CONCH: Implementation Report and Challenges (ver-1)

Dataset Description:

Dataset link: https://www.kaggle.com/datasets/purpleberrie/train-tcga-coad-msi-mss

The dataset contains patches of WSIs sorted by test, train and validation categories further divided into two classes MSI(named MSIMUT) and MSS.

Structure:

```
tcga_coad_msi_mss
├── train
│   ├── MSIMUT (11.3k)
│   └── MSS (10.8k)
├── test
│   ├── MSIMUT (7,505)
│   └── MSS (11.7k)
└── val
    ├── MSIMUT (7,503)
    └── MSS (11.7k)
```

We utilized the train directory for fine tuning the model and the validation directory for validation purposes.

Note: The dataset mentioned above has been used in all of the following activities, the only addition made to activities 1 and 2 was a prompt file used for captioning and gaining inference via textual assistance.

Stepwise breakdown (Activities):

1. Applied zeroshot classification with pretrained model weights, along with prompts mentioned in the prompt file present on github (bacc = 0.54)

2. Applied zeroshot classification with pretrained model weights, along with self made prompt file specific to msi and mss (bacc = 0.52)

3. Fine-tuned model:
    1. Froze pre-trained layers to retain learned features.

2. Added a new classification head
    - Applied a linear transformation followed by softmax activation for class prediction based on visual inputs.
3. Unfroze and fine-tuned only the classification head.
4. Used cross-entropy loss and Adam optimizer for training.

Results:

72.58% → only trained the classification head and evaluated it (10 epochs)
72.84% → only trained the classification head and evaluated it (30 epochs)
65.23% → trained model's hyperparams and evaluated classification head
68.07% → only trained head and changed transformations, evaluated head

4. Fine Tuned model once again, using same conditions as before, the only change was made to the classification head layers, mentioned below:
    a. Fully connected layer (nn.Linear(visual_output_dim, 256)) reduced dimensionality to 256.
    b. Activated by ReLU.
    c. Dropout layer (p=0.5) after the first linear layer.
    d. Fully connected layer (nn.Linear(256, 128)) further reduced dimensionality to 128.
    e. Activated by ReLU.
    f. Dropout layer (p=0.5) after the second linear layer.
    g. Final layer (nn.Linear(128, num_classes)) outputs predictions for num_classes classes

Summary:
Input dimension: 512
Layers:

- Fully connected layer (512 -> 256) with ReLU activation and 50% dropout
- Fully connected layer (256 -> 128) with ReLU activation and 50% dropout
- Final layer (128 -> 2) for binary classification

Training Parameters:

- Patience for Early Stopping: 10 epochs
- Minimum Delta for Early Stopping: 0.001

Results:

Test Accuracy: 81.99% → convergence at 18 epochs
Convergence time: ~3hrs

Environment Settings:

Dataset Transformation:

- Applied transformations:
  - Resize to (224, 224)
  - Convert to tensor
  - Normalize with mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225]
- Dataset details:
  - Number of training samples: 22117
  - Number of validation samples: 19230

Model Architecture:

- **Pretrained Model:**
  - Used a pretrained model (specified model conch_ViT-B-16) for feature extraction.
- **Custom Classification Head:**
  - Both methods mentioned above in detail activity 3 and 4)

Training Parameters:

- **Optimizer:**
  - Adam optimizer with learning rate: 1e-4
- **Loss Function:**
  - CrossEntropyLoss

Training Setup:

- **Batch Size:** 32
- **Number of Epochs:** 50
- **Device:** GPU P100

Challenges:

1.
2. `Total MSI paths: 28`
3. `Total MSS paths: 88`
4.

Prompts:

```
json_file = {
    "0": {
        "classnames": {
            'MSIMUT': ["microsatellite instable",
                "msi-h",
                "microsatellite instability high",
                "msi high",
                "microsatellite unstable",
                "msi mutant",
                "microsatellite instability",
                "high msi",
                "msi high cancer",
                "msi-high tumor",
                "msi-high"],

            'MSS': ["microsatellite stable",
                "non msi-h",
                "microsatellite stability",
                "msi low",
                "microsatellite stable cancer",
                "mss tumor",
                "microsatellite stable tumor",
                "low msi",
                "stable msi",
                "msi-stable",
                "mss cancer"]
        },
        "templates": [
            "CLASSNAME.",
            "a photomicrograph showing CLASSNAME.",
            "a photomicrograph of CLASSNAME.",
            "an image of CLASSNAME.",
            "an image showing CLASSNAME.",
            "an example of CLASSNAME.",
            "CLASSNAME is shown.",
            "this is CLASSNAME.",
            "there is CLASSNAME.",
            "a histopathological image showing CLASSNAME.",
            "a histopathological image of CLASSNAME.",
            "a histopathological photograph of CLASSNAME.",
            "a histopathological photograph showing CLASSNAME.",
            "shows CLASSNAME.",
```

```
            "presence of CLASSNAME.",
            "CLASSNAME is present.",
            "an H&E stained image of CLASSNAME.",
            "an H&E stained image showing CLASSNAME.",
            "an H&E image showing CLASSNAME.",
            "an H&E image of CLASSNAME.",
            "CLASSNAME, H&E stain.",
            "CLASSNAME, H&E."
        ]
    }
}

#           'MSIMUT': ["mucus",
#               "mucin",
#               "mucus pool",
#               "mucin pool"],
#           'MSS': ["smooth muscle",
#               "smooth muscle tissue",
#               "muscle",
#               "muscularis propria",
#               "muscularis mucosa"]
```

1. Preprocessing (clam)
2. Viv WIV
3. KAT
4. ConvNext
5. Conch (clip)
6. Uni
7. PathChat (MedLlava 2)

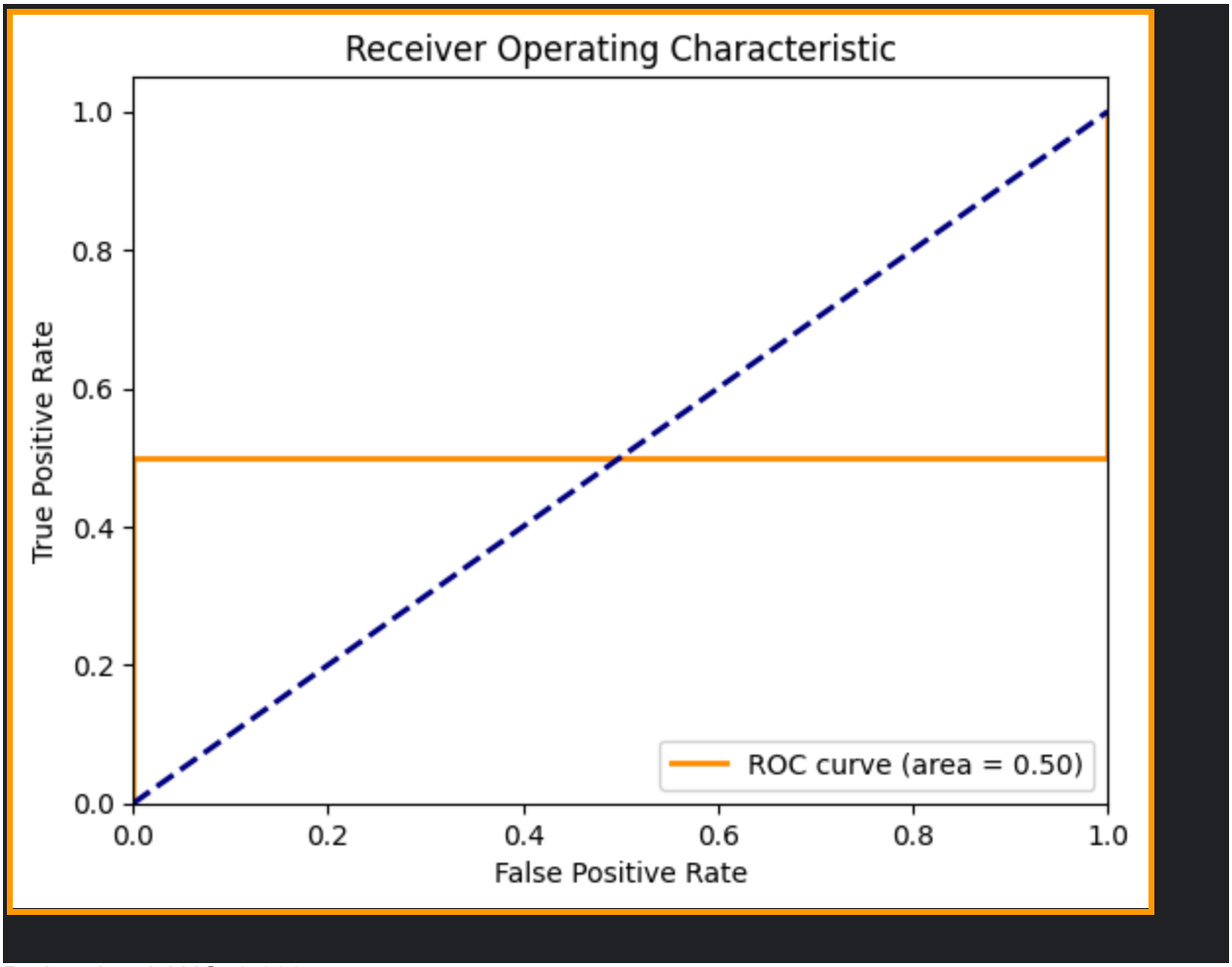
https://ceme.nust.edu.pk/icrai2024/index.html 31st july
https://etecte.uol.edu.pk/ 31st aug
https://hite.ucp.edu.pk/ 31st july
https://fit.edu.pk/ 31st july
https://conferences.ieee.org/conferences_events/conferences/conferencedetails/63607 22 july

1. https://g.co/kgs/eZFTbXY
2. https://www.linkedin.com/jobs/view/3925846649/?utm_source=theFreshDev&ref=theFreshDev /

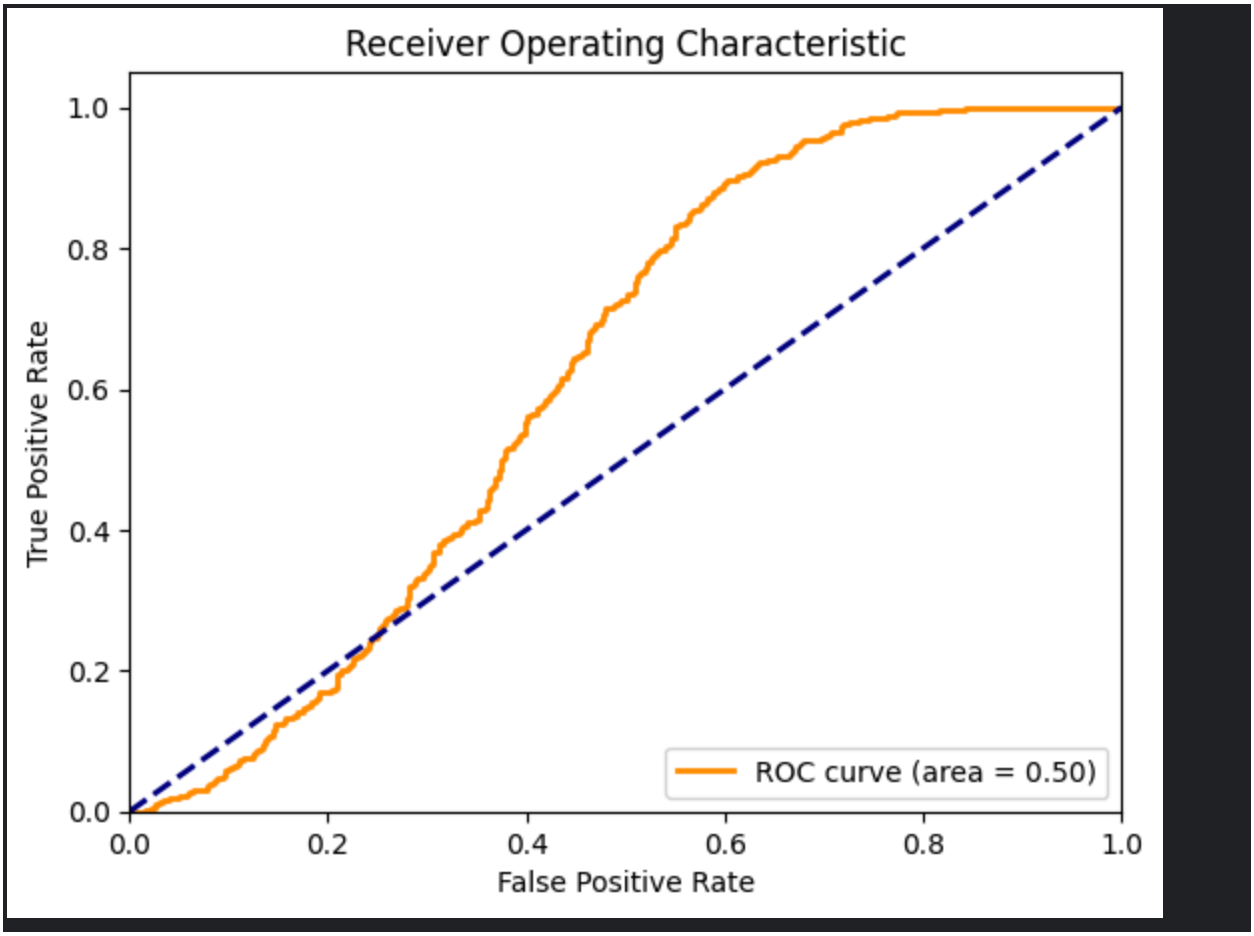
https://www.thefreshdev.com/job/ai-ml-engineer-intern-spring-2025-cranium-2267?utm_campaign=google_jobs_apply&utm_source=google_jobs_apply&utm_medium=organic

3. https://www.karkidi.com/job-details/4695-ai-intern-job?utm_campaign=google_jobs_apply&utm_source=google_jobs_apply&utm_medium=organic
4. https://www.karkidi.com/job-details/44364-internship-deep-learning-research-scientist-3d-generative-modeling-job?utm_campaign=google_jobs_apply&utm_source=google_jobs_apply&utm_medium=organic
5. https://www.linkedin.com/jobs/view/ml-research-intern-at-sahara-ai-3895924907/?utm_campaign=google_jobs_apply&utm_source=google_jobs_apply&utm_medium=organic
6. https://www.karkidi.com/job-details/4493-machine-learning-research-intern-remote-job?utm_campaign=google_jobs_apply&utm_source=google_jobs_apply&utm_medium=organic#google_vignette
7. https://www.karkidi.com/job-details/37445-machine-learning-engineer-intern-job?utm_campaign=google_jobs_apply&utm_source=google_jobs_apply&utm_medium=organic
8. https://www.linkedin.com/jobs/view/3942444937/?trk=li_ziprecruiter_Global_careers_jobsgtm_fa4659c3-0501-447d-878b-e79b6bc5b4d0_job-dist&utm_medium=jobdist&mcid=6810586802156523522&utm_source=ziprecruiter&ePP=CwEAAAGQwndfzyTEc-g7XVRx4Ja1Pid1lMJ8bsqG_wgWcJ9j0Wxg_rHUJudwHx7RHv1t6GRckA_JvzcoD5gCbZC8UApb7EzJNo1c7A7-OQ&ccuid=55335861366&cid=5150289b-3309-4e9e-bd97-c107ccdfd30f

mss percentage: 0.35984848484848486
msi percentage: 0.2314165497896213
Unique labels in all_patient_labels: [0 1]
Sample patient labels: [1 1 0 0]
Sample patient predictions: [0.55399061 0.04391892 0.35984848 0.23141655]
Patient-level AUC: 0.500
Patient-level F1 score: 0.667

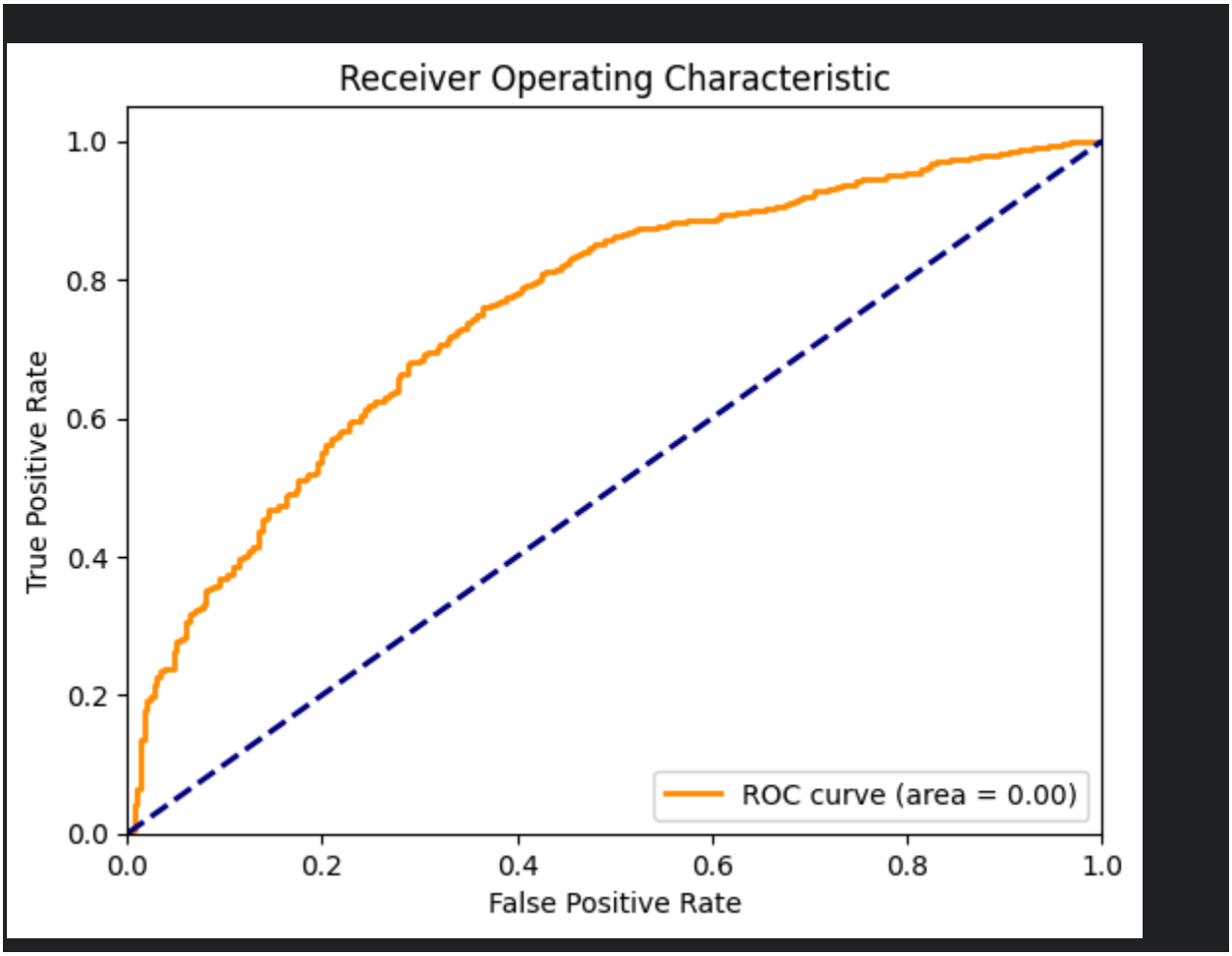


Patient-level AUC: 0.620
Patient-level F1 score: 0.399


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.50)

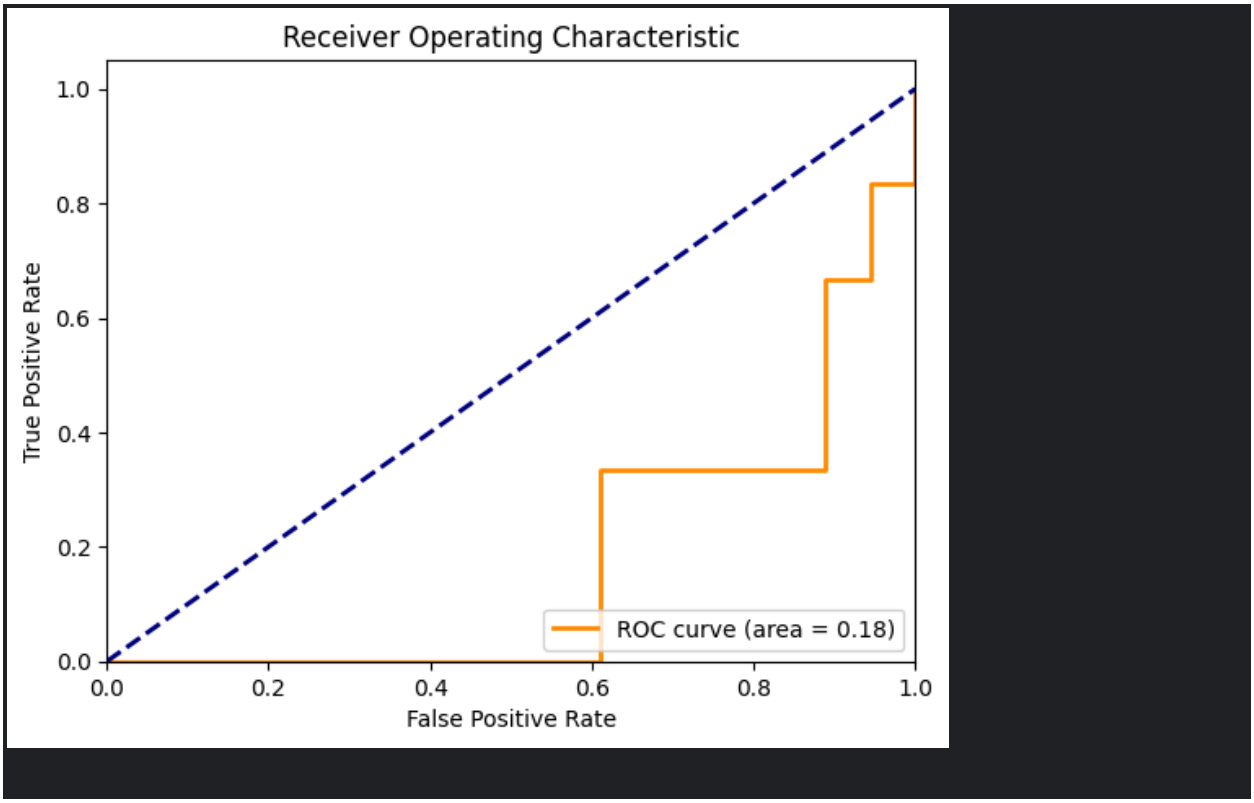
Smoller dataset

Patient-level AUC: 0.756
Patient-level F1 score: 0.530



Entire dataset: 80:20 train:test

Unique labels in all_patient_labels: [0 1]
Sample patient labels: [1 1 1 1 1 1 0 0 0 0]
Sample patient predictions: [0.75088339 0.25516224 0.90053286 0.68986384 0.5754717
0.89285714
 0.66985646 0.39393939 0.96333333 0.9378882]
Patient-level AUC: 0.176
Patient-level F1 score: 0.357



Patient-level AUC: 0.728

Patient-level F1 score: 0.847