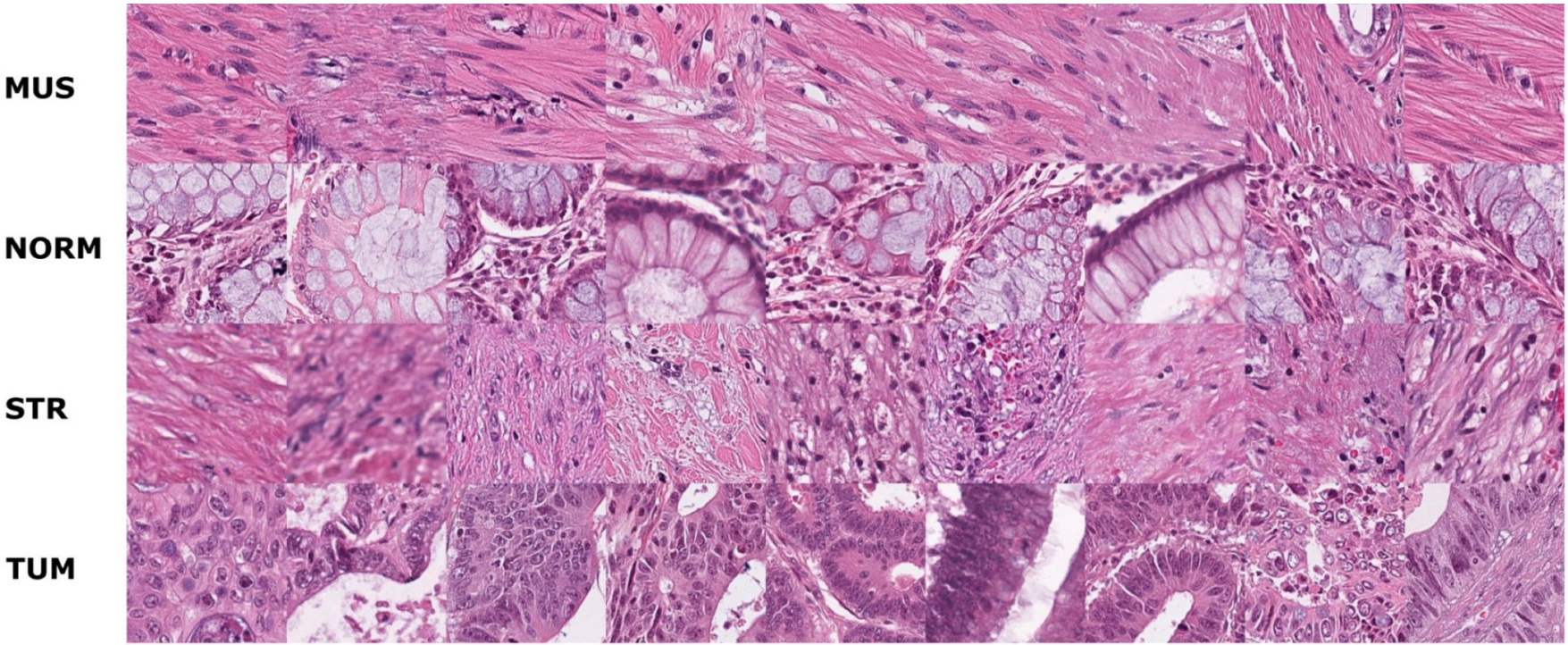
Colorectal Cancer Histology Classification

Youjin Park

DATASET

- The dataset consists of 100,000 individual image patches obtained from histological images (11.6GB)
- extracted from human cancer tissue slides at NCT Biobank in Heidelberg, Germany, and UMM Pathology Archive in Mannheim, Germany
- 224x224 pixels and color-normalized
- 9 Categories: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM)
    - About 10000 images in 9 folders of the categories

DATASET EXAMPLE

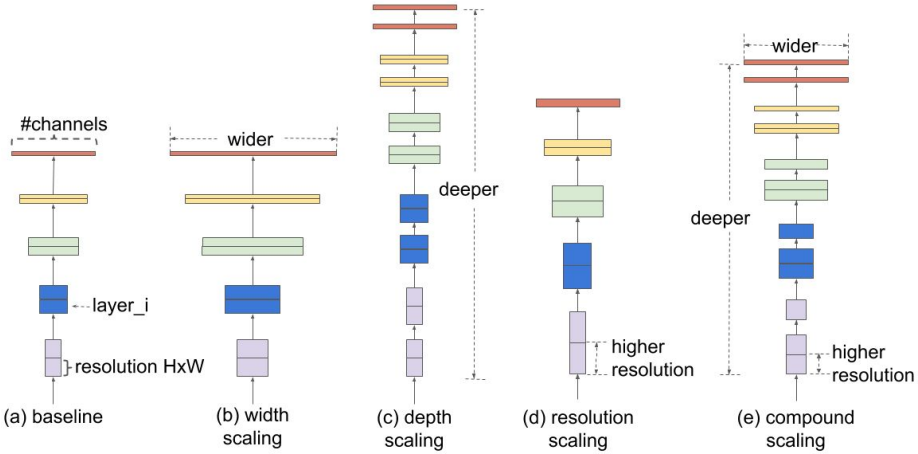


IMPORTANCE

- Colorectal cancer (CRC) is a pervasive and potentially life-threatening cancer that affects millions of individuals around the world.
- Automating histological tissue classification and increasing the accuracy of colon cancer diagnosis can have a direct and positive impact on the health and prediction of patients who are suffering from this condition.
- This project can lead to faster diagnosis, more personalized treatment plans, and increased likelihood of early detection

METHODOLOGY

- Aiming to predict tissue types from histological images and classify them into nine classes
- EfficientNetB5
  - EfficientNet is a mobile friendly pure convolutional model (ConvNet) that proposes a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple but highly effective compound coefficient.
- AWS cloud services (including AWS S3 for data storage and AWS SageMaker for computing performance)



TRAINING THE MODEL

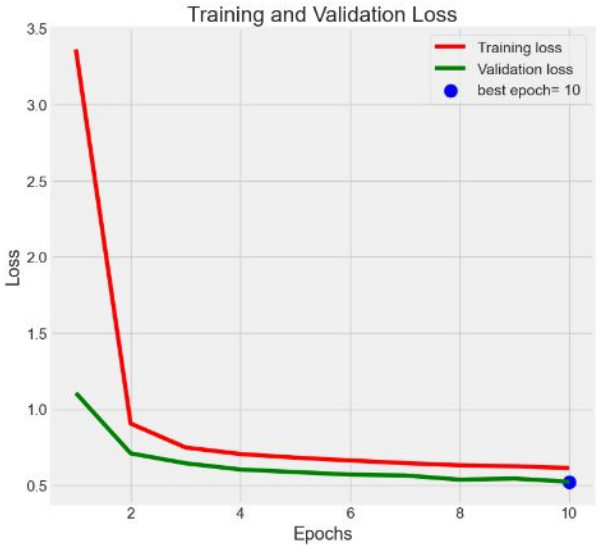
- Splitting the data: splitted into training, validation, and test sets with an 80-10-10 ratio (80000 - 10000 - 10000 images each)
- Model Compilation: compiled with the Adamax optimizer (learning rate = 0.001), categorical cross-entropy loss, and accuracy as the evaluation metric.
- Training Configuration: trained for 20 epochs using the training data with validation on the test data.
- Regularization Techniques: implements regularization techniques, such as L2 kernel regularization, L1 activity and bias regularization, and dropout (rate = 0.45), to enhance model generalization and to prevent overfitting.

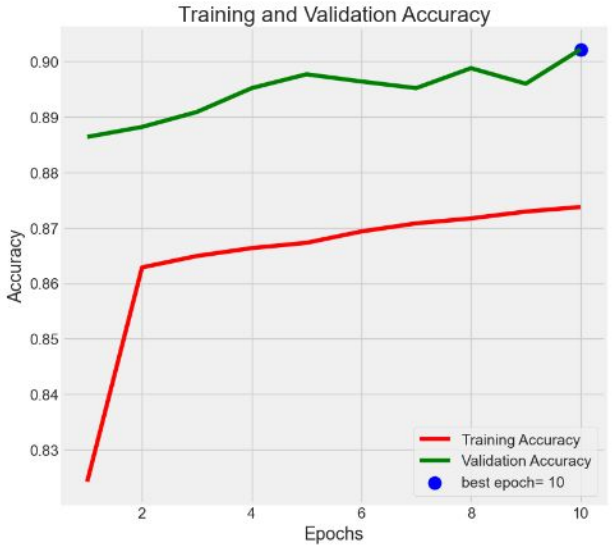
EVALUATION

- Train Loss: 0.50
- Train Accuracy: 0.91

- Valid Loss: 0.52
- Valid Accuracy: 0.90

- Test Loss: 0.52
- Test Accuracy: 0.90





CHALLENGES

- First time utilizing Cloud Computing
- AWS has limits for services - especially for SageMaker, the memory limit was so small that it was impossible to use 12GB of the dataset
    - Requested for a limit increase, but the request hasn't been accepted yet
    - Had to conduct the project on the local machine

- Still waiting for AWS on the limit increase for SageMaker - will try to conduct the project once more on SageMaker
- Will work on the evaluation and prediction parts by Friday

REFERENCES

- https://zenodo.org/records/1214456
- https://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.1002730&type=printable
- https://www.kaggle.com/code/abdallahwagih/human-cancer-tissues-classification-efficientnetb5/notebook
- https://www.kaggle.com/datasets/imrankhan77/NCT-CRC-HE-100K/code