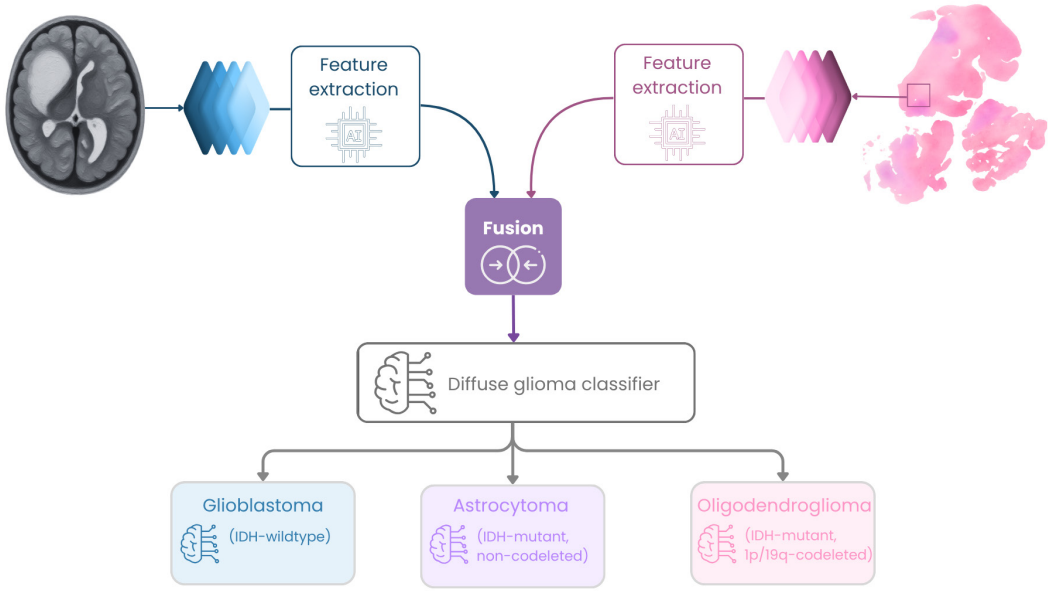
CHALMERS
UNIVERSITY OF TECHNOLOGY



Multimodal Classification of Adult-Type Diffuse Gliomas using Deep Learning on Whole-Slide Images and MRI

Ebba Fredlund, Emma Hedberg

Master's thesis in Electrical engineering

Department of Electrical Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026

Master's thesis 2026

Multimodal Classification of Adult-Type Diffuse Gliomas using Deep Learning on Whole-Slide Images and MRI

Ebba Fredlund, Emma Hedberg

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

Multimodal Classification of Adult-Type Diffuse Gliomas using Deep Learning on Whole-Slide Images and MRI Ebba Fredlund, Emma Hedberg

Supervisor: Ida Häggström, Electrical Engineering
Advisor: Asgeir Jakola, Department of Clinical Neuroscience at Sahlgrenska University Hospital
Examiner: Ida Häggström, Electrical Engineering

Master's Thesis 2026
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Overview of the multimodal deep learning framework for adult-type diffuse glioma classification using whole-slide images and MRI. Typeset in LaTeX
Gothenburg, Sweden 2026

Abstract

Adult-type diffuse gliomas are the most common malignant brain tumors and accurate molecular classification is essential for diagnosis and treatment planning. This thesis investigates deep learning approaches for classifying IDH mutation status and 1p/19q codeletion status using H&E-stained whole-slide images (WSIs) and magnetic resonance imaging (MRI). In a first step, foundation models (FMs) were used to extract feature vectors from the images, which were subsequently used as input to the models that performed the final classification. Both unimodal and multimodal models were evaluated, where different multimodal fusion techniques were explored to combine histopathology and MRI features. The study was conducted on data from Sahlgrenska University Hospital, including 543 WSIs, 528 MRI scans and 525 multimodal patient pairs. Results showed that multimodal models achieved the best overall performance, with the highest test AUC of 0.965 for IDH classification and 0.987 for the codeletion classification task. WSI-based models consistently outperformed MRI-based models, while MRI provided complementary information that improved certain multimodal models. Furthermore, for the WSI-based models, attention heatmaps could be generated, which may improve interpretability and strengthen their potential clinical applicability. The findings demonstrate that deep learning and FMs can enable reliable molecular classification of adult-type diffuse gliomas, while multimodal models offer modest improvements over approaches based only on histopathology.

Keywords: Adult-type diffuse gliomas, deep learning, multimodal, fusion techniques, whole-slide images, magnetic resonance imaging, foundation models, histopathology, IDH mutation, 1p/19q codeletion.

Acknowledgements

We would like to thank our supervisors, Ida Häggström and Asgeir Jakola, for their guidance and support throughout this project. We would also like to thank the members of the Division for Clinical Neuroscience at Sahlgrenska University Hospital for their support during the project. Finally, we thank Sahlgrenska AI Center for providing computational resources that made this work possible.

Ebba Fredlund & Emma Hedberg, Gothenburg, 2026-05-26

Contents

List of Acronyms

ABMIL Attention-Based Multiple Instance Learning.

AI Artificial Intelligence.

AUC Area Under the Curve.

FM Foundation Model.

H&E Hematoxylin and Eosin.

IDH Isocitrate Dehydrogenase.

MIL Multiple Instance Learning.

MLP Multilayer Perceptron.

MR Magnetic Resonance.

MRI Magnetic Resonance Imaging.

NPV Negative Predicted Value.

PPV Positive Predicted Value.

SSL Self-Supervised Learning.

ViT Vision Transformer.

WSI Whole Slide Image.

List of Figures

List of Tables

1

Introduction

Adult-type diffuse gliomas are the most common malignant brain tumors and pose significant challenges for diagnosis, treatment and prognosis [1]. Accurate classification of these tumors is essential, as it guides clinical decision-making and directly impacts patient outcomes. The current gold standard for classification is manual assessment according to the WHO Classification of Tumors of the Central Nervous System [2]. This classification distinguishes three main tumor types: IDH-mutant oligodendrogliomas with 1p/19q codeletion, IDH-mutant astrocytomas without 1p/19q codeletion and IDH-wildtype glioblastoma.

In current clinical practice, the assessment is performed by examining tumor tissue slides under a microscope. However, manual tumor classification is time-consuming, expensive and may be inconsistent between observers [3]. This variability can lead to differing or incorrect diagnoses and consequently, inappropriate treatment decisions. In addition, the high cost and need for specialized expertise limit access to such analyses, leading to unequal access to healthcare across regions.

The introduction of scanners in the 1990s enabled tissue slides to be digitized as an Whole Slide Image (WSI) and laid the foundation for digital pathology [3]. Digital pathology allows faster and more consistent diagnostics by enabling tissue samples to be scanned and analyzed digitally. This development has led to a growing interest in the application of Artificial Intelligence (AI) in pathology. Recent advances in AI and deep learning offer the potential to automate and improve the accuracy of tumor classification [4]. In particular, Self-Supervised Learning (SSL) has enabled the development of pretrained Foundation Model (FM), which can learn generalizable representations from large amounts of unannotated data and then be adapted to specific tasks.

Despite these advances, most AI models for medical image analysis rely on a single data modality, such as either histopathology or Magnetic Resonance Imaging (MRI) [5]. This creates a gap with clinical reality, where clinicians make decisions based on multiple sources of information. Combining different imaging modalities, such as histopathology and radiology, has therefore emerged as a promising direction for medical image analysis and clinical decision support [6].

Currently, multimodal approaches for glioma classification are not widely implemented in clinical practice and to the best of our knowledge, such methods have not been applied in Sweden. In this thesis, conducted in collaboration with Sahlgrenska

University Hospital, unimodal and multimodal classifiers are developed using WSIs and Magnetic Resonance (MR) images to classify adult-type diffuse gliomas. The aim of this thesis is to investigate whether combining these modalities can improve classification performance compared to single-modality approaches.

1.1 Related work

Several studies have compared and used histopathology FMs for classification [7], [8], [9], [10], [11], [12]. Pretraining strategies have been shown to be critical, with models trained on medical images consistently outperforming those trained on natural images [8]. In addition, studies have also demonstrated the importance of diverse tissue representation and the use of SSL [8], [12]. Comparative studies of models pretrained on WSIs further show that H-optimus-0, Prov-GigaPath and UNI achieve consistently strong performance across downstream tasks [9], [10].

In a review of 150 studies on medical FMs, V. van Veldhuizen et al. noted that most existing work focuses on pathology, while radiology and particularly MRI FMs remain limited [4]. They suggest that this limitation is largely due to the need to process 3D imaging data, whereas most current deep learning architectures are optimized for 2D inputs. Nevertheless, recent FMs such as BrainIAC and 3DINO-ViT, pretrained on brain MR images, have demonstrated strong performance across several downstream tasks [13], [14].

Several studies have developed multimodal models that integrate clinical data, pathology images, radiology images and extracted features from these images [15], [16], [17], [18], [19]. These approaches typically use late or mid-level fusion and consistently outperform unimodal models for tasks such as subtype classification. Hamidinekoo et al. compared different fusion strategies, highlighting transformer-based cross-modal attention as an emerging trend in medical imaging [17]. A recent study of Saueressig et al. demonstrates that combining histopathology and MRI FMs can outperform unimodal approaches [6]. This is the only study to date that includes feature extraction with FMs for both WSI and MR images, achieving an Area Under the Curve (AUC) of 0.94 on an independent test set.

Despite these advances, important gaps remain. Most studies focus on single modalities and although multimodal approaches have shown promise, the use of FMs for both WSIs and MR images remains limited, particularly for glioma classification. In addition, many studies rely on large public datasets, with less focus on institution-specific data that reflect real-world challenges such as limited samples and imaging artifacts. This motivates further investigation of multimodal pipelines in clinically representative settings.

1.2 Aim and objective

The primary goal of this project is to classify adult-type diffuse gliomas from a dataset of WSIs and MR images. The classification process will follow the 2021 WHO guidelines and is illustrated in Figure 1.1. The first step involves identifying IDH-mutant and IDH-wildtype tumors. Finally, for the IDH-mutant cases, 1p/19q codeletion is assessed as a part of the molecular classification.

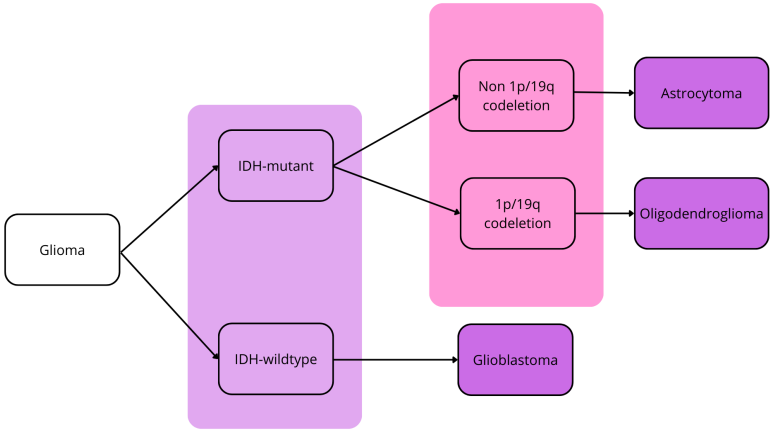


Figure 1.1: Classification pipeline of adult-type diffuse gliomas.

The study will develop both unimodal and multimodal models using WSIs and MR images for both classification tasks (IDH and 1p/19q codeletion). The WSI-based, MRI-based and multimodal models will be compared to evaluate their relative performance. Furthermore, the study aims to evaluate different deep learning architectures and FMs to determine which approaches achieve the best performance for this task. Another objective is to analyze which regions of the WSIs are most influential in the model's predictions using explainable AI methods.

The following research questions will be addressed for each model type:

- How accurately can the models classify adult-type diffuse gliomas, as measured by AUC, accuracy, positive predictive value, negative predictive value, recall and specificity?
- Which model and architecture provide the best performance for this classification task?
- Which regions of the WSIs are most influential in the models predictions and how can explainable AI methods be used to visualize them?

A successful outcome is defined as performance comparable to state-of-the-art methods reported in related studies. In this study, metrics above 90% are considered indicative of strong performance.

1.3 Limitations

This study has several limitations that should be considered when interpreting the results. First, the FMs are frozen and used as feature extractors in the initial stage. The study focuses on a limited set of in-domain histopathology FMs (H-optimus-0, Prov-GigaPath and UNI2-h), as well as BrainIAC and 3DINO-ViT for MRI. For the multimodal models, only mid-level fusion is considered and a simple Multilayer Perceptron (MLP) is used as the classifier for both the MRI and multimodal models. For the WSI models, only PyTorch Attention-Based Multiple Instance Learning (ABMIL) is used as the aggregator and classifier. Furthermore, the MRI models had limited interpretability, as they did not provide spatial attention maps indicating which image regions contributed to the model's decisions.

The dataset is relatively small and includes patients from a single institution, Sahlgrenska University Hospital. All WSIs were acquired using the same scanner, which may limit generalizability. Only one WSI and one MR image sequence per patient are used. The model is restricted to adult-type diffuse gliomas and focuses only on IDH mutation and 1p/19q codeletion, despite the existence of additional relevant molecular markers.

Compared to previous studies, this work uses a more limited biomarker set (IDH mutation and 1p/19q codeletion), resulting in weaker labels and does not select regions of interest or tiles but instead uses the whole WSI. It also restricts itself to in-domain FMs and does not consider tumor grade, due to overlapping grades within the same subtype and an imbalanced grade distribution in the dataset.

2

Theory

This chapter introduces the theoretical background relevant to this study, covering both the clinical and technological parts. First, adult-type diffuse gliomas are presented, including the classification steps and key molecular characteristics. This is followed by an overview of medical imaging modalities, with a focus on histopathology and Magnetic Resonance Imaging (MRI). The chapter then introduces Foundation Model (FM)s and their role in medical image analysis, before discussing multimodal learning approaches that combine information from multiple data sources. Finally, classification heads used for downstream prediction tasks are described.

2.1 Adult-type diffuse gliomas

The most prevalent malignant tumors in the central nervous system are adult-type diffuse gliomas [1]. Gliomas arise from glial progenitor cells and develop as a result of genetic changes that gradually accumulate as the tumor grows. Each year approximately six new cases are diagnosed per 100 000 individuals worldwide. Gliomas are more common in men, who are about 1.6 times more likely to be diagnosed than women [20]. The symptoms depend on how fast the tumor grows and where it is located. Symptoms can be either general or specific to the affected area of the brain [21]. The growth rate depends on the grade of the tumor, where lower-grade gliomas (grades 2-3) grow more slowly and are less aggressive. These give symptoms for years before being diagnosed and lead to subtle neurologic changes rather than sudden deficits, except when the first symptom is a first-time seizure (i.e., an epileptic event) [22]. High-grade gliomas (grade 4) are fast-growing, highly invasive and related to poor prognosis. They usually present symptoms only weeks prior to diagnosis. Common symptoms the year before diagnosis are usually non-specific and similar in both high- and lower-grade gliomas, including headache, mental tiredness and fatigue [20], [23].

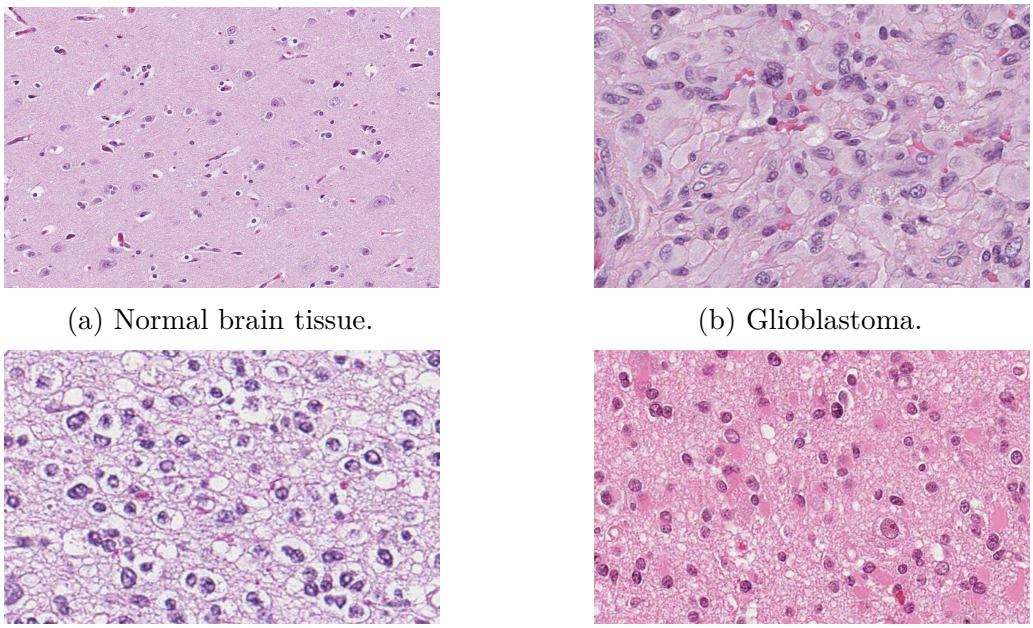
Diffuse gliomas are divided into lower- and high-grade gliomas and include the subtypes astrocytoma, oligodendroglioma and glioblastoma. Glioblastoma is the highest-grade (grade 4) and most aggressive type of brain tumor [24], [25]. It accounts for about 45% of all brain tumors and the median age of diagnosis is 64 years. It has a rapid growth rate and is highly invasive, leading to a poor prognosis [25]. Despite undergoing treatment such as radiotherapy and surgery, the average survival length is only 14.8 months.

Oligodendrogliomas are rare, slow-growing tumors accounting for approximately 5% of all brain tumors [26]. The median age at diagnosis is about 45 years. Compared to glioblastomas, oligodendrogliomas are associated with a relatively favorable prognosis with reported median overall survival ranging from 12 to over 17 years. Astrocytomas, on the other hand, occur across a range of grades and can exhibit both slow- and fast-growing behavior [27]. Depending on the grade the survival length can vary from months to years.

To determine tumor grade and establish a diagnosis based on tumor tissue, the 2021 WHO Classification of Tumors of the Central Nervous System is currently used as the standard framework [2]. Accurate classification is critical for diagnosis, treatment planning and prognosis. Previously, the classification only relied on histological features such as necrosis (dead tissue) and microvascular proliferation (formation of new, small blood vessels), but the new update from 2021 has integrated molecular profiles. The update places greater emphasis on biomarkers, such as IDH mutation and 1p/19q codeletion status [1]. IDH mutation is key to classify adult-typ diffuse gliomas and means that the gene encoding for Isocitrate Dehydrogenase (IDH) gets mutated. This mutation causes increased production and accumulation of the metabolite 2-hydroxyglutarate [28]. Genes that lack this mutation is called IDH-wildtype. The definition of 1p/19q codeletion is a combined loss of the short arm of chromosome 1 (1p) and the long arm of chromosome 19 (19q) [29]. Both IDH mutation and 1p/19q codeletion is associated with a favorable prognosis.

The update classification system from 2021 resulted in three main disease groups for adult-type diffuse gliomas: IDH-mutant oligodendrogliomas with 1p/19q codeletion (grades 2-3), IDH-mutant astrocytomas without 1p/19q codeletion (grades 2-4) and IDH-wildtype glioblastomas (grade 4). A simplified classification flow chart is shown in Figure 1.1. As discussed, accurate classification requires histopathological and molecular analysis of the tumor tissue. Therefore, surgery is necessary to obtain tissue for both definitive diagnosis and therapeutic purposes. A confirmed diagnosis can only be made after surgery, until then the tumor is classified and considered a suspected or presumed glioma.

For patients with a presumed glioma, the first step in clinical classification is brain MRI, which provides information about the tumor location and characteristics [20]. This is followed by a biopsy or resection of the tumor tissue, depending on its location and the condition of the patient. After that, the tissue undergoes histopathological and molecular examination by a pathologist. Techniques such as Hematoxylin and Eosin (H&E) staining and immunohistochemistry are used to evaluate the tumor tissue under the microscope [28]. Figure 2.1 illustrates H&E-stained images of normal brain tissue and the three main disease groups of adult-type diffuse gliomas. H&E-staining assesses both microvascular proliferation and necrosis, which are associated with higher-grade gliomas. Immunohistochemistry is used to test for IDH mutations, while 1p/19q codeletion is assessed using molecular methods. Together, these markers refine the diagnosis and determine the tumor subtype.



(a) Normal brain tissue.

(b) Glioblastoma.

(c) Oligodendroglioma.

(d) Astrocytoma.

Figure 2.1: H&E-stained WSIs of normal brain tissue and brain tumors (astrocytoma, oligodendroglioma and glioblastoma). Images adapted from the Human Protein Atlas [30]. Image credit: Human Protein Atlas. Available under CC BY-SA 4.0. Source images: `https://www.proteinatlas.org/learn/dictionary/normal/cerebral+cortex`, `https://www.proteinatlas.org/learn/dictionary/cancer/glioma#Glioma-1,-Astrocytoma`, `https://www.proteinatlas.org/learn/dictionary/cancer/glioma#Glioma-2,-Oligodendroglioma`, `https://www.proteinatlas.org/learn/dictionary/cancer/glioma#Glioma-3,-Glioblastoma-multiforme`.

2.2 Medical imaging modalities

As described above, the diagnostic workflow for classifying adult-type diffuse gliomas begins with brain MRI, followed by biopsy or surgical resection of tumor tissue for histopathological analysis [20]. The tissue is then stained using H&E and digitized using a slide scanner to produce a Whole Slide Image (WSI). In this work, the two imaging modalities are considered and their respective characteristics are described in the following subsections.

2.2.1 WSI

Whole Slide Image (WSI)s are high-resolution digital images of H&E-stained tissue [31]. These images are typically generated at resolutions of up to 0.25 micrometers per pixel over a standard tissue slide (approximately 20 mm x 15 mm), resulting in gigapixel-scale data. At this scale, histological structures become visible through characteristic staining patterns, where hematoxylin stains cell nuclei in shades of blue and purple, while eosin highlights cytoplasmic and extracellular components in shades of pink [32]. The accurate interpretation and classification of such large and

detailed images place substantial demands on pathologists [31].

The process of H&E-staining requires careful and standardized preparation to ensure consistent image quality for downstream analysis [32]. Despite standardized protocols, factors such as time, temperature and pH can be difficult to control in practice [33]. These factors can affect the staining of tissue sections, leading to variation in color across WSIs. When the slides are scanned different scanners may be used, which can have varying resolution and quality, further influencing the color of the digitized images. Therefore, color correction methods can be used to make image analysis practically useful.

2.2.2 MRI

Magnetic Resonance Imaging (MRI) is a widely used non-invasive imaging technique for visualizing and analyzing soft tissues in the human body [34]. It does not involve ionizing radiation, which makes it a relatively safe diagnostic tool. However, it is generally more expensive and cannot be used in patients with certain magnetizable implants or claustrophobia.

The underlying principle of MRI is based on the magnetic properties of hydrogen protons [35]. When a patient is placed in a strong external magnetic field, these protons align either parallel or antiparallel to the field. A radiofrequency pulse is then applied, which excites the protons and displaces them from their equilibrium state. As the pulse is turned off, the protons return to their original state through relaxation, emitting signals that are detected by receiver coils. To determine the spatial origin of the signals, small variations in the magnetic field, called gradients, are applied. The acquired data is transformed into the frequency domain and reconstructed to produce detailed 3D images.

There are two types of relaxation that can be measured, T1 (longitudinal) relaxation and T2 (transverse) relaxation [35]. T1-relaxation describes how quickly the protons realign with the main magnetic field, whereas T2-relaxation measures the time it takes for the protons to lose phase coherence induced by the radiofrequency pulse. Different tissues have distinct T1 and T2 values, which generate contrast in the resulting images. By adjusting MRI parameters such as the echo time (TE) and repetition time (TR), various types of sequences can be obtained.

One common sequence is T1-weighted, which uses short TE and TR times [36]. Gadolinium contrast is commonly injected intravenously with this sequence, referred to as T1-weighted contrast-enhanced (T1GD). The contrast agent shortens T1-relaxation, thereby brightening leaky blood vessels around tumors. Two other common sequences are T2-weighted and Fluid Attenuated Inversion Recovery (FLAIR). Both of these use longer TE and TR compared to T1-weighted images, with FLAIR employing the longest times, resulting in higher contrast. Figure 2.2 shows the same brain slice with T1-weighted, T2-weighted, FLAIR and T1GD sequences.

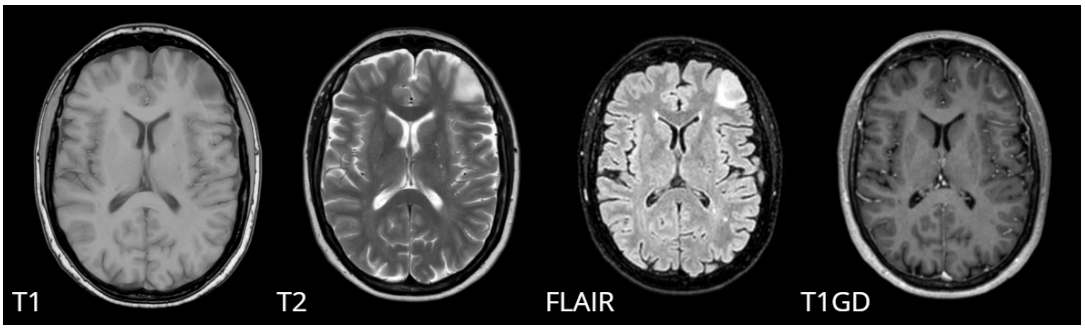


Figure 2.2: MRI scans of a patient with astrocytoma, showing T1-weighted, T2-weighted, FLAIR and T1GD sequences.

Magnetic Resonance (MR) images may contain artifacts and noise originating from the scanner, the patient or the imaging process [37]. Common artifacts include truncation, motion, aliasing, chemical shift artifacts, as well as distortions caused by the presence of metal objects. Intensity inhomogeneity artifacts may also occur due to imperfections in the magnetic field. These factors can reduce image quality, affect interpretation and lead to variations in image contrast.

2.3 Image preprocessing

To ensure standardized input data and reduce artifacts and noise, preprocessing is required before WSIs and MR images can be used in deep learning models. Common preprocessing techniques for each modality are described below.

2.3.1 WSI preprocessing

A challenge when it comes to use WSIs in AI models is the size since it places considerable demands on computational resources and hardware [9]. To address this, the WSI is normally divided into smaller regions, or tiles, which are then passed through a neural network separatly. This step is by far the most computationally intensive but only needs to be done once. All outputs from each tile are then aggregated to produce a comprehensive representation of the entire slide.

To further reduce the image size, Otsu's method can be employed [38]. This technique removes the background, which in turn reduces the overall size of the WSI. The method separates objects from the background based on pixel intensities by evaluating different threshold values. The threshold that maximizes the separation between the pixel groups below and above it is selected as optimal. Otsu's method is widely used in medical imaging for the identification of cells or tumor regions. The most common pipeline for the image preprocessing for WSIs is shown in Figure 2.3.

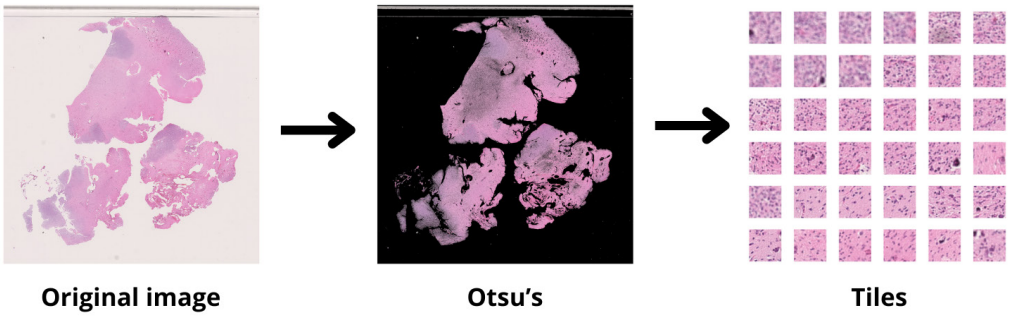


Figure 2.3: Overview of a common preprocessing pipeline for a WSI, where tissue regions are extracted using Otsu's method and subsequently divided into tiles.

2.3.2 MR image preprocessing

To reduce intensity inhomogeneity, the N4 bias field correction algorithm can be used [13]. This method iteratively estimates the bias field and removes it from the image, resulting in a more uniform intensity distribution [39]. Another step is resampling to ensure consistent spatial resolution across all images [13]. This is achieved by interpolating all volumes to an isotropic voxel size, ensuring that voxel dimensions are equal in all directions.

To enable spatial alignment and comparability across subjects, registration and skull stripping may be applied. In image registration, individual MR images are rigidly aligned to a standardized brain template (MNI brain atlas) and transformed into MNI space [13]. Brain extraction, or skull stripping, is used to remove non-brain tissue such as the skull and scalp [40]. This is a common preprocessing step in neuroimaging applications and may contribute to patient privacy by removing facial structures, that could potentially be reconstructed from MRI data. It can also improve the quality of downstream analysis by focusing the model on relevant brain structures. A tool for this is HD-BET, a deep learning-based method trained for automated brain extraction [13].

2.4 Foundation models

Historically, the shortage of comprehensive labeled datasets has posed a significant challenge to the development of reliable medical AI models [9]. This challenge has been addressed through the introduction of Self-Supervised Learning (SSL), which enables models to learn general and transferable visual representations that can subsequently be adapted to more specialized tasks. In recent years, several FMs, also known as pretrained models, have been introduced based on SSL. These models can be trained either in-domain, meaning on domain-specific data such as WSIs or out-of-domain, where the training data consist of different types of natural images [41]. Within healthcare, various types of in-domain FMs exist, which are predominantly trained using SSL on large-scale collections of medical images [4].

There are several types of SSL frameworks, which can mainly be categorized into discriminative and generative methods [4]. Instead of relying on labeled data, these methods use auxiliary tasks that the model learns to solve. Through this process, the model learns meaningful representations of the input data that can later be transferred to downstream tasks. Within discriminative SSL, contrastive learning is a commonly used framework, with methods such as SimCLR, MoCo and SwAV being popular examples. These methods learn to identify different augmented versions of the same image as similar, while distinguishing them from other images. Another class of discriminative SSL methods is distillation-based approaches, including BYOL, DINO and DINOv2. These methods typically employ a teacher-student architecture, where the teacher network generates target representations (soft labels) that the student network learns to reproduce. In contrast, generative SSL methods learn by reconstructing missing information from the input data. Two common examples are Masked Image Modeling (MIM) and Masked Autoencoders (MAE).

Medical FMs architecture are mainly based on convolutional neural networks (CNNs) or Vision Transformer (ViT) [9]. CNNs use sliding convolutional filters to extract hierarchical features from images [4]. This makes them efficient at learning local patterns such as edges and textures, especially when limited training data are available. In a ViT, the input image is divided into small, non-overlapping segments known as patches [42]. Each patch is transformed into a numerical representation, called an embedding, which serves as input to the transformer blocks. To preserve spatial information, a learnable positional embedding is added to each patch. The transformer blocks then apply self-attention mechanisms to model the relationships between patches. The outputs from all patches are subsequently aggregated and passed through a feedforward neural network, which produces the final classification. By combining patch-based representations with attention mechanisms, ViTs can learn both local patterns and relationships between patches [4]. On the other hand, they require larger datasets to achieve optimal performance.

This section further presents FMs for histopathology and MRI applications, including specific models, their underlying backbone architectures and the SSL methods used during pretraining.

2.4.1 Histopathology foundation models

There are several FMs developed for histopathology tasks. Table 2.1 presents three of them that are used in this study. These models were selected based on their performance in previous studies, see Section 1.1.

Table 2.1: Histopathology foundation models.

Model	**Parameters (M)**	**SSL**	**Training data Source**	**Tiles (M)**	**Slides (K)**	**Training Resolution**
H-optimus-0	1135	DINOv2	Proprietary	>100	>500	20x
Prov-GigaPath	1135	DINOv2, LongNet	PHS	1300	171	20x
UNI2-h	681	DINOv2	MGB	200	>350	20x

2.4.1.1 H-optimus-0

H-optimus-0 is a FM for histopathology trained on more than 500 000 H&E-stained WSIs [43]. The dataset includes tissue samples from multiple anatomical sites and has a high patient diversity, with an average of 1.5 slides per patient. The model is based on a ViT-Giant and is trained using discriminative SSL (DINOv2). It is one of the largest models used in computational pathology to date. The model has been evaluated on both tile-level and slide-level tasks. At the tile level, it shows strong performance in tissue classification and recognition of morphological features, outperforming several baseline and state-of-the-art approaches on standard benchmark datasets. For slide-level tasks, the model is combined with Attention-Based Multiple Instance Learning (ABMIL) to enable whole-slide analysis. In this setting, it achieves strong results in tasks such as lymph node metastasis detection in breast cancer and biomarker detection.

2.4.1.2 Prov-GigaPath

Prov-GigaPath is a FM trained on a dataset with 171 189 WSIs from Providence Health and Services (PHS) [44]. The dataset includes both H&E-stained and immunohistochemistry (IHC) images from 31 tissue types, including brain tumors. It uses ViT-based tile encoders to extract local tile features and a slide encoder to integrate information across the whole slide. For pretraining they used DINOv2 for the tile-encoder and LongNet combined with MAE for the slide encoder. For cancer subtype classification, Prov-GigaPath can be used together with ABMIL that aggregates tile embeddings to capture slide-level patterns from all tiles.

2.4.1.3 UNI2-h

UNI2-h is a FM that has been trained on a dataset with over 100 000 WSIs from 20 major tissue types [45]. It is based on a ViT-Large and uses discriminative SSL via DINOv2 to learn rich visual representations. The model has been evaluated across 34 clinical tasks, including metastasis detection, cancer grading and tumor subtyping, consistently outperforming previous state-of-the-art models such as CTransPath and REMEDIS.

For slide-level classification UNI2-h is combined with Multiple Instance Learning (MIL), specifically ABMIL. Studies show that using UNI2-h features with ABMIL outperforms many more complex MIL architectures, highlighting that a strong pretrained encoder can be more important than advanced MIL design [45]. UNI2-h is particularly effective for brain tumor subtyping, including IDH-mutated astrocytomas, IDH-mutated/1p19q-codeleted oligodendrogliomas and IDH-wildtype glioblastomas. It is also effective at predicting IDH mutation status, distinguishing between IDH-mutated and IDH-wildtype tumors. The model generalizes well to rare cancer types, shows strong data efficiency for few-shot learning and supports robust performance across varying image resolutions.

2.4.2 MRI foundation models

Compared to histopathological FMs, FMs for MRI are less widely available [4]. This is mainly due to the 3D of MR images, whereas most FM architectures are designed for 2D data, similarly to ViTs. To work with 3D, the images are often divided into multiple 2D slices, but this approach results in a loss of important spatial information. However, building FMs specifically for 3D images requires greater computational power, larger datasets and more advanced analysis methods in order to capture rich and meaningful representations. Table 2.2 presents the two FMs that are used in this study, which were selected based on their performance in previous studies, see Section 1.1.

Table 2.2: MRI foundation models.

Model	**Parameters (M)**	**SSL**	**Training data**	**MR images (K)**
BrainIAC	88	SimCLR	MRI	32
3DINO-ViT	307	DINOv2	MRI, CT, PET	70

2.4.2.1 BrainIAC

BrainIAC is a FM pre-trained on 32 015 MR images from 16 datasets across 10 medical conditions [13]. Prior to model training, the MR images were systematically preprocessed, including conversion from DICOM to NIfTI format, N4 bias field correction, resampling to $1 \times 1 \times 1\,\mathrm{mm}^3$ voxels, registration to MNI space and skull stripping to isolate brain tissue. The model was pretrained using two SSL strategies, SimCLR and MAE, with various encoder architectures evaluated for each. Benchmarking under limited data conditions showed that SimCLR with a ViT-Base encoder provided the most effective feature representations and was selected as the BrainIAC backbone.

The model processes standardized 3D MRI volumes of size $96 \times 96 \times 96$ voxels and learns representations through contrastive learning between augmented volumetric views. In addition, the pretraining pipeline incorporated extensive spatial augmentations to improve robustness and spatial feature learning.

2.4.2.2 3DINO-ViT

The 3DINO-ViT FM was pretrained on a dataset of approximately 100 000 unlabeled 3D scans, including 70 434 MRI volumes, from over ten different organs [14]. Prior to model training, the brain MR images were skull-stripped. The model uses a distillation-based SSL approach with a teacher-student framework that combines both global image-level and local patch-level objectives. To achieve this, the original volume is augmented multiple times to create two global crops and eight local crops. The model then learns to align representations of these augmented views originating from the same volume. In addition, masked patch representations from the student network are matched to corresponding unmasked representations from the teacher network. The model backbone is based on a ViT-Large architecture adapted for 3D

input, with $16 \times 16 \times 16$ voxel patches and a 3D ViT adapter to capture spatial information for improved segmentation performance.

2.5 Model architecture components

This section presents common model architecture components used following FM feature extraction. This study will use ABMIL for embedding aggregation, and both concatenation and cross-modal attention as fusion strategies, followed by an MLP as the final classification head.

2.5.1 Aggregation of embeddings

Multiple Instance Learning (MIL) is a supervised learning framework used when detailed annotations are not available for every image region [46]. In digital pathology, a WSI is treated as a *bag* of smaller patches, or *instances*, with a single label assigned at the slide level. The model learns to identify which instances contribute most to the bag label [47]. Each bag (WSI) is represented as a set of instances

$$X = \{x_1, x_2, \ldots, x_N\}, \tag{2.1}$$

where N is the number of patches in the WSI, and each instance x_n corresponds to a patch-level feature vector extracted by a pretrained model. The bag is associated with a binary label $Y \in \{0, 1\}$ that depends on the instance-level labels y_n. Under the standard MIL assumption, a bag is positive if at least one instance is positive:

$$Y = \max_n y_n. \tag{2.2}$$

From a probabilistic perspective, the model estimates the bag-level probability

$$\theta(X) = P(Y = 1 \mid X), \tag{2.3}$$

and assumes

$$Y \mid X \sim \text{Bernoulli}(\theta(X)). \tag{2.4}$$

The model is trained by maximizing the log-likelihood of the Bernoulli distribution over the training bags [47]. For this to work, $\theta(X)$ must be permutation-invariant, meaning that the prediction must be independent of the ordering of the instances in X. According to the fundamental result on permutation-invariant set functions, any symmetric function $S(X)$ defined on a set X can be decomposed as

$$S(X) = g\left(\sum_{x \in X} f(x)\right), \tag{2.5}$$

where f and g are suitable transformations. In the MIL setting, this corresponds to transforming patch-level feature vectors using

$$h_n = f_\psi(x_n), \tag{2.6}$$

aggregating them through a permutation-invariant pooling operator

$$z = \sigma(\{h_n\}_{n=1}^{N}), \tag{2.7}$$

and mapping the aggregated representation to a bag-level prediction via

$$\theta(X) = g_\phi(z). \tag{2.8}$$

Common permutation-invariant pooling operators include max pooling and mean pooling. While MIL is an effective approach for slide-level label assignment, it is prone to overfitting, particularly when applied to small datasets [47].

2.5.1.1 Attention-based MIL

Another way to construct the pooling operator is to use a trainable weighted sum that adapts to the transformed patch-level feature vectors, commonly referred to as Attention-Based Multiple Instance Learning (ABMIL) [47]. Instead of treating all instances equally (as in mean pooling) or selecting only the most dominant one (as in max pooling), attention-based pooling learns instance-specific weights through a neural network. Let $H = \{h_1, \ldots, h_N\}$ denote a bag of N low-dimensional embeddings. The bag representation is computed as a weighted average

$$z = \sum_{n=1}^{N} a_n h_n, \tag{2.9}$$

where the attention weights a_n are determined by a small neural network and satisfy

$$\sum_{n=1}^{N} a_n = 1. \tag{2.10}$$

The weights are typically defined using a softmax function to ensure positivity and normalization:

$$a_k = \frac{\exp\left(w^\top \tanh(V h_k^\top)\right)}{\sum_{j=1}^{K} \exp\left(w^\top \tanh(V h_j^\top)\right)}, \tag{2.11}$$

where V and w are learnable parameters.

An extension of this formulation is gated attention, also proposed in [47]. Here, an additional gating mechanism is used to increase the expressiveness of the attention network. In this case, the attention weights are computed as

$$a_k = \frac{\exp\left(w^\top \left[\tanh(V h_k^\top) \odot \sigma(U h_k^\top)\right]\right)}{\sum_{j=1}^{K} \exp\left(w^\top \left[\tanh(V h_j^\top) \odot \sigma(U h_j^\top)\right]\right)}, \tag{2.12}$$

where $\odot$ denotes element-wise multiplication, $\sigma(\cdot)$ is the sigmoid activation function, and U, V, and w are learnable parameters. The gating mechanism allows the model to capture more complex relationships between instances and has been shown to improve the flexibility of the attention-based pooling operator.

Both standard and gated attention formulations allow the model to assign higher weights to the most informative instances while remaining permutation-invariant and independent of bag size. Consequently, the network can focus on diagnostically relevant patches within a WSI.

2.5.2 Fusion strategies

Multimodal models combine information from multiple data sources or modalities, such as imaging, pathology and clinical data to improve predictive performance over unimodal models [6], [16], [17]. These models are useful in medical applications, where different modalities are able to capture complementary aspects of a disease. Multimodal models also reflects more closely how a clinician works in practice, integrating multiple sources of information before arriving at a final decision. A key component in multimodal models is the fusion strategy, which determines how information from different modalities is combined.

Fusion can be broadly categorized into three types: early, mid-level and late fusion [48]. Early fusion, also called input-level fusion, combines raw inputs or very early features from different modalities into a unified representation, which is then processed by a single network for prediction. This strategy allows the model to learn cross-modal interactions from the start, but it can be sensitive to differences in feature scale and modality quality.

Mid-level fusion or feature-level fusion integrates information after each modality has been processed through a modality-specific network. The resulting embeddings or high-level features are combined into a shared representation, which is then used for downstream tasks such as classification. This approach preserves modality-specific representations while enabling joint learning across modalities [6], [17]. Late fusion or decision-level fusion combines the outputs of independently trained unimodal models, typically through averaging, voting or another aggregation mechanism. While this approach is simpler and more robust to missing modalities, late fusion may not fully capture complex interactions between modalities [16].

2.5.2.1 Concatenation and cross-modal attention

This study will use mid-level fusion strategies, specifically concatenation and cross-modal attention, for combining multimodal representations. Concatenation is a simple fusion strategy where feature vectors from different modalities are combined into a single joint representation by concatenating them [48].

$$\mathbf{z} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_n] \tag{2.13}$$

where $\mathbf{x}_i$ represents the feature vector from modality (i), and $\mathbf{z}$ denotes the fused representation.

Cross-modal attention is a fusion method where one modality, for example WSI, can focus on and weight information from another modality, for example MRI, [49]. This allows the model to learn relationships between modalities by selecting the most relevant features instead of simply combining them through concatenation [50].

$$\mathbf{z} = \sum_{i=1}^{N} a_i \mathbf{x}_i \tag{2.14}$$

where N is the number of modalities, $\sum_{i=1}^{N} a_i = 1$ defines the learnable attention weights, $\mathbf{x}_i$ represents the feature vector and $\mathbf{z}$ denotes the fused representation.

2.5.3 Multilayer perceptron

A Multilayer Perceptron (MLP) is a network composed of fully connected layers, also called dense layers [51]. Fully connected layers consists of neurons where each neuron is connected to all neurons in the previous layer, see Figure 2.4. Each connection is associated with a learnable weight that determines the strength of the signal passed between neurons. In addition, each neuron has a bias term that shifts the output of the linear transformation. Mathematically, the output of neuron j in a fully connected layer is given by:

$$z_j = \sum_{i=1}^{N} w_{ij} x_i + b_j, \tag{2.15}$$

where x_i represents the input features, w_{ij} are the weights, b_j is the bias term and N is the number of input features. The linear output z_j can then be passed through a non-linear activation function $f(\cdot)$, such as ReLU, GeLU or Tanh:

$$a_j = f(z_j). \tag{2.16}$$

The inclusion of non-linear activation functions enables the network to learn complex, non-linear relationships in the data.

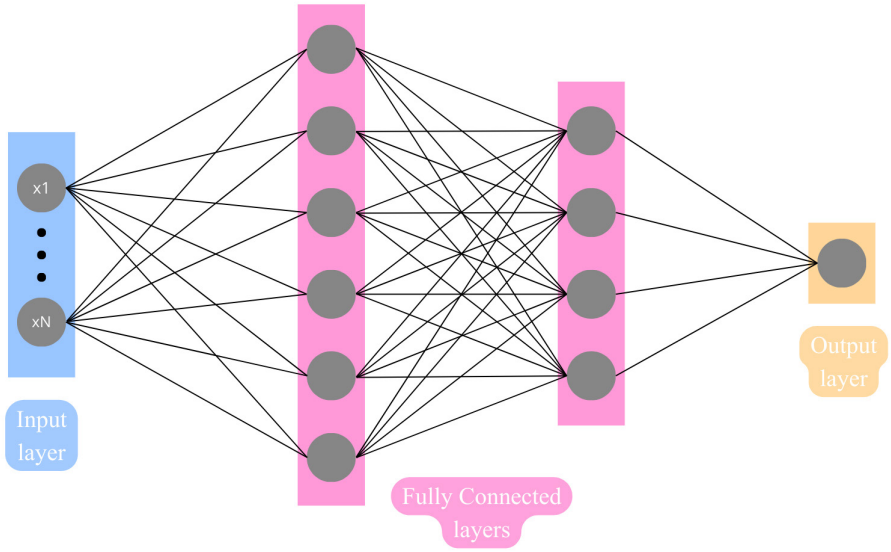


Figure 2.4: Graph of an input layer connected to fully connected layers, followed by a single-neuron output layer.

2.6 Evaluation Metrics

The model is evaluated using a set of performance metrics. These metrics provide different perspectives on the model's classification ability, including its overall accuracy and its performance in identifying positive and negative instances. Together, they present a comprehensive understanding of the model's strengths and limitations.

2.6.1 Accuracy

Accuracy is the proportion of correct predictions out of all predictions. For a binary classifier, it is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.17}$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative.

2.6.2 Positive Predictive Value

Positive Predicted Value (PPV), also called precision, is defined as:

$$PPV = \frac{TP}{TP + FP} \tag{2.18}$$

It measures the proportion of predicted positive instances that are actually positive. A high PPV indicates a low number of false positives.

2.6.3 Negative Predictive Value

Negative Predicted Value (NPV) is defined as:

$$NPV = \frac{TN}{TN + FN} \tag{2.19}$$

It measures the proportion of predicted negative instances that are actually negative. A high NPV indicates a low number of false negatives.

2.6.4 Recall

Recall, also called sensitivity or True Positive Rate (TPR), is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{2.20}$$

It measures the proportion of actual positive instances that are correctly identified by the model. A high recall indicates a low number of false negatives.

2.6.5 Specificity

Specificity, also called True Negative Rate (TNR), is defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{2.21}$$

It measures the proportion of actual negative instances that are correctly identified by the model. A high specificity indicates a low number of false positives.

2.6.6 AUC

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures the model's ability to distinguish between the two classes across all possible classification thresholds. For each threshold, one calculates the true positive rate (TPR) and the false positive rate (FPR), where $FPR = \frac{FP}{FP+TN}$, and plots the ROC curve. A higher AUC indicates better classification performance, meaning the model is more capable of separating positive and negative instances. An AUC of 0.5 indicates no ability to distinguish between the classes, equivalent to random guessing, while an AUC of 1.0 represents perfect separation, as shown in Figure 2.5.

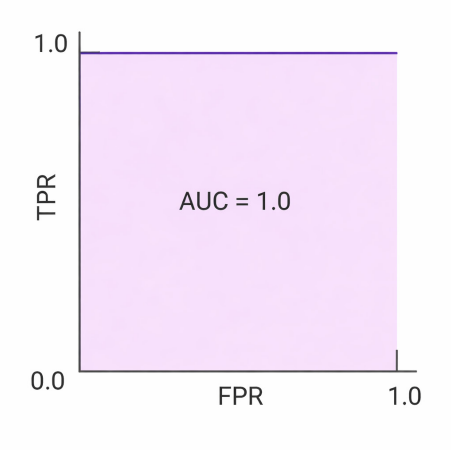


Figure 2.5: ROC curve showing a perfect classifier with an AUC of 1.0.

3

Method

The following sections describe the methodology used for the three classification pipelines: WSI-based, MRI-based and multimodal classification. For both unimodal pipelines, the general workflow consists of preprocessing the input data, extracting features using FMs and performing classification using a neural network. In the multimodal approach, the previously preprocessed and extracted features from the respective best-performing FMs are used. The resulting features are then fused and provided as input to a classifier. For each unimodal and multimodal setup, two separate classifiers were developed: one for the IDH classification task and one for the 1p/19q codeletion task. An overview of the three pipelines is illustrated in Figures 3.1, 3.2 and 3.3. The following sections describe each component of the methodology in detail.

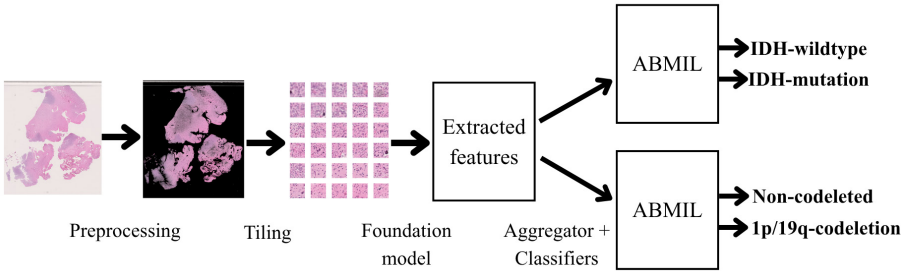


Figure 3.1: Overview of the classification pipeline using WSI.

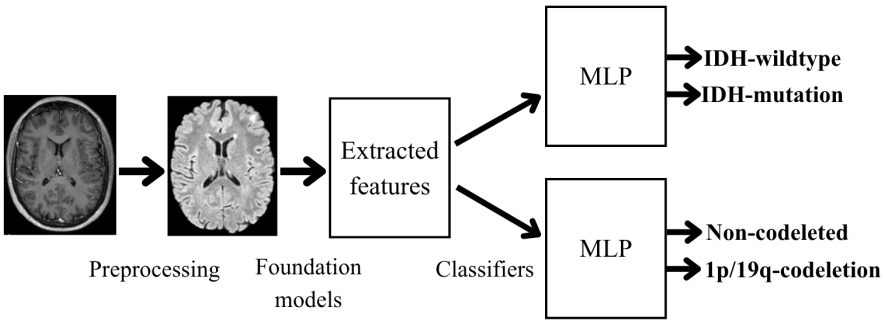


Figure 3.2: Overview of the classification pipeline using MR images.

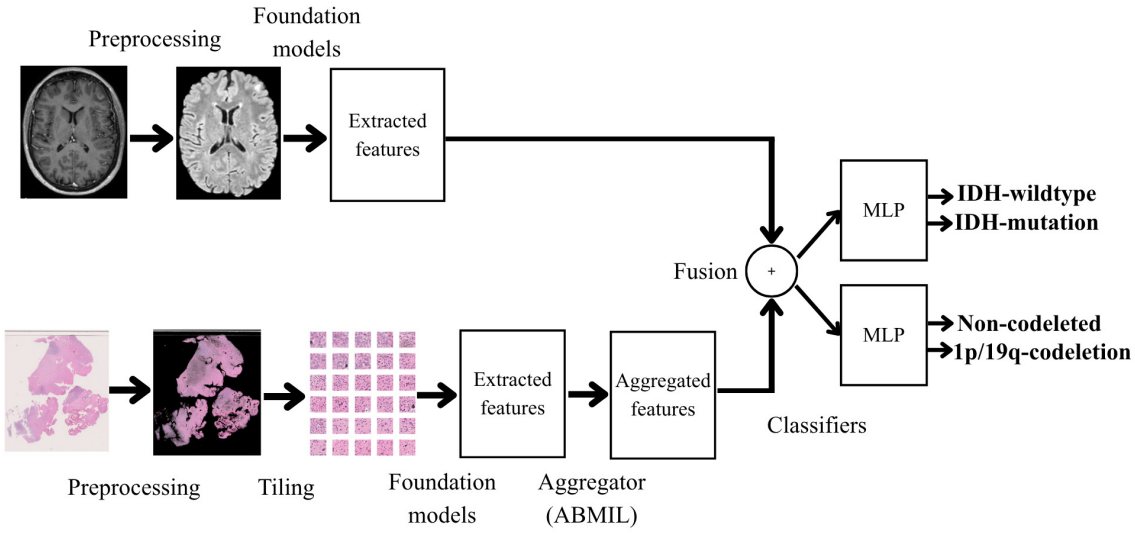


Figure 3.3: Overview of the multimodal classification pipeline using both WSIs and MR images.

3.1 Data

The WSI and MRI datasets used in this study were provided by Sahlgrenska University Hospital with ethical approval. The WSI dataset consisted of 546 patients and all patients had ground truth labels which were assigned by an expert pathologist, according to the 2021 WHO Classification of Tumors of the Central Nervous System. The patient cohort had a mean age of 54 years (range 18–82 years) and consisted of 214 female and 330 male patients.

The WSI slides were digitized using a scanner at 40× magnification. Before use, all WSIs and MR images were visually inspected to identify artifacts or noise that could negatively affect model performance. During this process, 3 WSIs were removed, resulting in a final WSI dataset of 543 patients, where one slide per patient was used. None of the MR images were removed, resulting in a final MRI dataset of 528 patients, using the T1GD MRI sequence. The multimodal dataset was created by combining the patients who had both WSI and MR images, resulting in a total of 525 patients. The class distribution was consistent for all datasets, with IDH-wildtype glioblastoma accounting for 54%, IDH-mutant astrocytoma for 26% and IDH-mutant 1p/19q codeleted oligodendroglioma for 20%. An overview of the datasets is shown in Table 3.1, and tumor localization across the brain regions and subtypes are presented in Table 3.2.

Table 3.1: Number of patients per molecular profile and subtype for the WSI, MRI and multimodal dataset.

Molecular profile	Subtype	WSI	MRI	Multimodal
IDH-wildtype	Glioblastoma	292	286	285
IDH-mutant, Non-codeleted	Astrocytoma	143	138	137
IDH-mutant, 1p/19q codeleted	Oligodendroglioma	108	104	103
Total		**543**	**528**	**525**

Table 3.2: Number of patient per tumor localization and subtype. Color intensity indicates frequency among the three most common locations (darkest = highest, lightest = third highest).

Subtype	Frontal	Temporal	Insular	Parietal	Occipital	Multifocal
Glioblastoma	54	88	6	19	4	121
Astrocytoma	75	36	9	21	2	0
Oligodendroglioma	80	12	7	8	1	0

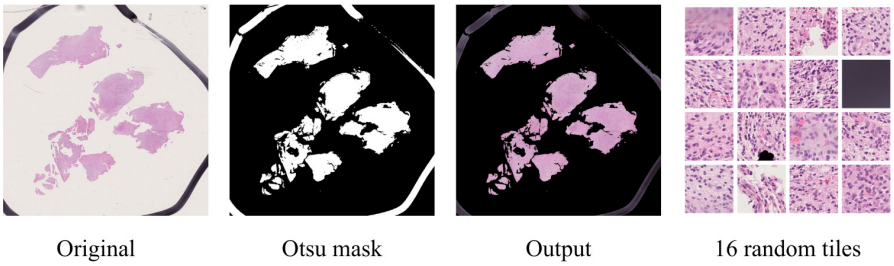
3.2 Data preprocessing

Since the FMs were trained on preprocessed images, both the WSIs and MR images were preprocessed. Following preprocessing, the images were visually reviewed again to verify that no significant artifacts had been introduced during the preprocessing pipeline. The preprocessing pipeline for each modality is described below.

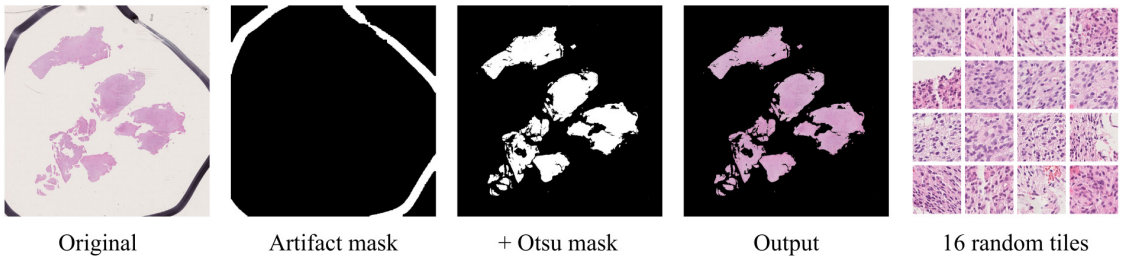
3.2.1 WSI preprocessing

The first step of the preprocessing was to extract tissue regions from each WSI, which was performed using Otsu's thresholding method. A region of interest was not selected, since the majority of the tissue regions in the WSIs consisted of tumor tissue, as reported by Sahlgrenska University Hospital. A downsampled thumbnail of each WSI was converted from RGB (red, green, blue) to HSV (hue, saturation, value) and the saturation channel was used to improve contrast between tissue and background. Otsu's method was then applied to this channel to automatically determine a threshold that separates the tissue from the background. Pixels with values above the threshold were classified as tissue, resulting in a tissue mask.

A total of 41 WSIs contained text or pen markings that interfered with Otsu's method. To remove these artifacts, an additional mask was created using the downsampled thumbnail, analyzing the HSV channels. Dark text was detected via the V channel and edge detection, while colored pen markings were identified from specific H channel ranges. The final artifact mask combined both detections and was applied before Otsu's thresholding. Figure 3.4 illustrates the results obtained with and without the artifact mask applied prior to Otsu's thresholding.



(a) Otsu mask



(b) Artifact and otsu mask

Figure 3.4: Comparison of preprocessing pipelines using only Otsu's thresholding and the combined artifact and Otsu mask for tissue extraction. For each method, the original WSI, the generated mask, the resulting masked tissue and 16 randomly sampled tiles are shown.

Based on the identified tissue regions, each WSI was divided into non-overlapping tiles and only tiles containing at least 50% tissue were selected. This threshold was chosen based on a previous study [11]. The tile size was chosen based on the input size required by each FM. For all three models, the input size was 224×224 pixels, which resulted in a total of approximately 12.2 million tiles, corresponding to approximately 22 499 tiles per WSI on average.

The tiles were analyzed based on their color intensities. For each tile the mean value and standard deviation were calculated for each RGB channel. Following this, the RGB values were normalized to the range of [0, 1] and the corresponding statistics were computed. This analysis was performed to measure the variability across the tiles and to enable comparison with the values used as input to the FM during tile-based feature extraction.

For Prov-GigaPath and UNI2-h, the official preprocessing pipelines include resizing and center-cropping to 224×224 pixels. Since the extracted WSI tiles were already 224×224, these steps did not change the tiles. Before the tiles were used as input to the chosen FMs, each tile was converted to a tensor and normalized per channel using the model-specific mean and standard deviation, listed in Table 3.3.

Table 3.3: Channel-wise RGB mean and standard deviation values used for normalizing input tiles for each foundation model and the WSI dataset.

Foundation model	Mean value	Standard deviation
H-Optimus-0	(0.707, 0.579, 0.704)	(0.212, 0.230, 0.178)
Prov-GigaPath	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)
UNI2-h	(0.485, 0.456, 0.406)	(0.229, 0.224, 0.225)
WSI dataset	**(0.878, 0.701, 0.809)**	**(0.121, 0.157, 0.108)**

3.2.2 MR images preprocessing

The preprocessing steps of the MR images were performed using the BrainIAC preprocessing script [52]. These steps included N4 bias field correction, resampling to isotropic $1 \times 1 \times 1$ mm^3 voxels, registration to MNI space and skull stripping to isolate brain tissue using the HD-BET package. For the 3DINO-ViT model, the images were additionally divided into $16 \times 16 \times 16$ tiles and normalized to the range [-1, 1].

3.3 Feature extraction

To extract features from the tiles, three different WSI FMs were used: H-optimus-0, Prov-GigaPath and UNI2-h. These models were selected based on their previously reported performance in Section 1.1. For each FM and WSI, a PyTorch file (.pt) was generated, containing the coordinates and feature vectors of all tiles in the image (see Appendix A.1 for file structure). Each vector contained 1536 elements.

To extract features from the MR images, two FMs were used: BrainIAC and 3DINO-ViT. These models were selected based on their previously reported performance in Section 1.1. The official implementations and pretrained weights were used to extract the features [52], [53]. For each FM, a PyTorch file (.pt) containing the extracted feature vectors was generated (see Appendix A.2 for file structure). Each vector from BrainIAC contained 768 elements, while the vectors from 3DINO-ViT contained 1024 elements.

3.4 Data split

The dataset was divided into training and test sets using an 85/15 stratified split, where the test set was reserved for the final evaluation of the models. The 85% training portion was used for stratified 5-fold cross-validation, meaning that it was internally split into training and validation sets during model development. To ensure a fair comparison between modalities, the held-out test set was constructed at the patient level such that the same patients were included across all datasets. Furthermore, all test patients included in the codeletion classification task were also included in the held-out test set for the IDH calssification task. This held-out test

set was never touched until the models were fully trained and the final evaluation of the models were analyzed.

For the IDH classification task, the split resulted in 463 WSI, 448 MRI and 445 combined (WSI and MRI) training and validation samples, with a held-out test set consisting of 80 samples. For the codeletion classification task, the split resulted in 215 WSI, 206 MRI and 204 combined (WSI and MRI) training and validation samples, with a held-out test set consisting of 36 samples.

3.5 Model architectures

Two separate classifiers were developed for each modality: one for the IDH classification task and one for the 1p/19q codeletion task. The models for the IDH classification task were trained on the entire dataset, while the models for the codeletion task were trained only on patients with an IDH mutation. For all models, the training samples were used in stratified 5-fold cross-validation. During training, the FM weights were frozen and and binary cross-entropy loss with logits (BCEWithLogitsLoss) was used, with class weights applied to account for class imbalance. Model parameters were optimized using the Adam optimizer and hyperparameters were tuned using Optuna in combination with cross-validation.

For each unimodal model and classification task, hyperparameter tuning was performed using stratified 5-fold cross-validation on the training set. Only the final classification layers were tuned, while the FM weights remained frozen. The mean and standard deviation of AUC and accuracy were computed across the folds. The final hyperparameters for each model and task were selected based on the highest mean AUC, with accuracy used as a complementary metric. The best-performing feature extractor was then identified based on these results. Using the selected feature extractor and its corresponding optimal hyperparameters, a final model was trained on the entire training set without cross-validation. This resulted in one final unimodal model for each modality and classification task. These final models were then used as feature extractors for the multimodal model. For the multimodal approach, three fusion techniques were evaluated for each classification task. The number of epochs used for final training of all models was chosen based on validation performance observed during cross-validation. The final models were then evaluated on the held-out test set. Below is each specific model architecture and training process described in more detail.

3.5.1 WSI feature aggregator and classifier

For the classification step, the TorchMIL ABMIL model was used as a binary classifier. The model was built as described in Section 2.5.1.1 with a dropout layer and a final linear classification layer that produced a single logit output. The extracted features from the FMs were used as input to the model, where only the ABMIL model was trained, while the FM weights remained frozen throughout both hyperparameter tuning and training. Optuna was used to find the best hyperparameters, which were selected based on the highest AUC. Hyperparameter tuning was performed for each

model and for the three different FMs: H-Optimus-0, Prov-GigaPath and UNI2-h. Optuna was run for 20 trials and 40 epochs, tuning the learning rate, weight decay, attention dimension, dropout, use of gated attention and the activation function in the attention mechanism (see Appendix B.1 for all ranges). The best Optuna-tuned hyperparameters are shown in Table 3.4.

Table 3.4: Final optuna-tuned hyperparameters for IDH and Codeletion models across the three WSI foundation models.

Task	Model	Activation	Attention dim	Dropout	Gated	Learning rate	Weight decay
IDH	H-optimus-0	relu	512	0.2742	True	7.910×10^{-4}	1.171×10^{-5}
IDH	Prov-GigaPath	tanh	384	0.3610	True	9.820×10^{-4}	9.110×10^{-4}
IDH	UNI2-h	tanh	256	0.06314	False	9.544×10^{-4}	1.642×10^{-5}
Codeletion	H-optimus-0	tanh	384	0.1732	True	4.403×10^{-4}	2.610×10^{-5}
Codeletion	Prov-GigaPath	relu	256	0.4318	True	7.084×10^{-5}	1.880×10^{-4}
Codeletion	UNI2-h	gelu	256	0.4890	False	9.814×10^{-4}	9.250×10^{-4}

After hyperparameter optimization with Optuna, the learning rate was further manually fine-tuned to minimize validation loss, resulting in a final value of 1×10^{-5}. The models for the IDH classification task were trained for 25 epochs, while the models for the codeletion classification task were trained for 40 epochs.

3.5.2 MRI classifier

For the classification step, a MLP with a single fully connected layer and a sigmoid output neuron was used for binary classification. The network consists of a linear transformation, activation function and dropout. The extracted features from the MRI FMs were used as input to train the MLP classification, while the FM weights remained frozen throughout hyperparameter tuning and training. Hyperparameter tuning was performed for each task and for the two different FMs: BrainIAC and 3DINO-ViT. The parameters tuned over 80 trials were learning rate, weight decay, batch size, dropout, activation function, hidden dimensions and the number of epochs (see Appendix B.2 for all ranges). Early stopping was applied with patience of 25% of the total number of epochs. The best hyperparameters and extracted features were selected based on the highest AUC. The best Optuna-tuned hyperparameters are shown in Table 3.5.

Table 3.5: Final optuna-tuned hyperparameters for IDH and Codeletion task across the two MRI foundation models.

Task	Model	Activation	Hidden dim	Dropout	Epochs	Batch size	Learning rate	Weight decay
IDH	BrainIAC	tanh	512	0.415	91	32	4.42×10^{-5}	4.73×10^{-6}
IDH	3DINO-ViT	gelu	128	0.347	64	64	1.50×10^{-4}	1.90×10^{-3}
Codeletion	BrainIAC	gelu	256	0.288	50	64	3.80×10^{-5}	6.90×10^{-3}
Codeletion	3DINO-ViT	relu	64	0.495	97	64	2.92×10^{-4}	4.42×10^{-3}

The final models for both tasks, IDH and codeletion, were trained for 14 epochs.

3.5.3 Multimodal fusion strategy and classifier

The WSIs and MR images were processed using the same preprocessing and feature extraction steps as described in Section 3.2 and 3.3. The best-performing FM for each modality was selected based on validation performance. This resulted in WSI features from H-Optimus-0 for the IDH task, WSI features from Prov-GigaPath for the codeletion task and MRI features from 3DINO-ViT for both IDH and codeletion task.

To aggregate the tile-level feature vectors into a single WSI feature vector, the final WSI models for the IDH and codeletion tasks were used. The weights were frozen and only the ABMIL feature aggregation component was used. This resulted in one feature vector from each modality per patient, which was used in the subsequent fusion step. This resulted in one WSI feature vector per patient (1536 elements) and one MRI feature vector per patient (1024 elements). A PyTorch file (.pt) containing the extracted feature vectors was generated (see Appendix A.3 for file structure).

The extracted features from the WSI and MR images were fused using three techniques: concatenation, cross-modal attention and projected concatenation. The models with cross-modal attention and projected concatenation, first projected the WSI and MRI features to the same dimension and then concatenated the vectors. In contrast, the model with only concatenation, concatenated the vectors to one dimension directly:

$$\mathbf{z} = \text{concat}(\mathbf{x}_{\text{wsi}}, \mathbf{x}_{\text{mri}})$$

In the cross-modal attention strategy, the concatenated features were further processed by an attention mechanism. This mechanism generated modality-specific weights using a single fully connected layer consisting of a linear transformation, activation function and dropout, followed by a final linear layer with a softmax producing two normalized weights. These weights were then used in a softmax and then multiplied with each modality feature vector, which resulted in the final feature vector:

$$\mathbf{z} = w_{wsi} \cdot \mathbf{x}_{wsi} + w_{mri} \cdot \mathbf{x}_{mri}$$

The feature vectors were subsequently used as input to an MLP with one single fully connected layer and a sigmoid output neuron was used for binary classification. The network consisted of a linear transformation, activation function dropout and a final linear layer projecting to one output. Hyperparameter tuning was performed for each model over 80 trials, including learning rate, weight decay, batch size, dropout, activation function, hidden dimension and number of epochs (see Appendix B.3 for all ranges). Early stopping was applied during training, with a patience of 25% of the total number of epochs. The best hyperparameters and extracted features were selected based on the highest AUC averaged over the cross-validation sets, which are shown in Table 3.6.

Table 3.6: Final optuna-tuned hyperparameters for IDH and codeletion task across the three multimodal fusion techniques.

Task	Fusion	Activation	Hidden dim	Dropout	Epochs	Batch size	Learning rate	Weight decay
IDH	Concatenation	relu	128	0.442	111	16	3.05×10^{-6}	5.03×10^{-3}
IDH	Attention	tanh	128	0.296	50	16	2.28×10^{-6}	2.04×10^{-3}
IDH	Projected concatenation	gelu	256	0.0049	78	64	9.29×10^{-6}	2.03×10^{-2}
Codeletion	Concatenation	gelu	128	0.311	72	16	3.43×10^{-6}	4.41×10^{-3}
Codeletion	Attention	relu	256	0.498	102	16	5.28×10^{-6}	8.65×10^{-2}
Codeletion	Projected concatenation	gelu	128	0.220	171	16	5.81×10^{-6}	6.28×10^{-4}

For the concatenation fusion technique 15 epochs were used for both classification tasks. For the cross-modal attention fusion technique 40 epochs were used for IDH task and 15 for the codeletion task. For the concatenation with projection fusion technique 30 epochs were used for the IDH task and 15 for the codeletion task.

3.6 Evaluation of the models

The models were evaluated using a held-out test set. For the IDH classification task, IDH-mutant cases were defined as positive class and IDH-wildtype cases as negative class. For the codeletion task, 1p/19q-codeleted cases were defined as positive class and non-codeleted cases as negative class. The following performance metrics were calculated: accuracy, AUC, PPV, NPV, specificity and recall, as defined in Section 2.6.

To further evaluate the model performance, heatmaps were generated for the two WSI models. These heatmaps were created using the attention weights from ABMIL, applied to each patch with a corresponding spatial coordinate. In the heatmaps, lighter yellow colors represent tiles assigned higher importance by the model, whereas darker purple colors indicate lower importance. The heatmaps were compared to the original images to identify general regions that the model focused on. Additionally, attention weights from the multimodal model with cross-modal attention were analyzed to evaluate how the model weighted features from WSI and MRI, respectively. Since the unimodal MRI model produced only a single feature vector per image, it was not possible to generate heatmaps to analyze which regions of the MR image contributed most to the model's decisions. The same limitation applied to the multimodal models using standard concatenation and projected concatenation, as these architectures did not preserve spatially interpretable attention information.

For all models, the sigmoid classification outputs (i.e., predicted probabilities) were analyzed to assess the model confidence. For the IDH task, outputs were grouped into correctly and incorrectly classified IDH-mutant cases, as well as correctly and incorrectly classified IDH-wildtype cases. Similarly, for the codeletion task, outputs were grouped into correctly and incorrectly classified non-codeleted and 1p/19q-codeleted cases. The predicted probabilities were then plotted according to their respective groups. In an ideal scenario, correctly classified samples would yield probabilities close to 1, whereas incorrectly classified samples would yield probabilities closer to 0.5.

4

Results

The following sections describe the results for the three classification pipelines: WSI-based, MRI-based and multimodal classification. The results are described for each task (IDH and codeletion) and lastly more detailed result about the model attention and confidence.

The total time required for feature extraction from the WSIs using the FMs was approximately 80 hours. In contrast, training the classifier (ABMIL) required only about 30 minutes. The total time required for feature extraction from the MRI images using the FMs and preprocessing was approximately 30 hours. In contrast, training the classifier (MLP) required only about 30 seconds. The total time required training the classifier (MLP) for the multimodal model required only about 30 seconds.

4.1 IDH classification

The following sections present the results for the IDH classification task. The final unimodal models used the best feature extractor combined with the optimal parameters obtained from Optuna.

4.1.1 Training and validation

For the WSI-based models, the mean and standard deviation across the five folds are presented under *WSI* in Table 4.1. The model based on H-optimus-0 achieved the highest performance, with an AUC of 0.961 ± 0.028 and an accuracy of 0.903 ± 0.012. The Prov-GigaPath model achieved second highest performance, while UNI2-h showed the lowest performance among the evaluated WSI-based models.

For the MRI-based models, the corresponding results are shown under *MRI* in Table 4.1. 3DINO-ViT achieved the best performance, with an AUC of 0.856 ± 0.015 and an accuracy of 0.781 ± 0.031. BrainIAC achieved lower performance and exhibited higher variation across folds compared to 3DINO-ViT.

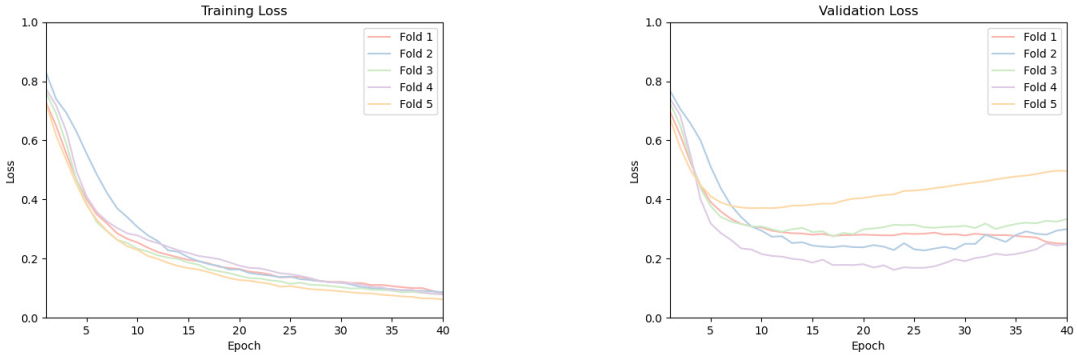
For the multimodal models, constructed using the best-performing unimodal models, the results for the three fusion strategies are presented under *Multimodal* in Table 4.1. All multimodal models achieved higher performance compared to the unimodal approaches. The cross-modal attention model achieved the highest AUC

of 0.985 ± 0.016, while all multimodal models achieved an accuracy of 0.953 ± 0.025. The multimodal models showed small differences in performance between fusion strategies.

Table 4.1: Mean and standard deviation of validation AUC and accuracy with the optimized parameters for the IDH classification across WSI, MRI and multimodal pipelines.

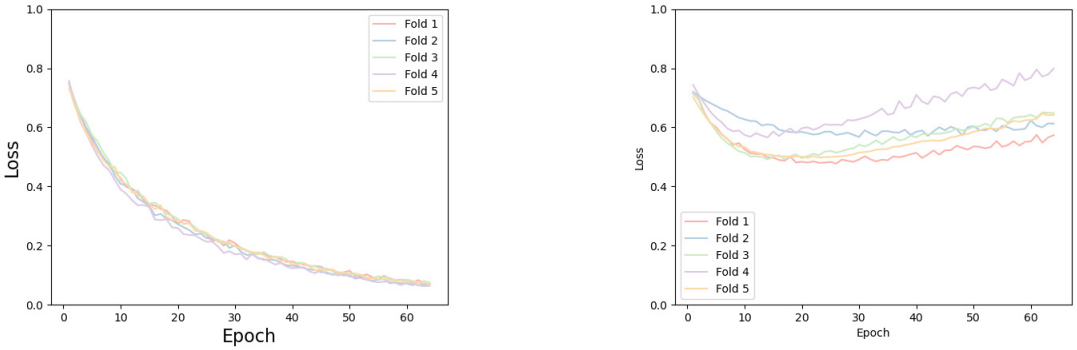
Foundation model	**Fusion**	**AUC** ($\mu \pm \sigma$)	**Accuracy** ($\mu \pm \sigma$)
WSI			
H-optimus-0	–	**0.961 ± 0.028**	**0.903 ± 0.012**
Prov-GigaPath	–	0.960 ± 0.043	0.898 ± 0.053
UNI2-h	–	0.947 ± 0.021	0.883 ± 0.032
MRI			
BrainIAC	–	0.732 ± 0.068	0.661 ± 0.066
3DINO-ViT	–	**0.856 ± 0.015**	**0.781 ± 0.031**
Multimodal			
H-optimus-0, 3DINO-ViT	Concatenation	0.984 ± 0.014	**0.953 ± 0.025**
H-optimus-0, 3DINO-ViT	Attention	**0.985 ± 0.016**	**0.953 ± 0.025**
H-optimus-0, 3DINO-ViT	Projected concatenation	0.984 ± 0.015	**0.953 ± 0.025**

The training and validation loss for each fold for the best-performing unimodal feature extractors are shown in Figures 4.1 and 4.2. The rest of the FMs training and validation loss plots are shown in Appendix C. As observed in the WSI validation loss plots, fold 5 has higher validation loss than the other folds across all three FM models. Otherwise the validation loss tends to plateau after 7 epochs. The training loss decreases sharply between 0-7 epochs and then decreases more slowly. As observed in the MRI validation loss plot, the minimum loss occurs at epoch 18 for Fold 1, 3 and 5, and around epoch 12 and 30 for Fold 4 respectively for Fold 2. After reaching these minima, the validation loss increases except for Fold 2, which plateaus. The MRI training loss plot shows a smooth, approximately exponential decay across all folds.



(a) Training loss. (b) Validation loss.

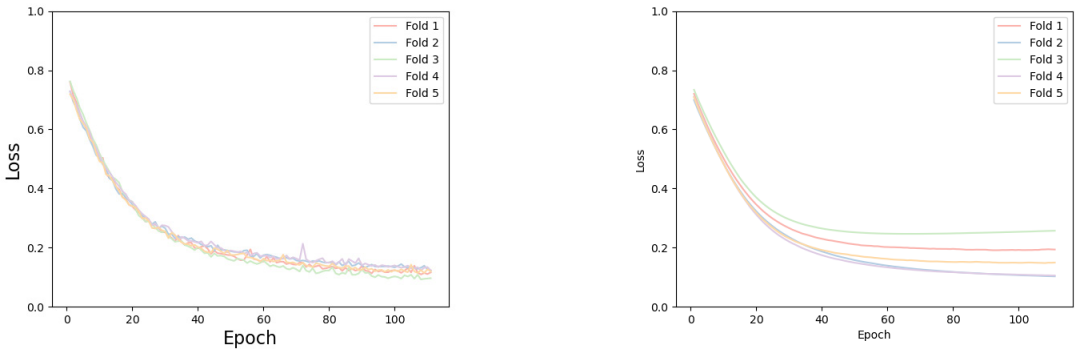
Figure 4.1: Training and validation loss for the IDH classification task using the WSI model with features extracted from H-Optimus-0.



(a) Training loss. (b) Validation loss.

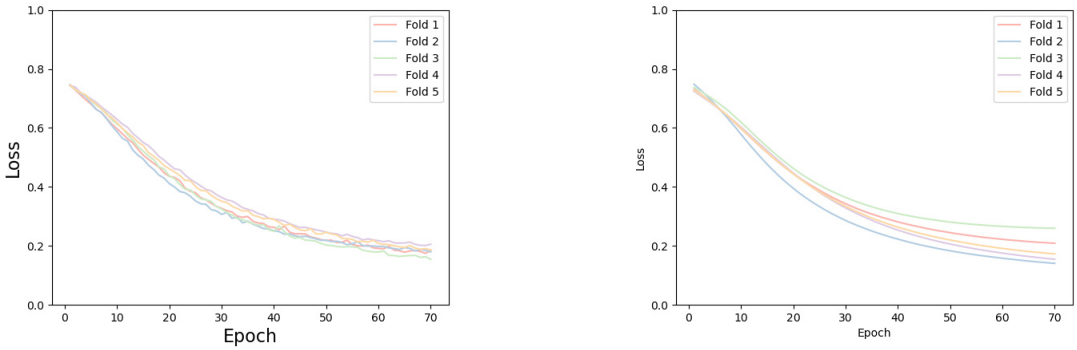
Figure 4.2: Training and validation loss for the IDH classification task using the MRI model with features extracted from 3DINO-ViT.

The training and validation loss for each fold for the three used fusion techniques used in the multimodal models, are shown in Figures 4.3–4.5. The H-optimus-0 was used as feature extractor for the WSIs and 3DINO-ViT for the MR images. As observed in the validation loss plots the multimodal with both normal concatenation and projected concatenation plateaus after 40 epochs and the multimodal with cross-modal attention after 50 epochs. Fold 3 has the highest final validation loss and Fold 2 has the lowest, in all plots. The training loss decreases exponentially and follows the same pattern as the validation loss curve.



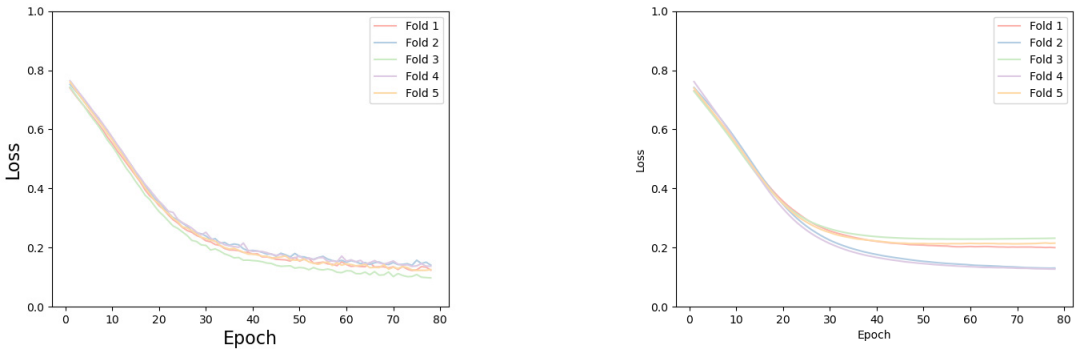
(a) Training loss. (b) Validation loss.

Figure 4.3: Training and validation loss for the IDH classification task using the multimodal concatenation model, with WSI features extracted from H-Optimus-0 and MRI features extracted from 3DINO-ViT



(a) Training loss. (b) Validation loss.

Figure 4.4: Training and validation loss for the IDH classification task using the multimodal cross-modal attention model, with WSI features extracted from H-Optimus-0 and MRI features extracted from 3DINO-ViT.



(a) Training loss.

(b) Validation loss.

Figure 4.5: Training and validation loss for the IDH classification task using the multimodal projected concatenation model, with WSI features extracted from H-Optimus-0 and MRI features extracted from 3DINO-ViT.

4.1.2 Held-out test set results

The results for the final models for the IDH classification task, evaluated on the held-out test set, are presented in Table 4.2. The multimodal model using standard concatenation achieved the highest performance across all metrics except AUC. The model achieved an accuracy of 0.938, PPV of 0.919, NPV of 0.953, specificity of 0.932 and recall of 0.944. The second-best model was the multimodal model with cross-modal attention which obtained the highest AUC (0.965) and the third-best model was the WSI model. The WSI model and multimodal models using cross-modal attention and standard concatenation achieved the same recall of 0.944. The multimodal model based on projected concatenation performed similarly to the WSI model across all metrics, except for recall where it achieved a lower value (0.925 compared to 0.944). The MRI model showed the lowest performance across all evaluated metrics. The MRI model and WSI model achieved accuracy and AUC that were consistent with, but slightly higher than, the performance observed during training (Table 4.1). On the other hand, all multimodal models achieved both higher AUC and accuracy during training.

The overall performance of the models was relatively similar, except for the MRI model. In general, the multimodal models achieved the best performance although the WSI model also showed strong results. The MRI model demonstrated the lowest overall performance. For all models except MRI, recall was higher than specificity, indicating that the models were more effective at identifying positive cases (IDH mutation) than negative cases (IDH-wildtype). In contrast, the MRI model showed higher specificity than recall, indicating that it was more effective at identifying IDH-wildtype cases than IDH-mutant cases. Additionally, for all models including MRI, NPV was higher than PPV, indicating that the models were more reliable when predicting negative cases (IDH-wildtype) than positive cases (IDH mutation).

Table 4.2: Comparison of IDH classification performance across WSI, MRI and multimodal models with concatenation (cat), cross-modal attention (attn) and projected concatenation (proj cat). IDH mutation is defined as positive classification and IDH-wildtype is defined as negative classification.

Metric	WSI	MRI	MM (cat)	MM (attn)	MM (proj cat)
Accuracy	0.925	0.788	**0.938**	0.925	0.925
AUC	0.964	0.852	0.960	**0.965**	0.964
PPV	0.895	0.771	**0.919**	0.895	0.895
NPV	0.952	0.800	**0.953**	0.952	0.952
Specificity	0.909	0.818	**0.932**	0.909	0.909
Recall	**0.944**	0.750	**0.944**	**0.944**	0.925

4.2 Codeletion classification

The following sections present the results for the codeletion classification task. The final unimodal models used the best feature extractor combined with the optimal parameters obtained from Optuna.

4.2.1 Training and validation

For the WSI-based models, the mean and standard deviation across the five folds are presented under WSI in Table 4.3. The model based on Prov-GigaPath achieved the highest AUC 0.990 ± 0.011 and accuracy 0.930 ± 0.015. UNI2-h achieved the lowest performance, particularly in terms of accuracy, while H-optimus-0 performed between these two models.

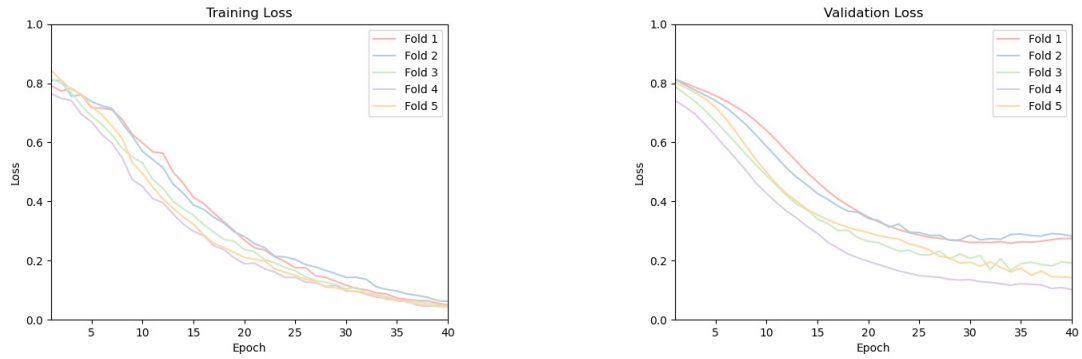
For the MRI-based model, the corresponding results are shown under MRI in Table 4.3. 3DINO-ViT achieved the best performance with an AUC of 0.702 ± 0.077 and accuracy of 0.627 ± 0.044. BrainIAC achieved substantially lower performance, with an AUC close to 0.5. Compared to the WSI-based models, the MRI-based models achieved lower performance overall.

For the multimodal models, the best-performing FMs from the unimodal experiments were combined using three different fusion strategies. The results are presented under Multimodal in Table 4.3. All multimodal models achieved identical AUC values of 1.000 ± 0.000, while the cross-modal attention approach obtained the highest mean accuracy of 0.995 ± 0.010.

Table 4.3: Mean and standard deviation of validation AUC and accuracy with the optimized parameters for the codeletion classification across WSI, MRI and multimodal pipelines.

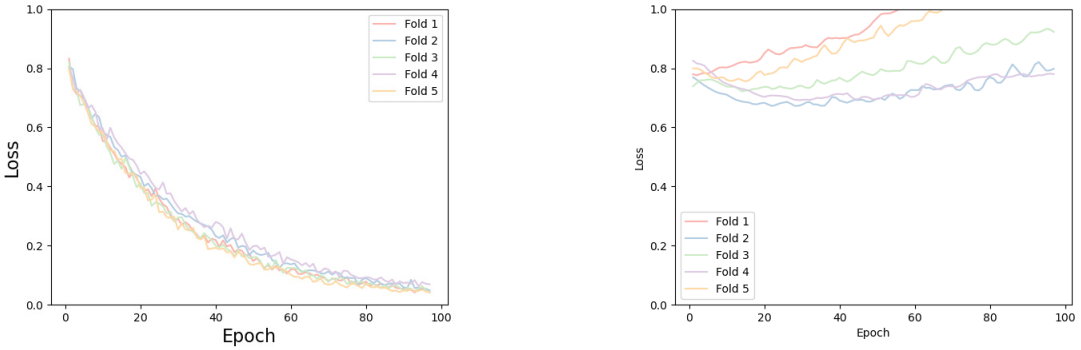
Foundation model	**Fusion**	**AUC** ($\mu \pm \sigma$)	**Accuracy** ($\mu \pm \sigma$)
WSI			
H-optimus-0	–	0.949 ± 0.030	0.916 ± 0.038
Prov-GigaPath	–	**0.990 ± 0.011**	**0.930 ± 0.015**
UNI2-h	–	0.939 ± 0.041	0.878 ± 0.050
MRI			
BrainIAC	–	0.515 ± 0.119	0.525 ± 0.110
3DINO-ViT	–	**0.702 ± 0.077**	**0.627 ± 0.044**
Multimodal			
Prov-GigaPath + 3DINO-ViT	Concatenation	**1.000 ± 0.000**	0.990 ± 0.012
Prov-GigaPath + 3DINO-ViT	Attention	**1.000 ± 0.000**	**0.995 ± 0.010**
Prov-GigaPath + 3DINO-ViT	Projected concatenation	**1.000 ± 0.000**	0.995 ± 0.098

The training and validation loss for each fold for the best-performing unimodal feature extractors are shown in Figures 4.6 and 4.7. The rest of the unimodal FMs training and validation loss plots are shown in Appendix C. As observed in the WSI validation loss plot, it plateaus after 23 epochs. Compared to the other FMs plot, the validation loss varies between the folds but none of them increase after the plateau. The training loss decreases exponentially. As observed in the MRI validation loss plot, the minimum loss occurs around epoch 18–22 for Fold 2, 3, 4 and 5, and around epoch 2 for Fold 1. After these points, the validation loss increases, while the training loss continues to decrease exponentially.



(a) Training loss. (b) Validation loss.

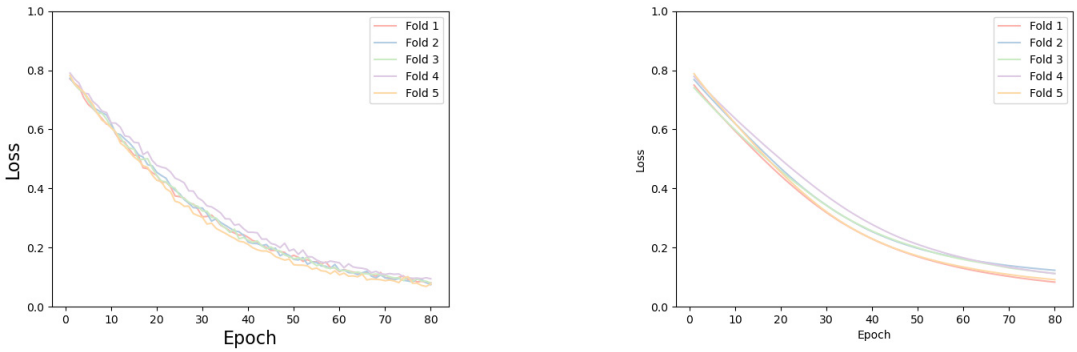
Figure 4.6: Training and validation loss for the codeletion classification task using the WSI model with features extracted from Prov-gigapath.



(a) Training loss. (b) Validation loss.

Figure 4.7: Training and validation loss for the codeletion classification task using the MRI model with features extracted from 3DINO-ViT.

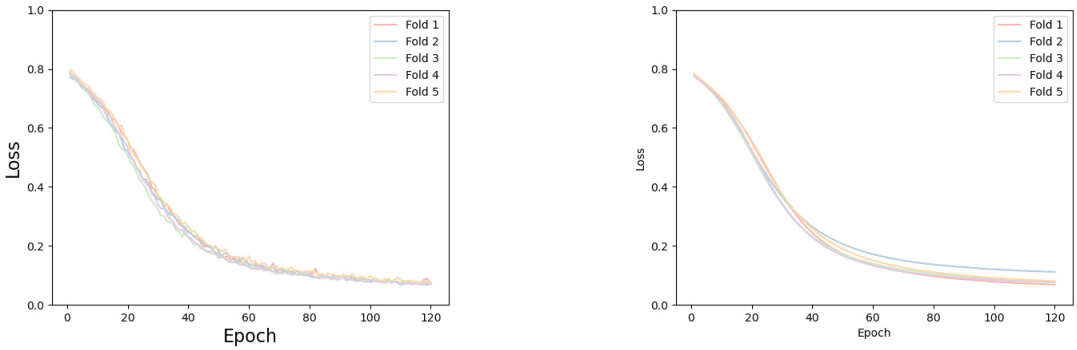
The training and validation loss for each fold for the three used fusion techniques used in the multimodal models, are shown in Figures 4.8–4.9. The Prov-GigaPath was used as feature extractor for the WSI and 3DINO-ViT for the MR images. As observed in the validation loss plots the multimodal with standard concatenation decrease exponentially, cross-modal attention and projected concatenation plateaus after 50 epochs. All folds follows each other and the training loss follows the same pattern as the validation loss curve.



(a) Training loss.

(b) Validation loss.

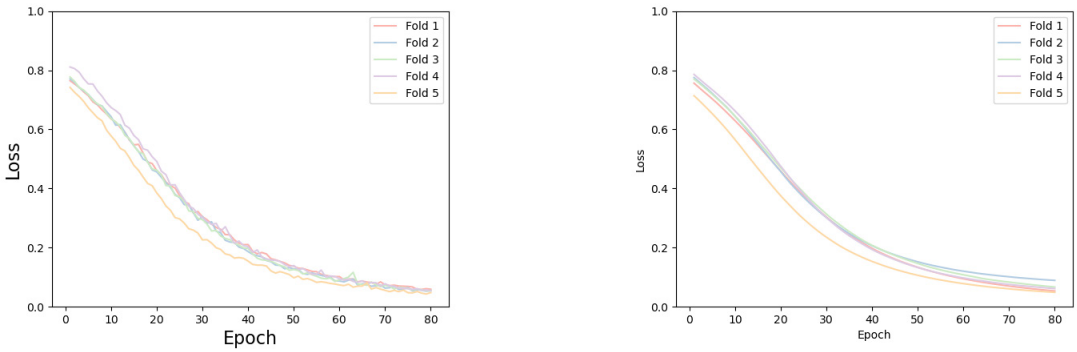
Figure 4.8: Training and validation loss for the codeletion classification task using the multimodal concatenation model, with WSI features extracted from Prov-GigaPath and MRI features extracted from 3DINO-ViT.



(a) Training loss.

(b) Validation loss.

Figure 4.9: Training and validation loss for the codeletion classification task using the multimodal cross-modal attention model, with WSI features extracted from Prov-GigaPath and MRI features extracted from 3DINO-ViT.



(a) Training loss.

(b) Validation loss.

Figure 4.10: Training and validation loss for the codeletion classification task using the multimodal projected concatenation model with, WSI features extracted from Prov-GigaPath and MRI features extracted from 3DINO-ViT.

4.2.2 Held-out test set results

The results for the final models for the codeletion classification task, evaluated on the held-out test set, are presented in Table 4.4. The WSI model and the three multimodal models performed similarly across all metrics, except for AUC. These models achieved an accuracy of 0.917, PPV of 0.929, NPV of 0.909, specificity of 0.952 and recall of 0.867. The highest AUC (0.987), and therefore also the highest overall performance, was achieved by the multimodal model using projected concatenation. The second-best models were the multimodal model with cross-modal attention, followed by the WSI-based model and then the multimodal model with standard concatenation. The MRI model showed the lowest overall performance across most metrics, but achieved the same recall (0.867) as the other models.

The WSI model and all multimodal models achieved lower AUC and accuracy on the test set compared to the training results shown in Table 4.3. In contrast, the MRI model achieved higher AUC and accuracy on the test set than during training.

Table 4.4: Comparison of codeletion classification performance across WSI, MRI and multimodal models with concatenation (cat), cross-modal attention (attn) and projected concatenation (proj cat). 1p/19q codeletion is defined as positive classification and non-codeletion is defined as negative classification.

Metric	WSI	MRI	MM (cat)	MM (attn)	MM (proj cat)
Accuracy	**0.917**	0.722	**0.917**	**0.917**	**0.917**
AUC	0.979	0.740	0.949	0.981	**0.987**
PPV	**0.929**	0.619	**0.929**	**0.929**	**0.929**
NPV	**0.909**	0.867	**0.909**	**0.909**	**0.909**
Specificity	**0.952**	0.619	**0.952**	**0.952**	**0.952**
Recall	**0.867**	**0.867**	**0.867**	**0.867**	**0.867**

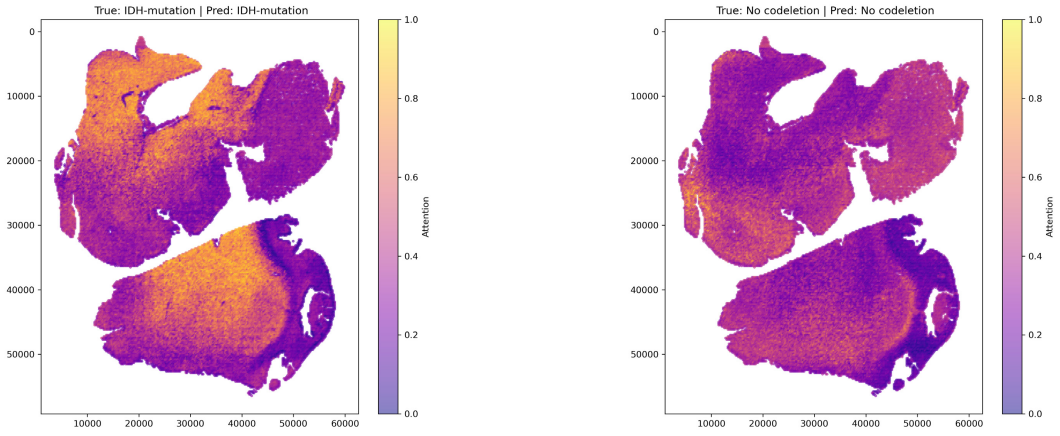
Overall, the models showed similar performance except for the MRI model, which achieved significantly lower performance compared with the other models. The multimodal models demonstrated only slight improvements, while the WSI model alone already achieved strong performance across most evaluation metrics. Among all models, the MRI model showed the lowest overall performance. For all models except the MRI model, specificity was higher than recall, indicating better identification of negative cases (non-1p/19q codeletion) than positive cases (1p/19q codeletion). In contrast, the MRI model showed higher recall than specificity, indicating improved identification of positive cases relative to negative cases. Additionally, PPV was higher than NPV across all models, indicating that positive predictions (1p/19q codeletion) were generally more reliable than negative predictions (non-1p/19q codeletion).

4.3 Model analysis

The following sections present the results from the final models evaluated on the held-out test set, including attention-based interpretations of the WSI models, analysis of attention weights from the multimodal model with cross-modal attention and an evaluation of model confidence.

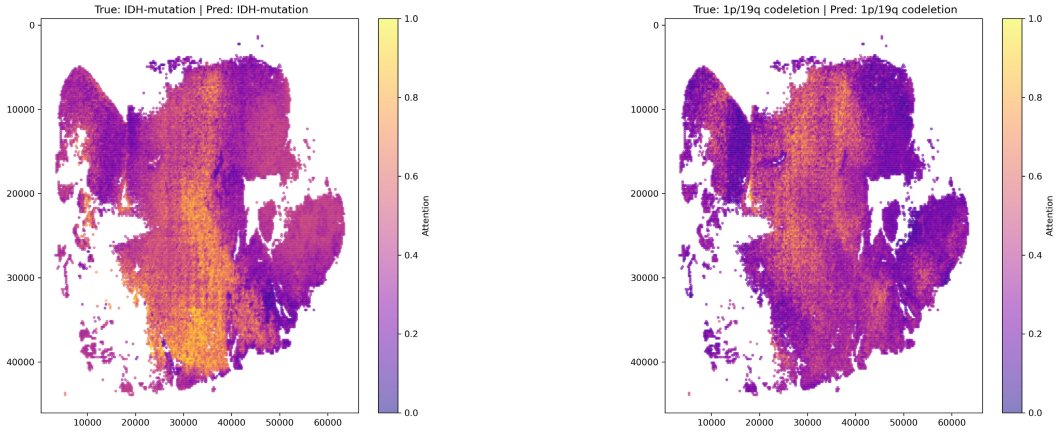
4.3.1 Attention analysis

Figures 4.11 and 4.12 show attention heatmaps for two patients correctly classified with IDH-mutated glioma. Figure 4.11 corresponds to the patient with IDH mutation and no 1p/19q codeletion (astrocytoma), while Figure 4.12 corresponds to the patient with IDH mutation and 1p/19q codeletion (oligodendroglioma). The figures illustrate the attention distributions produced by the WSI-based models for the two classification tasks: IDH prediction and 1p/19q-codeletion prediction. As seen in the figures, the models for the two tasks often focus on overlapping tumor regions for the same patient, but this is not consistently observed across all areas. In some regions, the models attended to different parts of the tissue. Furthermore, the model for the IDH task show more concentrated high-activation regions as seen in Figure 4.11a and Figure 4.12a, whereas the model for the codeletion task displays smoother and more diffuse activation patterns as observed in Figure 4.11b and Figure 4.12b. The high activation regions were observed in areas with high cellular density in the original WSI.



(a) IDH-mutation. (b) Non-codeleted.

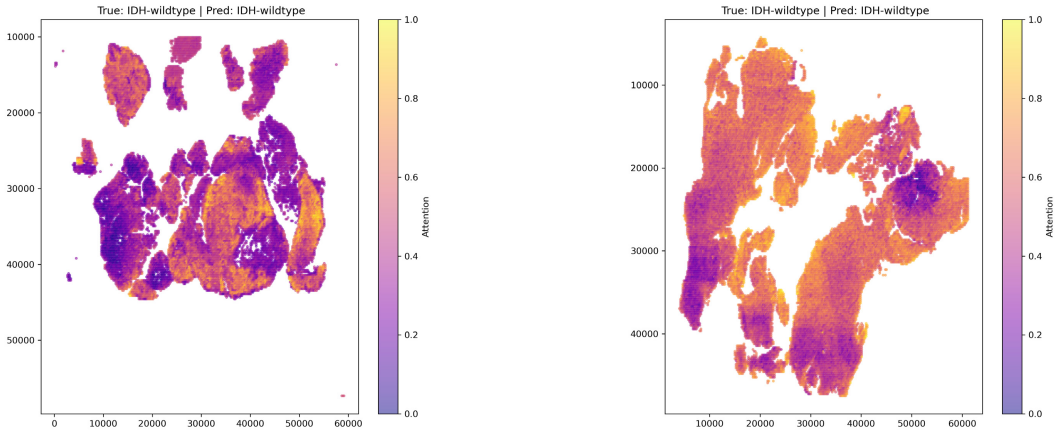
Figure 4.11: Heatmaps generated by the two WSI models for the same patient from the held-out test set. The patient was correctly classified as IDH-mutant and non-codeleted.



(a) IDH-mutation. (b) 1p/19q-codeletion.

Figure 4.12: Heatmaps generated by the two WSI models for the same patient from the held-out test set. The patient was correctly classified as IDH-mutant and 1p/19q-codeleted.

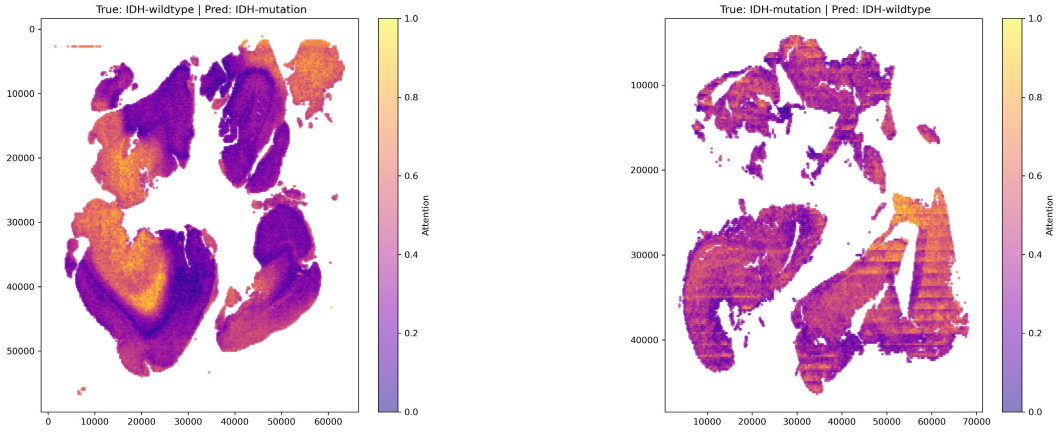
Figure 4.13 shows the resulting heatmaps for two patients correctly classified as IDH-wildtype (glioblastoma). For these patients the activation patterns appeared more diffuse and evenly distributed, with fewer distinct high-activation regions, as seen in Figure 4.13a. However, other cases also exhibit stronger high-activation regions, as illustrated in Figure 4.13b. As in the IDH mutation cases, high-activation regions were observed in areas with high cellular density in the original WSI.



(a) IDH-wildtype. (b) IDH-wildtype.

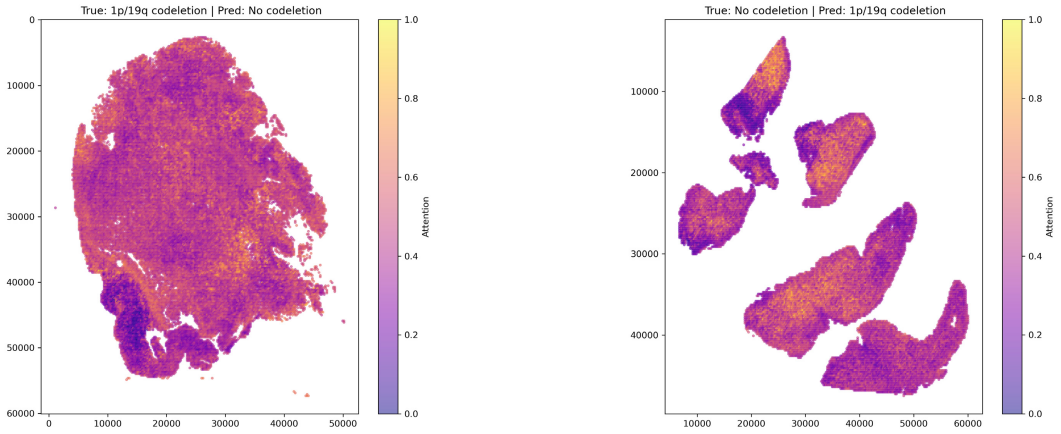
Figure 4.13: Heatmaps for two patients from the held-out test set, correctly classified with IDH-wildtype

Figure 4.14 and Figure 4.15 show heatmaps for patients that were incorrectly classified. Figure 4.14 shows more concentrated high-activation patterns, whereas Figure 4.15 shows smoother and more distributed activation patterns. However, visual inspection of all misclassified samples did not reveal any clear or consistent localization patterns that distinguish them from the correctly classified cases.



(a) Incorrect IDH-mutation. (b) Incorrect IDH-wildtype.

Figure 4.14: Heatmaps for two patients from the held-out test set, incorrectly classified with IDH-mutation and IDH-wildtype.



(a) Incorrect non-codeleted.

(b) Incorrect 1p/19q-codeletion.

Figure 4.15: Heatmaps for two patients from the held-out test set, incorrectly classified with no codeletion and 1p/19q-codeletion.

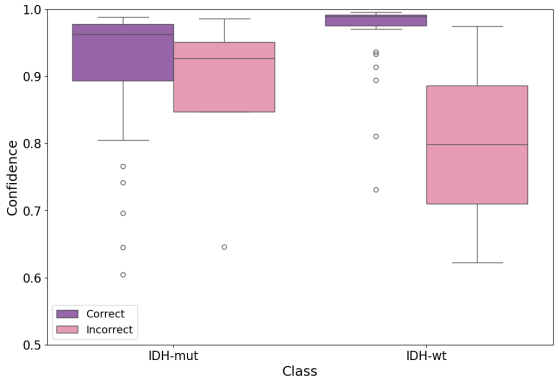
The attention weights from the multimodal model with cross-modal attention were analyzed to evaluate how the model weighted features from WSI and MRI, respectively. For IDH classification, the average weights were 0.909 for WSI and 0.091 for MRI, whereas the corresponding weights for codeletion classification were 0.670 and 0.330, respectively. Although WSI features dominated in both tasks, the contribution from MRI features was notably greater for the codeletion task compared to the IDH task.

4.3.2 Model confidence

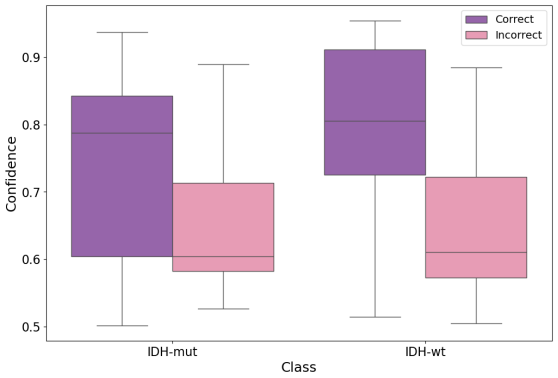
Figure 4.16 shows boxplots of the confidence distributions for all models for the IDH classification task, where the circles represent outliers. Higher probabilities correspond to higher model confidence. Overall, all models demonstrate higher confidence for correctly classified IDH-wildtype samples compared to correctly classified IDH-mutated samples. This is consistent with the higher NPV compared to PPV observed for the IDH task. Similarly, incorrectly classified IDH-wildtype samples generally show lower confidence values than incorrectly classified IDH-mutated samples. The WSI-based model exhibits the highest overall confidence and the clearest separation between correctly and incorrectly classified IDH-wildtype samples. In contrast, the MRI-based model shows lower confidence values overall, but demonstrates the largest separation between correctly and incorrectly classified IDH-mutated samples. This model also displays the largest variability, with a broader spread of confidence values.

For the multimodal models, different confidence patterns are observed. The model based on standard concatenation shows limited separation between correctly and incorrectly classified IDH-mutated samples, with higher confidence for incorrectly classified cases. However, a clearer separation is observed for the IDH-wildtype class.

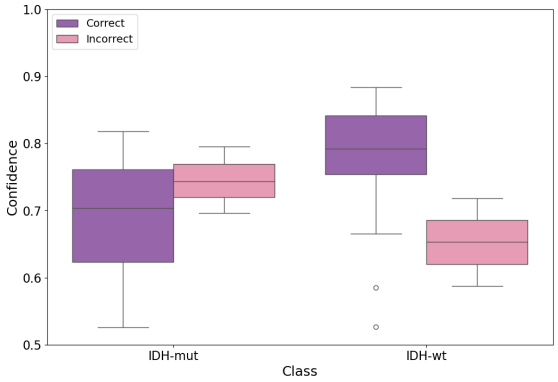
The cross-modal attention and projected concatenation models exhibit similar overall behavior, although the projected concatenation model demonstrates a narrower distribution of confidence values.



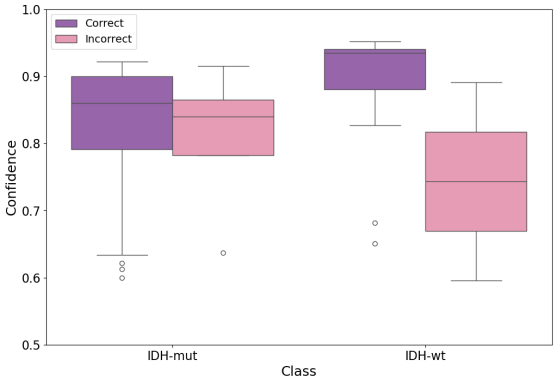
(a) WSI.



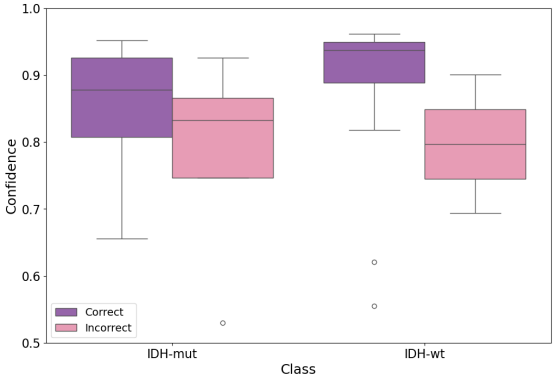
(b) MRI.



(c) Multimodal with concatenation.



(d) Multimodal with cross-modal attention.



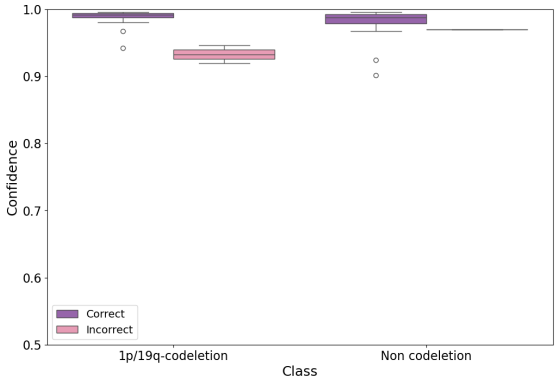
(e) Multimodal with projected concatenation.

Figure 4.16: Confidence distributions for the IDH classification task across all models, shown as boxplots of probabilities grouped by correct and incorrect classifications for each class.

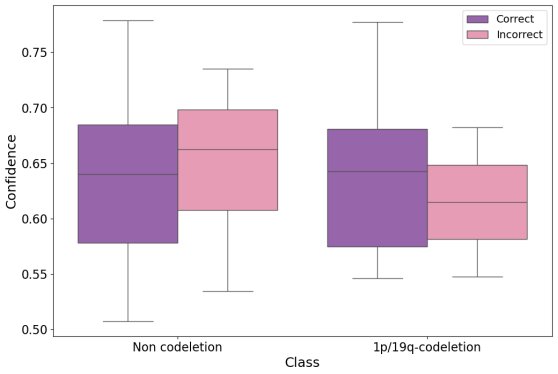
Figure 4.17 shows boxplots of the confidence distributions for all models for the

codeletion classification task. Higher probabilities correspond to higher model confidence, and the circles represent outliers. In general, correctly classified 1p/19q-codeleted samples exhibit higher confidence values than correctly classified non-codeleted samples, although this difference is relatively small for the MRI-based and cross-modal attention models. This is consistent with the higher PPV compared to NPV observed for the codeletion task. Furthermore, all models show lower confidence for incorrectly classified 1p/19q-codeleted samples compared to incorrectly classified non-codeleted samples. It can also be observed that, for all plots except the MRI model, the distribution of incorrectly classified cases is small. This is because there are only one to three patients in these groups.

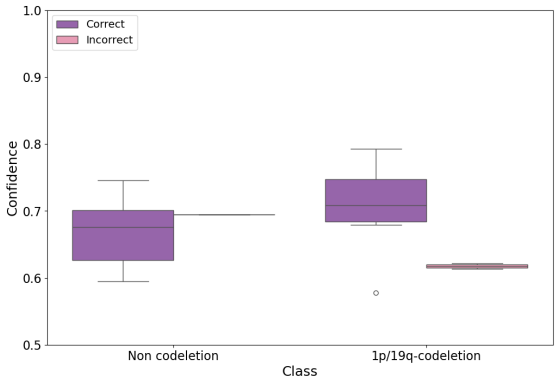
Among the evaluated models the WSI-based, cross-modal attention and projected concatenation models demonstrate the clearest separation between correctly and incorrectly classified non-codeleted samples. The cross-modal attention model shows the lowest confidence values for incorrectly classified 1p/19q-codeleted samples. In contrast, the projected concatenation model demonstrates the largest separation between correctly and incorrectly classified 1p/19q-codeleted samples. Overall, the multimodal models exhibit low confidence values for misclassified samples, particularly for the 1p/19q-codeleted class.



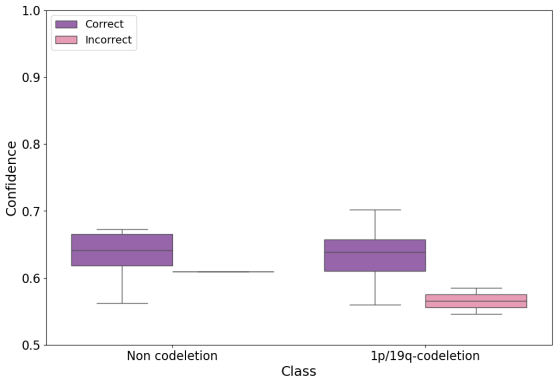
(a) WSI.



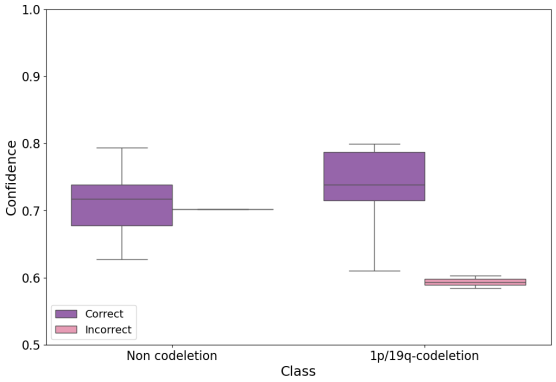
(b) MRI.



(c) Multimodal with concatenation.



(d) Multimodal with cross-modal attention.



(e) Multimodal with projected concatenation.

Figure 4.17: Confidence distributions for the codeletion classification task across all models, shown as boxplots of probabilities grouped by correct and incorrect classifications for each class.

5

Discussion

The following sections discuss the results, starting with an overview of the overall model performance. This is followed by a discussion of dataset limitations, then FMs and preprocessing. Thereafter, explainability and clinical implications are addressed, before concluding with a discussion of future work.

5.1 Overall model performance

When analyzing the best-performing models for the IDH classification task, the top three were the multimodal model with standard concatenation, the multimodal model using cross-modal attention and the WSI-based model. For the codeletion task, the top-performing models were the multimodal model using projected concatenation, the multimodal model using cross-modal attention and the WSI-based model. The multimodal models achieved the highest overall performance for both classification tasks, which is consistent with recent studies that have shown that multimodal models achieve improved performance [6]. However, the improvements over the unimodal WSI-based model were relatively modest. This suggests that most predictive information relevant to IDH status and codeletion status is already present in the histopathological images, while MRI may provide additional but limited complementary information. This is also consistent with clinical practice, where the final diagnosis is primarily based on histopathological assessment.

All models for the codeletion classification task performed better compared to the ones for IDH classification task. This indicates that the codeletion task is more easily separable than the IDH task for this dataset. For the MRI-based model on the other hand, the AUC was higher for the IDH task compared to the codeletion task. This may be explained by differences in the tumor location of each subtype, where the IDH-mutated gliomas were more frequently located in the frontal lobe, whereas IDH-wildtype tend to be more diffusely distributed or more commonly found in the temporal lobe. This could make the IDH classification task more spatially separable in MR image, while the codeletion task instead rely on histopathological differences that is less directly encoded in the image alone.

Furthermore, no single multimodal architecture consistently outperformed all other models across both tasks. Different multimodal fusion strategies performed best for the IDH and codeletion tasks, while the overall performance differences between the top models remained relatively small. This suggests that the choice of fusion archi-

tecture has a limited impact on performance within this dataset, or that different classification tasks benefit from different types of multimodal feature integration. On the other hand, it may be useful for the user to know which modality the model relies more heavily on, which can be more easily interpreted when using the multimodal model with cross-modal attention. However, these findings should be interpreted with caution due to the relatively small held-out test set size. It cannot be excluded that the choice of fusion architecture may have a greater impact when evaluated on larger and more heterogeneous datasets.

When comparing the two unimodal models, WSI-based model achieved the best overall performance for both tasks. This result was expected, as several histopathological FMs have previously demonstrated strong performance across multiple studies [9], [10]. As discussed earlier, FMs for MRI require more advanced architectures and training strategies due to the 3D nature of MR images. Due to this, the field of MRI-based FMs remains less mature and there is currently no clearly established optimal architecture, SSL strategy or FM consistently adopted across studies. While MRI primarily provides complementary information about tumor characteristics such as location and size, histopathology remains essential for definitive diagnosis. However, the small performance improvement observed in the multimodal models suggests that MRI still contributes complementary information to the classification task. Another interesting observation is that the multimodal model based on cross-modal attention assigns approximately 33% of the attention weights to MRI, although the modality appears to contribute only a relatively weak signal to the classification performance. In addition, the confidence values decrease when MRI features are incorporated. Therefore, it remains unclear whether the multimodal models would outperform the WSI-based model if a larger dataset were available.

5.2 Dataset limitations

One major limitation of this study is the relatively small dataset, particularly the held-out test set used for the codeletion classification. The final models for this task (except the MRI-based model) achieved an accuracy of 0.917, corresponding to only 3 misclassified cases out of 36. Similarly, the best-performing models for the IDH task achieved accuracies between 0.925 and 0.938, corresponding to 5-6 misclassified cases out of 80. These results may therefore be overly optimistic and may not generalize well to larger and more diverse datasets due to the limited sample size.

For the IDH classification task, the unimodal models performed better on the held-out test set, whereas the multimodal models achieved higher performance during cross-validation, indicating variability across evaluation splits. A similar pattern was observed for the codeletion task, where cross-validation performance was consistently higher than performance on the held-out test set. This variation suggests that cross-validation may overestimate generalization performance in this setting.

The models for the codeletion task, in particular the multimodal models, showed signs of overfitting during training, which is further supported by their perfect per-

formance during cross-validation. Since these models were trained only on patients with IDH mutations the total dataset size was reduced, which may increase the risk of overfitting. Although the multimodal models achieved perfect validation AUC, this performance did not fully transfer to the held-out test set, suggesting limited generalization capability.

Furthermore, the WSI representations are constructed using an ABMIL aggregator, which compresses spatial information into a single representation. This may lead to overly confident predictions and increase the risk of overfitting, particularly when applied to relatively small datasets. As discussed in Section 2.5.1, this limitation has previously been noted for MIL-based approaches [47]. Together, these factors suggest that the reported performance may be influenced by both dataset size and model architecture and may not fully reflect performance on larger, independent datasets. Future studies should therefore include larger and more diverse datasets, as well as external validation, to better assess model robustness and clinical generalisability.

5.3 Foundation models and preprocessing

The performance of FMs may be influenced by several factors, including the choice of SSL method, model architecture and preprocessing strategy. A majority of histopathological FMs are based on DINOv2 as the SSL method, whereas MRI models employ a broader range of SSL approaches. In this study, DINO3-ViT demonstrated substantially better overall performance on the cross-validation dataset compared to BrainIAC. This difference may partly be explained by the SSL method itself, but also by architectural differences between the models. DINO3-ViT contains a larger number of parameters, is pretrained on a larger collection of MR images, utilises a student-teacher framework that combines global and local representations and is based on a larger ViT architecture. In contrast, all histopathological FMs evaluated in this study are based on the same SSL method, which may explain the relatively small performance differences observed between them on the cross-validation dataset.

As discussed earlier, MR images can differ substantially between patients, even when acquired using the same MR sequence. Variations in acquisition protocols, scanners, patient positioning, image resolution, contrast and intensity distributions may influence the performance and generalisability of MRI FMs. In addition, anatomical differences such as tumour size, tumour location and overall brain morphology may further affect the extracted representations and final model predictions. Although the same preprocessing procedures recommended for the MRI FMs were applied in this study, the input data may still differ from the datasets used during pretraining.

In contrast, WSI data from a single institution are often more standardised, particularly when scanned using the same equipment and staining procedures. A factor that may have influenced the results of the histopathological FMs is the difference in RGB normalisation values used during preprocessing. Both Prov-GigaPath and UNI2-h apply standard normalisation values derived from natural images, whereas H-optimus-0 uses values that are more specifically adapted to histopathological im-

ages. Since the Sahlgrenska dataset exhibited higher mean RGB values compared to those used by the pretrained models, this mismatch may have affected the feature extraction process and reduced the generalisability of the extracted representations. One possible explanation for the higher mean values is a lighter staining appearance in the dataset images. Furthermore, the standard deviation of the RGB channels in the Sahlgrenska dataset was lower than the values used during preprocessing of the FMs. This may be explained by the fact that all images originated from the same institution and were acquired using the same scanner, resulting in lower variability within the dataset. This difference suggests a potential distribution shift between the training data of the FMs and the Sahlgrenska dataset. Further preprocessing steps, such as color standardization, could potentially reduce these effects and improve the generalizability of the model.

Among the evaluated models, H-optimus-0 had normalization statistics closest to those of the Sahlgrenska dataset, which may explain why this feature extractor achieved better performance on the IDH classification task. The heatmap analysis showed that the model for the IDH task exhibited more concentrated high-activation regions, suggesting a stronger reliance on localized histopathological details. Consequently, differences in color distribution and normalization statistics may have a greater influence on the IDH classification task. For the codeletion task, on the other hand, Prov-GigaPath achieved the best performance despite having normalization statistics that differed more from those of the Sahlgrenska dataset. The model for the codeletion task instead displayed smoother and more diffuse activation patterns, indicating a greater reliance on broader tissue structures and overall morphological patterns, which may make the task less sensitive to differences in color normalization.

In addition to the variability in staining, WSIs can also contain varying amounts of tissue present in each slide, leading to differences in the number of extracted training tiles per WSI. The FMs may also have been trained on datasets with different tissue distributions and tissue coverage. This could influence the extracted representations and downstream classification performance.

On the other hand, WSIs are images with very high resolution and extracting the features from the FM requires substantial computational resources. This cost can potentially be reduced by focusing on regions of interest, which would limit the number of tiles that need to be processed and improve both training efficiency and computational cost. In a clinical setting, such an approach could be advantageous by enabling faster classification and more efficient use of computational resources. In this study, the average feature extraction time for a single WSI was approximately 8 minutes.

5.4 Explainability

Since deep learning models are considered as a black box, it may be necessary to have explanations of how the models have made their decisions. This helps the clinicians to see which areas in the images that are important and also to evaluate

the reliability of the model. The histopathological FMs used in this study generated multiple tile feature vectors for each WSI, which were subsequently processed by ABMIL. The resulting attention weights could then be visualized as heatmaps to analyse which regions contributed most to the model predictions. However, the MRI FMs generated only a single feature vector per MRI, which limits the possibility of generating attention heatmaps for the MRI-based and multimodal models.

The differences observed in the WSI heatmaps suggest that the model for the IDH classification task relies more heavily on localized histopathological features, whereas the model for the codeletion task appears to capture broader tissue-level patterns. This may indicate that the two molecular classification tasks depend on partially distinct morphological characteristics, despite some overlap in the attended regions. Both models focus primarily on areas of high cellularity, which is consistent with regions typically assessed by pathologists in routine diagnostic practice. Misclassified cases showed no consistent attention patterns or shared morphological regions, indicating that errors result from tumor heterogeneity and limited dataset size rather than a systematic pattern of errors in the model.

The attention weights of the multimodal model with cross-modal attention indicate a stronger reliance on WSI features for both tasks. This is expected, as the tasks are generally assumed to benefit more from histopathological information, whereas MRI is expected to contribute complementary information regarding tumor location and morphology. Additionally, since the WSI FMs produce high-level embeddings that are subsequently aggregated by a frozen ABMIL, the fusion mechanism may inherently favor the modality with more stable representations, which in this case appears to be the histopathological. However, the weights on WSI differs between the tasks, which may reflect differences in the underlying signal strength of each modality for the respective classification task, as well as dataset-specific characteristics. To further investigate modality contributions, future work could explore models that provide spatially or voxel-level interpretable MRI representations, or alternative FMs with feature maps suitable for MIL aggregation.

Another explainable AI aspect that may contribute to understanding model behavior is the confidence value associated with each prediction. In general, higher confidence values indicate that the model is more certain about its prediction, whereas values closer to 0.5 reflect greater uncertainty. However, high confidence does not necessarily imply that the prediction is correct, as deep learning models can still produce overconfident incorrect predictions.

The observed confidence patterns in the IDH classification task may reflect underlying biological and radiological differences between tumor subtypes. IDH-wildtype gliomas are often associated with increased cellularity, necrosis and structural irregularities in histopathological tissue compared to IDH-mutant gliomas. Such features may create stronger discriminative patterns, particularly for the WSI-based models. Previous studies have also reported that different subtypes tend to occur in distinct anatomical brain regions and exhibit characteristic appearances on MR imaging, which could contribute to the separability observed in the confidence distributions.

The broader spread of confidence values observed for the MRI-based model may

indicate greater variability in the MR images and suggest that some cases are more difficult to classify from imaging alone. In contrast, the narrower confidence distributions observed for some multimodal approaches may reflect more stable feature representations after modality fusion. However, the presence of relatively confident incorrect predictions in certain multimodal models suggests that combining modalities does not automatically improve prediction calibration or uncertainty estimation.

For the codeletion classification task, the generally lower confidence associated with misclassified samples may indicate that the models were able to identify uncertain cases to some extent. From a clinical perspective, this behavior is important, as low-confidence predictions could potentially be flagged for additional expert review or complementary diagnostic testing. Furthermore, the clearer separation between correct and incorrect predictions observed for some multimodal approaches may suggest that integrating histopathological and radiological information improves the discriminative capability of the models.

Overall, the confidence analysis highlights that prediction certainty varies substantially across both models and molecular subtypes. This suggests that confidence values may provide complementary information that could support clinical decision-making in a future clinical setting by identifying uncertain predictions.

5.5 Clinical implications

An interesting observation for the IDH classification task is that, for all models except MRI, recall was higher than specificity. This indicates that the models were more effective at detecting IDH mutation cases than correctly identifying IDH-wildtype cases. In practice, this suggests a tendency toward minimizing false negatives, meaning that IDH mutations were rarely missed. However, this may come at the cost of increased false positives, where IDH-wildtype cases are incorrectly classified as IDH mutation. From a clinical perspective, prioritizing recall may be preferable, since missing an IDH mutation could potentially affect diagnosis, prognosis and treatment planning. However, missing an IDH-wildtype case may also have serious clinical consequences, since these tumours are generally associated with a poorer prognosis and may require more urgent treatment. Therefore, the clinical importance of recall versus specificity depends on the intended use of the model and the relative consequences of false negatives and false positives.

Furthermore, all models for the IDH classification task showed higher NPV than PPV, indicating that negative predictions (IDH-wildtype) were generally more reliable than positive predictions (IDH mutation). This suggests that the models were better at ruling out IDH mutations than confidently confirming them. The MRI model differed slightly from the other models by showing higher specificity than recall, indicating a greater ability to identify IDH-wildtype cases but a weaker ability to detect IDH mutation cases.

For the codeletion task, the specificity was higher than recall for all models except the MRI-based model, indicating better identification of non-codeleted cases than 1p/19q codeletion cases. Additionally, PPV was higher than NPV across all models,

suggesting that positive predictions of 1p/19q codeletion were generally more reliable than negative predictions. From a clinical perspective, this indicates that the models were more conservative in predicting 1p/19q codeletion, reducing the risk of falsely classifying non-codeleted tumours as codeleted. Since 1p/19q codeletion is associated with oligodendroglioma and generally a more favourable prognosis, a false positive prediction could potentially lead to overly optimistic prognostic assessment or influence treatment decisions. At the same time, the lower recall suggests that some truly codeleted tumours may have been missed, which could reduce the models usefulness as a screening tool. Therefore, the models appear to be more reliable for confirming the presence of 1p/19q codeletion than for excluding it.

Another important consideration is how the models would be integrated into a clinical pipeline. In practice, IDH-status prediction would likely be performed first, and only cases predicted as IDH-mutant would subsequently be analyzed by the model for the codeletion task. In such a sequential setup, falsely predicted IDH-mutant cases that are actually IDH-wildtype would also be passed to the model predicting codeletion status. Since the model has not been trained on IDH-wildtype, this would propagate errors through the pipeline and reducing the overall reliability of the system. The first step is therefore critical for the overall system performance, as errors at this stage propagate through the subsequent classification step.

5.6 Future work

To further evaluate the proposed models, future studies should include larger and more diverse test sets. Since the models in this study were developed for use at Sahlgrenska University Hospital, where the same WSI scanner and staining procedures are used, it would be beneficial to include additional data from the same institution to better reflect the intended clinical setting. At the same time, incorporating external open-source datasets would allow comparisons with previous studies and provide a more robust assessment of model generalizability. In that case, future work could also explore training on multi-institutional datasets to improve robustness across varying scanners, staining procedures and acquisition protocols.

Another important direction for future research would be to investigate alternative MIL approaches for the WSI-based models. As discussed earlier, ABMIL may increase the risk of overfitting and overestimation, particularly on small datasets, which was observed for some of the multimodal models for the codeletion task. Exploring alternative aggregation methods could therefore improve generalisation performance and model stability. Future studies could also investigate additional colour normalisation and stain standardisation techniques for WSI, which may reduce distribution shifts between datasets and improve model generalisability. Although the evaluated histopathological FMs already demonstrated strong performance, further optimisation of preprocessing strategies may still improve robustness across institutions.

Furthermore, it would be valuable to include additional MR sequences, such as T2-weighted and FLAIR images. Since tumour characteristics may be more visible in

certain sequences, combining multiple MR sequences could provide more comprehensive information about tumour morphology and extent. This may reduce the risk of missing relevant tumour regions and could potentially improve the performance of both the MRI-based and multimodal models. The evaluated MRI FMs in this study could then be used, as they were pretrained on multiple MR sequences. In addition, future studies could explore alternative MRI-based FMs, particularly models with improved explainability, such as architectures capable of generating attention heatmaps. Such approaches could provide further insight into how radiological information contributes to molecular classification. Further investigation of different SSL strategies and architectural designs for MRI-based FMs may also help identify more optimal approaches for this application.

Overall, continued improvements in dataset diversity, preprocessing, multimodal integration and explainability will likely be important steps toward future clinical implementation of deep learning-based molecular classification models. These future directions may contribute to the development of more robust, generalisable and clinically applicable models for molecular classification of adult-type diffuse gliomas. Furthermore, FMs represent a promising approach in medical imaging, as they enable the use of large-scale pretrained feature extractors without requiring extensive task-specific training from scratch. Since these models are pretrained on large and diverse datasets, they may also improve model generalisability and robustness. Continued developments in FMs and SSL are therefore likely to remain important areas of research within deep learning for clinical imaging and data analysis.

6

Conclusion

Overall, the models achieved strong classification performance for adult-type diffuse gliomas, as reflected by high AUC, accuracy, recall, specificity, PPV and NPV across both IDH and 1p/19q codeletion tasks. The models for the IDH classification generally showed higher recall, while the models for the codeletion task were characterized by a stronger emphasis on specificity relative to recall. Regarding model performance, multimodal architectures achieved the highest overall results for both tasks, although the improvements over WSI-based models were relatively modest. This suggests that histopathological images contain most of the predictive information, while MR images provide complementary but limited additional value. Although multimodal models, particularly those based on cross-modal attention, achieved the best overall performance, no consistent improvement was observed across all settings. This indicates that the optimal model design may depend on the specific molecular classification task.

Differences in performance may be more influenced by the underlying FMs architecture and the choice of SSL strategy, as these components form the basis of the feature extraction process. For the histopathological FMs, DINOv2 was the most commonly used SSL approach and was also present in the best-performing models. In this study, the MRI-based model DINO3-ViT also utilised a similar SSL strategy, which may partly explain its relatively strong performance. However, other factors, such as differences in model architecture, may also have contributed. In particular, DINO3-ViT is based on a larger ViT architecture compared to BrainIAC, which may have influenced the observed performance differences.

From an explainability perspective, attention-based heatmaps from WSI models showed that predictions were primarily driven by regions with high cellularity, which is consistent with clinically relevant tumor regions. The models for the IDH task tended to focus on more localized features, while the models for the codeletion task relied more on broader tissue-level patterns. MRI-based and multimodal models could not be visualized using spatial heatmaps due to their single-vector representation. The multimodal model with cross-modal attention indicated a stronger reliance on WSI features compared to MRI features, highlighting the dominant role of histopathology in the classification task. These findings support the clinical relevance of the learned representations and demonstrate that explainable AI methods can provide interpretable insights into model behaviour.

In addition to standard performance metrics, the analysis of prediction confidence

provided further insight into model behaviour. Overall, WSI-based models tended to produce more consistent confidence distributions, while MRI-based models exhibited greater variability, reflecting higher uncertainty in image-based classification. Multimodal models showed, in some cases, more stable confidence patterns, suggesting improved robustness through feature integration. However, the presence of high-confidence incorrect predictions across all model types indicates that confidence alone does not reliably indicate prediction correctness. Nevertheless, confidence values may still provide complementary information for identifying uncertain cases in a clinical setting.

Together, these findings demonstrate that reliable molecular classification of adult-type diffuse gliomas is feasible using deep learning and FMs applied to WSIs and MRI. Histopathology remained the dominant source of predictive information, while MRI contributed complementary features that improved performance in some multimodal settings. The study also showed that explainable AI methods can provide clinically meaningful insights into model behavior by highlighting relevant tumor regions and modality contributions.

Although the results are promising, further validation on larger and more heterogeneous multi-center datasets is required before clinical implementation. Future work should investigate additional molecular markers, larger multimodal datasets, more advanced fusion architectures and improved explainability methods for MRI-based models. With continued development and validation, multimodal deep learning systems have the potential to support pathologists and radiologists in achieving faster, more consistent and more accessible glioma classification in clinical practice.

Bibliography

[1] B. T. Whitfield and J. T. Huse, "Classification of adulttype diffuse gliomas: Impact of the world health organization 2021 update," *Brain Pathology*, vol. 32, no. 4, e13062, Mar. 2022, ISSN: 1015-6305. DOI: 10.1111/bpa.13062. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9245936/.

[2] W. C. of Tumours Editorial Board, Ed., *WHO Classification of Tumours of the Central Nervous System*, en, 5th. International Agency for Research on Cancer, 2021, vol. 6, ISBN: 978-92-832-4508-7. [Online]. Available: https://publications.iarc.who.int/Book-And-Report-Series/Who-Classification-Of-Tumours/Central-Nervous-System-Tumours-2021.

[3] C. Lancellotti et al., "Artificial intelligence tissue biomarkers: Advantages, risks and perspectives for pathology," en, *Cells*, vol. 10, no. 4, p. 787, Apr. 2021, ISSN: 2073-4409. DOI: 10.3390/cells10040787. [Online]. Available: https://www.mdpi.com/2073-4409/10/4/787.

[4] V. v. Veldhuizen et al., "Foundation models in medical imaging: A review and outlook," *arXiv preprint arXiv:2506.09095*, no. arXiv:2506.09095, Nov. 2025, arXiv:2506.09095. DOI: 10.48550/arXiv.2506.09095. [Online]. Available: http://arxiv.org/abs/2506.09095.

[5] S.-C. Huang, M. Jensen, S. Yeung-Levy, M. P. Lungren, H. Poon, and A. S. Chaudhari, "Multimodal foundation models for medical imaging - a systematic review and implementation guidelines," en, *medRxiv*, Oct. 2024. DOI: 10.1101/2024.10.23.24316003. [Online]. Available: http://medrxiv.org/lookup/doi/10.1101/2024.10.23.24316003.

[6] C. Saueressig, D. Scholz, P. Raffler, C. Delbridge, B. Wiestler, and P. Schüffler, "Multimodal fusion of pathology and radiology foundation models for who 2021 glioma subtyping," en, *npj Precision Oncology*, vol. 10, no. 1, p. 118, Mar. 2026, ISSN: 2397-768X. DOI: 10.1038/s41698-026-01366-5. [Online]. Available: https://www.nature.com/articles/s41698-026-01366-5.

[7] W. Wang et al., "Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images," en, *Nature Communications*, vol. 14, no. 1, p. 6359, Oct. 2023, ISSN: 2041-1723. DOI: 10.1038/s41467-023-41195-9. [Online]. Available: https://www.nature.com/articles/s41467-023-41195-9.

[8] V. Despotovic et al., "Glioma subtype classification from histopathological images using in-domain and out-of-domain transfer learning: An experimental study," en, *Heliyon*, vol. 10, no. 5, e27515, Mar. 2024, ISSN: 24058440. DOI: 10.

1016/j.heliyon.2024.e27515. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405844024035461.

[9] G. Campanella et al., "A clinical benchmark of public self-supervised pathology foundation models," en, *Nature Communications*, vol. 16, no. 1, p. 3640, Apr. 2025, ISSN: 2041-1723. DOI: 10.1038/s41467-025-58796-1. [Online]. Available: https://www.nature.com/articles/s41467-025-58796-1.

[10] R. Bareja et al., "Evaluating vision and pathology foundation models for computational pathology: A comprehensive benchmark study," en, *medRxiv*, May 2025. DOI: 10.1101/2025.05.08.25327250. [Online]. Available: https://www.medrxiv.org/content/10.1101/2025.05.08.25327250v1.

[11] L. Jose et al., "Artificial intelligenceassisted classification of gliomas using wholeslide images," *Archives of Pathology & Laboratory Medicine*, vol. 147, no. 8, pp. 916–924, 2023. DOI: 10.5858/arpa.2021-0518-OA. [Online]. Available: https://aplm.kglmeridian.com/view/journals/arpa/147/8/article-p916.xml.

[12] C. Saueressig et al., "From histology to diagnosis: Leveraging pathology foundation models for glioma classification," en, *Computers in Biology and Medicine*, vol. 197, p. 110 988, Oct. 2025, ISSN: 00104825. DOI: 10.1016/j.compbiomed.2025.110988. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S001048252501340X.

[13] D. Tak et al., "A generalizable foundation model for analysis of human brain mri," en, *Nature Neuroscience*, vol. 29, no. 4, pp. 945–956, Apr. 2026, ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-026-02202-6. [Online]. Available: https://www.nature.com/articles/s41593-026-02202-6.

[14] T. Xu et al., "A generalizable 3d framework and model for self-supervised learning in medical imaging," en, *npj Digital Medicine*, vol. 8, no. 1, p. 639, Nov. 2025, ISSN: 2398-6352. DOI: 10.1038/s41746-025-02035-w. [Online]. Available: https://www.nature.com/articles/s41746-025-02035-w.

[15] S. Shirae, S. S. Debsarkar, H. Kawanaka, B. Aronow, and V. B. S. Prasath, "Multimodal ensemble fusion deep learning using histopathological images and clinical data for glioma subtype classification," *IEEE Access*, vol. 13, pp. 57 780–57 797, 2025, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3556713. [Online]. Available: https://ieeexplore.ieee.org/document/10946995.

[16] S. Rathore, A. Chaddad, M. A. Iftikhar, M. Bilello, and A. Abdulkadir, "Combining mri and histologic imaging features for predicting overall survival in patients with glioma," en, *Radiology: Imaging Cancer*, vol. 3, no. 4, e200108, 2021, ISSN: 2638-616X. DOI: 10.1148/rycan.2021200108. [Online]. Available: http://pubs.rsna.org/doi/10.1148/rycan.2021200108.

[17] A. Hamidinekoo, T. Pieciak, M. Afzali, O. Akanyeti, and Y. Yuan, "Glioma classification using multimodal radiology and histology data," en, in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. Cham: Springer International Publishing, 2021, vol. 12659, pp. 508–518, ISBN: 9783030720865. DOI: 10.1007/978-3-030-72087-2_45. [Online]. Available: https://link.springer.com/10.1007/978-3-030-72087-2_45.

[18] A. M. Alessio, S. Khan, and S. Patel, "A multi-modal deep learning model integrates clinical, pathomic, and radiomic features for glioma classification and grading.," en, *Journal of Clinical Oncology*, vol. 40, no. 16$_s$*uppl*, e14038–e14038, 2022, ISSN: 0732-183X, 1527-7755. DOI: 10.1200/JCO.2022.40.16_suppl.e14038. [Online]. Available: https://ascopubs.org/doi/10.1200/JCO.2022.40.16_suppl.e14038.

[19] X. Wang et al., "Combining radiology and pathology for automatic glioma classification," English, *Frontiers in Bioengineering and Biotechnology*, vol. 10, Mar. 2022, ISSN: 2296-4185. DOI: 10.3389/fbioe.2022.841958. [Online]. Available: https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2022.841958/full.

[20] M. Weller et al., "Eano guidelines on the diagnosis and treatment of diffuse gliomas of adulthood," eng, *Nature Reviews. Clinical Oncology*, vol. 18, no. 3, pp. 170–186, Mar. 2021, ISSN: 1759-4782. DOI: 10.1038/s41571-020-00447-z.

[21] K. Lv et al., "Neuroplasticity of glioma patients: Brain structure and topological network," *Frontiers in Neurology*, vol. 13, p. 871 613, May 2022, ISSN: 1664-2295. DOI: 10.3389/fneur.2022.871613. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fneur.2022.871613/full.

[22] H. Colman, "Adult gliomas," en, *Continuum*, vol. 26, no. 6, pp. 1452–1475, Dec. 2020, ISSN: 1080-2371, 1538-6899. DOI: 10.1212/CON.0000000000000935. [Online]. Available: https://continuum.aan.com/doi/10.1212/CON.0000000000000935.

[23] M. C. M. Peeters et al., "Prediagnostic symptoms and signs of adult glioma: The patients view," en, *Journal of Neuro-Oncology*, vol. 146, no. 2, pp. 293–301, Jan. 2020, ISSN: 0167-594X, 1573-7373. DOI: 10.1007/s11060-019-03373-y. [Online]. Available: http://link.springer.com/10.1007/s11060-019-03373-y.

[24] J. P. Thakkar et al., "Epidemiologic and molecular prognostic review of glioblastoma," en, *Cancer Epidemiology, Biomarkers Prevention*, vol. 23, no. 10, pp. 1985–1996, Oct. 2014, ISSN: 1055-9965, 1538-7755. DOI: 10.1158/1055-9965.EPI-14-0275. [Online]. Available: https://aacrjournals.org/cebp/article/23/10/1985/14199/Epidemiologic-and-Molecular-Prognostic-Review-of.

[25] A. Pouyan et al., "Glioblastoma multiforme: Insights into pathogenesis, key signaling pathways, and therapeutic strategies," *Molecular Cancer*, vol. 24, p. 58, Feb. 2025, ISSN: 1476-4598. DOI: 10.1186/s12943-025-02267-0. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11863469/.

[26] Y. Zhao, Y. Yu, W. Chen, X. Zhang, J. Lv, and H. Zhao, "Oligodendroglioma: Advances in molecular mechanisms and immunotherapeutic strategies," en, *Biomedicines*, vol. 13, no. 5, p. 1133, May 2025, ISSN: 2227-9059. DOI: 10.3390/biomedicines13051133. [Online]. Available: https://www.mdpi.com/2227-9059/13/5/1133.

[27] F. Yang, Y. Zou, Q. Gong, J. Chen, W.-D. Li, and Q. Huang, "From astrocytoma to glioblastoma: A clonal evolution study," en, *FEBS Open Bio*, vol. 10,

no. 5, pp. 744–751, May 2020, ISSN: 2211-5463, 2211-5463. DOI: 10.1002/2211-5463.12815. [Online]. Available: https://febs.onlinelibrary.wiley.com/doi/10.1002/2211-5463.12815.

[28] Y. H. Byun and C.-K. Park, "Classification and diagnosis of adult glioma: A scoping review," en, *Brain Neurorehabilitation*, vol. 15, no. 3, e23, 2022, ISSN: 1976-8753, 2383-9910. DOI: 10.12786/bn.2022.15.e23. [Online]. Available: https://e-bnr.org/DOIx.php?id=10.12786/bn.2022.15.e23.

[29] A. Idbaih et al., "Two types of chromosome 1p losses with opposite significance in gliomas," en, *Annals of Neurology*, vol. 58, no. 3, pp. 483–487, 2005, ISSN: 0364-5134, 1531-8249. DOI: 10.1002/ana.20607. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ana.20607.

[30] The Human Protein Atlas, *Cancer*. [Online]. Available: https://www.proteinatlas.org/learn/dictionary%5C#cancer.

[31] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, and B. Srinivasan, "A generalized deep learning framework for whole-slide image segmentation and analysis," en, *Scientific Reports*, vol. 11, no. 1, p. 11 579, 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-90444-8. [Online]. Available: https://www.nature.com/articles/s41598-021-90444-8.

[32] A. T. Feldman and D. Wolfe, "Tissue processing and hematoxylin and eosin staining," en, in *Histopathology*, C. E. Day, Ed. New York, NY: Springer New York, 2014, vol. 1180, pp. 31–43, ISBN: 9781493910496. DOI: 10.1007/978-1-4939-1050-2_3. [Online]. Available: https://link.springer.com/10.1007/978-1-4939-1050-2_3.

[33] Y. Murakami et al., "Color correction in whole slide digital pathology," *Color and Imaging Conference*, vol. 20, no. 1, pp. 253–258, Jan. 2012, ISSN: 2166-9635. DOI: 10.2352/CIC.2012.20.1.art00045. [Online]. Available: https://library.imaging.org/cic/articles/20/1/art00045.

[34] National Institute of Biomedical Imaging and Bioengineering, en. [Online]. Available: https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri.

[35] M. T. Vlaardingerbroek and J. A. Boer, *Magnetic Resonance Imaging: Theory and Practice*, en. Springer Science Business Media, Mar. 2013, Google-Books-ID: 9D78CAAAQBAJ, ISBN: 9783662052525.

[36] Case Western Reserve University, Department of Neurology, *Mri basics*, en, 2026. [Online]. Available: https://case.edu/med/neurology/NR/MRI%20Basics.htm.

[37] M. Bekiesiska-Figatowska, "Artifacts in magnetic resonance imaging," en, *Polish Journal of Radiology*, vol. 80, pp. 93–106, 2015, ISSN: 0137-7183. DOI: 10.12659/PJR.892628. [Online]. Available: http://www.polradiol.com/abstract/index/idArt/892628.

[38] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979, ISSN: 0018-9472, 2168-2909. DOI: 10.1109/TSMC.1979.4310076. [Online]. Available: http://ieeexplore.ieee.org/document/4310076/.

[39] N. J. Tustison et al., "N4itk: Improved n3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010, ISSN: 0278-0062,

1558-254X. DOI: 10.1109/TMI.2010.2046908. [Online]. Available: http://ieeexplore.ieee.org/document/5445030/.

[40] F. Isensee et al., "Automated brain extraction of multisequence mri using artificial neural networks," en, *Human Brain Mapping*, vol. 40, no. 17, pp. 4952–4964, Dec. 2019, ISSN: 1065-9471, 1097-0193. DOI: 10.1002/hbm.24750. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/hbm.24750.

[41] Y. He et al., "Foundation model for advancing healthcare: Challenges, opportunities and future directions," *IEEE Reviews in Biomedical Engineering*, vol. 18, pp. 172–191, 2025, ISSN: 1941-1189. DOI: 10.1109/RBME.2024.3496744. [Online]. Available: https://ieeexplore.ieee.org/document/10750441/.

[42] A. Berroukham, K. Housni, and M. Lahraichi, "Vision transformers: A review of architecture, applications, and future directions," in *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, Agadir - Essaouira, Morocco: IEEE, Dec. 2023, pp. 205–210, ISBN: 9781665461337. DOI: 10.1109/CiSt56084.2023.10410015. [Online]. Available: https://ieeexplore.ieee.org/document/10410015/.

[43] C. Saillard et al., *H-optimus-0*, 2024. [Online]. Available: https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

[44] H. Xu et al., "A whole-slide foundation model for digital pathology from real-world data," en, *Nature*, vol. 630, no. 8015, pp. 181–188, 2024, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-024-07441-w. [Online]. Available: https://www.nature.com/articles/s41586-024-07441-w.

[45] R. J. Chen et al., "Towards a general-purpose foundation model for computational pathology," en, *Nature Medicine*, vol. 30, no. 3, pp. 850–862, Mar. 2024, ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-024-02857-3. [Online]. Available: https://www.nature.com/articles/s41591-024-02857-3.

[46] M. Gadermayr and M. Tschuchnig, "Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations future potential," en, *Computerized Medical Imaging and Graphics*, vol. 112, p. 102 337, Mar. 2024, ISSN: 08956111. DOI: 10.1016/j.compmedimag.2024.102337. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0895611124000144.

[47] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, no. arXiv:1802.04712, 2018, arXiv:1802.04712. DOI: 10.48550/arXiv.1802.04712. [Online]. Available: http://arxiv.org/abs/1802.04712.

[48] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," en, *Machine Vision and Applications*, vol. 32, no. 6, p. 121, Nov. 2021, ISSN: 0932-8092, 1432-1769. DOI: 10.1007/s00138-021-01249-8. [Online]. Available: https://link.springer.com/10.1007/s00138-021-01249-8.

[49] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. [Online].

Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[50] H. Kumar, M. Aruldoss, and M. Wynn, "Cross-modal attention fusion: A deep learning and affective computing model for emotion recognition," en, *Multimodal Technologies and Interaction*, vol. 9, no. 12, p. 116, Nov. 2025, ISSN: 2414-4088. DOI: 10.3390/mti9120116. [Online]. Available: https://www.mdpi.com/2414-4088/9/12/116.

[51] GeeksforGeeks, en, May 2024. [Online]. Available: https://www.geeksforgeeks.org/deep-learning/what-is-fully-connected-layer-in-deep-learning/.

[52] D. Tak, B. Gormosa, A. Zapaishchykova, et al., *A generalizable foundation model for analysis of human brain mri*, 2025. [Online]. Available: https://github.com/AIM-KannLab/BrainIAC.

[53] T. Xu, S. Hosseini, C. Anderson, A. Rinaldi, and Krishna, *3dino: Self-supervised 3d visual representation learning for medical images*, 2025. [Online]. Available: https://github.com/AICONSlab/3DINO?tab=readme-ov-file.

A

Appendix 1: Data structures

This appendix describes the data structures used in the implementation.

A.1 WSI extracted feature files

For each WSI and FM, a PyTorch file (.pt) was generated containing all extracted feature vectors and tile coordinates. The files were stored using the following dictionary structure:

```
{
    {
        coordinate:
        feature: tensor([])
    },
    ...
}
```

Here, `coordinate` denotes the tile coordinates, while `feature` contains the extracted feature vector for that tile.

A.2 MRI extracted feature files

For each MRI and FM, a PyTorch file (.pt) containing extracted feature vectors was generated. The files were stored using the following dictionary structure:

```
{
    pat_id: {
        idh_label:
        codel_label:
        subtype:
        feature: tensor([])
    },
    ...
}
```

Here, `pat_id` denotes the patient identifier, while `feature` contains the extracted feature vector for that patient.

A.3 Multimodal extracted feature files

For the multimodal dataset, a PyTorch file (.pt) containing extracted feature vectors from both modalities was generated. The files were stored using the following dictionary structure:

```
{
    pat_id: {
        idh_label:
        codel_label:
        subtype:
        mri_feature: tensor([])
        he_feature: tensor([])
    },
    ...
}
```

Here, `pat_id` denotes the patient identifier, `mri_feature` contains the extracted feature vector from 3DINO-ViT and `he_feature` contains the extracted feature vector from H-optimus-0 (IDH model) or Prov-Gigapath (Codeletion model) for that patient.

B

Appendix 2: Optuna hyperparameter intervals

This appendix lists the hyperparameter search ranges used during the Optuna optimization for each model setting (WSI, MRI and multimodal).

B.1 WSI

The following hyperparameter ranges were used for the WSI models during Optuna optimization.

- Activation function: [ReLU, Tanh, GELU]
- Hidden dimension: [128, 256, 384, 512]
- Dropout: 0.0 – 0.5
- Gated: [True, False]
- Learning rate: $1 \cdot 10^{-5}$ – $1 \cdot 10^{-3}$
- Weight decay: $1 \cdot 10^{-6}$ – $1 \cdot 10^{-3}$

B.2 MRI

The following hyperparameter ranges were used for the MRI models during Optuna optimization.

- Activation function: [ReLU, Tanh, GELU]
- Hidden dimension: [64, 128, 256, 512]
- Dropout: 0.0 – 0.5
- Epochs: 20 – 100
- Batch size: [8, 16, 32, 64]
- Learning rate: $1 \cdot 10^{-6}$ – $1 \cdot 10^{-2}$
- Weight decay: $1 \cdot 10^{-6}$ – $1 \cdot 10^{-1}$

B.3 Multimodal

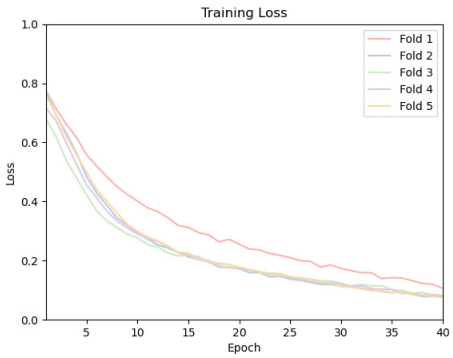
The following hyperparameter ranges were used for the multimodal models during Optuna optimization.

- Activation function: [ReLU, Tanh, GELU]
- Hidden dimension: [64, 128, 256, 512]
- Dropout: 0.0 – 0.5
- Epochs: 20 – 200
- Batch size: [16, 32, 64]
- Learning rate: $1 \cdot 10^{-7}$ – $1 \cdot 10^{-4}$
- Weight decay: $1 \cdot 10^{-4}$ – $1 \cdot 10^{-1}$

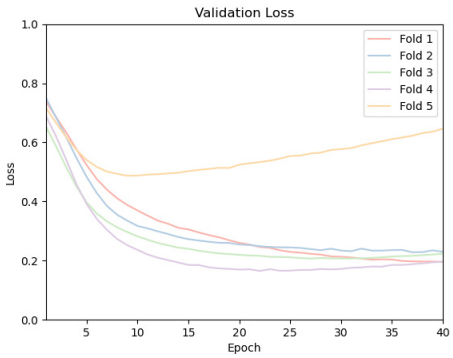
C

Appendix 3: Training and validation loss

In this appendix, the training and validation losses are presented for both the IDH and codeletion tasks using the FMs Prov-GigaPath, UNI2-h and BrainIAC.

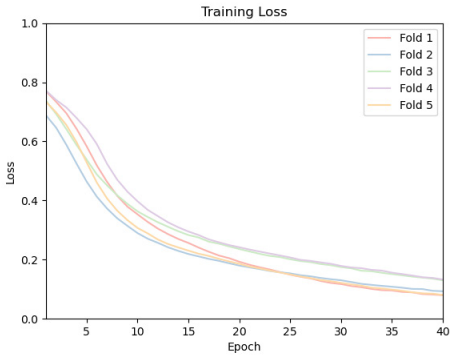


(a) Training loss.

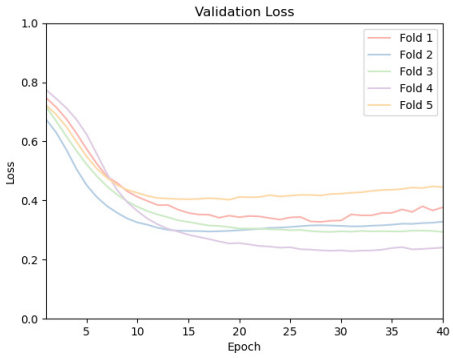


(b) Validation loss.

Figure C.1: Training and validation loss during training for the IDH classification task using WSI model with features extracted from Prov-gigapath.

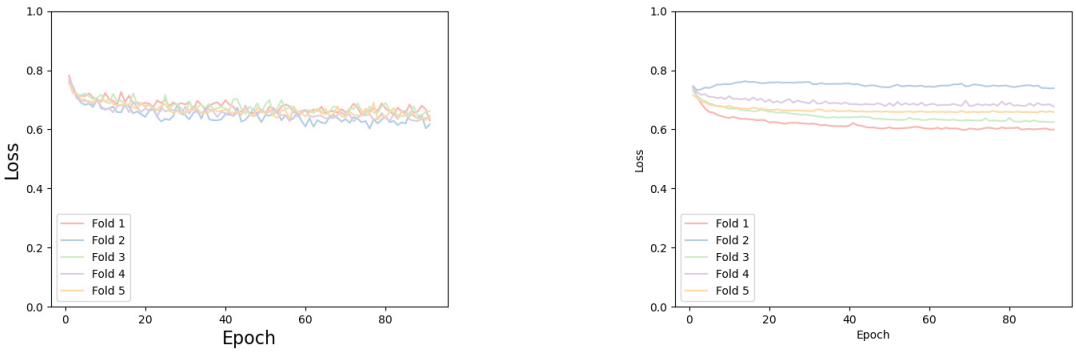


(a) Training loss.



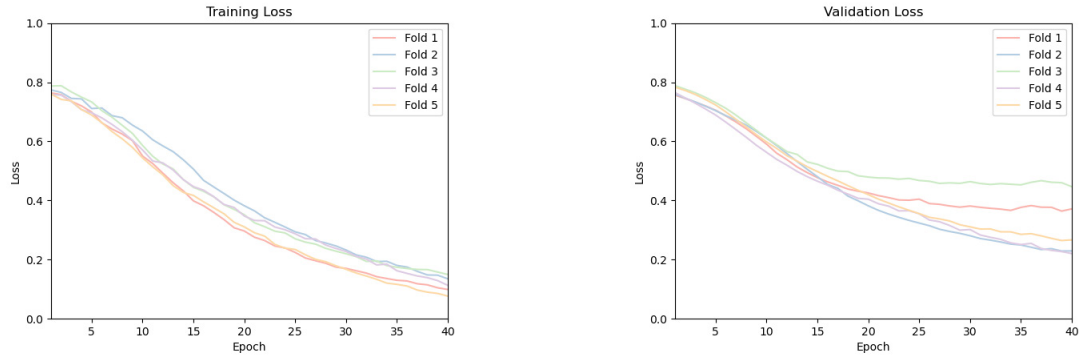
(b) Validation loss.

Figure C.2: Training and validation loss during training for the IDH classification task using WSI model with features extracted from UNI2-h.



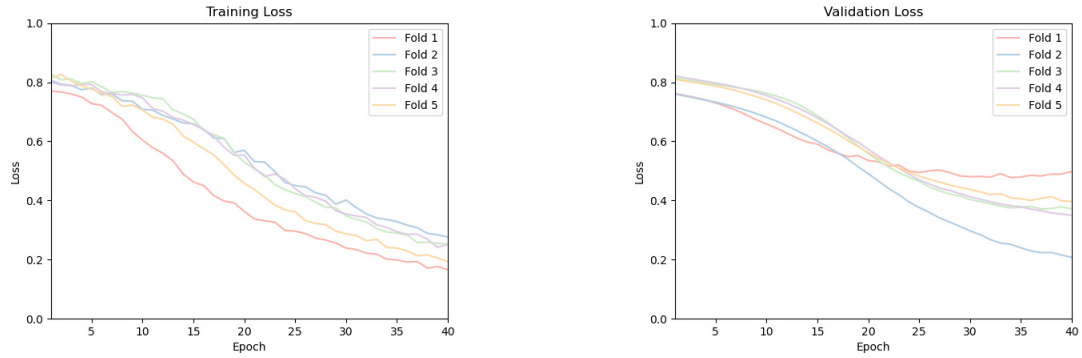
(a) Training loss. (b) Validation loss.

Figure C.3: Training and validation loss during training for the IDH classification task using MRI model with features extracted from BrainIAC.



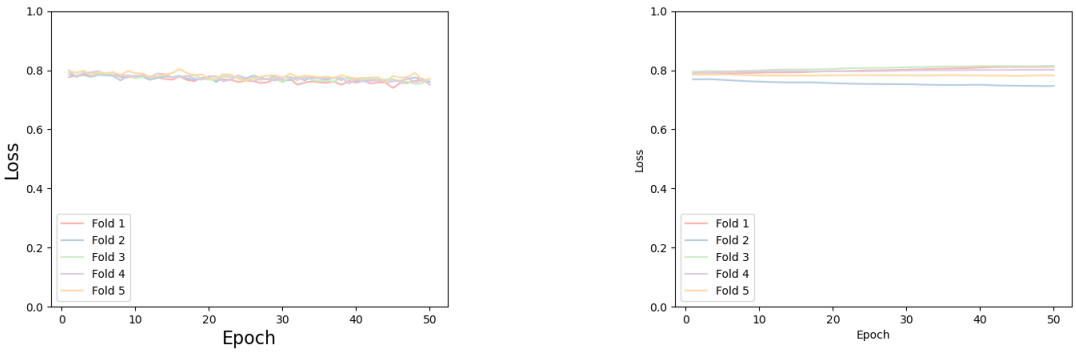
(a) Training loss. (b) Validation loss.

Figure C.4: Training and validation loss during training for the codeletion classification task using WSI model with features extracted from H-Optimus-0.



(a) Training loss. (b) Validation loss.

Figure C.5: Training and validation loss during training for the codeletion classification task using WSI model with features extracted from UNI2-h.



(a) Training loss.

(b) Validation loss.

Figure C.6: Training and validation loss during training for the codeletion classification task using a MRI model with features extracted from BrainIAC.