

Master's Thesis

**A Study on Prediction of IDH1  
Mutation Using Machine Learning**

**Riku Nakagaki**

Division of Electrical and Electronic Engineering  
Graduate School of Engineering  
Mie University

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Classification of Gliomas and IDH Gene Mutation . . . . .	2
1.3	Objective . . . . .	3
<b>2</b>	<b>Related Works</b>	<b>5</b>
2.1	Application of Deep Learning in Mutation Prediction from Pathological Images . . . . .	5
2.2	GAN-based Data Augmentation . . . . .	6
2.3	Combining Pathological Images and Clinical Data . . . . .	7
<b>3</b>	<b>Contrastive Learning Approach</b>	<b>8</b>
3.1	Outline & Proposed Approach . . . . .	8
3.2	Materials and Dataset . . . . .	8
3.3	Experimental Method . . . . .	10
3.3.1	PreProcessing . . . . .	10
3.3.2	Contrastive Learning with SimCLR . . . . .	10
3.3.3	Attention-based Multiple Instance Learning . . . . .	11
3.4	Results and Discussion . . . . .	13
3.5	Conclusion . . . . .	15
<b>4</b>	<b>Ensemble Approach Using Clinical Data</b>	<b>16</b>
4.1	Outline & Proposed Approach . . . . .	16
4.2	Materials and Dataset . . . . .	17
4.3	Experimental Method . . . . .	17
4.3.1	WSI-based Classification . . . . .	17
4.3.2	Clinical data-based Classification . . . . .	20
4.3.3	Ensemble Learning of WSI and Clinical data . . . . .	21
4.4	Results and Discussion . . . . .	23
4.4.1	Comparisons of WSI Feature Extractor . . . . .	23

4.4.2	Comparisons of Model Performance with the Ensemble of Image and Clinical Data . . . . .	23
4.4.3	Attention Visualization of WSI . . . . .	26
4.4.4	Interpretability of Clinical Data Model . . . . .	26
4.5	Conclusion . . . . .	30
<b>5</b>	<b>Conclusion</b>	<b>31</b>
5.1	Conclusion . . . . .	31
5.2	Future Works . . . . .	31
	<b>Acknowledgment</b>	<b>32</b>
	<b>Reference</b>	<b>33</b>
	<b>Publication List</b>	<b>38</b>
	<b>Appendix A Attention Heatmap</b>	<b>39</b>
	<b>Appendix B Feature Importance with SHAP</b>	<b>43</b>

## List of Figures

1.1	Example of HE-stained images of gliomas. . . . .	2
1.2	OncoMatrix of the top 20 most frequent gene mutations in TCGA-LGG and TCGA-GBM. . . . .	4
3.1	Overview of the Contrastive Learning approach. . . . .	9
3.2	Image preprocessing flow. . . . .	10
4.1	Overview of the ensemble approach. . . . .	17
4.2	Correlation matrix. Numerical-Categorical: Correlation ratio, Categorical-Categorical: Cramer's V. . . . .	19
4.3	ROC curve. The results are based on (a) WSI, (b) clinical data, and (c) ensemble approaches, respectively. The AUCs for each fold and the average ROC curves are plotted. . . . .	27
4.4	Example of attention heatmaps and attention patches (MaxViT). . . . .	28
4.5	Feature importance for fold 0. (a): Beeswarm Plot. (b)-(f): Dependence Plot. . . . .	29
A.1	Attention heatmaps and attention patches (MaxViT). Ground Truth: IDH1, Prediction: IDH1. . . . .	39
A.2	Attention heatmaps and attention patches (MaxViT). Ground Truth: IDH1, Prediction: WT. . . . .	40
A.3	Attention heatmaps and attention patches (MaxViT). Ground Truth: WT, Prediction: IDH1. . . . .	41
A.4	Attention heatmaps and attention patches (MaxViT). Ground Truth: WT, Prediction: WT. . . . .	42
B.1	Feature importance for fold 1. . . . .	43
B.2	Feature importance for fold 2. . . . .	44
B.3	Feature importance for fold 3. . . . .	44
B.4	Feature importance for fold 4. . . . .	45
B.5	Feature importance for fold 5. . . . .	45
B.6	Feature importance for fold 6. . . . .	46
B.7	Feature importance for fold 7. . . . .	46



B.8	Feature importance for fold 8. . . . .	47
B.9	Feature importance for fold 9. . . . .	47

## List of Tables

3.1	Dataset details of the contrastive learning approach. . . . .	9
3.2	Hyperparameters of SimCLR. . . . .	11
3.3	Hyperparameters of ABMIL. . . . .	12
3.4	Experimental results. The 5-fold average performance is shown with $\pm$ SD. . . . .	14
4.1	Dataset details of the ensemble approach. . . . .	18
4.2	ABMIL hyperparameters optimized for each encoder. . . . .	21
4.3	LightGBM hyperparameters. . . . .	22
4.4	Experimental results. The 10-fold average performance is shown with $\pm$ SD. . . . .	25

# Chapter 1

## Introduction

### 1.1 Background

In the field of histopathology, pathologists diagnose disease by observing pathological specimens obtained by biopsy and stained with Hematoxylin and Eosin (H & E) under a microscope. Pathological diagnosis, also known as “final diagnosis” or “definite diagnosis”, is essential for disease detection and classification.

Recently, results of genetic testing such as immunohistochemistry (IHC), fluorescence in situ hybridization (FISH), and next-generation sequencing (NGS) have become increasingly important in cancer diagnosis in addition to phenotypic testing [1, 2, 3]. The diagnosis with histology and molecular information is called integrated diagnosis. However, these genetic tests are expensive, take a long turnaround times, and require specialized facilities.

On the other hand, researches on cancer detection, classification, and gene mutation prediction using AI technology have been reported [4, 5, 6, 7, 8]. Furthermore, studies have reported the prediction of gene mutations using deep learning from pathological images. However, these studies only used pathology images, which means they might not accurately predict cases where gene mutations do not appear in the images. Therefore, there are limitations to accurately predicting gene mutations from the phenotype of pathological images.

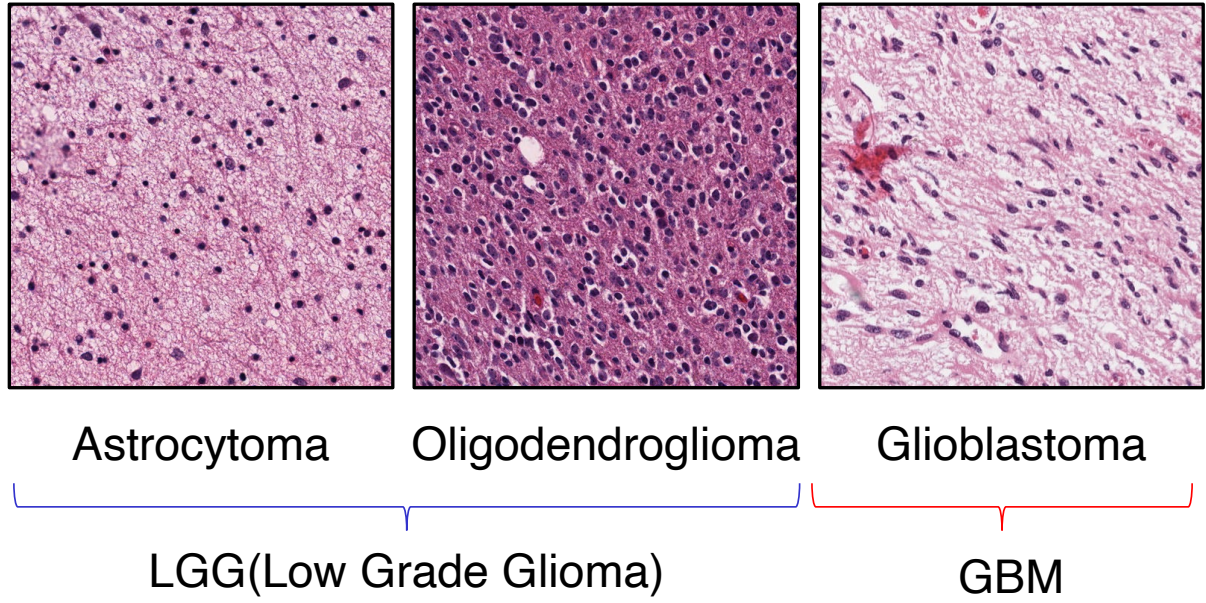


Figure 1.1: Example of HE-stained images of gliomas.

## 1.2 Classification of Gliomas and IDH Gene Mutation

According to the 2021 WHO classification of brain tumors [3], Adult-type diffuse gliomas are classified into three types as follows:

1. Astrocytoma, IDH-mutant,
2. Oligodendroglioma, IDH-mutant, and 1p/19q-codeleted,
3. Glioblastoma, IDH-wildtype.

Fig. 1.1 shows an example of pathological images of glioma. Astrocytoma and oligodendroglioma are usually called LGG (Low-Grade Glioma), and glioblastoma is called GBM. GBM is grade 4, and it is the highest grade of glioma [9]. In general, the 5-year survival rate of LGG patients is better than GBM patients [10].

IDH1 and IDH2 gene mutations are frequently found in LGG patients. Fig 1.2 shows an OncoMatrix of the top 20 most frequent gene mutations in glioma. In cases of glioma obtained from TCGA, the percentage of IDH mutations is 41.0% for IDH1 and 2.3% for IDH2 [11, 12]. Patients with IDH mutations have been reported to have a better prognosis than those without IDH mutations [13, 14]. Therefore, IDH mutations are helpful indicators for the classification of gliomas.

In pathological diagnosis, genotypic analysis is performed along with phenotypic analysis. If gene mutations can be identified without genetic testing, it will improve the efficiency of diagnosis and reduce the cost of treatment for patients.

### 1.3 Objective

IDH mutations include IDH1 and IDH2, but cases with IDH2 mutations are rare (Fig. 1.2). Therefore, this thesis focuses on the IDH1 mutation, the most frequent gene mutation in glioma patients. The purpose of this study is to accurately classify the presence or absence of IDH1 mutation from HE-stained pathological images using machine learning techniques. To improve classification performance, this study explores the effectiveness of applying contrastive learning using only pathological images and ensemble learning using both pathological images and clinical data. Large-scale H & E stained images and associated data from glioma patients are used [15].

The Contrastive Learning approach focused on the feature extractor and performed self-supervised learning using SimCLR. This approach confirmed the possibility of improving performance even with a limited amount of data.

The ensemble learning approach is an experiment focusing on the input dataset and interpretability. The ensemble approach uses a weakly supervised deep learning model and a LightGBM machine learning classifier to obtain IDH1 mutation versus wild-type. Even if histological features caused by gene mutations are unclear in pathological images, augmenting appropriate clinical metadata is expected to help improve the overall predictions. The ensemble approach of this thesis obtained accurate prediction in the fusion approach and was further amenable to interpretable features.

This paper is organized as follows. Chapter 2 introduces the related works to this study. Chapter 3 presents the approach with pre-training the feature extractor using the contrastive learning method. Chapter 4 introduces the proposed method that combines imaging and clinical data. Finally, Chapter 5 provides the overall conclusion.

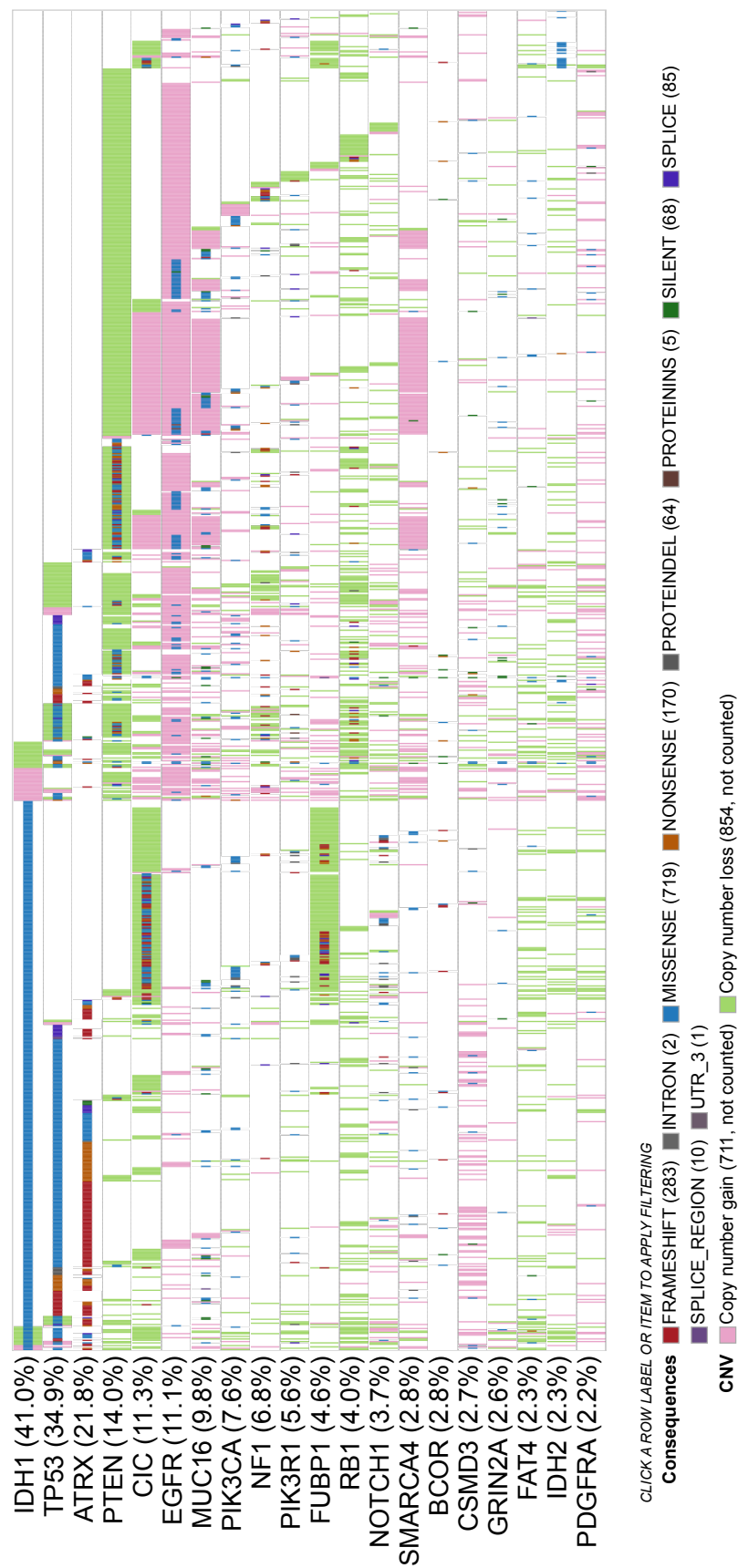


Figure 1.2: OncoMatrix of the top 20 most frequent gene mutations in TCGA-LGG and TCGA-GBM.

## Chapter 2

### Related Works

#### 2.1 Application of Deep Learning in Mutation Prediction from Pathological Images

Coudray *et al.* proposed a deep learning model to predict gene mutations directly from pathological images [4]. Although this study does not focus on gliomas, it is one of the early investigations to predict gene mutations from HE-stained pathological images.

Firstly, they trained InceptionV3 to classify lung adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and normal tissues. Subsequently, InceptionV3 was also trained to predict the status of ten common gene mutations in LUAD. Because a single slide image could have multiple gene mutations, the deep learning model was implemented as a multi-label classification task using a sigmoid activation function in the final layer instead of softmax. For training, slide images were divided into  $512 \times 512$  pixels patch images. The model predicted at the patch-level. Slide-level prediction results were obtained either by averaging the predicted probabilities of each patch or by counting the percentage of patches as positive.

Experimental results showed that the model was able to predict gene mutations, such as STK11, EGFR, FAT1, SETBP1, KRAS, and TP53, from image data alone. Their AUCs achieved 0.733 to 0.856. This study suggested that the deep learning model could detect cancer subtypes and gene mutations from image data alone, and this approach could potentially be applied to any cancer type.

## 2.2 GAN-based Data Augmentation

Liu *et al.* proposed a data augmentation method based on a Generative Adversarial Network (GAN) to improve the prediction performance of IDH mutations in glioma pathological images [5]. Deep learning requires a large amount of image data to develop models with good performance. However, it is difficult to collect disease cases in the medical field, leading to inevitably small sample size. Small sample size practically limit the performance of the model. In order to overcome this problem, this study introduced a GAN-based data augmentation. GAN consists of a generator and a discriminator, and the generator generates data resembling real samples, while the discriminator discriminates between the real and the fake. The generator generates more realistic data through this competition, and the discriminator tries to identify more accurately, enhancing each other's performance.

In this paper, two GAN models were designed to model IDH-wildtype and IDH-mutant data distributions separately. Progressive Growing of GAN (PG-GAN) was utilized. The size of the patch image was  $256 \times 256$  pixels. For GAN training, 12,000 patch images were randomly selected from each class. The network was configured to generate images with six levels of resolution. For IDH mutation prediction, ResNet50 was used as the backbone. To improve the prediction performance over the baseline model trained only with real images, GAN-generated images were gradually fed into ResNet50.

As a result, the experiment achieved an accuracy of 0.765 (AUC = 0.823) on the validation set and 0.794 (AUC = 0.927) on the test set without data augmentation. For the GAN-based data augmentation, 3,000 generated images were added to the training set. GAN-based data augmentation improved the accuracy to 0.853 (AUC = 0.868) on the validation set and 0.853 (AUC = 0.927) on the test set. In addition, age information was integrated into the image-based classification using a logistic regression classifier. It achieved an accuracy of 0.853 (AUC = 0.882) on the validation set and 0.882 (AUC = 0.931) on the test set.

This paper showed that augmenting the training dataset with GAN-generated fake images could improve model performance. It also suggested that better models could be built by integrating not only pathological images but also other information, such as age and MRI images.



## 2.3 Combining Pathological Images and Clinical Data

Jiang *et al.* focused only on grade 2 glioma. They developed end-to-end IDH mutation prediction models using ResNet18 as a feature extractor [6]. The impact of combining pathological images and clinical information on prediction performance was also investigated. In the preprocessing of slide images, patch images were extracted with a size of  $224 \times 224$  pixels.

The experiment resulted in an AUC of 0.667 for the model based on pathological images. As mentioned in section 2.2, age is a strong predictor for IDH mutations. The AUC for age was 0.689. Race was a weak predictor, with an AUC of 0.567. When race and pathological images were combined, the AUC increased to 0.687. Combining age and race resulted in an AUC of 0.711. Furthermore, when pathological images were combined with age and race, the AUC increased to 0.739.

The model in this study was trained solely on grade 2 glioma. The AUC was lower than other studies, which might be attributed to 85% of the grade 2 glioma patients in the dataset having IDH mutations.

On the other hand, it was found that combining clinical information improved classification performance. It was also found that the type of clinical information trained affects performance. These results suggest further improving performance by adding appropriate clinical information rather than predicting from pathological images alone.

## Chapter 3

# Contrastive Learning Approach

### 3.1 Outline & Proposed Approach

This chapter discusses the learning feature representations using Contrastive Learning and its effectiveness. As mentioned in section 2.2, deep learning requires a large amount of data, but the number of cases that can be collected is limited in the medical field. In addition, data annotation requires a significant amount of time and effort.

Therefore, this study proposes an efficient learning method for pathological images using self-supervised learning to address such constrained datasets. Figure 3.1 shows an overview of this experiment. The proposed approach has two steps. In step 1, Contrastive Learning was conducted as the pre-training method. Contrastive Learning has the advantage of being independent of the architecture of the feature extractor. The SimCLR [16] was employed in this experiment. In step 2, feature extraction and binary classification were performed. Feature representations were extracted using a feature extractor trained in step 1. Attention-based Multiple Instance Learning (ABMIL) [17] was used for slide-level binary classification. This method can be trained without patch-level annotations.

### 3.2 Materials and Dataset

This study utilized Whole-Slide Images (WSIs) and gene mutation data for TCGA-GBM and TCGA-LGG obtained from the NCI Genomic Data Commons (GDC) [12]. The Cancer Genome Atlas (TCGA) is a large-scale cancer genome analysis project that collects and makes open data such as cancer genomes, epigenomes, and transcriptomes [11].

The WSIs were annotated at the slide-level with the presence or absence of IDH1 mutation (IDH1 or Wild-Type). Table 3.1 summarizes the number of data. The data were collected from 324 patients with IDH1 mutation and 225 patients with Wild-Type, and 607 WSIs were used for each label.

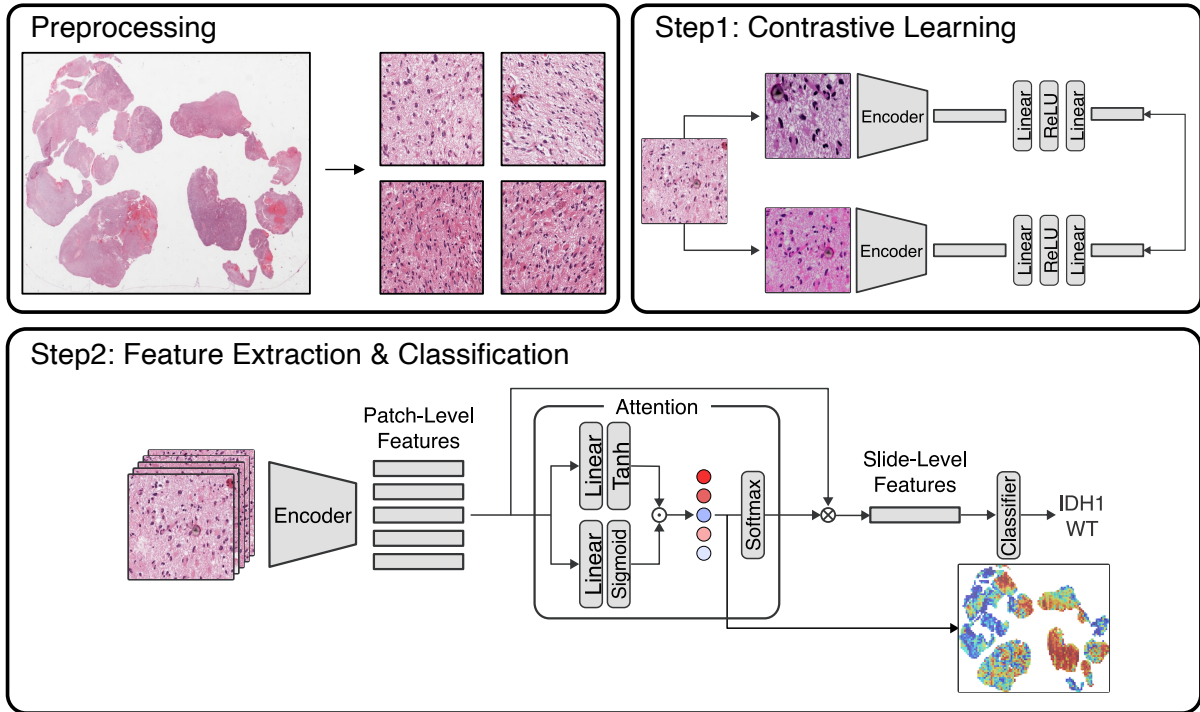


Figure 3.1: Overview of the Contrastive Learning approach.

Table 3.1: Dataset details of the contrastive learning approach.

	IDH1	Wild-Type	Total
# of cases	324	225	549
# of WSIs	607	607	1,214
# of patches	921,937	776,269	1,698,206

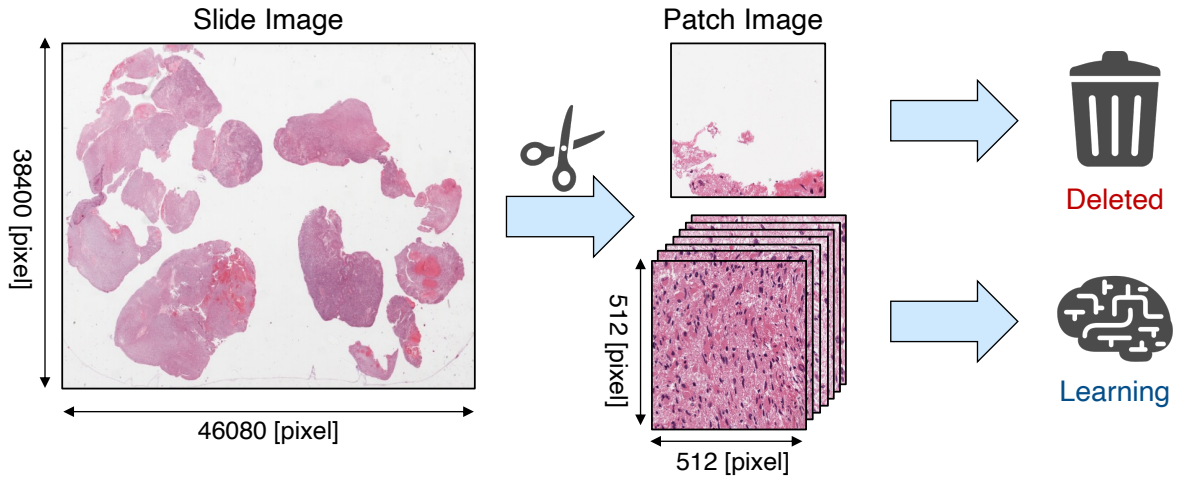


Figure 3.2: Image preprocessing flow.

### 3.3 Experimental Method

#### 3.3.1 PreProcessing

Figure 3.2 shows the preprocessing flow. In the preprocessing, the WSIs were divided into  $512 \times 512$  pixels patches because the WSIs size was too large to analyze. The white background regions were removed, and the regions containing sufficient cellular tissue were used as the dataset for training image model.

The dataset was split into a training set (90%) and a test set (10%). The dataset was divided carefully so that images of one patient were not sampled into the same set. The classification performance was evaluated using 5-fold stratified cross-validation. The test set was referenced only once after training to evaluate the model in each fold.

#### 3.3.2 Contrastive Learning with SimCLR

Contrastive Learning is a self-supervised learning method. Contrastive learning aims to extract better feature representations for the classification task.

In this work, SimCLR [16] was applied for Contrastive Learning and EfficientNetV2 B0 as the encoder for feature extraction [18]. This method trains models to bring similar images close in feature space and push different images apart. In SimCLR, random transformations are applied to one input image to obtain a pair of two transformed images. This pair is input to an encoder and projected into the latent space using MLP. The parameters are updated to maximize the similarity of the same pair of images.

Contrastive Learning was conducted using only the training set. The total number of patches in the training set is 1,581,795. If all of these were used for training, training would take too much time due to disk I/O. Therefore, a portion of the training set was

Table 3.2: Hyperparameters of SimCLR.

Hyperparameter	SimCLR
Batch size	128
Loss	NT-Xent
Epochs	100
Optimizer	LARS
Weight decay	1.0e-6
Scheduler	CosineAnnealingLR
Learning rate	0.15 ( $=0.3 \times \text{Batch size} / 256$ )
Augmentation	ResizedCrop ( $256 \times 256$ ), ColorJitter, HorizontalFlip, GrayScale
Temperature	0.5
Encoder	EfficientNetV2 B0

randomly sampled for training. 1% and 10% of the number of patches in the training set were randomly sampled and used, limiting the number of patches used for training. 15,818 patch images were utilized for 1% of the training set, and 158,180 patch images were used for 10%. For random sampling, the same number of patches were selected for each label. The impact of the different amounts of data, 1% and 10%, on classification performance, was compared. Additionally, it was also compared whether the encoder, EfficientNetV2 B0, had been pre-trained on ImageNet. Table 3.2 summarizes the hyperparameters of SimCLR.

### 3.3.3 Attention-based Multiple Instance Learning

Pathological images do not generally have uniform features throughout the image. However, dataset of this study had no patch-level labels; only slide-level labels were available. Therefore, in this study, attention-based deep multiple instance learning (AB-MIL) [17] was employed. ABMIL is a weakly supervised learning method that classifies only at the slide-level labels without requiring patch-level annotations. Attention-based learning automatically identifies the regions of interest in the WSI. Thus, the DL model is expected to capture histological features caused by IDH1 mutation.

ABMIL works as follows. First, the patch images are input to the encoder and embedded into patch-level feature vectors. In this experiment, feature representations

Table 3.3: Hyperparameters of ABMIL.

Hyperparameter	ABMIL
Batch size	1
Loss	Cross Entropy
Epochs	100
GatedAttention Hidden layer	1280-512-256-1
Dropout rate	0.25
Optimizer	SGD
Weight decay	3.0e-7
Scheduler	CosineAnnealingLR
Learning rate	3.0e-5

were extracted from all patch images using EfficientNetV2-B0 pre-trained by Contrastive Learning. A feature vector of 1280-dimensions was obtained for each patch. Next, each patch-level feature is weighted using the attention mechanism, and the patch-level features are aggregated into slide-level features. This weight is called attention score. The attention scores can also be used to visualize an attention heatmap.

The gated attention mechanism is represented as:

$$\mathbf{z} = \sum_{k=1}^K a_k \mathbf{h}_k \quad (3.1)$$

$$a_k = \frac{\exp \{ \mathbf{w}^\top (\tanh(\mathbf{V} \mathbf{h}_k^\top) \odot \text{sigm}(\mathbf{U} \mathbf{h}_k^\top)) \}}{\sum_{j=1}^K \exp \{ \mathbf{w}^\top (\tanh(\mathbf{V} \mathbf{h}_j^\top) \odot \text{sigm}(\mathbf{U} \mathbf{h}_j^\top)) \}} \quad (3.2)$$

where  $\mathbf{h}$  is the input vector,  $\mathbf{w}$ ,  $\mathbf{V}$ , and  $\mathbf{U}$  are trainable parameters and  $\odot$  is an element-wise multiplication.  $K$  is the number of patches in one slide image. Slide-level prediction results are obtained by inputting the features  $\mathbf{z}$  into the classifier, aggregated from the patch-level to the slide-level by attention mechanism.

Classification performance was evaluated with stratified 5-fold cross-validation. Table 3.3 summarizes the hyperparameters of ABMIL.

### 3.4 Results and Discussion

Table 3.4 summarizes the experimental results, including the mean and standard deviation from the 5-fold cross-validation.

The model without pre-training had the worst scores on all evaluation metrics. The average ROC-AUC was 0.500, indicating that it was not capable of predicting IDH1.

The model pre-trained on ImageNet and not trained on SimCLR showed the third-best classification performance, as shown in the “ImageNet” row of table 3.4. This model exhibited good performance even though it was not trained on pathological images of gliomas. This can be attributed to the diversity and complexity of the ImageNet dataset, which made it robust against out-of-distribution pathological images.

“SimCLR (1%)” and “SimCLR (10%)” mean models pre-trained with SimCLR using only pathological images, without ImageNet pre-training. Although the scores were lower than “ImageNet”, it was found that a certain level of performance could be achieved by learning only pathological images.

“SimCLR (1%) + ImageNet” and “SimCLR (10%) + ImageNet” are fine-tuned models with SimCLR training using weights pre-trained on ImageNet as the initial values. The experimental results showed that “SimCLR (10%) + ImageNet” had the best performance. The average scores for this model were as follows: accuracy of 0.794, precision of 0.816, recall of 0.762, f1-score of 0.787, ROC-AUC of 0.900, PR-AUC(WT) of 0.919, and PR-AUC(IDH1) of 0.888. The second-best performance was achieved by “SimCLR (1%) + ImageNet”. As mentioned the above, the “ImageNet” model, despite not being trained on pathological images, achieved the third-best performance. However, when comparing “ImageNet” and “SimCLR (1%) + ImageNet”, the SimCLR-trained model had all higher metrics except for recall and F1-score. Furthermore, comparing “ImageNet” and “SimCLR (10%) + ImageNet”, it was observed that the scores was improved for all metrics except for recall. These results suggest that pre-training with Contrastive Learning can effectively improve prediction performance on small dataset.

Additionally, it was confirmed that increasing the amount of training data from 1% to 10% improved evaluation scores on all metrics. If all patches were used for model training, disk I/O would become a bottleneck, and processing would take too long. Therefore, in this experiment, the amount of training data for SimCLR was limited to 10%. However, there is the potential for further improvement in classification performance by increasing the training data for SimCLR.

Table 3.4: Experimental results. The 5-fold average performance is shown with  $\pm$  SD.

Pre-train Method	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC(WT)	PR-AUC(IDH1)
No Pre-training	0.500 $\pm$ 0.000	0.200 $\pm$ 0.245	0.400 $\pm$ 0.490	0.267 $\pm$ 0.327	0.500 $\pm$ 0.032	0.502 $\pm$ 0.014	0.511 $\pm$ 0.006
ImageNet	0.765 $\pm$ 0.017	0.747 $\pm$ 0.029	<b>0.808 <math>\pm</math> 0.034</b>	0.775 $\pm$ 0.013	0.841 $\pm$ 0.016	0.851 $\pm$ 0.016	0.842 $\pm$ 0.018
SimCLR (1%)	0.658 $\pm$ 0.018	0.661 $\pm$ 0.023	0.650 $\pm$ 0.045	0.654 $\pm$ 0.023	0.730 $\pm$ 0.009	0.734 $\pm$ 0.005	0.756 $\pm$ 0.013
SimCLR (10%)	0.712 $\pm$ 0.031	0.713 $\pm$ 0.029	0.708 $\pm$ 0.051	0.710 $\pm$ 0.036	0.817 $\pm$ 0.016	0.846 $\pm$ 0.009	0.814 $\pm$ 0.020
SimCLR (1%) + ImageNet	0.779 $\pm$ 0.014	0.801 $\pm$ 0.018	0.742 $\pm$ 0.029	0.770 $\pm$ 0.016	0.872 $\pm$ 0.013	0.894 $\pm$ 0.011	0.856 $\pm$ 0.015
SimCLR (10%) + ImageNet	<b>0.794 <math>\pm</math> 0.017</b>	<b>0.816 <math>\pm</math> 0.027</b>	0.762 $\pm$ 0.015	<b>0.787 <math>\pm</math> 0.015</b>	<b>0.900 <math>\pm</math> 0.009</b>	<b>0.919 <math>\pm</math> 0.009</b>	<b>0.888 <math>\pm</math> 0.005</b>



### 3.5 Conclusion

In this chapter, the effectiveness of pre-training with Contrastive Learning was investigated to improve the prediction performance for IDH1 mutation. SimCLR was employed for Contrastive Learning and attention-based multiple instance learning for slide-level classification. The experimental results suggested that pre-training with SimCLR, even with small amounts of data, contributed to improved classification performance.

The goal is to establish the optimal learning method for predicting IDH1 mutations. In the future, prediction performance should be compared using other self-supervised learning methods such as DINOv2, MoCo v3, SimSiam, and BYOL [19, 20, 21, 22].

## Chapter 4

# Ensemble Approach Using Clinical Data

### 4.1 Outline & Proposed Approach

This chapter discusses the effectiveness of ensemble learning by fusing pathological images with clinical data.

Whole-Slide Images, the research material, are generally very high-resolution image data. Their data size ranges from several MB to several GB per image. Therefore, analysis of pathological images using deep learning models requires a huge amount of computational resources and time. In addition, the information obtained from pathological images is mainly limited to histological features. Therefore, it is difficult to obtain accurate prediction results from the images lacking histological features caused by gene mutations. For these reasons, there are certain limitations to accurately predict gene mutations from the phenotype of pathological images. On the other hand, as mentioned in sections 2.2 and 2.3, the overall prediction performance can be expected to improve by adding clinical information such as age.

Therefore, this study focuses on improving prediction performance by adding appropriate clinical information and investigates a prediction method for the multimodal model. Fig. 4.1 shows the overall proposed approach of this study. In the WSI-based approach, a weakly supervised learning method called ABMIL performs slide-level classification. In the clinical data-based approach, classification is performed using LightGBM. The average confidence of both approaches is calculated and used as the final output. Thus, separate deep learning and machine learning models were trained on the image and clinical data. The final prediction results were obtained by ensembling their outputs. This approach is expected to provide a broader coverage of the information necessary for predicting gene mutations.

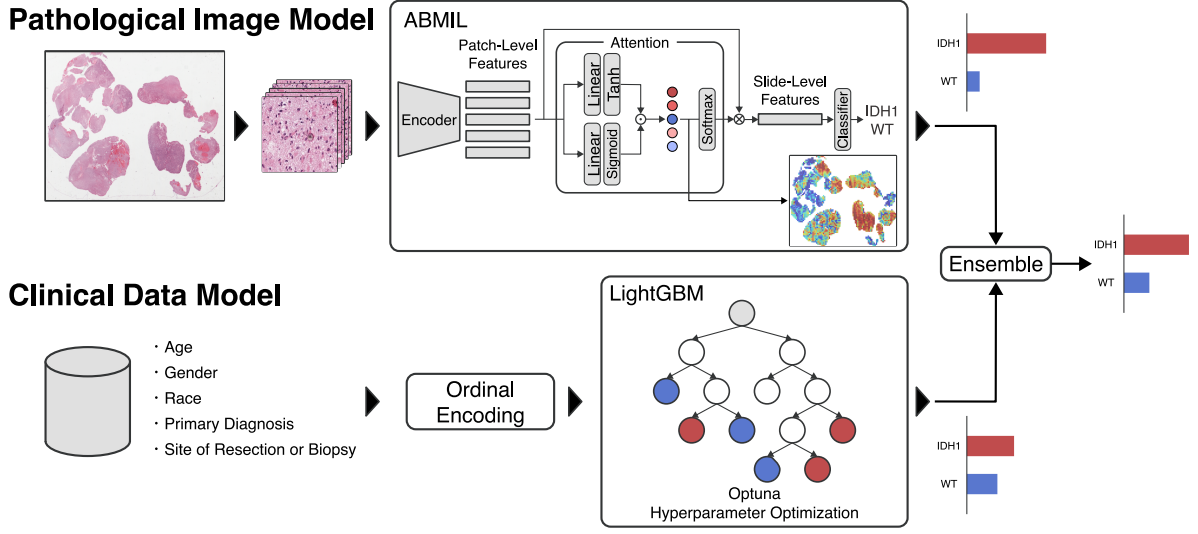


Figure 4.1: Overview of the ensemble approach.

## 4.2 Materials and Dataset

Whole-Slide Images (WSIs), gene mutation data, and clinical data for gliomas were used in this experiment. The data were obtained from the NCI Genomic Data Commons (GDC) [12, 11]. The WSIs and clinical data were annotated with the presence or absence of IDH1 mutation (IDH1 or Wild-Type). Table 4.1 shows the summary of the dataset. This thesis chose five non-redundant clinical variables and omitted the variables with too many missing values. The data were collected from 321 patients with IDH1 mutation and 225 patients with Wild-Type, and 603 WSIs were used for each label. Only age is a numerical variable. All variables except age are categorical variables. Fig. 4.2 shows the correlation matrix of the variables. The correlations between age and categorical variables were calculated as the correlation ratio, and the associations between categorical variables were calculated as the Cramer's V.

## 4.3 Experimental Method

### 4.3.1 WSI-based Classification

In the preprocessing of the WSIs, unlike chapter 3, the WSIs were divided into  $256 \times 256$  pixels patch images. The white background regions were removed, and the regions containing sufficient cellular tissue were used as the dataset for training the image model. The patch images were extracted where the area ratio of the tissue to the total area of the image was more than 80%, using the python library histolab [23]. Attention-based deep multiple instance learning (ABMIL) [17] was used for slide-level classification.

Table 4.1: Dataset details of the ensemble approach.

	IDH1	Wlid-Type	Total
# of cases	321	225	546
# of WSIs	603	603	1206
# of patches	7,212,076	4,492,145	11,704,221
Mean <b>Age</b> $\pm$ SD (years)	41.1 $\pm$ 12.2	54.2 $\pm$ 15.7	46.5 $\pm$ 15.2
<b>Gender</b>			
Male	186 (57.9%)	128 (56.9%)	314 (57.5%)
Female	135 (42.1%)	97 (43.1%)	232 (42.5%)
<b>Race</b>			
White	297 (92.5%)	202 (89.8%)	499 (91.4%)
Black or african american	12 (3.7%)	15 (6.7%)	27 (4.9%)
Asian	5 (1.6%)	4 (1.8%)	9 (1.6%)
American indian or alaska native	0 (0%)	1 (0.4%)	1 (0.2%)
Not reported	7 (2.2%)	3 (1.3%)	10 (1.8%)
<b>Primary diagnosis</b>			
Astrocytoma, anaplastic	63 (19.6%)	50 (22.2%)	113 (20.7%)
Astrocytoma, NOS	39 (12.1%)	13 (5.8%)	52 (9.5%)
Oligodendroglioma, anaplastic	48 (15%)	18 (8%)	66 (12.1%)
Oligodendroglioma, NOS	78 (24.3%)	20 (8.9%)	98 (17.9%)
Glioblastoma	4 (1.2%)	100 (44.4%)	104 (19%)
Mixed glioma	89 (27.7%)	24 (10.7%)	113 (20.7%)
<b>Site of resection or biopsy</b>			
Cerebrum	265 (82.6%)	109 (48.4%)	374 (68.5%)
Brain, NOS	46 (14.3%)	114 (50.7%)	160 (29.3%)
Temporal lobe	5 (1.6%)	0 (0%)	5 (0.9%)
Frontal lobe	3 (0.9%)	2 (0.9%)	5 (0.9%)
Occipital lobe	1 (0.3%)	0 (0%)	1 (0.2%)
Parietal lobe	1 (0.3%)	0 (0%)	1 (0.2%)

WSI: Whole-Slide Image, SD: Standard Deviation, NOS: Not Otherwise Specified

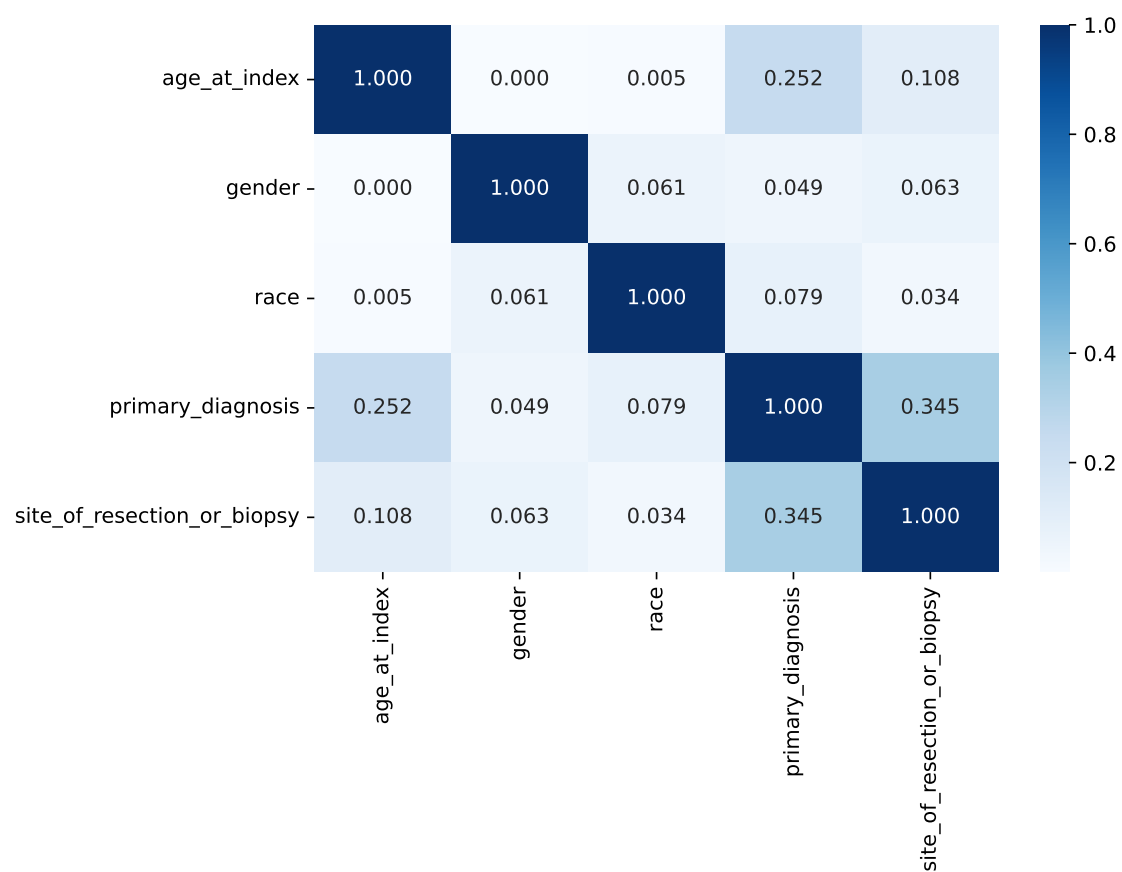


Figure 4.2: Correlation matrix. Numerical-Categorical: Correlation ratio, Categorical-Categorical: Cramer's V.

First, eight DL models were compared to identify the optimal encoder for the purpose. The DL models selected for comparison included Convolutional Neural Network (CNN) models, Transformer models, and hybrid models combining Convolution and Transformer architecture. All these encoders were pre-trained on ImageNet. The output of average pooling was used as feature vectors. The patch images were input to the encoder without data augmentation, such as flipping, rotation, or cropping. However, the CoAtNet was an exception. The patch images were resized into  $224 \times 224$  pixels because an error occurred with  $256 \times 256$  pixels image. The patch images were embedded into patch-level feature vectors of 512-4096 dimensions by each encoder.

Table 4.2 summarizes the ABMIL hyperparameters for each encoder. All models were trained for at least 50 epochs and up to a maximum of 100 epochs. Early stopping is used when validation loss has not decreased continuously by more than 20 epochs. Other hyperparameters, such as learning rate, were optimized by Optuna [24]. Hyperparameters with high ROC-AUC were adopted and used for ensemble learning.

In addition, an experiment using Clustering-constrained Attention Multiple Instance Learning (CLAM) [25] was also performed for comparison. Note that the number of parameters is different between CLAM’s ResNet50 and the ResNet50 used in this thesis because the final layer used is different.

The classification performances were evaluated by 10-fold monte carlo cross-validation. The dataset was randomly divided into a training set (80%), a validation set (10%), and a test set (10%) for each fold. If one patient had multiple WSIs, they were carefully divided so that all would be in the same set. The same cross-validation was used to evaluate model performance in the clinical data-based classification described in the next section.

Finally, Attention Heatmaps were generated by visualizing ABMIL’s attention score. By visualizing the trained attention heatmap, it is also possible to identify the regions important for classification.

### 4.3.2 Clinical data-based Classification

Five clinical features were used, including age, gender, race, primary diagnosis, and site of resection or biopsy. “Site of Resection or Biopsy” represents the anatomical site of malignant tumors in the patient’s brain (Table 4.1). Because all features were categorical variables except age, they were converted into numerical values using the Ordinal Encoder. LightGBM [26] was employed as the classification model for the clinical data. LightGBM can handle missing values directly and is a lightweight operation. The hyperparameters of LightGBM were optimized by Optuna. Table 4.3 summarizes the hyperparameters for LightGBM. In addition, feature importance was examined using SHAP (SHapley Additive exPlanations) [27].

Table 4.2: ABMIL hyperparameters optimized for each encoder.

Hyperparameter	VGG16	InceptionV3	ResNet50	EfficientNetV2 B0
Batch size	1	1	1	1
Loss	CrossEntropy	CrossEntropy	CrossEntropy	CrossEntropy
Epochs	100	100	100	100
Early stopping patience	20	20	-	-
GatedAttention Hidden layer	4096-512-256-1	2048-512-256-1	2048-512-256-1	1280-512-256-1
Dropout rate	0.25	0.25	0.25	0.25
Optimizer	SGD	SGD	SGD	SGD
Weight decay	3.0e-7	1.6e-6	1.6e-6	1.6e-6
Scheduler	CosineLRScheduler	CosineLRScheduler	CosineLRScheduler	CosineLRScheduler
Peak lr	3.0e-5	3.0e-4	3.0e-4	3.0e-4
Min lr	2.0e-8	1.5e-8	1.5e-8	1.5e-8
Warmup epochs	8	12	12	12
Warmup lr init	5.4e-8	7.6e-8	7.6e-8	7.6e-8

---

Hyperparameters	EfficientNetV2 B1	SwinTransformerV2	CoAtNet	MaxViT
Batch size	1	1	1	1
Loss	CrossEntropy	CrossEntropy	CrossEntropy	CrossEntropy
Epochs	100	100	100	100
Early stopping patience	20	20	20	20
GatedAttention Hidden layer	1280-512-256-1	768-512-256-1	768-512-256-1	512-512-256-1
Dropout rate	0.25	0.25	0.25	0.25
Optimizer	SGD	SGD	SGD	SGD
Weight decay	1.6e-6	1.2e-7	3.0e-7	1.6e-6
Scheduler	CosineLRScheduler	CosineAnnealingLR	CosineAnnealingLR	CosineLRScheduler
Peak lr	3.0e-4	1.6e-4	3.0e-5	3.0e-4
Min lr	1.5e-8	8.7e-7	1.0e-7	1.5e-8
Warmup epochs	12	-	-	12
Warmup lr init	7.6e-8	-	-	7.6e-8

### 4.3.3 Ensemble Learning of WSI and Clinical data

After training each model separately, ensemble learning was conducted by taking the average of the probabilities from both models to obtain a final prediction. This is called “Soft Voting”. Soft Voting has the advantage that any classifier can be used. This ensemble approach was employed since the pipeline of this thesis is a fusion of a deep learning model and a decision tree model. For example, consider a case where the WSI-based model predicts a probability of IDH1 to be 0.7 while the clinical data-based model predicts a probability of 0.5. Using Soft Voting, the ensemble results would be calculated as  $(0.7 + 0.5)/2 = 0.6$ . In this way, final prediction can be obtained by considering pathological images and clinical data.

Table 4.3: LightGBM hyperparameters.

Hyperparameter	LightGBM
objective	binary
boosting_type	gbdt
importance_type	split
learning_rate	0.08171120044551923
reg_alpha	1.2405777881849258e-08
reg_lambda	1.5937337120820927
num_leaves	27
max_depth	-1
colsample_bytree	0.6215618710838064
subsample	0.4438242926449302
subsample_freq	2
subsample_for_bin	200000
min_child_samples	26
min_child_weight	0.001
min_split_gain	0.0
n_estimators	100
num_iterations	268

Ensemble Learning is a method of combining multiple models to achieve higher performance than individual models. Such a late fusion approach of combining pathological images and clinical data is expected to complement each other and improve the prediction accuracy of gene mutations.



## 4.4 Results and Discussion

### 4.4.1 Comparisons of WSI Feature Extractor

The classification performance of the WSI-based, the clinical data-based, and the ensemble approaches were evaluated using 10-fold monte carlo cross-validation. Accuracy, precision, recall, F1-score, the area under the ROC curve (ROC-AUC), and the area under the precision-recall curve (PR-AUC) were calculated for each approach. Table 4.4 summarizes the experimental results, including each metric's mean and standard deviation across the 10-fold cross-validation. The classification performance of each encoder is shown in the WSI row of Table 4.4.

As a result, CoAtNet had the highest scores for many evaluation metrics, achieving an average accuracy of 0.752 over 10-fold cross-validation. In contrast, VGG16 showed the lowest performance. EfficientNet showed performance comparable to other encoders despite having fewer parameters. The AUC is an important metric as it allows evaluating model performance independently of the threshold. Additionally, because the dataset is balanced, not imbalanced, ROC-AUC is the most interesting metric in this study. MaxViT and CoAtNet had ROC-AUCs of 0.823 and 0.821, respectively. These models achieved higher results than 0.820. These models are hybrids of CNN and Vision Transformer architecture. Their prediction performance was found to be relatively higher than that of purely convolutional or transformer-based models. Apart from MaxViT and CoAtNet, the other encoders showed almost identical ROC-AUC scores.

### 4.4.2 Comparisons of Model Performance with the Ensemble of Image and Clinical Data

The classification performance of the clinical data-based and the ensemble approach is shown in Table 4.4. The LightGBM for clinical data achieved a ROC-AUC of 0.782. The ensemble of WSI and clinical data improved the average performance across all models and all metrics, except for the precision of CoAtNet.

Next, the best encoders of WSI and ensemble were compared for each metric. For accuracy and precision, CoAtNet from WSI was lesser than MaxViT from the ensemble. The WSI's CoAtNet had an accuracy of 0.752, while the ensemble's MaxViT had an accuracy of 0.778. As for precision, the both had an average of 0.765, but MaxViT from the ensemble had a smaller standard deviation. For recall, WSI's ResNet50(CLAM) was 0.819, and the ensemble's ResNet50(ABMIL) was 0.859. It was also found that the ensemble substantially improved the recall for all encoders. For the F1-score, ResNet50(ABMIL) from the ensemble was also higher than CoAtNet from WSI. As for ROC-AUC, MaxViT was the highest score compared to the other encoders for both WSI and the ensemble,

with the ROC-AUC improving to 0.852 for the ensemble. PR-AUC(WT) was the value when WT was treated as a positive class and IDH1 as a negative, and PR-AUC(IDH1) was the value when IDH1 was positive. For PR-AUC(WT), the highest result in WSI was 0.830 for CoAtNet, and the highest result in the ensemble was 0.868 for MaxViT. For PR-AUC(IDH1), the highest result in WSI was 0.819 for MaxViT, and the highest result in the ensemble was 0.833 for ResNet50(CLAM). In particular, MaxViT showed the best performance, achieving an accuracy of 0.778 and a ROC-AUC of 0.852. These scores were the highest scores in this thesis.

The obtained results suggested that both WSI and clinical data could be used to predict IDH1 mutation status, and combining them could enhance the prediction performance. It was considered that WSI-based and clinical data-based predictions complemented each other, leading to improved performance.

Fig. 4.3 shows the ROC curves. The ROC curves for each fold, the averaged ROC curve, and the confidence band were plotted. The confidence band represents  $\pm 1$  standard deviation of the averaged ROC curve. Fig. 4.3(a-c) displays the ROC curves obtained from the WSI, clinical data, and the ensemble approach, respectively. For the ROC curve of the WSI, the MaxViT results with the highest AUC was used.

Table 4.4: Experimental results. The 10-fold average performance is shown with  $\pm$  SD.

Models	#Params	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC(WT)	PR-AUC(IDH1)
<b>Clinical</b>								
LightGBM [26]	-	0.719 $\pm$ 0.073	0.715 $\pm$ 0.087	0.774 $\pm$ 0.070	0.741 $\pm$ 0.066	0.782 $\pm$ 0.076	0.818 $\pm$ 0.056	0.759 $\pm$ 0.082
<b>WSI</b>								
VGG16 [28]	134.26M	0.700 $\pm$ 0.059	0.698 $\pm$ 0.088	0.747 $\pm$ 0.057	0.720 $\pm$ 0.064	0.780 $\pm$ 0.063	0.777 $\pm$ 0.064	0.771 $\pm$ 0.083
InceptionV3 [29]	21.79M	0.735 $\pm$ 0.059	0.743 $\pm$ 0.088	0.775 $\pm$ 0.057	0.752 $\pm$ 0.064	0.798 $\pm$ 0.063	0.808 $\pm$ 0.064	0.777 $\pm$ 0.083
ResNet50 [30] (ABMIL)	23.51M	0.733 $\pm$ 0.048	0.738 $\pm$ 0.074	0.753 $\pm$ 0.058	0.743 $\pm$ 0.058	0.805 $\pm$ 0.048	0.805 $\pm$ 0.046	0.789 $\pm$ 0.077
EfficientNetV2 B0 [18]	5.86M	0.736 $\pm$ 0.049	0.744 $\pm$ 0.073	0.758 $\pm$ 0.070	0.747 $\pm$ 0.053	0.799 $\pm$ 0.052	0.811 $\pm$ 0.044	0.769 $\pm$ 0.082
EfficientNetV2 B1 [18]	6.86M	0.743 $\pm$ 0.074	0.750 $\pm$ 0.087	0.762 $\pm$ 0.108	0.751 $\pm$ 0.084	0.803 $\pm$ 0.064	0.818 $\pm$ 0.045	0.772 $\pm$ 0.098
SwinTransformerV2 [31]	27.58M	0.730 $\pm$ 0.065	0.739 $\pm$ 0.088	0.736 $\pm$ 0.097	0.734 $\pm$ 0.084	0.799 $\pm$ 0.059	0.800 $\pm$ 0.049	0.784 $\pm$ 0.082
CoAtNet [32]	40.96M	<b>0.752 <math>\pm</math> 0.044</b>	<b>0.765 <math>\pm</math> 0.082</b>	0.769 $\pm$ 0.077	<b>0.761 <math>\pm</math> 0.048</b>	0.821 $\pm$ 0.054	<b>0.830 <math>\pm</math> 0.046</b>	0.802 $\pm$ 0.082
MaxViT [33]	28.64M	0.739 $\pm$ 0.051	0.743 $\pm$ 0.066	0.757 $\pm$ 0.111	0.745 $\pm$ 0.070	<b>0.823 <math>\pm</math> 0.049</b>	0.821 $\pm$ 0.050	<b>0.819 <math>\pm</math> 0.068</b>
ResNet50 (CLAM) [25]	8.54M	0.725 $\pm$ 0.080	0.710 $\pm$ 0.098	<b>0.819 <math>\pm</math> 0.076</b>	0.755 $\pm$ 0.068	0.806 $\pm$ 0.063	0.810 $\pm$ 0.047	0.780 $\pm$ 0.095
<b>Ensemble</b>								
VGG16	134.26M	0.749 $\pm$ 0.067	0.727 $\pm$ 0.084	0.834 $\pm$ 0.057	0.775 $\pm$ 0.062	0.820 $\pm$ 0.068	0.841 $\pm$ 0.056	0.800 $\pm$ 0.096
InceptionV3	21.79M	0.768 $\pm$ 0.067	0.750 $\pm$ 0.084	0.845 $\pm$ 0.057	0.791 $\pm$ 0.062	0.835 $\pm$ 0.068	0.855 $\pm$ 0.056	0.814 $\pm$ 0.096
ResNet50 (ABMIL)	23.51M	0.777 $\pm$ 0.054	0.751 $\pm$ 0.071	<b>0.859 <math>\pm</math> 0.045</b>	<b>0.800 <math>\pm</math> 0.049</b>	0.844 $\pm$ 0.059	0.860 $\pm$ 0.052	0.832 $\pm$ 0.078
EfficientNetV2 B0	5.86M	0.767 $\pm$ 0.060	0.752 $\pm$ 0.082	0.831 $\pm$ 0.054	0.787 $\pm$ 0.055	0.839 $\pm$ 0.063	0.857 $\pm$ 0.055	0.824 $\pm$ 0.092
EfficientNetV2 B1	6.86M	0.773 $\pm$ 0.056	0.761 $\pm$ 0.082	0.831 $\pm$ 0.046	0.791 $\pm$ 0.050	0.842 $\pm$ 0.066	0.863 $\pm$ 0.055	0.818 $\pm$ 0.100
SwinTransformerV2	27.58M	0.767 $\pm$ 0.061	0.745 $\pm$ 0.081	0.845 $\pm$ 0.045	0.790 $\pm$ 0.056	0.839 $\pm$ 0.064	0.856 $\pm$ 0.053	0.829 $\pm$ 0.085
CoAtNet	40.96M	0.766 $\pm$ 0.073	0.753 $\pm$ 0.086	0.824 $\pm$ 0.069	0.784 $\pm$ 0.068	0.846 $\pm$ 0.067	0.866 $\pm$ 0.056	0.828 $\pm$ 0.086
MaxViT	28.64M	<b>0.778 <math>\pm</math> 0.062</b>	<b>0.765 <math>\pm</math> 0.075</b>	0.827 $\pm$ 0.088	0.792 $\pm$ 0.065	<b>0.852 <math>\pm</math> 0.059</b>	<b>0.868 <math>\pm</math> 0.050</b>	0.831 $\pm$ 0.080
ResNet50 (CLAM)	8.54M	0.771 $\pm$ 0.056	0.745 $\pm$ 0.076	0.857 $\pm$ 0.070	0.794 $\pm$ 0.056	0.843 $\pm$ 0.057	0.857 $\pm$ 0.050	<b>0.833 <math>\pm</math> 0.081</b>

### 4.4.3 Attention Visualization of WSI

Attention scores were visualized to identify important regions that contributed to classification. It is expected to lead to interpreting morphological features associated with IDH1 mutation.

Although non-overlapping patches were used for the training and evaluation of the model, the WSIs were divided into patches with 70% overlap to generate a smooth attention heatmap. The patch images were embedded into lower dimensions by the encoder and inferred by ABMIL to obtain attention scores. The attention score was calculated by equation 3.2. The more patch images, the smaller the value. Therefore, attention scores were converted to values ranging from 0 to 1 and normalized. If multiple patches overlapped in the same region, the scores for that region were averaged. In addition, gaussian blur was applied to draw the heatmap smoothly.

Fig. 4.4 illustrates a heatmap visualizing the attention score of ABMIL, employing MaxViT as the encoder. Fig. 4.4 displays cases with a positive IDH1 mutation, which ABMIL also correctly predicted as IDH1 was positive. This visualization helps to identify regions of interest that contribute to the prediction without requiring patch-level annotations. Regions with high-attention scores are colored in red, while those with low-attention scores are colored in blue.

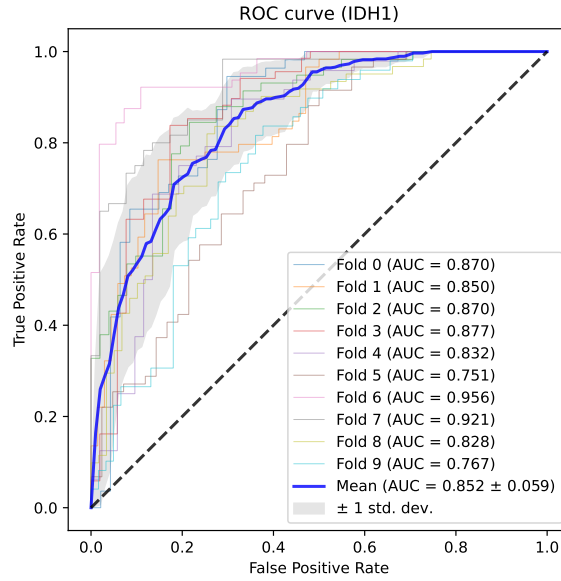
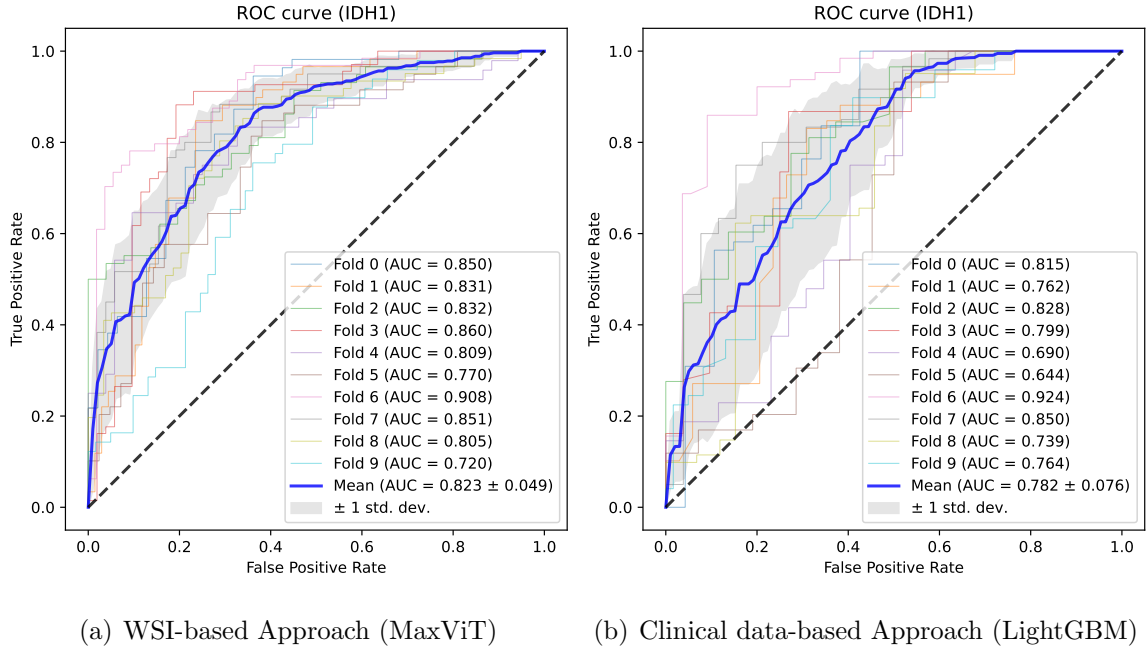
Examples of high and low attention patches in the heatmap are shown in “Attention Patches” of Fig. 4.4. The high-attention patches have “perinuclear halo” (fried egg appearance) that appears bright and clear around the cell nucleus. This “perinuclear halo” is a common morphological feature found in oligodendroglioma. In contrast, this feature was absent in the low-attention patches. Thus, it was seen that the approach could generate interpretable heatmaps. Therefore, it is expected to identify morphological features related to IDH1 gene mutation. In addition to the perinuclear halo, there may be other morphological features due to IDH1 mutation.

### 4.4.4 Interpretability of Clinical Data Model

Finally, the contribution of clinical variables was examined using SHAP (SHapley Additive exPlanations) values [27]. Fig. 4.5 shows the feature importance by SHAP for fold 0. A higher SHAP value means a higher impact on the prediction of IDH1 mutation positive.

The beeswarm plot illustrates which features affect the model’s output as shown in Fig. 4.5(a). It reveals that the primary diagnosis and age were essential, while gender, site of resection or biopsy, and race were less critical.

The dependence plot is a scatter plot between the feature value and its corresponding SHAP values in Fig. 4.5(b)-(f). For the age, it could be seen that the influence on IDH1



(c) Ensemble Approach

Figure 4.3: ROC curve. The results are based on (a) WSI, (b) clinical data, and (c) ensemble approaches, respectively. The AUCs for each fold and the average ROC curves are plotted.

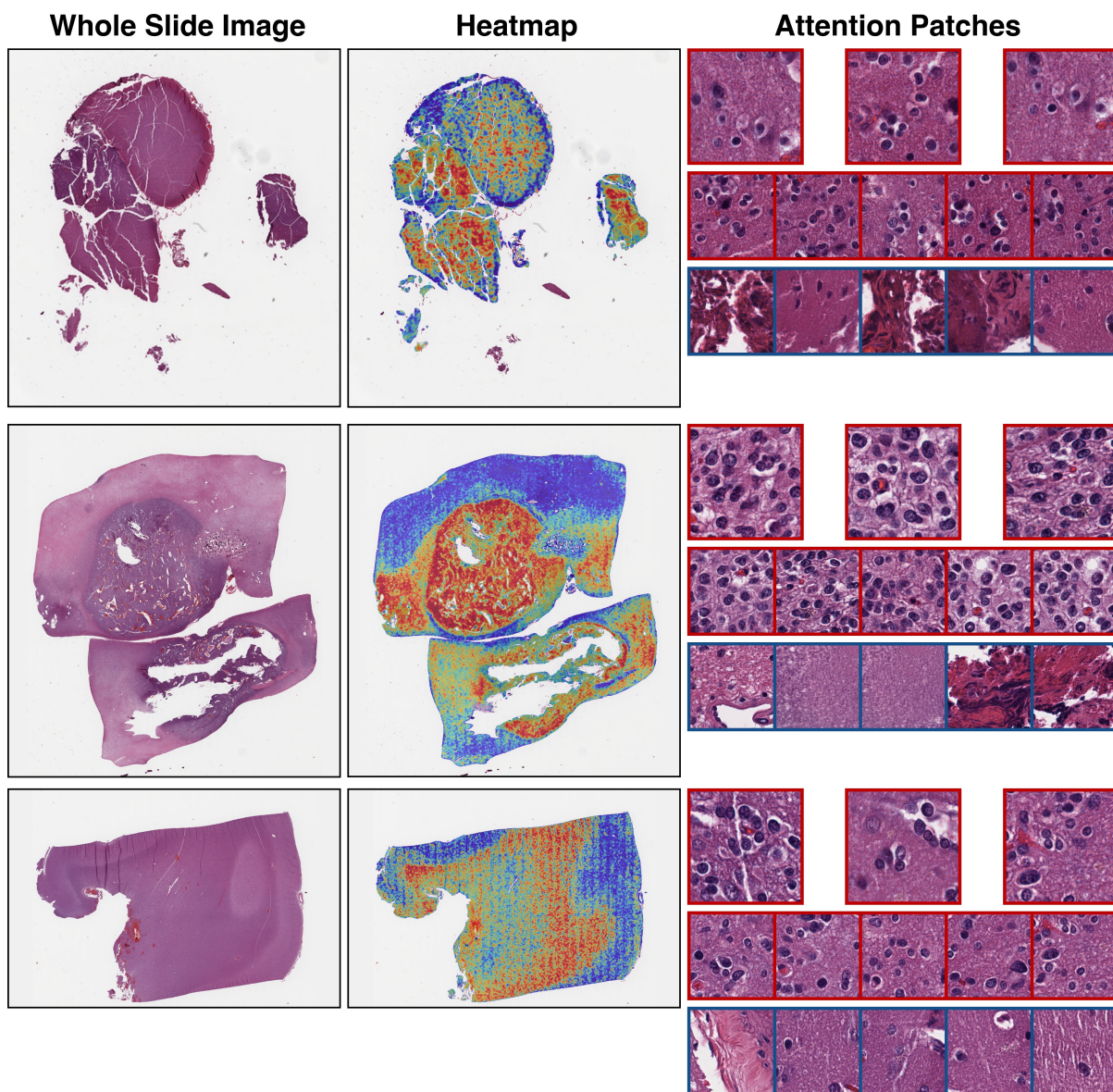


Figure 4.4: Example of attention heatmaps and attention patches (MaxViT).

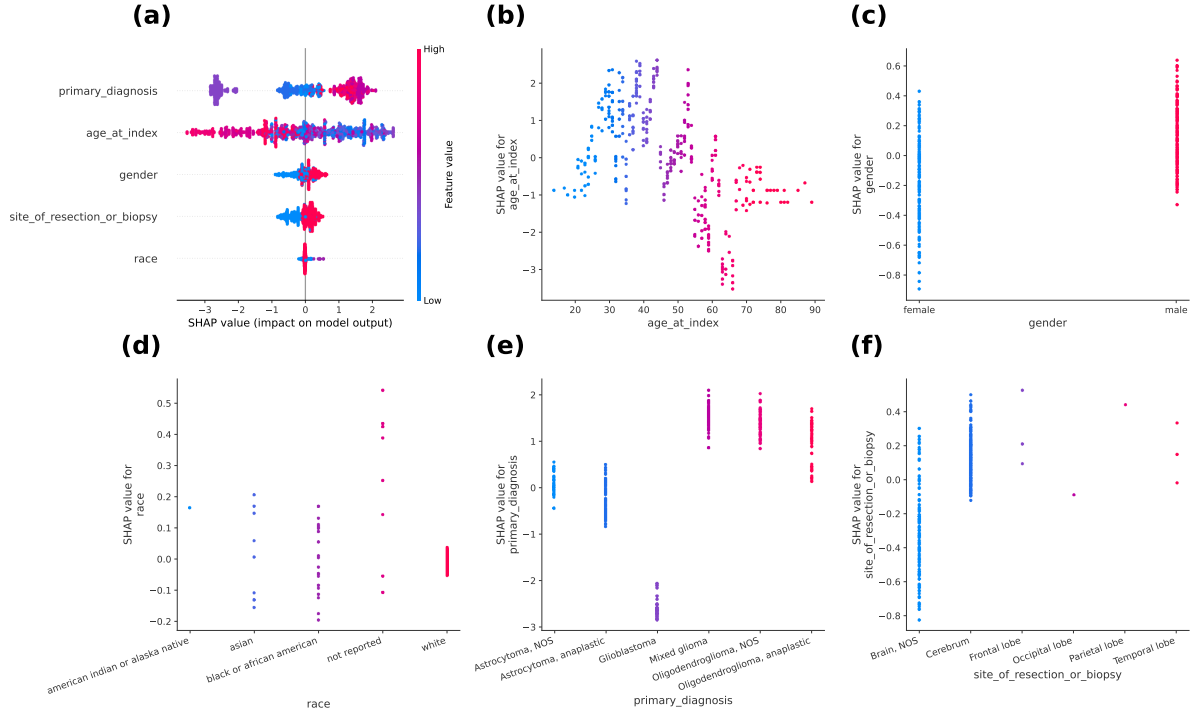


Figure 4.5: Feature importance for fold 0. (a): Beeswarm Plot. (b)-(f): Dependence Plot.

mutation prediction was higher for patient in their 20s to 40s and lower in those aged 50 or older. It has been reported that IDH mutations are frequently encountered in patients under 55 years of age [34]. Therefore, age was an essential variable for predicting IDH1 mutation. For the primary diagnosis, the influence was particularly low for “Glioblastoma” as shown in Fig. 4.5(e). This is probably due to the low frequency of IDH1 mutation in glioblastoma patients. For the site of resection or biopsy, the influence on IDH1 mutation prediction of “Brain, NOS” was lower than other brain sites (Fig. 4.5(f)). In other words, it is suggested that “Brain, NOS” contributes to wild-type prediction. “Brain, NOS” tends to be more common in the wild-type (Table 4.1). However, “Brain, NOS” means that the site of the brain is unknown, and it is a term used provisionally when the primary site cannot be identified clearly. Therefore, it may not be appropriate to conclude that “Brain, NOS” contributes to wild-type prediction since it doesn’t have accurate information about the brain site like other categories. Furthermore, the site of resection or biopsy was correlated with the primary diagnosis (Fig. 4.2). For these reasons, it seems that the site of resection or biopsy should not be used due to the fact the brain site is unclear and to avoid multicollinearity.

## 4.5 Conclusion

This chapter proposed a method to predict IDH1 mutation by combining pathological images and clinical data. The proposed approach trained separate models on pathological images and clinical data, and then the final prediction results were obtained by taking the average of their outputs. The obtained results suggested that the ensemble method achieved higher classification performance than the individual models. In addition, the results of feature importance analysis in the clinical data model (LightGBM) showed that primary diagnosis and age were the most important variables. Visualization of attention score in the pathological image model (ABMIL) suggested that the model can recognize histological features such as perinuclear halo.

Generalizing the proposed late fusion for discerning other gene mutations and studying their results constitutes an important future direction.



## Chapter 5

## Conclusion

### 5.1 Conclusion

This study aimed to predict IDH1 mutation status more accurately using pathological images of gliomas. In chapter 3, the effectiveness of pre-training using contrastive learning was examined. The experimental results suggested that pre-training the feature extractor with contrastive learning, even with a small amount of data, improves the prediction performance of IDH1. In chapter 4, an ensemble learning approach combining pathological images with clinical data was proposed. This fusion model was found to improve overall prediction performance compared to models using only pathological images. Furthermore, this experiment also focused on the interpretability of the models. For the WSI-based model, visualizing attention maps suggested that histological features associated with IDH1 mutations could be identified. Meanwhile, for the clinical data-based model, the analysis of feature importance by SHAP indicated that primary diagnosis and age were the most influential variables in predicting IDH1 mutations.

### 5.2 Future Works

All experiments in this study utilized data from the TCGA. Therefore, the robustness of the model has not been fully evaluated. Model performance should also be evaluated on data from independent cohorts. Future work will include validation with external datasets. Furthermore, the aim is to develop multimodal models that utilize MRI, transcriptomes, and other data types. It is hoped that this research will expand the potential applications of AI technology in classifying gliomas and predicting IDH1 mutations.

## Acknowledgement

First of all, I really would like to express my deepest gratitude to Prof. Hiroharu Kawanaka at the Graduate School of Engineering, Mie University, who offered continuing support and constant encouragement.

I am also grateful to Prof. Bruce J. Aronow, Prof. V. B. Surya Prasath, and Shyam Sundar Debsarkar at Cincinnati Children's Hospital Medical Center, USA. They provided a lot of technical help and encouragement. Prof. Aronow kindly accepted me as a short-term study abroad student at his laboratory in Cincinnati Children's Hospital Medical Center. Thanks to his help and encouragement, I was able to make good progress in my project. Prof. Surya and Shyam Sundar gave me a lot of technical and language help in my research project and paper writing. Their direct guidance during my short-term study abroad was particularly invaluable and greatly appreciated.

In addition, this study was partially supported by Mie University study abroad program. Thanks to this program, I could stay Cincinnati Children's Hospital Medical Center for a month to obtain great progress and discuss the project with experts. I would like to sincerely thank this scholarship program. And finally, I would like to appreciate other members related my research project.

## Reference

- [1] D. N. Louis, A. Perry, P. Burger, D. W. Ellison, G. Reifenberger, A. von Deimling, K. Aldape, D. Brat, V. P. Collins, C. Eberhart, D. Figarella-Branger, G. N. Fuller, F. Giangaspero, C. Giannini, C. Hawkins, P. Kleihues, A. Korshunov, J. M. Kros, M. Beatriz Lopes, H.-K. Ng, H. Ohgaki, W. Paulus, T. Pietsch, M. Rosenblum, E. Rushing, F. Soylemezoglu, O. Wiestler, and P. Wesseling, “International society of neuropathology-haarlem consensus guidelines for nervous system tumor classification and grading,” *Brain Pathology*, vol. 24, no. 5, pp. 429–435, 2014. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/bpa.12171>
- [2] D. N. Louis, A. Perry, G. Reifenberger, A. von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison, “The 2016 world health organization classification of tumors of the central nervous system: a summary,” *Acta Neuropathologica*, vol. 131, no. 6, pp. 803–820, Jun 2016. [Online]. Available: <https://doi.org/10.1007/s00401-016-1545-1>
- [3] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffietti, A. von Deimling, and D. W. Ellison, “The 2021 WHO Classification of Tumors of the Central Nervous System: a summary,” *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 06 2021. [Online]. Available: <https://doi.org/10.1093/neuonc/noab106>
- [4] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyő, A. L. Moreira, N. Razavian, and A. Tsirigos, “Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning,” *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, Oct 2018. [Online]. Available: <https://doi.org/10.1038/s41591-018-0177-5>
- [5] S. Liu, Z. Shah, A. Sav, C. Russo, S. Berkovsky, Y. Qian, E. Coiera, and A. Di Ieva, “Isocitrate dehydrogenase (idh) status prediction in histopathology images of gliomas using deep learning,” *Scientific Reports*, vol. 10, no. 1, p. 7733, May 2020. [Online]. Available: <https://doi.org/10.1038/s41598-020-64588-y>
- [6] S. Jiang, G. J. Zanazzi, and S. Hassanpour, “Predicting prognosis and idh mutation status for patients with lower-grade gliomas using whole slide images,”

- Scientific Reports*, vol. 11, no. 1, p. 16849, Aug 2021. [Online]. Available: <https://doi.org/10.1038/s41598-021-95948-x>
- [7] Y. Shimada, S. Okuda, Y. Watanabe, Y. Tajima, M. Nagahashi, H. Ichikawa, M. Nakano, J. Sakata, Y. Takii, T. Kawasaki, K.-i. Homma, T. Kamori, E. Oki, Y. Ling, S. Takeuchi, and T. Wakai, “Histopathological characteristics and artificial intelligence for predicting tumor mutational burden-high colorectal cancer,” *Journal of Gastroenterology*, vol. 56, no. 6, pp. 547–559, Jun 2021. [Online]. Available: <https://doi.org/10.1007/s00535-021-01789-w>
- [8] T. Hayakawa, V. B. S. Prasath, H. Kawanaka, B. J. Arronow, and S. Tsuruoka, “Computational nuclei segmentation methods in digital pathology : A survey,” *Archives of Computational Methods in Engineering*, vol. 28, no. 1, pp. 1–13, 2021. [Online]. Available: <https://doi.org/10.1007/s11831-019-09366-4>
- [9] A. Yonekura, H. Kawanaka, V. B. S. Prasath, B. J. Aronow, and H. Takase, “Improving the generalization of disease stage classification with deep CNN for glioma histopathological images,” in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Kansas, MO, USA, November 2017, pp. 1222–1226, international Workshop on Deep Learning in Bioinformatics, Biomedicine, and Healthcare Informatics (DLB2H). [Online]. Available: <https://doi.org/10.1109/BIBM.2017.8217831>
- [10] A. Yonekura, H. Kawanaka, V. B. S. Prasath, B. J. Aronow, and H. Takase, “Automatic disease stage classification of glioblastoma multiforme histopathological images using deep convolutional neural network,” *Biomedical Engineering Letters*, vol. 8, no. 3, pp. 321–327, 2018. [Online]. Available: <https://doi.org/10.1007/s13534-018-0077-0>
- [11] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt, “Toward a shared vision for cancer genomic data,” *New England Journal of Medicine*, vol. 375, no. 12, pp. 1109–1112, 2016, pMID: 27653561. [Online]. Available: <https://doi.org/10.1056/NEJMp1607591>
- [12] “GDC Data Portal,” <https://portal.gdc.cancer.gov/>, (Accessed on 4 August 2023).
- [13] H. Yan, D. W. Parsons, G. Jin, R. McLendon, B. A. Rasheed, W. Yuan, I. Kos, I. Batinic-Haberle, S. Jones, G. J. Riggins, H. Friedman, A. Friedman, D. Reardon, J. Herndon, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, and D. D. Bigner, “IDH1 and IDH2 Mutations in Gliomas,” *New England Journal of Medicine*, vol. 360, no. 8, pp. 765–773, 2009, pMID: 19228619. [Online]. Available: <https://doi.org/10.1056/NEJMoa0808710>

- [14] A. L. Cohen, S. L. Holmen, and H. Colman, “IDH1 and IDH2 mutations in gliomas,” *Current Neurology and Neuroscience Reports*, vol. 13, no. 5, p. 345, Mar 2013. [Online]. Available: <https://doi.org/10.1007/s11910-013-0345-4>
- [15] A. Yonekura, H. Kawanaka, V. B. S. Prasath, B. J. Arronow, and S. Tsuruoka, “Glioma subtypes clustering method using histopathological image analysis,” in *7th International Conference on Informatics, Electronics and Vision (ICIEV), and 2nd International Conference on Imaging, Vision and Pattern Recognition (icIVPR)*, Fukuoka, Japan, June 2018, pp. 442–446. [Online]. Available: <https://doi.org/10.1109/ICIEV.2018.8641031>
- [16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: <https://proceedings.mlr.press/v119/chen20j.html>
- [17] M. Ilse, J. Tomczak, and M. Welling, “Attention-based deep multiple instance learning,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2127–2136. [Online]. Available: <https://proceedings.mlr.press/v80/ilse18a.html>
- [18] M. Tan and Q. Le, “Efficientnetv2: Smaller models and faster training,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 10 096–10 106. [Online]. Available: <https://proceedings.mlr.press/v139/tan21a.html>
- [19] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “DINOv2: Learning robust visual features without supervision,” *Transactions on Machine Learning Research*, 2024. [Online]. Available: <https://openreview.net/forum?id=a68SUt6zFt>
- [20] X. Chen, S. Xie, and K. He, “An empirical study of training self-supervised vision transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9640–9649.

- [21] X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15 750–15 758.
- [22] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, “Bootstrap your own latent - a new approach to self-supervised learning,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 271–21 284. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf)
- [23] A. Marcolini, N. Bussola, E. Arbitrio, M. Amgad, G. Jurman, and C. Furlanello, “histolab: A python library for reproducible digital pathology preprocessing with automated testing,” *SoftwareX*, vol. 20, p. 101237, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352711022001558>
- [24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” 2019.
- [25] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, “Data-efficient and weakly supervised computational pathology on whole-slide images,” *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.
- [26] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)
- [27] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf)
- [28] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.

- [29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” 2015.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.
- [31] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, “Swin transformer v2: Scaling up capacity and resolution,” 2022.
- [32] Z. Dai, H. Liu, Q. V. Le, and M. Tan, “Coatnet: Marrying convolution and attention for all data sizes,” 2021.
- [33] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, “Maxvit: Multi-axis vision transformer,” 2022.
- [34] C. Andrews and R. A. Prayson, “IDH mutations in older patients with diffuse astrocytic gliomas,” *Annals of Diagnostic Pathology*, vol. 49, p. 151653, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1092913420301994>

## Publication List

### Journal Paper

- (1) Riku Nakagaki, Shyam Sundar Debsarkar, Hiroharu Kawanaka, Bruce J. Aronow, V. B. Surya Prasath, “Deep Learning-based IDH1 Gene Mutation Prediction Using Histopathological Imaging and Clinical Data”, *Computers in Biology and Medicine* (in review).

### International Conferences

- (1) Riku Nakagaki, Hiroharu Kawanaka, V. B. Surya Prasath, Bruce J. Aronow, “A Study on Mutation Prediction Using Deep Learning for Histopathological Images”, *Proc. of the 12th International Symposium for Sustainability by Engineering at Mie University (Research Area C)*, pp. 58-59, 2022.
- (2) Riku Nakagaki, Hiroharu Kawanaka, Shyam Sundar Debsarkar, V. B. Surya Prasath, Bruce J. Aronow, “A Study on IDH1 Mutation Prediction Using Contrastive Learning and Attention-based MIL”, *Proc. of the 13th International Symposium for Sustainability by Engineering at Mie University (Research Area C)*, pp. 83-84, 2023.

### Domestic Conferences

- (1) 中垣梨久, 川中普晴, V. B. Surya Prasath, Bruce J. Aronow, “深層学習を用いた Glioma 病理画像における IDH1 変異予測に関する一検討”, 令和 4 年度電気・電子・情報関係学会東海支部連合大会講演論文集, K5-8, 2022.
- (2) 中垣梨久, 川中普晴, V. B. Surya Prasath, Bruce J. Aronow, “CLAM を用いた病理画像における遺伝子変異予測に関する一検討”, 2022 年度日本生体医工学会東海支部大会抄録集, p. 19, 2022.
- (3) 中垣梨久, 川中普晴, V. B. Surya Prasath, Bruce J. Aronow, “Attention 機構を用いた Glioma 組織病理画像における遺伝子変異予測に関する一検討”, 令和 5 年度電気・電子・情報関係学会東海支部連合大会講演論文集, F5-7, 2023.
- (4) 中垣梨久, 川中普晴, Shyam Sundar Debsarkar, V. B. Surya Prasath, Bruce J. Aronow, “Glioma 病理画像における遺伝子変異予測のための特徴抽出器の検討”, 2023 年度日本生体医工学会東海支部大会抄録集, p. 24, 2023.



# AppendixA

## Attention Heatmap

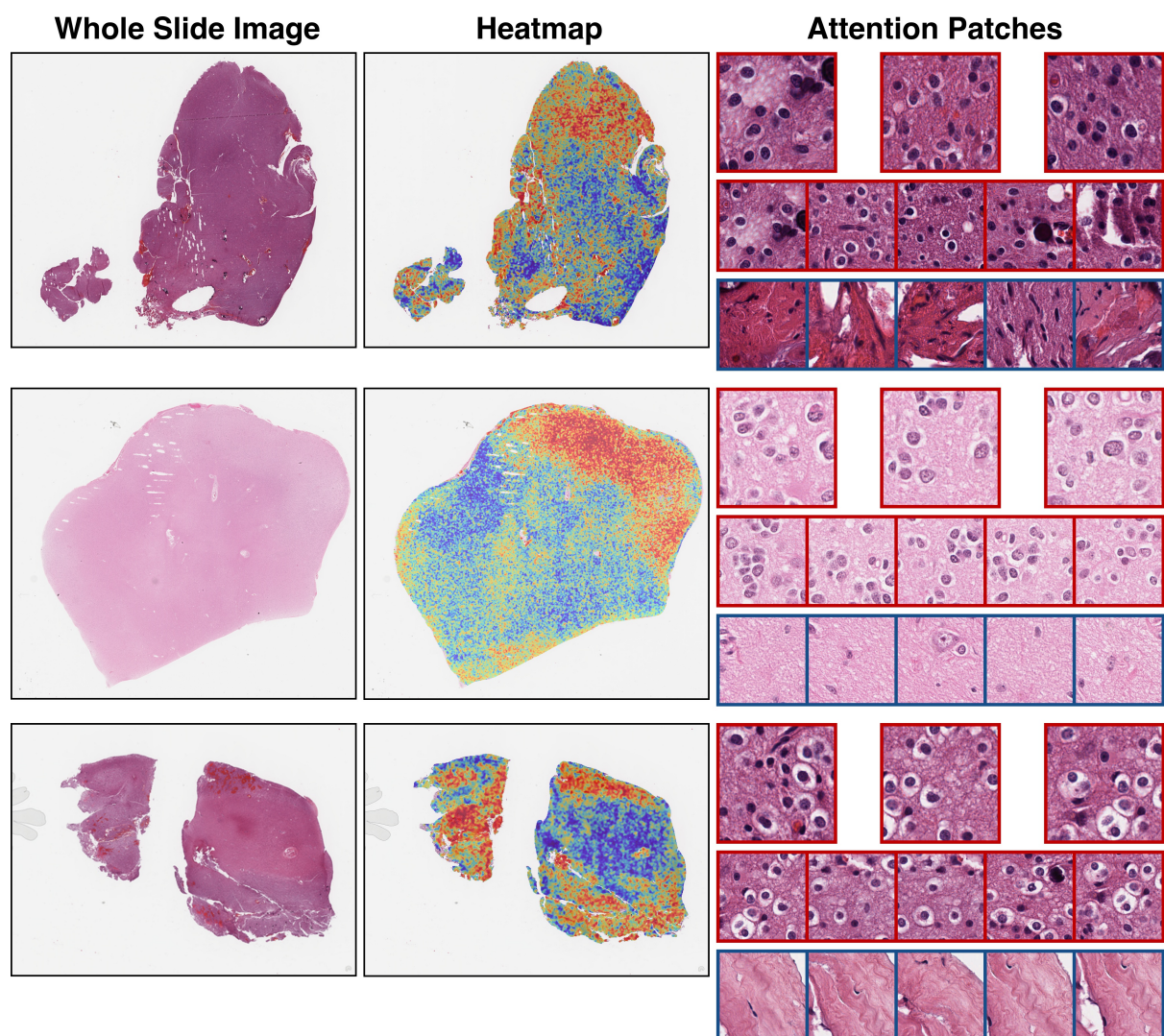


Figure A.1: Attention heatmaps and attention patches (MaxViT). Ground Truth: IDH1, Prediction: IDH1.

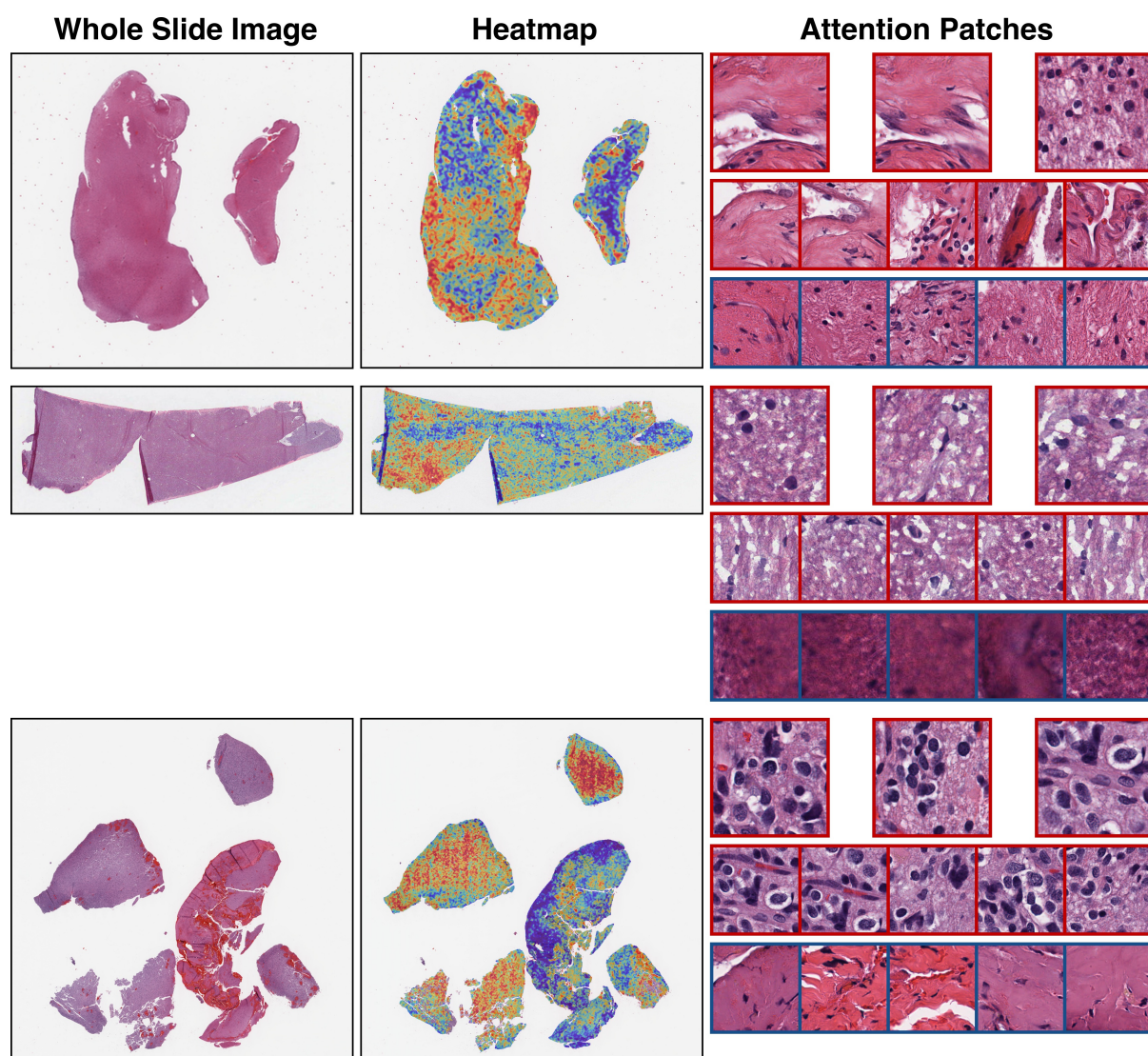


Figure A.2: Attention heatmaps and attention patches (MaxViT). Ground Truth: IDH1, Prediction: WT.



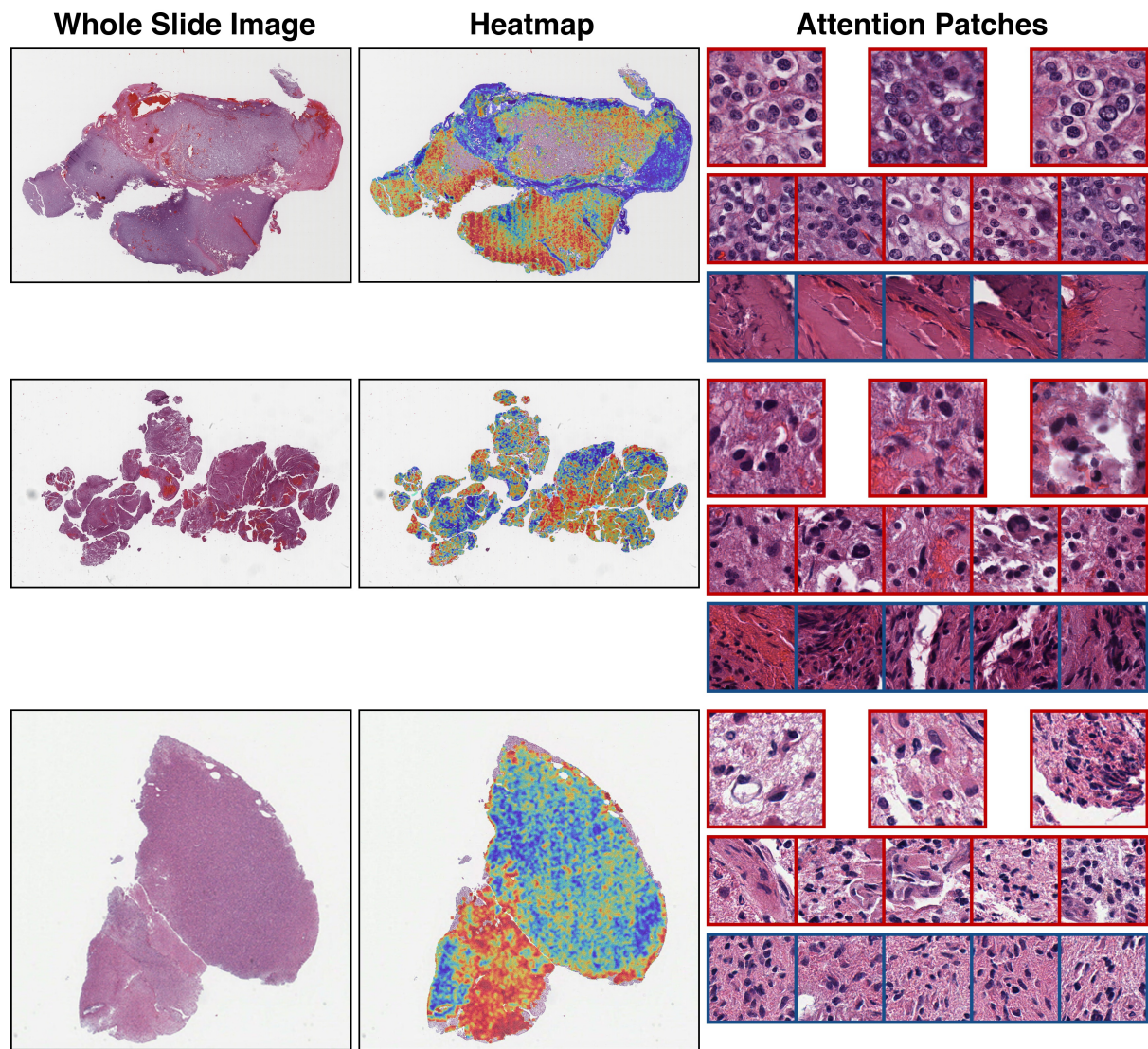


Figure A.3: Attention heatmaps and attention patches (MaxViT). Ground Truth: WT, Prediction: IDH1.

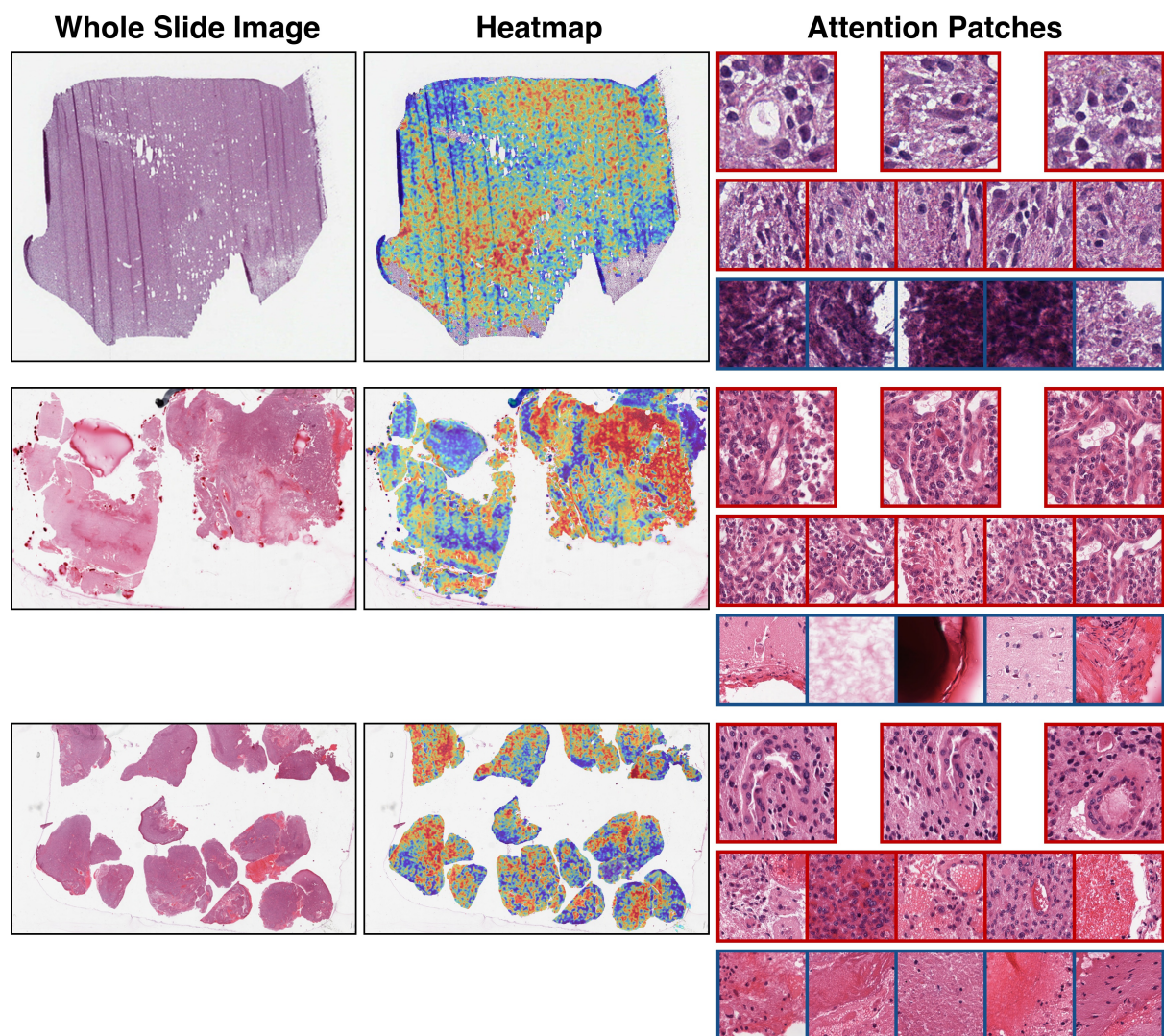


Figure A.4: Attention heatmaps and attention patches (MaxViT). Ground Truth: WT, Prediction: WT.

# AppendixB

## Feature Importance with SHAP

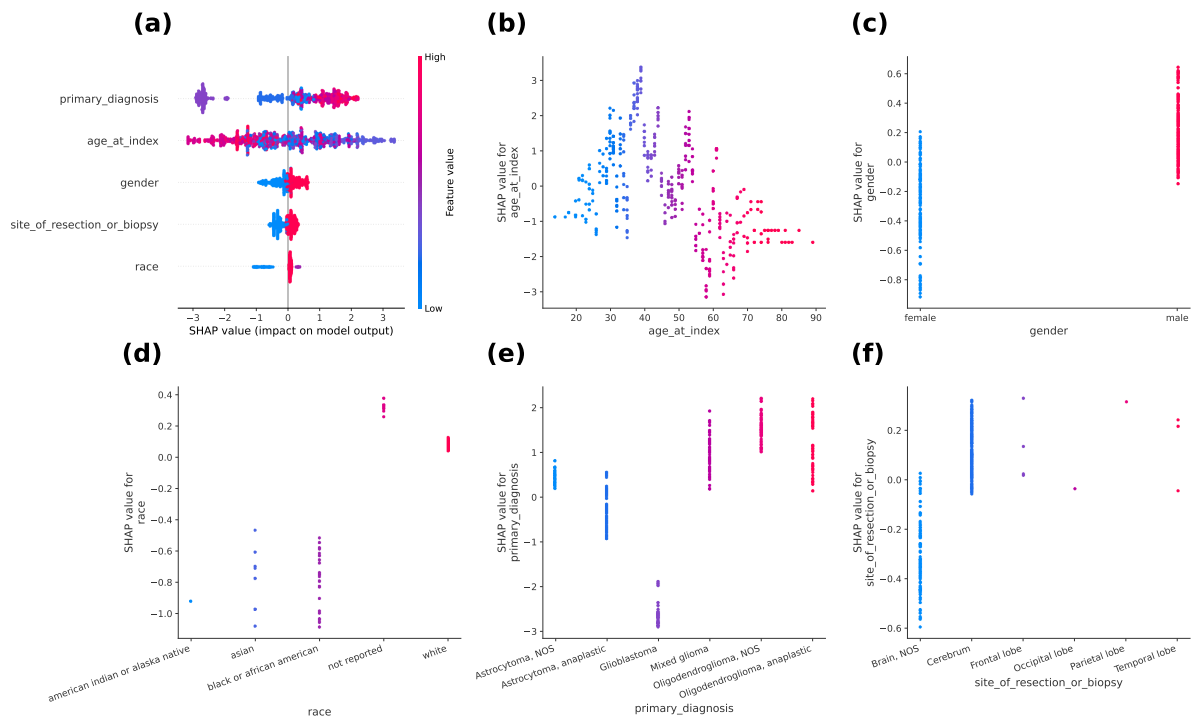


Figure B.1: Feature importance for fold 1.



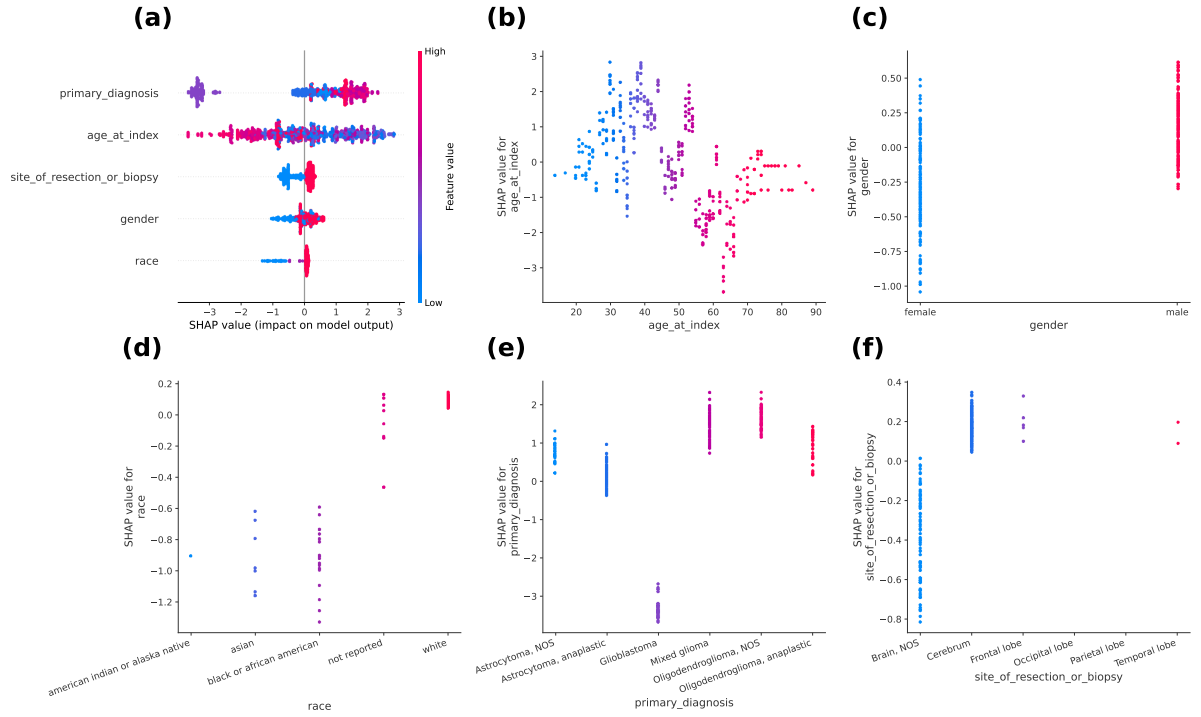


Figure B.2: Feature importance for fold 2.

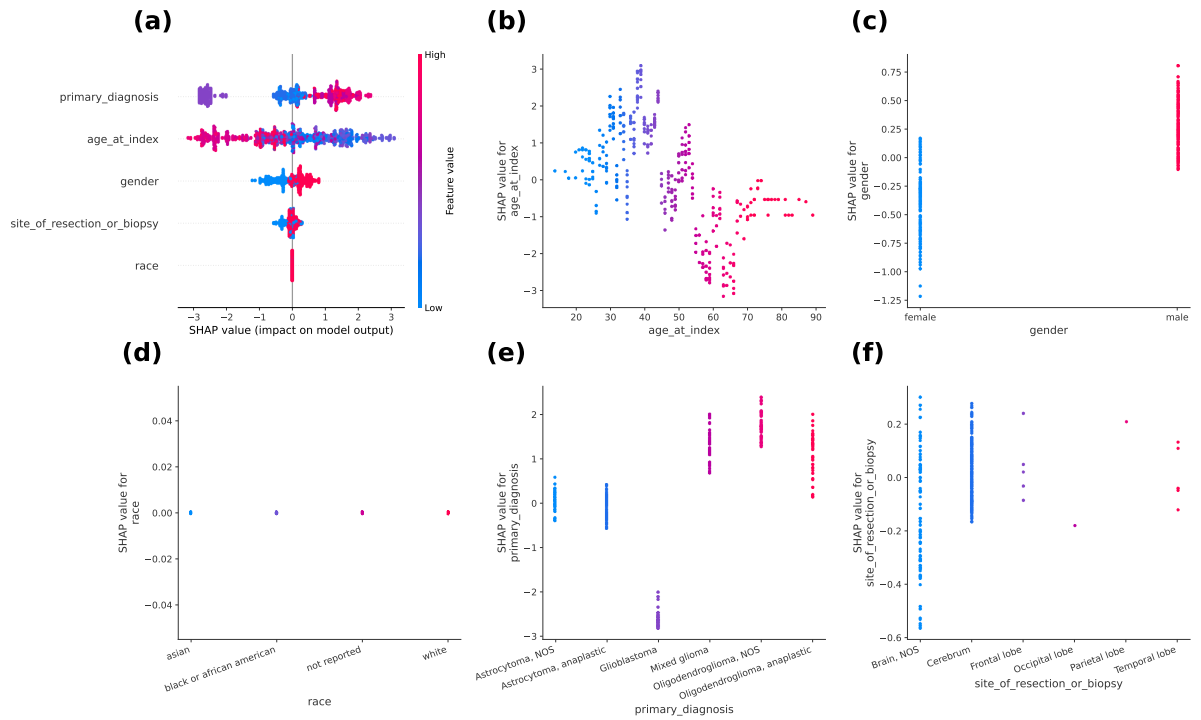


Figure B.3: Feature importance for fold 3.

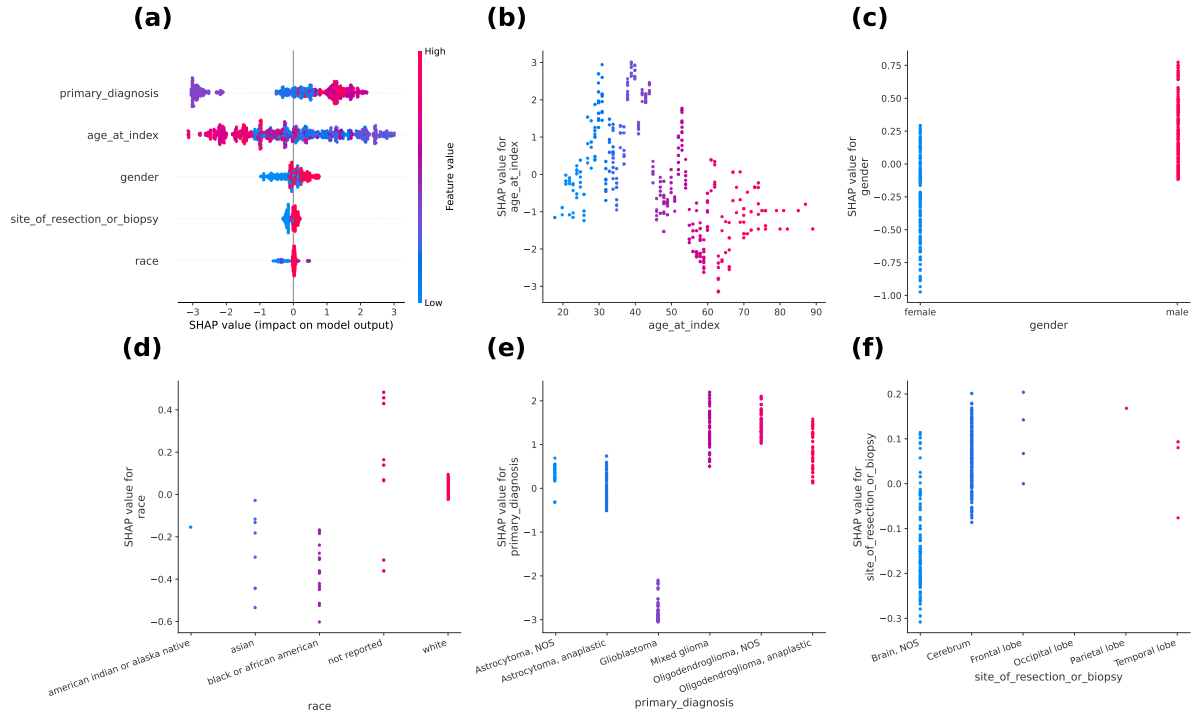


Figure B.4: Feature importance for fold 4.

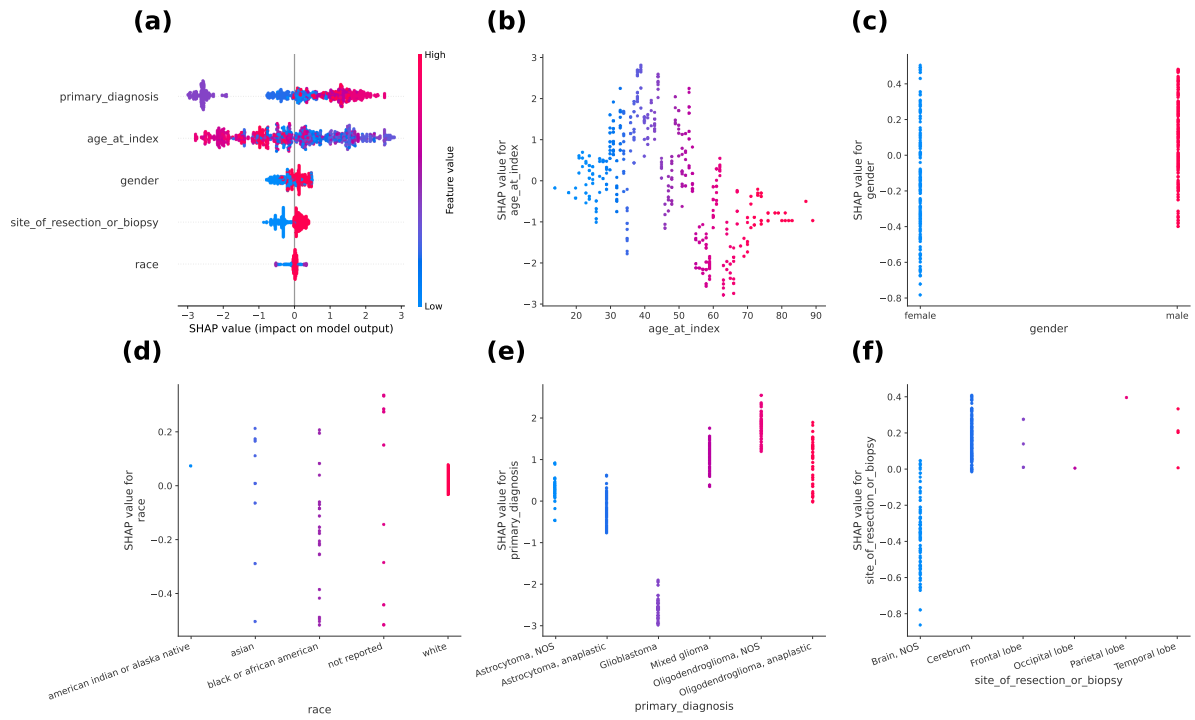


Figure B.5: Feature importance for fold 5.

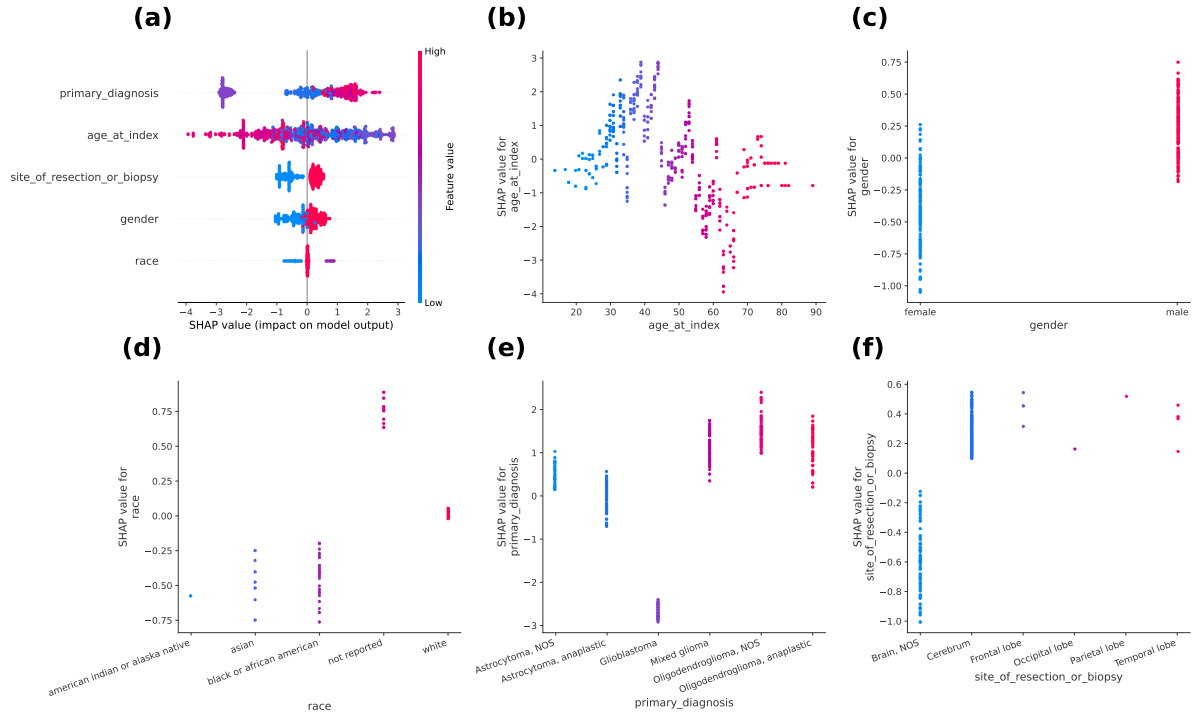


Figure B.6: Feature importance for fold 6.

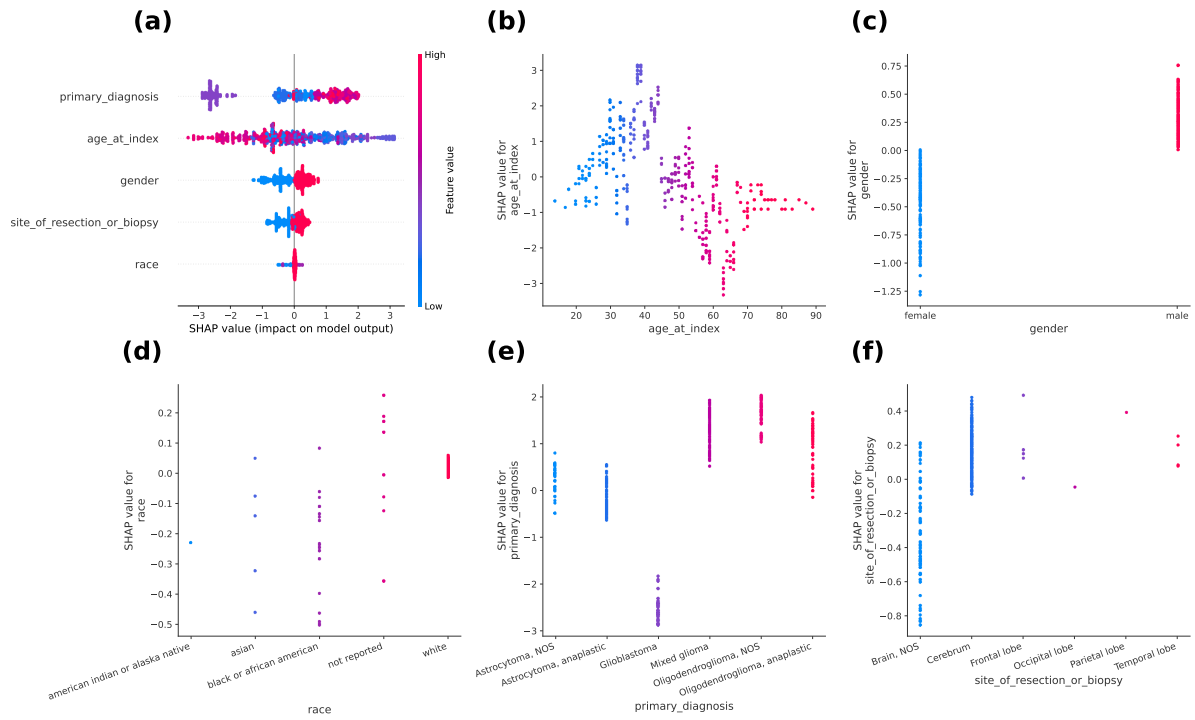


Figure B.7: Feature importance for fold 7.



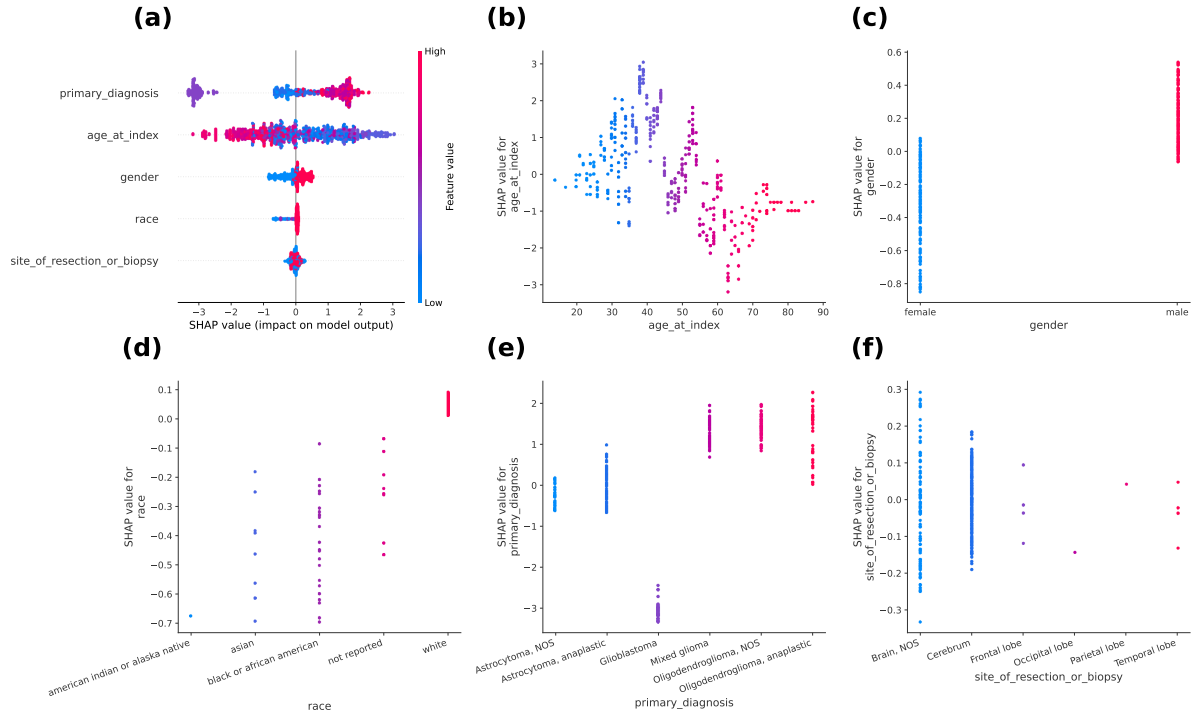


Figure B.8: Feature importance for fold 8.

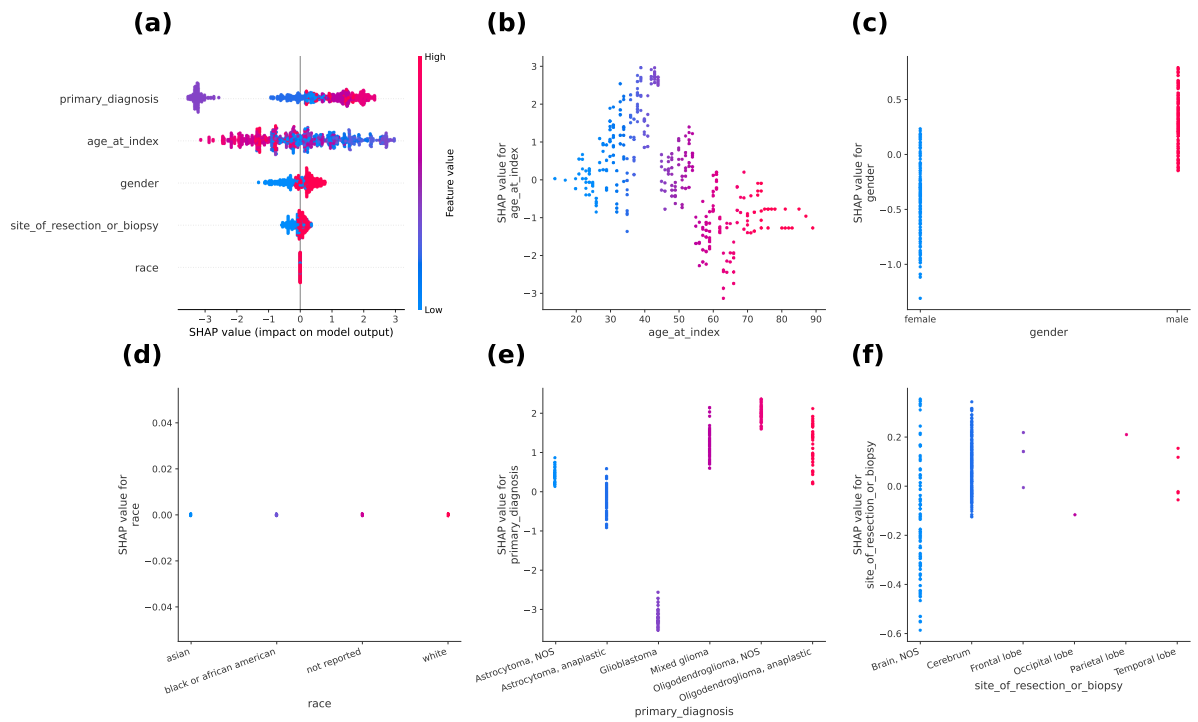


Figure B.9: Feature importance for fold 9.