uc3m | Universidad **Carlos III** de Madrid

Degree of Data Science and Engineering

2022 – 2023

Bachelor Thesis

“Diagnostic Aid System for Diffuse Gliomas using Deep Learning: Determination of IDH Status through Magnetic Resonance Imaging”

Sergio Aizcorbe Pardo

Supervisor: Fernando Díaz de María

Leganés, September 2023

ABSTRACT

Gliomas are the most common malignant primary brain tumors in adults. Their accurate classification plays an essential role in early treatments and guiding possible interventions. This research, carried out in collaboration with Hospital Gregorio Marañón of Madrid, explores how deep learning could be utilized to determine the status of the IDH gene in gliomas in Magnetic Resonance Imaging (MRI). The ultimate aim of this project is to contribute to the advancements of medical imaging research and develop a classification system that can accurately identify the presence of the IDH gene mutation in gliomas.

Throughout the study, several deep learning architectures have been tested and compared, including both Convolutional Neural Networks (CNN) and Transformers. The models achieved accuracies of 82-93% and AUC scores of 82-92%.

The findings demonstrate the advantages of these cutting-edge computational techniques in MRI analysis. Not only do they have the potential to improve the accuracy and efficiency of detecting the IDH gene mutation in gliomas, but they also present themselves as valuable support tools for neuroradiologists in their diagnostic processes.

Key Words: Deep learning; Neural Networks; IDH Classification

ACKNOWLEDGEMENTS

Firstly, I would like to dedicate this work to my family for being an unwavering support at all times and for always believing in me. Their love and encouragement have been decisive.

Secondly, to my friends, for all the good times we have shared and for making this journey even more memorable. Their laughter and companionship have brought me joy.

Lastly, to my supervisor Fernando Díaz de María and the neuroradiologists Juan Adan Guzmán-De-Villoria and Pilar Fernández from Hospital Gregorio Marañón, who have guided me and never hesitated to help at any time.

Thank you to all of you.

INDEX OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

GLOSSARY OF ABBREVIATIONS

ADC	Apparent diffusion coefficient maps
ADG	Adult diffuse glioma
aKG	a-ketoglutarate
ANTS	Advanced Normalization Tools
ATRX	Alpha thalassemia/mental retardation syndrome X-linked
AUC	Area Under the ROC Curve
BCE	Binary Cross-Entropy
BN	Batch Normalization
BRATS	Multimodal brain tumor segmentation
CAD	Computer-aided detection
CaPTk	Cancer Imaging Phenomics Toolkit
CNN	Convolutional Neural Network
CNS	Central nervous system
CPU	Central Processing Unit
DICOM	Digital Imaging and Communications in Medicine
DTI	Diffusion tensor imaging
DWI	Diffusion weighted imaging
EGFR	Epidermal growth factor receptor
GBM	Glioblastoma
GLCM	Gray level co-occurrence matrix
GPU	Graphics Processing Unit
HARDI	High Angular Resolution Diffusion Imaging
HGG	High grade glioma
HGUGM	Hospital General Universitario Gregorio Marañón
IDH	Isocitrate dehydrogenase
kNN	k-Nearest Neighbors
FLAIR	Fluid Attenuated Inversion Recovery
LGG	Low grade glioma
MGMT	Methyl-guanine methyl transferase
MLP	Multi-layer Perceptron
MR / MRI	Magnetic Resonance / Magnetic Resonance Imaging
NIFTI	Neuroimaging Informatics Technology Initiative
NLP	Natural Language Processing

NN	Neural Network
RAM	Random Access Memory
ReLU	Rectified Linear Unit
REMBRANDT	The Repository of Molecular Brain Neoplasia Data
RF	Random Forest
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
SWI	Susceptibility weighted imaging
TCIA	The Cancer Imaging Archive
T1CE / T1-CE	T1-weighted with contrast enhancement
T1 / T1-W	T1-weighted
T2 / T2-W	T2-weighted
TERT	Telomerase reverse transcriptase
TP53	Tumor suppressor protein p53
UNETR	Unet Transformer
ViT	Vision Transformer
WandB	Weight & Biases
WBCE	Weighted Binary-Cross Entropy
WHO	World Health Organization
WSL	Windows Subsystem for Linux
WT	Wild Type

1. INTRODUCTION

1.1. Background

Gliomas are tumors that develop from the anomalous growth of the brain's glial cells, which are essential for supporting and protecting neurons. About 80% of Adult Diffuse Gliomas (ADG) are malignant brain neoplasms [1]. According to the World Health Organization (WHO) classification system, tumors can be classified into four different grades: Grades 1 and 2 are low-grade gliomas (LGG) and Grades 3 and 4 are high-grade diffuse gliomas (HGG). In addition to this system, tumors can be categorized by their origin cell. Oligodendrogliomas originate from oligodendrocytes and are typically categorized as Grades 2 and 3. On the other hand, astrocytomas originate from astrocytes and their degree of severity can vary from harmless Grade 1 to aggressive Grade 4 glioblastomas (GBM).

Recent breakthroughs in molecular genetics have changed the understanding of the biology of gliomas, which has led to the incorporation of molecular markers into the classification system. Among these biomarkers, isocitrate dehydrogenase mutations (IDH1 or IDH2) are essential in determining prognosis and making therapeutic decisions. IDH mutations, primarily observed in LGG, are associated with favorable outcomes and increased treatment sensitivity [2] [3].

Nonetheless, IDH status determination is invasive and requires a tissue sample obtained either by a stereotactic biopsy or a surgical resection. Both of these procedures have operating risks and the potential possibility for sampling error. Magnetic Resonance Imaging (MRI) is a non-invasive imaging modality widely used for the detection, diagnosis and monitoring of gliomas. Advancements in MRI techniques provide insights of the molecular and metabolic morphology of tumors. Despite this, detecting information about IDH mutations from MRI images is complex and presents a challenge [2] [3].

Traditionally, radiomics methods have been used to extract features that facilitate IDH classification. However, in the recent years, deep learning, a subfield of machine learning, has revolutionized the medical image domain. The brain-inspired models known as neural networks have the ability to recognize patterns and features in imaging data that might be imperceptible to the human eye. The development of these new model architectures has the potential to lead to more accurate diagnostics and predictive analytics in the context of MRI.

Early detection and classification of gliomas is crucial as they can affect brain functions, leading to cognitive, motor and sensory impairments. The type and stage of the tumor determine treatment modalities, such as surgery, radiation or chemotherapy. An accurate diagnosis allows medical professionals to set realistic expectations and therapeutic strategies, which can improve patient outcomes and quality of life significantly [4].

1.2. Project Motivation

The field of medical diagnostics has always been an area of investigation and innovation. The brain, with its complex architecture and functions, presents several challenges when it comes to the treatment of diseases. Current methods for analyzing gliomas and, specifically, the determination of the IDH status, can cause many complications and risks, which brings us up with the question: Could there be a safer and effective alternative?

Magnetic Resonance Imaging (MRI) is fundamental in neuroimaging. Its ability to provide detailed images of the brain make it an invaluable tool. Recent innovations in deep learning offer a level of precision previously unthinkable in medical applications. The potential of neural networks to improve the diagnostic capabilities of MRI is worth exploring.

In conclusion, the development of a diagnostic aid system that uses the potential of deep learning to determine IDH status through MRI can potentially change the way gliomas are diagnosed and treated.

1.3. Objectives of the Study

The main objective of this project is to provide neuroradiologists with a decision support tool, with the ultimate goal of improving patient outcomes and reducing procedural risks.

Moreover, with the increasing volume of imaging data and the complexity of tumor biology, manual interpretation becomes time-consuming and might not be consistent. An automated, deep learning-driven system would not only mitigate this, but also affect personalized treatment strategies.

To achieve this, a series of novel deep learning models have been adapted and evaluated to automatically extract image features from magnetic resonance scans and reveal patterns or signals associated with the IDH status.

Once the primary goal of the work has been established, it is possible to identify the proposed steps to achieve that objective:

- Analyze the databases and pre-process the images using the latest methods.
- Conduct an analysis of the most appropriate techniques for image classification.
- Train and test different model and analyze which architecture provides better results and is more efficient in line with the needs of the problem.
- Study the results obtained to determine in which situations the system makes the most mistakes and try to find possibilities for improvement.
- Conclusions of the work and potential objectives for future research.

1.4. Socio-economic Environment

Currently, this project is not supported by external funds. Neither employees nor patients have received financial compensation as a result of the study. Furthermore, there are no direct costs attributed to the hospital for this research.

- **Economic Impact**: The use of deep learning in a diagnostic aid system may result in more precise and effective diagnosis. An early and accurate diagnosis can save prolonged hospital stays, repeated testing and associated medical costs.

- **Impact on society:** A better diagnostic system can significantly improve patient treatment since they can start receiving targeted therapy earlier. Additionally, this may increase the hospital's standing and establish it as a pioneer in cutting-edge medical care and use of technology.

- **Impact on the environment:** The use of digital tools and AI-based diagnostic techniques may reduce the need for physical resources and consumables, resulting in a lower environmental footprint. Reduced redundancy in MRI scans can also save energy and decrease the impact that MRI machines have on the environment.

- **Ethical Impact**: Deep learning systems for medical diagnosis have ethical implications that should be considered. To guarantee the confidentiality of patient data, the strictest measures have been taken.

1.5. Regulatory Framework

An analysis of the applicable legislation to this thesis has been conducted, taking into account the ethical and professional responsibilities, as well as the potential risks that this work could involve. The integration of deep learning models into the medical field in Spain requires understanding the applicable regulatory frameworks. This section explores the regulatory context relative to this project.

The imaging data collected for the study is fully pseudo-anonymized, ensuring that it contains no information that could identify the patient, as mandated by the Spanish Constitution, specifically in Article 18 [5]. Access to the pseudo-anonymized data is restricted to the researchers of the study from the GU Gregorio Marañón Hospital and the Carlos III University of Madrid. The development of this project has been approved by health authorities (Spanish Agency for Medicines and Health Products, foreign health authorities) [6] and the Research Ethics Committee with Medicines (CEIm) [7].

Confidentiality of this data is always maintained in accordance with the regulations on information handling as set out in the legislation related to the European General Data Protection Regulation (GDPR) [8] and the Organic Law 3/2018 "Protection of Personal Data and guarantee of digital rights" (LOPDGDD) [9].

Finally, when using machine learning algorithms, a series of ethical principles have been taken into account for the responsible use of this technology [10]. The following are included among the principles of responsibility:

- **Human enhancement:** People involved with this technology must understand the consequences of incorrect predictions, especially when automating critical processes that can have a significant impact on human lives.
- **Reproducibility:** The built infrastructure must allow a reasonable level of reproducibility in all operations of Machine Learning systems.
- **Displacement strategy:** It is necessary to identify and document the information needed to develop business processes in a way that mitigates the automation of workers.
- **Practical accuracy:** Accuracy and cost metrics must be aligned with the specific domain applications.
- **Data risk awareness:** It is necessary to commit to developing and improving processes and infrastructures that ensure the security of the data and models used.

1.6. Project Structure

Below is a summary of the chapters found in this report:

- **Introduction:** An overview of the problem at hand is presented. Additionally, objectives, socio-economic impact and the encompassing regulatory framework are included.
- **Literature Review:** This section describes existing knowledge on the subject. It provides a detailed review of relevant literature and the latest advancements related to the research topic.
- **Data Collection and Processing:** This section details the methodologies employed to gather and refine the MRI datasets used. It describes the sources, selection criteria and preprocessing steps to ensure the MRIs are prepared for deep learning-based analysis of IDH status.
- **Modeling:** Examines several convolutional neural networks (CNN) and attention models used for the current study. Each model's architecture, main features and functioning is studied. The primary goal is to offer a clear and precise understanding, allowing the reader to fully grasp the experimentation and the analysis process that has been carried out.
- **Evaluation and Results:** A comparative analysis of the trained models and their respective evaluation metrics. The main objective of this section is to accurately display the tests conducted and their results, thus providing an overall understanding of the work accomplished.

- **Conclusions and Future Work:** An in-depth analysis of all findings is conducted, giving a complete view of the entire study. A segment dedicated to potential future research is also included, identifying less-explored research areas and suggesting possible directions for upcoming investigations.

2. LITERATURE REVIEW

2.1. Preliminary Concepts

2.1.1. Types of MRI Sequences

Magnetic Resonance Imaging (MRI) is a central tool in brain imaging, providing an in-depth view of the brain's structure and activities. In this section, the main types of MRI sequences are briefly covered, providing insights on their technical basis and clinical applications:

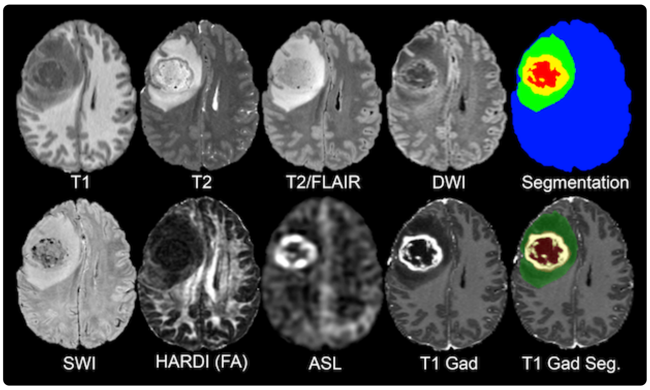


Fig. 2.1. MRI Sequences of TCIA Database [38]

- **T1-weighted (T1):** Produces high-contrast anatomical images using the longitudinal relaxation time of tissues; fat appears bright and cerebrospinal fluid dark.
- **T1-weighted Post-Contrast (T1CE):** A T1 image post-gadolinium-based contrast agent administration, highlighting vascular structures and potential abnormalities.
- **T2-weighted (T2):** Utilizes transverse relaxation time to make fluid appear bright. Ideal for spotting edemas, injuries or inflammations.
- **Fluid-Attenuated Inversion Recovery (FLAIR):** A T2 sequence with cerebrospinal fluid signal nulling, enhancing the detection of lesions adjacent to fluid-filled spaces.
- **Diffusion-Weighted Imaging (DWI):** Captures water molecule diffusion, aiding in immediate detection of ischemic strokes due to cellular swelling restricting diffusion and computes apparent diffusion coefficient maps (ADC).
- **Susceptibility Weighted Imaging (SWI):** Magnifies magnetic susceptibility differences to visualize blood products and vascular malformations.
- **High Angular Resolution Diffusion Imaging (HARDI):** An enhanced DWI, recording complex water molecule diffusion to better capture neural pathways in the white matter.
- **Arterial Spin Labeling (ASL):** A method that uses magnetically labeled blood water to measure cerebral blood flow for vascular-based brain studies.

2.1.2. Glioma Characterization

Since 2016, WHO classification includes molecular markers as a crucial part of the diagnosis of LGG due to an increasing awareness of their prognostic and therapeutic significance [11].

“Currently, one of the key known factors in predicting survival is the presence or absence of the isocitrate dehydrogenase (IDH) (IDH 1 and IDH 2) genotype [12] [13] [14]. In both low-grade and high-grade gliomas, the mutation is detectable as early as the oncogenesis phase, and it is frequently linked to elevated methylation in gliomas. Based on this mutation, glioma subtypes are classified either as IDH positive (mutant) or IDH negative (wildtype) [15] [4].”

IDH mutant types of grade 2 and 3 can be further sub-classified based on the presence or absence of 1p/19q co-deletion [16] [17] [18]. Other genetic alterations include mutations in telomerase reverse transcriptase (TERT) gene promoter [19] [20], methylation of the methyl-guanine methyl transferase (MGMT) gene promoter [21], epidermal growth factor receptor variant III (EGFRvIII) [22], alpha thalassemia/mental retardation syndrome X-linked (ATRX) loss [23], tumor suppressor protein p53 (TP53), and phosphatase and tensin homolog mutations etc. [24]

2.1.3. Neural Networks

Neural networks, inspired by the structure of the human brain, form the foundation of modern artificial intelligence, particularly deep learning. These networks consist of interconnected nodes or "neurons" organized in layers. An input layer accepts data, which is processed through one or more hidden layers. Ultimately, an output layer produces the final output.

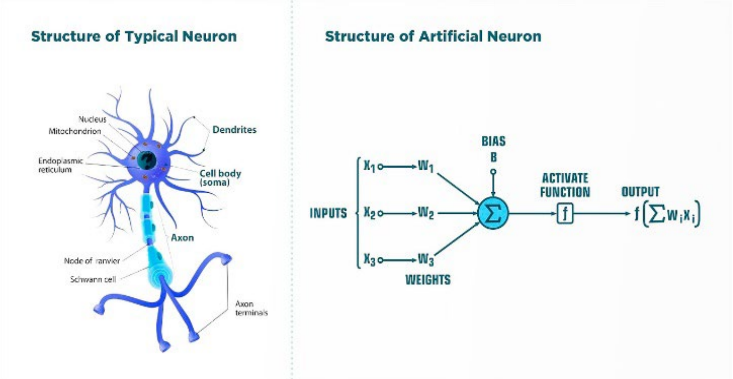


Fig. 2.2. Comparison of the Structure of a Brain and a Neural Network *[79]*

Each connection between the neurons is associated with a weight, which adjusts over time during training. When data flows through the network, these weights are adjusted dynamically and determine how much influence each input has on the final output. The process of adjusting these weights, based on the difference between the network's prediction and the actual result is called “backpropagation”.

Activation functions, like the sigmoid or ReLU, introduce non-linearity into the system, allowing neural networks to capture complex patterns and relationships in the data. This capacity to learn and generalize from vast amounts of data makes neural networks particularly powerful for tasks like image recognition or natural language processing.

Despite their complexity, the interpretation of neural networks' internal decision-making remains uncertain. This concept is known as “black box”.

2.2. Deep Learning in Medical Imaging

These days, deep learning algorithms are receiving a lot of attention as a solution to various problems in the medical imaging fields. New developments are continuously published and the various contributions make the current models more efficient [25].

In the past, computer-aided detection systems (CAD) have been used to solve image-related issues. Nonetheless, CAD systems produce more false positives than doctors do, which has increased assessment times and led to unnecessary biopsies [26] [27]. Deep learning technology allowed for the resolution of these issues, saving up time for humans to work on other useful activities. “However, the advent of this technology does not mean the ultimate replacement of physicians, especially radiologists. Instead, it helps radiologists to diagnose patients more accurately.” [28]

In the following sections, we will analyze the key neural networks used for image-related tasks, such as segmentation, object detection, and, specifically for this project, classification.

2.2.1. Convolutional Neural Networks

The convolutional neural network (CNN) is a specialized kind of neural network used in the field of image analysis. Unlike standard feedforward neural networks, CNNs are designed to adaptively learn spatial hierarchies of features from input images. They employ a mathematical operation called convolution to process data from image pixels.

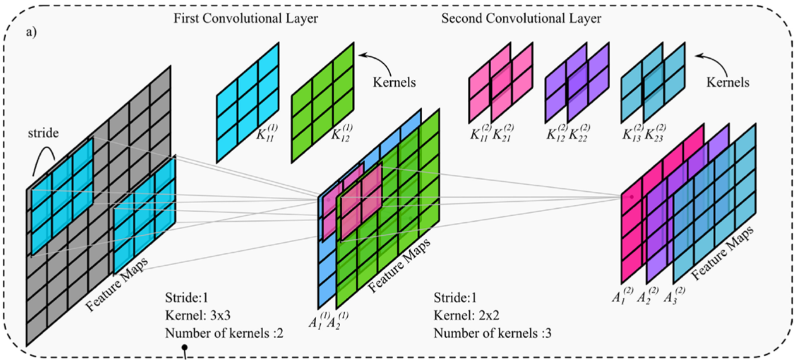


Fig. 2.3. Example of a two-layer CNN with two and three filters

The architecture is distinctive for its pattern of connecting neurons, where input information is processed by convolutional layers, pooling layers and fully connected layers.

Due to their design, CNNs are exceptionally good at identifying patterns and features in images, making them suitable for tasks such as object detection, image classification and segmentation. The convolution process in the CNN layers is governed under the following mathematical expression:

$$A_1 = f\left(\sum_{i=1}^{M} X_i * K_{i1} + b_1\right) \tag{1}$$

Where, for M feature maps, $*$ represents the convolution between the i-th map X_i and the filter K_{i1} at the map equivalent depth. b_i and f are the bias and the activation function, respectively [29].

2.2.2. Attention Models

The principle of attention models is based on the optic nerve of the visual system on the idea that not all perceived information has the same degree of importance. Following these premises, scientists have developed new attention mechanisms. Many papers have proven that these models can achieve as much or even more accuracy than convolutional neural networks [29].

The attention mechanism assigns different significance levels to various parts of an input sequence during the production of an output. It operates based on the concept of queries (Q), keys (K) and values (V), which are mathematical representations of the input. Their interaction determines the weighting or attention score. The attention weights (A) are computed by measuring the similarity between two elements and their key-value pairs according to:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Where Q, K and V matrices contain the query and key vectors, respectively and d_k is the dimension of the key vectors. The values (V) are then weighted according to these scores. With this approach, the model can dynamically prioritize which parts of the input data are crucial for a particular context.

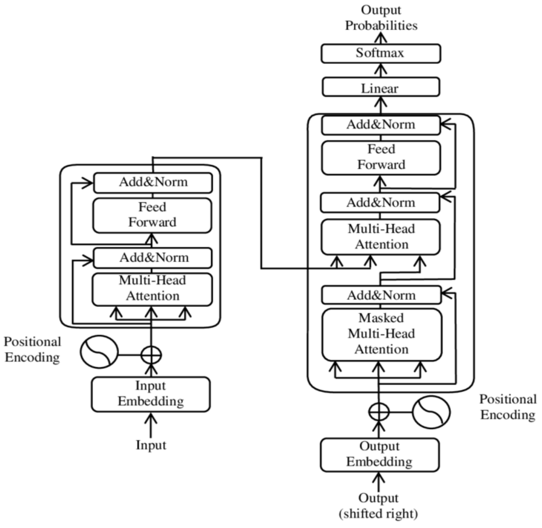


Fig. 2.4. Overview of Transformer architecture *[30]*

On the other hand, the multi-head attention mechanism, introduced in the seminal paper

"Attention Is All You Need" [30], takes this concept further. Instead of processing the input data as a whole, it divides it into multiple segments or "heads." Each segment processes the data independently, enabling the mechanism to differentiate relationships and patterns at the same time. These architectures have revolutionized deep learning tasks, particularly in natural language processing (NLP), by providing a framework that can capture complex patterns within data sequences.

2.3. Methods for Determining IDH Status

2.3.1. Radiomic Methods

Radiomics is a research field in which radiographic images are used to compute attributes such as geometry, intensity and histogram features, as well as image textures, in order to capture phenotypic patterns. The features include shape features such as volume, sphericity, mesh surface area etc., histogram features that may include mean, median, energy, entropy, 10/90 percentile intensity, etc., and texture features of Gray-Level Co-occurrence Matrix or Gray-Level Dependence Matrix among others [31]. VASARI (Visually AcceSAble Rembrandt Images) MRI feature set [32, 33, 34] is a widely adopted feature set which consists of 24 observations familiar to neuroradiologists, to describe the morphology of brain tumors

These methods generally make use of supervised learning methods for classification like logistic regression, support vector machines (SVM), random forests (RF), k-nearest neighbors (kNN) or multi-layer perceptrons (MLP).

2.3.2. Deep Learning Methods

Deep neural nets have been widely used to classify gliomas with IDH mutation, generally, CNNs. These types of models extract the image features automatically and diminish preprocessing steps required in radiomics [2]. Nevertheless, their “black-box” nature represents a significant challenge with these models, which often complicates the decision-making process and impedes interpretability.

Chang et al. [35], Liang et al. [36], and Nalawade et al. [37], all of which classified the IDH status and used the same dataset as this project, achieved accuracy of 89.1%, 91.1%, and 83.5%, respectively. Some common examples of neural network architectures include ResNet, DenseNet, EfficientNet and Inception.

2.3.3. Discussion

The majority of the studies to this date have mainly focused on the online open-source TCIA (The Cancer Imaging Archive) data [38] [39, 40, 41, 42, 43] while remaining studies have employed local datasets [44, 45, 46, 47, 48, 49, 50] . Studies on TCIA have demonstrated IDH predictive accuracies ranging from 72-92% while on other datasets the accuracies ranged from 73-90% based on multimodality features (T1, T1-CE, FLAIR, and T2). A consensus from all these studies shows that the attributes computed from T1-CE and FLAIR have been highly distinctive of IDH mutation than the ones computed from T1 and T2 weighted MRI [2].

The features of importance have been highly dependent on the grade of the tumor under consideration. For example, Javier et al. reported that ADC feature maps were more discriminative in grade 2 gliomas where IDH wildtype were highly associated with lower ADC values with poor clinical outcomes [51], however a similar trend was not observed in high grade gliomas [52].

The systematic literature review found that the highest classifier for IDH status was conventional radiomics with CNN deep learning derived features, which achieved an AUC of 0.95 (sensitivity of 94.4%, specificity of 86.7%) [53]. According to this study, a standard sequence image acquisition, use of texture-based features (gray-level co-occurrence matrix, followed by the gray-level run-length matrix) with deep learning-derived features and an SVM machine learning model may result in an optimal radiomic pipeline.

3. DATA COLLECTION AND PROCESSING

3.1. Databases

One of the goals of our project was to use real-life MR sequences to train a classifier capable of determining the IDH status. However, there was a limitation in the patient sample size, having only 41 patients in our initial dataset from Gregorio Marañón. Due to the need for substantial data samples to effectively train machine learning models – ensuring they generalize well and reduce the risk of overfitting –an additional database comprising 501 patients from “The Cancer Imaging Archive” [54] was used to supplement the original dataset.

3.1.1. TCIA Database

The main and largest database of the project is “The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM)” [38], which was incorporated to augment the training data and thereby enhance the robustness and reliability of the classifier. The database includes the following information:

1. **Magnetic Resonances (MRs):** Preoperative Magnetic Resonances performed on a 3.0 T scanner (Discovery 750, GE Healthcare, Waukesha, Wisconsin, USA) and a dedicated 8-channel head coil (Invivo, Gainesville, Florida, USA). The imaging protocol included 3D T2-weighted, T2/FLAIR-weighted, susceptibility-weighted (SWI), diffusion-weighted (DWI), pre- and post-contrast T1-weighted images, 3D arterial spin labeling (ASL) perfusion images, and 2D 55-direction high angular resolution diffusion imaging (HARDI). Over the study period, two gadolinium-based contrast agents were used: gadobutrol (Gadovist, Bayer, LOC) at a dose of 0.1 mL/kg and gadoterate (Dotarem, Guerbet, Aulnay-sous-Bois, France) at a dose of 0.2 mL/kg.

2. **Molecular Data:** The subjects have undergone IDH mutation testing through genetic sequencing of tissue obtained during biopsy or resection. Additionally, all grade 3 and 4 tumors were tested for MGMT methylation status using a sensitive quantitative PCR assay. Information regarding the presence of 1p/19q codeletion has also been collected.

3. **Additional Information:** The dataset includes additional information such as age, sex, overall survival and tumor segmentation. This information is crucial for analyzing and understanding the relationship between molecular data, clinical characteristics, and the outcomes of the glioma patients.

3.1.2. HGUGM Database

The second major subset of images is from the Hospital General Universitario Gregorio Marañón (HGUGM) and includes MR scanners from patients who underwent surgery for medium and high-grade gliomas. The patients have a confirmed histopathological diagnosis and have undergone an MRI with intravenous contrast between March 2022 and March 2023. It includes the following information:

1. **Magnetic Resonance Imaging (MRIs):** Pre-operative MRI images of patients with histologically confirmed gliomas. The scans include multiple MRI sequences, such as T1-weighted images (with and without contrast), T2-weighted images, T2-FLAIR, and susceptibility-weighted images (SWI) with the administration of contrast agent gadoterate (Dotarem, Guerbet, Aulnay-sous-Bois, France) at a dose of 0.1 mL/kg. All these sequences are 3D with isotropic voxel size and were obtained using a single 1.5 T MRI machine (Ingenia 1.5T Omega, Philips Healthcare, Best, The Netherlands).
2. **Molecular Data:** The status of IDH mutation is available for each patient to serve as a reference for model training and evaluation.
3. **Additional Information:** Demographic and clinical information, such as age, sex, and tumor location, is included to explore the possible influence of these factors on glioma classification and stratify the analysis accordingly.

3.2. Database Comparison

The following table provides a comparative overview of the two databases, the TCIA Database and the HGUGM Database, each of which contains a diverse range of patient information pertaining to sex, glioma diagnosis and other pathological data:

TABLE 1.
DATABASE COMPARISON

	TCIA Database		**HGUGM Database**	
	N	%	N	%
Patients	501		42	
Sex	501	100.0	42	100.0
Male	299	59.7	26	61.9
Female	202	40.3	16	38.1
IDH status	501	100.0	42	100.0
Mutated (positive)	103	20.6	4	9.5
Wildtype (negative)	398	79.4	38	90.5
Unknown	0	0.0	0	0.0
1p/19q codeletion	410	81.8	41	97.6

Co-deleted	15	3.0	1	2.4
Intact	395	78.8	40	95.2
Unknown	91	18.2	1	2.4
WHO Grade	501	100.0	42	100.0
2	56	11.2	2	4.8
3	43	8.6	1	2.4
4	402	80.2	39	92.8
Diagnosis	501	100.0	42	100.0
Glioblastoma, IDH-mutant	0	0.0	0	0.0
Glioblastoma, IDH-wildtype	374	74.6	34	83.4
Astrocytoma, IDH-mutant	90	18.0	3	7.1
Astrocytoma, IDH-wildtype	24	4.8	3	7.1
Oligodendroglioma	13	2.6	1	2.4

3.3. Pre-Processing Pipeline

To ensure consistency and streamline data processing, both the HGUGM and TCIA databases, comprising MRI scans from T1, T1CE, T2 and FLAIR modalities, were reorganized using the Brain Imaging Data Structure (BIDS) [55]. This standardization facilitated a more seamless integration and analysis of the datasets, integrating best practices in neuroimaging data management.

Magnetic Resonance Imaging (MRI) data can exhibit significant variations when collected from different scanners, especially when these scanners operate at distinct magnetic field strengths. The TCIA Database, derived from a 3.0 T scanner whereas the HGUGM Database was from a 1.5 T scanner. Scanners with a higher field strength, like the 3.0 T, typically provide images with higher signal-to-noise ratios and finer spatial resolution. Conversely, 1.5 T scanners could offer robust clinical utility but may present lower resolution and contrast in their images.

Preprocessing and standardizing MR images is fundamental. Without such normalization, the heterogeneity in the data can lead to biased or erroneous interpretations, especially when integrating datasets for unified analyses or applying machine learning models. The methodological process used for the preprocessing of two databases, HGUGM Database and TCIA Database, is described hereunder.

3.3.1. DICOM to NIFTI Conversion

The HGUGM Database images were converted from DICOM format to NIFTI format using the dcm2niix tool [56]. NIFTI format offers advantages over DICOM, including reduced file size, improved 3D image sequence handling and incorporation of voxel-wise statistical data. These benefits facilitate more efficient computational processing and compatibility across diverse software platforms.

3.3.2. Spatial Normalization

Spatial normalization is a fundamental practice in MRI data processing. It ensures that images from different subjects align to a common anatomical template. In order to standardize the spatial attributes of the HGUGM Database, the characteristics of the TCIA Database served as benchmarks and Advance Normalization Tools was used [57] [58]:

- **Image resampling:** Images have been reshaped to the size of 240x240x155 voxels. The interpolation method used was Windowed Sinc Interpolation**,** as it tends to retain sharp edges better than other methods like B-Spline or Gaussian, which could be important in capturing the details of the tumor regions. This is especially relevant in this case, as IDH status is known to be associated with morphological characteristics on MRI.
- **Isotropic resampling:** The spacing of every image was reconfigured to 1 mm isotropic resolution using automated non-linear registration.
- **Image reorientation:** Images were reoriented to RAI (Right, Anterior Interior) orientation, which provides a consistent anatomical viewpoint and facilitates comparative analysis. The origin and direction of the images were also unified.

3.3.3. BraTS Pre-processing Pipeline

The HGUGM Database was subjected to the BraTS preprocessing pipeline using CaPTk (Cancer Imaging Phenomics Toolkit) [59]. BraTS (Brain Tumor Segmentation Challenge) [60] is an annual competition that focuses on the evaluation of state-of-the-art methods for the segmentation of brain tumors in multimodal MRI scans. The pre-processing pipeline is as follows:

- **Co-registration:** It is crucial for aligning multiple images or sequences of a given subject into a common spatial framework. This alignment ensures that voxel-wise comparisons or computations are made on anatomically equivalent regions. CaPTk uses SRI-24 Atlas [61] as reference for image registration:
  - **N4 Bias Field Correction** was applied to minimize lighting artifacts.
  - **Rigid Registration** was employed to align the T1, T2, FLAIR images with the T1CE.
  - A further **rigid registration** was carried out to map the T1CE with the SRI-Atlas [61].
  - The derived transformations were then applied to the reoriented images.
- Deep-Learning based **Skull-stripping** [62].
- Deep-Learning based **Tumor Segmentation** using DeepMedic [63] [64].

3.3.4. Intensity Matching

Given that the HGUGM Database was acquired from a 1.5 T scanner, its images inherently differ in intensity distributions compared to the TCIA database, which was obtained from a 3.0 T scanner. It was crucial to harmonize these intensity differences to ensure that the deep learning model is not biased.

For every MR image, its histogram is computed and KMeans clustering is applied to group histograms of similar intensity distributions. This step simplifies the task of finding a close match for a given image's histogram from a large set of reference histograms like the TCIA database. The cluster for the input histogram is predicted using the trained KMeans model to find the closest match in reference histograms. Then, it computes the distances between the cluster center and all histograms in the reference set, choosing the histogram with the smallest distance as the closest match.

3.3.5. Intensity normalization

Intensity normalization is a very important step in preprocessing MRI data. It reduces variability in image intensities among different scans and subjects. This step ensures that intensity values of MR images are standardized which facilitates comparisons. The histograms images before and after the normalization process can be seen below.

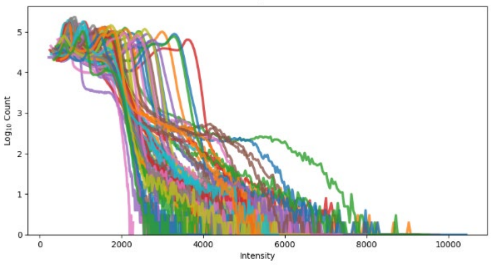


Fig. 3.1. T1 without Normalization

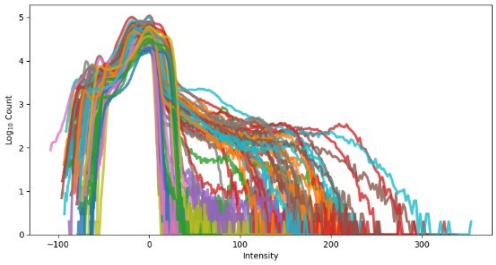


Fig. 3.2. T1 with Normalization

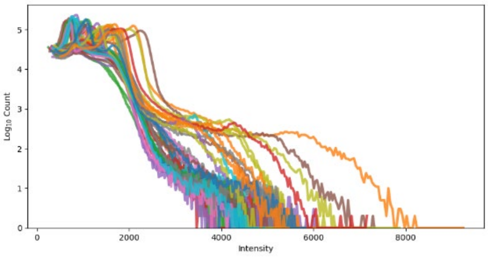


Fig. 3.3. T1CE without Normalization

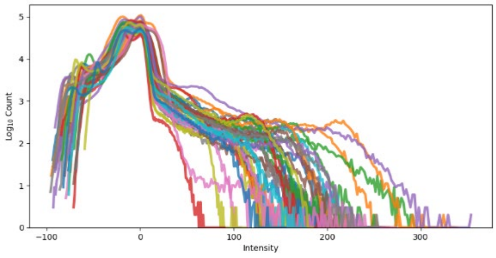


Fig. 3.4. T1CE with Normalization

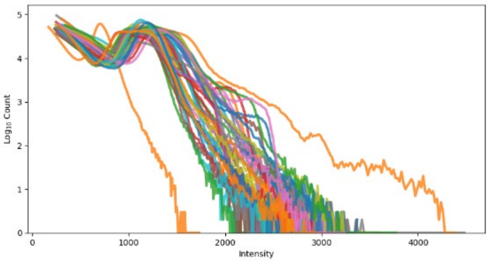


Fig. 3.5. FLAIR without Normalization

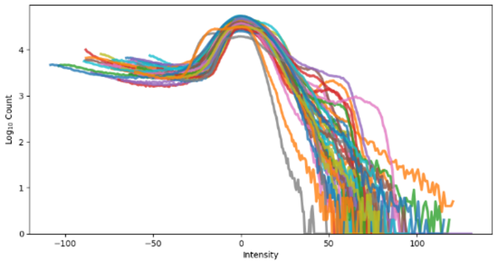


Fig. 3.6. FLAIR with Normalization

WhiteStripe normalization method has been used. The WhiteStripe method [65], is a technique for intensity normalization that operates by identifying white matter “stripes” in MR images, which are relatively unaffected by pathology. By using these stripes as a reference, the method scales the intensity values of the entire image.

3.3.6. Tumor cropping

One of the pivotal steps in the pre-processing pipeline of the project is the tumor cropping phase. Original MR 3D images have a dimensionality of 240x240x155 voxels, which could be extensive for model training and might contain extraneous information that does not necessarily contribute to the diagnosis. To minimize the computational cost, the tumors are cropped using their respective segmentation masks. This cropped region ensures that the primary focus remains on the tumor and facilitates the neural network's understanding.

A 3D bounding box is computed around the center of the tumor based on the segmentation mask (previously computed), with dimensions 128x128x64. These choices as the sizes along the axes represent the closest power-of-2 of the average sizes of the tumors and facilitate faster data access for GPU memory architectures.

In the case of tumors which are very small in comparison to the designated cropping box, the entire tumor and brain remain within the cropped region, ensuring that no relevant information is left out.

3.3.7. Conversion to NPZ

Prior to model training, the NIFTI images were converted to the NPZ format. This conversion was motivated by the efficiency and convenience offered by this format. The NPZ format, native to the NumPy library [66], allows for seamless storage and access of large numerical arrays, making it particularly suitable for high-dimensional imaging data. Additionally, NPZ files are compressed resulting in faster read-write operations.

3.4. Data Cleaning

Before the model training phase, the decision was made to focus only on T1CE and FLAIR images. This choice was based on the consensus of many studies [2], which demonstrated that the attributes computed from T1CE and FLAIR images were more distinctive of IDH mutation than the ones computed from T1 and T2 weighted sequences.

Moreover, while T2 images were accessible, sequences from the Gregorio Marañón Images were bi-dimensional and obtained in the coronal plane, resulting in the exclusion of some brain regions. Regarding T1 images, they are similar to T1CE images with the main difference being that tumor regions within T1 sequences had less enhancement.

Additionally, a review revealed that some of the data did not contain essential IDH information, was corrupted or was missing. Such inconsistencies were removed to avoid introducing bias or errors to the deep learning model. The refined datasets stand as follows: HGUGM Database now consists of 40 patients, while the TCIA Database has been narrowed down to 494 patients.

3.5. Auxiliary Features

To enhance the diagnostic precision of our model, several clinically significant features were integrated alongside the MRI data. Three features were selected: age, sex and tumor grades.

- **Age:** Age is an influential factor in medical diagnostic processes. To ensure that the age variable is on a similar scale with other features it was normalized.

- **Sex:** Gender differences can sometimes play a role in diseases, making it a relevant feature. To make this categorical variable usable, it was encoded, turning it into a format suitable for machine learning algorithms.

- **Tumor Grades:** The grading of tumors provides crucial information about the malignancy and aggressiveness of the tumor. This categorical feature, like sex, was encoded to ensure its integration into the training process.

Ultimately, these features were concatenated into a feature vector.

3.6. Data Splitting

Given the two primary datasets: HGUGM Database and TCIA Database. Both databases contain information on gliomas, categorized as either IDH positive or IDH negative. There exists a data imbalance between the two databases:

- HGUGM Database comprises 40 patients.
- TCIA Database encompasses 494 patients.

Moreover, there also exists an imbalance in the IDH status within each database:

- HGUGM Database: 4 patients are IDH Positive, while 38 are IDH Negative.
- TCIA Database: 103 patients are IDH Positive, and 398 are IDH Negative.

To ensure a balanced evaluation, each dataset was split based on the IDH status, resulting in four subsets: HGUGM-IDH-Negative, HGUGM-IDH-Positive, TCIA-IDH-Negative, and TCIA-IDH-Positive. For each of these subsets, the data was divided as follows:

1. 80% was allocated to the **training set**.
2. From this training set, 25% was further designated as the **validation set**.

3. 20% of the entire subset was reserved for the **testing set**.

Post these divisions, the subsets from both databases were integrated, merging all the respective training, validation and testing sets.

The following table includes the distribution of data splits across the two databases and their respective training, validation and testing sets:

TABLE 2.
DATASET SPLIT DISTRIBUTION

Root Training Set (80%)				**Testing Set (20%)**	
Training Set (75%)		**Validation Set (25%)**			
318 patients		107 patients		109 patients	
295 patients	23 patients	99 patients	8 patients	100 patients	9 patients
TCIA	**HGUGM**	**TCIA**	**HGUGM**	**TCIA**	**HGUGM**

3.7. Data Augmentation

Data augmentation plays a pivotal role in our diagnostic aid system and aids in minimizing the impact of data‘s imbalance. The scarcity of images can lead to overfitting the training data, limiting their ability to generalize to unseen cases. Introducing variations through data augmentations can synthetically expand the datasets, lending the model a more robust and diverse training. For this purpose, the MONAI library [67] [68], designed for deep learning in healthcare imaging, was employed to facilitate augmentation techniques.

3.7.1. Spatial Transformations

Spatial transformations manipulate the position or orientation of pixels in an image. They adjust the geometry of the image without altering its content by applying variations in scale, rotation and position.

- *CropForeground:* Crops an image using a bounding box. An arbitrary function and a margin can be defined to select the expected foreground from the whole image or specified channels.
- *RandRotate90:* Randomly rotates the input data by multiples of 90 degrees. This technique helps the model become invariant to the orientation of features.

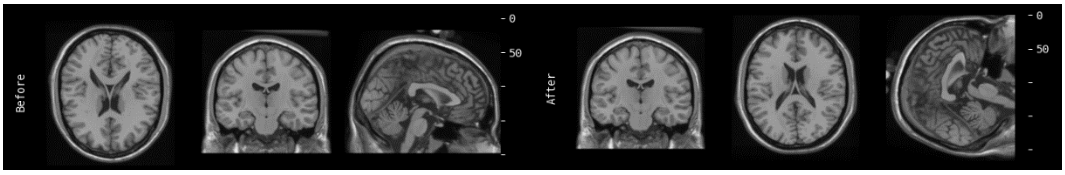


Fig. 3.7. Example of RandRotate90

- *RandZoom:* Randomly zooms into or out of the input data. By adjusting the zoom levels, this technique introduces scale variations, ensuring the model can recognize features at different scales and improves its ability to generalize across various image sizes.

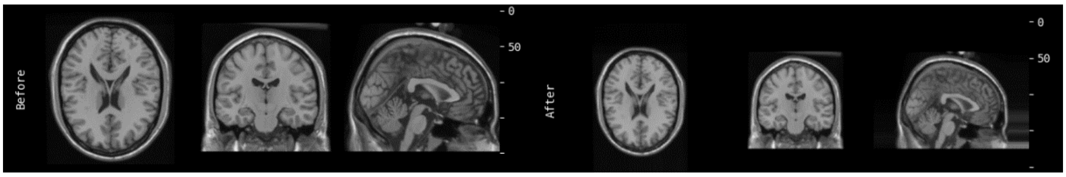


Fig. 3.8. Example of RandZoom

- *RandAffine:* Applies random affine transformations to the input data. Affine transformations maintain lines and parallelism, but can change the angles and lengths of lines. This method introduces variations in scale, rotation and translation.

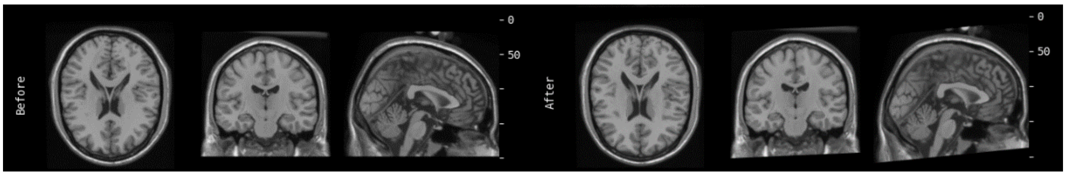


Fig. 3.9. Example of RandAffine

3.7.2. Intensity Transformations

Intensity transformations alter the pixel values of an image without changing their positions. These manipulations adjust the brightness, contrast or noise levels, enhancing image features with different lighting conditions.

- *RandScaleIntensity:* Randomly scales the intensity values of the input image. This transformation can help the model handle variations in brightness and enhance its performance under different lighting conditions.

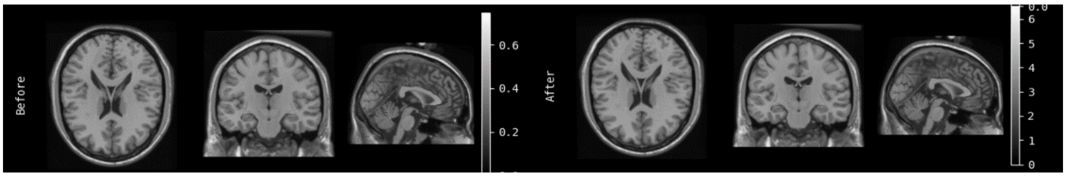


Fig. 3.10. Example of RandScaleIntensity

- *RandAdjustContrast:* Randomly adjusts the contrast of the input image. It either increases or reduces the difference in intensity between objects and their background.

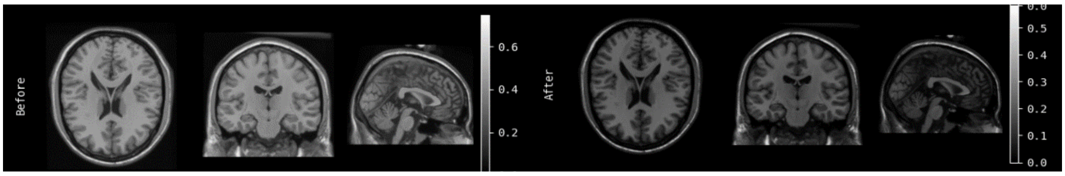


Fig. 3.11. Example of RandAdjustContrast

- *RandShiftIntensity:* Randomly shifts the intensity values of the input image. This technique can assist in adjusting the baseline intensity level, ensuring the model is not overly sensitive to specific brightness levels.

- *RandGaussianSharpen:* Applies a random Gaussian sharpening filter to the input image. This technique enhances the edges and features, making them more distinct, and trains the model to recognize structures even when they are blurred.

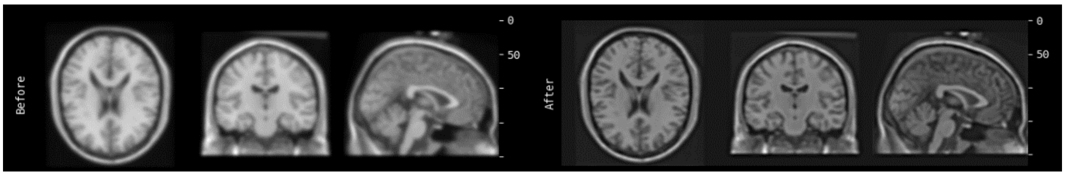


Fig. 3.12. Example of RandGaussianSharpen

- *RandGaussianNoise:* Introduces random Gaussian noise to the input image. By adding this noise, the technique ensures the model remains resilient to unexpected or random interference, optimizing its performance on potentially noisy real-world data.

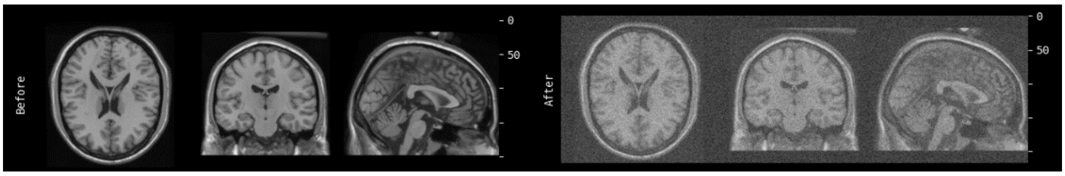


Fig. 3.13. Example of RandGaussianNoise

- *RandKSpaceSpikeNoise:* Introduces random spike noise into the k-space of image data, simulating potential artifacts that can arise during MRI acquisition. By randomly perturbing the k-space data, the resulting image can exhibit unique noise patterns.

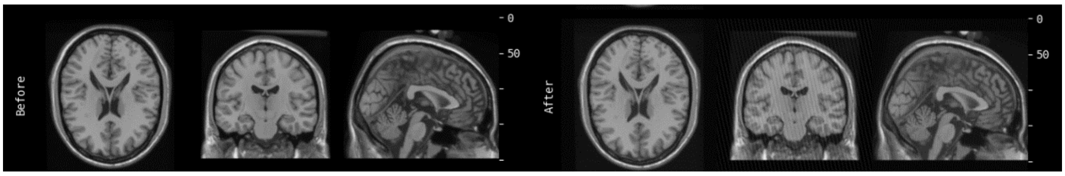


Fig. 3.14. Example of RandKSpaceSpikeNoise

4. MODELING

4.1. Deep Learning Architectures and Design

During the modeling phase, several architectures from the MONAI library have been tested and evaluated for the classification task. These architectures represent some of the most novel and cutting-edge designs in deep learning for medical imaging. All of them have been modified to incorporate a technique known as “late fusion”. This adaptation ensures that features such as age and sex are utilized during the training process. Further details will be explained in the following sections.

The following models have been utilized for classifying IDH mutations in MR images: DenseNet, HighResNet, Attention Unet, ViT Autoencoder and UNETR.

4.1.1. DenseNet

DenseNet, or Densely Connected Convolutional Networks [69] is a deep learning architecture introduced to address two main challenges: the vanishing gradient problem and the need for computational efficiency. These properties make DenseNet particularly beneficial for medical imaging tasks.

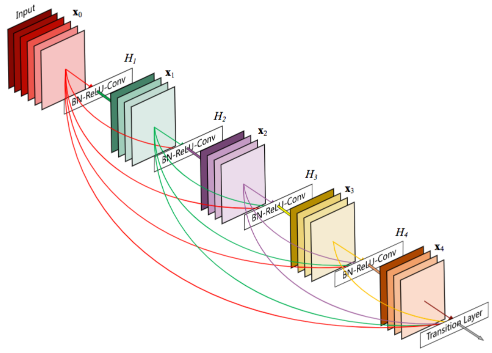


Fig. 4.1. Overview of DenseNet architecture *[69]*

The main distinction of DenseNet from traditional convolutional networks is its dense connectivity pattern. The layers are connected directly with every subsequent layer in a feed-forward manner. This means that each of them receives feature maps from all preceding layers, allowing for more efficient feature reuse and a reduced number of parameters. Mathematically, if $x_0, x_1, \dots, x_{L-1}$ are the feature maps produced in layers $0,1, \dots, L$, then the feature maps at layer L is:

$$x_L = H_L([x_0, x_1,, \dots, x_{L-1}]) \tag{4}$$

Where $[x_0, x_1, \dots, x_{L-1}]$ denotes the concatenation of feature maps and H_L is a composite function of operations like Batch Normalization (BN), Rectified Linear Activation (ReLU) and Convolution (Conv) [69]. Within each dense block, every output of the layers is concatenated with the outputs of all previous layers.

The transition layers change the number of feature maps and reduce dimensions. The growth rate (denoted as k) determines the number of filters (feature maps) added at each layer. Finally, a bottleneck layer, consisting of a 1x1 convolution, is inserted before each 3x3 convolution in dense layers. This reduces the number of input feature maps, restructuring the computational process.

4.1.2. HighResNet

HighResNet is an advanced convolutional neural network specifically created for medical imaging tasks, especially those requiring 3D images. The architecture aims to preserve high-resolution properties across the network in tasks that need precise spatial information, such as the detection of diseased areas in MRI images [70].

It stands out due to its deep lightweight structure, which is mostly composed of 3x3x3 convolutions. When the spatial resolution is decreased across the layers, typical CNNs lose critical information necessary for medical diagnosis; whereas HighResNet avoids this by maintaining the original resolution throughout the network. This feature makes this architecture perfect for catching the minuscule anatomical features in medical scans.

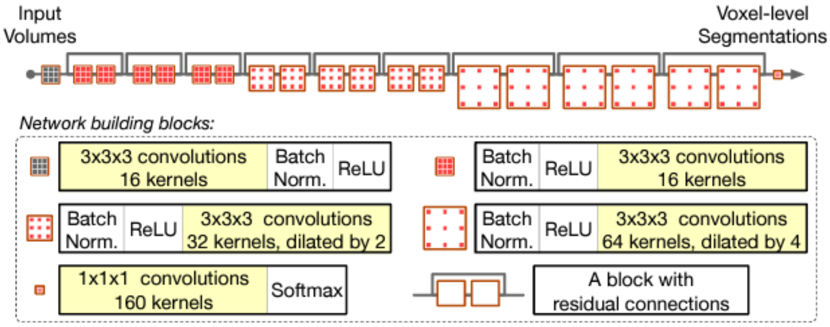


Fig. 4.2. Overview of HighResNet architecture (segmentation) *[70]*

The architecture of HighResNet is composed of a series of convolutional blocks. Each of these blocks is made up of residual units. The use of dilated convolutions instead of pooling layers or strides allows the network to preserve spatial resolution to access larger contextual information. These types of convolutions provide a dense feature extraction mechanism by enlarging the receptive field without adding extra parameters or losing resolution.

Additionally, HighResNet utilizes instance normalization technique. In medical imaging, where data often comes from various sources with different acquisition protocols, instance normalization helps in mitigating the effects of variability across scans.

The final layers of the net contain a combination of convolutional layers and fully connected layers.

4.1.3. Attention Unet

The Attention UNet architecture [71] is an adaptation of the traditional UNet model, which has been extensively utilized for various biomedical image segmentation tasks. The UNet architecture [72], provides an efficient approach for segmenting medical images using its unique symmetric expanding path, allowing precise localization. Attention UNet incorporates an attention mechanism into the U-Net structure.

Attention UNet's attention gates capture and suppress the irrelevant regions of feature maps by dynamically measuring their importance. This is achieved using the learned attention coefficients and ensures that only the significant features are passed onto the next layers. This specific design choice allows for more refined feature maps and can potentially reduce the risk of overfitting, as the model is less likely to focus on noise or other irrelevant patterns in the data [73].

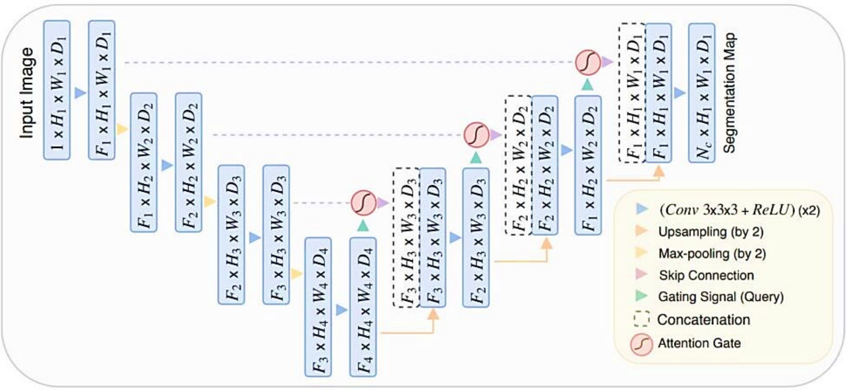

Input Image
$1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_4 \times W_4 \times D_4$
$F_4 \times H_4 \times W_4 \times D_4$
$F_3 \times H_3 \times W_3 \times D_3$
$F_3 \times H_3 \times W_3 \times D_3$
$F_2 \times H_3 \times W_3 \times D_3$
$F_2 \times H_2 \times W_2 \times D_2$
$F_2 \times H_2 \times W_2 \times D_2$
$F_1 \times H_2 \times W_2 \times D_2$
$F_1 \times H_1 \times W_1 \times D_1$
$F_1 \times H_1 \times W_1 \times D_1$
$N_c \times H_1 \times W_1 \times D_1$
Segmentation Map
$(Conv\ 3x3x3 + ReLU)$ (x2)
Upsampling (by 2)
Max-pooling (by 2)
Skip Connection
Gating Signal (Query)
Concatenation
Attention Gate

Fig. 4.3. Overview of Attention UNET architecture *[71]*

Moreover, attention gates can change how neurons are activated in both forward and backward traversal. By giving less weight to gradients that come from background regions during the backward pass, these gates give the model a higher level of specificity. This enables attention units of varying sizes to adapt their responses to a wide range of foreground features in images.

4.1.4. ViT Autoencoder

The Vision Transformer (ViT) architecture [74] has outperformed traditional convolutional neural networks (CNNs) performance in a variety of visual tasks. The key innovation behind ViT is the application of transformers to visual data.

ViT takes an image, splits it into fixed-size non-overlapping patches and linearly embeds each patch to generate a flat vector. To transfer spatial information of these patches, positional embeddings are added to the flat vectors, which encode the location of each patch within an image. This makes sure the model understands the spatial context and relationships between patches. After being fed into a chain of transformer blocks, these augmented vectors will become the focus of the attention mechanism.

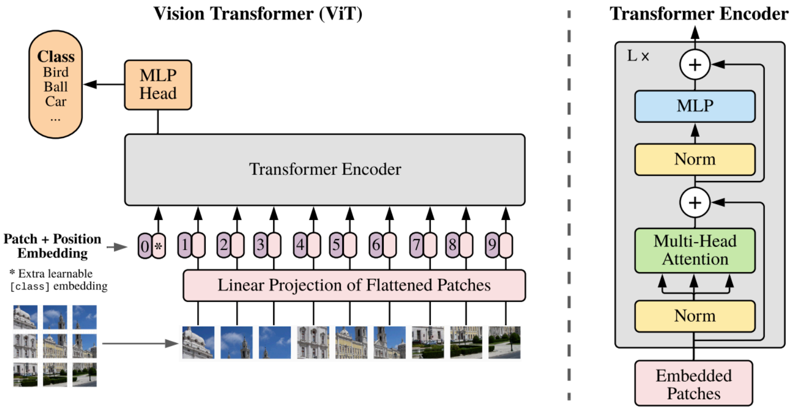


Fig. 4.4. Overview of Vision Transformer architecture *[74]*

When applied to a ViT autoencoder, the architecture is adjusted to work in two main phases: encoding and decoding. The encoding phase is similar to the initial steps of ViT, however, instead of leading to a classification token, the encoded representation tries to capture the information of the input image in a compressed form. This compressed representation is the latent space of the autoencoder.

During decoding, the original image is reconstructed from the hidden information. This is achieved by computing a series of inverse operations on the latent vectors, which are like the encoding steps but in reverse. Finally, the image is reconstructed by piecing together the decoded patches.

4.1.5. UNETR

UNETR represents an evolution in medical imaging models by combining the strengths of transformer and UNet architectures [75]. It was originally proposed for biomedical image segmentation.

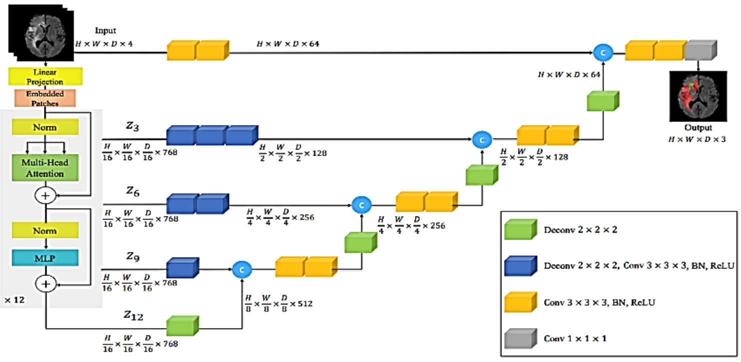


Fig. 4.5. Overview of the UNETR architecture (segmentation) *[75]*

In addition to the conventional UNet characteristics, which have an encoder-decoder structure, UNETR makes use of the benefits of transformers to capture long-range contextual information more effectively. Nevertheless, the primary difference lies in the replacement of the conventional convolutional blocks with transformer blocks. These transformer blocks operate on fixed-size non-overlapping patches of the input images, just as in the Vision Transformer (ViT) approach [74].

In order to provide a sequence input to the transformers, image patches are linearly embedded into flat vectors. Encoders are constructed from multiple transformer layers. Each layer is made up of feedforward neural networks and multi-head self-attention mechanisms. On the other hand, the decoder uses transformer layers and up-sampling layers to guarantee that the resolution of the segmented output is identical to that of the original image.

Another distinguishing feature of UNETR is the integration of positional encoding, which adds information about the relative or absolute position of the patches in the sequence.

4.2. Training and Validation Process

During the development phase of deep learning models, special attention was put into the data input structure due to our decision of using both T1CE and FLAIR images. Each of these images has dimensions of 128x128x64.

To efficiently feed them into the models, a strategy of concatenation along a new axis was employed. The two types of images were stacked as a result, giving the new input image a dimension of 2x128x128x64. This allowed the models to process the combined data from both MRI sequences at once.

4.2.1. Late Fusion

Late fusion is a technique in deep learning where information from different sources is combined in the later stages of a model. Instead of merging all data at the start, each type of data, like demographic details, is processed separately. These are then combined, or “fused” in the later layers of the neural network.

In our work, late fusion was used to add more depth to the models. The feature vector

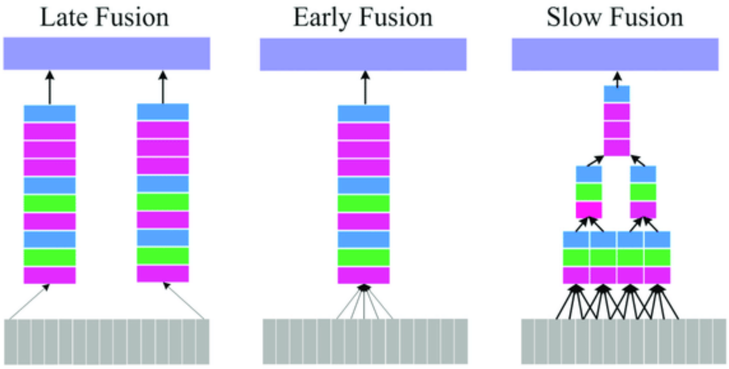


Fig. 4.6. Comparison between Late Fusion, Early Fusion and Slow Fusion *[81]*

includes a patient's age, sex and tumor grade. These factors are vital in understanding gliomas and can provide added context to the MRI data.

To achieve this, all of the previously mentioned models have been modified and their final layers have been adapted to include the new feature vector. This change allows the models to consider both the MRI data and the patient's background information before making a decision on the IDH status.

4.2.2. Loss Functions

The choice of a loss function plays a pivotal role. Given the binary nature of our problem - categorizing IDH status as positive or negative – Weighted Binary Cross-Entropy (WBCE) has been chosen as loss function. Mathematically, it can be represented as:

$$WBCE(y, \hat{y}) = -w_1 \cdot y \cdot \log(\hat{y}) - w_2 \cdot (1 - y) \cdot \log(1 - \hat{y}) \tag{5}$$

Where:

- y is the true label (0 or 1).
- $\hat{y}$ is the predicted probability of the label being 1.
- w_1 and w_2 are weights for positive and negative classes, respectively.

The WBCE loss quantifies the difference between the predicted probability and the actual label and, at the same time, attributes different weights to positive and negative instances to minimize class imbalance. This ensures that the model does not overly in favor the majority class.

4.2.3. Optimization Algorithms

Adam optimizer was utilized in the framework, in combination with the ReduceLROnPlateau learning rate scheduler to fine-tune the model.

Adam (Adaptive Moment Estimation) optimizer [76] brings together the benefits of two popular optimization methods - AdaGrad and RMSProp. It is preferred in many applications due to its efficient computation and memory requirements. Adam fine-tunes the model's weights by computing adaptive learning rates for each parameter, which helps in reaching convergence faster and more stably.

However, there may be times when the model reaches a plateau and learning ceases during the training phase. ReduceLROnPlateau scheduler was implemented to address this issue. When a model's validation performance reaches a plateau, this scheduler dynamically modifies the learning rate. By doing this, the model is able to avoid any potential inactivity stages and achieve a more precise optimization, increasing the accuracy of the diagnosis.

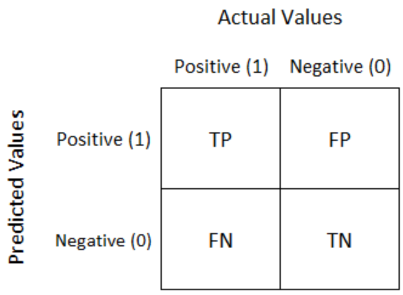
Together, the Adam optimizer and the ReduceLROnPlateau scheduler ensure efficient convergence and robust diagnostic performance in the models.

5. EVALUATION AND RESULTS

5.1. Performance Metrics

5.1.1. Confusion Matrix

One of the most useful and well-known metrics is the confusion matrix. This matrix displays the number of samples that have been correctly classified in contrast to those misclassified for each category. The scheme corresponds to the following figure:



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 5.1. Confusion Matrix

In the case of this problem, the corresponding parameters [55] would be as follow:

1. TP (True Positives): Samples of adult diffuse gliomas with a wildtype IDH status that our model correctly classifies.

2. FP (False Positives): Samples of adult diffuse gliomas with a mutated IDH status that our model incorrectly classifies as having a wildtype IDH status.

3. FN (False Negatives): Samples of adult diffuse gliomas with a wildtype IDH status that our model incorrectly classifies as having a mutated IDH status.

4. TN (True Negatives): Samples of adult diffuse gliomas with a mutated IDH status that our model correctly classifies.

From the confusion matrix, other metrics can be derived that are useful when assessing the effectiveness of the models. Among them, the following stand out:

- **Accuracy:** The ratio of correctly predicted instances to the total number of instances. It gives an overall effectiveness of the model.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} = \frac{No. of\ Correct\ Predictions}{Total\ No. of\ Predictions} \tag{6}$$

- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives. It indicates the reliability of the model's positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

- **TPR (Recall / Sensitivity):** The ratio of correctly predicted positive observations to all the actual positives (true positive rate). It demonstrates the model's capability to correctly identify all relevant instances.

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

- **FPR:** The ratio of negative instances that are incorrectly classified as positive (false positive rate).

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

- **F1 Score:** The weighted average of Precision and Recall. It provides a balance between the two, especially useful when the class distribution is uneven.

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

5.1.2. AUC Score

The AUC Score or, Area Under the Receiver Operating Characteristic Curve (ROC-AUC), is a metric used to evaluate the performance of binary classification models. It quantifies the model’s ability to distinguish between the positive and negative cases.

The ROC curve is a graphical representation of the true positive rate (sensitivity) against the false positive rate (1-specificity) for different threshold values

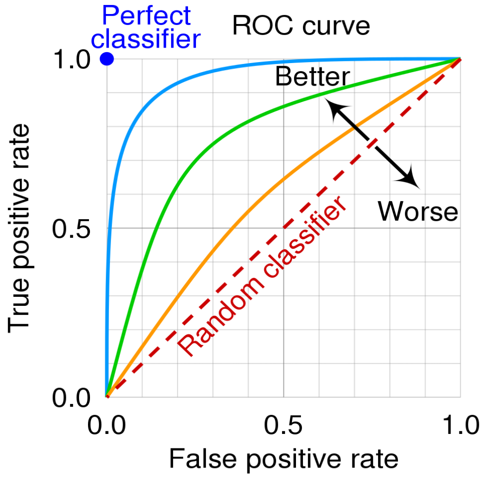


Fig. 5.2. Example of ROC Curve *[82]*

The AUC score calculates the area under this ROC curve. An AUC score of 1.0 indicates that the model perfectly distinguishes between the two classes, while an AUC of 0.5 suggests that the model performs no better than random guessing.

5.2. Hyperparameter Tuning

Given the complexity of deep learning models and the computational resources they demand, exhaustive hyperparameter tuning for every parameter was infeasible, especially when considering the 3D nature of our image inputs. It is necessary to understand that the parameters were selected to represent an equilibrium between the best model performance and the practical computational cost.

- **Batch Size:** Influences both training speed and model's performance, the batch size was limited by the available GPU memory. An initial batch size of 64 was utilized as default value for the models, nevertheless, there were models where this setting led to memory overflow. In such scenarios, the batch size was scaled down.

- **Learning Rate:** The learning rate determines the step size at each iteration while moving towards a minimum of the loss function. Given its importance in model convergence, three values for hyperparameter tuning were tuned (0.001, 0.0001, and 0.00001).

- **Epochs:** By default, all models were set to train for 100 epochs. However, to avoid overfitting an early stopping mechanism was integrated during training. This mechanism monitors the validation loss and stops the training process if no improvement is observed after a predefined number of epochs.

The following table shows the best hyperparameters for each of the evaluated models:

TABLE 3.
HYPERPARAMETERS

Model	Batch Size	Learning Rate	Number of Epochs
DenseNet	32	0.0001	100
HighResNet	2	0.001	31
Attention Unet	8	0.001	100
Vit Autoencoder	64	0.0001	53
UNETR	8	0.0001	80

5.3. Comparative Analysis

This section analyzes and compares the performance metrics of the models: DenseNet, HighResNet, Attention Unet, ViT Autoencoder and UNETR. Each model has been evaluated to determine its performance in identifying the IDH status in Magnetic Resonances.

The training loss serves as an indicator for the model's learning capability, with a lower loss indicating better performance. The following graph represents the error rates of each model during its training phase.

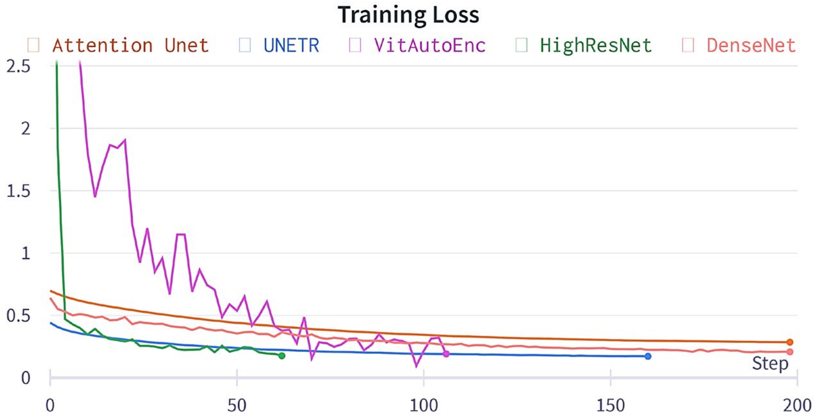


Fig. 5.3. Training loss comparison

The validation loss, evaluates the models' ability of learning during validation. The graph captures the evolution of this loss:

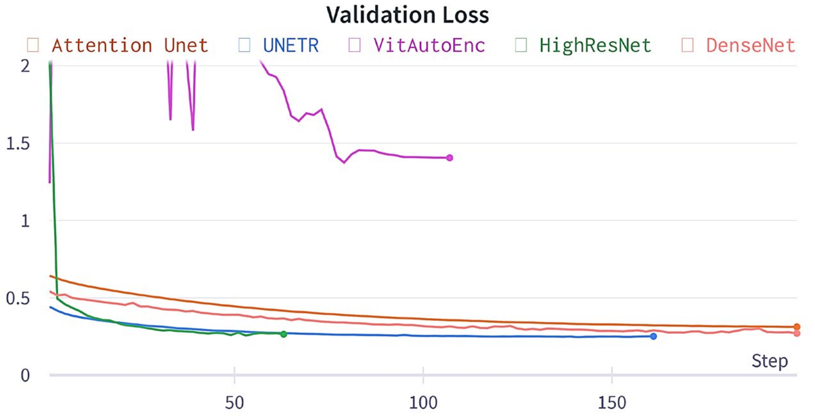


Fig. 5.4. Validation loss comparison

The validation accuracy graph shows the overall accuracy rates of the models throughout their training. It stands as a verification to the models' ability in correctly classifying the IDH status.

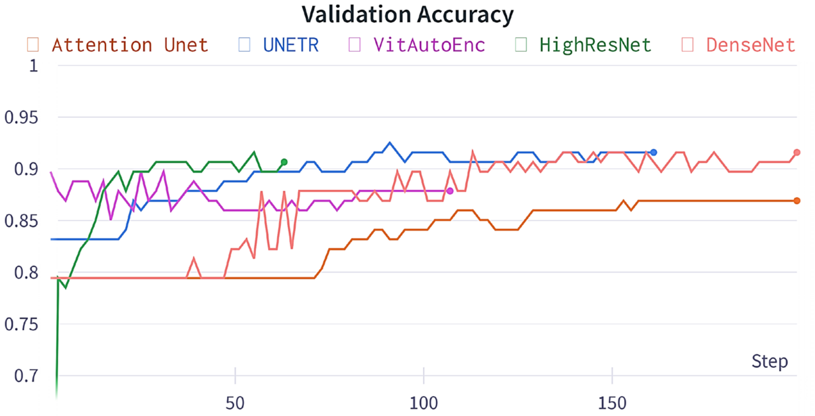


Fig. 5.5. Validation accuracy comparison

It is important to present the confusion matrices before analyzing the specific performance of each model. These matrices compare the actual classifications with the predicted ones and offer an overall view of the model's performance. By examining these matrices, new insights can be gained to determine where the model might require further refinement. The confusion matrices for DenseNet, HighResNet, Attention Unet, ViT Autoencoder and UNETR are displayed below.

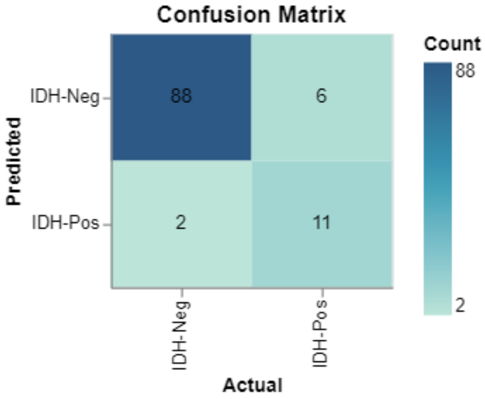


Fig. 5.6. Confusion Matrix of DenseNet

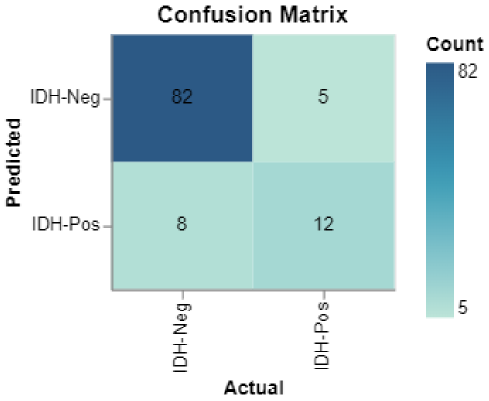


Fig. 5.7. Confusion Matrix of HighResNet

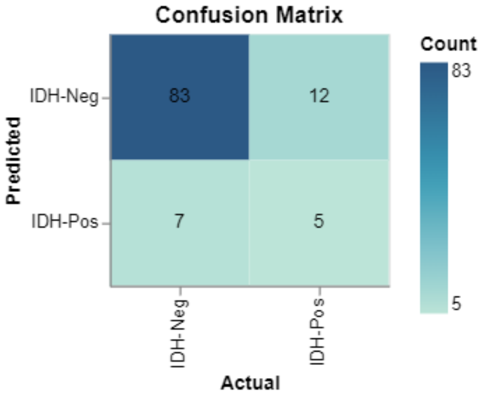


Fig. 5.8. Confusion Matrix of Attention Unet

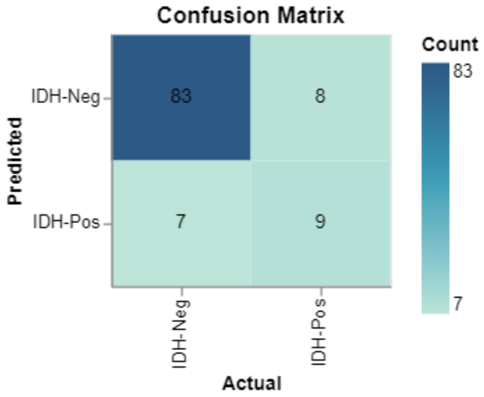


Fig. 5.9. Confusion Matrix of Vit Autoencoder

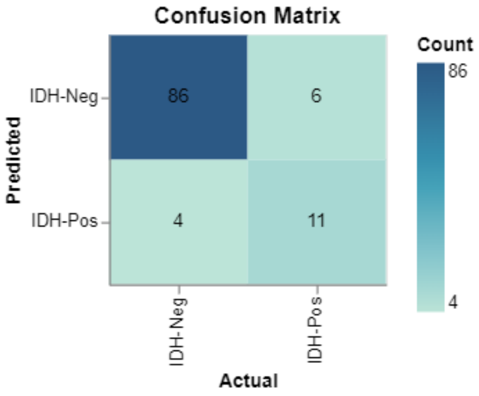


Fig. 5.10. Confusion Matrix of UNETR

The following table includes the performance measures of the analyzed models during the testing phase in order to offer a systematic comparison. It compares their performance in terms of accuracy, precision, recall, F1 score and AUC. Such insights are fundamental in guiding the selection of the most appropriate model:

TABLE 4.
MODEL PERFORMANCE METRICS

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**AUC (%)**
DenseNet	92.52	84.62	64.71	73.33	91.18
HighResNet	87.85	60.00	70.59	64.86	91.18
Attention Unet	82.24	41.67	29.41	34.48	82.22
Vit Autoencoder	85.98	56.25	52.94	54.55	86.11
UNETR	90.65	73.33	64.71	68.75	91.24

- **DenseNet:** DenseNet stands out as the top-performing model, with an accuracy of 92.52%. Although its recall of 64.71% indicates some room for improvement in recognizing true positives, its precision of 84.62% is remarkable. Its robustness is further supported by the F1 Score and AUC, which makes it an appropriate option for this classification task.

- **HighResNet:** The accuracy rate for HighResNet is 87.85%. While its recall is considerably higher at 70.59% compared to its precision of 60.00%, this suggests that it can accurately identify a significant fraction of positive cases. Its AUC is the same as DenseNet's, indicating that it is effective in differentiating between the classes.

- **Attention Unet:** The least accurate of the models tested, this model has an accuracy of 82.24%. With a precision and recall of 41.67% and 29.41%, respectively, it is clear that there are difficulties in correctly identifying and predicting positive cases. The AUC of 82.22% suggests that there is potential for enhancement.

- **Vit Autoencoder:** With an accuracy of 85.98%, the Vit Autoencoder performs reasonably well. Its precision and recall, demonstrate a balanced but moderate performance in terms of predicting and identifying positive cases. An AUC of 86.11% indicates a fair distinction between classes.

- **UNETR:** UNETR presents a significant accuracy of 90.65%, making it one of the top contenders alongside DenseNet. Its precision of 73.33% and recall of 64.71% demonstrate a balanced performance and its AUC of 91.24% confirms its capability to effectively differentiate between the classes.

5.4. Computational Cost and Efficiency

The following table provides an overview of the computational footprint of each model, detailing their respective sizes, GPU utilization rates, memory allocation percentages and number of images per second. This information helps us to understand the computational efficiency and requirements of each model, resulting in better decisions about which architecture is best suited for a given deployment scenario:

TABLE 5.
MODEL RESOURCE UTILIZATION

Model	**Size (MB)**	**GPU Power Usage (~ %)**	**GPU Memory Allocation (~ %)**	**No. Images per Second**
DenseNet	43.6	80	65	0.549
HighResNet	3.1	100	90	0.059
Attention Unet	1.4	90	85	1.481
Vit Autoencoder	359.2	60	100	2.447
UNETR	366.1	95	100	0.825

- **DenseNet**: With a size of 43.6 MB, DenseNet allocates about 65% of memory and uses 80% of GPU's power. However, its processing speed is moderate, handling only about 0.549 images per second.

- **HighResNet**: HighResNet is a lightweight model (3.1 MB). It demands a high memory allocation of 90% and makes full use of the GPU at 100%. Despite these resource demands, its processing rate is the slowest, analyzing 0.059 images per second.

- **Attention Unet**: This model stands out with its minimal size of 1.4 MB and a notable processing speed of 1.481 images per second. Attention Unet offers a good balance between resource usage and performance with 90% GPU utilization and 85% memory allocation.

- **Vit Autoencoder**: The Vit Autoencoder is one of the largest models (359.2 MB) and requires full memory allocation. It utilizes 60% of the GPU and manages the highest processing speed of 2.447 images per second.

- **UNETR**: UNETR, the largest model at 366.1 MB, fully utilizes memory and operates at a high GPU utilization rate of 95%. Its processing speed is moderate, handling 0.825 images per second.

The detailed graphics for Memory Allocation and GPU Utilization can be located in the appendix (A.2).

5.5. Understanding Misclassifications

By analyzing the misclassifications of DenseNet, the model with the highest performance, insights can be provided into potential areas of improvement and the challenges faced by the model. The confusion matrix for this model revealed a total of two false negatives (FN) and six false positives (FP). Two examples of each category of incorrect labeling are shown below. For each example, the left image represents the T1CE modality, while the right image is the FLAIR modality.

False negatives samples are MR that have been classified as having a mutated IDH status (positive) but in reality have wildtype IDH status (negative).

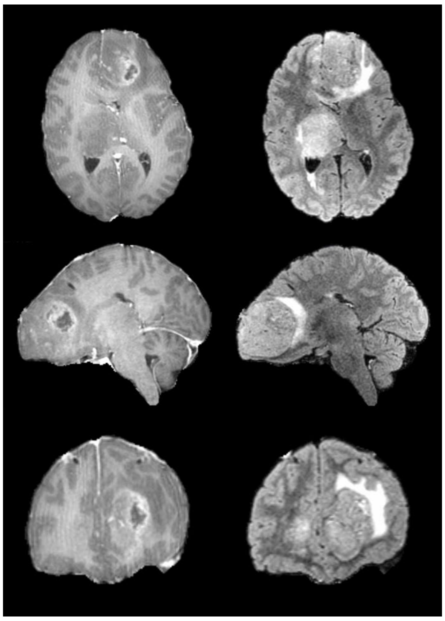


Fig. 5.12. MRI of FN sample (471)

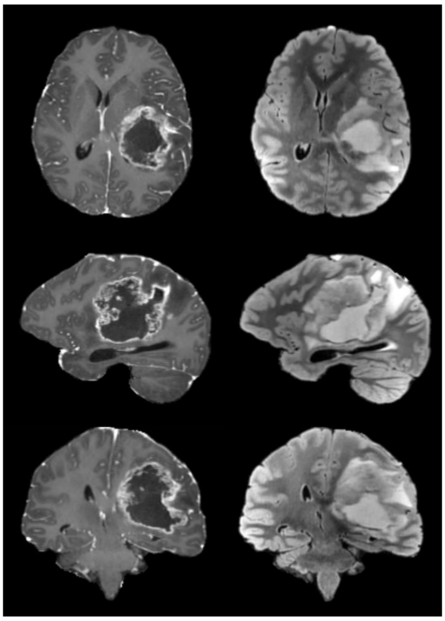


Fig. 5.12. MRI of FN sample (487)

- Sample 471 is a Grade 4 Glioblastoma from a 26-year old female from the TCIA Dataset (Fig. 5.11). Upon close observation, one can discern the presence of k-space spikes, which are anomalies in the frequency domain of MRI data, potentially affecting the model's prediction.
- Sample 487 is a Grade 4 Glioblastoma from a 21-year-old male from the TCIA Dataset (Fig. 5.12).

A commonality between these two samples is that both are glioblastomas derived from younger patients and are notably large in size. This shared characteristic might provide insight into certain challenges the model faces when classifying glioblastomas in younger demographics.

The false positives samples are MR that have been classified as having a wildtype IDH status (negative) but in reality have mutated IDH status (positive).

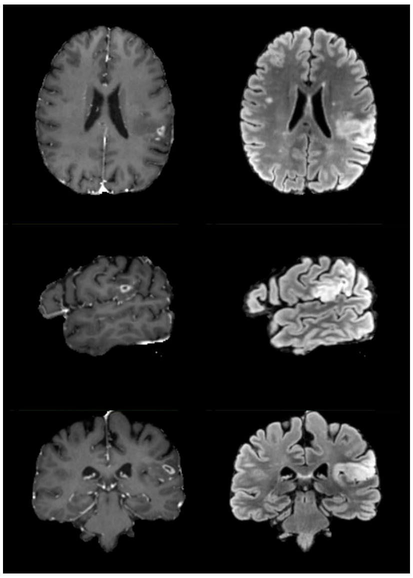


Fig. 5.14. MRI of FP sample (304)

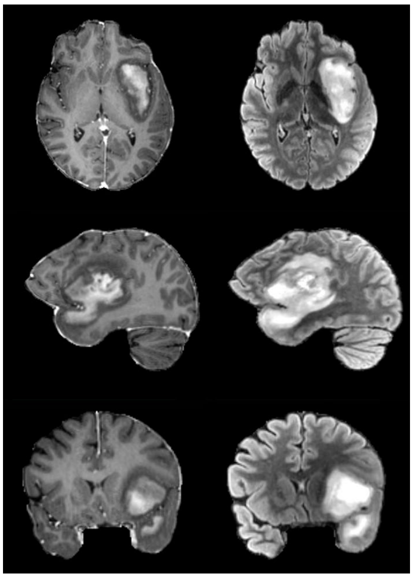


Fig. 5.14. MRI of FP sample (517)

- Sample 073 is a Grade 4 astrocytoma from a 54-year old female from the TCIA Dataset.
- Sample 147 is a Grade 4 astrocytoma from a 58-year old male from the TCIA Dataset.
- Sample 304 is a Grade 3 astrocytoma from a 52-year old male from the TCIA Dataset (Fig. 5.13).
- Sample 407 is a Grade 4 astrocytoma from a 42-year old male from the TCIA Dataset.
- Sample 485 is a Grade 3 astrocytoma from a 65-year old male from the TCIA Dataset.
- Sample 517 is a Grade 4 astrocytoma from a 53-year-old male from the TCIA Dataset (Fig. 5.14).

All of the false positive samples are astrocytomas from patients of a similar age bracket. Additionally, a shared characteristic of most of these samples is their location in the left hemisphere of the brain.

6. CONCLUSIONS AND FUTURE WORK

6.1. Summary of Findings

For the purpose of developing a deep learning-based diagnostic aid system for diffuse gliomas, several models were evaluated based on their performance metrics and computing requirements. The results of this study demonstrate the benefits and limitations of each model when determining the IDH status of gliomas in MRI images.

DenseNet, a convolutional neural network (CNN) architecture, emerged as the most proficient, registering an accuracy of 92.52%. UNETR, a fusion of the traditional U-Net and transformer mechanisms, followed closely with an accuracy of 90.65%. Both models demonstrated remarkable precision, recall, F1 score and AUC values, demonstrating their potential applicability in clinical scenarios.

It is noteworthy to mention the performance of transformer-based models: UNETR, Vit Autoencoder and Attention Unet. Despite their innovative architectures, they did not outperform the traditional CNN model, DenseNet. This is in line with the common consensus that large amounts of data are often necessary for transformer-based models to achieve optimal generalization. Considering the limitations that are typically associated with medical imaging datasets, this is an important challenge for these designs to overcome.

In conclusion, deep learning has the potential to substantially improve presurgical medical diagnoses. Nonetheless, the option of architecture is still critical. This study shows that traditional CNN designs remain robust, even in data-limited circumstances, while transformer-based models can encounter difficulties.

6.2. Challenges and Limitations

The research encountered several challenges and limitations that are worth noting:

- **Data Imbalance and Size:** The HGUGM database, being relatively smaller compared to the TCIA database, presented an imbalance. When combined, the total number of patients from both databases amounted to 494. This number, although substantial, is not particularly large for deep learning models, which typically require vast amounts of data to generalize effectively.

- **Scanner Intensity Variations:** The data sourced from HGUGM came from a 1.5 scanner, whereas the TCIA database utilized a 3.0 scanner. This discrepancy led to variations in intensities, which could potentially affect the model's performance.

- **Pre-processing Concerns:**

  - **Interpolation Methods:** The research employed Windowed Sinc

  - interpolation. Nevertheless, the impact of this specific method in comparison to others, such as B-spline, remains uncertain.

  - **Intensity Matching and Normalization:** The chosen method for intensity matching was KMeans clustering and the normalization method used was WhiteStripe. The direct impact of these choices, in relation to other techniques has not been studied.

- **Modeling Limitations:**

  - **Image Sequences:** The research utilized T1-CE and FLAIR images. The inclusion or exclusion of other sequences might yield different results.

  - **Image Processing:** T1-CE and FLAIR images were combined into two channels and processed as a single tensor. Exploring alternative processing methods might offer different insights.

  - **Feature Impact:** Features such as age, sex and tumor grade were incorporated into the model using late fusion. Their direct impact on the model's performance as well as the utilized fusion strategy have not been compared to other alternatives.

  - **Hyperparameter Tuning:** Due to time constraints and computational limitations, an exhaustive hyperparameter search was not conducted.

  - **Validation Techniques:** Owing to time restrictions and the complexity of the task, K-fold cross-validation was not implemented, limiting the exploration of different train, test and validation set configurations.

  - **Attention Models:** Transformer-based models, despite their recent success in various domains, did not outperform traditional CNNs in this research. This could be attributed to the sample size.

6.3. Potential Enhancements to the System

To address the aforementioned challenges and limitations, the following enhancements could be considered:

- **Data Augmentation:** To mitigate the issue of data imbalance and limited sample size, more data augmentation techniques could be employed to artificially increase the dataset's size and diversity.

- **Scanner Calibration:** Implementing calibration techniques might help in harmonizing the intensity variations between different scanners.

- **Exploratory Pre-processing:** Conducting a systematic study on the effects of different interpolation, intensity matching and normalization methods could provide clearer insights into their impacts.

- **Modeling Enhancements:**
  - **Sequence Exploration:** Investigating the inclusion of additional or alternative MRI sequences might enhance the model's performance.
  - **Feature Analysis:** To better understand the value of each feature, an in-depth feature importance analysis could be performed.
  - **Hyperparameter Optimization:** Given more time and computational resources, a more exhaustive hyperparameter search could be beneficial.
  - **Cross-Validation:** Implementing K-fold cross-validation could provide a more robust evaluation of the model's performance.

6.4. Recommendations for Future Research

Future research investigations could focus on:

- **Larger Datasets:** Acquiring and integrating more data from diverse sources could help in building more robust models.
- **Advanced Pre-processing Techniques:** Exploring other pre-processing methods or pipelines might lead to better data quality and, consequently, improved model performance.
- **Feature Engineering:** More research into feature engineering and selection may provide details on which features are most important and why.
- **Model Architectures:** Investigating newer or alternative deep learning architectures could be beneficial.

7. BIBLIOGRAPHY

[1] Y. Rajesh, I. Pal, P. Banik, S. Chakraborty, S. A. Borkar, G. Dey, A. Mukherjee y M. Mandal, «Insights into molecular therapy of glioma: current challenges and next generation blueprint,» *Acta Pharmacologica Sinica,* vol. 38, pp. 591-613, 2017.

[2] S. Gore, T. Chougule, J. Jagtap, J. Saini y M. Ingalhalikar, «A review of radiomics and deep predictive modeling in glioma characterization,» *Academic Radiology,* vol. 28, nº 11, pp. 1599-1621, 2021.

[3] A. Bhandari, R. Liong, J. Koppen, S. Murty y A. Lasocki, «Noninvasive Determination of IDH and 1p19q Status of Lower-grade Gliomas Using MRI Radiomics: A Systematic Review,» *American Journal of Neuroradiology,* vol. 42, nº 1, pp. 94-101, Jan 2021.

[4] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffietti, A. von Deimling and D. W. Ellison, "The 2021 WHO Classification of Tumors of the Central Nervous System: a summary," *Neuro-oncology,* vol. 23, no. 8, pp. 1231-1251, 2021.

[5] «Constitución Española,» Boletín Oficial del Estado, 1978.

[6] «Agencia Española de Medicamentos y Productos Sanitarios,» [En línea]. Available: https://www.aemps.gob.es/.

[7] «Comité de Ética de la Investigación con Medicamentos (CEIm),» [En línea]. Available: https://www.iisgm.com/organizacion/comisiones/comite-de-etica-de-la-investigacion-con-medicamentos-ceim/.

[8] «European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of,» Official Journal of the European Union, L 119, 2016, pp. 1-88.

[9] Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, Boletín Oficial del Estado, 2018.

[10] «The Institute for Ethical AI Machine Learning,» [En línea]. Available: https://ethical.institute/principles.html.

[11] D. Louis, A. Perry y G. Reifenberger, «The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary,» *Acta Neuropathol ,* vol. 131, pp. 803-820, 2016.

[12] A. Cohen, S. Holmen y H. Colman, «IDH1 and IDH2 mutations in gliomas.,» *Curr Neurol Neurosci Rep.,* vol. 13, nº 5, p. 345, 2013.

[13] A. Olar, K. Wani y K. Alfaro-Muñoz, «IDH mutation status and role of WHO grade and mitotic index in overall survival in grade II-III diffuse gliomas,» *Acta Neuropathol,* vol. 129, nº 4, pp. 585-596, 2015.

[14] Q. SongTao, Y. Lei y G. Si, «IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma,» *Cancer Sci,* vol. 103, nº 2, pp. 269-273, 2012.

[15] M. Waitkus, B. Diplas y H. Yan, «Isocitrate dehydrogenase mutations in gliomas,» *Neuro Oncol,* vol. 18, nº 1, pp. 16-26, 2015.

[16] N. Hu, R. Richards y R. Jensen, «Role of chromosomal 1p/19q co-deletion on the prognosis of oligodendrogliomas: A systematic review and meta-analysis,» *Interdisciplinary Neurosurgery,* vol. 5, pp. 58-63, 2016.

[17] S. Boots-Sprenger, A. Sijben y J. Rijntjes, «Significance of complete 1p/19q co-deletion, IDH1 mutation and MGMT promoter methylation in gliomas: use with caution,» *Modern Pathology,* vol. 26, nº 7, pp. 922-929, 2013.

[18] A. Idbaih, Y. Marie y G. Pierron, «Two types of chromosome 1p losses with opposite significance in gliomas,» *Annals of neurology,* vol. 58, nº 3, pp. 483-487, 2005.

[19] S. Beck, X. Jin y Y. Sohn, «Telomerase activity-independent function of TERT allows glioma cells to attain cancer stem cell characteristics by inducing EGFR expression,» *Mol Cells,* vol. 31, nº 1, pp. 9-15, 2011.

[20] J. Eckel-Passow, D. Lachange y A. Molinaro, «Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors,» *N Engl J Med,* vol. 372, nº 26, pp. 2499-2508, 2015.

[21] A. von Deimling, A. Korshunov y C. Hartmann, «The next generation of glioma biomarkers: MGMT methylation, BRAF fusions and IDH1 mutations,» *Brain Pathol,* vol. 21, nº 1, pp. 74-87, 2011.

[22] R. Verhaak, K. Hoadley y E. Purdom, «Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1,» *Cancer cell,* vol. 17, nº 1, pp. 98-110, 2010.

[23] K. Kannan, A. Inagaki y J. Silber, «Whole-exome sequencing identifies ATRX mutation as a key molecular determinant in lower-grade glioma,» *Oncotarget,* vol. 3, nº 10, pp. 1194-1203, 2012.

[24] H. Zheng, H. Ying y H. Yan, «p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation,» *Nature,* vol. 455, nº 7216, pp. 1129-1133, 2008.

[25] R. Wason, «Deep learning: Evolution and expansion,» *Cognitive Systems Research,* vol. 52, pp. 701-708, 2018.

[26] J. Fenton, S. Taplin y P. Carney, «Influence of computer-aided detection on performance of screening mammography,» *N Engl J Med,* vol. 356, pp. 1339-1409, 2007.

[27] C. Lehman, R. Wellman y D. Buist, «Diagnostic accuracy of digital screening mammography with and without computer-aided detection,» *JAMA Intern Med,* vol. 175, pp. 1828-1837, 2015.

[28] K. Mingyu, Y. Jihye, C. Yongwon, S. Keewon, J. Ryoungwoo, B. Hyun-jin y K. Namkug, «Deep Learning in Medical Imaging,» *Neurospine,* vol. 16, nº 4, pp. 657-668, 2019.

[29] A. Anaya-Isaza, L. Mera-Jiménez y M. Zequera-Diaz, «An overview of deep learning in medical imaging,» *Informatics in Medicine Unlocked,* vol. 26, p. 100723, 2021.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser y I. Polosukhin, «Attention Is All You Need,» de *31st Conferenceon Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[31] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper y H. J. Aerts, «Computational radiomics system to decode the radiographic phenotype,» *Cancer Research,* vol. 77, nº 21, pp. e104-e107, 2017.

[32] O. Gevaert , L. Mitchell y A. Achrol, «Glioblastoma Multiforme: Exploratory Radiogenomic Analysis by Using Quantitative Image Features,» *Radiology,* vol. 273, nº 1, pp. 168-174, 2014.

[33] E. Rios Velazquez, R. Meier y W. Dunn Jr, «Fully automatic GBM segmentation in the TCGA-GBM dataset: Prognosis and correlation with VASARI features,» *Sci Rep,* vol. 5, nº 1, p. 16822, 2015.

[34] D. Gutman, W. Dunn y P. Grossmann, «Somatic mutations associated with MRI-derived volumetric features in glioblastoma,» *Neuroradiology,* vol. 57, nº 12, pp. 1227-1237, 2015.

[35] K. Chang, H. Bai y H. Zhou, «Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas from MR Imaging,» *ClinCancerRes,* vol. 24, nº 5, p. 1073–1081, 2018.

[36] S. Liang, «Multimodal 3D DenseNet for IDH Genotype Prediction in Gliomas,» *Genes,* vol. 9, nº 8, p. 382, 2018.

[37] S. Nalawade, G. Murugesan y M. Vejdani-Jahromi, «Classification of brain tumor isocitrate dehydrogenase status using MRI and deep learning,» *JournalofMedicalImaging,* vol. 6, nº 4, p. 046003, 2019.

[38] E. Calabrese, J. Villanueva-Meyer, J. Rudie, A. Rauschecker, U. Baid, S. Bakas, S. Cha, J. Mongan y C. Hess, «The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM) (Version 4),» The Cancer Imaging Archive, 2022.

[39] J. Jagtap , J. Saini y V. Santosh, «Proceedings of the 2nd International Conference on Data Engineering and Communication Technology,» de *Springer*, 2019.

[40] H. Zhou , K. Chang y H. Bai, «Machine learning reveals multimodal MRI patterns predictive of isocitrate dehydrogenase and 1p/19q status in diffuse low-and high-grade gliomas,» *J Neurooncol,* vol. 142, nº 2, pp. 299-307, 2019.

[41] C. Lu, F. Hsu y K. Hsieh, «Machine Learning-Based Radiomics for Molecular Subtyping of Gliomas,» *Clin Cancer Res,* vol. 24, nº 18, pp. 4429-4436, 2018.

[42] Z. Li, H. Bai y Q. Sun, «Multiregional radiomics profiling from multiparametric MRI: Identifying an imaging predictor of IDH1 mutation status in glioblastoma,» *Cancer Med,* vol. 7, nº 12, pp. 5999-6009, 2018.

[43] S. van der Voort, F. Incekara y M. Wijnenga, «Predicting the 1p/19q Codeletion Status of Presumed Low-Grade Glioma with an Externally Validated Machine Learning Algorithm,» *Clin Cancer Res,* vol. 25, nº 24, pp. 7455-7462, 2019.

[44] B. Zhang, K. Chang y S. Ramkissoon, «Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas,» *Neuro Oncol,* vol. 19, nº 1, pp. 109-117, 2017.

[45] J. Yu, Z. Shi y Y. Lian, «Noninvasive IDH1 mutation estimation based on a quantitative radiomics approach for grade II glioma,» *Eur Radiol,* vol. 27, nº 8, pp. 3509-3522, 2017.

[46] K. Hsieh, C. Chen y C. Lo, «Radiomic model for predicting mutations in the isocitrate dehydrogenase gene in glioblastomas,» *Oncotarget,* vol. 8, nº 28, pp. 45888-45897, 2017.

[47] X. Zhang, Q. Tian y L. Wang, «Radiomics Strategy for Molecular Subtype Stratification of Lower-Grade Glioma: Detecting IDH and TP53 Mutations Based on Multimodal MRI,» *J Magn Reson Imaging,* vol. 48, nº 4, pp. 916-926, 2018.

[48] B. Joo, K. Han y S. Ahn, «Amide proton transfer imaging might predict survival and IDH mutation status in high-grade glioma,» *Eur Radiol,* vol. 29, nº 1, pp. 1-10, 2019.

[49] S. Jiang, T. Zou y C. Eberhart, «Predicting IDH mutation status in grade II gliomas using amide proton transfer-weighted (APTw) MRI,» *Magn Reson Med,* vol. 78, nº 3, pp. 1100-1109, 2017.

[50] A. Jakola , Y. Zhang y A. Skjulsvik , «Quantitative texture analysis in the prediction of IDH status in low-grade gliomas,» *Clin Neurol Neurosurg,* vol. 164, pp. 114-120, 2018.

[51] J. Villanueva-Meyer, M. Wood, B. Choi, M. Mabray, N. Butowski, T. Tihan y S. Cha, «MRI Features and IDH Mutational Status of Grade II Diffuse Gliomas: Impact on Diagnosis and Prognosis,» *AJR Am J Roentgenol,* vol. 210, nº 3, pp. 621-628, 2018.

[52] Y. Kang, S. Choi, Y. Kim, K. Kim, C. Sohn, J. Kim, T. Yun y K. Chang, «Gliomas: Histogram Analysis of Apparent Diffusion Coefficient Maps with Standard- or High-b-Value Diffusion-weighted MR Imaging—Correlation with Tumor Grade,» *Radiology,* vol. 261, nº 3, pp. 882-890, 2011.

[53] Z. Li, Y. Wang y J. Yu, «Deep Learning-Based Radiomics (DLR)and its usage in noninvasive IDH1 prediction for low grade glioma,» *Sci Rep,* vol. 7, p. 5467, 2017.

[54] «The Cancer Imaging Archive,» [En línea]. Available: https://www.cancerimagingarchive.net/.

[55] «Brain Imaging Data Structure,» [En línea]. Available: https://bids.neuroimaging.io/.

[56] «dcm2niix: DICOM to NIfTI converter: compiled versions available from NITRC,» [En línea]. Available: https://github.com/rordenlab/dcm2niix.

[57] «Advanced Normalization Tools,» [En línea]. Available: https://stnava.github.io/ANTs/.

[58] «Advanced Normalization Tools in Python,» [En línea]. Available: https://github.com/ANTsX/ANTsPy.

[59] «Cancer Imaging Phenomics Toolkit (CaPTk),» [En línea]. Available: https://www.med.upenn.edu/cbica/captk/.

[60] «Brain Tumor Segmentation (BraTS) Challenge,» [En línea]. Available: https://www.med.upenn.edu/cbica/brats/.

[61] T. Rohlfing, N. M. Zahr, E. V. Sullivan y A. Pfefferbaum, «The SRI24 multichannel atlas of normal adult human brain structure,» *Human brain mapping,* vol. 31, nº 5, p. 798–819, 2010.

[62] S. Thakur, J. Doshi, S. Pati, S. Ha, C. Sako, S. Talbar, U. Kulkarni, C. Davatzikos, G. Erus y S. Bakas, «Skull-Stripping of Glioblastoma MRI Scans Using 3D Deep Learning,» *Springer - BrainLes 2019 - LNCS,* vol. 11992, pp. 57-68, 2020.

[63] K. Kamnitsas, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, D. Rueckert y B. Glocker, «Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation,» Medical Image Analysis, 2016.

[64] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert y B. Glocker, «Multi-Scale 3D CNNs for segmentation of brain Lesions in multi-modal MRI,» in proceeding of ISLES challenge, MICCAI, 2015.

[65] R. Shinohara, E. Sweeney, J. Goldsmith, N. Shiee, F. Mateen, P. Calabresi, S. Jarso, D. Pham, D. Reich and C. Crainiceanu, "Statistical normalization techniques for magnetic resonance imaging," *Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing, & Alzheimer's Disease Neuroimaging Initiative, NeuroImage. Clinical,* vol. 6, pp. 9-19, 2014.

[66] «NumPy: The fundamental package for scientific computing with Python,» [En línea]. Available: https://numpy.org/. [Último acceso: 2023].

[67] «MONAI: Medical Open Network for Artificial Intelligence,» [En línea]. Available: https://monai.io/.

[68] «MONAI Docs: Transforms,» [En línea]. Available: https://docs.monai.io/en/stable/transforms.html#.

[69] G. Huang, Z. Liu, L. Van Der Maaten y Q. Weinberger, «Densely Connected Convolutional Networks,» *Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2261-2269, 2017.

[70] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso y T. Vercauteren, «On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task,» *ArXiv,* pp. 348-360, 2017.

[71] O. Otkay, «Attention U-Net: Learning Where to Look for the Pancreas,» *arXiv,* 2018.

[72] O. Ronneberger, «U-Net: Convolutional networks for biomedical image segmentation,» *MICCAI,* pp. 234-241, 2015.

[73] S. Jetley, N. A. Lord, N. Lee y P. H. Torr, «Learn to Pay Attention,» de *ICLR 2018 Conference Blind Submission*, 2018.

[74] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit y N. Houlsby, «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,» *Arxiv,* 2020.

[75] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. Roth y D. Xu, «UNETR: Transformers for 3D Medical Image Segmentation,» de *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[76] D. P. Kingma y J. Ba, «Adam: A Method for Stochastic Optimization,» de *3rd International Conference on Learning Representations (ICLR)*, San Diego, 2015.

[77] A. M. Stark, J. van de Bergh, J. Hedderich, H. M. Mehdorn y A. Nabavi, «Glioblastoma: Clinical characteristics, prognostic factors and survival in 492 patients,» *Clinical Neurology and Neurosurgery,* vol. 114, nº 7, pp. 840-845, 2012.

[78] «A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way,» 2018. [En línea]. Available: https://saturncloud.io/blog/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way/.

[79] «Fundamentals: Artificial Neural Networks are to Deep Learning, What Atoms are to Matter,» 2020. [En línea]. Available: https://acsicorp.com/blogs/fundamentals-artificial-neural-networks-are-to-deep-learning-what-atoms-are-to-matter/.

[80] «Transformer (machine learning model),» [En línea]. Available: https://en.wikipedia.org/wiki/Transformer_%28machine_learning_model%29. [Último acceso: 2023].

[81] M. Peng, C. Wang, T. Chen, G. Liu y X. Fu, «Dual Temporal Scale Convolutional Neural Network for Micro-Expression Recognition,» *Front Psychol.,* 2017.

[82] «Receiver operating characteristic,» Wikipedia, [En línea]. Available: https://en.wikipedia.org/wiki/Receiver_operating_characteristic. [Último acceso: 2023].

A. APPENDIX

A.1. Tools, Frameworks and Source Code

The project utilized Python as the primary programming language due to its wide use in data science and machine learning and its extensive libraries and community support. For the development environment, Jupyter Notebook was employed, an interactive platform for coding and visualization. Pycharm was also utilized as an integrated development environment (IDE) for Python.

For preprocessing, pyANTS library was used. ANTS (Advanced Normalization Tools) provides algorithms for neuroimaging, ensuring image quality. For compatibility reasons, Windows Subsystem for Linux (WSL) was engaged, which allows Linux programs to run on a Windows platform. The Cancer Imaging Phenomics Toolkit (CaPTk) was another tool designed for MRI processing. For this phase, a system equipped with an AMD Ryzen 7 5700 8-Core 3.40GHz processor, 16.0GB RAM and an NVIDIA GeForce GTX 1650 GPU was employed.

For the modeling phase, PyTorch was the chosen deep learning framework, often used in AI research, complemented by the use of PyTorch Lightning to streamline and simplify the training process. The MONAI library, designed specifically for healthcare imaging, facilitated the exploration of various model architectures. During the process, "Weight and Biases" (WandB) was integrated for experiment tracking and visualization, providing real-time insights into model performance and metrics. Modeling operations were conducted on Google Colab, which was equipped with an Intel(R) Xeon(R) CPU @ 2.30GHz, 51.0GB RAM and an NVIDIA Tesla T4 GPU with a dedicated 16.0GB of RAM.

The source code of this project can be found at https://github.com/saizk/GlioScan.

A.2. Graphics Resources Utilization

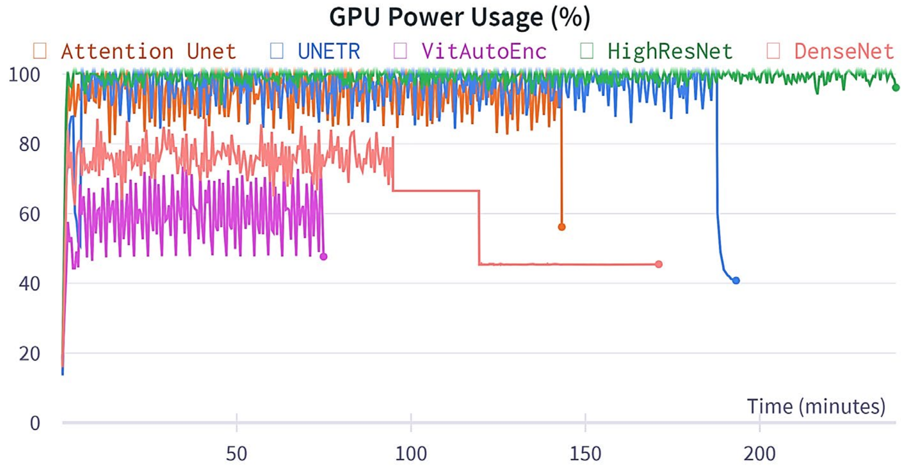


Fig. A.1. GPU Power Usage Comparison

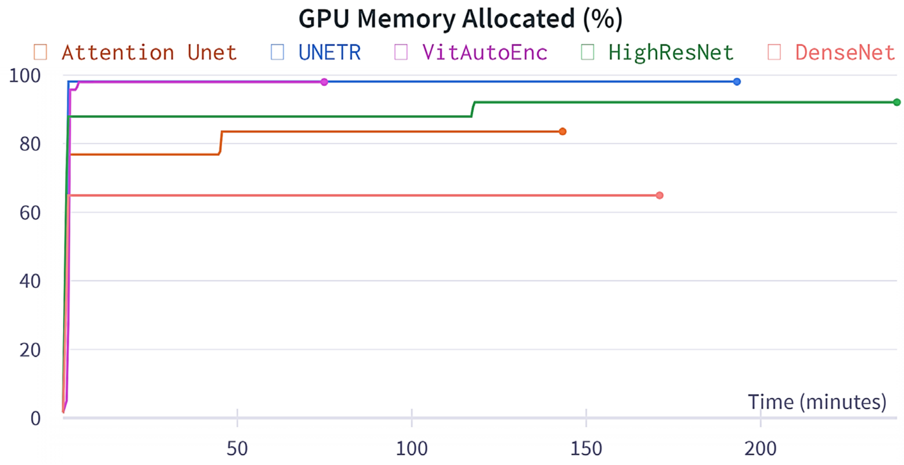


Fig. A.2. GPU Memory Allocation Comparison

A.3. Budget

A.3.1. Material Costs

The primary material expenses correspond to the computational resources for model training and analysis. The breakdown of this cost is as follows:

- Google Colab Pro Subscription: 2 months at 11.19€/month
- Google Colab Pro + Subscription: 1 month at 51.12€/month

TOTAL COST: 73.50€

A.3.2. Personnel Costs

Quantifying the cost associated with human resources represents a challenge, especially as there was no exact registry of the hours dedicated by each professional involved. It is imperative to acknowledge the valuable contributions made by the following:

- Fernando Díaz de María, Professor at University Carlos III of Madrid, who established the initial contact with the hospital and offered guidance and support in the academic and research phases of the project.
- Juan Adan Guzmán-De-Villoria and Pilar Fernández, neuroradiologists of Hospital Gregorio Marañón who put forward the idea of this project, helped in the acquisition of the data and provided expertise in the interpretation of MRI scans and clinical perspectives.

It is crucial to note that, although the personnel costs are not explicitly mentioned in monetary terms, represent a significant intellectual and time investment that was vital for the success of this project.