A Data Release and Source Code Technical Details

Preprocessing As described in Section 4, SRH images are acquired in two Raman shift frequencies. They each correspond to a registered channel in an RGB image (green and blue for 2845 cm^{-1} and 2930 cm^{-1}, respectively). The co-registration utilizes discrete Fourier transform implemented using `imreg_dft` python package [61]. The third channel (red) is obtained by subtracting the first two channels, in their original 16-bit depth. The three-channel images are then converted to floats between 0 and 1, and are used as model input. A panel of paired raw and RGB images is shown in Figure 6.

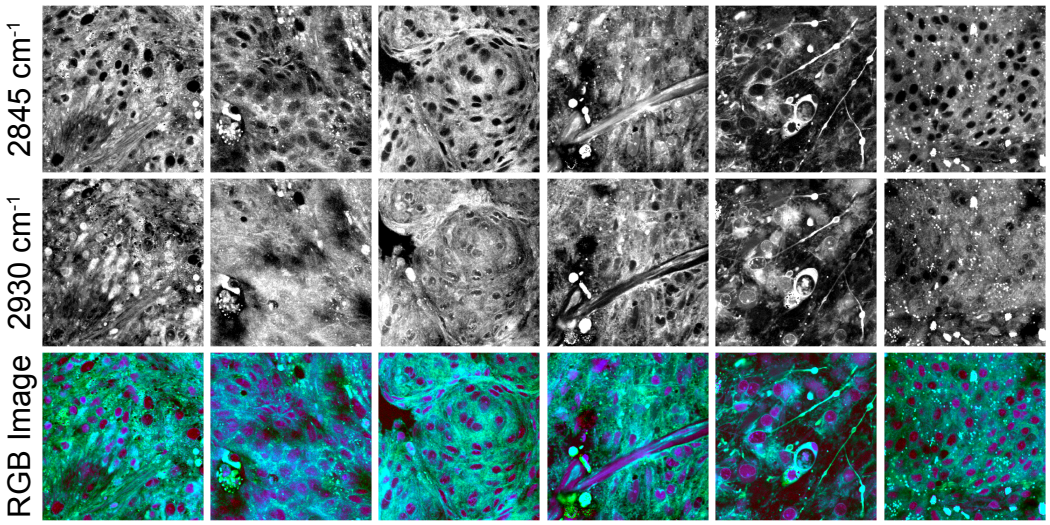

Figure 6: A panel of constructed RGB images and their respective raw images. The RGB images are normalized and contrast adjusted for better viewing.

Segmentation The data included in OpenSRH have been approximately segmented by two steps. They were first divided into non-overlapping 300×300 pixel patches and then each patch will be classified into three categories (tumor, nondiagnostic and normal) using the previously trained model. The model was trained on about 6.65 million patches using data collected from multiple institutions. An ImageNet pretrained ResNet50 model was used to train on first manually labelled 300,000 images by certified pathologists. Then, the predictions on the part of unlabelled data were manually checked and used for fine-tuning the model. This process was applied to the rest of the data iteratively until all data are well labelled. The model's predictions are included in the metadata for each slide.

Data release Data is available through a multitude of options. Primarily, they will be available via a Google Drive and Amazon AWS S3 upon completion of a short data usage agreement. The link to google drive will be available automatically after completing the survey, without human approval. Please contact the authors if you wish to download the data from AWS in case of regional unavailability through other means. The data available through Google Drive is compressed (∼ 364 GB) and can be downloaded directly and uncompressed afterward. A list of checksums is also made available for data integrity checks. Please note, data available through AWS is uncompressed (∼ 449 GB) and but it will require you to have an AWS account.

Dataset directory organization and metadata OpenSRH is intended to be assembled into the following directory structure in figure 7. The metadata for OpenSRH is stored in `meta/opensrh.json`. It provides the patient-level ground truth label, relative paths to patches and their segmentation prediction using the format described in figure 8. The `meta/train_val_split.json` file contains the default random split between the training and validation set. It's format is described in figure 9.

```
OpenSRH
├── meta
│   ├── opensrh.json
│   └── train_val_split.json
└── studies
    ├── NIO_001
    │   ├── 6
    │   │   ├── img_6_1.dcm # Psuedo H&E-colored image
    │   │   ├── patches
    │   │   │   ├── NIO_001_6_0_0_0_0.tif
    │   │   │   └── ... # additional patches
    │   │   └── strips
    │   │       ├── CH2 # 2845 cm^-1 channel
    │   │       │   ├── img_6_10.dcm
    │   │       │   └── ...
    │   │       └── CH3 # 2930 cm^-1 channel
    │   │           ├── img_6_11.dcm
    │   │           └── ...
    │   └── ... # additional slides
    └── NIO_... # additional patients
```

Figure 7: Intended tree directory structure for the OpenSRH dataset. All the associated data for training has been included.

```
{
    "NIO_001": {
        "patient_id": "NIO_001",
        "class": "hgg",
        "slides": {
            "6": {
                "slide_id": "6",
                "tumor_patches": [
                    "NIO_001/6/patches/NIO_001-6-0_0_0_0.tif",
                    ...
                ],
                "normal_patches": [...],
                "nondiagnostic_patches": [...],
            }, ...
        }
    }, ...
}
```

Figure 8: Metadata file format. A patient may have several slides. The number of patches in a slide is variable. The last four numbers included in the patch file name indicate their location in the whole slide image.

```
{
    "train": ["NIO_001", "NIO_002", "NIO_005", "NIO_006", ...],
    "val": ["NIO_003", "NIO_004", "NIO_007", "NIO_009", ...]
}
```

Figure 9: Training validation split metadata file. It consists of a dictionary with 2 lists of strings representing patient IDs.

B Patient Age Distribution

OpenSRH data includes patients ranging from newborn to 87 years old. Different classes of tumors are well-distributed among the age groups shown in figure 10. The distribution is skewed to the left as we expect with an increasing incidence of brain tumors with age. We did not observe a difference in model performance or notice differences in tumor cytologic or histoarchitectural features between age groups.

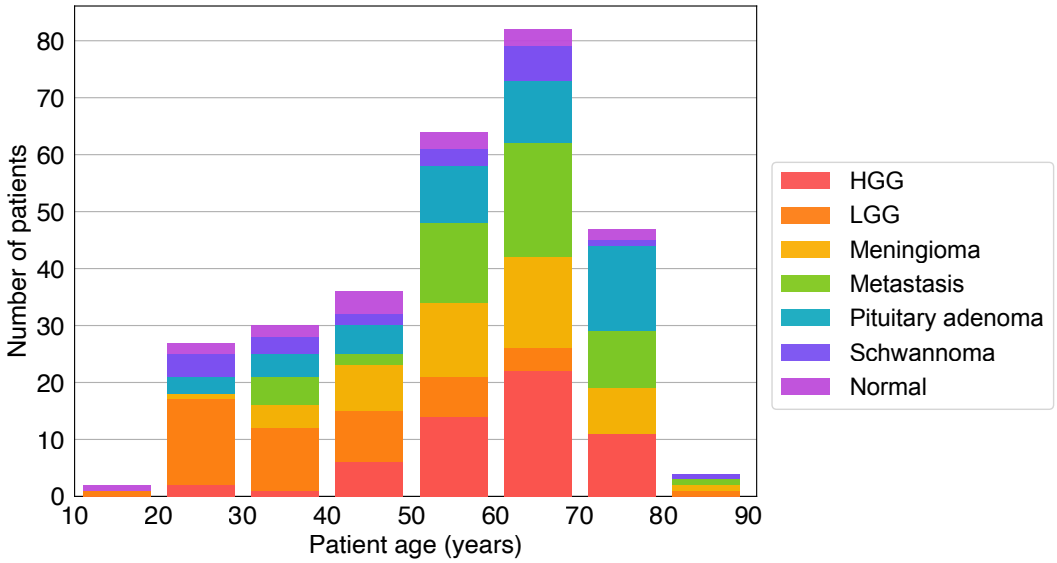

Figure 10: Distribution of patient age in OpenSRH. We can observe that high grade gliomas, metastasis, and meningiomas are more common in older patients, while low grade gliomas are more common in younger patients. HGG, high grade glioma; LGG, low grade glioma.

C Training Protocol and Details

The experiments were trained using PyTorch Lightning, which is a wrapper for PyTorch and allows for efficient multi-GPU training (if available). Associated code and example configuration file to train the benchmark models is available at https://opensrh.mlins.org.

Training protocol is described in sections 5.1 and 6.1 for our histological classification and contrastive learning benchmarks, respectively. In this section, we provide detailed parameters and more details on the augmentations used for contrastive learning.

C.1 Histological Classification Training Protocols

Table 3 shows detailed training parameters for cross entropy experiments. The augmentations used for these experiments are vertical flipping and horizontal flipping, each with a probability of 0.5. For ViT-S training, we used a cosine learn rate scheduler, with a 0.3 cosine period and the first 10% steps as a linear warmup stage.

Backbone	ResNet50	ViT-S
Classification head	Linear(2048, 7)	Linear(384, 7)
Augmentations	Flipping	Flipping, Resize 224
Batch size	96	256
# GPUs	2 $\times$ Nivida 2080Ti	
Loss	Cross Entropy	
# Epochs	20	
Pretrained	{Random, ImageNet}	
Optimizer	AdamW	
Initial learn rate	1E-3	1E-4
Scheduler	Step, half @ epoch	Cosine w/ warmup
Seeds	{1000, 2000, 3000}	
Time (hrs)	9.5	9
# Parameters	23.5M	21.7M

Table 3: Training parameters of histological classification benchmarks. Training time is an estimate and should be used for reference only.

C.2 Contrastive Learning Training Protocols

Table 4 shows detailed training parameters for contrastive learning experiments. The cosine learn rate scheduler follows the same parameters as described in section C.1.

Backbone	ResNet50	ViT-S
Projection head	Linear(2048, 128)	Linear(384, 24)
Augmentations	Strong	Strong, Resize 224
Batch size	448	512
# GPUs	8 $\times$ Nivida 2080Ti	
Method	{SimCLR, SupCon}	
# Epochs	40	
Optimizer	AdamW	
Initial learn rate	1E-2	5E-4
Scheduler	None	Cosine w/ warmup
Seeds	{1000, 2000, 3000}	
Time (hrs)	15.5	9.8

Table 4: Contrastive learning pre-training parameters. Training time is an estimate and should be used for reference only.

C.2.1 Augmentations

The strong augmentations used in these contrastive learning experiments consist of multiple random augmentations applied sequentially:

- Random Horizontal Flip
- Random Vertical Flip
- Gaussian Noise
- Color Jittering
- Random Autocontrast
- Random Solarize with threshold 0.2
- Random Adjust Sharpness with sharpness factor 2
- Gaussian Blur with kernel size 5 and sigma 1
- Random Erasing
- Random Affine Transformation with max 10 degrees rotation and 10-30% image translation
- Random Resized Crop

All augmentations are applied with a 0.3 probability and use the default PyTorch parameter unless otherwise noted above. In ViT experiments, Resize is applied before all other augmentations. A panel of randomly augmented images is shown in figure 11.

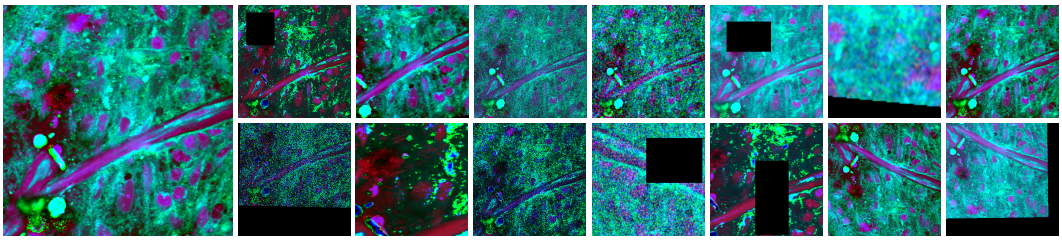

Figure 11: A sample panel of randomly augmented patches. The left patch is the original, and the rest of the patches are augmented using the protocol in section C.2.1.

C.3 Contrastive Learning Linear Evaluation Protocols

Table 5 shows detailed training parameters for contrastive learning linear evaluation. The cosine learn rate scheduler follows the same parameters as described in section C.1.

Backbone	ResNet50	ViT-S
Classification head	Linear(2048, 7)	Linear(384, 7)
Augmentations	Flipping	Flipping, Resize 224
Batch size	96	256
# GPUs	2 $\times$ Nivida 2080Ti	
Loss	Cross Entropy	
# Epochs	20	
Initialization	{SimCLR, SupCon, ImageNet}	
Optimizer	AdamW	
Initial learn rate	1E-3	1E-4
Scheduler	Step, half @ epoch	Cosine w/ warmup
Seeds	{1000, 2000, 3000}	
Linear training time (hrs)	20	6
Finetune training time (hrs)	13	5

Table 5: Cross entropy experiment training parameters. Training time is an estimate and should be used for reference only.

D Additional Histological Classification Results

In addition to the results in Table 1, we also computed average precision. Average precision is computed using one-vs-all and averaged over all classes. In addition to patch- and patient-level metrics, we also report the metric computed at the slide level. These results are shown in Table 6:

	Backbone	Pretrain	Accuracy	Top 2	MCA	MAP	FNR
Patch level metrics	ResNet50	Random	84.4 (0.4)	93.5 (0.2)	83.8 (0.5)	89.5 (0.5)	**0.9 (0.1)**
	ResNet50	ImageNet	**86.5 (0.4)**	**94.4 (0.1)**	**85.6 (0.3)**	**91.2 (0.3)**	1.0 (0.0)
	ViT-S	Random	77.2 (0.5)	90.0 (0.4)	76.8 (0.8)	82.3 (0.5)	1.8 (0.1)
	ViT-S	ImageNet	83.7 (0.5)	93.4 (0.2)	82.7 (0.9)	88.8 (0.1)	1.9 (0.2)
Slide level metrics	ResNet50	Random	**88.7 (0.8)**	95.3 (0.3)	**88.1 (1.0)**	93.6 (0.1)	**0.5 (0.5)**
	ResNet50	ImageNet	**88.8 (0.5)**	95.7 (0.3)	**88.4 (0.5)**	**94.4 (0.1)**	1.2 (0.0)
	ViT-S	Random	83.7 (0.3)	95.5 (0.4)	83.8 (0.9)	92.0 (0.6)	1.8 (0.2)
	ViT-S	ImageNet	**88.7 (0.8)**	**97.1 (0.2)**	**88.3 (0.9)**	93.9 (0.4)	1.9 (0.0)
Patient level metrics	ResNet50	Random	**90.0 (0.0)**	95.0 (0.0)	**91.4 (0.0)**	92.8 (0.2)	**0.0 (0.0)**
	ResNet50	ImageNet	88.9 (0.8)	94.4 (0.8)	90.5 (0.6)	**94.0 (0.1)**	**0.6 (0.8)**
	ViT-S	Random	85.0 (1.4)	95.0 (0.0)	87.2 (1.1)	93.2 (0.4)	1.7 (0.0)
	ViT-S	ImageNet	88.9 (0.8)	**96.1 (0.8)**	90.5 (0.6)	**93.9 (0.4)**	1.7 (0.0)

Table 6: Extended metrics for histologic classification benchmarks for ResNet50 and ViT-S. Pretrain refers to the pretraining strategy. Each experiment included three random initial seeds. The mean value and standard deviation (in parentheses) for each metric are reported. MCA, mean class accuracy; MAP, mean average precision; FNR, false negative rate.

In Figure 4, we showed the confusion matrices at the patient level for the first random seed. We compared different initialization strategies (random and ImageNet) and different architectures (ResNet and ViT). The confusion matrices at the patient level for the second and third seeds are shown in Figure 12 below.

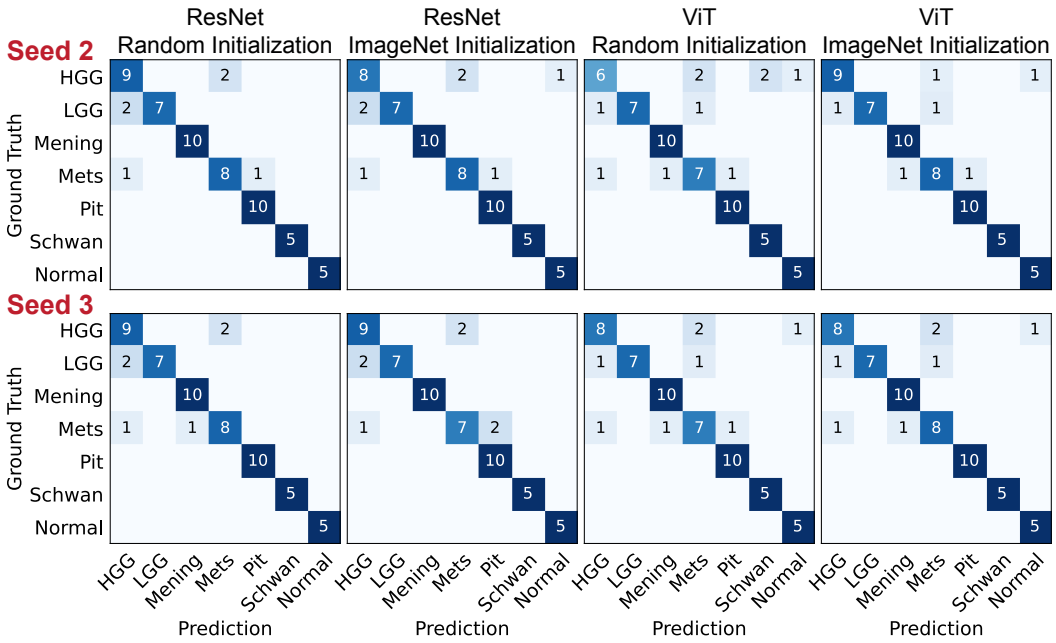

Figure 12: Patient-level confusion matrices for the four different training strategies on the validation set. Seeds 2 and 3 are shown. Mening, meningioma; Mets, metastasis; Pit, pituitary adenoma; Schwan, schwannoma.

We also include the confusion matrix for the first random seed at patch, slide, and patient level in figure 13 below. They correspond to the confusion matrices in figure 4.

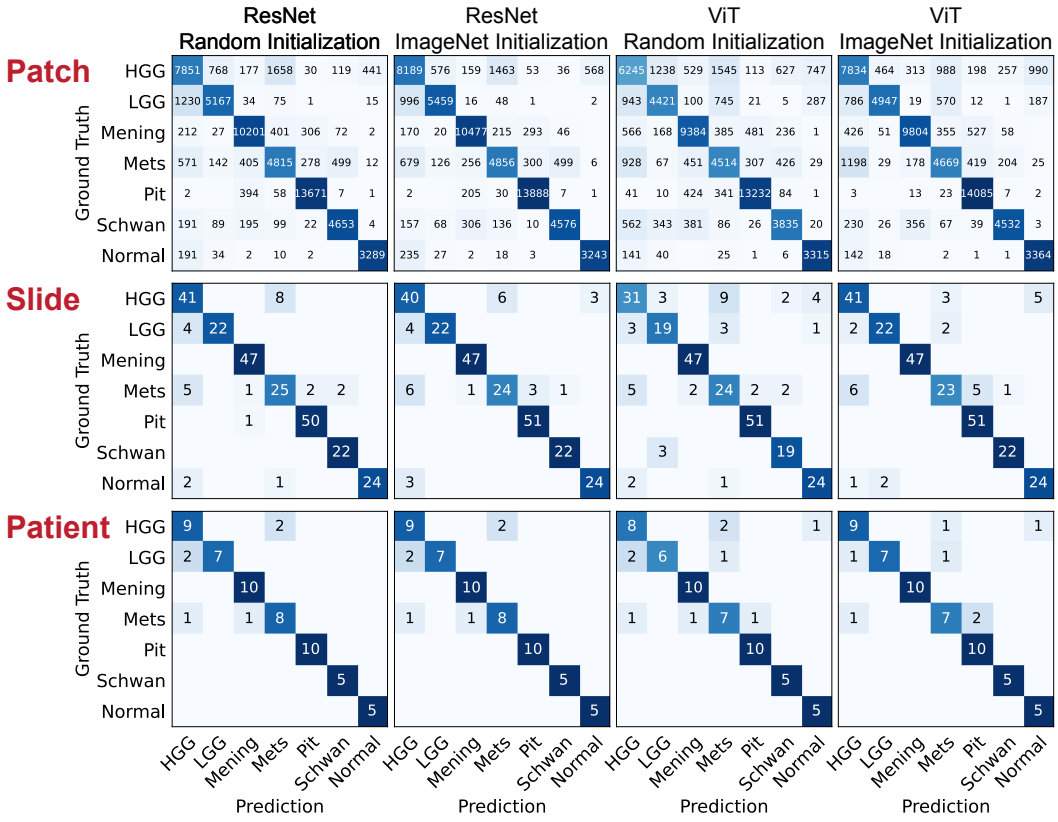

Figure 13: Patch, slide, and patient-level confusion matrices for the four different training strategies on the validation set. Seed 1 is shown. Mening, meningioma; Mets, metastasis; Pit, pituitary adenoma; Schwan, schwannoma.

E Additional Linear Evaluation Results for Contrastive Representation Learning

In table 2, we reported contrastive representation benchmarks evaluated using a linear classifier. Table 7 shows the extended results, including mean average precision and slide-level classification metrics.

	Backbone	Methods	Accuracy	Top 2	MCA	MAP	FNR
Patch level metrics	ResNet50	ImageNet	68.3 (0.0)	84.1 (0.0)	67.9 (0.0)	72.9 (0.1)	1.5 (0.0)
	ResNet50	SimCLR	79.1 (0.4)	92.8 (0.3)	78.9 (0.4)	84.2 (0.6)	1.5 (0.0)
	ResNet50	SupCon	**87.5 (0.3)**	**94.8 (0.2)**	**86.8 (0.3)**	**91.5 (0.5)**	**1.4 (0.2)**
	ViT-S	ImageNet	71.8 (0.1)	87.0 (0.0)	71.1 (0.1)	77.1 (0.1)	**1.4 (0.0)**
	ViT-S	SimCLR	76.8 (0.5)	90.7 (0.2)	76.3 (0.5)	82.5 (0.3)	**1.2 (0.2)**
	ViT-S	SupCon	81.4 (0.2)	92.2 (0.3)	80.2 (0.3)	85.6 (0.5)	1.7 (0.0)
Slide level metrics	ResNet50	ImageNet	80.9 (0.3)	92.2 (0.0)	81.2 (0.3)	86.1 (0.1)	0.8 (0.0)
	ResNet50	SimCLR	84.4 (1.6)	**96.8 (0.4)**	84.3 (1.4)	91.9 (0.2)	1.8 (0.2)
	ResNet50	SupCon	**91.1 (0.4)**	**97.0 (0.2)**	**90.6 (0.4)**	**95.3 (0.3)**	1.4 (0.2)
	ViT-S	ImageNet	89.1 (0.3)	**96.8 (0.2)**	88.6 (0.4)	92.7 (0.2)	**0.3 (0.2)**
	ViT-S	SimCLR	83.0 (0.4)	95.7 (0.8)	83.0 (0.7)	90.1 (0.6)	1.0 (0.9)
	ViT-S	SupCon	87.4 (0.2)	96.6 (0.2)	86.5 (0.4)	92.2 (0.4)	1.9 (0.0)
Patient level metrics	ResNet50	ImageNet	80.0 (0.0)	93.3 (0.0)	82.9 (0.0)	88.8 (0.1)	**0.0 (0.0)**
	ResNet50	SimCLR	83.9 (1.0)	**97.2 (1.0)**	86.1 (0.9)	92.4 (0.1)	1.7 (0.0)
	ResNet50	SupCon	**90.0 (0.0)**	95.0 (0.0)	**91.4 (0.1)**	94.6 (0.5)	1.7 (0.0)
	ViT-S	ImageNet	88.3 (0.0)	95.0 (0.0)	89.8 (0.0)	**93.9 (0.0)**	**0.0 (0.0)**
	ViT-S	SimCLR	80.0 (1.7)	96.1 (1.0)	83.0 (1.3)	92.3 (0.0)	1.1 (1.0)
	ViT-S	SupCon	87.8 (1.0)	96.7 (0.0)	89.4 (0.7)	94.0 (0.4)	1.7 (0.0)

Table 7: Extended metrics for linear evaluation protocol results in contrastive representation learning. Each experiment included three random initial seeds. Mean value and standard deviation (in parentheses) for each metric are reported. MCA, mean class accuracy; MAP, mean average precision; FNR, false negative rate.

.

In addition to the classification metric, we also include the confusion matrix in figure 14 below.

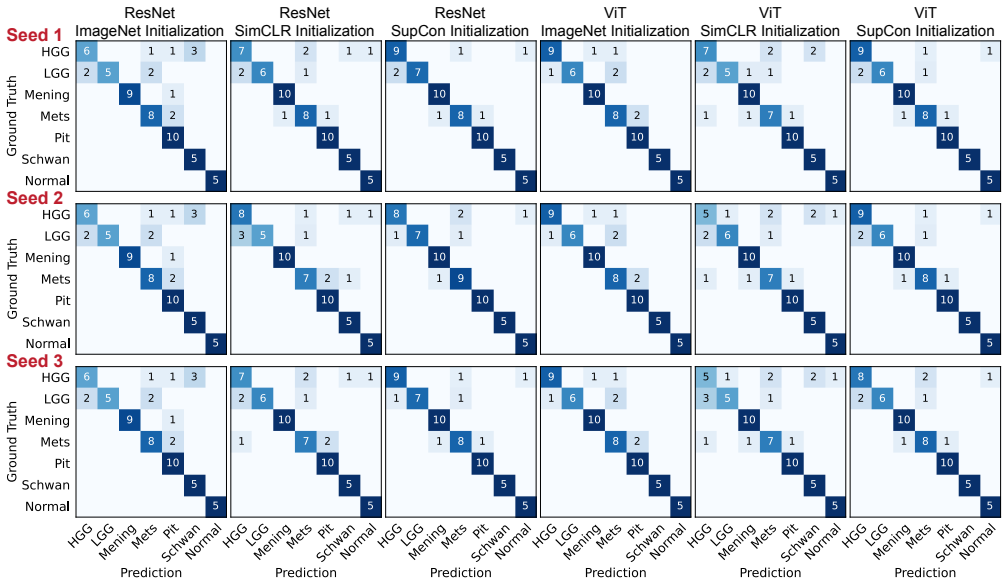

Figure 14: Patient-level confusion matrices of linear evaluations on the validation set. We included initialization from ImageNet, SimCLR, and Supcon, for both ResNet and ViT architecture, and all 3 seeds. Mening, meningioma; Mets, metastasis; Pit, pituitary adenoma; Schwan, schwannoma.

F Contrastive Representation Learning Fine-tuning Evaluation Results

In addition to evaluating our contrastive learning benchmarks using the linear evaluation protocol (section 6.1), we also performed evaluation using fine-tuning. These experiments have the same protocol as the linear evaluation (as described in section 6.1 and C.3), except for trainable (unfrozen) weights in the model backbone. Finetuning metrics are reported in table 8 below:

	Backbone	Methods	Accuracy	Top 2	MCA	MAP	FNR
Patch level metrics	ResNet50	SimCLR	86.3 (0.3)	**94.5 (0.2)**	85.2 (0.4)	90.6 (0.1)	**1.1 (0.1)**
	ResNet50	SupCon	**87.8 (0.3)**	**94.8 (0.2)**	**86.5 (0.4)**	**91.4 (0.4)**	**1.1 (0.1)**
	ViT-S	SimCLR	81.4 (0.1)	92.3 (0.2)	80.7 (0.4)	86.2 (0.2)	2.0 (0.1)
	ViT-S	SupCon	81.2 (0.4)	92.2 (0.1)	80.0 (0.5)	85.3 (0.6)	2.0 (0.2)
Slide level metrics	ResNet50	SimCLR	**89.8 (0.2)**	**96.4 (1.2)**	**89.2 (0.2)**	**94.8 (0.2)**	**1.0 (0.2)**
	ResNet50	SupCon	**90.1 (0.4)**	**96.0 (0.2)**	**89.5 (0.5)**	**95.0 (0.3)**	**0.9 (0.4)**
	ViT-S	SimCLR	85.6 (0.4)	97.1 (0.6)	84.9 (0.4)	92.7 (0.8)	2.2 (0.2)
	ViT-S	SupCon	86.8 (0.7)	97.1 (0.6)	86.2 (0.8)	91.7 (1.6)	2.2 (0.2)
Patient level metrics	ResNet50	SimCLR	89.4 (1.0)	95.6 (1.0)	90.9 (0.7)	94.2 (0.3)	**1.1 (1.0)**
	ResNet50	SupCon	**91.7 (0.0)**	95.0 (0.0)	**92.7 (0.0)**	**94.9 (0.3)**	**0.6 (1.0)**
	ViT-S	SimCLR	87.2 (1.0)	**96.7 (0.0)**	88.9 (0.9)	94.1 (0.6)	**1.7 (0.0)**
	ViT-S	SupCon	86.7 (0.0)	**96.7 (0.0)**	88.3 (0.1)	**94.0 (0.7)**	**2.8 (1.0)**

Table 8: Metrics for finetuning evaluation protocol results for contrastive representation learning. Each experiment included three random initial seeds. Mean value and standard deviation (in parentheses) for each metric are reported. MCA, mean class accuracy; MAP, mean average precision; FNR, false negative rate.

We also include confusion matrices for these experiments in figure 15 below.

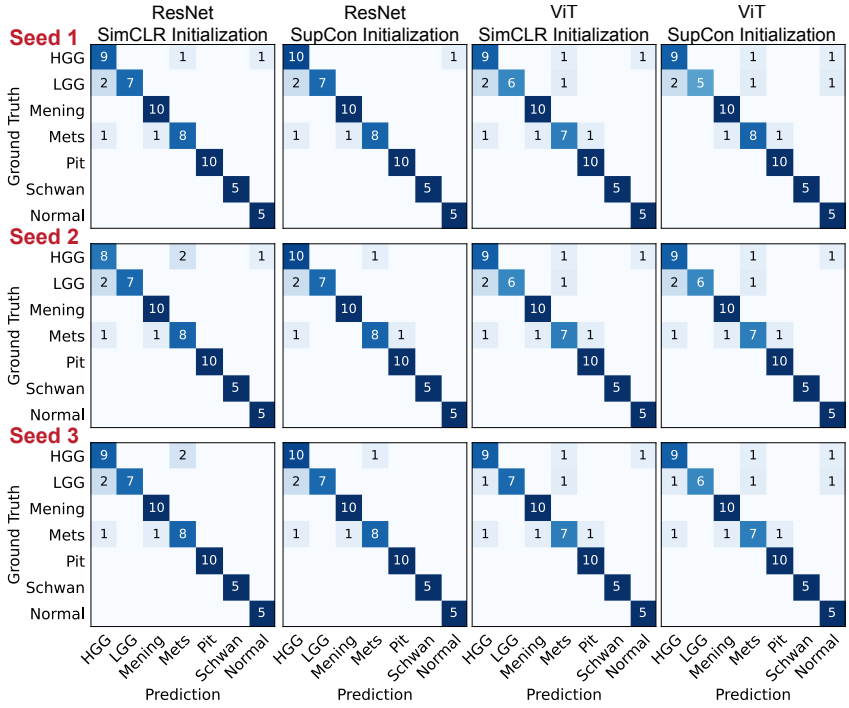

Figure 15: Patient-level confusion matrices of finetuning evaluations on the validation set. We included initialization from SimCLR and Supcon, for both ResNet and ViT architecture, and all 3 seeds. Mening, meningioma; Mets, metastasis; Pit, pituitary adenoma; Schwan, schwannoma. Mening, meningioma; Mets, metastasis; Pit, pituitary adenoma; Schwan, schwannoma.