UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

# PH.D. THESIS

IN

INFORMATION AND COMMUNICATION TECHNOLOGY FOR HEALTH

DATA FOR HEALTH – DELIVERY MANAGER OF COGNITIVE COMPUTING FOR NEURONCOLOGY

# FUSEMEDML: EXPLORING THE POTENTIAL FOR A NEW COGNITIVE COMPUTING TOOL IN IDENTIFICATION AND CLASSIFICATION OF BRAIN TUMOURS STARTING FROM MAGNETIC RESONANCE IMAGING

CANDIDATE: DR. CAMILLA RUSSO

TUTOR: PROF. PAOLO MARESCA

Information and Communication Technology for Health (ICTH) – XXXV cycle

# Abstract

This document describes preliminary results of the research activities carried out within the PhD in Information and Communication Technology for Health (ICTH) on the topic of Delivery Manager of Cognitive Computing for Neuroncology at the Department of Electrical Engineering and Information Technologies (DIETI) of the University of Naples “Federico II”. Primary endpoint of this research line is the evaluation of an innovative computational system based on cognitive computing technologies, namely the open-source PyTorch-based deep learning framework for medical data named FuseMedML, for the analysis of magnetic resonance imaging derived data in patients with brain neoplasm of unknown origin; main goal is to obtain a semi-supervised binary classification model which allows the timely identification of the two most common malignant brain tumours of the adulthood, requiring therapeutic strategies and clinical-radiological monitoring different the one from the other. Secondary endpoint is to test whether the proposed binary classification model can predict brain tumour classification more accurately than conventional assessment carried out by trained human readers, in order to determine if the prediction model based on cognitive computing technologies can be able to supplement information and support neuroradiologists in decision-making for daily clinical practice in Neuroncology.

# Keywords

Neuroncology; Glioblastoma; Solitary Brain Metastasis; Neuroradiology; Magnetic Resonance Imaging; Computer-aided diagnosis; Diagnostic performance; Human performance; Artificial Intelligence; Machine Learning; Convolutional Neural Network; FuseMedML.

# Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

# Index

# List of Figures and Tables

### Figures

### Tables

# Abbreviations list

(Alphabetical order)

ADC: apparent diffusion coefficient

AI: artificial intelligence

AUC: area under the curve

BraTS: brain tumour segmentation dataset

CE: contrast enhancement

CNN: convolutional neural network

CNS: central nervous system

DICOM: digital imaging and communications in medicine

DWI: diffusion weighted imaging

ETL: extract transform and load

FLAIR: fluid attenuated inversion recovery

GBM: glioblastoma

IDH: isocitrate dehydrogenase

JPEG: joint photographic experts group

MRI: magnetic resonance imaging

ROC: receiver-operating characteristic curves

SBM: single brain metastasis

TCIA: the cancer imaging archive

TE: echo time

TI: inversion time

TR: repetition time

T1W SE: T1-weighted spin echo

# Chapter 1

# Introduction

Machine learning, defined as a subfield of Artificial Intelligence (AI) in which computers can learn and iteratively improve a task performance without being explicitly programmed but only based on the data collected, is increasingly gaining ground in the medical field thanks to its rapid progresses and over-increasing impact on clinical care, education and research. Most recent successes encompass a wide spectrum of medical sub-specialties, ranging from very early identification of laboratory alterations associated with high morbidity and mortality conditions to automated interpretation of imaging findings [1]–[3]. It can therefore be reasonably assumed that AI and machine learning can progressively revolutionize image processing pipelines and be integrated into routinary clinical workflows; especially in the radiological field, as we are moving into the era of computer-aided diagnosis, the application of AI tools based on machine learning and cognitive computing in imaging pattern recognition may implement diagnostic procedures and support physicians in daily clinical practice for what concerns imaging-guided procedures or imaging-related decisions. From a practical standpoint, AI technologies supplement information for humans to make decisions, by resorting to algorithms learning from large data sets to solve problems and continuously apprehending from constantly changing data and/or self-correction mechanisms. These processes mimic the human cognitive functions in stratified learning and problem solving, in order to perform higher-order complex data synthesis in a shorter time.

Due to its intrinsic properties, one of the most promising fields of application of cognitive computing and AI techniques to medical imaging is represented by brain imaging, more specifically in the field of Neuroncology [[4]]. Neuroncology is the branch of medicine that focuses on the study of tumours arising in or from the central nervous system (CNS). Neuroncology imaging consists of diagnostic methods that non-invasively evaluate brain tumours before, during and after treatments, with Magnetic Resonance Imaging (MRI) representing the most reliable and widespread used technique. Indeed, MRI plays a key role in the identification of brain tumours and has become a mandatory step during preoperative evaluation to aid determination of tumour type, grade and overall prognosis; information provided by MRI examination can also influence surgical choices and radiation treatment planning, help in predicting drug efficacy, and monitoring treatment response and efficacy [[5]]. Among different brain tumour types, from an epidemiological standpoint the two most prevalent entities in western adult population are represented by brain metastases and high-grade gliomas [6]. Brain metastases, defined as secondary localizations to the CNS of primary neoplastic lesions arising in another organ or system, represent the most common malignant brain tumour of the adulthood; generally multiple and more frequently diagnosed in advanced disease stages in patients with a known cancer history, metastases can rarely be single and represent the first clinical manifestation in patients with no previous history of systemic cancer. Primary brain tumours, defined as neoplastic lesions originating from the cells of the nervous tissue, are far less common. Among these lesions, the most frequent subtype in the adulthood is represented by glioblastoma (GBM), a primary high-grade neoplasm arising from glial cells located within the nervous tissue; this kind of lesion is characterized by significant local biological aggressiveness, low tendency to dissemination outside the CNS and overall poor prognosis. Patients with clinical signs and symptoms suspected of the presence of an expansive brain lesion usually undergo brain imaging, largely relying on MRI examination; the results of neuroimaging examinations are required to guide the most appropriate diagnostic-therapeutic workflow for each patient, in order to confirm diagnostic suspicion and optimize treatment planning as well as long-term follow-up. However, at MRI examination GBMs and single brain metastases (SBMs) may present with similar features on conventional imaging, thus raising important issues in terms of early identification and differentiation; in these cases, further and more invasive investigations, which are both time-consuming and cost-consuming at the same time, are frequently required [6]. With this background, in recent years several strategies of automated computer image analysis have been explored as a potential support for physicians to provide higher diagnostic accuracy both in tumour recognition and classification [7].

The core of the presented research activity is to evaluate the role for innovative computational systems based on visual recognition and cognitive computing technologies in enhancing MRI characterization of brain neoplasms, with specific reference to the recently introduced open-source PyTorch-based deep learning framework for medical data FuseMedML [8]; FuseMedML is a comprehensive machine learning library developed by IBM research group on Artificial Intelligence on Healthcare and Life Sciences Discovery, that mainly focuses on the biomedical domain. It offers several tools for accelerated machine learning, with the goal of simplifying and streamlining medical research activities, especially in terms of model training and evaluation [9].

Such technology harbours the potential to aid neuroradiologists for diagnostic-prognostic-therapeutic purposes and limiting inter/intra-observer variability either in qualitative or quantitative images interpretation. As ancillary but not less important purpose, the paper also explores the difference in diagnostic performances between the above-mentioned computational system and human visual raters with different expertise in the field of Neuroncology imaging, thus analysing whether they could benefit from assistance by pre-trained image classification tool.

# Chapter 2

# Background, aims and scopes

## 2.1 Clinical and Radiological Background

Adult intra-axial brain neoplasms are expansive lesions affecting the CNS, originating within the brain or the spinal cord (differently from extra-axial lesions which instead originate outside the brain from nerves, meninges, etc.). Intra-axial tumours can be mainly classified into primary and secondary; primary lesions derive from nervous cells within the nervous tissue, with variable biological aggressiveness; conversely, secondary lesions are malignant intra-axial localizations resulting from the metastatic spread to the nervous tissue of a primary neoplasm of a different organ or tissue.

Regardless of lesion type, the clinical manifestations due to the presence of an intra-axial mass are extremely variable and strongly influenced by location and size, with no substantial difference in the case of primary or secondary CNS tumours. Symptoms and signs largely rely on infiltrative phenomena on the nervous tissue in eloquent areas, compression and distortion of the healthy nervous tissue, uncontrolled raise in intracranial pressure, obstacle to normal cerebrospinal fluid circulation, and infiltration of vascular structures close to the tumour. The onset of patient-reported symptoms may be insidious and non-specific, ranging from headache to nausea and vomiting, from vertigo to unexplained behavioural changes; epilepsy may be another presenting symptom, therefore all patients with a first epileptic seizure episode in adulthood must be directed to instrumental examinations of second level to rule out an unrecognized brain neoplasm. Focal neurological symptoms such as speech disturbance, motor or sensitive deficits are less usual and more commonly observed in case of lesions located in eloquent cortical areas; finally intra-axial lesions can also be asymptomatic, thus represent an occasional finding in case of radiological examinations carried out for other reasons, as well as during disease staging when brain examinations are performed to rule out possible secondary CNS lesions [6]. In case of brain lesion suspected for GBM or SBM, contrast-enhanced MRI still represents the golden standard first-line examination for both non-invasive characterization and pre-surgical planning [10]–[12].

### 2.1.1 Conventional MRI: basic considerations

Contrast-enhanced MRI is the diagnostic modality of choice in patients suspected with brain tumour, as it provides both bidimensional and tridimensional imaging of the lesion on the three orthogonal planes, allowing for its location, size and extent depiction with optimal tissue characterization and nearby structures discrimination; moreover, the use of advanced MRI techniques (such as for example perfusion weighted imaging, functional MRI, magnetic resonance spectroscopy, or diffusion tensor tractography) can provide additional information, crucial for both pre-operative planning and response to therapy monitoring. Minimum requirements for conventional MRI examination in patients with brain tumour should always include at least one volumetric acquisition on T2w and on pre/post-contrast T1w (before and after gadolinium-based contrast media administration); the use of T2* acquisitions or, even more, susceptibility weighted imaging have completely replaced computed tomography for the identification of haemorrhagic foci and inner calcifications; diffusion weighted imaging (DWI) with relative apparent diffusion coefficient (ADC) maps is essential to define tumour cellularity and exclude vascular complications. Despite today perfusion acquisitions should always represent an integral part of basic MRI examination for intra-axial tumour first diagnosis, it is still not routinely used, regardless of the efforts to standardize their acquisition and provide comparable results [13]; conversely, advanced techniques are used from time to time by the neuroradiologist in accordance with the neurosurgeon, depending on anatomical characteristics and on the contribution they can provide for therapeutic planning.

Brain computed tomography is currently reserved to patients with formal counter indication to MRI (i.e., incompatible pacemakers and implantable devices), in case of emergencies (i.e., haemorrhagic transformation, ischemia, etc), or in case of patient refusal. Finally, radiometabolic techniques and nuclear medicine investigations (such as brain scintigraphy or positron emission tomography) are used as integration to MRI examination in selected or doubtful cases. Although MRI is crucial for brain tumour diagnostic workout, the final diagnosis still largely relies on histopathological examination after biopsy or tumour resection, coupled to immunohistochemical and molecular biology analysis [14].

### 2.1.2 Glioblastoma

Among primary lesions of the CNS, the most frequent are those arising from the glial cell-line, with GBM representing the most common histologic type in adults (as well as the one associated with the worst prognosis); despite its relatively low incidence compared to other tumour types (between 5-20 cases per 100,000), due to its biological aggressiveness GBM is responsible for about 2% of all cancer mortality, a trend that has been increasing over the past three decades. GBM can be further divided in two sub-categories depending on the absence or presence of a mutation in the two isocitrate dehydrogenase enzymes involved in cytoplasmic (IDH1) and mitochondrial (IDH2) conversion of alpha-ketoglutarate to D-2-hydroxyglutarate. In most cases, GBM adult-variant is represented by IDH wild-type lesions (i.e., de novo putative mutations, with brain tumour probably resulting from the rapid malignant transformation of an astrocytic precursor). In a minority of GBM patients, a mutation in IDH genes can be detected at pathological examination, thus suggesting GBM arising from malignant transformation of a lower grade astrocytoma (of which GBM represents the pathological continuum).

On MRI IDH-mutated and wild-type GBMs are indistinguishable, with evidence of diffusely infiltrating lesion with irregular margins and large necrotic-haemorrhagic core; solid components usually show weak DWI restriction and intense post-contrast enhancement, with a variable raise in perfusion parameters due to increased angiogenesis; calcification on T2* imaging is sporadic. Intense peri-lesional infiltrative oedema is also present, with variable invasion of the adjacent structures. An example of wild-type GBM MRI appearance is shown in Figure 1.

Depending on the number of observed lesions, GBM can also be divided into monofocal (single enhancing lesion with variable infiltrative oedema halo), multifocal (multiple enhancing lesions with a single halo of infiltrative oedema) and multicentric (multiple enhancing lesions, each one with its own halo of infiltrative oedema). The most important differential diagnoses should include brain metastases (specially SBM, typically large and hypervascular), as well as the far uncommon gliosarcoma.

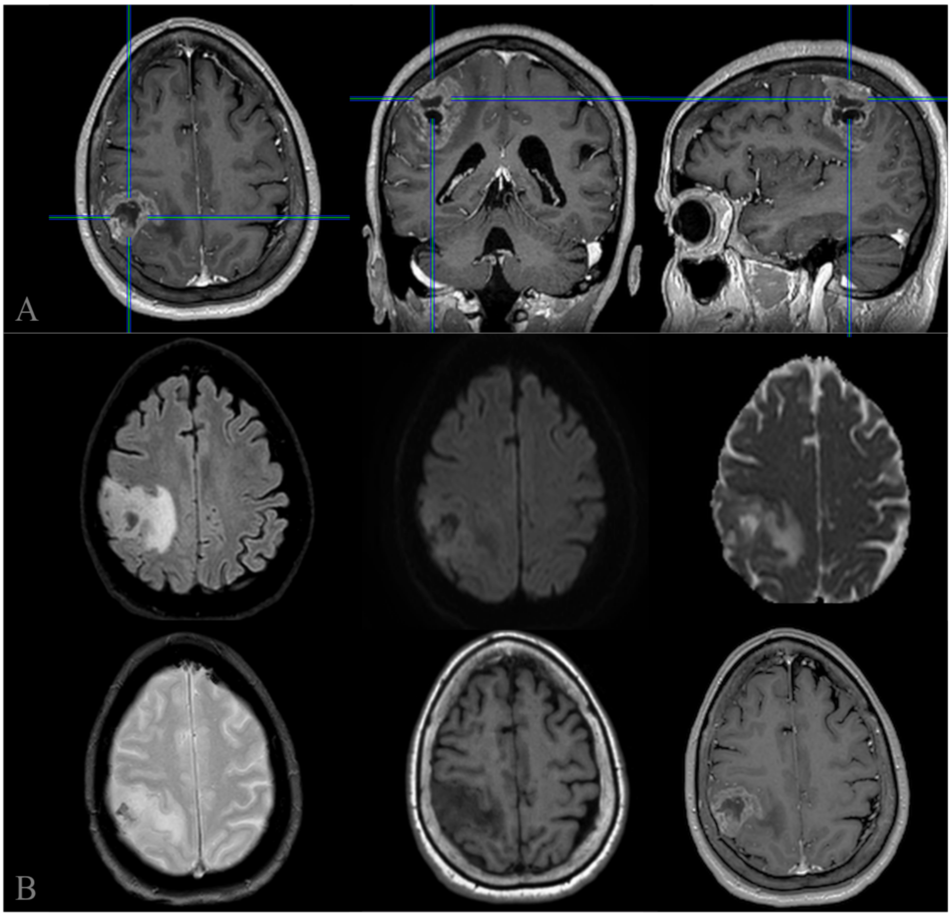


*Figure 1. Conventional MRI imaging in a 59 years-old patient newly diagnosed with suspected right frontal-parietal high-grade glioma. (A) Axial post-contrast T1w 3D imaging with orthogonal reconstruction. (B) Multiparametric 2D imaging at the same level of the lesion: upper row (left to right), FLAIR, DWI (b=1000) and relative ADC map; bottom row (left to right), T2*, T1w GRE and contrast-enhanced T1w GRE. Final histopathological analysis after partial tumour resection confirmed the diagnosis of glioblastoma, wild-type.*

### 2.1.3 Solitary brain metastasis

Metastases account for over a half of all brain neoplasms, far more common in case of primary lung, breast and skin tumours. Their incidence increased in recent years (prevalence ranging between 20-40% depending on the primary lesion type); such an increase is to be attributed both to the overall longer survival rate of cancer patients due to new therapeutic discoveries, as well as to technical progresses in the field of prevention and diagnostic technologies. Although the most common, intra-axial spread is only one of the possible patterns of secondary diffusion of primary neoplasm to the CNS, along with leptomeningeal-subarachnoid, dural, intra-ventricular, peri-neural and peri-vascular spreading (however these variants go far beyond the purposes of this dissertation). About 80% cases have supratentorial localization to the brain, mainly at the cortico-subcortical junction level due to hemodynamic factors related to hematogenous dissemination. Generally multiple and associated with further metastatic localizations to other organs and systems, CNS metastases can also be isolated; in case of SBM and in case of silent patient's previous medical history, differential diagnosis between SBM and GBM may be very challenging.

Indeed, imaging features of SBM may often resemble to the one observed in primary lesion (i.e., intra-lesional calcifications, cystic-necrotic components, micro-haemorrhagic foci, and so on); involvement and distortion of ventricular system resulting in hydrocephalus, dislocation of nervous or vascular structures with subsequent mass effect in an inextensible closed space and vascular complications are common to both SBM and GBM. One of the most distinctive features of brain metastases compared to GBM is the large vasogenic oedema peritumoral halo, usually disproportionate to the actual size of the lesion. On MRI, SBM appears as a focal hypointense area on T1w (except for melanin-rich and haemorrhagic metastases) and T2w sequences (except for cystic and mucinous tumours, which often show intralesional hyperintense components); peri-tumoral vasogenic oedema is usually disproportionate to the actual size of the lesion, while post-contrast enhancement of the solid components is typically intense. On DWI signal restriction depends on tumour cellularity (lower for adenocarcinomas, higher in sarcomas and neuroendocrine tumours); on PWI, an overall increase in perfusion parameters is generally observed, the greater the more the tumour is provided with anarchical blood vessel supply. An example of SBM from lung microcitoma MRI appearance is shown in Figure 2.

The main differential diagnosis is represented by GBM, both in terms of frequency and MRI appearance; rarely, SBMs can be confused with infectious-abscess lesions, demyelinating tumefactive lesions or vascular alterations simulating haemorrhagic metastases.

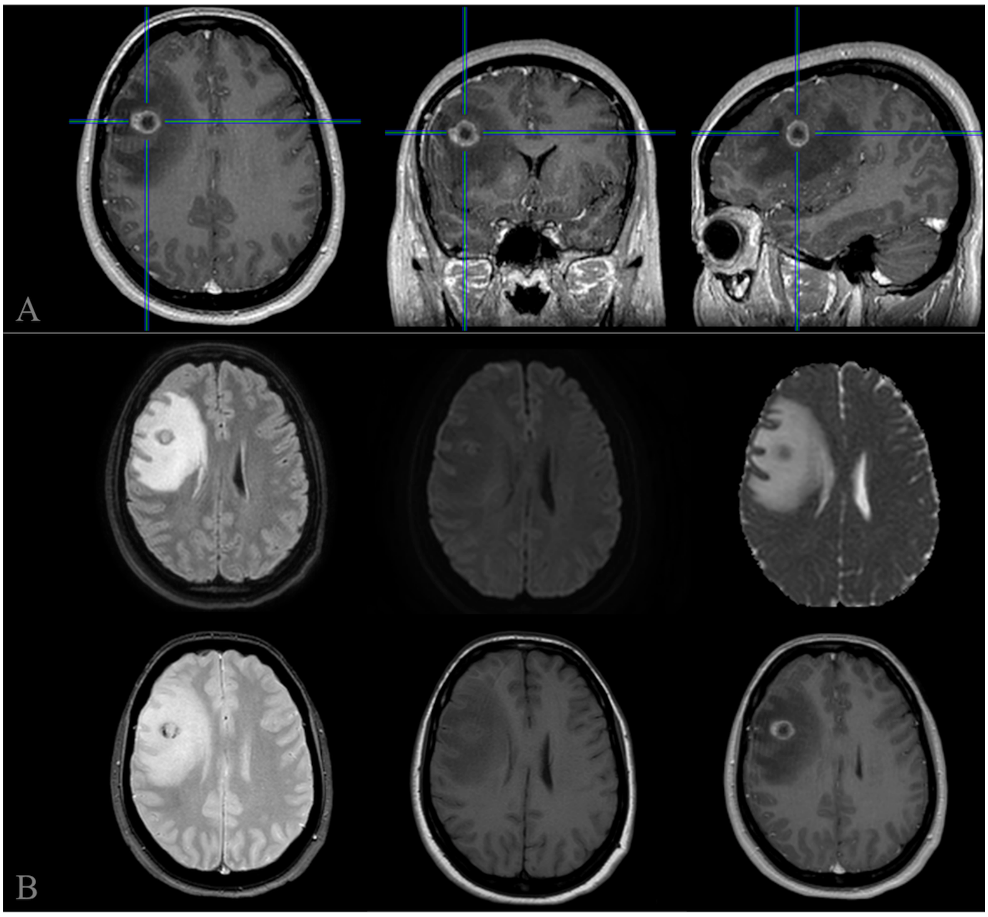


*Figure 2. Conventional MRI imaging in a 67 years-old patient diagnosed with suspected single right frontal metastasis from unknown primary malignancy. (A) Axial post-contrast T1w 3D imaging with orthogonal reconstruction. (B) Multiparametric 2D imaging at the same level of the lesion: upper row (left to right), FLAIR, DWI (b=1000) and relative ADC map; bottom row (left to right), T2*, T1w SE and contrast-enhanced T1w SE. Multiphase total-body CT scan revealed the presence of a primary lung malignancy with multiple nodal metastases; primary tumour biopsy confirmed the diagnosis of pulmonary microcitoma.*

## 2.2 Rationale

As previously stated, at MRI examination GBM and SBM may present with similar features on multiparametric conventional imaging, thus raising important issues in terms of early identification and differentiation; in these cases, further and more invasive investigations, which are both time-consuming and cost-consuming at the same time, are frequently required.

With this background, as primary endpoint we aim to test the application of visual recognition and cognitive computing for transfer-learning-based pre-trained models' application in MRI pattern recognition of these two different brain tumour types, highlighting possible decision-support implications in clinical practice. Visual recognition, defined as a subcategory of cognitive computing and AI technologies, represents a set of methods for detecting and analysing images in order to identify objects and other elements within a pool of provided images, and then performing higher-order complex data synthesis to solve a given problem. As we previously stated, medical innovations focusing on precision medicine for cancer diagnosis and treatment are considered one of the most challenging domains to be revolutionized by AI and, in addition to many academic efforts, companies are getting increasingly involved in the process. In our experimental setting we tested the ability of such deep learning method, based on transfer learning and pre-trained convolutional neural networks (CNN), to increase the probability of correct brain neoplasms allocation and identification starting from multiparametric conventional MRI, in order to ensure prompt diagnostic workout and optimize treatment planning. Among possible available tools we decided to concentrate our attention on the recent IBM developed open-source product named FuseMedML, a Python framework designed for machine learning applications to medical domain, specifically projected with the goal of promoting flexibility, code reusability and easy collaboration. The use of FuseMedML library gives the chance to study the benefits of this new library over the direct use of the mainstream frameworks it is based on. Currently publicly available, FuseMedML was preliminarily proved successful in different medical scenarios before open-sourcing it [9], [15]–[20] and at present offers a range of tools covering the entire development process, from data organisation to model training and evaluation.

Coupled to this first experimental pilot project to test usefulness and feasibility of FuseMedML application to classify brain tumour types based on MRI images, we also aim to compare the neural network performances in classifying these tumours to human-level classification performances. Indeed, comparison between machine learning versus human reader performances still represents a heretofore unexplored sector, with few evidence collected in different medical fields [21]. To date, thanks to the overall lower variability and increased interpretation consistency reached by AI tools, the use of automated classifications seems promising in improving diagnostic accuracy and the predicting outcomes; however only few studies tested direct head-to-head comparisons between AI and conventional human interpretation of MRI images, especially in the field of Neuroncology [22]. Therefore, as a secondary endpoint, we tested whether the proposed binary classification model based on FuseMedML can predict brain tumour diagnostic category more accurately than conventional analysis by human readers. To this purpose, radiologists with different expertise in brain MRI interpretation were involved in the rating, in order to reproduce real-world clinical practice framework and expected variability among human observers; with the same aim, we also included in the analysis extreme cases such as MRI images degradation or distortion, speculating on diagnostic confidence levels importance when making clinical decisions.

## 2.3 Project evolution over time

Among several tools developed for image classification and allocation by companies involved in AI medical industry, different computer vision applications released over time by IBM have been thoroughly and progressively tested for image recognition, analysis and classification in the field of Neuroncology. We first started from IBM Watson™ Visual Recognition system, a tool developed for several health applications that had been already used as a prototype for pattern recognition in precision oncology [23]; IBM Watson™ Visual Recognition could be defined as a service (subject to charges) previously available on IBM Cloud, which enabled images' tagging, classification, and inspection by means of machine learning algorithms. IBM Watson™ initially received considerable attention for its focus on precision medicine, with specific reference to cancer diagnosis and treatment. However, early enthusiasm for this application has given way to a certain scepticism due to some difficulties in training and integrating Watson™ into actual diagnostic processes, clearly indicating cancer diagnosis support as an overly ambitious objective for this tool [24]. Therefore, in a preliminary approach we tested the application of Watson™ for MRI pattern recognition of GBMs versus SBM, with promising results despite a relatively low number of instances collected at the very beginning of this preliminary analysis. However, due to the costs related to the use of a cloud repository, the over-increasing amount of input imaging data due to progressive data pool enrichment and the imminent discontinuation of the above-mentioned IBM product, we then moved to IBM Maximo® Visual Inspection (that however shared some of the weaknesses of its predecessor, including technology usage costs, and was not specifically designed for medical imaging). Similarly to Watson™, also Maximo® Visual Inspection was rapidly discontinued; it is reasonable to think that, besides implementation issues and difficulties in adapting these technologies to medical knowledge, both products also suffered from competition with other similar open-source programs provided by different vendors [7].

Consequently, we finally moved to FuseMedML, a fully open-source platform for machine learning mainly relying on PyTorch syntax and on TensorFlow+Keras frameworks, in order to deploy a cognitive computing-based support for brain tumours MRI differential diagnosis [8]. FuseMedML was specifically developed by IBM research group on Artificial Intelligence on Healthcare and Life Sciences Discovery with the purpose of simplifying and streamlining medical research projects; built on top of popular machine learning frameworks such as PyTorch, it also includes domain-specific capabilities to complement these frameworks [9]. This choice has been guided by the will of experimenting this recently released framework on new data (no evidence collected since now in the Neuroncology field) and to exploit reproducibility and reusability for medical research purposes. Additional details on FuseMedML are provided in the project description below.

# Chapter 3

# Materials and Methods

This section illustrates the characteristics of the dataset used for the analysis and its pre-processing, as well as the technologies adopted for developing the proposed neural network architecture. Procedures adopted for human readers comparison analysis are described in the remainder of the section.

## 3.1 Dataset selection

We retrospectively identified patients who had undergone brain MRI as suspected of brain neoplasm, who had been subsequently confirmed to arbor monofocal GBM or SBM. MRI examinations were acquired between 2015 and 2021 in different centres with 6 different MRI scanners for clinical, diagnostic and pre-surgical purposes, in accordance with the 1964 Helsinki Declaration and its subsequent amendments; MRI scans characteristics are listed in Table 1.

| Vendor | System | Field | Head coil |
|---|---|---|---|
| Philips | Ingenia | 1.5T | 8 channels |
| GE | Signa Voyager | 1.5T | 8 channels |
| Siemens | Magnetom Trio | 3T | 16 channels |
| Philips | Achieva | 1.5T | 8 channels |
| GE | Explorer | 1.5T | 8 channels |
| Philips | Ingenia | 3T | 16 channels |

*Table 1. main features of the 6 scanners used for MRI examinations.*

The dataset was then enriched by using two publicly available MRI images repository, namely the BraTS (BRAin Tumour Segmentation) dataset that includes scans from 19 institutions [25], and the TCIA (The Cancer Imaging Archive) dataset that includes scans from 8 institutions [26]— consisting of two main sources, namely the Ivy Glioblastoma Altas Project, and The Cancer Genome Atlas Glioblastoma Multiforme collection. We eventually collected an overall number of 947 unique patients (mean age 59,7±14,1; M:F 1,3:1), of which 59% received histopathological diagnosis of GBM (after biopsy or partial/total tumour resection) and 41% of SBM from an unknown primary tumour (as confirmed by further instrumental investigations and subsequent direct histopathological assessment); despite the final ratio of number of subjects shows a minimum imbalance (1,4:1), such distribution between the two groups can be considered representative of the actual distribution of the two disorders in the general population.

Images' preliminary evaluation and patients' inclusion/exclusion was performed by a 10-years experienced neuroradiologist. Patients with multiple brain lesions (N≥2), as well as patients with imaging evidence of significant brain comorbidities (i.e., previous major ischemic stroke, leptomeningeal diseases, venous thrombosis with brain infarction, etc.) were excluded from the analysis. Similarly, MRI images severely vitiated by motion- or device-related artifacts were also excluded; conversely MRI images only partly vitiated by motion artifacts even though still usable for clinical diagnosis (a common scenario faced by neuroradiologist in daily clinical practice) were included in the analysis. Subjects with incomplete examination and absence of post-contrast T1w imaging were also excluded. Flow-chart for patients' selection and exclusion criteria is shown in Figure 3.

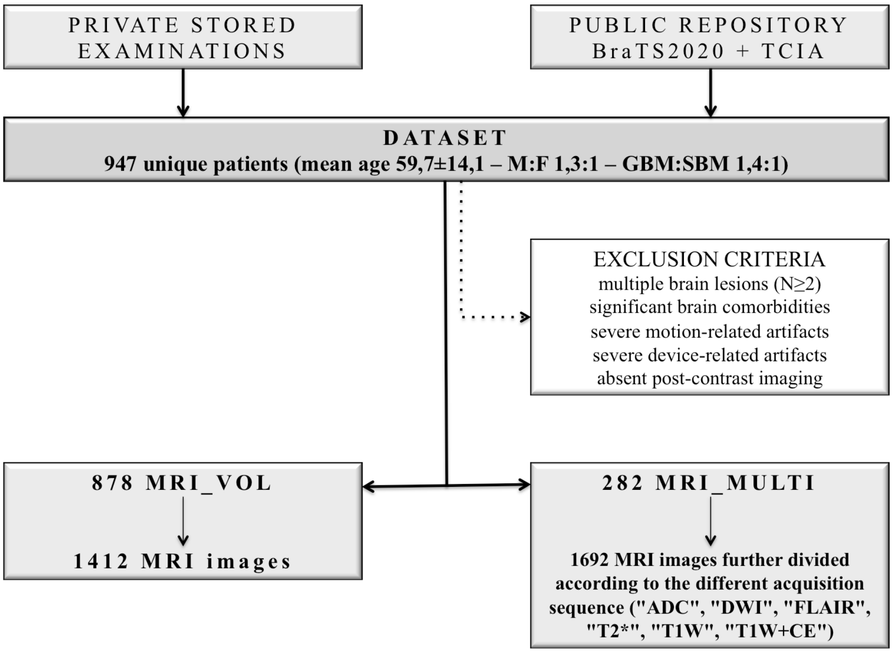

PRIVATE STORED EXAMINATIONS
PUBLIC REPOSITORY BraTS2020 + TCIA
DATASET
947 unique patients (mean age 59,7±14,1 – M:F 1,3:1 – GBM:SBM 1,4:1)
EXCLUSION CRITERIA
multiple brain lesions (N≥2)
significant brain comorbidities
severe motion-related artifacts
severe device-related artifacts
absent post-contrast imaging
878 MRI_VOL
1412 MRI images
282 MRI_MULTI
1692 MRI images further divided according to the different acquisition sequence ("ADC", "DWI", "FLAIR", "T2*", "T1W", "T1W+CE")

*Figure 3. Flow chart of patients' selection; data extraction was performed between 2021 and 2022.*

For each patient, the most representative MRI slice of the brain neoplasm was selected by the same 10 years-experienced neuroradiologist, then exported from DICOM (Digital Imaging and Communications in Medicine) format to compressed JPEG (Joint Photographic Experts Group) format; to avoid possible oversampling, lesions with maximum diameter <5cm were sampled once at the most representative level, while larger lesions with maximum diameter >5cm were sampled twice at the level of the two most representative slices (minimum between-slices distance 3cm). An overall number of 878 patients was used to populate a group of subjects with available post-contrast volumetric T1W MRI imaging obtained for pre-operative neuronavigation purposes (named MRI_VOL), and 282 patients were used to populate a subgroup of subjects from the same studies with complete multiparametric MRI obtained for diagnostic purposes (named MRI_MULTI). From these two groups, we finally acquired and stored 1412 anonymized post-contrast volumetric T1W MRI images (MRI_VOL group), and 1692 anonymized multimodal MRI images further subdivided according to the different acquisition sequence (MRI_MULTI group – based on the sequence type, respectively divided into "ADC", "DWI", "FLAIR", "T2*", "T1W", "T1W+CE") and grouped in the above-mentioned 282 actual samples. Minimum sequences' requirements both for MRI_VOL and MRI_MULTI images inclusion are listed in Table 2.

| | **T1w** | **DWI** | **T2*** | **FLAIR** | **T1w+CE** |
|---|---|---|---|---|---|
| **TE** | 2-19ms | variable | variable | 120-155ms | 2-20ms |
| **TR** | 300-3600ms | variable | 1-30ms | 6000-11000ms | 5-3300ms |
| **TI** | - | - | - | 2000-2200ms | - |

*Table 2. Details of the MRI acquisition parameters for the structural weighted images (T1w, DWI+ADC; T2*, FLAIR, and a T1w+contrast enhancement, CE) used for clinical-diagnostic purposes: echo time (TE), repetition time (TR), and inversion time (TI).*

In both datasets the samples (single in the case of MRI_VOL and multiple in the case of MRI_MULTI) have been divided into two classes, according to the underlying tumour type (i.e., GBM and SBM).

## 3.2 Data pre-processing

MRI data have been anonymized, extracted, uploaded and stored in a private repository; images of the most representative MRI slice or slices (depending on tumour size) of the brain neoplasm, preliminarily selected as described before and originally acquired in DICOM format, were converted to supported JPEG format. This phase was based on a manual process at present (still need for supervised ETL – extract, transform, and load – automatization and weakly supervised target lesions' segmentation). All selected MRI slices in our datasets contained undesired uninformative spaces around the salient and informative portion of the image, whose presence could potentially affect classification performances. Hence, it was necessary to remove unwanted areas before further proceeding in the analysis. A pre-processing including resizing and cropping based on extreme point calculation was carried out on these images to eliminate uninformative areas around the region of interest and ensure homogeneous figures' dimension, by means of OpenCV and Python. First, each image is resized using the PyTorch function:

```
torchvision.transforms.Resize(size)(image)
        # transform for square resize
     transform = T.Resize(224 x 224 pixels)
```

This function does not modify the used image, but instead returns another squared image with newly defined dimensions (homogeneous compared to the remaining collected items). Edge point calculation for object detection process (Canny edges detection) was then performed to separate desirable foreground image objects from the background based on the difference in pixel intensities of each region [27], [28]. We loaded the resized images, then we used the function for object segmentation from extreme points in contours with OpenCV:

```
extreme_points.py
```

We first imported the required packages:

```
import imutils
import cv2
```

then loaded the image, convert it to grayscale to enhance contour detection, and blur it slightly in order to ease subsequent thresholding:

```
image = cv2.imread(“MRI_001.png")
gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
gray = cv2.GaussianBlur(gray,(5,5),0)
```

We then applied thresholding to convert them into binary images, performing dilations and erosions operations to remove the noise of images, by using the following Python methods:

```
cv2.threshold()
cv2.erode()
cv2.dilate()
```

After that, we used the:

```
findContours()
```

method of OpenCV library to find all boundary points of the threshold area of interest in the image and select the largest contour by calculating the four extreme points (extreme top, extreme bottom, extreme right, and extreme left):

```
 extLeft = tuple(c[c[:, :, 0].argmin()][0])
extRight = tuple(c[c[:, :, 0].argmax()][0])
  extTop = tuple(c[c[:, :, 1].argmin()][0])
  extBot = tuple(c[c[:, :, 1].argmax()][0])
```

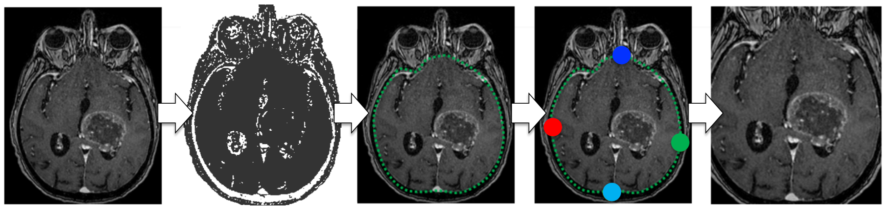
to then draw the outline of the object and each of the extreme points (where the left-most is red, right-most is green, top-most is blue, and bottom-most is light blue):

```
              cv2.drawContours() function
 cv2.circle(image, extLeft, radius, color, thickness, lineType, shift)
cv2.circle(image, extRight, radius, color, thickness, lineType, shift)
  cv2.circle(image, extTop, radius, color, thickness, lineType, shift)
  cv2.circle(image, extBot, radius, color, thickness, lineType, shift)
```

As final step, we crop the image based on the information on contours and extreme points, using the following PyTorch function:

```
torchvision.transforms.functional.crop()
```

An example of the result of Canny edges detection is shown in Figure 4.



1. load the original image

2. apply smoothing and thresholding

3. find the outer contour

4. find the edge points

5. crop the image

*Figure 4. MRI images pre-processing: example of figure smoothing, thresholding and cropping by using extreme points calculation with zero-parameter automatic Canny edge detection (Finding Extreme Points in Contours with OpenCV. In PyImageSearch; see references in the main text).*

The decision to adopt Canny edge detection relates to the fact that in brain tumour the most representative slice generally contains several useful clues not only within the lesion itself, but also in the area around the lesion (i.e., peritumoral oedema, ventricular system distortion, sulci effacement, and so on); these additional features may differ across different tumour types. Therefore, a cropping approach preserving the brain parenchyma/cerebrospinal fluid spaces around the main lesion allows for background removal (uninformative areas around the skull) but foreground with possible ancillary imaging features preservation, that could be useful at the time of model training and evaluation (brain parenchyma/cerebrospinal fluid spaces inside the skull, within and around the monofocal brain lesion).
Although the corpus of collected evidence could already be considered relatively large compared to what previously reported in scientific literature, nonetheless we performed image augmentation on the dataset. Image augmentation is a technique that artificially modify the dataset, creating multiple copies of the original images composing the dataset itself, with different orientation, brightness, rotation, and so on; this approach is intended to improve the classification accuracy of the selected predictive model by augmenting the existing data rather than collecting new ones. To the purpose, we adopted three augmentation strategies and preliminary performed data augmentation generating image translations, brightness adjustment and horizontal flipping, by using the module `FuseAugmentorBase` within FuseMedML (see below).

## 3.3 Model definition and design

Our method analysed MRI data by using the deep learning framework for medical data FuseMedML [8], [29]. As anticipated in introduction section, FuseMedML is an open-source python-based framework designed to improve code reuse and accelerate discoveries in in the biomedical field through advanced technologies of machine learning; its initial release supports Python 3.6 and PyTorch 1.5. The best way to install FuseMedML is to clone the Github repository and install it in an editable mode using `pip`:

```
!git clone https://github.com/IBM/fuse-med-ml.git
%cd fuse-med-ml
!pip install -e .
```

This mode installs all the currently publicly available domain extensions. As an alternative, it is possible to install FuseMedML from PyPI by using the command:

```
$ pip install fuse-med-ml[all]
```

Import of Python and Fuse Libraries from Project home page (`https://github.com/libfuse/python-fuse`) was then performed, and the metric definition and calculation were then carried on by using the module `FuseMetricBase` within FuseMedML. The MetricBase basic class defines the interface for a metric implementation, and it consists of collect, set, reset and eval methods.

```
import os
from typing import OrderedDict
import torch
import torch.nn.functional as F
import torch.optim as optim
import torchvision
from torch.utils.data.dataloader import DataLoader
from torchvision import transforms, datasets
from fuse.eval.evaluator import EvaluatorDefault
from fuse.data.dataset.dataset_wrapper import
FuseDatasetWrapper
from fuse.data.sampler.sampler_balanced_batch import
FuseSamplerBalancedBatch
from fuse.losses.loss_default import FuseLossDefault
from fuse.managers.callbacks.callback_tensorboard import
FuseTensorboardCallback
from fuse.managers.manager_default import FuseManagerDefault
from
fuse.eval.metrics.classification.metrics_classification_commo
n import␣
,!MetricAccuracy, MetricAUCROC, MetricROCCurve, MetricAUCPR,␣
,!MetricConfusionMatrix
from
fuse.eval.metrics.classification.metrics_thresholding_common
import␣
,!MetricApplyThresholds
from fuse.models.model_wrapper import FuseModelWrapper
from fuse_examples.tutorials.hello_world.hello_world_utils
import LeNet,␣
,!perform_softmax
```

```
from fuse.data.augmentor.augmentor_toolbox import
aug_image_default_pipeline
```

Among implemented metrics for classification problems that can be used we mainly referred to:

- `MetricROCCurve` → it can be used to calculate the receiver operating characteristic (ROC) curve, and not just the area under the curve (AUC) underneath it.
- `MetricAUC` → it can be used to calculate AUC from ROC curve.
- `MetricAUCPR` → it calculates the area under the Precision-Recall curve.
- `MetricAccuracy` → it computes the accuracy
- `MetricConfusion` → it computes the following of the possible metrics: sensitivity, recall, true positive ratio, specificity, precision, and F1 score.
- `MetricConfusionMatrix` → it computes a multi-class confusion matrix.

Some of the advantages of the software design of FuseMedML framework we can list encompass:

- good modularity;
- rapid, flexible and scalable development;
- encouragement to share and collaborate;
- standard ratings;
- expertise in medical imaging;
- interoperability between different frameworks.

CNNs are perhaps the most used Deep Learning algorithm in computer vision for object recognition and identification. For our purpose, we decided to adopt a pre-trained CNN model (created and trained to solve a problem that share structural similarities with our problem) based on very conspicuous datasets (in this specific case, ImageNet) to create a large neural network for image classification (namely, VGG16). CNNs have the advantage of extracting meaningful information directly from data, eliminating the need to extract features manually. The use of CNN is widespread due to three important factors:

- features are learned directly from CNN;
- CNN produces highly accurate results;
- CNN can be retrained for new visual recognition activities on increasingly growing amount of image data.

From an architectural standpoint, a CNN consists of an input layer, an output layer and many intermediate “hidden” layers, which perform operations that connect and interpolate the data in order to infer their specific features;

intermediate layers of nodes between the input and output layers are also named “hidden” because they are not directly observable from the systems inputs and outputs. Some of the most important layers in this architecture include N number of convolution and pooling, used for ordering layers within the CNN and iteratively repeated over many layers. The CNN has neurons with weights and biases, used to learn during training and constantly updated after each iteration. After learning the features, the architecture of a CNN moves on to classification: the penultimate layer is represented one or more fully connected layers that emits a vector of dimensions equal to the number of classes that the network will be able to predict. That vector expresses the probability associated to each image classification.

Among the considered pre-trained CNNs, the best candidate for this type of analysis is the VGG16 network. VGG16 is one of the most widely used pre-trained CNN for image classification, developed by the Visual Graphics Group of the Oxford University [30]. This model achieves 92.7% accuracy in ImageNet's 5 top-tests and consists of 13 convolutional levels, 5 pooling layers, and 3 dense layers. The version of VGG16 that has been used for our purposes is the one provided by the Keras submodule of TensorFlow [31]:

```
tf.keras.applications.vgg16.VGG16
```

The default input size for this model was 224x224. VGG block diagram is shown hereafter in Figure 5.

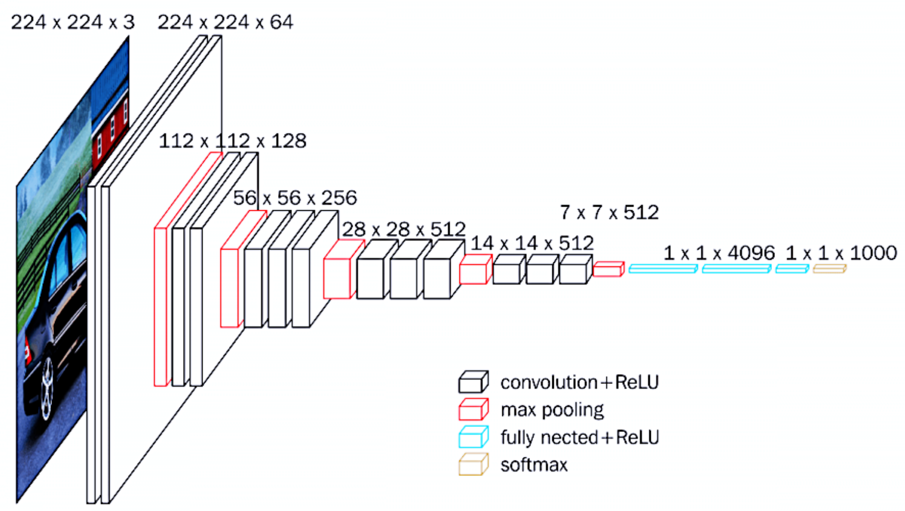


*Figure 5. Structural details of a VGG16 network (source:neurohive.io).*

A diagrammatic illustration of the transfer learning for the proposed CNN-based architecture is shown in Figure 6.

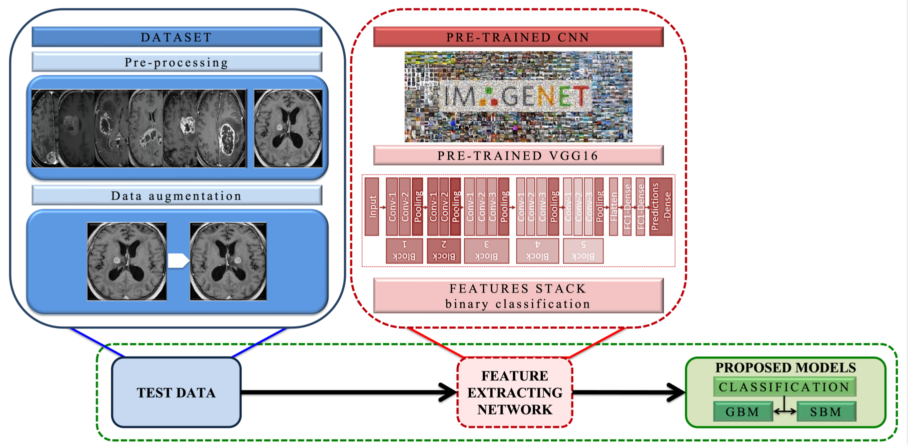


*Figure 6. Stepwise illustration of the transfer learning for CNN-based architectures: CNN models are pre-trained on images from ImageNet (VGG16) and used as feature extractors; pre-processed MRI images underwent data augmentation to increase the number of the data sample.*

Hereafter, we finally describe the models that have been used and their specific architecture. For the MRI_VOL group, in which the samples to be classified are represented by a single image, the proposed architecture is shown in Figure 7.

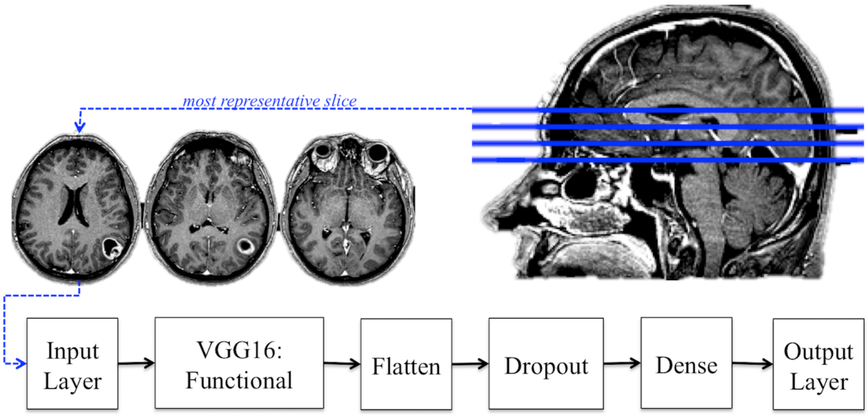


*Figure 7. Model design and proposed architecture for MRI_VOL group.*

For the MRI_MULTI group, the six above-mentioned images categories (namely, "ADC", "DWI", "FLAIR", "T2*", "T1W", "T1W+CE") are provided as input to a single instance of the CNN, which in turn provides the relevant features for each, which are then concatenated to perform a single prediction model based on the information obtained on the whole dataset; the proposed architecture is shown in Figure 8.

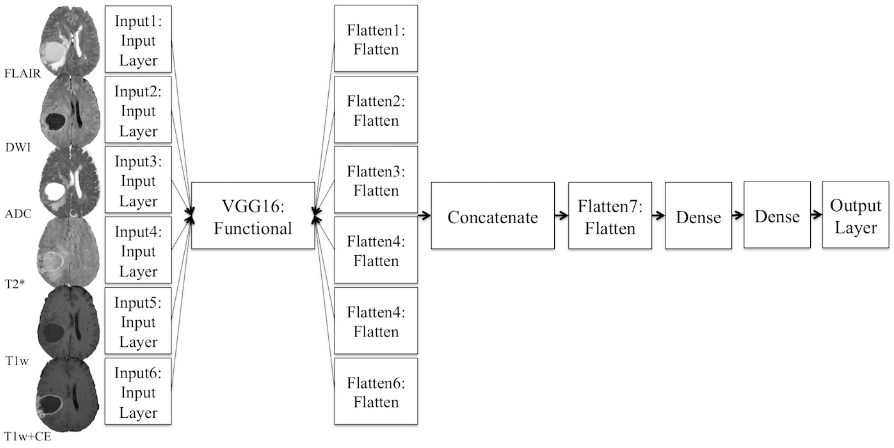


*Figure 8. Model design and proposed architecture for MRI_MULTI group.*

## 3.4 Model training, testing and validation

Once the architecture of the models was established, we proceeded to the learning phase. For neural networks, learning mainly consists of two different steps: feed-forward and error back-propagation. These steps have been repeated iteratively until a configuration of the weights able to offer satisfactory performances. Using the tools provided by the Keras framework, this process can be summarized in establishing a pipeline for entering the network and configuring the hyperparameters for training (the definition of which was made using the Keras Tuner library). In particular, it was necessary to define the optimization algorithm, the loss (i.e., the function that estimates the distance between expected and real values), the number of epochs (or iterations) of the training cycle on the entire training set, the batch size (i.e., the number of inputs processed in parallel by the calculator), and the additional metrics required to evaluate and compare the performance of the models.

For the correct validation of the proposed models, each dataset (MRI_VOL and MRI_MULTI) was initially divided into three subsets: a training set (80% of the sample) used for neural network learning and weights update, a test set (10% of the sample) for validating and evaluating the proposed model, and a validation set (10% of the sample) for testing the ability of the neural network in replicating/generalizing the performances on new data/images never encountered during training process. This partition was preliminarily performed by using different sources for training, validation and test sets in a manual step, in order to balance lesion types (GBM and SBM) across the three groups; this choice comes from the unbalanced distribution between privately stored images (including both GBMs and SBMs) and public repositories (including almost exclusively GBMs). Therefore, before proceeding to data augmentation, from the whole MRI data pool we casually extracted the desired number of MRI images for each subgroup, balancing the proportion of GBMs and SBMs according to lesions' epidemiology, preferring for the validation group external data from public repositories when available. Risk of oversampling has been averted by the decision taken at the moment of significant slices selection (lesions with maximum diameter <5cm were sampled once; lesions with maximum diameter >5cm were sampled twice at the level of the two most representative slices, with a minimum between-slices distance 3cm; see Materials and Methods section – Dataset Selection subsection). A schematic representation of MRI_VOL and MRI_MULTI samples' distribution among training, testing and validation group is shown in Table 3.

| | | Training set (80%) | Test set (10%) | Validation set (10%) | Total (100%) |
|---|---|---|---|---|---|
| MRI_VOL | GBM | 638 | 98 | 98 | 834 |
| | SBM | 474 | 52 | 52 | 578 |
| | ALL | 1112 | 150 | 150 | 1412 |
| MRI_MULTI | GBM | 130 | 18 | 18 | 166 |
| | SBM | 92 | 12 | 12 | 116 |
| | ALL | 222 | 30 | 30 | 282 |

*Table 3. Detailed distribution of selected GBM and SBM cases among training set (80% of the sample), test set (10% of the sample) and validation set (10% of the sample), both in MRI_VOL group and MRI_MULTI group.*

Training process was stopped when the loss on the validation set got larger than or was equal to the previous lowest loss for 10 times; the MRI_VOL model training took 5'5", whereas the MRI_MULTI model training took 2'20". Model training and evaluation in FuseMedML framework were carried on between May-July 2022 (referral period for FuseMedML library updating).
For each sub-dataset several metrics were extracted, and the following ones specifically analysed for the evaluation of the prediction model:

- Accuracy, defined as the percentage of correct predictions
- Precision, defined as the ratio between true positive and the sum of true positive and false positive instances
- Recall, defined as the ratio of positive instances correctly identified by the model
- F1, defined as the harmonic mean of precision and recall that assume high values only if accuracy and recovery are both high.

## 3.5 Comparison with human diagnostic performances

The same above-mentioned anonymized MRI images converted to JPEG format, distributed in the same three groups used for neural network model training and evaluation as previously described (see Table 3), have also been used for training and subsequent evaluation of human classification performances. For this purpose, three observers with different experience in the field of Neuroncology imaging were involved, in order to reproduce real-world clinical practice framework and expected variability among human observers. The characteristics of each observer in terms of experience in the field of oncological neuroimaging are briefly listed below:

- one neuroradiologist with 10 years of experience in neuroimaging and oncological neuroimaging (arbitrarily named observer #1);
- one general radiologist with 3 years of neuroimaging experience and with rudiments of conventional/advanced oncological neuroimaging (arbitrarily named observer #2);
- one general radiologist, with few/no experience in the field of oncological neuroimaging (arbitrarily named observer #3).

The three observers were asked to classify the MRI images in the two previously described categories, respectively GBM and SBM. Like what was done for the training of the neural network, all the observers independently gained experience on the training set (80% of the sample), freely disposing of MRI images for about one month; after this suitable period of time to elapse for formation, each rater was asked to perform a first phase of testing on the test set (10% of the sample), to refine the skills acquired on the training set and eventually deepen their knowledge on most complex radiological presentations. After two months, a final phase of evaluation of the blinded data was independently performed on the validation set (10% of the sample) by each observer under the supervision of the principal medical investigator.
The same evaluation was performed both for the MRI_VOL sample and MRI_MULTI sample at two different time-points; for each item of the validation sub-set and for both samples (MRI_VOL and MRI_MULTI), the three observers were also asked to express their diagnostic confidence level by using a 3-point rating scale:

- 1=low confidence level;
- 2=intermediate confidence level;
- 3=high confidence level.

These results were analysed in terms of sensitivity, specificity, diagnostic accuracy and AUC, and graphically represented with ROC curves. Agreement between observers and neural network results in each subgroup was assessed by using Kappa statistics, with the following result stratification according to the strength of agreement:

- 0.00-0.20 = poor agreement
- 0.21-0.40 = fair agreement
- 0.41-0.60 = moderate agreement
- 0.61-0.80 = substantial agreement
- 0.81-1.00 = almost perfect agreement

The significance of the difference between the areas that lie under the curves between observers and neural network results in each subgroup was tested by using non-parametric Mann-Whitney U test. For this statistical analysis, the additional statistical software XLSTAT available in Excel (Xlstat package, 2019.7) was used.

# Chapter 4

# Results

Regarding FuseMedML-based predictive model, we evaluated separately the two model designs and proposed architectures; as part of our evaluation, we report AUC with a 95% confidence interval, as well as diagnostic accuracy, sensitivity, specificity, precision and F1 score. The MRI_VOL model performed better than the MRI_MULTI model in all metrics, with results obtained on volumetric post-contrast images acquired for neuronavigational purposes more robust by virtue of a larger sample size (N=1412, see Table 3). However, it should be emphasized that the robustness of the results from the analysis of multimodal imaging, somehow influenced by the smaller sample size available (N=282 samples, see Table 3), is balanced by the amount of information contained within the sample itself (higher informative value of multiparametric MRI acquisition technique compared to single sequence 3D T1w acquired for surgical neuronavigation purposes). Results concerning prediction model of the trained neural network on MRI_VOL and MRI_MULTI images within the validation sub-group are summarized in Table 4.

| | **ACCURACY** | **PRECISION** | **RECALL** | **F1 SCORE** |
|---|---|---|---|---|
| **MRI_VOL** | 0.96 | 0.96 | 0.97 | 0.97 |
| **MRI_MULTI** | 0.94 | 0.96 | 0.92 | 0.94 |

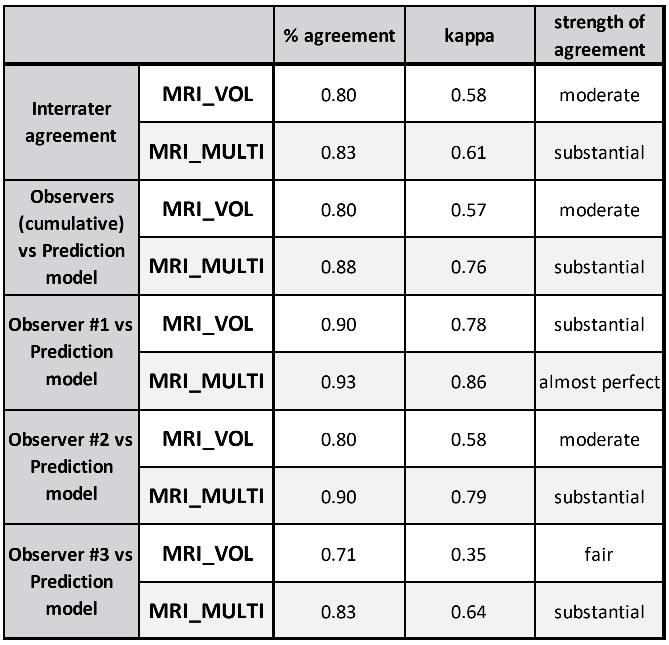
*Table 4: Schematic representation of final neural network results both in MRI_VOL and MRI_MULTI subgroups, expressed in terms of accuracy, precision, recall and F1 score.*

Concerning the comparison with human diagnostic performances, we analysed both individual data from each reader assessment and cumulative data obtained by merging the three readers classification performances. In this ancillary analysis, we report AUC with a 95% confidence interval, diagnostic accuracy, sensitivity and specificity in each sub-group (MRI_VOL and MRI_MULTI), compared with analogous metrics resulting from neural network predictive model. As expected, it should be noted that the three independent observers' assessment was considerably affected by readers expertise, with higher performances obtained for the more experienced observer #1 and lower performances for the less experienced observer #3. Results concerning human performances on MRI_VOL and MRI_MULTI images within the validation sub-group are summarized in Table 5.

| | | Observer #1 | Observer #2 | Observer #3 | Observers (cumulative) | Prediction model |
|---|---|---|---|---|---|---|
| Accuracy | MRI_VOL | 0.87 | 0.77 | 0.67 | 0.77 | 0.96 |
| | MRI_MULTI | 0.93 | 0.83 | 0.77 | 0.84 | 0.94 |
| Sensitivity | MRI_VOL | 0.87 | 0.78 | 0.69 | 0.78 | 0.97 |
| | MRI_MULTI | 1 | 0.89 | 0.89 | 0.92 | 0.92 |
| Specificity | MRI_VOL | 0.86 | 0.75 | 0.61 | 0.74 | 0.94 |
| | MRI_MULTI | 0.83 | 0.75 | 0.58 | 0.72 | 0.92 |
| ROC curve AUC | MRI_VOL | 0.94 | 0.89 | 0.84 | 0.89 | 0.99 |
| | MRI_MULTI | 0.98 | 0.90 | 0.89 | 0.94 | 0.99 |

*Table 5: Schematic representation of human readers results (individual and cumulative) both in MRI_VOL and MRI_MULTI subgroups, expressed in terms of accuracy, sensitivity, specificity and area under the curve (compared to neural network results - last column).*

Measurement of the extent to which the three observers independently assigned the same diagnosis to the same MRI image/set of MRI images in the two groups (MRI_VOL and MRI_MULTI, respectively) was performed by using Kappa statistics; interrater agreement was also measured between observers and FuseMedML-based prediction model. Results concerning interrater reliability are summarized hereafter in Table 6. When comparing agreement between observers and FuseMedML-based prediction model, what is immediately clear is that the strength of agreement is remarkably affected by readers expertise, with very good results when comparing the performances of more experienced observers (i.e., observer #1) and lower percentage of agreement for less experienced ones (i.e., observer #3). The same trend applies both to MRI_VOL sub-group and MRI_MULTI sub-group, although with higher percentage of agreement in this latter probably due to the lower sample size and the higher informative value of multiparametric MRI acquisition technique compared to post-contrast neuronavigation MRI images alone.



| | | % agreement | kappa | strength of agreement |
|---|---|---|---|---|
| Interrater agreement | MRI_VOL | 0.80 | 0.58 | moderate |
| | MRI_MULTI | 0.83 | 0.61 | substantial |
| Observers (cumulative) vs Prediction model | MRI_VOL | 0.80 | 0.57 | moderate |
| | MRI_MULTI | 0.88 | 0.76 | substantial |
| Observer #1 vs Prediction model | MRI_VOL | 0.90 | 0.78 | substantial |
| | MRI_MULTI | 0.93 | 0.86 | almost perfect |
| Observer #2 vs Prediction model | MRI_VOL | 0.80 | 0.58 | moderate |
| | MRI_MULTI | 0.90 | 0.79 | substantial |
| Observer #3 vs Prediction model | MRI_VOL | 0.71 | 0.35 | fair |
| | MRI_MULTI | 0.83 | 0.64 | substantial |

*Table 6: Interrater reliability both for MRI_VOL and MRI_MULTI subgroups.*

Variation between AUCs measuring the accuracy of the two classification systems (human observers versus FuseMedML-based prediction model) was reported in terms of area difference (area$_{\text{prediction model}}$ – area$_{\text{observer(s)}}$), standard error of the difference, *z*-values and *p*-values. Results concerning the difference between AUCs among observers and neural network performances in each subgroup are summarized in Table 7. Significance of the difference between AUCs showed meaningful results when comparing the performances obtained on the MRI_VOL validation group ($p$<0.01), with more striking differences and smaller *p*-values the lower the readers' experience is. Conversely, no significant results were obtained on the MRI_MULTI validation group ($p$>0.1) regardless of readers' experience, although with a trend showing lower differences for the more experienced observer (i.e., observer #1) compared to the less experienced one (i.e., observer #3).

| | | Observers (cumulative) vs Prediction model | Observer #1 vs Prediction model | Observer #2 vs Prediction model | Observer #3 vs Prediction model |
|---|---|---|---|---|---|
| difference area$_{(prediction\ model)}$ - area$_{(obs)}$ | MRI_VOL | 0.1 | 0.05 | 0.1 | 0.15 |
| | MRI_MULTI | 0.05 | 0.01 | 0.09 | 0.1 |
| standard error of the difference | MRI_VOL | 0.027 | 0.020 | 0.027 | 0.032 |
| | MRI_MULTI | 0.056 | 0.036 | 0.068 | 0.071 |
| Z | MRI_VOL | -3.7 | -2.5 | -3.7 | -4.6 |
| | MRI_MULTI | -0.9 | -0.2 | -1.3 | -1.4 |
| P | MRI_VOL | 0.0001 | 0.01 | 0.0001 | 0.000004 |
| | MRI_MULTI | 0.36 | 0.78 | 0.18 | 0.16 |

*Table 7: Significance of the difference between the areas under ROC Curves, both for MRI_VOL and MRI_MULTI subgroups.*

The probability ROC curves obtained for the MRI_VOL and MRI_MULTI cohorts respectively, graphically reporting the performance of the human readers (both individual and cumulative) and of the proposed binary classification model, are shown in Figure 9 (for MRI_VOL group) and Figure 10 (for MRI_MULTI group).

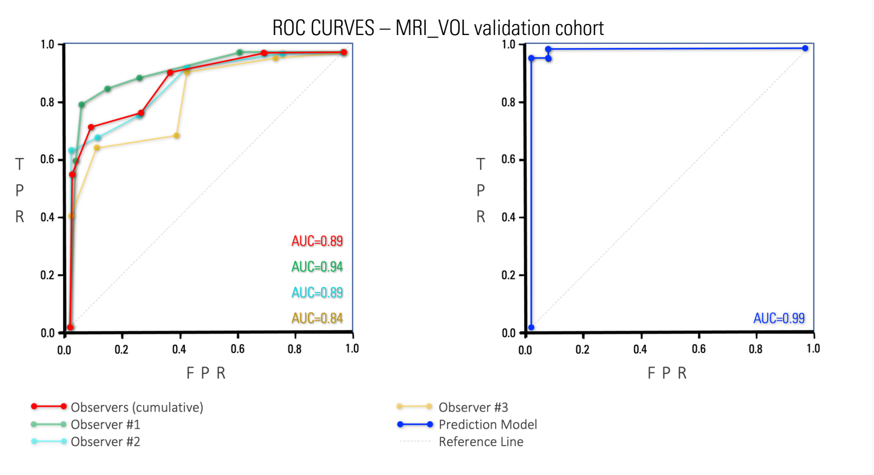


*Figure 9. Receiver-operating characteristic (ROC) curves and corresponding area under the curve (AUC) statistics for the classification performance in the validation cohort of MRI_VOL subset, both in human observers (on the left: observers, cumulative – red line; observer #1, high experience – green line; observer #2, intermediate experience – light blue line; observer #3, low experience – yellow line) and FudeMedML (on the right; dark blue line). The X-axis represents the false positive rate and the Y-axis the true positive rate.*

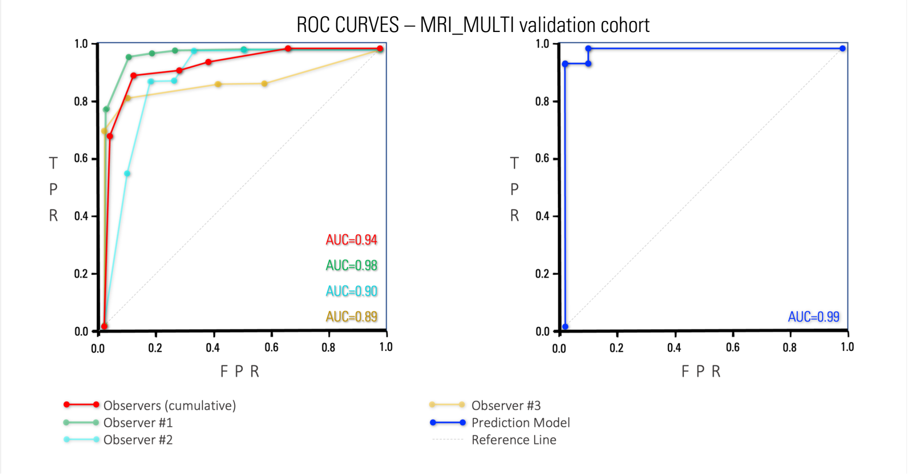


*Figure 10. Receiver-operating characteristic (ROC) curves and corresponding area under the curve (AUC) statistics for the classification performance in the validation cohort of MRI_MULTI subset, both in human observers (on the left: observers, cumulative – red line; observer #1, high experience – green line; observer #2, intermediate experience – light blue line; observer #3, low experience – yellow line) and FudeMedML (on the right; dark blue line). The X-axis represents the false positive rate and the Y-axis the true positive rate.*

# Chapter 5

# Discussion

According to the World Health Organization's most recent 2021 classification [14], brain tumours are one of the most common causes of cancer-related morbidity and mortality worldwide. MRI is the diagnostic tool most extensively used for the initial differential diagnosis of tumour type; based on MRI imaging, tumour location and identification typically depends on the radiologist's expertise. Final diagnosis is confirmed on histopathological analysis; however, it must be noticed that biopsy, that is carried out to determine the tissue's benignity or malignancy as well as the hystotype, is a highly invasive procedure which is not performed prior to end-of-the-brain surgery (in contrast to tumours discovered elsewhere in the body). In this light, it is critical to develop a viable alternative diagnostic tool for tumour classification and segmentation from MRI images in order to obtain accurate diagnosis and minimize the resort to invasive medical procedures prone to procedural difficulties and interpretative subjectivity [32], [33]. The ability for physicians to quickly and accurately classify CNS tumours based on brain images is gathering an over increasing importance [32], [33]. Traditional diagnostic approaches are unable to effectively manage the substantial growth in data volume in the medical sector, with the problem of storage and interpretation of big medical data still representing a challenge in the field of medical image analysis. To the purpose the emergence of new technologies, particularly machine learning and CNN, had probably the most significant impact on the medical sector since their introduction, especially in the field of medical imaging and even more of neuroimaging. At present, to implement interpretation of MRI images and to support radiologist's choice, a variety of machine learning algorithms have been used both for segmentation and classification purposes. However, these techniques require specialized knowledge on how to extract the best features and still need more extensive experimentations for ensuring results generalization and reproducibility.

In this light, deep learning-based models with semi-supervised and unsupervised approaches have attracted researchers' attention due to both their high performance and the chance to automatically generate features extraction. The described research activity is part of this over-increasing number of studies that have tested the use of AI methods for the identification or classification of brain tumours [34]–[37], mainly focusing in the proposed setting on a binary classification of the two most common brain neoplasms of the adulthood, which are actually similar with each other at imaging, and which raise the bar in terms of differential diagnosis difficulty. Indeed, the issue of differential diagnosis between these two selected entities based on MRI has been faced by resorting to both conventional and advanced MRI sequences [38]–[42], as well as to the analysis of quantitative information about internal tumours' structure by mean of radiomics [37], [43]–[45] and texture analysis tools [28], [46], [47]; these studies, despite confirming the existence of qualitative features and quantitative/semi-quantitative findings to distinguish between GBM and SBM, were carried on very small populations and were significantly affected by readers interpretation (also in terms of manual steps for perfusion/spectroscopy acquisition/processing) or by individual variability in images segmentation (for radiomics/texture extrapolation).

In recent times the application of machine learning pattern recognition techniques has been proposed as a possible tool to overcome some of these barriers, and has been tested on neurooncological imaging with promising results in the differentiation of common intra-axial brain tumours (also encompassing GBM and SBM) from MRI images obtained for advanced diagnostic purposes (i.e., perfusion or spectroscopy). First experimental studies proved machine learning to have a substantial incremental diagnostic value for brain tumour differentiation, provided that the MRI acquisition technique is highly standardized and dedicated to the purpose [48]. Similarly, features derived from the peri-enhancing oedema region surrounding the intra-axial lesion also showed moderate value in differentiating supratentorial SBMs from GBMs; the peri-enhancing oedema's distinct tumour signatures, identified using deep learning-based algorithms, have been proved to distinguish the peritumoral microenvironment of GBM from SBM with an accuracy ranging from 0.56 to 0.85 depending on the case study [49], [50], but also in these cases results are always obtained by applying CNN on advanced and highly standardized MRI sequences obtained for research purposes and by performing time consuming semi-automated segmentation. In 2020, Amin et al. [51] proposed a model based on the fusion of different MRI sequences (T1w+CE, T1w, FLAIR and T2w) using CNN to differentiate tumour from non-tumour areas in glioma patients and applying this model on publicly available datasets; the results showed that fused MRI images provided better results than single sequences, thus supporting the higher informative value of multiparametric MRI compared to individual MRI acquisitions. However, the proposed methodology was limited to the distinction between neoplastic vs non-neoplastic tissue, with no further classificative ambition. Afterwards, several papers explored the potential for machine learning and CNN to enhance the distinction among different tumour types, however always with results affected by sample selection and size [52]–[54]. Similarly, Badža et al. moved a little bit forward by proposing a CNN architecture with 10-fold cross-validation methods for brain tumour classification using different and dissimilar neoplasm, but with the main strength of amplifying the role of a large and representative sample size; the study was based on a very large amount of T1w post-contrast imaging and reached an overall accuracy of 0.96 [55], thus pointing out the pivotal importance of large datasets for machine learning performance tuning. Similar approaches have been used for different classification purposes such as distinction into low- vs high-grade glioma [56], IDH mutated vs IDH non-mutated glioblastoma [57], intra- vs extra-axial neoplasm [33] or benign vs malignant brain tumor [58]; in particular in this latter study, the research project was carried out on more than 7000 images from different sources, and deep learning feature extraction by mean of transfer learning was performed using the pre-trained CNN model VGG16 on public data (as per our study), reaching an overall accuracy of 0.97 in determining the presence and the extent of the lesion.

In our study, we moved forward compared to this previous experience obtained by using the VGG16 pre-trained CNN model for brain tumour classification; here we proposed a binary classification by introducing the two most common brain tumour types of the adulthood; we operated on a large dataset (also considering the lower sample heterogeneity due to the inclusion only of GBM and SBM), obtained by merging private repositories (mainly used for training and testing) with public repositories (mainly used for validation purposes). We also proposed two different architectures, one for the analysis limited to volumetric post-contrast imaging and one for multimodal MRI acquisitions, in order to test the weight of more informative multiparametric MRI compared to individual MRI acquisitions as well as their ability in mitigating numerical difference in terms of sample size for final classification purposes. The observed overall accuracy in the two cases (respectively 0.96 for MRI_VOL subset and 0.94 from MRI_MULTI subset) complies with what expected and at least in part confirms previous literature evidences. Analysing the results in detail, the performances obtained on volumetric post-contrast images acquired for neuronavigational purposes are particularly robust by virtue of the large sample size and the representative number of input data; however, despite a significantly smaller training data set, results derived from multiparametric MRI are more than satisfactory and perfectly in line with the ones observed for neuronavigational volumetric post-contrast MRI. It should therefore be emphasized that the robustness of the results from the analysis of multimodal imaging, somehow influenced by the lower sample size available, is balanced by the amount of information contained within the sample itself (since the neural network takes advantage of the higher informative value of multimodal MRI acquisition technique). To the best of our knowledge, an approach similar to the one described in our paper has previously only been adopted by Shin et al. in 2021 by using a 2D convolutional neural network (namely the ResNet-50 model) on internal and external test sets, with the aim to distinguish GBM from SBM and with the secondary goal to compare its results to two expert neuroradiologists' performances [59], although limiting the analysis to pre-operative contrast-enhanced T1W/T2W images; compared to the results of the VGG16 architecture and the proposed FuseMedML-based model (that showed excellent generalization capabilities on unseen testing data), lower diagnostic performances were observed, probably due to the smaller sample size and to the lower number of collected MRI sequences. However, to be confirmed, these results would benefit of a larger dataset and of a direct comparison with the most widely used CNN architectures; such comparison of different architectures on the same data would be worthy of attention to determine reproducibility and reliability of these networks when applied to medical imaging. Neither must be forgotten, when interpreting CNN performances, that real-world problems are often more complex than the one simulated in experimental designs and involve massive amounts of data other than imaging ones. Consequently, it can be assumed that a single machine learning tool cannot address all diagnostic dilemmas, whereas is more likely that a group of tools appropriately integrated with one another can provide better prospective solutions [60].

Another consideration applying to the proposed research activity is that we analysed imaging data through the open source FuseMedML, that has never been used in neuro-oncological imaging since our very first experience [29], but that has been used for similar clinical purposes in few preliminary studies on breast tumours, liver and kidney masses [9], [15]–[20], [61]. FuseMedML was designed to simplify and streamline medical research projects; it is a Python framework designed to accelerate Machine Learning based discovery in the medical domain. Flexible and designed for easy collaboration, it encourages code reuse, and allows to efficiently process and fuse information from multiple modalities (for example, different MRI modalities, but also different imaging techniques or imaging techniques with biochemistry, clinical data, and so on) [9]. Therefore, in the light of recent literature evidence, the described models could be a promising innovation and an effective support for the differential diagnosis of single brain neoplasms, especially in case of classification uncertainty on conventional MR imaging.

Of note, the proposed approach also harbours the potential to provide the basis for including in the predictive model data other than imaging, once available. However, it should always be considered that FuseMedML is still an evolving tool (given the rapid transformations and updates provided for this open-source library) and some content, steps, functions and commands are changing over time; this is both a challenge and a strength of FuseMedML, as this flexible framework is continuously adapting to researchers' and AI specialists' needs and requirements, in order to guarantee the best service possible and increase the performance reliability over time.

As a secondary endpoint we compared the neural network performances in classifying monofocal GBM vs SBM brain lesions to human-level image interpretation, showing how under controlled conditions (such as those created for our research purpose) FuseMedML-based predictive model may equal and somehow even exceed radiologists' performances. Indeed, oncology healthcare providers first and foremost rely on accurate imaging interpretation for shared decision-making, and sometimes critical imaging interpretation can make the difference between invasive, minimally invasive or non-invasive medical approaches. This imaging interpretation is largely entrusted to radiologists. However it is well known that, depending on readers' experience and case-specific difficulty coefficient, a certain variability among readers may be observed; attended discrepancy could be further amplified by image acquisition differences, image quality degradation (i.e., motion, device-related artifacts, etc.), reader fatigue or inconsistent reporting (i.e., not modulated on the specific diagnostic issue) [62], [63]. Such assumptions are the bases for the need to further optimize image interpretation for cancer detection; potentially attesting machine learning methods superiority compared to human readers' performances would indeed open the way for further systematic studies to improve clinical decision-making and implement computer vision image analysis in clinical practice, not only in Neuroncology but also in other medical fields.

At today state-of-the-art, this comparison between machine learning versus human reader performances still represents a very little explored sector, with few evidence collected in different fields and with a relatively low number of published papers (with sometimes controversial results) [21], [64]; this lack of information in this field is also due to the absence/difficulty in finding satisfactory or reliable historical key performance indicators to use for direct comparison, largely not available at present. Some preliminary experiences have been newly collected regarding electrocardiograms [2] and echocardiography [65] interpretation, whose automated reading by mean of machine learning-based algorithms demonstrated stronger correlation with different biomarkers of acute disease compared to manual reads. For neurological applications, machine learning and deep learning have mainly been compared to human experts' interpretation in the domain of electroencephalography analysis, where automated technologies seemed to allow for better generalization capabilities and more flexible applications, reaching competitive performance on selected target tasks [66]–[68]; however, more realistic approaches concerning the same issue (on larger amounts of data and with fewer restrictions on data input) suggested that hybrid approaches, where human raters applied international consensus diagnostic criteria to automated detections of AI-based algorithms, was far more accurate and suitable for clinical implementation compared to AI alone [69]. Concerning radiological imaging, analogous testimonies have been collected in emergency radiology, breast imaging and nuclear medicine [70]–[72], where machine learning methods performances were proved comparable to the one obtained by radiologists, with relatively small datasets used for machine learning training. Also in these cases, such reports represented isolated non-systematic attempts of comparing the two scenarios (human vs machine) and are probably strongly influenced by low samples sizes used for training, testing and validating purposes. To date, more systematic and comprehensive data to compare the accuracy of human readers versus machine-learning algorithms have only been collected for skin lesions classification based on dermatoscopic images; coherently to what expected in the light of the above-mentioned literature elements, dermatology experts were mostly outperformed by machine-learning algorithms in the assessment of benign vs malignant pigmented skin lesions, confirming how human readers would benefit from automated image classification assistance [73]–[76].

Our findings on neurooncological imaging tie in with the corpus of research projects aiming to document interpretation consistency reached by AI tools and provide head-to-head comparisons between AI vs conventional human image interpretation. To the best of our knowledge, only few direct comparisons between machine learning approaches and diagnostic human performances have been applied to MRI imaging in Neuroncology; to date, despite a cautious optimism for machine learning-enhanced image interpretation, human visual classification performances still seem to be more robust in complex situations, such as image manipulation, weaker MRI signal and contrast reduction, or additive noise and image distortions/artifacts [77]. In 2019 Molina-García et al. compared machine-learning predictive models for GBM prognosis based on clinical information and most meaningful morphological MRI data (according to literature: age, enhancing lesion volume, enhancement rim width and surface regularity) to prognostic models based on human performances, documenting similar discriminatory capability

[22]; however, tested MRI images belonged to the same diagnostic category (histologically proved GBM), human interpretation was also based on simple (but not exhaustive) clinical data, prognostic implications were inevitably influenced by external factors (i.e., extent of resection, treatment choice, treatment discontinuation due to patients' related factors, etc.) that are not fully explored in the paper, and preliminary semi-automatic MRI image segmentation was required (manually performed, then reviewed by experienced radiologists). Apart from the different diagnostic purpose (differential diagnosis instead of prognostic stratification), in our experimental setting we aimed to overcome some of these possible sources of bias that affected previous comparisons: we have voluntarily waived any clinical data on patients (as not exhaustive, especially for public repositories' ones) and limited the analysis on first-diagnosed patients, and we have opted for machine learning approaches not requiring complex and time-consuming pre-processing (such as the semi-automated segmentation described by Molina-Garcia et al.). Moreover, we have not limited the morphological MRI data that can be selected by the tested machine learning prediction model (as key biomarkers for the specific purpose have not been formally defined); although FuseMedML-based approach may be seen as a black box model, we tried to limit potential spurious associations by resorting to independent datasets for validation and by comparing the results to human-level performances.

Because of the lack of gold-standard benchmarks, human performances have been newly collected by resorting to three different raters with variable experience in neuroimaging and oncological neuroimaging. This choice is motivated by the will to reproduce real-world clinical practice framework and expected variability among human observers, and by the attempt to compare FuseMedML-based prediction model to human readers with different experience in neurooncological imaging. In our experience, human diagnostic accuracy was considerably affected by readers expertise, with higher performances obtained for the most experienced observer and lower performances for the less experienced one. Similarly, when moving to compare the agreement between observers with FuseMedML-based prediction model, the strength of agreement is remarkably affected by readers expertise; indeed cumulative observers agreement with prediction model is moderate, but when we individually analyse the results from each rather it becomes evident a substantial difference among observers depending on their specific background (with the highest agreement for the most experienced observer and the lowest agreement for less experienced one). The same trend in diagnostic performances and interrater agreement applies both to MRI_VOL sub-group and MRI_MULTI sub-group, although with less significant differences in the latter group regardless of human readers experience; this is probably attributable to the higher informative value of multiparametric MRI acquisition technique compared to post-contrast neuronavigation MRI images alone, however this point requires further in-depth exploration on larger data sets to be elucidated.

Despite machine learning models should be conceptually based on datasets as wide as possible (ideally, on millions different images), this study confirm that robust performances may be achieved with a relatively limited but heterogeneous dataset (i.e., few hundred images), provided that image distribution among diagnostic classes is balanced and not redundant [71]. Nevertheless, it must not be forgotten that (as in the proposed research activity) machine learning models are usually tested in artificial preselected settings (in this specific case, concerning binary differential diagnosis between two given tumour types), thus not reflecting the complexity of real diagnostic workflows and not taking into account any further patient's information. Automatic classification strategies based on AI or CNN should more faithfully replicate the high complex human decision processes, rather than solely and simplistically provide binary decisions between few options (i.e., benignant vs malignant; high-grade vs low-grade; good prognosis vs poor prognosis; and so on), which is futile in most circumstances. Moreover, real-life radiologists are aware of the whole clinical picture, of patient's personal or familiar anamnesis, of laboratory findings and of previous imaging data when available; therefore, their final diagnosis depends on much more than imaging itself. Therefore, to date such information multiplicity and clinical complexity are not reflected in this study setting, and imperatively require to be explored before integrating deep learning into the existing clinical workflows. FuseMedML framework gathers its momentum and has the perspective to combine such information, however its exploitation is still at an embryonal phase and the knowledge on its potentialities is yet incomplete. Therefore at present, although the promising results obtained with AI, its routine application for oncology assessment within cancer trials is still an aspiration and needs to be more extensively explored before considering hybrid approaches to be implemented into real-life clinical practice; final decisions on image interpretation should still be left to experienced radiologists, aware of the complete and complex patient's information [7], [64], [78].

# Chapter 6

# Limitations and strengths

Possible limitations to this study include:
(1) the relative lack of direct comparison with previous FuseMedML applications, due to its relative novelty and to the limited corpus of published papers on this library (especially when considering the neuroradiological field, as stated before);
(2) the absence of direct comparative study on the benefits of adopting FuseMedML instead of other mainstream frameworks (representing the basis for reflection on future research solutions);
(3) the still relatively limited sample size, mainly for multimodal analysis (which more closely represents the actual clinical practice, and is expected to become more performing than the volumetric one as the sample size grows);
(4) the choice of not limiting data inclusion to standardized MRI protocols for scan acquisition, with consequent high variability of image resolution, voxel size and image contrast dynamics, resulting in an extensive variety within the datasets (which could be viewed both as a limitation and a strength at the same time, see below);
(5) the lack of non-radiological information on tumour's characteristics (laboratory examinations, pathological features or molecular-genetical data), thus limiting the possibility to integrate different structured and unstructured information to improve the quality of the classifier;
(6) the choice to focus the analysis only on the two most common monofocal brain tumours of the adulthood, GBM and SBM respectively (while other more uncommon histopathological diagnoses, as well as possible brain tumour mimics are not included in this preliminary model). This latter procedural decision, also related to contingent data availability, could be viewed both as a limitation and a strength at the same time; indeed limiting the research sample to these two tumour types allows to be more rigorous in the MRI data selection, only including brain lesions that are actually similar with each other in terms of localization (intra-axial only), age of onset (adulthood: >25y and <75y), clinical presentation (neurological signs and symptoms), MRI acquisition protocol (similar MRI sequences available), MRI signal changes (infiltrative behaviour, enhancement, perilesional oedema, possible presence of necrosis/haemorrhage, etc), number of lesions

(monofocal only), and so on; other papers in scientific literature generally included a larger amount of different lesion types that usually have more clues guiding the neuroradiologist (and thus also the machine learning model) to the correct diagnosis. In our case, we aimed to focus on a relevant differential diagnostic issue in daily clinical practice, thus pointing out possible future implications in difficult and uncertain classificative grey areas at MRI imaging; therefore, we aimed to reproduce a setting in which neuroradiologists could took interest in receiving an external support to diagnosis, with neither detracting from the importance of overall clinical information on the patients nor diminishing the real-life experience factor;

(7) the choice to train human observers before the validation phase, as to reproduce the learning capabilities of humans' vs artificial agents. During the experiment, training and testing images are submitted to human experts in advance to prepare the validation phase. It would also be interesting to compare the performance of naive human observers with the performance of the same operator after specific training with pre-labelled images, to further clarify the impact of the ground truth.

Alongside possible limitations, major strengths of this research activity include:

(1) the representative number of collected evidence compared to the vast majority of previously published papers;

(2) the choice to select only newly diagnosed single brain lesions, limiting the effects of possible confounders on the classification model (such as chemo/radiotherapy-related brain changes, post-surgical changes, relapsing/residual diseases differences, etc.);

(3) the inclusion of somehow heterogeneous MRI images (acquired on different scanners by different vendors, with different coils and sequences, sometimes vitiated by minor artifacts, and so on), which represents an important novelty element that reflects the everyday reality and the actual variability in diagnostic imaging, thus reproducing a real-life work setting (and not only an "ideal" experimental setting in which MRI images that do not strictly respond to the experimental inclusion/criteria are generally dropped out);

(4) the choice to limit the analysis to the most significant slice/the two most significant slices per patient and to use JPEG rather than most complex file formats (thus preferring raster image file formats with lossy compression, more suitable for data storage or sharing), that probably is a core difference from previous machine learning/radiomics studies based on standardized MRI acquisitions performed on patients with brain neoplasm also encompassing GBM and metastases [43], [44], [53];

(5) the availability of a representative and largely independent external validation group, to further increase the confidence in generalisation across different sources;
(6) the comparison with human performances, to reinforce knowledge on possible real-life advantages of these network models' implementation in daily clinical practice.

# Chapter 7

# Conclusions

In conclusion, this type of semi-supervised FuseMedML-based approach has the potential to gain physicians' and researchers' interest thanks to the relatively low computational costs and the efficiency in extracting/inferring imaging-related features. The experimental results demonstrate that brain tumour classification based on conventional MRI by using the proposed model could achieve high diagnostic accuracy and low error rate in classifying tumour types, consistently with the final histopathological diagnoses.

Moreover, in daily practice oncologists and surgeons largely rely on the interpretation provided by neuroradiologists when making clinical decisions, and variability among readers in imaging interpretation (inherently subjective) represents a potentially significant source of bias. The two most important sources of variability, namely case difficulty and single reader's skills, are accompanied by other minor causes of discrepancy in image assessment (different image acquisition techniques, specific training in a given sub-specialty, and so on); proposed experimental results from ancillary analysis on human diagnostic performances highlight a possible role for AI technologies as emerging solution to this problem, providing higher standardization level, more consistent or predictable behaviour in image allocation, and lower susceptibility to human biasing factors.

It should be also noted that these classification performances have been obtained adopting MRI images acquired for daily clinical purposes in different centres and with different MRI tools (a major strength of this work, thus not strongly influenced by the specific clinical setting and hypothetically reproducible on a wider scale). All these elements taken together map out an interesting route to overcome diagnostic challenges in imaging interpretation and human fatigue barriers (in our specific case referred to Neuroncology, but potentially extendable to other imaging fields), not only without compromising but also improving the quality of care for patients.

Future goal of the research activity is to replicate and upgrade the proposed models on a larger population by acquiring and storing new MRI studies, in order to re-test their goodness and robustness; moreover, further validation using larger public and free repository datasets is still ongoing and will provide new external evidence concerning the reproducibility of the described results. An additional goal is to develop a more sophisticated and comprehensive models for providing classifications other than binary (i.e., including different tumour types other than GBM versus SBM, not presently considered in the proposed prototype); it is also possible to assume that these models could also be used for other-than-oncological classes of pathology, as well as for other applications within the field of Neuroncology (i.e., genetic-molecular analyses, nuclear imaging, histological preparations, and so on).

As a final remark, considering the promising performances achieved by the algorithm and the ensemble of scientific evidences on similar topics, this embryonal research activity may represent a foundation stone to in-depth exploring the potential of FuseMedML to start a line of research both for academic purposes, as well as for real-world clinical practice through possible industrial implications for the development of automated diagnostic-support systems based on deep learning.

# Acknowledgements

Thanks to Prof. P. Maresca and Prof. A. Marinelli for co-tutoring the planning and the developmental phases of this research project.

Thanks to Prof. A. Brunetti and Prof. E. M. Covelli for the precious contribution to technical aspects, especially concerning data collection/transmission.

Thanks to D. Benfenati and V. Carandente for the help in handling the technical aspects of this research work.

Thanks to Dr. F. Giordano, Dr. R. Ronza and Dr. A. Tucci for the commitment in the traditional classification operational aspects, representing a core part of this research work.

## Selected bibliography

[1] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, “A targeted real-time early warning score (TREWScore) for septic shock,” *Sci Transl Med*, vol. 7, no. 299, 2015, [Online]. Available: www.ScienceTranslationalMedicine.org

[2] A. Y. Hannun *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nat Med*, vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/s41591-018-0268-3.

[3] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA - Journal of the American Medical Association*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/jama.2016.17216.

[4] K. W. Noh, R. Buettner, and S. Klein, “Shifting gears in precision oncology—challenges and opportunities of integrative data analysis,” *Biomolecules*, vol. 11, no. 9. MDPI, Sep. 01, 2021. doi: 10.3390/biom11091310.

[5] H. Nandu, P. Y. Wen, and R. Y. Huang, “Imaging in neuro-oncology,” *Therapeutic Advances in Neurological Disorders*, vol. 11. SAGE Publications Ltd, Jan. 01, 2018. doi: 10.1177/1756286418759865.

[6] F. Barkhof, H. R. Jäger, M. M. Thurnher, and À. Rovira, *Clinical Neuroradiology - The ESNR Textbook*. Springer Cham, 2019. doi: 10.1007/978-3-319-68536-6.

[7] T. Davenport and R. Kalakota, “DIGITAL TECHNOLOGY The potential for artificial intelligence in healthcare,” 2019.

[8] “IBM Research, Haifa: FuseMedML. https://github.com/IBM/fuse-med-ml(2021).”

[9] A. Golts, M. Raboh, Y. Shoshan, S. Polaczek, S. Rabinovici-Cohen, and E. Hexter, “FuseMedML: a framework for accelerated discovery in machine learning based biomedicine,” *J Open Source Softw*, vol. 8, no. 81, p. 4943, Jan. 2023, doi: 10.21105/joss.04943.

[10] J. Faehndrich, S. Weidauer, U. Pilatus, A. Oszvald, F. E. Zanella, and E. Hattingen, “Neuroradiological viewpoint on the

diagnostics of space-occupying brain lesions," *Clinical Neuroradiology*, vol. 21, no. 3. pp. 123–139, Sep. 2011. doi: 10.1007/s00062-011-0073-6.

[11] D. She, Z. Xing, and D. Cao, "Differentiation of Glioblastoma and Solitary Brain Metastasis by Gradient of Relative Cerebral Blood Volume in the Peritumoral Brain Zone Derived from Dynamic Susceptibility Contrast Perfusion Magnetic Resonance Imaging," *J Comput Assist Tomogr*, vol. 43, no. 1, pp. 13–17, Jan. 2019, doi: 10.1097/RCT.0000000000000771.

[12] C. H. Suh, H. S. Kim, S. C. Jung, C. G. Choi, and S. J. Kim, "Perfusion MRI as a diagnostic biomarker for differentiating glioma from brain metastasis: a systematic review and meta-analysis," *European Radiology*, vol. 28, no. 9. Springer Verlag, pp. 3819–3831, Sep. 01, 2018. doi: 10.1007/s00330-018-5335-0.

[13] N. Anzalone *et al.*, "Brain Gliomas: Multicenter Standardized Assessment of Dynamic Contrast-enhanced and Dynamic Susceptibility Contrast MR Images," *Radiology*, vol. 287, no. 3, pp. 933–943, Jun. 2018, doi: 10.1148/radiol.2017170362.

[14] D. N. Louis *et al.*, "The 2021 WHO classification of tumors of the central nervous system: A summary," *Neuro Oncol*, vol. 23, no. 8, pp. 1231–1251, Aug. 2021, doi: 10.1093/neuonc/noab106.

[15] V. Barros *et al.*, "Virtual Biopsy by Using Artificial Intelligence-based Multimodal Modeling of Binational Mammography Data," *Radiology*, vol. 306, no. 3, p. e220027, Mar. 2023, doi: 10.1148/radiol.220027.

[16] Tal Tlusty *et al.*, "Pre-biopsy Multi-class Classification of Breast Lesion Pathology in Mammograms," in *12th International Workshop, MLMI 2021 Held in Conjunction with MICCAI 2021*, 2021, pp. 277–286. doi: 10.1007/978-3-030-87589-3.

[17] M. Raboh, D. Levanony, P. Dufort, and A. Sitek, "Context in medical imaging: the case of focal liver lesion classification," in *Proc.SPIE*, Apr. 2022, p. 120320O. doi: 10.1117/12.2609385.

[18] Ibrahim Jubran, Moshiko Raboh, Shaked Perek, David Gruen, and Efrat Hexter, "A Glimpse into the Future: Disease Progression Simulation for Breast Cancer in Mammograms," in *6th International Workshop, SASHIMI 2021 Held in Conjunction*

*with MICCAI 2021*, 2021, pp. 34–43. doi: 10.1007/978-3-030-87592-3.

[19] S. Rabinovici-Cohen, T. Tlusty, X. M. Fernández, and B. G. Rejo, “Early prediction of metastasis in women with locally advanced breast cancer,” in *Proc.SPIE*, Apr. 2022, p. 120330F. doi: 10.1117/12.2613169.

[20] S. Rabinovici-Cohen *et al.*, “Multimodal Prediction of Five-Year Breast Cancer Recurrence in Women Who Receive Neoadjuvant Chemotherapy,” *Cancers (Basel)*, vol. 14, no. 16, Aug. 2022, doi: 10.3390/cancers14163848.

[21] G. E. Cacciamani *et al.*, “Is Artificial Intelligence Replacing Our Radiology Stars? Not Yet!,” *European Urology Open Science*, vol. 48. Elsevier B.V., pp. 14–16, Feb. 01, 2023. doi: 10.1016/j.euros.2022.09.024.

[22] D. Molina-García, L. Vera-Ramírez, J. Pérez-Beteta, E. Arana, and V. M. Pérez-García, “Prognostic models based on imaging findings in glioblastoma: Human versus Machine,” *Sci Rep*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-42326-3.

[23] S. H. Yu *et al.*, “Early experience with Watson for Oncology: a clinical decision-support system for prostate cancer treatment recommendations,” *World J Urol*, vol. 39, no. 2, pp. 407–413, Feb. 2021, doi: 10.1007/s00345-020-03214-y.

[24] Casey Ross and Ike Swetlitz, “IBM_pitched_Watson_as_a_revolution_in_cancer_care,” *STAT*, 2017.

[25] B. H. Menze *et al.*, “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE Trans Med Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[26] K. Clark *et al.*, “The cancer imaging archive (TCIA): Maintaining and operating a public information repository,” *J Digit Imaging*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.

[27] J. Kang, Z. Ullah, and J. Gwak, “Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers,” *Sensors*, vol. 21, no. 6, pp. 1–21, Mar. 2021, doi: 10.3390/s21062222.

[28] G. Zhang *et al.*, "Discrimination Between Solitary Brain Metastasis and Glioblastoma Multiforme by Using ADC-Based Texture Analysis: A Comparison of Two Different ROI Placements," *Acad Radiol*, vol. 26, no. 11, pp. 1466–1472, Nov. 2019, doi: 10.1016/j.acra.2019.01.010.

[29] C. Russo, P. Maresca, and A. Marinelli, "Cognitive computing tools for identification and classification of brain tumors starting from Magnetic Resonance Imaging: preliminary results," in *MELECON 2022 - IEEE Mediterranean Electrotechnical Conference, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 595–599. doi: 10.1109/MELECON53508.2022.9842916.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." [Online]. Available: http://code.google.com/p/cuda-convnet/

[31] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR 2015*, Sep. 2014. doi: 10.48550/arXiv.1409.1556.

[32] A. Z. Shirazi *et al.*, "The application of deep convolutional neural networks to brain cancer images: A survey," *Journal of Personalized Medicine*, vol. 10, no. 4. MDPI AG, pp. 1–27, Nov. 01, 2020. doi: 10.3390/jpm10040224.

[33] I. A. El Kader, G. Xu, Z. Shuai, S. Saminu, I. Javaid, and I. S. Ahmad, "Differential deep convolutional neural network model for brain tumor classification," *Brain Sci*, vol. 11, no. 3, Mar. 2021, doi: 10.3390/brainsci11030352.

[34] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Classification using deep learning neural networks for brain tumors," *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 68–71, Jun. 2018, doi: 10.1016/j.fcij.2017.12.001.

[35] A. J. Fordham *et al.*, "Differentiating glioblastomas from solitary brain metastases: An update on the current literature of advanced imaging modalities," *Cancers*, vol. 13, no. 12. MDPI, Jun. 02, 2021. doi: 10.3390/cancers13122960.

[36] L. Jekel *et al.*, "Machine Learning Applications for Differentiation of Glioma from Brain Metastasis—A Systematic

Review," *Cancers*, vol. 14, no. 6. MDPI, Mar. 01, 2022. doi: 10.3390/cancers14061369.

[37] A. Stadlbauer *et al.*, "Differentiation of Glioblastoma and Brain Metastases by MRI-Based Oxygen Metabolomic Radiomics and Deep Learning," *Metabolites*, vol. 12, no. 12, Dec. 2022, doi: 10.3390/metabo12121264.

[38] S. Cha *et al.*, "Differentiation of glioblastoma multiforme and single brain metastasis by peak height and percentage of signal intensity recovery derived from dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging," *American Journal of Neuroradiology*, vol. 28, no. 6, pp. 1078–1084, Jun. 2007, doi: 10.3174/ajnr.A0484.

[39] P. Lehmann *et al.*, "Cerebral peritumoral oedema study: Does a single dynamic MR sequence assessing perfusion and permeability can help to differentiate glioblastoma from metastasis?," *Eur J Radiol*, vol. 81, no. 3, pp. 522–527, Mar. 2012, doi: 10.1016/j.ejrad.2011.01.076.

[40] M. H. Maurer *et al.*, "Glioblastoma multiforme versus solitary supratentorial brain metastasis: Differentiation based on morphology and magnetic resonance signal characteristics," *RoFo Fortschritte auf dem Gebiet der Rontgenstrahlen und der Bildgebenden Verfahren*, vol. 185, no. 3, pp. 235–240, 2013, doi: 10.1055/s-0032-1330318.

[41] G. Crisi, L. Orsingher, and S. Filice, "Lipid and Macromolecules Quantitation in Differentiating Glioblastoma From Solitary Metastasis: A ShortYEcho Time Single-Voxel Magnetic Resonance Spectroscopy Study at 3 T," *J Comput Assist Tomogr*, vol. 37, no. 2, pp. 265–271, 2013, [Online]. Available: www.jcat.org

[42] A. Romano *et al.*, "Single brain metastasis versus glioblastoma multiforme: a VOI-based multiparametric analysis for differential diagnosis," *Radiologia Medica*, vol. 127, no. 5, pp. 490–497, May 2022, doi: 10.1007/s11547-022-01480-x.

[43] M. Artzi, I. Bressler, and D. Ben Bashat, "Differentiation between glioblastoma, brain metastasis and subtypes using radiomics analysis," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 2, pp. 519–528, Aug. 2019, doi: 10.1002/jmri.26643.

[44] S. Bae *et al.*, “Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation,” *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-68980-6.

[45] J. Li, S. Liu, Y. Qin, Y. Zhang, N. Wang, and H. Liu, “High-order radiomics features based on T2 FLAIR MRI predict multiple glioma immunohistochemical features: A more precise and personalized gliomas management,” *PLoS One*, vol. 15, no. 1, Jan. 2020, doi: 10.1371/journal.pone.0227703.

[46] N. Mouthuy, G. Cosnard, J. Abarca-Quinones, and N. Michoux, “Multiparametric magnetic resonance imaging to differentiate high-grade gliomas and brain metastases,” *Journal of Neuroradiology*, vol. 39, no. 5, pp. 301–307, Dec. 2012, doi: 10.1016/j.neurad.2011.11.002.

[47] K. Skogen, A. Schulz, E. Helseth, B. Ganeshan, J. B. Dormagen, and A. Server, “Texture analysis on diffusion tensor imaging: discriminating glioblastoma from single brain metastasis,” *Acta radiol*, vol. 60, no. 3, pp. 356–366, Mar. 2019, doi: 10.1177/0284185118780889.

[48] E. Tsolaki *et al.*, “Automated differentiation of glioblastomas from intracranial metastases using 3T MR spectroscopic and perfusion data,” *Int J Comput Assist Radiol Surg*, vol. 8, no. 5, pp. 751–761, 2013, doi: 10.1007/s11548-012-0808-0.

[49] F. Dong *et al.*, “Differentiation of supratentorial single brain metastasis and glioblastoma by using peri-enhancing oedema region–derived radiomic features and multiple classifiers,” *Eur Radiol*, vol. 30, no. 5, pp. 3015–3022, May 2020, doi: 10.1007/s00330-019-06460-w.

[50] Z. R. Samani, D. Parker, R. Wolf, W. Hodges, S. Brem, and R. Verma, “Distinct tumor signatures using deep learning-based characterization of the peritumoral microenvironment in glioblastomas and brain metastases,” *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-93804-6.

[51] J. Amin, M. Sharif, N. Gul, M. Yasmin, and S. A. Shad, “Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network,” *Pattern Recognit Lett*, vol. 129, pp. 115–122, Jan. 2020, doi: 10.1016/j.patrec.2019.11.016.

[52] G. S. Tandel, A. Tiwari, O. G. Kakde, N. Gupta, L. Saba, and J. S. Suri, "Role of Ensemble Deep Learning for Brain Tumor Classification in Multiple Magnetic Resonance Imaging Sequence Data," *Diagnostics*, vol. 13, no. 3, Feb. 2023, doi: 10.3390/diagnostics13030481.

[53] L. Tariciotti *et al.*, "A Deep Learning Model for Preoperative Differentiation of Glioblastoma, Brain Metastasis, and Primary Central Nervous System Lymphoma: An External Validation Study," *NeuroSci*, vol. 4, no. 1, pp. 18–30, Dec. 2022, doi: 10.3390/neurosci4010003.

[54] S. Chakrabarty, A. Sotiras, M. Milchenko, P. Lamontagne, M. Hileman, and D. Marcus, "MRI-based identification and classification of major intracranial tumor types by using a 3D convolutional neural network: A retrospective multi-institutional analysis," *Radiol Artif Intell*, vol. 3, no. 5, Sep. 2021, doi: 10.1148/ryai.2021200301.

[55] M. M. Badža and M. C. Barjaktarović, "Classification of brain tumors from mri images using a convolutional neural network," *Applied Sciences (Switzerland)*, vol. 10, no. 6, Mar. 2020, doi: 10.3390/app10061999.

[56] H. Mzoughi *et al.*, "Deep Multi-Scale 3D Convolutional Neural Network (CNN) for MRI Gliomas Brain Tumor Classification," *J Digit Imaging*, vol. 33, no. 4, pp. 903–915, Aug. 2020, doi: 10.1007/s10278-020-00347-9.

[57] C. Ge, I. Y. H. Gu, A. S. Jakola, and J. Yang, "Enlarged Training Dataset by Pairwise GANs for Molecular-Based Brain Tumor Classification," *IEEE Access*, vol. 8, pp. 22560–22570, 2020, doi: 10.1109/ACCESS.2020.2969805.

[58] M. A. Khan *et al.*, "Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists," *Diagnostics*, vol. 10, no. 8, Aug. 2020, doi: 10.3390/diagnostics10080565.

[59] I. Shin *et al.*, "Development and validation of a deep learning-based model to distinguish glioblastoma from solitary brain metastasis using conventional MR images," *American Journal of Neuroradiology*, vol. 42, no. 5, pp. 838–844, May 2021, doi: 10.3174/AJNR.A7003.

[60] A. Stadlbauer *et al.*, “Radiophysiomics: Brain Tumors Classification by Machine Learning and Physiological MRI Data,” *Cancers (Basel)*, vol. 14, no. 10, May 2022, doi: 10.3390/cancers14102363.

[61] E. Barkan, C. Porta, S. Rabinovici-Cohen, V. Tibollo, S. Quaglini, and M. Rizzo, “Artificial intelligence-based prediction of overall survival in metastatic renal cell carcinoma,” *Front Oncol*, vol. 13, 2023, doi: 10.3389/fonc.2023.1021684.

[62] A. M. Schmid *et al.*, “Radiologists and Clinical Trials: Part 1 The Truth About Reader Disagreements,” *Therapeutic Innovation and Regulatory Science*, vol. 55, no. 6. Springer Science and Business Media Deutschland GmbH, pp. 1111–1121, Nov. 01, 2021. doi: 10.1007/s43441-021-00316-6.

[63] S. H. Yoon *et al.*, “Interobserver variability in Lung CT Screening Reporting and Data System categorisation in subsolid nodule-enriched lung cancer screening CTs Abbreviations LDCT Low-dose chest CT Lung-RADS Lung CT Screening Reporting and Data System NLST National Lung Screening Trial,” *Eur Radiol*, vol. 31, pp. 7184–7191, 2021, doi: 10.1007/s00330-021-07800-5/Published.

[64] I. Sánchez Fernández and J. M. Peters, “Machine learning and deep learning in medicine and neuroimaging,” *Annals of the Child Neurology Society*, Feb. 2023, doi: 10.1002/cns3.5.

[65] F. M. Asch *et al.*, “Human versus Artificial Intelligence–Based Echocardiographic Analysis as a Predictor of Outcomes: An Analysis from the World Alliance Societies of Echocardiography COVID Study,” *Journal of the American Society of Echocardiography*, vol. 35, no. 12, pp. 1226-1237.e7, Dec. 2022, doi: 10.1016/j.echo.2022.07.004.

[66] J. Jing *et al.*, “Development of Expert-Level Automated Detection of Epileptiform Discharges during Electroencephalogram Interpretation,” *JAMA Neurol*, vol. 77, no. 1, pp. 103–108, Jan. 2020, doi: 10.1001/jamaneurol.2019.3485.

[67] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, “Deep learning-based electroencephalography analysis: A systematic review,” *Journal of Neural Engineering*,

vol. 16, no. 5. Institute of Physics Publishing, Aug. 14, 2019. doi: 10.1088/1741-2552/ab260c.

[68] M. Abou Jaoude *et al.*, "Detection of mesial temporal lobe epileptiform discharges on intracranial electrodes using deep learning," *Clinical Neurophysiology*, vol. 131, no. 1, pp. 133–141, Jan. 2020, doi: 10.1016/j.clinph.2019.09.031.

[69] M. A. Kural *et al.*, "Accurate identification of EEG recordings with interictal epileptiform discharges using a hybrid approach: Artificial intelligence supervised by human experts," *Epilepsia*, vol. 63, no. 5, pp. 1064–1073, May 2022, doi: 10.1111/epi.17206.

[70] C. Blüthgen, A. S. Becker, I. Vittoria de Martini, A. Meier, K. Martini, and T. Frauenfelder, "Detection and localization of distal radius fractures: Deep learning system versus radiologists," *Eur J Radiol*, vol. 126, May 2020, doi: 10.1016/j.ejrad.2020.108925.

[71] A. Ciritsis, C. Rossi, M. Eberhard, M. Marcon, A. S. Becker, and A. Boss, "Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making," *Eur Radiol*, vol. 29, no. 10, pp. 5458–5468, Oct. 2019, doi: 10.1007/s00330-019-06118-7.

[72] P. Lovinfosse *et al.*, "Distinction of Lymphoma from Sarcoidosis on 18F-FDG PET/CT: Evaluation of Radiomics-Feature-Guided Machine Learning Versus Human Reader Performance," *J Nucl Med*, vol. 63, no. 12, pp. 1933–1940, Dec. 2022, doi: 10.2967/jnumed.121.263598.

[73] H. A. Haenssle *et al.*, "Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[74] H. A. Haenssle *et al.*, "Skin lesions of face and scalp – Classification by a market-approved convolutional neural network in comparison with 64 dermatologists," *Eur J Cancer*, vol. 144, pp. 192–199, Feb. 2021, doi: 10.1016/j.ejca.2020.11.034.

[75] H. A. Haenssle *et al.*, "Man against machine reloaded: performance of a market-approved convolutional neural network

in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions," *Annals of Oncology*, vol. 31, no. 1, pp. 137–143, Jan. 2020, doi: 10.1016/j.annonc.2019.10.013.

[76] P. Tschandl *et al.*, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *Lancet Oncol*, vol. 20, no. 7, pp. 938–947, Jul. 2019, doi: 10.1016/S1470-2045(19)30333-X.

[77] R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, "Comparing deep neural networks against humans: object recognition when the signal gets weaker," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.06969

[78] N. M. Singh *et al.*, "How Machine Learning is Powering Neuroimaging to Improve Brain Health," *Neuroinformatics*. Springer, 2022. doi: 10.1007/s12021-022-09572-9.