Ma e ial a d Me h d

S d c h

We conducted a retrospective study across two independent institutions (Stanford University [Site 1] and Vanderbilt University [Site 2]) a er the IRB approval. e inclusion criteria were:

Pathologically con rmed diagnosis of the following pediatric posterior fossa tumors: medulloblastoma, pilocytic astrocytoma, or ependymoma,

Patients were aged 1 day to 19 years, and

Hematoxylin and Eosin (H&E) glass histology slides were available for review by a neuropathologist. Patients were excluded if the tumor histology diagnosis was unclear.

Hi l g da a e

Neuropathologists from each site independently viewed individual histology slides under a microscope at 20× and captured 4800 × 3600 pixel screenshot images with 72 × 72 dpi resolution centred over a tissue region representative of the brain tumor. E ort was made to reduce image capture of normal tissue, white space, and processing artifacts.

Da a e di ib i f ai i g

e data were strati ed by tumor type to ensure an equal distribution of tumor types in both the training set and validation set. For each site, 80% of the data served as training and 20% was withheld from the training set to serve as a test set to assess the nal model performance.

E e ime al e ie

We conducted the following experimental approaches:

Pha e 1: Develop a deep learning algorithm using solely Site 1 data and test its performance on test sets from Site 1 and Site 2.

Pha e 2: Fine-tune the best performing model from Phase 1 using a subset of the Site 2 cohort and assess model performance on test sets from Site 1 and Site 2.

M del a chi ec e

We used ResNet architectural backbone pretrained on the Image Net dataset, a compilation of over 14 million images of everyday objects [7,8]. Due to the relatively small cohort size, we used the smallest available pretrained architecture with 18 layers to reduce risk of over tting. e pretrained ResNet-18 architecture was modi ed to classify the three PF tumor classes: PA, EP, MB.

Image e- ce i g

Pixel values were normalized per PyTorch pretrained model guidelines [9]. All images contained three (i.e., RGB) color channels. We performed several data augmentations for training. Each image used for model training had a 50% probability of rescaling to 224 × 224 dimensions or random cropping of an unmagni ed 224 × 224 sized original image. In addition to these rescaling options, each image in the train set had a 50% probability of vertical or horizontal ip. Validation and test set images were rescaled to 224 × 224 to allow the model to analyze the image but were not otherwise manipulated; no data augmentations were applied to validation or test set images.

M del ai i g

All models were trained using the Python 3.6 programming language and the PyTorch deep learning framework and a NVIDIA TitanXp Graphic Processing Unit with 12 GB of memory [9]. During training, all layers of the model, including the pretrained convolutional layers, were ne-tuned on the histology training data and trained to minimize classi cation cross entropy loss. e Adam optimizer was used to update the weights of the model with each iteration [10]. We conducted a two-phase experimental approach, as shown in Figure 1.

Pha e 1: Develop a deep learning algorithm using solely Site 1 data and test its performance on test sets from Site 1 and Site 2. In Phase 1, during which the model only had access to Site 1 training data, the model was trained for 10 epochs with a batch size of 64 images and a learning rate of 0.001. Random majority subset (80%) of data from Site 2 served as the test set.

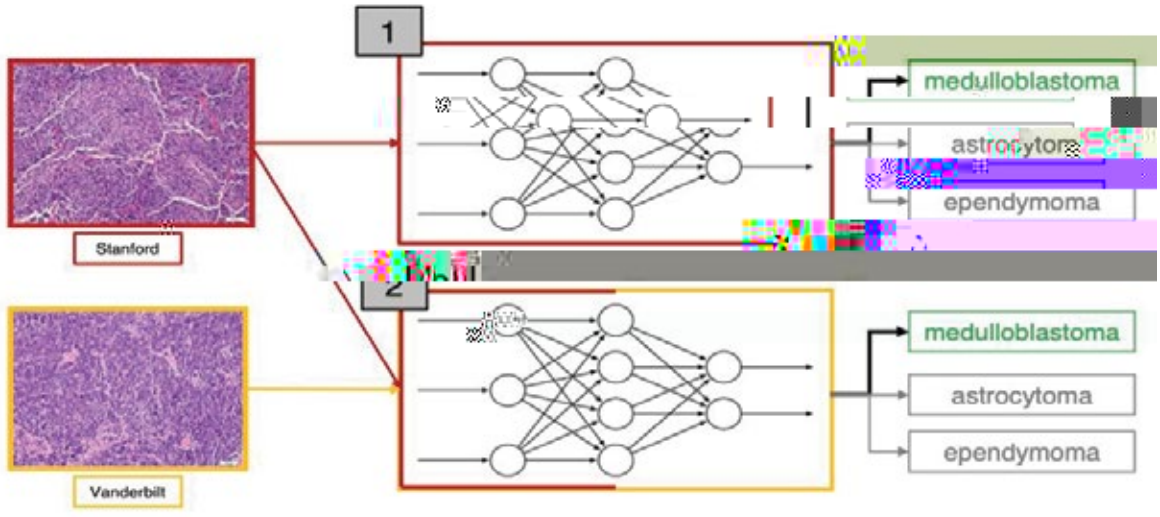

Figure 1: Model architectures of medulloblastoma; astrocytoma; ependymoma.

Pha e 2: Fine-tune the best performing model from Phase 1 using a subset cohort from Site 2 and assess model performance on test sets from Site 1 and Site 2. In Phase 2, during which the model was further ne-tuned on Site 2 training data, the model was trained for 5 epochs with a batch size of 64 images and a learning rate of 0.0001. Here, a random minority subset (20%) of data from Site 2 was used to ne-tune the best performing model from Phase 1. Similar to Phase 1, the majority subset (80%) from Site 2 served as test set to determine the model performance.

Each model was trained with 5-fold cross validation, using a 20% proportion of the training set as the validation set. During nal evaluation, the model with the lowest total loss was evaluated on the test set to gauge performance.