan, Youbao Tang, Adam P Harrison, Jinzheng Cai, Le Lu, and Jingjing Lu. Interpretable medical image classification with self-supervised anatomical embedding and prior knowledge. *OpenReview*, 2021.
- [106] Esther Puyol-Antón, Chen Chen, James R Clough, Bram Ruijsink, Baldeep S Sidhu, Justin Gould, Bradley Porter, Marc Elliott, Vishal Mehta, Daniel Rueckert, et al. Interpretable deep models for cardiac resynchronisation therapy response prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 284–293. Springer, 2020.
- [107] Xiaodan Chen, Yun Liu, Huazhang Tong, Yonghai Dong, Dongyang Ma, Lei Xu, and Cheng Yang. Meta-analysis of computed tomography angiography versus magnetic resonance angiography for intracranial aneurysm. *Medicine*, 97(20), 2018.
- [108] Monique HM Vlak, Ale Algra, Raya Brandenburg, and Gabriël JE Rinkel. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. *The Lancet Neurology*, 10(7):626–636, 2011.
- [109] Robert D Brown Jr and Joseph P Broderick. Unruptured intracranial aneurysms: epidemiology, natural history, management options, and familial screening. *The Lancet Neurology*, 13(4):393–404, 2014.
- [110] Alexis Hadjiathanasiou, Patrick Schuss, Simon Brandecker, Thomas Welchowski, Matthias Schmid, Hartmut Vatter, and Erdem Güresir. Multiple aneurysms in subarachnoid hemorrhage-

## REFERENCES

---

- identification of the ruptured aneurysm, when the bleeding pattern is not self-explanatory-development of a novel prediction score. *BMC neurology*, 20(1):1–12, 2020.
- [111] C Kouskouras, A Charitanti, C Giavroglou, N Foroglou, P Selviaridis, V Kontopoulos, and AS Dimitriadis. Intracranial aneurysms: evaluation using cta and mra. correlation with dsa and intraoperative findings. *Neuroradiology*, 46(10):842–850, 2004.
- [112] Jacoba P Greving, Marieke JH Wermer, Robert D Brown Jr, Akio Morita, Seppo Juvela, Masahiro Yonekura, Toshihiro Ishibashi, James C Torner, Takeo Nakayama, Gabriël JE Rinkel, et al. Development of the phases score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. *The Lancet Neurology*, 13(1):59–66, 2014.
- [113] Philip M White, Evelyn M Teasdale, Joanna M Wardlaw, and Valerie Easton. Intracranial aneurysms: Ct angiography and mr angiography for detection—prospective blinded comparison in a large patient cohort. *Radiology*, 219(3):739–749, 2001.
- [114] Alexander Keedy. An overview of intracranial aneurysms. *McGill Journal of Medicine: MJM*, 9(2):141, 2006.
- [115] Balaji Rao, Vahe Zohrabian, Paul Cedeno, Atin Saha, Jay Pahade, and Melissa A Davis. Utility of artificial intelligence tool as a prospective radiology peer reviewer—detection of unreported intracranial hemorrhage. *Academic radiology*, 28(1):85–93, 2021.
- [116] Yukihiro Nomura, Yoshitaka Masutani, Soichiro Miki, Mitsutaka Nemoto, Shouhei Hanaoka, Takeharu Yoshikawa, Naoto Hayashi, and Kuni Ohtomo. Performance improvement in computerized detection of cerebral aneurysms by retraining classifier using feedback data collected in routine reading environment. *Journal of Biomedical Graphics and Computing*, 4(4):12, 2014.
- [117] Daiju Ueda, Akira Yamamoto, Masataka Nishimori, Taro Shimono, Satoshi Doishita, Akitoshi Shimazaki, Yutaka Katayama, Shinya Fukumoto, Antoine Choppin, Yuki Shimahara, et al. Deep learning for mr angiography: automated detection of cerebral aneurysms. *Radiology*, 290(1):187–194, 2019.
- [118] Bio Joo, Sung Soo Ahn, Pyeong Ho Yoon, Sohi Bae, Beomseok Sohn, Yong Eun Lee, Jun Ho Bae, Moo Sung Park, Hyun Seok Choi, and Seung-Koo Lee. A deep learning algorithm may automate intracranial aneurysm detection on mr angiography with high diagnostic performance. *European Radiology*, 30(11):5785–5793, 2020.

## REFERENCES

---

- [119] Takahiro Nakao, Shouhei Hanaoka, Yukihiro Nomura, Issei Sato, Mitsutaka Nemoto, Soichiro Miki, Eriko Maeda, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe. Deep neural network-based computer-assisted detection of cerebral aneurysms in mr angiography. *Journal of Magnetic Resonance Imaging*, 47(4):948–953, 2018.
- [120] Joseph N Stember, Peter Chang, Danielle M Stember, Michael Liu, Jack Grinband, Christopher G Filippi, Philip Meyers, and Sachin Jambawalikar. Convolutional neural networks for the detection and measurement of cerebral aneurysms on magnetic resonance angiography. *Journal of digital imaging*, 32(5):808–815, 2019.
- [121] Michael Baumgartner, Paul F Jäger, Fabian Isensee, and Klaus H Maier-Hein. nndetection: A self-configuring method for medical object detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 530–539. Springer, 2021.
- [122] Kimberley M Timmins, Irene C van der Schaaf, Edwin Bennink, Ynte M Ruigrok, Xingle An, Michael Baumgartner, Pascal Bourdon, Riccardo De Feo, Tommaso Di Noto, Florian Dubost, et al. Comparing methods of detecting and segmenting unruptured intracranial aneurysms on tof-mras: The adam challenge. *Neuroimage*, 238:118216, 2021.
- [123] Krzysztof J Gorgolewski, Tibor Auer, Vince D Calhoun, R Cameron Craddock, Samir Das, Eugene P Duff, Guillaume Flandin, Satrajit S Ghosh, Tristan Glatard, Yaroslav O Halchenko, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, 3(1):1–9, 2016.
- [124] Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncalves, et al. Openneuro: An open resource for sharing of neuroimaging data. *BioRxiv*, 2021.
- [125] Paul A Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C Gee, and Guido Gerig. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.
- [126] Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

## REFERENCES

---

- [127] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.
- [128] Pauline Mouches and Nils D Forkert. A statistical atlas of cerebral arteries generated using multi-center mra datasets from healthy subjects. *Scientific data*, 6(1):1–8, 2019.
- [129] Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009.
- [130] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.
- [131] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [132] Dev P Chakraborty and Kevin S Berbaum. Observer studies involving detection and localization: modeling, analysis, and validation. *Medical physics*, 31(8):2313–2330, 2004.
- [133] Janice Ward, Katherine S Naik, J Ashley Guthrie, Daniel Wilson, and Philip J Robinson. Hepatic lesion detection: comparison of mr imaging after the administration of superparamagnetic iron oxide with dual-phase ct by using alternative-free response receiver operating characteristic analysis. *Radiology*, 210(2):459–466, 1999.
- [134] M McHugh. The chi-square test of independence. *biochemiamedica*. 2013;: 143–149.
- [135] Philip M White, Joanna M Wardlaw, and Valerie Easton. Can noninvasive imaging accurately depict intracranial aneurysms? a systematic review. *Radiology*, 217(2):361–370, 2000.
- [136] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.
- [137] Michael Weller, Wolfgang Wick, Ken Aldape, Michael Brada, Mitchell Berger, Stefan M Pfister, Ryo Nishikawa, Mark Rosenthal, Patrick Y Wen, Roger Stupp, et al. Glioma. *Nature reviews Disease primers*, 1(1):1–18, 2015.



## REFERENCES

---

- [138] Paul Kleihues, David N Louis, Bernd W Scheithauer, Lucy B Rorke, Guido Reifenberger, Peter C Burger, and Webster K Cavenee. The who classification of tumors of the nervous system. *Journal of Neuropathology & Experimental Neurology*, 61(3):215–225, 2002.
- [139] Gwenaëlle Marquet, Olivier Dameron, Stephan Saikali, Jean Mosser, and Anita Burgun. Grading glioma tumors using owl-dl and nci thesaurus. In *AMIA Annual Symposium Proceedings*, volume 2007, page 508. American Medical Informatics Association, 2007.
- [140] Melissa L Bondy, Michael E Scheurer, Beatrice Malmer, Jill S Barnholtz-Sloan, Faith G Davis, Dora Il’Yasova, Carol Kruchko, Bridget J McCarthy, Preetha Rajaraman, Judith A Schwartzbaum, et al. Brain tumor epidemiology: consensus from the brain tumor epidemiology consortium. *Cancer*, 113(S7):1953–1968, 2008.
- [141] Jennifer M Connelly and Mark G Malkin. Environmental risk factors for brain tumors. *Current neurology and neuroscience reports*, 7(3):208–214, 2007.
- [142] Quinn T Ostrom and Jill S Barnholtz-Sloan. Current state of our knowledge on brain tumor epidemiology. *Current neurology and neuroscience reports*, 11(3):329–335, 2011.
- [143] Hiroko Ohgaki, Young-Ho Kim, and Joachim P Steinbach. Nervous system tumors associated with familial tumor syndromes. *Current opinion in neurology*, 23(6):583–591, 2010.
- [144] David N Louis, Hiroko Ohgaki, Otmar D Wiestler, Webster K Cavenee, Peter C Burger, Anne Jouvot, Bernd W Scheithauer, and Paul Kleihues. The 2007 who classification of tumours of the central nervous system. *Acta neuropathologica*, 114(2):97–109, 2007.
- [145] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.
- [146] Bjoern Menze, Fabian Isensee, Roland Wiest, Bene Wiestler, Klaus Maier-Hein, Mauricio Reyes, and Spyridon Bakas. Analyzing magnetic resonance imaging data from glioma patients using deep learning. *Computerized medical imaging and graphics*, 88:101828, 2021.
- [147] Harpreet Hyare, Steffi Thust, and Jeremy Rees. Advanced mri techniques in the monitoring of treatment of gliomas. *Current treatment options in neurology*, 19(3):1–15, 2017.

## REFERENCES

---

- [148] Yannick Suter, Urspeter Knecht, Mariana Alão, Waldo Valenzuela, Ekkehard Hewer, Philippe Schucht, Roland Wiest, and Mauricio Reyes. Radiomics for glioblastoma survival analysis in pre-operative mri: exploring feature robustness, class boundaries, and machine learning techniques. *Cancer Imaging*, 20(1):1–13, 2020.
- [149] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.
- [150] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific data*, 4(1):1–8, 2017.
- [151] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.
- [152] Martin J van den Bent, Jeffrey S Wefel, David Schiff, Martin JB Taphoorn, Kurt Jaeckle, L Junck, Terri Armstrong, A Choucair, Adam D Waldman, Thierry Gorlia, et al. Response assessment in neuro-oncology (a report of the rano group): assessment of outcome in trials of diffuse low-grade gliomas. *The lancet oncology*, 12(6):583–593, 2011.
- [153] David R Macdonald, Terrance L Cascino, S Clifford Schold Jr, and J Gregory Cairncross. Response criteria for phase ii studies of supratentorial malignant glioma. *Journal of clinical oncology*, 8(7):1277–1280, 1990.
- [154] Michael A Vogelbaum, Sarah Jost, Manish K Aghi, Amy B Heimberger, John H Sampson, Patrick Y Wen, David R Macdonald, Martin J Van den Bent, and Susan M Chang. Application of novel response/progression measures for surgically delivered therapies for gliomas: Response assessment in neuro-oncology (rano) working group. *Neurosurgery*, 70(1):234–244, 2012.
- [155] Patrick Y Wen, David R Macdonald, David A Reardon, Timothy F Cloughesy, A Gregory Sorensen, Evanthia Galanis, John DeGroot, Wolfgang Wick, Mark R Gilbert, Andrew B Lassman, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *Journal of clinical oncology*, 28(11):1963–1972, 2010.

## REFERENCES

---

- [156] Benjamin M Ellingson, Patrick Y Wen, and Timothy F Cloughesy. Modified criteria for radiographic response assessment in glioblastoma clinical trials. *Neurotherapeutics*, 14(2):307–320, 2017.
- [157] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O’Reilly Media, Inc.”, 2009.
- [158] Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning.* Springer Science & Business Media, 2011.
- [159] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.
- [160] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [161] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.
- [162] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.
- [163] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.
- [164] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.
- [165] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [166] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [167] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

## REFERENCES

---

- [168] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.
- [169] Andriy I Bandos, Howard E Rockette, and David Gur. A permutation test sensitive to differences in areas for comparing roc curves from a paired design. *Statistics in medicine*, 24(18):2873–2893, 2005.
- [170] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.
- [171] Heinz Pampel, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf. Making research data repositories visible: the re3data. org registry. *PloS one*, 8(11):e78080, 2013.
- [172] Bram van Ginneken, Sjoerd Kerkstra, and James Meakin. Grand challenges in biomedical image analysis, 2018.
- [173] Caroline Bivik Stadler, Martin Lindvall, Claes Lundström, Anna Bodén, Karin Lindman, Jeronimo Rose, Darren Treanor, Johan Blomma, Karin Stacke, Nicolas Pinchaud, et al. Proactive construction of an annotated imaging database for artificial intelligence training. *Journal of digital imaging*, 34(1):105–115, 2021.
- [174] Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.
- [175] Jackson M Steinkamp, Charles Chambers, Darco Lalevic, Hanna M Zafar, and Tessa S Cook. Toward complete structured information extraction from radiology reports using machine learning. *Journal of digital imaging*, 32(4):554–564, 2019.
- [176] Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, et al. A systematic review of natural language processing applied to radiology reports. *BMC medical informatics and decision making*, 21(1):1–18, 2021.

## REFERENCES

---

- [177] Danilo Dessi, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. Tf-idf vs word embeddings for morbidity identification in clinical notes: An initial study. *arXiv preprint arXiv:2105.09632*, 2021.
- [178] Michal Marcinczuk, Mateusz Gniewkowski, Tomasz Walkowiak, and Marcin Bedkowski. Text document clustering: Wordnet vs. tf-idf vs. word embeddings. In *Proceedings of the 11th Global Wordnet Conference*, pages 207–214, 2021.
- [179] Sabri Eyuboglu, Geoffrey Angus, Bhavik N Patel, Anuj Pareek, Guido Davidzon, Jin Long, Jared Dunnmon, and Matthew P Lungren. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body fdg-pet/ct. *Nature communications*, 12(1):1–15, 2021.
- [180] Basil Mustafa, Aaron Loh, Jan Freyberg, Patricia MacWilliams, Megan Wilson, Scott Mayer McKinney, Marcin Sieniek, Jim Winkens, Yuan Liu, Peggy Bui, et al. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021.
- [181] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [182] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.
- [183] Md Mamunur Rahaman, Chen Li, Yudong Yao, Frank Kulwa, Mohammad Asadur Rahman, Qian Wang, Shouliang Qi, Fanjie Kong, Xuemin Zhu, and Xin Zhao. Identification of covid-19 samples from chest x-ray images using deep learning: A comparison of transfer learning approaches. *Journal of X-ray Science and Technology*, 28(5):821–839, 2020.
- [184] Jasha Droppo and Oguz Elibol. Scaling laws for acoustic models. *arXiv preprint arXiv:2106.09488*, 2021.
- [185] Tatsunori Hashimoto. Model performance scaling with multiple data sources. In *International Conference on Machine Learning*, pages 4107–4116. PMLR, 2021.
- [186] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

## REFERENCES

---

- [187] Tommaso Di Noto, Meritxell Bach Cuadra, Chirine Atat, Eduardo Gamito Teiga, Monika Hegi, Andreas Hottinger, Patric Hagmann, and Jonas Richiardi. Transfer learning with weak labels from radiology reports: application to glioma change detection, October 2022.
- [188] Allison Park, Chris Chute, Pranav Rajpurkar, Joe Lou, Robyn L Ball, Katie Shpanskaya, Rashad Jabarkheel, Lily H Kim, Emily McKenna, Joe Tseng, et al. Deep learning–assisted diagnosis of cerebral aneurysms using the headxnet model. *JAMA network open*, 2(6):e195600–e195600, 2019.
- [189] Zhao Shi, Chongchang Miao, U Joseph Schoepf, Rock H Savage, Danielle M Dargis, Chengwei Pan, Xue Chai, Xiu Li Li, Shuang Xia, Xin Zhang, et al. A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. *Nature communications*, 11(1):1–11, 2020.
- [190] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.
- [191] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.
- [192] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022.
- [193] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- [194] David A Wood, Sina Kafiabadi, Aisha Al Busaidi, Emily Guilhem, Jeremy Lynch, Matthew Townend, Antanas Montvila, Juveria Siddiqui, Naveen Gadapa, Matthew Bengler, et al. Labelling imaging datasets on the basis of neuroradiology reports: a validation study. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 254–265. Springer, 2020.

## REFERENCES

---

- [195] Who - ageing and health. <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>, note = Accessed: 2022-11-15.
- [196] Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359, 2017.
- [197] Hedvig Hricak, May Abdel-Wahab, Rifat Atun, Miriam Mikhail Lette, Diana Paez, James A Brink, Lluís Donoso-Bach, Guy Frija, Monika Hierath, Ola Holmberg, et al. Medical imaging and nuclear medicine: a lancet oncology commission. *The Lancet Oncology*, 22(4):e136–e172, 2021.
- [198] Robert J McDonald, Kara M Schwartz, Laurence J Eckel, Felix E Diehn, Christopher H Hunt, Brian J Bartholmai, Bradley J Erickson, and David F Kallmes. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic radiology*, 22(9):1191–1198, 2015.
- [199] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [200] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.
- [201] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584, 2017.
- [202] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.
- [203] Ge Wang. A perspective on deep imaging. *IEEE access*, 4:8914–8924, 2016.
- [204] Peter Imming, Christian Sinning, and Achim Meyer. Drugs, their targets and the nature and number of drug targets. *Nature reviews Drug discovery*, 5(10):821–834, 2006.
- [205] Heinz Mehlhorn. *Encyclopedia of parasitology: AM*. Springer Science & Business Media, 2008.

## REFERENCES

---

- [206] Benedetta Franceschiello, Tommaso Di Noto, Alexia Bourgeois, Micah M Murray, Astrid Minier, Pierre Pouget, Jonas Richiardi, Paolo Bartolomeo, and Fabio Anselmi. Machine learning algorithms on eye tracking trajectories to classify patients with spatial neglect. *Computer Methods and Programs in Biomedicine*, page 106929, 2022.
- [207] Tommaso Di Noto, Guillaume Marie, Sébastien Tourbier, Yasser Alemán-Gómez, Guillaume Saliou, Meritxell Bach Cuadra, Patric Hagmann, and Jonas Richiardi. An anatomically-informed 3d cnn for brain aneurysm classification with weak labels. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, pages 56–66. Springer, 2020.



# Papers

In the following pages, the PDF versions of the papers included in the PhD thesis are attached. Following the same order described throughout the thesis, first we report the work [35], then [47], and finally [36]. Although not directly described in the thesis, here we also report some secondary contributions of this PhD thesis:

- In [206], we compared classic ML architectures and a CNN to distinguish saccadic eye-movement trajectories of healthy patients from eye-movement trajectories of patients with spatial neglect, a neurological syndrome characterized by a lack of awareness of contralesional stimuli following right hemispheric damage.
- In [207], we addressed the task of patch-wise classification on TOF-MRA patches (with and without aneurysms), investigating the impact of negative sampling and prior anatomical knowledge. This work was a precursor of [35] where instead we addressed patient-wise detection of aneurysms.
- In [122], we helped writing the manuscript related to the ADAM challenge.



# Towards Automated Brain Aneurysm Detection in TOF-MRA: Open Data, Weak Labels, and Anatomical Knowledge

Tommaso Di Noto<sup>1</sup> · Guillaume Marie<sup>1</sup> · Sebastien Tourbier<sup>1</sup> · Yasser Alemán-Gómez<sup>1,2</sup> · Oscar Esteban<sup>1</sup> · Guillaume Saliou<sup>1</sup> · Meritxell Bach Cuadra<sup>3</sup> · Patric Hagmann<sup>1</sup> · Jonas Richiardi<sup>1</sup>

Accepted: 1 August 2022  
© The Author(s) 2022

## Abstract

Brain aneurysm detection in Time-Of-Flight Magnetic Resonance Angiography (TOF-MRA) has undergone drastic improvements with the advent of Deep Learning (DL). However, performances of supervised DL models heavily rely on the quantity of labeled samples, which are extremely costly to obtain. Here, we present a DL model for aneurysm detection that overcomes the issue with “weak” labels: oversized annotations which are considerably faster to create. Our weak labels resulted to be four times faster to generate than their voxel-wise counterparts. In addition, our model leverages prior anatomical knowledge by focusing only on plausible locations for aneurysm occurrence. We first train and evaluate our model through cross-validation on an in-house TOF-MRA dataset comprising 284 subjects (170 females / 127 healthy controls / 157 patients with 198 aneurysms). On this dataset, our best model achieved a sensitivity of 83%, with False Positive (FP) rate of 0.8 per patient. To assess model generalizability, we then participated in a challenge for aneurysm detection with TOF-MRA data (93 patients, 20 controls, 125 aneurysms). On the public challenge, sensitivity was 68% (FP rate = 2.5), ranking 4th/18 on the open leaderboard. We found no significant difference in sensitivity between aneurysm risk-of-rupture groups ( $p = 0.75$ ), locations ( $p = 0.72$ ), or sizes ( $p = 0.15$ ). Data, code and model weights are released under permissive licenses. We demonstrate that weak labels and anatomical knowledge can alleviate the necessity for prohibitively expensive voxel-wise annotations.

**Keywords** Model robustness · Weak annotation · Domain knowledge · Deep learning · Magnetic resonance angiography · Aneurysm detection

## Introduction

Time-Of-Flight Magnetic Resonance Angiography (TOF-MRA) is a non-invasive and non-contrast imaging technique sensitive to the blood flow in brain vessels. TOF-MRA has found widespread clinical application to identify Unruptured Intracranial Aneurysms (UIAs) which are small (typical diameter  $\cong 5$  mm) abnormal focal dilatations in cerebral arteries (Chen et al., 2018). If untreated, UIAs can rupture

and lead to subarachnoid hemorrhages which have a mortality rate of 40% and usually cause severe disability for patients (Frösen et al., 2012).

Manually assessing a TOF-MRA is a costly process: radiologists detect aneurysms by selectively scrolling through the TOF-MRA volumes in different planes—for instance, they check in the axial plane the most recurrent locations where aneurysms can occur. Then, the sagittal view permits better views of areas like the basilar trunk; afterwards, the coronal view can be used for areas like the anterior cerebral arteries or the Sylvian segments. In addition, Maximum Intensity Projection (MIP) images can be used to search for stenoses, or to confirm potential aneurysms that were spotted.

Considering that the workload of radiologists is steadily increasing (Rao et al., 2021) and the detection of UIAs is a meticulous and non-trivial task (Nakao et al., 2018), the development of automated algorithms that aid clinicians in detecting aneurysms with high sensitivity is an active line of

✉ Tommaso Di Noto  
tommaso.di-noto@chuv.ch

<sup>1</sup> Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

<sup>2</sup> Center for Psychiatric Neuroscience, Department of Psychiatry, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

<sup>3</sup> Center for Biomedical Imaging, CIBM, Lausanne, Switzerland

research which holds the promise of improving care while reducing radiologists' assessment times.

Before the popularization of Deep Learning (DL), (Arimura et al., 2004) detected aneurysms by means of image filtering, and later, (Yang et al., 2011) used candidate points of interest in the brain arteries to locate aneurysms. Then, starting from 2016, there was a shift towards DL algorithms, which have now become the de facto standard for UIA detection. Table 1 illustrates several recent studies that use DL for UIA detection. Despite their success, these DL approaches are still constrained by a major bottleneck common to several medical applications: the lack of large, labeled datasets. This is mainly due to two factors: first, the creation of voxel-wise labels for medical images is tedious and time-consuming for radiologists (Razzak et al., 2018); second, none of the TOF-MRA studies to date made their dataset publicly available (Joo et al., 2020; Nakao et al., 2018; Sichtermann et al., 2019; Stember et al., 2019; Ueda et al., 2019). This hampers reproducibility and multi-site analyses that are paramount for building robust DL architectures. The lack of openly available data, such as the TOF-MRA challenge dataset (Timmins et al., 2021), also

hinders comparisons across models. Of all reviewed studies of Table 1, only (Baumgartner et al., 2021) evaluated their models on the challenge dataset.

In this work, we develop a fully automated DL network for UIA detection and propose to mitigate the data availability bottleneck as follows: we explore the use of “weak” labels (Abousamra et al., 2020; Ezhov et al., 2018; Ke et al., 2020). These can be coarse or oversized annotations that are less precise, but considerably faster to create for medical experts. In addition, we release our annotated in-house dataset to the community. To the best of our knowledge, this will be the largest openly available TOF-MRA aneurysm dataset to date.

Furthermore, we constrain the DL analysis only to the areas of the brain where aneurysm occurrence is plausible. This anatomically-informed approach aims at simulating the selective analysis that radiologists perform on the TOF-MRA scans. Then, we assess multi-site robustness by evaluating our algorithm on the external TOF-MRA challenge dataset (Timmins et al., 2021). Last, since every aneurysm can have a different prognosis, we investigate how the performances of our model change with respect to aneurysm

**Table 1** Summary of papers that use deep learning models to tackle automated brain aneurysm detection/segmentation

Paper	Modality	Task(s)	N. Sub	N. Aneurysms	DL Model	Model input	Voxel-wise labels	Use anatomical information	Multi-Site
(Ueda et al., 2019)	MRA	Detection	1271	1477	ResNet	2D patches	Not specified	No	Yes
(Joo et al., 2020)	MRA	Detection	744	761	3D ResNet	3D patches	Yes	Yes	Yes
(Nakao et al., 2018)	MRA	Detection	450	508	CNN	2D MIP patches	Yes	Yes	No
(Stember et al., 2019)	MRA	Detection	302	336	RCNN	2D MIP patches	Yes	No	No
(Baumgartner et al., 2021)	MRA	Detection	254	N/A	nnDetection	3D patches	Yes	No	No
(Sichtermann et al., 2019)	MRA	Detection (via segmentation)	85	115	DeepMedic	3D patches	Yes	Yes	No
(Shi et al., 2020)	CTA	Detection + Segmentation	1177	1099	3D UNET	3D patches	Yes	Yes	Yes
(Yang et al., 2020)	CTA	Detection	1068	1337	ResNet	3D patches	Not specified	No	Yes
(Park et al., 2019)	CTA	Segmentation + CAD assessment	662	358	HeadXNet	3D patches	Yes	Yes	No
(Dai et al., 2020)	CTA	Detection	311	352	RCNN	2D NP images	Not specified	No	Yes
(Liu et al., 2021)	DSA	Detection + Segmentation	451	485	3D UNET	3D DSA volumes	Yes	Yes	No
(Duan et al., 2019)	DSA	Detection	281	261	2D CNN	2D DSA images	Bounding Boxes	Yes	No
(Hainc et al., 2020)	DSA	Detection	240	187	2D CNN	2D DSA images	ROI circle	No	No

Use anatomical information: whether the method uses some sort of anatomical prior knowledge during training, patch sampling or inference (more details in Online Resources – Section A)

*MRA* Magnetic Resonance Angiography, *CTA* Computed Tomography Angiography, *DSA* Digital Subtraction Angiography, *N* number, *Sub* subjects

risk-of-rupture groups (defined in “[Aneurysm Annotation, Size, Location and Risk Groups for In-house Dataset](#)” section), location and size.

## Materials and Methods

### In-house Dataset

This study was approved by the regional ethics committee; written informed consent was waived. In this retrospective work, we included consecutive patients that underwent TOF-MRA between 2010 and 2015, and for which the corresponding radiological reports were available. Patients with ruptured/treated aneurysms or with other vascular pathologies were excluded. Totally thrombosed aneurysms and infundibula (dilatations of the origin of an artery) were likewise excluded. In total, we retrieved 284 TOF-MRA subjects: 157 had one (or more) UIAs, while 127 did not present any. Table 2 illustrates the main demographic information for our study group. A 3D gradient recalled echo sequence with Partial Fourier technique was used for all subjects (acquisition parameters are reported in Online Resources—Table 1). 214 subjects of this study were also used in (Di Noto et al., 2020). This prior article dealt with patch-wise classification, whereas here we address patient-wise aneurysm detection. The dataset was anonymized and organized according to the Brain Imaging Data Structure (BIDS) standard (Gorgolewski, 2008). It is available on OpenNeuro (Markiewicz et al., 2021) at <https://openneuro.org/datasets/ds003949> under the CC0 license.

### Aneurysm Annotation, Size, Location and Risk Groups for In-house Dataset

Aneurysms were annotated by one radiologist with 2 years of experience in neuroimaging, and double-checked by a senior neuroradiologist with over 15 years of experience to exclude potential false positives or false negatives. Two annotation schemes were followed:

1. **Weak labels:** for most subjects (246/284), the radiologist used the Multi-image Analysis GUI (Mango) software (v. 4.0.1) to create the aforementioned weak labels. These correspond to spheres that enclose the whole aneurysm, regardless of the shape. In other words, the size of the spheres was chosen manually by our radiologist on a case-by-case basis ensuring that the whole aneurysm was always entirely enclosed within the sphere. A visual example of one weak label is provided in Fig. 1.
2. **Voxel-wise labels:** for the remaining subjects (38/284), the radiologist used ITK-SNAP (v. 3.6.0) (Yushkevich et al., 2006) to create voxel-wise labels drawn slice by slice scrolling in the axial plane. No specific selection criterion was used to select the 38 subjects, which were consecutive to the 246 of the first group.

The overall number of aneurysms included in the study is 198 (178 saccular, 20 fusiform). Table 3 shows their locations and sizes grouped according to the PHASES score (Greving et al., 2014). This is a clinical score used to assess the 5-year risk of rupture of aneurysms. Although using the PHASES sizes leads to a very skewed distribution (e.g. the category size  $d \leq 7$  mm contains 91% of the aneurysms), we decided to stick to this grouping since it is the one used in the clinic.

In addition, for post-hoc analysis and stratification purposes, we divided the aneurysms into two groups based on their risk of rupture: *low-risk* and *medium-risk*. Aneurysms in the *low-risk* group are those that are monitored over time, but do not require any intervention. Instead, aneurysms in the *medium-risk* group can be considered for treatment. We computed for each aneurysm a partial PHASES score that only considered size, location, and patient’s age, thus neglecting population, hypertension, and earlier aneurysmal hemorrhage, since this information was not available for all patients. If an aneurysm had partial PHASES score  $\leq 4$ , it was assigned to the *low-risk* group, while if it had a partial score  $> 4$ , it was assigned to the *medium-risk* group. Each aneurysm was reviewed by our senior neuroradiologist to assess whether the partial PHASES score was reasonable.

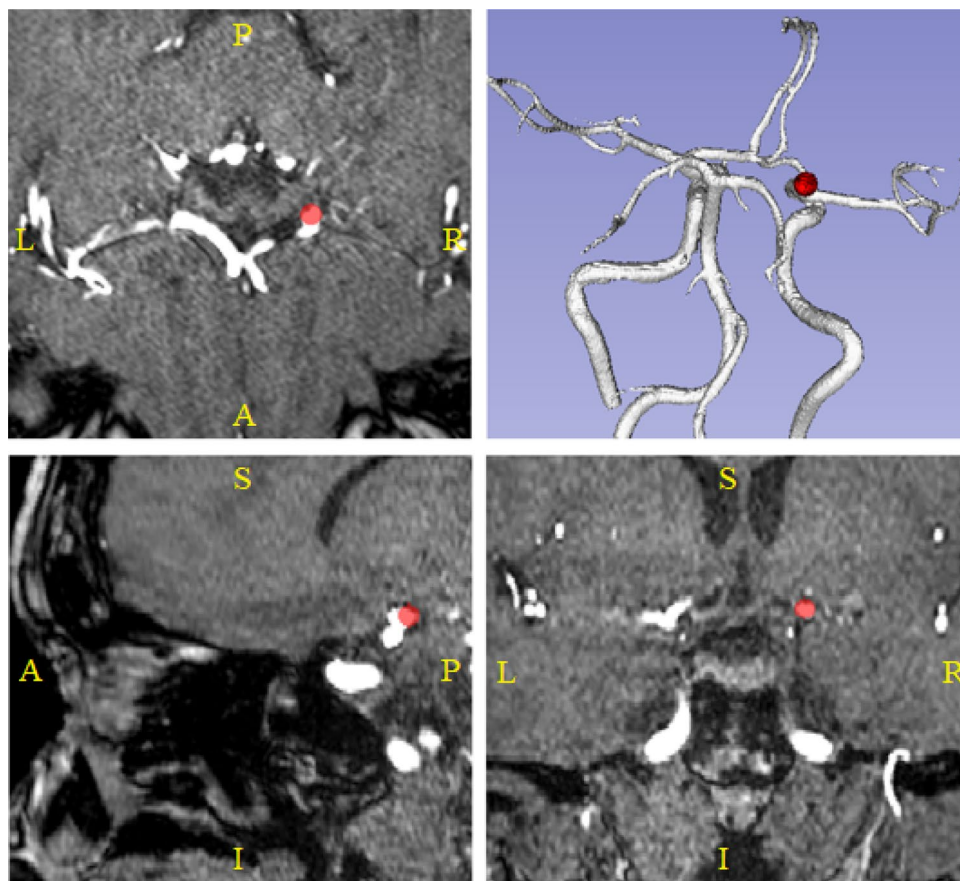
**Table 2** Demographics of the study sample

	Patients	Controls	Test, <i>p</i> value	Whole Sample
<b>N</b>	157	127	/	284
<b>Age (y)</b>	56 ± 14	46 ± 17	$t = -4, 3, p = 7.6 \times 10^{-7}$	51 ± 16
<b>Sex</b>	53 M, 104F	61 M, 66F	$\chi^2 = 5.9 p = 0.01$	114 M, 170F
<b># UIA</b>	198	0	/	198

Patients = subjects with aneurysm(s). Controls = subjects without aneurysms. Age calculated in years and presented as mean ± standard deviation. Two-sided t-test to compare age between patients and controls. Chi-squared test to compare sex counts between patients and controls

*N* number of samples, *M* males, *F* females, *UIA* Unruptured Intracranial Aneurysms

**Fig. 1** TOF-MRA orthogonal views of a 62-year-old female patient. Red areas correspond to our spherical weak labels. Top-left: axial plane; top-right: 3D posterior reconstruction of the cerebral arteries; bottom-left: sagittal plane; bottom-right: coronal plane



Fusiform aneurysms were excluded from the risk analysis since the PHASES score was built for saccular aneurysms. Similarly, extracranial carotid artery aneurysms were excluded since they do not bleed in the subarachnoid space. This resulted in 141 *low-risk* and 23 *medium-risk* aneurysms. A table summarizing aneurysm shape, size, location, associated PHASES score and risk groups is provided as Supplementary Material.

**Table 3** Locations and sizes of aneurysms according to the PHASES score for the in-house dataset

		Count	%
<b>Location</b>	ICA	59	29.8 (59/198)
	MCA	57	28.8 (57/198)
	ACA/Pcom/Posterior	82	41.4 (82/198)
<b>Size</b>	$d \leq 7$ mm	180	91.0 (180/198)
	7 – 9, 9 mm	7	3.5 (7/198)
	10 – 19, 9 mm	10	5.0 (10/198)
	$d \geq 20$ mm	1	0.5 (1/198)

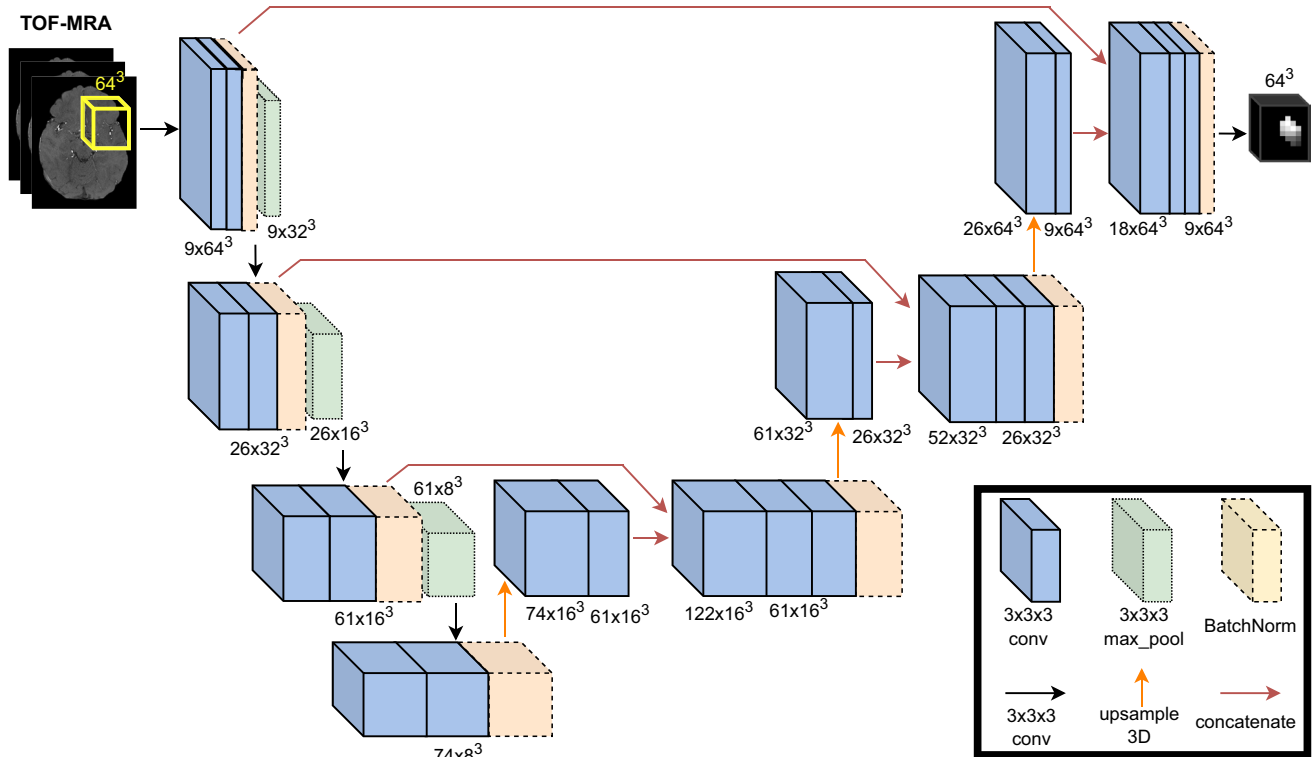
ICA Internal Carotid Artery, MCA Middle Cerebral Artery, ACA Anterior Cerebral Arteries, Pcom Posterior communicating artery, Posterior posterior circulation,  $d$  maximum diameter

## Data Processing

Several preprocessing steps were carried out for each subject. First, we performed skull-stripping with the FSL Brain Extraction Tool (v. 6.0.1) (Smith, 2002). Second, we applied N4 bias field correction with SimpleITK (v. 1.2.0) (Tustison et al., 2010). Third, we resampled all volumes to a median voxel spacing, again with SimpleITK. This effectively normalizes nonuniform voxel sizes (Isensee et al., 2021). Last, a probabilistic vessel atlas built from multi-center MRA datasets (Mouches & Forkert, 2014) was co-registered to each patient's TOF-MRA using ANTS (v. 2.3.1) (Avants et al., 2014) (details in Online Resources – Section B). The atlas was used both during training and inference (see “Use of Anatomical Information” section).

## Network, Cross-Validation, Metrics and Statistics

**Network** The deep learning model used in this study is a custom 3D UNET, inspired by the original work (Özgün et al., 2016). We used upsample layers in the decoding branch rather than transpose convolutions since these led to faster model convergence. Figure 2 illustrates the structure of our network. We used 3D TOF-MRA patches as input to our



**Fig. 2** Proposed variant of the 3D UNET. The input corresponds to a 64x64x64 voxels TOF-MRA patch. The output is a probabilistic patch with the same size of the input, but where each voxel corre-

sponds to the probability of either belonging to foreground (i.e., aneurysm) or background. *Conv* convolutional, *Max\_pool* max pooling, *BatchNorm* batch normalization

network. We set the side of the input patches to 64x64x64 voxels to include even the largest aneurysms. All patches were Z-score normalized, as is common practice (Bengio et al., 2016). A kernel size of 3x3x3 was used in all convolutional layers, with padding and stride = 1. We applied the ReLU activation function for all layers, except for the last layer which is followed by a sigmoid function. To fit the model, the Adam optimization algorithm (Kingma & Ba, 2015) was applied with adaptive learning rate (initial learning rate = 0.0001). We trained the model for 100 epochs, and we adopted the Combo loss function (Taghanaki et al., 2019) with  $\alpha = \beta = 0.5$ . This function combines Dice and Cross-entropy, and has proven to be effective for imbalanced segmentation tasks. We used Xavier initialization (Glorot & Bengio, 2010) for all layers. Biases were initialized to 0 and a batch size of 8 was chosen. Batch normalization (Ioffe & Szegedy, 2015) was used to prevent overfitting. The number of convolutional filters, the batch size, the value of  $\alpha$  (and therefore  $\beta = 1 - \alpha$ ) and the learning rate were chosen using the Optuna algorithm (Akiba et al., 2019) on an internal validation set (20% of training cases of external cross-validation fold 1, see below for cross-validation details). The total number of trainable parameters in our network

is 855,111. Training and evaluation were performed with Tensorflow 2.4.0 and a GeForce RTX 2080TI GPU with 11 GB of SDRAM.

**Cross-validation** To evaluate detection performances, we conducted a fivefold cross-validation on the 246 subjects with weak labels. At each cross-validation split, 80% ( $\approx 197/246$ ) of the subjects are used for training the network, while the remaining 20% ( $\approx 49/246$ ) of the subjects are used to compute patient-wise results (i.e. for inference). This division occurs 5 times (as the number of folds) and every time a different 80%-20% split is created, meaning that all 246 patients are ultimately used for evaluation. At each cross-validation split, the 38 patients with voxel-wise labels were always added to the training set to increase the effect size of label quality in further analyses (see experiments in “Use of Weak Labels”). To avoid over-optimistic results, we ensured that patients with multiple sessions were not split between training and test set. In order to make results comparable across experiments, we always used the same cross-validation split and we released this split for reproducibility on [https://github.com/connectomicslab/Aneurysm\\_Detection](https://github.com/connectomicslab/Aneurysm_Detection).

In all experiments on the in-house dataset, we always pre-trained our network on the whole ADAM training dataset (Timmins et al., 2021) and then fine-tuned it on the in-house training data. To validate the effectiveness of pre-training on ADAM, we performed ablation experiments of domain adaptation across the two datasets (in-house and ADAM). As these experiments are beyond the main focus of the manuscript, we added them in the Online Resources – Section F. When performing pre-training on the ADAM dataset, we applied both anatomically-informed expedients described below in “Use of Anatomical Information” section.

**Metrics and Statistics** In line with the ADAM challenge (presented in “Participation to the ADAM Challenge” section), we used sensitivity and false positive (FP) rate as detection metrics. A detection was considered correct if the center-of-mass of the predicted aneurysm was located within the maximum aneurysm size of the ground truth mask. In addition, we computed the Free-response Receiver Operating Characteristic (FROC) curve (Chakraborty & Berbaum, 2004). To compare different model configurations, we used a two-sided Wilcoxon signed-rank test of the areas under the FROC curves across test subjects, as similarly performed in (Ward et al., 1999). To compare the performances of a configuration with respect to aneurysm rupture risk, location and size we performed several Chi-squared tests (McHugh, 2012). The statistical tests were performed using SciPy (v.1.4.1), setting a significance threshold  $\alpha=0.05$ .

## Experiments

In this section, we will present the four experiments that we conducted: in “Use of Weak Labels” section, we investigate the use of weak labels in terms of difference in annotation time and in detection performances, when comparing to voxel-wise labels; in “Use of Anatomical Information” section, we present our anatomically-informed approach for tackling UIA detection; in “Participation to the ADAM Challenge” section, we describe the participation to the ADAM challenge to investigate the generalization of our model; in “Performances With Respect to Risk-of-rupture, Location and Size” section, we analyze

the changes in detection performances with respect to aneurysm risk-of-rupture groups, location and size.

### Use of Weak Labels

The goal of this experiment was to answer the following questions: 1) how much faster is the creation of weak labels with respect to the creation of voxel-wise labels? 2) what is the impact of using weak labels in terms of detection performances when comparing to voxel-wise labels?

To answer the first question, we selected a subset of 14 patients (mean aneurysm size (s.d.)=5.2 (1.0) mm), and we assessed the time difference between the two annotation schemes (i.e. all 14 patients were annotated first with weak labels, and then with voxel-wise labels). These 14 patients were chosen randomly among the 284 TOF-MRA subjects, but we ensured that the mean aneurysm size was representative of the whole cohort.

To answer the second question, we used the 38 subjects with voxel-wise labels and for these patients we artificially generated corresponding weak spherical labels (‘weakened’ labels, details in Online Resources – Section C). Then, to evaluate the influence of annotation quality (weakened vs. voxel-wise) in terms of detection performances, we conducted 3 experiments in which we used an increasing number of patients with voxel-wise labels: (i) all 38 patients with weakened labels (*Model 1*, Table 4), (ii) 19 patients with weakened labels and 19 with voxel-wise labels (*Model 2*, Table 4), and (iii) all 38 patients with voxel-wise labels (*Model 3*, Table 4). Results related to the use of weak labels are presented in “Weak Labels Allow Fourfold Annotation Speedup Without Degrading Performances” section.

### Use of Anatomical Information

Because the task of aneurysm detection is extremely spatially constrained, we exploit the prior information that aneurysms a) must occur in vessels, and b) tend to occur in specific locations of the vasculature. To include this anatomical knowledge, one of our radiologists pinpointed in the vessel atlas (described in “Aneurysm Annotation, Size,

**Table 4** Average detection results on the in-house dataset across test folds when changing the ratio of voxel-wise/weakened labels. Sensitivity values are reported as mean and 95% Wilson confidence interval inside parentheses

Model Configuration	Anatomically-informed patch selection	Anatomically-informed sliding window	Labels of 38 added subs	Avg. Sensitivity (CI)	Avg. FP rate
<i>Model 1</i>	Yes	Yes	38 weakened	95/127 = 75% (65%, 80%)	1.3
<i>Model 2</i>	Yes	Yes	19 weakened, 19 voxel-wise	99/127 = 78% (68%, 82%)	<b>0.9</b>
<i>Model 3</i>	Yes	Yes	38 voxel-wise	101/127 = <b>80%</b> (72%, 85%)	1.2

Bold values represent the best performances

Avg average, FP false positive, CI confidence interval, voxel-wise labels drawn slice by slice on the axial plane, weakened voxel-wise labels that are artificially converted to weak spherical labels, subs subjects

Location and Risk Groups for In-house Dataset” section) the location of 20 landmark points where aneurysm occurrence is most frequent (list in Online Resources – Table 2). These points were chosen according to the literature (Brown & Broderick, 2014) and were co-registered to the TOF-MRA space of each subject, as illustrated in Fig. 3.

**Training** We apply an anatomically-informed selection of training patches to sample both negative (without aneurysms) and positive (with aneurysms) samples. Specifically, 8 positive patches per aneurysm were randomly extracted in a non-centered fashion. Then, we extracted 50 negative patches per TOF-MRA volume. Out of these, 20 were centered in correspondence with the landmark points, 20 were patches containing vessels (details in Online Resources – Section D), and 10 were extracted randomly. Overall, this sampling strategy allows us to extract most of the negative patches (i.e., all but the random ones) which are comparable to the positive ones in terms of average intensity. To mitigate class imbalance, we applied data augmentations on positive patches: namely, rotations (90°, 180°, 270°), flipping (horizontal, vertical), contrast adjustment, gamma correction, and addition of gaussian noise.

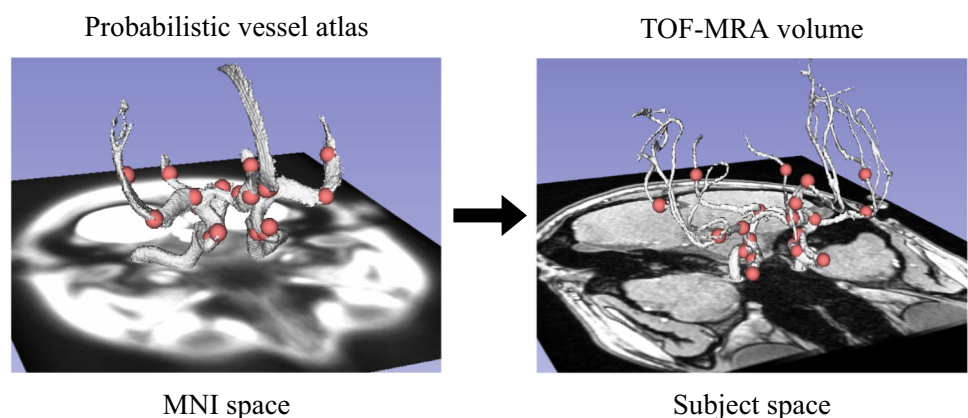
**Inference** The patient-wise evaluation was performed following the sliding window approach (details in Online Resources – Section E). We exploited again the prior anatomical information described above by retaining only the patches which are both within a minimum distance from the landmark points and fulfill specific intensity criteria (details in Online Resources – Section D). The rationale behind this choice was to only focus on patches located in the main cerebral arteries, as shown in Fig. 4. Two post-processing steps were adopted: first, we kept a maximum of 5 candidate aneurysms per patient (only the 5 most probable). Second, we applied test-time augmentation to increase sensitivity.

**Validation** To validate the effectiveness of our two anatomically-informed expedients (patch sampling and sliding window), we first evaluated an anatomically-agnostic baseline where none of the two expedients is used and the 38 added subjects have *weakened* labels (*Model 4*, Table 5). Second, we evaluated the same anatomically-agnostic baseline (none of the two expedients used) but with the 38 subjects having voxel-wise labels (*Model 5*, Table 5). Third, we tested one model where only the anatomically-informed patch sampling is carried out (*Model 6*, Table 5). Last, we computed performances when only the anatomically-informed sliding window is performed (*Model 7*, Table 5). Results related to the use of anatomical information are shown in “Anatomically-informed Sliding Window Increases Detection Performances” section.

### Participation to the ADAM Challenge

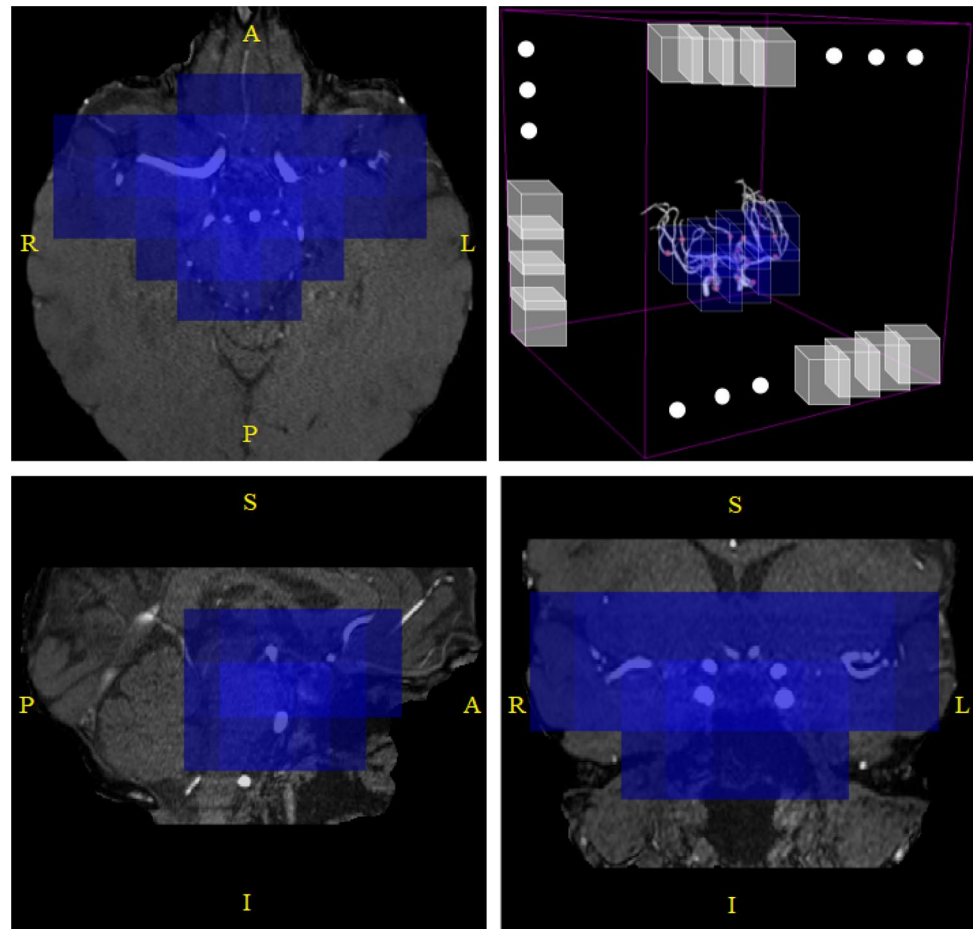
To evaluate model performances in data coming from a different institution, we participated to the Aneurysm Detection And segmentation (ADAM) challenge (<http://adam.isi.uu.nl/>) (Timmins et al., 2021). The ADAM training dataset is composed of 113 TOF-MRA (93 patients with UIAs, 20 controls). The total number of UIAs is 125 and the voxel-wise annotations were drawn in the axial plane by two radiologists. Instead, the unreleased test dataset is made of 141 cases (117 patients, 26 controls) and it is solely used by the organizers to compute patient-wise results. To improve detection performances on the ADAM test set, we pre-trained our network on the whole in-house dataset and then fine-tuned it on the ADAM training dataset. Results related to our model submission to the ADAM challenge are presented in “The Proposed Model Ranked At the Top of the ADAM Challenge” section.

**Fig. 3** **left:** 20 landmark points (in red) located in specific positions of the cerebral arteries (white segmentation) in MNI space. **right:** same landmark points co-registered to the TOF-MRA space of a 21-year-old, female subject without aneurysms





**Fig. 4** TOF-MRA orthogonal views of a 62-year-old female subject after brain extraction: blue patches are the ones which are retained in the *anatomically-informed* sliding-window approach. (top-right): 3D schematic representation of sliding-window approach; out of all the patches in the volume (white patches), we only retain those located in the proximity of the main brain arteries (blue ones)



#### Performances with Respect to Risk-of-rupture, Location and Size

Each aneurysm has a different prognosis and, depending on its risk-of-rupture group (defined in “[Aneurysm Annotation, Size, Location and Risk Groups for In-house Dataset](#)” section), it will be either monitored over time (low risk) or considered for treatment (medium risk). Therefore, we investigated how detection performances would vary with respect to the risk-of-rupture groups. In

addition, we also explored how performances would vary with respect to aneurysm location and size. Although the latter analysis is less relevant from a clinical perspective, it is still interesting from a methodological point of view and it is also frequent in the literature. Results related to the detection performances with respect to aneurysm risk-of-rupture groups, location and size are described in “[Detection Performances Across Rupture Risk, Location, and Size](#)” section.

**Table 5** Average detection results on the in-house dataset across test folds when applying none, or one of the two anatomically-informed expedients. Sensitivity values are reported as mean and 95% Wilson confidence interval inside parentheses

Model Configuration	Anatomically-informed patch selection	Anatomically-informed sliding window	Labels of 38 added subs	Avg. Sensitivity (CI)	Avg. FP rate
Model 4	No	No	38 <i>weakened</i>	83/127 = 65% (55%, 71%)	4.6
Model 5	No	No	38 voxel-wise	95/127 = 74% (63%, 78%)	4.5
Model 6	Yes	No	38 voxel-wise	61/127 = 48% (38%, 55%)	4.8
Model 7	No	Yes	38 voxel-wise	106/127 = <b>83%</b> (75%, 88%)	<b>0.8</b>

Bold values represent the best performances

Avg average, *FP* false positive, *CI* confidence interval, *voxel-wise* labels drawn slice by slice on the axial plane, *weakened* voxel-wise labels that are artificially converted to weak spherical labels, *subs* subjects

## Results

### Weak Labels Allow Fourfold Annotation Speedup Without Degrading Performances

When measuring the time needed to create weak vs. voxel-wise annotations on the 14 subjects described in “Use of Weak Labels” section, we noticed a significant difference (two-sided Wilcoxon signed-rank test – annotation timings,  $W=0$ ,  $p=0.001$ ): creating weak annotations (average  $23\text{ s} \pm 6$  per aneurysm) resulted to be approximately 4 times faster than creating voxel-wise annotations (average  $93\text{ s} \pm 25$ ). A more detailed stratification of the timings with respect to location and size is provided in Supplementary Figs. 1 and 2.

Subsequently, to investigate the effect that voxel-wise labels can have for detection performances with respect to weak labels, we conducted several experiments where an increasing ratio of voxel-wise/*weakened* labels was used for the 38 patients described in “Use of Weak Labels” section. Table 4 shows detection performances when the ratio is gradually increased.

The configuration with all voxel-wise labels (*Model 3*) had higher sensitivity with respect to the other two configurations with *weakened* labels (*Model 1* and *Model 2*). However, this difference was not significant (two-sided Wilcoxon signed-rank test on the areas under the FROC curves,  $W=14.0$ ,  $p=0.054$  when comparing to *Model 1* and  $W=685.5$ ,  $p=0.977$  when comparing to *Model 2*).

### Anatomically-informed Sliding Window Increases Detection Performances

In Table 5, we report detection results when adopting zero, one, or both anatomically-informed expedients presented in “Use of Anatomical Information” section. In the anatomically-agnostic baseline with the 38 subjects having *weakened* labels (*Model 4*), the negative patch sampling is random and all non-zero patches of the TOF-MRA volumes are retained in the sliding window approach, thus disregarding any anatomical constrain. Similarly, row 2 (*Model 5*) shows detection results when using neither the anatomically-informed patch sampling nor the anatomically-informed sliding window, but this time with the 38 subjects having voxel-wise labels. Row 3 (*Model 6*) illustrates detection performances when only the anatomically-informed patch sampling is applied, but the sliding window is still anatomically-agnostic. Instead, row 4 (*Model 7*) shows the inverse scenario (i.e. random negative patch sampling, but anatomically-informed sliding window). *Model 7* statistically outperformed *Model 5* ( $W=74.5$ ,  $p=2 \times 10^{-6}$ ), thus indicating that the anatomically-informed sliding window is

an effective expedient to increase detection results. In fact, sensitivity is increased and the average FP rate is drastically reduced. In addition, we compared *Model 5* and *Model 6* and we saw that *Model 5* significantly outperforms *Model 6* ( $W=202.0$ ,  $p=8 \times 10^{-6}$ ). This finding shows that the anatomically-informed patch sampling is detrimental for detection performances when the sliding window is anatomically-agnostic. Last, when comparing *Model 3* and *Model 7* we found no significant difference ( $W=81.5$ ,  $p=0.24$ ): this result indicates that the anatomically-informed patch sampling is not detrimental when we are also applying the anatomically-informed sliding window.

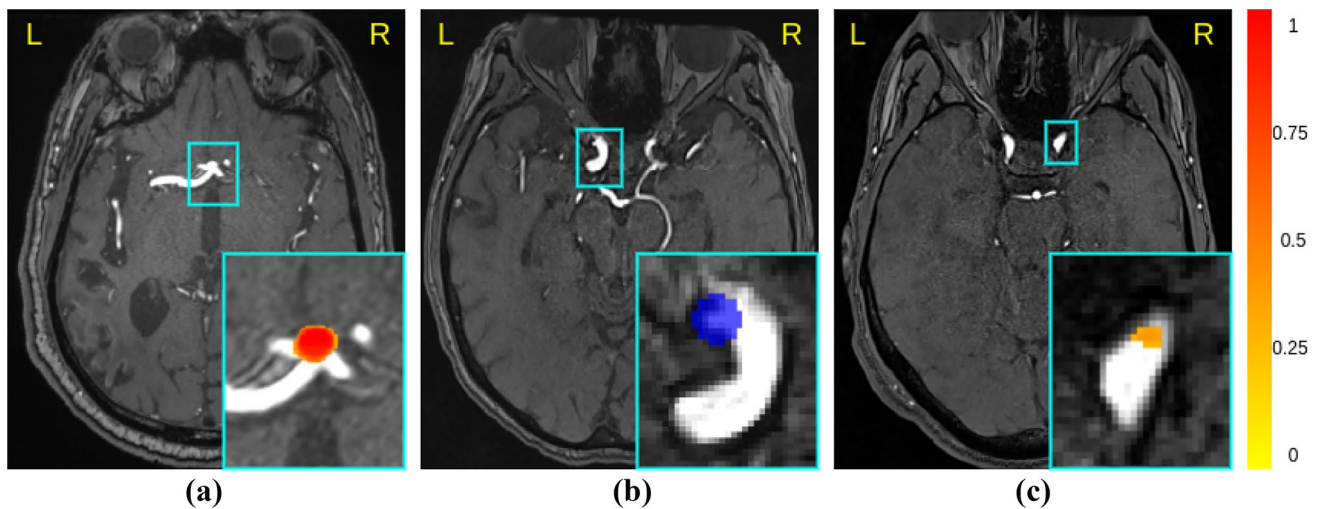
To provide a visual interpretation of our network predictions, we show in Fig. 5 one correctly identified aneurysm (true positive), one small, missed aneurysm (false negative) and one false positive prediction. Also, in Fig. 6 we report the FROC curves corresponding to *Model 5*, *Model 6*, and *Model 7*. This figure reflects the statistical tests: *Model 7* (green curve) outperforms the anatomically-agnostic *Model 5* (red curve) at all operating points. Similarly, *Model 5* (red curve) significantly outperforms *Model 6* (blue curve), confirming the effectiveness of the anatomically-informed sliding window and the ineffectiveness of the anatomically-informed patch sampling.

### The Proposed Model Ranked At the Top of the ADAM Challenge

Table 6 illustrates detection results on the ADAM test dataset. Our algorithm ranked in 4th/18 position for detection and in 4th/15 position for segmentation (with highest volumetric similarity). Interested readers can check the methods proposed by other teams on the challenge website (<https://adam.isi.uu.nl/>) and in the paper (Timmins et al., 2021).

### Detection Performances Across Rupture Risk, Location, and Size

Supplementary Fig. 3 illustrates performances achieved by one of our top-performing models (*Model 3*, Table 4) stratified according to the two risk groups presented in “Aneurysm Annotation, Size, Location and Risk Groups for In-house Dataset” section. For the *low-risk* group, our model reaches a mean sensitivity of 80%, while for the *medium-risk* group it reaches a mean sensitivity of 73%. The difference was not significant ( $\chi^2=0.09$ ,  $DoF=1$ ,  $p=0.75$ ). In Supplementary Figs. 4 and 5, we also report the model sensitivity stratified according to aneurysm location and size of the PHASES score, respectively. No significant difference was found across different locations ( $\chi^2=0.64$ ,  $DoF=2$ ,  $p=0.72$ ) or sizes ( $\chi^2=0.92$ ,  $DoF=2$ ,  $p=0.15$ , excluding  $n=1$  huge aneurysm with  $s > 20$  mm). Regarding aneurysm



**Fig. 5** Qualitative analysis of predictions and errors. The heatmap generated by the network ranges from 0 (low probability, yellow color) to 1 (high probability, red color) **(a)** True positive prediction in the anterior communicating artery. **(b)** False negative (i.e., missed

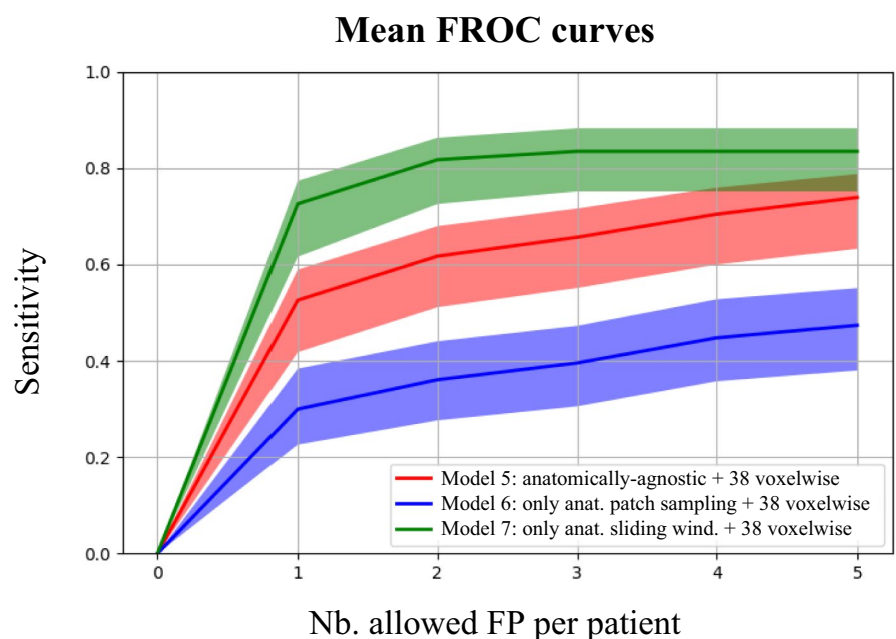
aneurysm) in the internal carotid artery. The ground truth label mask is shown in blue. **(c)** False positive prediction in the internal carotid artery

size, we conducted a further stratification of performances since most of the aneurysms lied in the group ( $<7$  mm). Thus, we divided this group into subgroups, namely  $\leq 3$ ,  $3 < s \leq 5$ , and  $5 < s < 7$ . Detection results with this more granular stratification are shown in Supplementary Fig. 6. The model sensitivity was significantly lower for the tiny aneurysms ( $\leq 3$ ) with respect to the other two subgroups ( $\chi^2 = 27.57$ ,  $DoF = 2$ ,  $p = 10^{-6}$ ).

## Discussion

This work shows that competitive results can be obtained in automated aneurysm detection for TOF-MRA data even with rapid data annotation. To this end, we proposed a fully-automated, deep learning algorithm that is trained using weak labels and exploits prior anatomical knowledge.

**Fig. 6** Mean Free-response Receiver Operating Characteristic (FROC) curves across the five test folds of the cross-validation. Shaded areas represent the 95% Wilson confidence interval. The three models correspond to *Model 5*, *Model 6*, and *Model 7*. *Anatomically-agnostic model*: none of the two anatomically-informed expedients are used. *Anat*: Anatomically-Informed



**Table 6** Detection results on the ADAM dataset. Our team (in bold) ranked in 4th position in the open leaderboard out of 18 participating groups

Ranking	Team	Detection	
		Sens	Avg. FP rate
1	abc	68%	0.40
2	xlim	70%	4.03
3	mibaumgartner	67%	0.13
<b>4</b>	<b>unil-chuv3</b>	<b>68%</b>	<b>2.50</b>
5	joker	63%	0.16
...			
18	ibbm	2%	0.01

Sens sensitivity, FP false positive

Despite being less accurate, weak labels are drastically faster to create for medical experts reducing fourfold the annotation time. Although the configuration with all voxel-wise labels (*Model 3*, Table 4) had higher sensitivity, we found no statistical difference when comparing with the configurations with some (*Model 2*) or all *weakened* labels (*Model 1*). This finding indicates that weak labels are sufficient to obtain satisfactory detection results on our in-house dataset. If reasoning in terms of larger datasets (e.g., thousands of patients), the weak annotation proposed in this work is a scalable solution which can significantly alleviate the annotation bottleneck in medical ML applications.

In addition to the use of weak labels, our model leverages the underlying anatomy of the brain vasculature (i.e., we “*anatomically-informed*” our network) in order to simulate the radiologists’ exploration of the TOF-MRA scans. First, most of the negative patches (i.e. patches without aneurysms) extracted during training either contained a vessel or were located in correspondence with the aneurysm landmark points. Second, we limited the sliding window approach only to regions of the brain that are plausible for aneurysm occurrence. These constraints reflect the radiologists’ behavior in the sense that only regions containing vessels, or at higher risk for aneurysms are scanned, while the rest of the brain is neglected. The experiments in “[Anatomically-informed Sliding Window Increases Detection Performances](#)” section showed that the anatomically-informed sliding window is an effective expedient since it increases sensitivity, while reducing the average FP rate. Instead, the anatomically-informed patch sampling proved to be negligible when combined with the anatomically-informed sliding-window (*Model 3* vs. *Model 7*), or even detrimental when the sliding window was anatomically-agnostic (*Model 5* vs. *Model 6*). We hypothesize that applying only the anatomically-informed patch sampling leads to a domain shift issue: specifically, the model is trained using intensity-matched patches, but then is tested with any patch in the brain (because there is

no anatomically-informed sliding window). We think this difference between training and test domain is what causes the decrease in performances in the comparison *Model 5* vs. *Model 6*.

Nevertheless, the anatomically-informed sliding window expedient suggests that injecting prior anatomical knowledge in the pipeline can improve detection performances. We believe this general principle is also applicable to other pathologies with sparse spatial extent.

The state-of-the-art for automated brain aneurysm detection in TOF-MRA has been rapidly advancing in the last five years, especially due to the advent of deep learning algorithms. However, further multi-site validation is needed before safely applying these algorithms during routine clinical practice. Although (Joo et al., 2020; Ueda et al., 2019) did publish results obtained from multiple institutions, none of them released their dataset publicly which makes comparisons between algorithms unfeasible. The comparisons between methods are further hindered by the use of non-standardized evaluation metrics (e.g. FROC/lesion-wise sensitivity/subject-wise specificity) or by the fact that not all related studies include both patients (subjects with aneurysms) and controls (subjects without aneurysms). By openly releasing our dataset, we aim to bridge the data availability gap and foster reproducibility in the medical imaging community. The combination of our in-house dataset and the ADAM dataset will allow researchers to assess the realistic robustness of proposed algorithms on heterogeneous data generated from different scanners, acquisition protocols and study population. In addition, it could help increasing detection performances which are still too far from being clinically useful, considering that even the team with highest sensitivity on the ADAM test set (team *xlim*) only reaches a value of 70% (i.e., 30% of aneurysms still not detected), with 4 FPs per case.

In a separate analysis, we also computed the sensitivity of our model with respect to the PHASES score risk of rupture, location, and size. No significant differences were found across the three groups indicating that our model is robust to different aneurysm types. However, when stratifying the aneurysm sizes into finer subgroups, we noticed that sensitivity for extremely tiny aneurysms ( $\leq 3$  mm) was significantly lower, which confirms a known trend (Timmins et al., 2021).

Our work has several limitations. First, even combining our in-house dataset with the ADAM dataset, the number of subjects is still limited when compared to some related TOF-MRA (Joo et al., 2020; Ueda et al., 2019) or Computed Tomography Angiography (Park et al., 2019; Shi et al., 2020; Yang et al., 2020) studies. Second, we acknowledge that the number of patients for whom we compared the different annotations schemes (i.e., weak vs. voxel-wise) is limited ( $N = 38$ ); it is possible that statistically significant

performance differences could be found with a larger sample size. Third, we have to further increase detection performances if we plan to deploy our model as a second reader for radiologists, especially to detect tiny aneurysms which are more frequently overlooked (Keedy, 2006).

In future works, we aim at enlarging the TOF-MRA dataset and experiment new variants of the 3D encoding–decoding UNET. For instance, we might consider a multi-scale approach with patches of larger (or smaller) scales. Alternatively, we are considering combining our anatomically-driven approach with the novel nnUnet model (Isensee et al., 2021) which has proven to be effective not only for aneurysm detection (it was adopted by 2 of the top-performing teams in the ADAM challenge), but also for several other segmentation tasks. We believe this combination holds potential to boost detection performances. Also, the ablation study performed in the Online Resources – Section F showed that pre-training on the ADAM dataset did not increase detections results. Therefore, future works should investigate a different transfer learning approach to better leverage knowledge acquired from the ADAM dataset. Last, we plan to conduct further error analyses to identify common patterns for both false positive and false negative cases.

In conclusion, our study presented an anatomically-informed 3D UNET that tackles brain aneurysm detection across different sites. The combination of time-saving weak labels and anatomical prior knowledge allowed us to build a robust deep learning model. We believe our approach and dataset (both openly available) can foster the development of clinically applicable automated systems for the task at hand.

## Information Sharing Statement - Data Availability

Our open-access dataset is available on OpenNeuro under the CC0 license at <https://openneuro.org/datasets/ds003949>. The ADAM dataset can be downloaded from the challenge website <https://adam.isi.uu.nl/data/> after signing a confidentiality agreement. The code used for this study is available at [https://github.com/connectomicslab/Aneurysm\\_Detection](https://github.com/connectomicslab/Aneurysm_Detection) under the Apache-2.0 license, together with the configuration files to replicate all the experiments, and the weights of the trained model if users simply want to perform inference.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s12021-022-09597-0>.

**Acknowledgements** We would like to thank the organizing team of the ADAM challenge for their great effort and availability.

**Funding** Open access funding provided by University of Lausanne.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abousamra, S., Fassler, D., Hou, L., Zhang, Y., Gupta, R., Kurc, T., Escobar-Hoyos, L. F., Samaras, D., Knudson, B., Shroyer, K., Saltz, J., & Chen, C. (2020). Weakly-supervised deep stain decomposition for multiplex IHC images. *Proceedings - International Symposium on Biomedical Imaging*, 481–485. <https://doi.org/10.1109/ISBI45749.2020.9098652>
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: a next-generation hyperparameter optimization framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/3292500.3330701>
- Arimura, H., Li, Q., Korogi, Y., Hirai, T., & Abe, H. (2004). Automated computerized scheme for detection of unruptured intracranial aneurysms in three-dimensional magnetic resonance angiography 1. *Academic Radiology*. <https://doi.org/10.1016/j.acra.2004.07.011>
- Avants, B. B., Tustison, N., & Johnson, H. (2014). Advanced Normalization Tools (ANTS). *Insight J*, 2(365), 1–35. <https://brianavants.wordpress.com/2012/04/13/updated-ants-compile-instructions-april-12-2012/>. Accessed January 2021.
- Baumgartner, M., Jäger, P. F., Isensee, F., & Maier-Hein, K. H. (2021). nnDetection: a self-configuring method for medical object detection. *MICCAI*. <https://github.com/MIC-DKFZ/nnDetection>. Accessed July 2021.
- Bengio, Y., Goodfellow, I., & Courville, A. (2016). Deep learning. *MIT Press*, 29(7553).
- Brown, R. D., & Broderick, J. P. (2014). Unruptured intracranial aneurysms: Epidemiology, natural history, management options, and familial screening. *The Lancet Neurology*, 13(4), 393–404. [https://doi.org/10.1016/S1474-4422\(14\)70015-8](https://doi.org/10.1016/S1474-4422(14)70015-8)
- Chakraborty, D. P., & Berbaum, K. S. (2004). Observer studies involving detection and localization: Modeling, analysis, and validation. *Medical Physics*, 31(8), 2313–2330. <https://doi.org/10.1118/1.1769352>
- Chen, X., Liu, Y., Tong, H., Dong, Y., Ma, D., Xu, L., & Yang, C. (2018). Meta-analysis of computed tomography angiography versus magnetic resonance angiography for intracranial aneurysm. *Medicine (United States)*, 97(20). <https://doi.org/10.1097/MD.0000000000010771>
- Dai, X., Huang, L., Qian, Y., Xia, S., Chong, W., Liu, J., Di Ieva, A., Hou, X., & Ou, C. (2020). Deep learning for automated cerebral aneurysm detection on computed tomography images. *International Journal of Computer Assisted Radiology and Surgery*, 15(4), 715–723. <https://doi.org/10.1007/s11548-020-02121-2>
- Di Noto, T., Marie, G., Tourbier, S., Alemán-Gómez, Y., Saliou, G., Cuadra, M. B., Hagmann, P., & Richiardi, J. (2020). An anatomically-informed 3D CNN for brain aneurysm classification with weak labels. *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology*. <http://arxiv.org/abs/2012.08645>. Accessed January 2021.
- Duan, H., Huang, Y., Liu, L., Dai, H., Chen, L., & Zhou, L. (2019). Automatic detection on intracranial aneurysm from digital subtraction angiography with cascade convolutional neural networks.









- BioMedical Engineering Online*, 18(1). <https://doi.org/10.1186/s12938-019-0726-2>
- Ezhov, M., Zakirov, A., & Gusarev, M. (2019). Coarse-to-fine volumetric segmentation of teeth in cone-beam CT. *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*.
- Frösen, J., Tulamo, R., Paetau, A., Laaksamo, E., Korja, M., Laakso, A., Niemelä, M., & Hernesniemi, J. (2012). Saccular intracranial aneurysm: Pathology and mechanisms. *Acta Neuropathologica*, 123(6), 773–786. <https://doi.org/10.1007/s00401-011-0939-3>
- Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 9, 249–256.
- Gorgolewski, K. J. (2008). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*. [https://doi.org/10.1007/978-1-4020-6754-9\\_1720](https://doi.org/10.1007/978-1-4020-6754-9_1720)
- Greving, J. P., Wermer, M. J. H., Brown, R. D., Morita, A., Juvela, S., Yonekura, M., Ishibashi, T., Torner, J. C., Nakayama, T., Rinkel, G. J. E., & Algra, A. (2014). Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: A pooled analysis of six prospective cohort studies. *The Lancet Neurology*, 13(1), 59–66. [https://doi.org/10.1016/S1474-4422\(13\)70263-1](https://doi.org/10.1016/S1474-4422(13)70263-1)
- Hainc, N., Mannil, M., Anagnostakou, V., Alkadhi, H., Blüthgen, C., Wacht, L., Bink, A., Husain, S., Kulcsár, Z., & Winklhofer, S. (2020). Deep learning based detection of intracranial aneurysms on digital subtraction angiography: A feasibility study. *Neuroradiology Journal*, 33(4), 311–317. <https://doi.org/10.1177/1971400920937647>
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning*. PMLR, 2015.
- Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. <https://doi.org/10.1038/s41592-020-01008-z>
- Joo, B., Ahn, S. S., Yoon, P. H., Bae, S., Sohn, B., Lee, Y. E., Bae, J. H., Park, M. S., Choi, H. S., & Lee, S. K. (2020). A deep learning algorithm may automate intracranial aneurysm detection on MR angiography with high diagnostic performance. *European Radiology*, 30(11), 5785–5793. <https://doi.org/10.1007/s00330-020-06966-8>
- Ke, R., Bugeau, A., Papadakis, N., Schuetz, P., & Schönlieb, C.-B. (2020). Learning to segment microscopy images with lazy labels. *ArXiv*. [https://doi.org/10.1007/978-3-030-66415-2\\_27](https://doi.org/10.1007/978-3-030-66415-2_27)
- Keedy, A. (2006). An overview of intracranial aneurysms. *McGill Journal of Medicine: MJM*, 9(2).
- Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.
- Liu, X., Feng, J., Wu, Z., Neo, Z., Zhu, C., Zhang, P., Wang, Y., Jiang, Y., Mitsouras, D., & Li, Y. (2021). Deep neural network-based detection and segmentation of intracranial aneurysms on 3D rotational DSA. *Interventional Neuroradiology*. <https://doi.org/10.1177/15910199211000956>
- Markiewicz, C. J., Gorgolewski, K. J., Feingold, F., Blair, R., Halchenko, Y. O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncalves, M., Jwa, A., & Poldrack, R. A. (2021). OpenNeuro: An open resource for sharing of neuroimaging data. *BioRxiv*. <https://doi.org/10.1101/2021.06.28.450168>
- McHugh, M. L. (2012). The chi-square test of independence. *Biochemistry Medica*, 23(2), 143–149. <https://doi.org/10.11613/BM.2013.018>
- Mouches, P., & Forkert, N. D. (2014). A statistical atlas of cerebral arteries generated using multi-center MRA datasets from healthy subjects. *Scientific Data*, 6(1), 1–8. <https://doi.org/10.1038/s41597-019-0034-5>
- Nakao, T., Hanaoka, S., Nomura, Y., Sato, I., Nemoto, M., Miki, S., Maeda, E., Yoshikawa, T., Hayashi, N., & Abe, O. (2018). Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. *Journal of Magnetic Resonance Imaging*, 47(4), 948–953. <https://doi.org/10.1002/jmri.25842>
- Özgül, Ç., Abdulkadir, A., Lienkamp, S., Brox, T., & Ronneberg, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. *ArXiv*. <https://doi.org/10.1007/978-3-319-46723-8>
- Park, A., Chute, C., Rajpurkar, P., Lou, J., Ball, R. L., Shpanskaya, K., Jabarkheel, R., Kim, L. H., McKenna, E., Tseng, J., Ni, J., Wishah, F., Wittber, F., Hong, D. S., Wilson, T. J., Halabi, S., Basu, S., Patel, B. N., Lungren, M. P., & Yeom, K. W. (2019). Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA Network Open*, 2(6), e195600. <https://doi.org/10.1001/jamanetworkopen.2019.5600>
- Rao, B., Zohrabian, V., Cedeno, P., Saha, A., Pahade, J., & Davis, M. A. (2021). Utility of artificial intelligence tool as a prospective radiology peer reviewer — detection of unreported intracranial hemorrhage. *Academic Radiology*, 28(1), 85–93. <https://doi.org/10.1016/j.acra.2020.01.035>
- Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. *Lecture Notes in Computational Vision and Biomechanics*, 26, 323–350. [https://doi.org/10.1007/978-3-319-65981-7\\_12](https://doi.org/10.1007/978-3-319-65981-7_12)
- Shi, Z., Miao, C., Schoepf, U. J., Savage, R. H., Dargis, D. M., Pan, C., Chai, X., Li, X. L., Xia, S., Zhang, X., Gu, Y., Zhang, Y., Hu, B., Xu, W., Zhou, C., Luo, S., Wang, H., Mao, L., Liang, K., & Zhang, L. J. (2020). A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. *Nature Communications*. <https://doi.org/10.1038/s41467-020-19527-w>
- Sichtermann, T., Faron, A., Sijben, R., Teichert, N., Freiherr, J., & Wiesmann, M. (2019). Deep learning – based detection of intracranial aneurysms in 3D TOF-MRA. *American Journal of Neuroradiology*. <https://doi.org/10.3174/ajnr.A5911>
- Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, 17(3), 143–155. <https://doi.org/10.1002/hbm.10062>
- Stember, J. N., Chang, P., Stember, D. M., Liu, M., Grinband, J., Filippi, C. G., Meyers, P., & Jambawalikar, S. (2019). Convolutional neural networks for the detection and measurement of cerebral aneurysms on magnetic resonance angiography. *Journal of Digital Imaging*, 32(5), 808–815. <https://doi.org/10.1007/s10278-018-0162-z>
- Taghanaki, S. A., Zheng, Y., Kevin Zhou, S., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., & Hamarneh, G. (2019). Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75, 24–33. <https://doi.org/10.1016/j.compmedimag.2019.04.005>
- Timmins, K. M., van der Schaaf, I. C., Bennink, E., Ruigrok, Y. M., An, X., Baumgartner, M., Bourdon, P., De Feo, R., Noto, T., Di Dubost, F., Fava-Sanches, A., Feng, X., Giroud, C., Group, I., Hu, M., Jaeger, P. F., Kaiponen, J., Klimont, M., Li, Y., & Kuijf, H. J. (2021). Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge. *NeuroImage*, 238, 118216. <https://doi.org/10.1016/j.neuroimage.2021.118216>
- Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6), 1310–1320. <https://doi.org/10.1109/TMI.2010.2046908>

- Ueda, D., Doishita, S., & Choppin, A. (2019). Deep learning for MR angiography : automated detection of cerebral aneurysms. *Radiology*. <https://doi.org/10.1148/radiol.2018180901>
- Ward, J., Naik, K. S., Guthrie, F. J. A., Wilson, D., & Robinson, P. J. (1999). Hepatic lesion detection: comparison of MR imaging after the administration of superparamagnetic iron oxide with dual-phase CT by using alternative-free response receiver operating characteristic analysis 1. *Radiology*. <https://doi.org/10.1148/radiology.210.2.r99fe05459>
- Yang, J., Xie, M., Hu, C., Alwalid, O., Xu, Y., Liu, J., Jin, T., Li, C., Tu, D., Liu, X., Zhang, C., Li, C., & Long, X. (2020). Deep learning for detecting cerebral aneurysms with CT angiography. *Radiology*, 298(1), 155–163. <https://doi.org/10.1148/RADIOL.2020192154>
- Yang, X., Blezek, D. J., Cheng, L. T. E., Ryan, W. J., Kallmes, D. F., & Erickson, B. J. (2011). Computer-aided detection of intracranial aneurysms in MR angiography. *Journal of Digital Imaging*, 24(1), 86–95. <https://doi.org/10.1007/s10278-009-9254-0>
- Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31(3), 1116–1128. <https://doi.org/10.1016/j.neuroimage.2006.01.015>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



# Diagnostic Surveillance of High-Grade Gliomas: Towards Automated Change Detection Using Radiology Report Classification

Tommaso Di Noto<sup>1</sup> , Chirine Atat<sup>1</sup> , Eduardo Gamito Teiga<sup>1</sup> ,  
Monika Hegi<sup>2,3,4</sup> , Andreas Hottinger<sup>5</sup> , Meritxell Bach Cuadra<sup>1,6</sup> ,  
Patric Hagmann<sup>1</sup> , and Jonas Richiardi<sup>1</sup> 

<sup>1</sup> Department of Radiology, Lausanne University Hospital  
and University of Lausanne, Lausanne, Switzerland  
[tommaso.di-noto@chuv.ch](mailto:tommaso.di-noto@chuv.ch)

<sup>2</sup> Neuroscience Research Center, Lausanne University Hospital  
and University of Lausanne, Lausanne, Switzerland

<sup>3</sup> Neurosurgery, Lausanne University Hospital and University of Lausanne,  
Lausanne, Switzerland

<sup>4</sup> Swiss Cancer Center Léman (SCCL), Lausanne, Switzerland

<sup>5</sup> Department of Clinical Neurosciences; Department of Oncology,  
Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

<sup>6</sup> Medical Image Analysis Laboratory, Center for Biomedical Imaging,  
Lausanne, Switzerland

**Abstract.** Natural Language Processing (NLP) on electronic health records (EHRs) can be used to monitor the evolution of pathologies over time to facilitate diagnosis and improve decision-making. In this study, we designed an NLP pipeline to classify Magnetic Resonance Imaging (MRI) radiology reports of patients with high-grade gliomas. Specifically, we aimed to distinguish reports indicating changes in tumors between one examination and the follow-up examination (treatment response/tumor progression versus stability). A total of 164 patients with 361 associated reports were retrieved from routine imaging, and reports were labeled by one radiologist. First, we assessed which embedding is more suitable when working with limited data, in French, from a specific domain. To do so, we compared a classic embedding techniques, TF-IDF, to a neural embedding technique, Doc2Vec, after hyperparameter optimization for both. A random forest classifier was used to classify the reports into stable (unchanged tumor) or unstable (changed tumor). Second, we applied the post-hoc LIME explainability tool to understand the decisions taken by the model. Overall, classification results obtained in repeated 5-fold cross-validation with TF-IDF reached around 89% AUC and were significantly better than those achieved with Doc2Vec (Wilcoxon signed-rank test,  $P = 0.009$ ). The explainability toolkit run on TF-IDF revealed some interesting patterns: first, words indicating change such as *progression* were rightfully frequent for reports classified as unstable; similarly, words indicating no change such as *not* were frequent for reports classified as stable. Lastly,



the toolkit discovered misleading words such as *T2* which are clearly not directly relevant for the task. All the code used for this study is made available.

**Keywords:** Natural Language Processing (NLP) · Term Frequency - Inverse Document Frequency (TF-IDF) · Doc2Vec · Diagnostic surveillance · LIME model explainability

## 1 Introduction

In the last decade, Machine Learning (ML) has reshaped research in radiology. ML models yield state-of-the-art results for numerous medical imaging tasks such as segmentation, anomaly detection, registration, and disease classification [1]. In addition to images, ML models have also been increasingly applied to radiology reports and more generally to data coming from Radiology Information Systems (RIS) [2]. However, even though radiology reports contain valuable, high-level insights from trained physicians, they also come with some associated drawbacks; in particular, most reports are stored as unstructured, free-text documents. Consequently, they exhibit a strong degree of ambiguity, uncertainty and lack of conciseness [3].

Natural Language Processing (NLP) is a branch of ML that helps computers understand, interpret, and manipulate human language [4]. In the case of radiology reports, NLP has the goal of extracting clinically relevant information from unstructured texts. As recently illustrated in one extensive review [5], one frequent application of NLP for radiology reports is diagnostic surveillance. Its objective is to monitor the evolution of a pathology in order to extrapolate useful knowledge and improve decision-making. In line with this trend, our work focuses on oncology patients with high-grade gliomas that are scanned longitudinally for frequent follow-up.

According to [5], the majority (86%) of studies published up until 2019 focused on medical reports written in English, while only 1% of the reviewed studies utilized French reports. This language gap is understandable given that a substantial portion of NLP tools was developed using English texts. Nonetheless, in medical NLP, researchers need to adapt their models to the language of the radiology reports. This entails custom precautions and expedients to take since languages are often syntactically and/or semantically different from English. In this work, we investigate NLP methods for radiology reports written in French.

In addition, [5] concluded that although a growing number of Deep Learning (DL) NLP methods has been applied in recent years, “conventional ML approaches are still prevalent”. To assess which technique is more suitable for our dataset, we compare two traditional embedding strategies, namely Term Frequency-Inverse Document Frequency (TF-IDF) [6] and Doc2Vec [7].

The task that we address is binary document classification. Specifically, we aim to identify the main conclusion of the medical reports deciding among the following groups: tumor *stability* vs. tumor *instability*. Details about these classes

are provided in Sect. 2.2. The potential applications of our report classifier are twofold: first, it could help referring physicians to focus the attention on the main conclusion of the report, thus accelerating subsequent decisions. Second, the predicted classes could be used as weak labels for a downstream machine learning task (e.g. automated cohort creation). In addition, most clinically relevant images in RIS are associated with a radiology report, and thus offer potential access to several hundred thousands of weakly labelled images in medium to large hospitals.

In this work we also conduct an interpretability analysis of the model's decisions [8,9], based on the post-hoc interpretation technique LIME [10]. Its main objective is to identify the most important words that influenced the final prediction, by creating a surrogate linear model that performs local input perturbation (details in Sect. 2.4).

In summary, this study presents a classifier for French radiology reports in the context of diagnostic surveillance, while comparing two embedding techniques and providing a visual interpretation of the model's decisions.

## 1.1 Related Works

Here, we present the works most similar to ours. In [11], the authors compared several embedding techniques and five different classifiers for detecting the radiologist's intent in oncologic evaluations. Similarly, [12] investigated a DL model to identify oncologic outcomes from radiology reports. The authors in [13] utilized a combination of ML and rule-based approaches to highlight important changes and identify significant observations that characterize radiology reports. [14] devised a model that extracts radiological measurements and the corresponding core descriptors (e.g. temporality, anatomical entity, ...) from Magnetic Resonance (MR), Computed Tomography (CT) and mammography reports. The work of [15] describes an NLP pipeline that identifies patients with (pre)cancer of the cervix and anus from histopathologic reports. Last, [16] detected thromboembolic diseases and incidental findings from angiography and venography reports.

Among all these works, only [16] used French reports, while the others worked with English documents. Moreover, only [12] addressed the issue of model explainability which we believe is paramount for the ML community, especially in the medical domain.

## 2 Materials and Methods

### 2.1 Dataset

We retrospectively included 164 subjects that underwent longitudinal MR glioma follow-up in the university hospital of Lausanne (CHUV) between 2005 and 2019. 71% of the patients in the cohort had Glioblastoma Multiforme (GBM), while the remaining 29% had either an oligoastrocytoma or an oligodendroglioma. At

every session, a series of MR scans were performed including structural, perfusion and functional imaging. For the sake of this study, we only focused on the native T1-weighted (T1w) scan, the T2-weighted (T2w) scan and the T1w-gad (post gadolinium injection, a contrast agent). For 25 patients, we collected images and reports across multiple sessions (on average, 9 sessions per subject). For the remaining 139 patients, we only retrieved images and reports from 1 random session. This latter sampling strategy was adopted to increase the chance of having cases of tumor progression and tumor response, since multiple sessions of the same subject mostly showed tumor stability and thus led to a very imbalanced data set. Overall, we ended up with a dataset of 361 radiology reports to use for the NLP pipeline. Every report was written (dictated) during routine clinical practice by a junior radiologist after exploring all sequences of interest. Then, a senior radiologist reviewed each case amending the final report when necessary. The extracted reports have varying length ranging from 114 to 533 words (average 255, standard deviation 68). The MR acquisition parameters for the cohort are provided in Table 1. The protocol of this study was approved by the regional ethics committee; written informed consent was waived.

**Table 1.** MR acquisition parameters of scans used for the study population.

# sessions $\equiv$ # reports	Vendor	Scanner	Field strength [T]
174	Siemens Healthcare	Skyra	3.0
73	Philips	Intera	3.0
46	Siemens Healthcare	Prisma	3.0
32	Siemens Healthcare	Symphony	1.5
21	Siemens Healthcare	TrioTim	3.0
10	Siemens Healthcare	Aera	1.5
5	Siemens Healthcare	Verio	3.0

## 2.2 Report Tagging

In order to build a supervised document classifier, one radiologist (4 years of experience in neuroimaging) tagged the reports with labels of interest. For each report, the annotator was instructed to perform two separate tasks: first, she had to assign 3 classes to the reports; one class that indicated the global conclusion of the report, one class to indicate the evolution of the enhanced part of the lesion (T1w conclusion) and the last one to indicate the evolution of the lesion on T2-weighted sequences (T2w conclusion). For each of these three groups, the annotator could choose between the following labels:

- **Stable:** assigned when the tumor did not change significantly with respect to the previous comparative exam.

- **Progression**: assigned when the tumor worsened with respect to the previous comparative exam. This class included cases where the enhanced part of the tumor increased in size or when the T2 signal anomalies surrounding the tumor increased in extension.
- **Response**: assigned when the tumor responded positively to the treatment (either chemotherapy or radiotherapy).
- **Unknown**: used when the annotator was not able to assign any of the three classes above.

The second task of the annotator was to highlight the most recent comparative date in the reports. Since the reports are not structured, this helped linking the current report being tagged with the most meaningful previous one. For simplicity, in this work we only focused on the global conclusion of the reports, and not on the T1 and T2 conclusions. Also, we removed all cases that were tagged as **unknown** (21 reports) and we merged **progression** and **response** into one unique class which we denote as **unstable**. By doing this, we narrowed the task to a binary classification problem where the model tries to distinguish between **stable** and **unstable** reports. After these modifications, we ended up with 191 **stable** reports and 149 **unstable** reports.

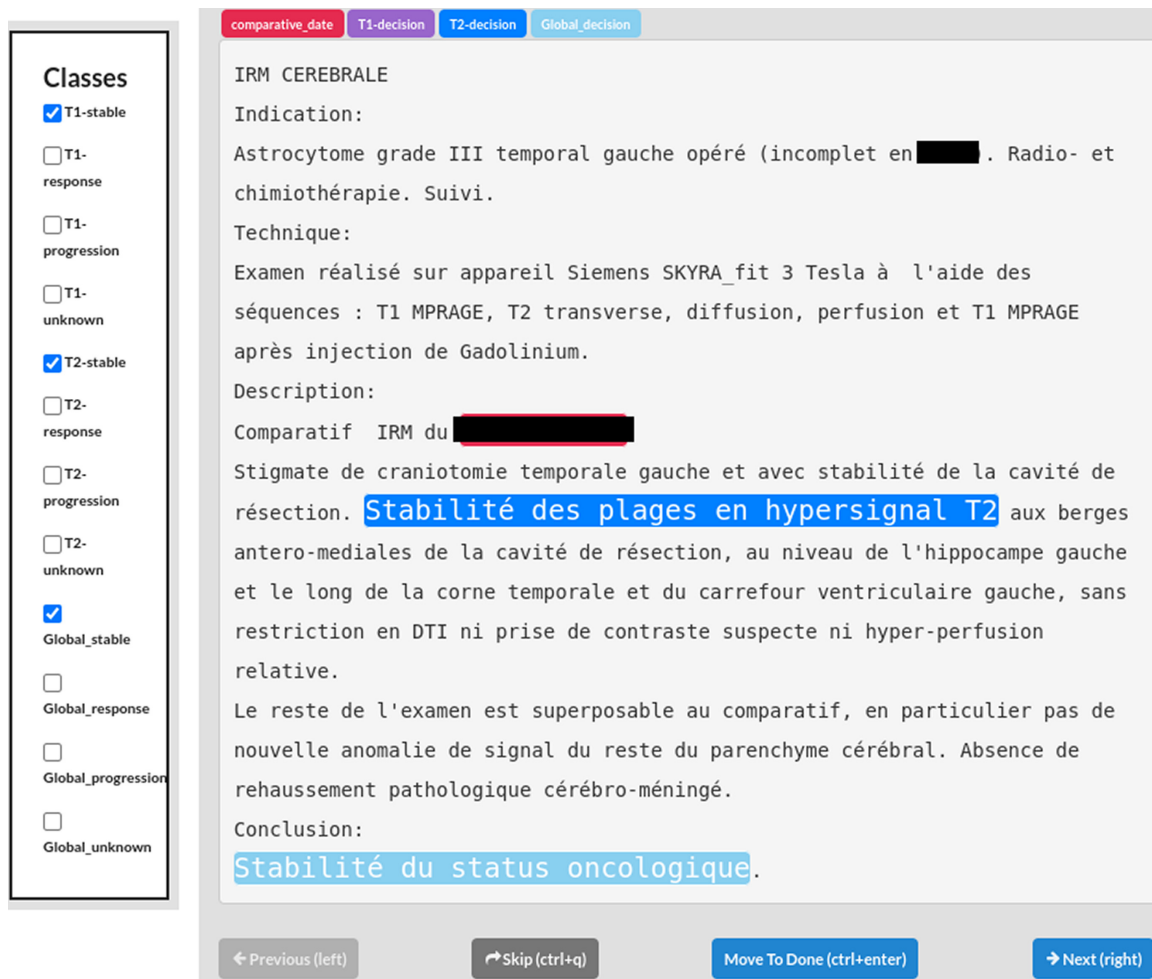
To facilitate the annotation process, we utilized the open-source software *Dataturks*<sup>1</sup>. This provided a graphic interface to the annotator which allowed her to tag, skip, highlight, and review the reports in a user-friendly way. Moreover, it automatically generated machine-readable labels once the annotation process was over. One exemplary report is illustrated in Fig. 1, together with the corresponding annotations.

### 2.3 Text Preprocessing and Embedding

Several preprocessing steps were carried out to reduce the vocabulary size. First, we removed all proper nouns such as physicians' and patients' names. This was performed using a pre-trained French Part-Of-Speech tagger from the Spacy library (version 3.0.6) [19]. Second, all the words in the reports were converted to lowercase. This operation is typical when there are no words that indicate a specific meaning when expressed with capital letters. Third, we removed punctuation and the most common French stop words, namely ['de', 'la', 'en', 'et', 'du', 'd', 'le', 'l', 'un', 'une', 'les', 'des', 'ces', 'á', 'au', 'aux']. Among these, we ensured to keep the French negation 'pas' (*not*) since it is very frequent in the reports, and reverses the meaning of the sentence. Fourth, all reports were tokenized using the *wordpunct* class of the Natural Language Toolkit framework (version 3.6.1) [20]. As last step, since all the reports contain the three sections '*indications*', '*description*' and '*conclusion*', we removed all content before the '*indication*' section, which is either useless (e.g. department phone number) or sensitive (e.g. patient identifier).

---

<sup>1</sup> OpenSource Data Annotation tool - <http://github.com/DataTurks/DataTurks>.



**Fig. 1.** Daturks annotation interface. The annotator can select the classes in the left box and highlight the text of interest. Sensitive information has been blacked out for privacy.

A key step in any NLP pipeline is text embedding. This corresponds to the conversion of tokenized text into numerical vectors. Historically, many embedding techniques have been proposed in literature. In this work, we compare two of the most widespread, namely TF-IDF [6] and Doc2Vec [7]. While the former is a standard term-weighting embedding scheme (traditional ML) that preserves the length of the tokenized documents, the latter is a DL-based technique that creates dense vectors which encode word order and context. TF-IDF was performed at the word level with the sklearn package (version 0.24.1) [21], whereas Doc2Vec was performed using the gensim library (version 4.0.1) [22].

## 2.4 Experiments

All experiments were run in a 5-fold, nested, stratified cross validation (CV). The internal CV was used to tune the hyperparameters of the pipeline with a custom Grid Search algorithm. Instead, the external CV was used to compute results on hold-out test samples. For TF-IDF, two hyperparameters were tuned:

first, the types of retained N-grams were searched in the range [3,5]. Second, the percentage of vocabulary size to use was varied between 100% (all words are used) and 90% (the 10% rarest words are removed). The other parameters were fixed: the minimum document frequency was set to 2 and the maximum document frequency was set to 0.9 (indicating 90% of the documents). For Doc2Vec, the algorithm type (PV-DM or PV-DBOW) and the vector dimensionality [10] were tuned with the validation set. The context window was set to 5 words. Five “noise” negative words were drawn. Words with a total frequency lower than 2 were ignored. The model was trained for 100 epochs. Since stop words are not necessarily useless for Doc2Vec, we also tried to run the Doc2Vec pipeline preserving them.

The stratification of the CV guaranteed that both training and test sets contained approximately the same percentage of reports indicating tumor **stability** and tumor **instability**. To avoid overoptimistic predictions, we also ensured that the reports from multiple sessions of the same subject were not present some in the train set and some in the test set. Furthermore, to reduce the bias introduced by the random choice of patients at each CV split, the whole nested CV was repeated 10 times, each time performing the splitting anew, and results were averaged.

For all experiments, we adopted the Random Forest algorithm [23] to classify the embedded documents, using once again the sklearn package. As hyperparameters, we set a fixed number of 501 trees and we tuned the maximum retained features in the internal CV, choosing between 0.8 (only 80% of the features are used) and 1.0 (all features are used).

To compare the two pipelines (Doc2Vec vs. TF-IDF embedding), we computed all standard classification metrics, namely accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score. Moreover, we also plotted the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. The reports indicating tumor stability were considered as negative samples, whereas those indicating a change in the tumor were considered as positive samples. The classification metrics and the curves were averaged across the 10 runs. To statistically compare the classification results, a Wilcoxon signed-rank test was performed [24]. For simplicity, the test only accounted for the area under the ROC curve (AUC) across the 10 runs. A significance threshold level  $\alpha = 0.05$  was set for comparing P values.

The explainability analysis was performed with the LIME toolkit on the TF-IDF pipeline only since it resulted in higher performances (see Table 2). We set the best hyperparameters obtained across the random runs and we ran LIME over all test reports. For each report, the toolkit performs a post-hoc interpretation following a two-step approach: first, it randomly generates neighborhood data in the vicinity of the example being explained; then, it “learns locally weighted linear models on this neighborhood data to explain each of the classes in an interpretable way”. The user can choose how many features (words) are shown in the explanation. For this work, we set a maximum of 6 features per document. These weighted features represent the linear model which approximates

the behaviour of the random forest classifier in the vicinity of the explained test example.

All the Python 3.6 code developed for this study is available on [github](https://github.com)<sup>2</sup>.

### 3 Results

#### 3.1 Classification Performances

The nested CV with the Doc2Vec embedding took 50 min per run, while the one with TF-IDF took 2 h. The most frequent hyperparameters chosen in the internal CV for Doc2Vec across the 10 random runs were a vector size of 10 and the PV-DV version of the algorithm. Instead, for TF-IDF, n-grams in the range (1, 3) were the most frequent, and the optimal percentage of vocabulary size was 90%. For the Random Forest classifier, the configuration with 80% of the features was most frequent.

We report in Table 2 the classification results of the two pipelines (TF-IDF vs. Doc2Vec), averaged over the 10 runs. Similarly, Figs. 2 and 3 illustrate the average ROC and PR curves. When comparing the two pipelines across the 10 random runs with the Wilcoxon signed-rank test, the AUC values of TF-IDF were significantly higher than those of Doc2Vec ( $P = 0.009$ ). Last, classification results of the Doc2Vec pipeline run preserving the stop words led to higher results (average AUC =  $.85 \pm .03$ ). However, these were still significantly lower than the TF-IDF pipeline.

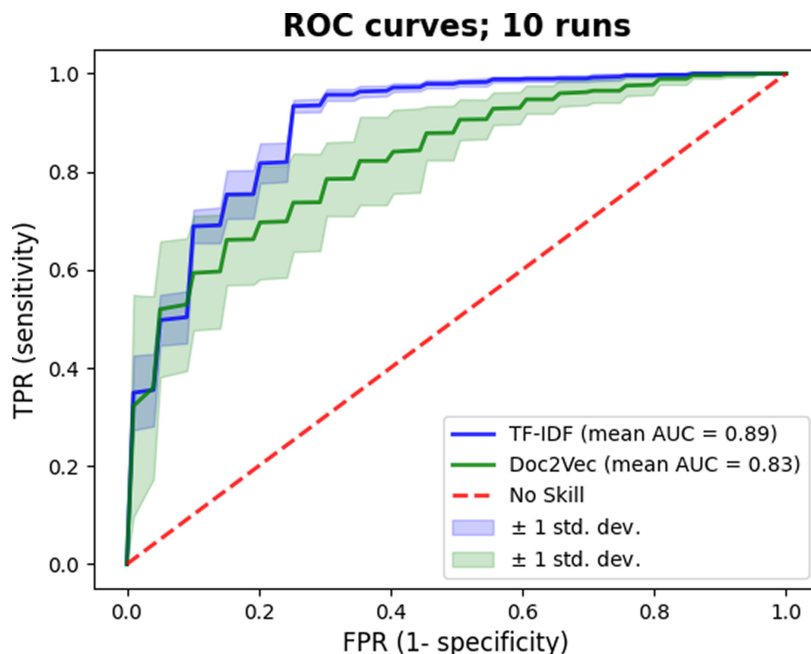
**Table 2.** Classification results across the 10 random runs. Values are presented as mean  $\pm$  standard deviation. Bold values indicate the highest performances. Acc = accuracy; Sens = sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value; F1 = F1-score; AUC = area under the ROC curve; AUPR = area under the PR curve.

Embedding	Acc %	Sens %	Spec %	PPV %	NPV %	F1 %	AUC	AUPR
TF-IDF	<b>88</b> $\pm$ 1	91 $\pm$ 1	<b>75</b> $\pm$ 0	<b>95</b> $\pm$ 0	<b>60</b> $\pm$ 2	<b>93</b> $\pm$ 0	<b>.89</b> $\pm$ .01	<b>.97</b> $\pm$ .00
Doc2Vec	86 $\pm$ 2	<b>94</b> $\pm$ 3	38 $\pm$ 10	89 $\pm$ 1	57 $\pm$ 10	92 $\pm$ 1	.83 $\pm$ .05	.96 $\pm$ .01

#### 3.2 Error Analysis and Model Interpretation

To further understand the decisions taken by the random forest algorithm, we applied the LIME post-hoc interpretability toolkit. Specifically, we investigated both the explanations created for the correctly classified reports and for the false positive and false negative reports. Table 3 shows the most frequent words used by the linear classifier created by LIME. We notice that most of the words intuitively make sense for the True Positive and True Negative samples. For instance, words like ‘*progression*’, ‘*augmentation*’ and ‘*diminution*’ that all indicate some

<sup>2</sup> [https://github.com/connectomicslab/Glioma\\_NLP](https://github.com/connectomicslab/Glioma_NLP).



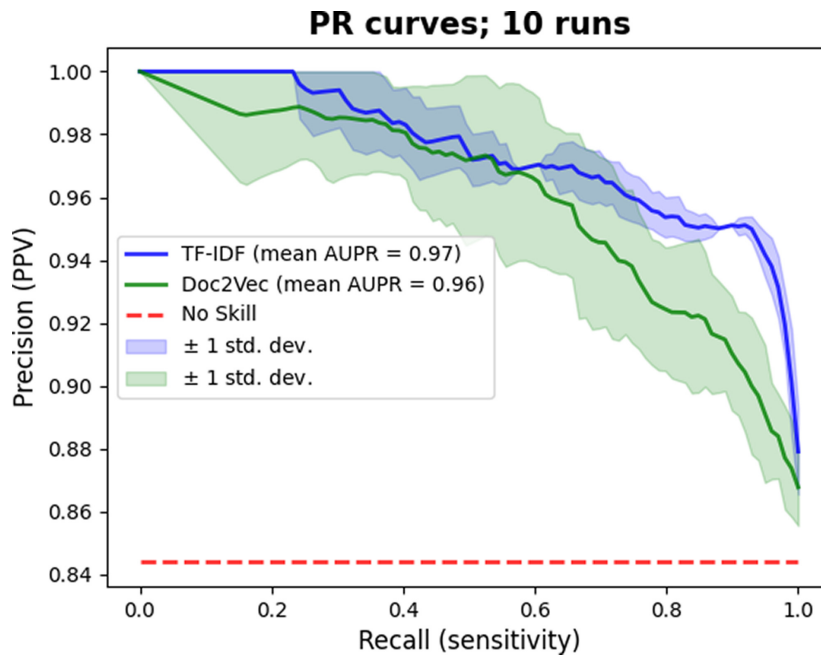
**Fig. 2.** Receiver operating characteristic (ROC) curves of the two pipelines (TF-IDF vs. Doc2Vec) averaged across the 10 runs.

sort of change are recurrent for predicting TP samples and outweigh the corresponding words indicating tumor stability such as ‘*sans*’ (*without*) or ‘*récidive*’ (*recurrence*). A similar trend can be observed for TN samples where words like ‘*pas*’ (*not*), ‘*stabilité*’ (*stability*) and ‘*inchangé*’ (*unchanged*) outweigh words indicating instability like ‘*apparition*’ (*appearance*). However, the error analysis also highlighted some recurrent mistakes, such as the importance given to the words ‘*t2*’ and ‘*axial*’ in the FN samples or ‘*2007*’ in the FP which ultimately deteriorate the predictions. To have a qualitative idea of the output of the LIME toolkit, we show in Figs. 4 and 5 one TP and one FN example, respectively.

## 4 Discussion

In this work, we explored the potential of NLP for the task of diagnostic surveillance in patients with high-grade gliomas. As pointed out in [5], and subsequently shown in other works [25,26], traditional ML embedding techniques can lead to comparable results with respect to DL techniques when properly tuned. Moreover, they are still frequent when the dataset size is limited such as in medical imaging applications. Our work confirms this trend since, given the same classifier, the TF-IDF pipeline statistically outperformed the Doc2Vec one. The explainability analysis highlighted interesting trends. For the correctly classified reports, it confirmed that the model is focusing on relevant words.





**Fig. 3.** Precision-Recall (PR) curves of the two pipelines (TF-IDF vs. Doc2Vec) averaged across the 10 runs.

When investigating reports indicating instability, most of the recurrent terms indeed indicate a status of change such as ‘*diminution*’, ‘*progression*’ or ‘*plus*’ (*more*). Similarly, the recurrent words for the reports indicating tumor stability reflect a status of no-change (e.g. ‘*pas*’ (French negation)). Regarding the errors of the model, the LIME toolkit also uncovered some misleading words which obfuscate the final predictions. For instance, the words ‘*appareil*’ (*MR scanner*), ‘*t2*’, ‘*axial*’ or ‘*transverse*’ are recurrent in the explanations of FP and FN even though they are related to the acquisition process rather the status of the tumor.

The following limitations must be acknowledged. First, the annotations were performed by one single radiologist which is not the optimal scenario for ambiguous NLP tasks. Second, the dataset size is still limited with respect to similar studies [11, 12, 14].

In future works we are planning to enlarge the dataset and add a second annotator to assess inter-rater variability (and ideally intra-rater variability as well). Also, we would like to investigate which part of the report is the most important with respect to the final prediction. For instance, we would like to evaluate classification performances when using only *description* and *conclusion* of the reports, or even just the *conclusion*. In addition, we are planning to

**Table 3.** Six most frequent features (words) used by the linear model generated by LIME to predict the class of the reports, sorted in descending order. For instance, the word ‘*progression*’ is the most frequent word indicating instability used by the linear classifier for the TP test documents, whereas ‘*pas*’ (French negation) is the most frequent word indicating stability used for the TN test documents. TP = True Positive (i.e. reports indicating tumor instability and predicted as such); TN = True Negative; FP = False Positive; FN = False Negative.

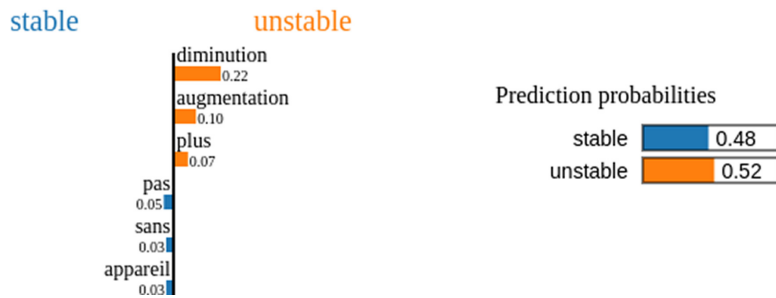
	Stable	Unstable		Stable	Unstable
TP	sans	progression	FP	sans	progression
	récidive	augmentation		depuis	axial
TN	pas	oedème	appareil	diminution	
	signe	plus	réalisé	plus	
	anomalie	diminution	inchangé	oedème	
	ou	spectroscopie	pondération	2007	
	pas	apparition	FN	récidive	apparition
	récidive	augmentation		pas	spectroscopie
sans	axial	sans		augmentation	
stabilité	spectroscopie	transverse		diminution	
transverse	plus	t2	axial		
inchangé	postérieure	stabilité	dans		

experiment different classifiers, or French pre-trained embedding models developed with larger corpora. Next, we will investigate what happens when shifting from a binary problem (*stable* vs. *unstable*) to a more granular task. Last, we will leverage the information extracted by the explainability toolkit to further preprocess the documents, for instance removing terms related to the acquisition protocol.

In conclusion, this work presented an NLP pipeline for the classification of radiology reports for patients with high-grade gliomas. The top-performing model (TF-IDF + Random Forest) attained satisfactory performances (AUC = .89) that lays a good foundation for generating weak labels, and the post-hoc explainability toolkit that we used holds promise for the development of a robust and transparent ML analysis.

**Text with highlighted words**

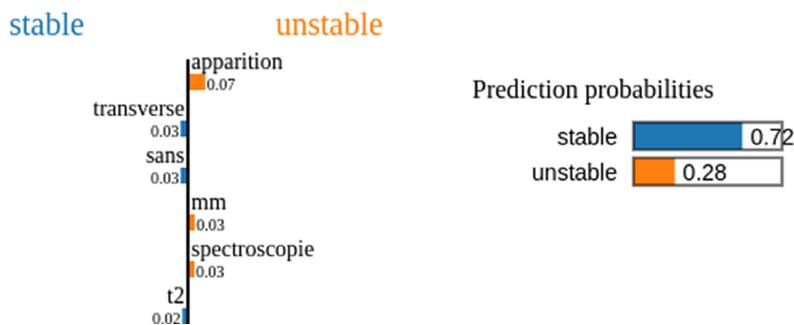
glioblastome fronto pariétal gauche avec status post traitement par radiothérapie bilan 5 mois post arrêt traitement description examen réalisé sur appareil 3 aide séquences t1 sagittales t2 axiales perfusion gadolinium t1 sagittales axiales ainsi que spectroscopie comparatif irm [redacted] on retrouve status post craniotomie fronto pariétale gauche avec cavité exérèse hypersignal t2 bordée hémossidérine sans modification pas modification non plus hypersignal t2 prédominant niveau corona radiata centre semi ovale gauche ainsi qu rétro atrial bilatéral partie lié radiothérapie également sans modification persistance rehaussement prédominant dans portion inférieure interne cavité résection très discrète diminution par rapport comparatif mois [redacted] séries perfusions qualité sub optimales ne permettant pas interprétation séries spectroscopies montrent discrète augmentation choline associée baisse retrouve quelques hyper intensités signal substance blanche bi hémisphériques nature aspécifique dégénérescence xantho granulomateuse plexus choroides citerne optochiasmatique niveau loge sellaire muqueux cadre sinus maxillaire droit sphénoïde droit conclusions discrète diminution prises contraste localisées face inférieure interne cavité exérèse glioblastome fronto pariétal gauche pas autre changement significatif [redacted]



**Fig. 4.** LIME toolkit explanations for a TP report. Words such as ‘*diminution*’ and ‘*augmentation*’ correctly outweigh words indicating stability like ‘*pas*’ (French negation) or ‘*sans*’ (*without*). Sensitive information has been blacked out for privacy.

**Text with highlighted words**

suivi évolution glioblastome réséqué [redacted] traité par radio chimiothérapie adjuvante technique examen réalisé sur appareil 3 t avec séquences t1 sagittale t2 transverse dti injection gadolinium suivie perfusion séquences t1 sagittale transverse spectroscopie description examen comparatif [redacted] status post thérapeutique avec résection tumorale frontale droite sans modification taille aspect cavité résection t2 frontal droit extension inchangée pas modification également extension hypersignal frontal controlatéral série injectée démontre apparition nouvelle prise contraste nodulaire frontale antérieure droite arrière partie antéro interne cavité résection tumorale mesurant 10x4 mm allure suspecte surveiller par ailleurs persistance dilatation ventriculaire modérée ainsi que déformation corne frontale droite secondaire status postopératoire sans évolution persistance séquelle hémorragique arrière corne postérieure ventricule latéral droit inchangée conclusions apparition nouvelle prise contraste nodulaire arrière partie antéro interne cavité résection frontale droite suspecte surveiller associé [redacted]



**Fig. 5.** LIME toolkit explanations for a FN report. Words such as ‘*sans*’ and ‘*transverse*’ incorrectly outweigh the key word indicating instability in this report which is ‘*apparition*’ (*appearance*). Sensitive information has been blacked out for privacy.

## References

1. Shen, D., Wu, G., Suk, H.-I.: Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017)
2. Lakhani, P., et al.: Machine learning in radiology: applications beyond image interpretation. *J. Am. Coll. Radiol.* **15**(2), 350–359 (2018)
3. Schwartz, L.H., et al.: Improving communication of diagnostic radiology findings through structured reporting. *Radiology* **260**(1), 174–181 (2011)
4. Chowdhury, G.G.: Natural language processing. *Annu. Rev. Inf. Sci. Technol.* **37**(1), 51–89 (2003)
5. Casey, A., et al.: A Systematic Review of Natural Language Processing Applied to Radiology Reports. arXiv preprint [arXiv:2102.09553](https://arxiv.org/abs/2102.09553) (2021)
6. Sammut, C., Webb, G.I. (eds.): *Encyclopedia of Machine Learning*. Springer, Boston (2011). <https://doi.org/10.1007/978-0-387-30164-8>
7. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: *International Conference on Machine Learning*. PMLR (2014)
8. Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* **16**(3), 31–57 (2018)
9. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint [arXiv:1702.08608](https://arxiv.org/abs/1702.08608) (2017)
10. Ribeiro, M.T., Singh, S., Guestrin, C.: “Why should I trust you?” Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016)
11. Chen, P.-H., et al.: Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *J. Digit. Imaging* **31**(2), 178–184 (2018)
12. Kehl, K.L., et al.: Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncol.* **5**(10), 1421–1429 (2019)
13. Hassanpour, S., Bay, G., Langlotz, C.P.: Characterization of change and significance for clinical findings in radiology reports through natural language processing. *J. Digit. Imaging* **30**(3), 314–322 (2017)
14. Bozkurt, S., et al.: Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *J. Digit. Imaging* **32**(4), 544–553 (2019)
15. Oliveira, C.R., et al.: Natural language processing for surveillance of cervical and anal cancer and precancer: algorithm development and split-validation study. *JMIR Med. Inform.* **8**(11), e20826 (2020)
16. Pham, A.-D., et al.: Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC Bioinform.* **15**(1), 1–10 (2014)
17. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. arXiv preprint [arXiv:cmp-lg/9602004](https://arxiv.org/abs/cmp-lg/9602004) (1996)
18. Gwet, K.L.: *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC (2014)
19. Honnibal, M., Montani, I., et al.: spaCy: industrial-strength natural language processing in Python. Zenodo (2020). <https://doi.org/10.5281/zenodo.1212303>
20. Bird, S., Klein, E., Loper, E.: *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, Inc., Sebastopol (2009)
21. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)

22. Rehurek, R., Sojka, P.: Gensim-python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3.2 (2011)
23. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
24. Wilcoxon, F.: Individual comparisons by ranking methods. In: Kotz, S., Johnson, N.L. (eds.) *Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics)*, pp. 196–202. Springer, New York (1992). [https://doi.org/10.1007/978-1-4612-4380-9\\_16](https://doi.org/10.1007/978-1-4612-4380-9_16)
25. Dessi, D., et al.: TF-IDF vs word embeddings for morbidity identification in clinical notes: an initial study. arXiv preprint [arXiv:2105.09632](https://arxiv.org/abs/2105.09632) (2021)
26. Marcińczuk, M., et al.: Text document clustering: wordnet vs. TF-IDF vs. word embeddings. In: *Proceedings of the 11th Global Wordnet Conference* (2021)

# Transfer learning with weak labels from radiology reports: application to glioma change detection<sup>1</sup>

Tommaso Di Noto, Meritxell Bach Cuadra, Chirine Atat, Eduardo Gamito Teiga, Monika Hegi, Andreas F. Hottinger, Patric Hagmann, Jonas Richiardi

**Abstract**— Creating large annotated datasets represents a major bottleneck for the development of deep learning models in radiology. To overcome this, we propose a combined use of weak labels (imprecise, but fast-to-create annotations) and Transfer Learning (TL). Specifically, we explore *inductive* TL, where source and target domains are identical, but tasks are different due to a label shift: our target labels are created manually by three radiologists, whereas the source weak labels are generated automatically from textual radiology reports. We frame knowledge transfer as hyperparameter optimization, thus avoiding heuristic choices that are frequent in related works. We investigate the relationship between model size and TL, comparing a low-capacity VGG with a higher-capacity SEResNeXt. The task that we address is change detection in follow-up glioma imaging: we extracted 1693 T2-weighted magnetic resonance imaging difference maps from 183 patients, and classified them into stable or unstable according to tumor evolution. Weak labeling allowed us to increase dataset size more than 3-fold, and improve VGG classification results from 75% to 82% Area Under the ROC Curve (AUC) ( $p=0.04$ ). Mixed training from scratch led to higher performance than fine-tuning or feature extraction. To assess generalizability, we also ran inference on an open dataset (BraTS-2015: 15 patients, 51 difference maps), reaching up to 76% AUC. Overall, results suggest that medical imaging problems may benefit from smaller models and different TL strategies with respect to computer vision problems, and that report-generated weak labels are effective in improving model performances. Code, in-house dataset and BraTS labels are released.

**Index Terms**— Change Detection, Deep Learning, High-grade Glioma, Transfer Learning, Weak Labels

## I. INTRODUCTION

Change detection aims at spotting the parts of an image that change over time. In recent years, this task has attracted increasing attention for medical imaging applications such as multiple sclerosis [1], chest x-ray [2], retinal fundus images [3] and glioma [4]–[6], as well as in other areas such as remote sensing and video processing [7]. Change detection is particularly relevant for evolving diseases monitored longitudinally, such as gliomas, the most frequent primary brain tumors occurring in the adult population. Their most aggressive

form (high-grade glioma) has a low survival (median  $\leq 2$  years) and requires prompt and dedicated treatment [8]. Magnetic Resonance Imaging (MRI) is the gold standard modality to monitor the evolution of gliomas since it allows the acquisition of diverse sequences, which provide complementary information to clinicians [9], and potentially avoids the need for serial biopsies [10]. Glioma change detection is a clinically-relevant, meticulous and non-trivial task for radiologists whose goal is to visually detect relevant tumor-related changes in order to detect early progressions or responses, and tailor treatment. Throughout the rest of the paper, we denote the longitudinal monitoring of gliomas via MRI as “glioma change detection”.

In this work, we address glioma change detection via a Deep Learning (DL) pipeline that leverages weak labels and transfer learning. In the following paragraphs, we explain why these two expedients are useful in medical imaging, and how they can be exploited to increase performances.

The need for large amounts of manual annotations is arguably the major bottleneck for the development of supervised DL models in medical imaging. Not only is the creation of manual labels tedious for medical experts, but it is also extremely time-consuming [11]; combined with the increasing workload of radiologists [12], the creation of manual labels is expected to become more and more expensive in the coming years. This has prompted interest and advances in more sample-efficient learning methods. In this respect, weak labels are an interesting alternative to manual labels: they correspond to noisy, limited, or imprecise labels that are adopted to guide the learning process [13]–[15]. One under-explored approach for generating weak labels in medical imaging is the use of radiology reports. Most of the time, reports are stored as unstructured free-text and exhibit a strong degree of ambiguity and lack of conciseness [16]. However, recent advances in Natural Language Processing (NLP) enable the extraction of clinically-relevant labels [17], [18], [27], [28], [19]–[26] from radiology reports. Although these weak labels are inherently imperfect, they are drastically faster to obtain with respect to manual labels, and are also more scalable since they potentially allow to leverage tens of thousands of retrospective exams that would otherwise remain unused in hospital PACS (Picture

T.D.N is with the Department of Radiology, Lausanne University Hospital (CHUV) and University of Lausanne (Unil) (abbreviation: RAD-CHUV-Unil) (e-mail: [tommaso.di-noto@chuv.ch](mailto:tommaso.di-noto@chuv.ch)). M.B.C. is with CIBM, Center for Biomedical Imaging, Lausanne, and with RAD-CHUV-Unil (e-mail: [meritxell.bachcuadra@unil.ch](mailto:meritxell.bachcuadra@unil.ch)). C.A. is with RAD-CHUV-Unil (email: [chirine.atat@chuv.ch](mailto:chirine.atat@chuv.ch)). E.G.T. is with RAD-CHUV-Unil (e-mail: [eduardo.gamito-teiga@chuv.ch](mailto:eduardo.gamito-teiga@chuv.ch)). J.R. is with RAD-CHUV-Unil (e-mail: [jonas.richiardi@chuv.ch](mailto:jonas.richiardi@chuv.ch)). P.H. is with RAD-CHUV-Unil, and with Lundin Brain Tumor Research Center, Lausanne University Hospital and University of Lausanne (abbreviation: LBTRC-CHUV-Unil) (e-mail: [patric.hagmann@chuv.ch](mailto:patric.hagmann@chuv.ch)). A.H. is with LBTRC-CHUV-Unil, with Department of Clinical Neurosciences, and with Department of Oncology, CHUV (e-mail: [andreas.hottinger@chuv.ch](mailto:andreas.hottinger@chuv.ch)). M.H. is with LBTRC-CHUV-Unil, with the Neuroscience Research Center, CHUV, with Neurosurgery, CHUV, and with Swiss Cancer Center Léman (SCCL), Lausanne (e-mail: [monika.hegi@chuv.ch](mailto:monika.hegi@chuv.ch)).

Archiving and Communication Systems).

Transfer Learning (TL) is the branch of machine learning where knowledge acquired from a specific task or domain (source) is exploited to solve a downstream, related task (target) [29]. Since datasets used by most research groups in medical imaging are typically small [30] (especially compared to datasets in Computer Vision (CV)), TL holds great potential to overcome data scarcity in the field. Adopting the notation from [31], we can formally define a domain  $D$  and a task  $T$  as  $D = \{X, P(x)\}$  and  $T = \{Y, F(\cdot)\}$ , where  $X$  is the feature space,  $P(x)$  is the corresponding marginal probability distribution,  $Y$  is the label space, and  $f(\cdot)$  is the objective predictive function. Moreover, we use the notations  $D_s$ ,  $D_t$ ,  $T_s$ , and  $T_t$  to indicate source domain, target domain, source task and target task, respectively. Most papers dealing with medical TL focused on the choice of the source domain  $D_s$ , trying to understand which is the best  $D_s$  from which we should transfer knowledge. For instance, several works investigated the use of natural images (e.g. the ImageNet dataset [32]) for pre-training [33]–[36]. Conversely, more recent works showed that the use of natural images could lead to negligible performance improvements [37], and rather suggested that using a medical domain as source is preferable [38]–[40]. Differently from these previous studies, in our work we explore the TL scenario where source and target domain are identical ( $D_s = D_t$ ), but the tasks are different ( $T_s \neq T_t$ ) because of distinct label spaces ( $Y_s \neq Y_t$ ). In other words, we aim to understand to what extent it is possible to transfer knowledge from a source domain which has a different label distribution from the target domain, a scenario called *inductive* TL [31]. More specifically, we address the task of glioma change detection with difference maps as input samples ( $D_s = D_t$ , details in section II-D), with  $Y_s$  consisting of the above-mentioned weak labels generated automatically from radiology reports, and  $Y_t$  consisting of manual labels created by human experts, again from radiology reports.

Once domains ( $D_s$ ,  $D_t$ ) and tasks ( $T_s$ ,  $T_t$ ) have been defined, TL can be further subdivided into three main types [41]:

- **Fine-tuning**: the DL model is pre-trained on the source domain and then all its weights are fine-tuned on the target domain.
- **Feature Extraction**: the DL model is pre-trained on the source domain and then only some of its weights (typically the last linear layers) are fine-tuned on the target domain. Instead, the convolutional backbone layers are usually “frozen” (i.e. not trained again).
- **Mixed Training**<sup>2</sup>: the DL model is trained only once on a mixed dataset composed of source domain and the training portion of the target domain.

Similarly to the discussion about the choice of the source domain, there is also a lack of consensus regarding which type of TL is the most effective (e.g. is fine-tuning better than feature extraction?), with most of the works trying several combinations empirically [41]. In this paper, we develop an automated pipeline which treats the TL type as just another

hyperparameter to optimize. Because the optimal value of other hyperparameters (such as the learning rate) depends on the TL type, our approach avoids the arbitrary choice of a TL type which can be potentially suboptimal.

Beside the choice of the source domain and the TL type to adopt, it is also unclear how much model size influences classification results during TL. For instance, [37] found that large networks that yield state-of-the-art results for ImageNet are not necessarily the top performing networks for medical datasets. Moreover, the authors in [37] also showed that in the small data regime (i.e. few thousands of samples or below) large ImageNet models benefit more from TL with respect to smaller networks. This behavior is frequent in CV where larger networks tend to maintain a performance edge over smaller networks even in the low-data regime [42]. Subsequent work [43] instead found that using extremely large architectures (380M parameters) and massive pre-training datasets (300M images) from the natural images domain can actually improve results on target medical domains. In the same line as these studies, we compare a low-capacity model to a higher-capacity model (details in section II-E) to investigate the impact of model size, but in the scenario of inductive TL.

In summary, the goal of this paper is to tackle glioma change detection within an inductive TL scenario ( $D_s = D_t$ ,  $Y_s \neq Y_t$ ). The main contributions of our work are the following: (i) we propose a TL approach that leverages inexpensive and fast-to-create weak labels generated from radiology reports; (ii) we automate the choice of TL type, treating it as another hyperparameter to optimize, and thus avoiding manual empirical trials; (iii) we assess the impact of model size on TL for medical imaging and (iv) we release new expert labels for the longitudinal subjects of the public BraTS 2015 dataset [44]–[46], as well as our in-house 1693 longitudinal difference images for glioma, the largest such dataset currently available.

#### A. Related works

Previous works have addressed the task of glioma change detection. For instance, [4] used difference maps after contrast midway mapping to monitor tumor growth with FLuid Attenuated Inversion Recovery (FLAIR) images. Instead, [5] monitored low-grade glioma growth via a dedicated segmentation pipeline, again on FLAIR images. Last, [6] tried to distinguish radiation-induced pseudo progression from real tumor progression using 3D shape features and a support vector machine.

Several works have explored the potential of NLP for classifying radiology reports [17], [18], [27], [28], [19]–[26]. However, only a few studies later investigated the application of their trained report classifier for a downstream imaging task [28], [47]–[49]. The authors in [47] showed that their report classifier could be used to triage head MRI scans and identify relevant abnormalities. Instead, authors in [48] used labels generated from reports of FDG-PET/CT to detect and estimate the location of abnormalities in whole-body scans. The work [49] described an NLP model that is used to generate weak

<sup>2</sup> Although strictly speaking there is no transfer of knowledge for this subgroup, we loosely include Mixed Training among the TL types.

image-level labels which are later integrated into a semi-supervised framework for mass detection in mammography images. Last, the authors in [28] trained an NLP classifier to create weak labels from pathology reports and later used these weak labels to train a DL model on colon Whole Slide Images. Similarly to [28], [47]–[49], we investigate whether the report classifier built in [17] can be useful to generate weak labels which are then used for the downstream imaging task of glioma change detection.

Regarding the automation of TL, most works have focused on measuring *transferability* between domains and tasks: for instance [50] define transferability as the difference in performance between models trained on source and target tasks, and use this information to improve several downstream tasks using logistic regression models. Similarly, [51] proposed a computational approach to discover transferability between 26 CV tasks, yielding optimal combinations of deep learning features for each target task. In terms of domain choice, the authors in [52] propose an information theoretic framework which permits to rank convolutional neural networks trained on different source domains and understand which is the most suitable for knowledge transfer. Alternatively, [53] proposed an adversarial multi-armed bandit that automatically decides which (if any) are the features of the source network that are useful for the target network. Our work differs from the above since the automation of our TL pipeline is focused on the type of TL (fine-tuning, feature extraction, or mixed training) rather than the selection of the most relevant source domain or task.

Finally, previous works have already explored the influence of model size on TL [37], [38], [43]. However, these works focused on the *transductive* TL scenario [31] where  $D_s \neq D_t$ , whereas we found no work that assessed the impact of model size for the *inductive* TL scenario ( $D_s = D_t, Y_s \neq Y_t$ ).

## II. MATERIALS AND METHODS

### A. In-house Dataset

We retrieved 2100 MR scans belonging to 183 retrospective patients with high-grade gliomas who were scanned between 2004 and 2019 at the Lausanne University Hospital (average number of scans per subject 5, standard deviation 4.5). At every session, a series of MR scans including structural, perfusion and functional imaging were performed. For simplicity, in this work we only focused on the T2-weighted (T2w) scans. The MR acquisition parameters for the cohort are provided in Table I. Scans that were too close to surgery (within 4 weeks) were excluded since they contained exaggerated intensity changes and brain deformations around the resection cavity, due to edema. We deem these changes irrelevant since they are not related to the tumor itself. In addition, we extracted the radiology reports associated with each session. These were written (or dictated) in French during routine clinical practice by a junior radiologist after exploring all sequences of interest. Then, a senior radiologist reviewed each case amending the report when necessary. The extracted reports have varying length ranging from 121 to 751 words (average 325, standard deviation 84). The protocol of this study was approved by the regional ethics committee; written informed consent was

waived. We release an anonymized version of our in-house dataset on Zenodo (DOI: 10.5281/zenodo.7214605) under the permissive CC BY 4.0 license [54].

### B. BraTS Dataset

To assess the generalization of our pipeline to an external dataset, we ran inference on the longitudinal subjects of the Multimodal Brain Tumor Segmentation (BraTS) 2015 multi-institutional dataset. We selected the 2015 edition because it is the only one that contains patients with multiple scans (i.e. longitudinal patients). Out of the 20 available longitudinal patients, we discarded 5 because they only contained two scans, namely the one before tumor resection and the one right after. For the remaining 15 subjects, we used 59 MR scans (average of  $\sim 4$  scans per subject), again only focusing on T2w scans. From these 59 scans, we generated 51 difference maps (creation process described in section II-D) which were tagged by one radiologist with over 18 years of experience (P.H.), using the labels presented in section II-C. We openly release these labels ([https://github.com/connectomicslab/Glioma\\_Change\\_Detecti\\_on\\_T2w/blob/master/extra\\_files/df\\_dates\\_and\\_t2\\_labels\\_brats\\_tcia\\_2015.csv](https://github.com/connectomicslab/Glioma_Change_Detecti_on_T2w/blob/master/extra_files/df_dates_and_t2_labels_brats_tcia_2015.csv)) for other researchers.

### C. Report Tagging

From the 183 glioma patients of the in-house dataset, we created two sub-datasets: a Human-Annotated Dataset and a Weakly-Annotated Dataset.

**Human Annotated Dataset (HAD)** - For this sub-dataset, three radiologists tagged the MR radiology reports with labels of interest. For each report, the annotators were instructed to assign 3 classes: one class that indicated the global conclusion of the report (global conclusion), one to indicate the evolution of the enhanced part of the tumor (T1w conclusion) and the last class to indicate the evolution of the tumor on T2-weighted sequences (T2w conclusion). For each of these classes, the annotator could choose between the following labels:

- **stable**: assigned when the tumor did not change significantly with respect to the previous comparative exam
- **progression**: assigned when the tumor worsened with respect to the previous comparative exam. This class included cases where the enhanced part of the tumor increased in size or when the T2 signal anomalies surrounding the tumor increased in extension
- **response**: assigned when the tumor responded positively to the treatment (either chemotherapy or radiotherapy)
- **unknown**: assigned if the annotator was not able to assign any of the three classes when reading the report



TABLE I  
MR ACQUISITION PARAMETERS OF THE 2100 T2w SCANS BELONGING TO THE IN-HOUSE GLIOMA PATIENTS USED FOR THE STUDY.

# scans	Vendor	Model	Field Strength [T]	Median TR [ms]	Median TE [ms]	Median Voxel Spacing [mm <sup>3</sup> ]
800	Philips	Intera	3.0	3000	80	0.45x0.45x4.0
445	Siemens Healthineers	Skyra	3.0	5000	77	0.45x0.45x3.3
304	Siemens Healthineers	TrioTim	3.0	4700	84	0.45x0.45x3.9
254	Siemens Healthineers	Symphony	1.5	4370	103	0.45x0.45x6.5
127	Siemens Healthineers	Verio	3.0	5000	85	0.45x0.45x3.3
85	Siemens Healthineers	Aera	1.5	6220	84	0.6x0.6x3.3
77	Siemens Healthineers	Prisma	3.0	4881	77	0.45x0.45x3.3
2	Siemens Healthineers	Espreo	1.5	6000	93	0.53x0.53x7.2
2	Philips	Ingenia	1.5	3448	80	0.34x0.34x3.6
1	Siemens Healthineers	Vida	3.0	5320	77	0.45x0.45x3.3
1	Philips	Achieva	1.5	3659	50	0.39x0.39x4.55
1	Philips	Panorama HFO	1.0	5300	100	0.33x0.33x5.3
1	GE HealthCare	Discovery MR750	3.0	7955	100	0.47x0.47x4.0

381 reports (belonging to 169 distinct patients) were manually annotated by our three experts.

Out of these 381, 39 reports (belonging to 39 distinct patients) were tagged by a senior radiologist with over 18 years of experience in neuroimaging (P.H), while 342 reports (belonging to 162 patients) were tagged by two radiologists both with 4 years of experience (C.A, E.G.T). Cohen's kappa coefficient between the two readers for the T2w conclusion was  $k=0.80$  which is considered a "substantial agreement" [55]. The 41/342 reports for which the two annotators disagreed were discarded. Also, we discarded 90 reports for which the T2w conclusion was different from the global conclusion. The rationale behind this choice was to exclude misleading cases for which the report was ambiguous (e.g. T2w conclusion = progression, global conclusion = stable), discordant (e.g. T2w conclusion = progression, global conclusion = response), or cases for which signs of progression were visible only on T1w scans (e.g. T1w conclusion = progression, T2w conclusion = stable). Last, we also excluded the 17 reports for which the T2w conclusion was tagged as unknown. This left a total of 91 patients, 378 scans, and 233 difference maps (see Figure 1).

**Weakly Annotated Dataset (WAD)** - For this sub-dataset, reports were tagged with the classifier proposed in our previous work [17]. Briefly, this consists of an NLP pipeline in which we preprocess (e.g. removed proper nouns, stopwords, punctuation), embed (with Doc2Vec [56]) and then classify (Random Forest with 501 trees) the radiology reports precisely into the classes mentioned above (i.e. **stable**, **progression** and **response**). We denote the labels generated from the report classifier as weak because the classifier will commit errors, and because, differently from human readers, it cannot abstain when the reports are unclear (i.e. there is no **unknown** label).

Both for **HAD** and **WAD** we merged **progression** and **response** into one unique class which we denote as **unstable**. By doing this, we narrowed the task to a binary classification problem where we try to distinguish between **stable** and **unstable** reports.

This scenario corresponds to a worklist prioritization in radiology departments, where we would want to prioritize examinations that are unstable and require more attention. After these modifications, HAD contained 233 reports (159 stable, 74 unstable), whereas WAD contained either 795 (333 stable, 462 unstable) or 361 (165 stable, 196 unstable) reports, depending on the probabilistic output of the random forest (hyperparameter *fraction\_of\_WAD*, details in section II-F). A detailed overview of the dataset is provided in Figure 1.

#### D. Image-based change detection

While our former study [17] focused on report-based glioma change detection, this work deals with image-based glioma

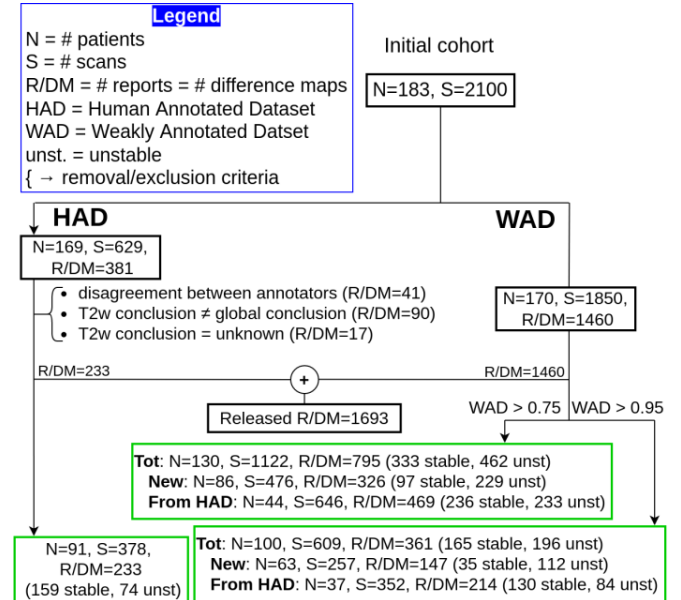


Fig. 1. Dataset overview. Each report corresponds to one T2w difference map since they both link two time points (i.e. two MR scans). The branches WAD > 0.75 and WAD > 0.95 depend on the hyperparameter *fraction\_of\_WAD* described in section II-F. Difference maps in WAD come both from new patients (distinct from HAD patients) and HAD patients since not all reports from HAD have been tagged manually. Green rectangles indicate the final sets of patients/scans/difference maps used for the downstream analyses.

change detection. We know that every radiology report links two time points, namely the current scan and a previous scan which is used as baseline for comparison and longitudinal monitoring. Thus, for each report, we generated a corresponding T2w absolute difference map as illustrated in Figure 2. The rationale behind these difference maps is that parts of the tumor that either progress or respond to treatment (unstable) should appear as hyper-intense (examples (a) and (c) in Figure 2); instead, if the tumor is stable across the two time-points, the difference map will likely be hypo-intense overall (examples (b) and (d)). To generate the difference maps, we first applied N4 bias field correction with ANTs [57] both to the previous and to the current T2w volumes. Second, we registered the previous scan to the current scan, again with ANTs. Third, we skull-stripped both volumes (previous warped and current) with HD-BET [58]. Fourth, we applied z-score normalization on both volumes. Last, we computed the absolute voxel-wise difference of the normalized volumes.

### E. Classification Networks

The image-based change detection is treated as a binary classification problem: as for the reports, we try to classify the difference maps into stable and unstable in order to prioritize more urgent patients. We used two Convolutional Neural Networks (CNNs) for the classification of the T2w difference maps: a custom 3D-VGG [59] (henceforth called VGG) and a 3D-ResNeXt [60] with Squeeze-and-Excitation [61] (henceforth called SEResNeXt). The VGG was written in PyTorch and contains 4 convolutional blocks followed by 4 fully-connected blocks. We used the ReLU activation function for all layers, except for the last layer which is followed by a sigmoid function. Batch normalization [62] was added in the VGG to prevent overfitting. The SEResNeXt was implemented with the MONAI framework [63]. For both CNNs, we used the cross-entropy loss function and the ADAM optimizer [64] to guide the learning process. During the training, we applied online data augmentations, namely flip, addition of Gaussian noise, zoom (from 0.7 to 1.3, 1 being the original volume size) and elastic deformation, each with probability of 20%. The total number of trainable parameters in our networks is  $\sim 7.5$  M for the VGG and  $\sim 19.4$  M for SEResNeXt. Training and evaluation were performed with PyTorch 1.11.0 and a GeForce RTX 3090 GPU.

### F. Experiments and Hyperparameter Tuning

**Creation of weak labels with report classifier** - For this work, we adapted the report classifier [17] and trained it to classify the T2w conclusion (in [17] it was trained to classify the global conclusion). We ran a nested 5-fold cross-validation on the 233 HAD reports, selected the best hyperparameters, and finally performed inference with the best model on all WAD reports to obtain the weak labels later used for the image-based change detection.

**Image-based glioma change detection** - Because of computational constraints, we decided to fix some hyperparameters, and tune others. Among the fixed (not tuned) hyperparameters we chose a batch size of 4, and 60 training epochs with early stopping. Depending on the experiments

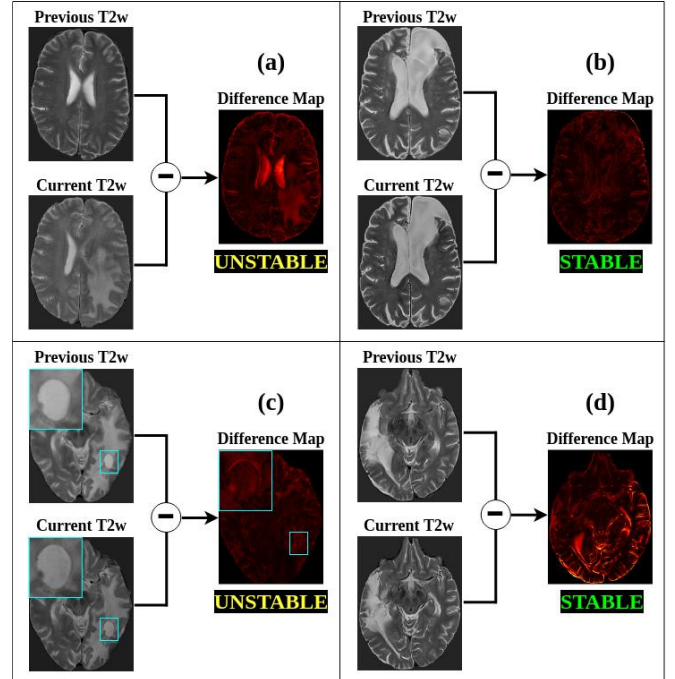


Fig. 2. Creation of T2w difference maps. After registration and normalization of the previous and current T2w volumes, the maps are computed via voxel-wise absolute difference. (a) 58-year-old male patient with a progressing (i.e. unstable) gliosarcoma. (b) 59-year-old male patient with a stable astrocytoma. (c) 60-year-old male patient with seemingly stable glioblastoma, but with enlarging cystic lesion (zoom inset in cyan color). (d) 60-year-old male patient with a (less evident) stable oligodendroglioma.

detailed below, other hyperparameters were tuned using the Optuna framework [65] with default arguments (Tree-structured Parzen Estimator as sampler, and Median pruner), and maximizing the Area Under the Receiver Operating Characteristic Curve (AUC) of a dedicated validation set composed of 25% of the training subjects (details in section II-G).

To understand which TL type is the most appropriate to improve classification performances and how model capacity can influence TL results, we performed two experiments (called **Baseline** and **TL**) with the two DL models described above (VGG and SEResNeXt): in the **Baseline** experiment, we conducted a 5-fold cross-validation only on HAD (details in section II-G), and WAD was intentionally not used (i.e. no TL). Evaluation was performed on the test subjects of each cross-validation fold and then results were aggregated. The only two hyperparameters that were tuned for the **Baseline** experiment were *learning\_rate* and *weight\_decay*. The former was chosen from  $\{1 \times 10^{-4}, 1 \times 10^{-5}, 1 \times 10^{-6}\}$ , whereas the latter was chosen from  $\{0, 0.01\}$ . Since only two hyperparameters were tuned, both for the VGG-Baseline experiment and the SEResNeXt-Baseline experiment we ran all the six hyperparameter combinations. In the **Transfer Learning (TL)** experiment, we still performed a 5-fold cross-validation on HAD, but this time we also exploited the WAD difference maps. In addition to *learning\_rate* and *weight\_decay* (which are tuned identically to the **Baseline**), here we also searched for the best transfer learning configuration. Specifically, we tuned 3

additional hyperparameters: *mixed\_training*, *feature\_extraction* and *fraction\_of\_WAD*.

- *mixed\_training* can either be True or False: if True, we use for training a mixed shuffled dataset that is composed of WAD difference maps and the difference maps of the training HAD patients (scenario 3, section 1); if instead *mixed\_training* is False, we either
  - perform feature extraction if *feature\_extraction* is True (scenario 2, section 1), or
  - fine-tuning if *feature\_extraction* is False (scenario 1, section 1)
- *fraction\_of\_WAD* indicates which portion of WAD to use. We added this hyperparameter because not all weakly-labeled data is necessarily useful. In other words, by tuning *fraction\_of\_WAD* we wanted to understand whether some reports (and hence some difference maps) are more informative than others. The tunable values that we chose for *fraction\_of\_WAD* were  $\{WAD > 0.75, WAD > 0.95\}$  where 0.75 and 0.95 are the output probabilities (soft labels) of the report classifier from [17]. For instance, when using  $WAD > 0.95$  we only use a small portion of WAD, namely only the reports for which the report classifier is highly confident (output probability  $> 0.95$ ). Instead, when using  $WAD > 0.75$  we also include reports for which the NLP classifier is less confident<sup>3</sup>.

Figure 3 illustrates one branch of the tree containing all possible hyperparameter combinations for the TL experiment. Since running all combinations would have been computationally impractical, we only ran each TL experiment (VGG-TL and SEResNeXt-TL) for 4 days.

To summarize, we ran 4 experiments: VGG-Baseline, VGG-TL, SEResNeXt-Baseline, and SEResNeXt-TL. The comparisons between **Baseline** and **TL** aimed to assess the effectiveness of the weak labels in WAD. Results related to the impact of weak labels and TL are reported in section III-A. Instead, comparisons between the two CNNs (e.g. VGG-Baseline vs. SEResNeXt-Baseline) aimed to understand the influence that model capacity can have on TL strategies for our task. Results related to the impact of model size are reported in section III-B. The most frequent hyperparameter combinations are then reported in section III-C, and finally in section III-D we report inference results of our 4 models (VGG-Baseline, VGG-TL, SEResNeXt-Baseline, SEResNeXt-TL) on the longitudinal patients of the external BraTS 2015 dataset. For each model, we ran inference with the five trained model of the cross-validation, and then performed majority voting of the five predictions.

### G. Cross-Validation and evaluation

**Cross-Validation** - For the **Baseline** experiments, we performed a 5-fold cross-validation on HAD. At each cross-validation split, 80% (72/91 subjects, 166 difference maps) of the subjects are used to train the CNN (either VGG or SEResNeXt), while the remaining 20% (19/91 subjects; 67

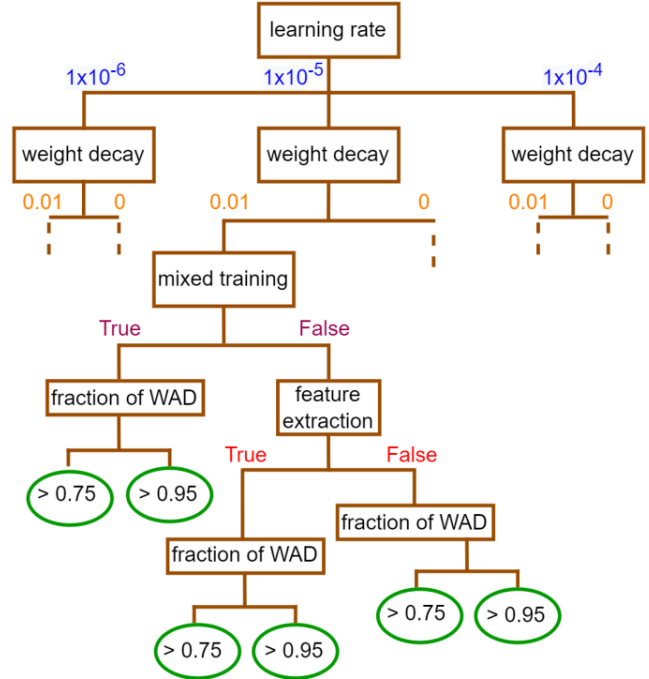


Fig. 3. Tree of hyperparameter combinations for the transfer learning experiments. Dashed lines indicate branches that are not shown because of limited figure space. Green ellipses are the leaves of the hyperparameter tree. If *mixed\_training* is True, we end up in scenario 3 of section 1; if *mixed\_training* is False and *feature\_extraction* is True we perform feature extraction (scenario 2, section 1); if False and *feature\_extraction* is False, we perform fine-tuning (scenario 1, section 1).

difference maps) of the subjects are used to compute test results. Within each cross-validation fold, we also used 25% (18/72 subjects, 49 difference maps) of the training subjects as validation set for tuning the hyperparameters. To avoid over-optimistic results, the cross-validation splits were always performed at the subject-level to prevent multiple difference maps of the same subjects being assigned some to the training and some to the test (or validation) set. For the **TL** experiments, we performed the same 5-fold cross-validation on HAD, but then adapted the learning strategy according to the hyperparameters chosen during hypertuning: if *mixed\_training* was True, then WAD subjects were added to the training subjects of HAD, while if *mixed\_training* was False, the model was first pre-trained on WAD and then fully (fine-tuning) or partially (feature extraction) fine-tuned on the training subjects of HAD. We ensured that the same splits were performed on HAD both for the **Baseline** and **TL** experiments, so that an exact comparison could be carried out. Also, for the HAD patients with overlapping scans (some manually annotated and some automatically annotated), we made sure to never assign some scans to training and some to test (or validation) set.

**Metrics** - The task that we address is binary classification of the T2w difference maps which are labeled either as **stable** or **unstable**. We report in the Results section accuracy, sensitivity (recall), specificity, F1 score, AUC, and Area Under the

<sup>3</sup> In the beginning, we tried using all WAD, but this consistently led to lower performances (results not shown).

Precision-Recall curve (AUPR). We consider the class **unstable** as “positive”, and the class **stable** as “negative”.

**Statistics** - To statistically compare the four different models presented in section III-F, we ran permutations tests using the difference in AUCs as test statistic, as similarly performed in [66]. We set a significance threshold  $\alpha = 0.05$  and we ran 10,000 permutations for each test.

**Code availability** - All the code used for this paper is available at [https://github.com/connectomicslab/Glioma\\_Change\\_Detection\\_T2w](https://github.com/connectomicslab/Glioma_Change_Detection_T2w), together with corresponding configuration files to reproduce the experiments.

### III. RESULTS

In cross-validation, the report classifier reached an accuracy of 93%, a sensitivity of 91% and a specificity of 94% on the 233 HAD reports. When running inference on WAD, 795 reports were associated with a class probability  $> 0.75$ , while 361 were associated with a class probability  $> 0.95$ .

The upper part of Table II shows test classification results of the VGG and the SEResNeXt for the task of image-based glioma change detection on the in-house dataset. The **Baseline** experiments are those where only HAD is used, while in the **TL** experiments we also leverage WAD. To visually summarize classification results, we also report in Figures 4 and 5 the Receiver Operating Characteristic (ROC) and the Precision-Recall (PR) curves, respectively.

#### A. Weak labels and TL improve classification results for VGG

We found a significant difference in AUC between the models VGG-Baseline and VGG-TL ( $p=0.05$ ). This finding indicates the superiority of the TL pipeline with respect to the Baseline, which is visually confirmed in Figures 4 and 5 where the VGG-TL consistently outperforms VGG-Baseline.

Conversely, the permutation test indicated that the SEResNeXt-Baseline and SEResNeXt-TL had no significant difference ( $p=0.18$ ), even though SEResNeXt-TL showed higher AUC and AUPR, and the corresponding PR curve (yellow, Figure 5) outperforms the one from SEResNeXt-Baseline (black, Figure 5) for most operating points. Overall,

the two experiments show that only the VGG model benefits significantly from TL with the weakly-labeled dataset WAD.

#### B. Model size is negligible for the task at hand

To assess the impact of model size, we compared the VGG-Baseline vs. the SEResNeXt-Baseline model and found no significant difference between the two ( $p=0.17$ ). Then, we also compared the VGG-TL to the SEResNeXt-TL model and again we found no significant difference ( $p=0.39$ ). These experiments indicate that, for the task at hand, model size does not influence classification results, even though the SEResNeXt has  $\sim 2.5X$  more trainable parameters than the VGG (19.4M vs. 7.5M) and is slower to train (e.g. 1 epoch of the Baseline experiment takes 120 seconds for SEResNeXt vs. 90 seconds for VGG). Overall, these results suggest that the VGG is preferable for the task at hand because it is simpler and more computationally efficient.

#### C. Most frequent hyperparameters

Here, we report the most frequent hyperparameters that were chosen by the Optuna optimizer across the 5 training folds. For the VGG-Baseline experiment, the most frequent *learning\_rate* was  $1 \times 10^{-4}$  (3 folds out of 5) and the most recurrent *weight\_decay* was 0.01, while for the SEResNeXt-Baseline the most frequent *learning\_rate* was  $1 \times 10^{-5}$  (3/5 folds) and the most frequent *weight\_decay* was 0. More interestingly, we found a peculiar pattern in the hyperparameters of the TL pipeline: both for the VGG-TL (5/5 folds) and for the SEResNeXt-TL (4/5 folds), the hyperparameter *mixed\_training* was always True. This means that training from scratch with a mixed dataset (WAD + training HAD) consistently leads to higher performances with respect to either fine-tuning or feature extraction. Regarding the hyperparameter *fraction\_of\_WAD*, the most frequent value for the VGG-TL experiment was  $WAD > 0.95$  (3/5 folds), whereas the most recurrent value for SEResNeXt-TL was  $WAD > 0.75$  (4/5 folds).

#### D. Inference on BraTS

Out of the 51 difference maps that we extracted from BraTS 2015, 12/51 (23%) were tagged as stable, while 39/51 (76%) were tagged as unstable, by our senior radiologist. The lower

TABLE II

CLASSIFICATION TEST RESULTS. **UPPER PART**: IN-HOUSE DATASET . **LOWER PART**: BRATS-2015 DATASET . BOLD VALUES INDICATE THE HIGHEST PERFORMANCES. N=NUMBER OF DIFFERENCE MAPS; BASELINE = PIPELINE WHERE ONLY HAD DATA IS USED. TL = TRANSFER LEARNING: PIPELINE WHERE BOTH HAD AND WAD ARE USED. ACC=ACCURACY; SENS=SENSITIVITY; SPEC=SPECIFICITY; F1=F1 SCORE; AUC=AREA UNDER THE ROC CURVE; AUPR=AREA UNDER THE PRECISION-RECALL CURVE; PARAMS=NUMBER OF PARAMETERS IN THE MODEL.

Dataset	N	MODEL	ACC	SENS	SPEC	F1	AUC	AUPR	# PARAMS
In-house	233	VGG-Baseline	70	55	77	54	.74	.55	7.5M
		VGG-TL	<b>79</b>	<b>80</b>	79	<b>71</b>	.82	.72	
		SEResNeXt-Baseline	76	50	<b>88</b>	57	.79	.63	19.4M
		SEResNeXt-TL	77	78	76	68	<b>.83</b>	<b>.73</b>	
BraTS 2015	51	VGG-Baseline (inference)	75	82	50	83	.66	.90	7.5M
		VGG-TL (inference)	76	<b>92</b>	25	86	.59	.89	
		SEResNeXt-Baseline (inference)	73	69	<b>83</b>	79	<b>.76</b>	<b>.93</b>	19.4M
		SEResNeXt-TL (inference)	<b>78</b>	95	25	<b>87</b>	.60	.60	

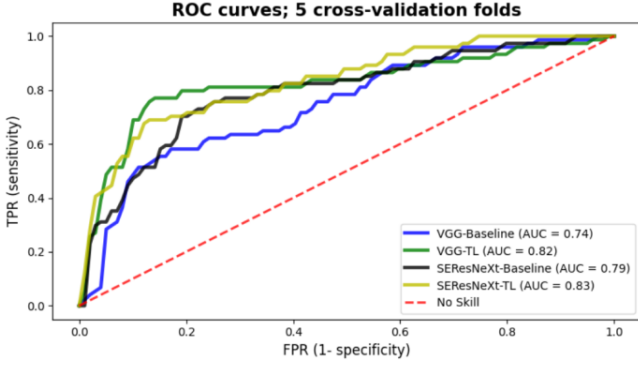


Fig. 4. Receiver Operating Characteristic Curve (ROC) curves aggregated over the five test folds of HAD. AUC = Area Under the ROC Curve. TL = Transfer Learning.

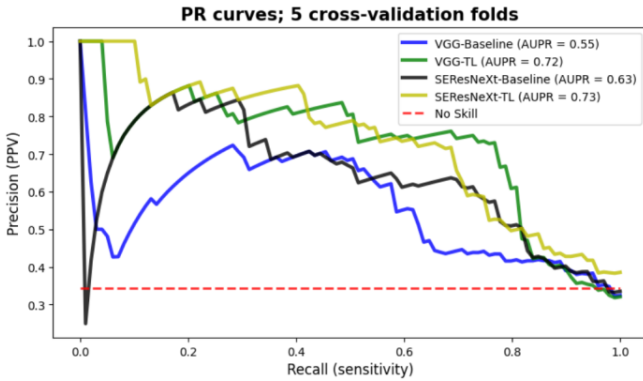


Fig. 5. Precision-Recall (PR) curves aggregated over the five test folds of HAD. AUPR = Area Under the PR curve. TL = Transfer Learning.

part of Table II illustrates inference results of the trained models after majority voting among the five splits of the cross-validation. Although the SEResNeXt-Baseline model showed the highest AUC, it did not significantly outperform the SEResNeXt-TL ( $p=0.46$ ), or the VGG-Baseline ( $p=0.39$ ).

#### IV. DISCUSSION

This work investigated the effectiveness of inductive TL for the task of image-based glioma change detection. To this end, we compared an automated TL pipeline that leverages weakly annotated data with a baseline that uses only human-annotated data. The experiments were run with two CNNs (VGG and SEResNeXt) to assess the impact that model size can have on classification performances and finally the pipeline was validated on the external BraTS dataset to assess model generalizability.

Despite being less accurate, weak labels extracted from radiology reports hold great potential for mitigating the manual annotation bottleneck in medical imaging. The main advantage of NLP-generated weak labels is that report classifiers are normally fast to train (e.g. the one presented in [17] takes  $\sim 10$  minutes). Therefore, labeling hundreds (or even thousands) of new subjects becomes extremely fast and inexpensive. In this work, the weak annotation process allowed us to obtain a more than 3-fold increase in sample size (233 difference maps for

HAD vs. 795 for the TL pipeline with  $WAD > 0.75$ ) at very little added cost. Results in section III-A showed that the automatically-labeled dataset WAD helps improving classification results, although the difference in performance between Baseline and TL was only significant for the VGG model. This result differs from [37] since in the small data regime we found the smaller network (VGG) to benefit more from TL with respect to the larger SEResNeXt. Nonetheless, as similarly reported in [48], we expect performances of both models to increase even further as more weakly-labeled samples are added.

When studying the impact of model size in classification performances (section III-B), we found no significant difference between VGG and SEResNeXt neither for the Baseline nor for the TL experiment. Therefore, for our application, we conclude that the VGG model is preferable because it is simpler and faster to train. A similar result was found in [67] where a VGG19 model outperformed much deeper networks in a TL pipeline for COVID-19 detection. Both our results and the ones in [67] indicate that the high-capacity networks and transfer learning strategies typically used for computer vision tasks in the high-data regime are not necessarily optimal for medical imaging tasks, where models often operate in the low-data regime. Given that deep learning scaling studies typically show log-linear or power laws relating loss to dataset size [68], [69], including for transfer learning [70], it is possible that the higher-capacity SEResNeXt model in our study would be superior if much more data were available, but this is not visible with our small dataset as we are far from the performance asymptote.

Another contribution of this work is the automation of the TL pipeline. Instead of searching for the best TL type manually, we framed the TL experiments as a hyperparameter optimization problem. We believe that our pipeline can be adopted by similar works that aim to automate TL for image classification. Surprisingly, we found that mixed training TL led to the highest classification performances. From a computational and environmental point of view, this finding is alarming because it indicates that the longest-running, least resource-efficient TL pipeline could be preferable with respect to feature extracting or fine-tuning.

As last contribution, we also evaluated our four models on the external BraTS dataset in order to assess model generalizability. Although the sample size is limited (51 difference maps) and no significant differences were found with the permutation tests, results on the lower part of Table II seem to indicate that the two Baseline models (VGG and SEResNeXt) can better cope with class imbalance (higher specificity and AUC).

Our work has several limitations. First, we narrowed the classification problem to a binary scenario **stable** vs. **unstable** tumor, mainly because we do not have enough cases of tumor **response** in our cohort. This is a simplification because **progression** and **response** are distinct clinical indicators. In future works, we are planning to extract new patients and adapt the classification towards a 3-class problem (**stable**, **progression**, **response**). As shown in [71], this will require careful analyses since results might change significantly when the labels of the task become more granular. The second limitation of this study is that we only focused on T2w MRI volumes, even though a multi-modal assessment of glioma

evolution would be more accurate [10]. Third, the reports from the HAD for which the two annotators disagreed were discarded, while in the future we plan to use them after a consensus between the readers has been reached. Additionally, we only evaluated one approach for fine-tuning, whereas other strategies, including freezing different layers for different number of epochs [36], remain to be explored.

## V. CONCLUSION

This study presented a TL pipeline that uses weakly-labeled data generated from radiology reports to improve classification performances for the task of glioma change detection. We found that a custom VGG model benefits more from transfer learning (and has similar performances) with respect to a more complex ResNet-like model. We hope this finding raises awareness regarding the potentially misleading translation between computer vision and medical imaging applications, and that our automated pipeline can be replicated for similar TL tasks in the field.

## ACKNOWLEDGMENT

We thank the Lundin Brain Tumor Research Center for support.

## REFERENCES

- [1] M. Bosc, F. Heitz, J. P. Armspach, I. Namer, D. Gounot, and L. Rumbach, "Automatic change detection in multimodal serial MRI: Application to multiple sclerosis lesion evolution," *Neuroimage*, vol. 20, no. 2, pp. 643–656, Oct. 2003, doi: 10.1016/S1053-8119(03)00406-3.
- [2] D. Y. Oh, J. Kim, and K. J. Lee, "Longitudinal Change Detection on Chest X-rays Using Geometric Correlation Maps," in *MICCAI*, 2019, vol. 11769, doi: 10.1007/978-3-030-32226-7.
- [3] Y. Fu, Y. Wang, Y. Zhong, D. Fu, and Q. Peng, "Change detection based on tensor RPCA for longitudinal retinal fundus images," *Neurocomputing*, vol. 387, pp. 1–12, Apr. 2020, doi: 10.1016/j.neucom.2019.12.104.
- [4] E. D. Angelini, J. Delon, A. B. Bah, L. Capelle, and E. Mandonnet, "Differential MRI analysis for quantification of low grade glioma growth," *Med. Image Anal.*, vol. 16, no. 1, pp. 114–126, Jan. 2012, doi: 10.1016/j.media.2011.05.014.
- [5] H. M. Fathallah-Shaykh *et al.*, "Diagnosing growth in low-grade gliomas with and without longitudinal volume measurements: A retrospective observational study," *PLoS Med.*, vol. 16, no. 5, May 2019, doi: 10.1371/journal.pmed.1002810.
- [6] M. Ismail *et al.*, "Shape features of the lesion habitat to differentiate brain tumor progression from pseudoprogression on routine multiparametric MRI: A multisite study," *Am. J. Neuroradiol.*, vol. 39, no. 12, pp. 2187–2193, Dec. 2018, doi: 10.3174/ajnr.A5858.
- [7] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005, doi: 10.1109/TIP.2004.838698.
- [8] M. Weller *et al.*, "Glioma," *Nature Reviews Disease Primers*, vol. 1. Nature Publishing Group, Jul. 16, 2015, doi: 10.1038/nrdp.2015.17.
- [9] S. Bauer, R. Wiest, L. P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," *Physics in Medicine and Biology*, vol. 58, no. 13, Jul. 07, 2013, doi: 10.1088/0031-9155/58/13/R97.
- [10] B. Menze *et al.*, "Analyzing magnetic resonance imaging data from glioma patients using deep learning," *Comput. Med. Imaging Graph.*, vol. 88, Mar. 2021, doi: 10.1016/j.compmedimag.2020.101828.
- [11] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," *Lect. Notes Comput. Vis. Biomech.*, vol. 26, pp. 323–350, 2018, doi: 10.1007/978-3-319-65981-7\_12.
- [12] B. Rao, V. Zohrabian, P. Cedeno, A. Saha, J. Pahade, and M. A. Davis, "Utility of Artificial Intelligence Tool as a Prospective Radiology Peer Reviewer — Detection of Unreported Intracranial Hemorrhage," *Acad. Radiol.*, vol. 28, no. 1, pp. 85–93, 2021, doi: 10.1016/j.acra.2020.01.035.
- [13] S. Abousamra *et al.*, "Weakly-Supervised Deep Stain Decomposition for Multiplex IHC Images," in *Proceedings - International Symposium on Biomedical Imaging*, 2020, vol. 2020-April, pp. 481–485, doi: 10.1109/ISBI45749.2020.9098652.
- [14] Ezhov, Zakirov, and Gusarev, "Coarse-to-fine volumetric segmentation of teeth in cone-beam CT," *arXiv*, pp. 0–4, 2018.
- [15] R. Ke, A. Bugeau, N. Papadakis, P. Schuetz, and C.-B. Schönlieb, "Learning to Segment Microscopy Images with Lazy Labels," *arXiv*, pp. 411–428, 2020, doi: 10.1007/978-3-030-66415-2\_27.
- [16] L. H. Schwartz, D. M. Panicek, A. R. Berk, M. Yuelin Li, and H. Hricak, "Improving Communication of Diagnostic Radiology Findings through Structured Reporting 1," *Radiology*, vol. 260, no. 1, 2011, doi: 10.1148/radiol.11101913/-DC1.
- [17] T. Di Noto *et al.*, "Diagnostic surveillance of high-grade gliomas: towards automated change detection using radiology report classification," 2021, doi: 10.1101/2021.09.24.21264002.
- [18] P. H. Chen, H. Zafar, M. Galperin-Aizenberg, and T. Cook, "Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports," *J. Digit. Imaging*, vol. 31, no. 2, pp. 178–184, Apr. 2018, doi: 10.1007/s10278-017-0027-x.
- [19] K. L. Kehl *et al.*, "Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes from Radiology Reports," *JAMA Oncol.*, vol. 5, no. 10, pp. 1421–1429, Oct. 2019, doi: 10.1001/jamaoncol.2019.1800.
- [20] S. Hassanpour, G. Bay, and C. P. Langlotz, "Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing," *J. Digit. Imaging*, vol. 30, no. 3, pp. 314–322, Jun. 2017, doi: 10.1007/s10278-016-9931-8.
- [21] S. Bozkurt, E. Alkim, I. Banerjee, and D. L. Rubin, "Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm," *J. Digit. Imaging*, vol. 32, no. 4, pp. 544–553, Aug. 2019, doi: 10.1007/s10278-019-00237-9.
- [22] C. R. Oliveira *et al.*, "Natural language processing for surveillance of cervical and anal cancer and precancer: Algorithm development and split-validation study," *JMIR Med. Informatics*, vol. 8, no. 11, Nov. 2020, doi: 10.2196/20826.
- [23] A.-D. Pham *et al.*, "Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings," *BMC Bioinformatics*, 2014, [Online]. Available: <http://www.biomedcentral.com/1471-2105/15/266>.
- [24] I. Banerjee *et al.*, "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification," *Artif. Intell. Med.*, vol. 97, pp. 79–88, Jun. 2019, doi: 10.1016/j.artmed.2018.11.004.
- [25] M. C. Chen *et al.*, "Deep learning to classify radiology free-text reports," *Radiology*, vol. 286, no. 3, pp. 845–852, Mar. 2018, doi: 10.1148/radiol.2017171115.
- [26] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," 2019, [Online]. Available: [www.aaii.org](http://www.aaii.org).
- [27] H. C. Shin, L. Lu, and R. M. Summers, "Natural Language Processing for Large-Scale Medical Image Analysis Using Deep Learning," in *Deep Learning for Medical Image Analysis*, Elsevier Inc., 2017, pp. 405–421.
- [28] N. Marini *et al.*, "Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations," *npj Digit. Med.*, vol. 5, no. 1, Dec. 2022, doi: 10.1038/s41746-022-00635-4.
- [29] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, Dec. 2016, doi: 10.1186/s40537-016-0043-6.
- [30] M. J. Willemink *et al.*, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1. Radiological Society of North America Inc., pp. 4–15, 2020, doi: 10.1148/radiol.2020192224.
- [31] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE, TKDE*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.
- [32] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*,

- Mar. 2010, pp. 248–255, doi: 10.1109/cvpr.2009.5206848.
- [33] Y. Xie and D. Richmond, “Pre-training on Grayscale ImageNet Improves Medical Image Classification,” 2018.
- [34] P. Rajpurkar *et al.*, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” *arXiv*, Nov. 2017, [Online]. Available: <http://arxiv.org/abs/1711.05225>.
- [35] V. Igloukov and A. Shvets, “TernausNet: U-Net with VGG11 Encoder Pre-Trained on ImageNet for Image Segmentation,” *arXiv*, Jan. 2018, [Online]. Available: <http://arxiv.org/abs/1801.05746>.
- [36] W. H. K. Chiu *et al.*, “Detection of COVID-19 Using Deep Learning Algorithms on Chest Radiographs,” *J. Thorac. Imaging*, vol. 35, no. 6, pp. 369–376, Nov. 2020, doi: 10.1097/RTI.0000000000000559.
- [37] M. Raghu, C. Zhang, G. Brain, J. Kleinberg, and S. Bengio, “Transfusion: Understanding Transfer Learning for Medical Imaging,” 2019.
- [38] L. Alzubaidi *et al.*, “Towards a better understanding of transfer learning for medical imaging: A case study,” *Appl. Sci.*, vol. 10, no. 13, Jul. 2020, doi: 10.3390/app10134523.
- [39] H.-Y. Zhou, S. Yu, C. Bian, Y. Hu, K. Ma, and Y. Zheng, “Comparing to Learn: Surpassing ImageNet Pretraining on Radiographs By Comparing Image Representations,” Jul. 2020, [Online]. Available: <http://arxiv.org/abs/2007.07423>.
- [40] M. R. H. Taher, F. Haghighi, R. Feng, M. B. Gotway, and J. Liang, “A Systematic Benchmarking Analysis of Transfer Learning for Medical Image Analysis,” *arXiv*, Aug. 2021, [Online]. Available: <http://arxiv.org/abs/2108.05930>.
- [41] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, “Transfer learning for medical image classification: a literature review,” *BMC Med. Imaging*, vol. 22, no. 1, p. 69, Dec. 2022, doi: 10.1186/s12880-022-00793-7.
- [42] J. Bornschein, F. Visin, and S. Osindero, “Small Data, Big Decisions: Model Selection in the Small-Data Regime,” 2020.
- [43] B. Mustafa *et al.*, “Supervised Transfer Learning at Scale for Medical Imaging,” *arXiv*, Jan. 2021, [Online]. Available: <http://arxiv.org/abs/2101.05913>.
- [44] B. H. Menze *et al.*, “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE TMI*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.
- [45] S. Bakas *et al.*, “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Sci. Data*, vol. 4, Sep. 2017, doi: 10.1038/sdata.2017.117.
- [46] S. Bakas *et al.*, “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge,” *arXiv*, Nov. 2018, [Online]. Available: <http://arxiv.org/abs/1811.02629>.
- [47] D. A. Wood *et al.*, “Automated labelling using an attention model for radiology reports of MRI scans (ALARM),” Feb. 2020, [Online]. Available: <http://arxiv.org/abs/2002.06588>.
- [48] S. Eyuboglu *et al.*, “Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT,” *Nat. Commun.*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/s41467-021-22018-1.
- [49] Y. Tang *et al.*, “Leveraging Large-Scale Weakly Labeled Data for Semi-Supervised Mass Detection in Mammograms,” 2021.
- [50] E. Eaton, M. Desjardins, and T. Lane, “Modeling Transfer Relationships Between Learning Tasks for Improved Inductive Transfer,” 2008.
- [51] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese, “Taskonomy: Disentangling Task Transfer Learning,” 2018, [Online]. Available: <http://taskonomy.vision/>.
- [52] M. J. Afridi, A. Ross, and E. M. Shapiro, “On automated source selection for transfer learning in convolutional neural networks,” *Pattern Recognit.*, vol. 73, pp. 65–75, Jan. 2018, doi: 10.1016/j.patcog.2017.07.019.
- [53] K. Murugesan, V. Sadashivaiah, R. Luss, K. Shanmugam, P.-Y. Chen, and A. Dhurandhar, “Auto-Transfer: Learning to Route Transferrable Representations,” Feb. 2022, [Online]. Available: <http://arxiv.org/abs/2202.01011>.
- [54] T. Di Noto *et al.*, “Transfer learning with weak labels from radiology reports: application to glioma change detection,” Oct. 2022, doi: 10.5281/ZENODO.7214605.
- [55] Mary L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem. Medica*, 2012.
- [56] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *31st International Conference on Machine Learning, ICML 2014*, May 2014, vol. 4, pp. 2931–2939, [Online]. Available: <http://arxiv.org/abs/1405.4053>.
- [57] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, “A reproducible evaluation of ANTs similarity metric performance in brain image registration,” *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011, doi: 10.1016/j.neuroimage.2010.09.025.
- [58] F. Isensee *et al.*, “Automated brain extraction of multisequence MRI using artificial neural networks,” *Hum. Brain Mapp.*, vol. 40, no. 17, pp. 4952–4964, Dec. 2019, doi: 10.1002/hbm.24750.
- [59] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2015, [Online]. Available: <http://arxiv.org/abs/1409.1556>.
- [60] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, and U. San Diego, “Aggregated Residual Transformations for Deep Neural Networks,” 2017, [Online]. Available: <https://github.com/facebookresearch/ResNeXt>.
- [61] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation Networks,” 2018, [Online]. Available: <http://image-net.org/challenges/LSVRC/2017/results>.
- [62] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *Int. Conf. Mach. Learn.*, 2015.
- [63] M. Consortium, “MONAI: Medical Open Network for AI,” Jul. 2022, doi: 10.5281/ZENODO.6903385.
- [64] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.
- [65] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Jul. 2019, pp. 2623–2631, doi: 10.1145/3292500.3330701.
- [66] A. I. Bandos, H. E. Rockette, and D. Gur, “A permutation test sensitive to differences in areas for comparing ROC curves from a paired design,” *Stat. Med.*, vol. 24, no. 18, pp. 2873–2893, Sep. 2005, doi: 10.1002/sim.2149.
- [67] M. M. Rahaman *et al.*, “Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches,” *J. Xray. Sci. Technol.*, vol. 28, no. 5, pp. 821–839, 2020, doi: 10.3233/XST-200715.
- [68] J. Droppo and O. Elibol, “Scaling Laws for Acoustic Models,” *arXiv*, Jun. 2021, [Online]. Available: <http://arxiv.org/abs/2106.09488>.
- [69] T. Hashimoto, “Model Performance Scaling with Multiple Data Sources,” 2021.
- [70] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish, “Scaling Laws for Transfer,” *arXiv*, Feb. 2021, [Online]. Available: <http://arxiv.org/abs/2102.01293>.
- [71] D. A. Wood *et al.*, “Labelling Imaging Datasets on the Basis of Neuroradiology Reports: A Validation Study,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Jul. 2020, vol. 12446 LNCS, pp. 254–265, doi: 10.1007/978-3-030-61166-8\_27.