**4.Datasets**

**4.1Internal dataset**

For the Internal dataset,we collected 550 patients from Shanghai Tenth People's Hospital as an internal dataset, and divided them into training and validation sets at a 4:1 ratio. The ratio of T0 and T1 stage prostate cancer patients in both the training and validation sets is close to 1:1, ensuring a relatively balanced distribution.

**4.2 External validation datasets**

For the External validation datasets.....

|  |  |  |  |
| --- | --- | --- | --- |
| datasets | Internal datasets  （train） | Internal datasets  （validation） | External validation |
| All patients | 440 | 110 | ... |
| T0 stage patients | 230 | 58 | ... |
| T1 stage patients | 210 | 52 | ... |

Table1.Composition of the Internal datasets and External validation datasets

**5.Implementation**

**5.1 Pretraining**

We implemented our experiments using PyTorch,We used Kinetics-710 pretrained weights for MAE weight initialization.Although there is a domain gap in natural and medical image data, studies show that pretraining on natural image data improves network performance on medical imaging tasks.

We chose ViT-Base as the back bone.We set the patch size as 2x3x16x16,which means our token generation will give out N = tokens for an input video clip for 16x3x224x224.We set the mask ratio as 0.9,resulting in 1568\*(1-0.9) =157 tokens that are visible for vision transformer encoder. The pretraining phase is trained with an AdamW optimizer with LR 0.0004,for minimizing the MSE loss over 500 epochs. The batch size is 4. Warm-up was done for 5 epochs with LR 0.001.

**5.2 Fintuning**

As mentioned in Sec. 3,we divide the video into 16frames to constitute a clip.From each video, we sampled 5 clips uniformly.During inference, we predict the labels for each of the clips. If any of the clips is predicted as malignant, the entire video is labelled as malignant.We minimized a cross entropy loss using an AdamW optimizer with LR 0.0004.

We process our experiments with 2 NVIDIA GeForce RTX 3090 GPUs for both pretraining and fintuning.

**6.Experiments and Results**

**6.1Experiment**

For other works that can use our dataset, we classify them into image-based and video-based works. For video-based works, we directly obtain a score from the model for the input validation set videos, and calculate the accuracy, AUC, and ROC curve based on this score and the actual disease status of the videos. For image-based works, we split the video into frames and obtain the model's evaluation of each frame. The final score for a video is obtained by averaging the scores of all frames in the video, and then we calculate the accuracy, AUC, and ROC curve based on this score and the actual disease status of each video.

The result is shown in Table2.

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Method | Acc | AUC |
| Image-Based |  |  |  |
|  |  |  |
|  |  |  |
| RadFormer |  |  |
| Video-Based |  |  |  |
| VideoMAEv2 | 0.914 | 0.942 |
| Nzk-MAE(ours) | **0.9236** | **0.963** |

**6.2 Ablation Experiment mask**

As ablation analysis,We present the final accuracy and ROC curve based on the random masking pretraining method, as well as the final accuracy and ROC curve based on the segmentation region masking method in the Figure 3 and Table 3.

|  |  |  |
| --- | --- | --- |
| Method | Accuracy | AUC |
| Random Mask | 0.8207 | 0.875 |
| Segment Guided Mask (our work) | **0.9236** | **0.963** |

Table 3.Accuracy of Random Mask and Segment Guided Mask (our work)

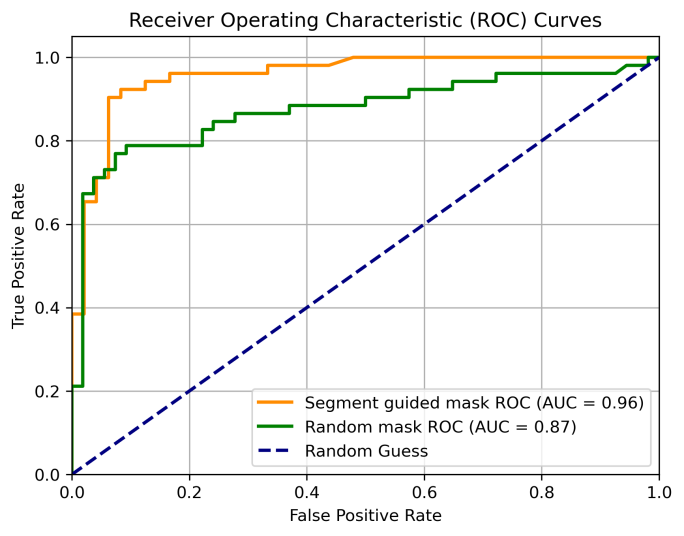


Figure 4.Performance of our work and random mask in AUC.(Internal datasets)