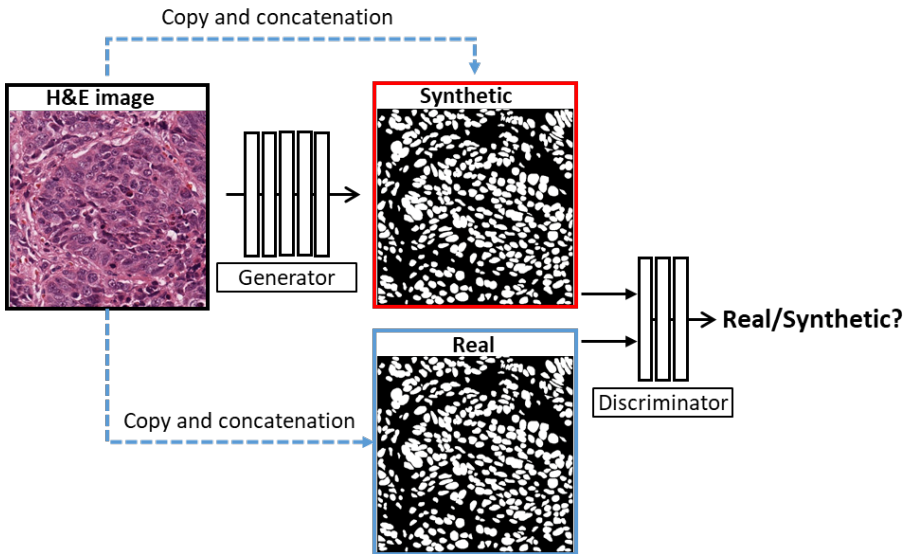
Supplemental Material: Computerized tumor multinucleation index (MuNI) is prognostic in p16+ oropharyngeal carcinoma: A multi-site validation study

Supplementary Method 1: Network architecture and validation of the model

Two cGAN models were utilized for MuNI calculation, each of which was built to learn different tasks, segmentation of MN events (GAN_{MN}), and EP cells (GAN_{EP}). After the models were built, the MuNI calculation step starts with the extraction of tiles from slide tissue regions. Then, the tiles were inputted to MN and EP models to generate their corresponding masks. Both segmentation models generate output images with the same size as the input images. Output of GAN_{MN} is a colored image where multinucleated cells, other segmented cells, and background regions appear blue, white, and black, respectively. Output of GAN_{EP} is a binary map, where black and white pixels illustrate epithelium and other regions, respectively.

The GAN segmentation model used in this study is shown in Suppl. Figure 1. Our developed model contains a generator and a discriminator. The generator is a basic end-to-end codec structure. In order to make the training of model stable, we used a structure similar to encoder of the generator in the discriminator. Both generator and discriminator use modules of the form convolution-BatchNorm-ReLu (1). In the generator, we used convolutional and deconvolutional layers to downsample and upsample features, so the deep features can be fused. Our loss function contains two parts: cGAN loss (2) and feature matching loss (3).

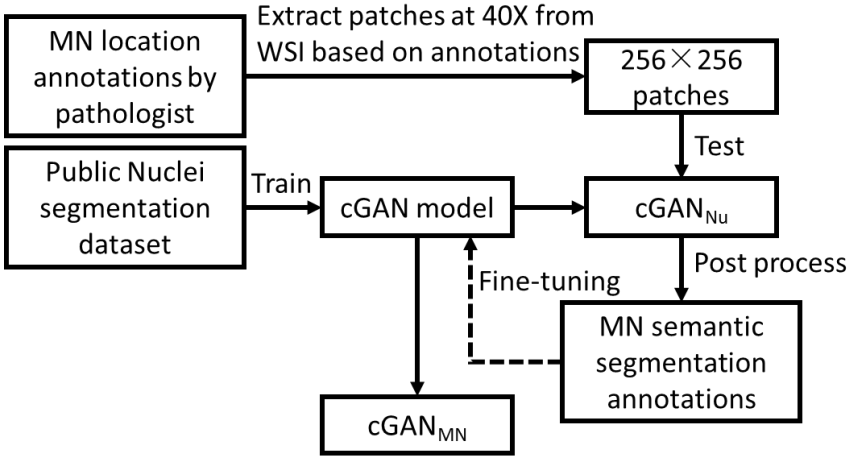


Suppl. Figure 1 – GAN model architecture

MNs were annotated by the collaborating pathologist (JSL) using 12 WSIs from S_{TR}, which resulted in 1,002 annotations. For the sake of efficiency, the pathologist located the center of multinucleated cells instead of drawing entire cell boundaries. Before fine-tuning the model with MN annotations, the boundaries of the annotated MNs were delineated automatically by using the model trained for the cell segmentation using a publicly available dataset (Suppl. Figure 2). The training of GAN_{MN} started with learning nuclear segmentation using 30 images from a publicly available nuclear segmentation data set. The image size is 1,000×1,000 at 40x magnification. We resized these images to 1,024×1,024, and then tiled into 480 256×256 images to be fed into the network. We used a 7:3 ratio to randomly divide the dataset into training set and test set for GAN_{MN} training and validation. The pixel-level F1-score of GAN_{MN} on validation dataset was 0.93. In the second step, we automatically colored region of annotated MN with blue to generate the MN segmentation data

set. GAN_{MN} that had been trained to detect any cells independent of the cell type, it was fine-tuned for differentiating MNs from other types of cells such as epithelial cells and lymphocytes using the MN segmentation data set. 9 WSIs having a total of 668 MNs were used for training and 3WSIs having a total of 334 MNs for validation. The GAN_{MN} model yielded a pixel-level F1-score of 0.76 on the validation images.

Efforts were made to reduce the false positive rate of GAN_{MN}. Given that MNs should be distributed in the epithelial regions only, we trained GAN_{EP} to ignore MNs identified outside the epithelial regions. GAN_{EP} was built and evaluated using a set of 6 cases from S_{TR}. A total of 153 image patches each corresponding to 512x512 pixels were cropped at 10x magnification and then annotated by a pathologist. 102 of them were used for training GAN_{EP}. Its performance was then evaluated quantitatively on the remaining 51 images and yielded a pixel-level F1-score of 0.88.



Suppl. Figure 2 – G_{MN} training pipeline

Supplementary Method 2: Different variants of the MuNI

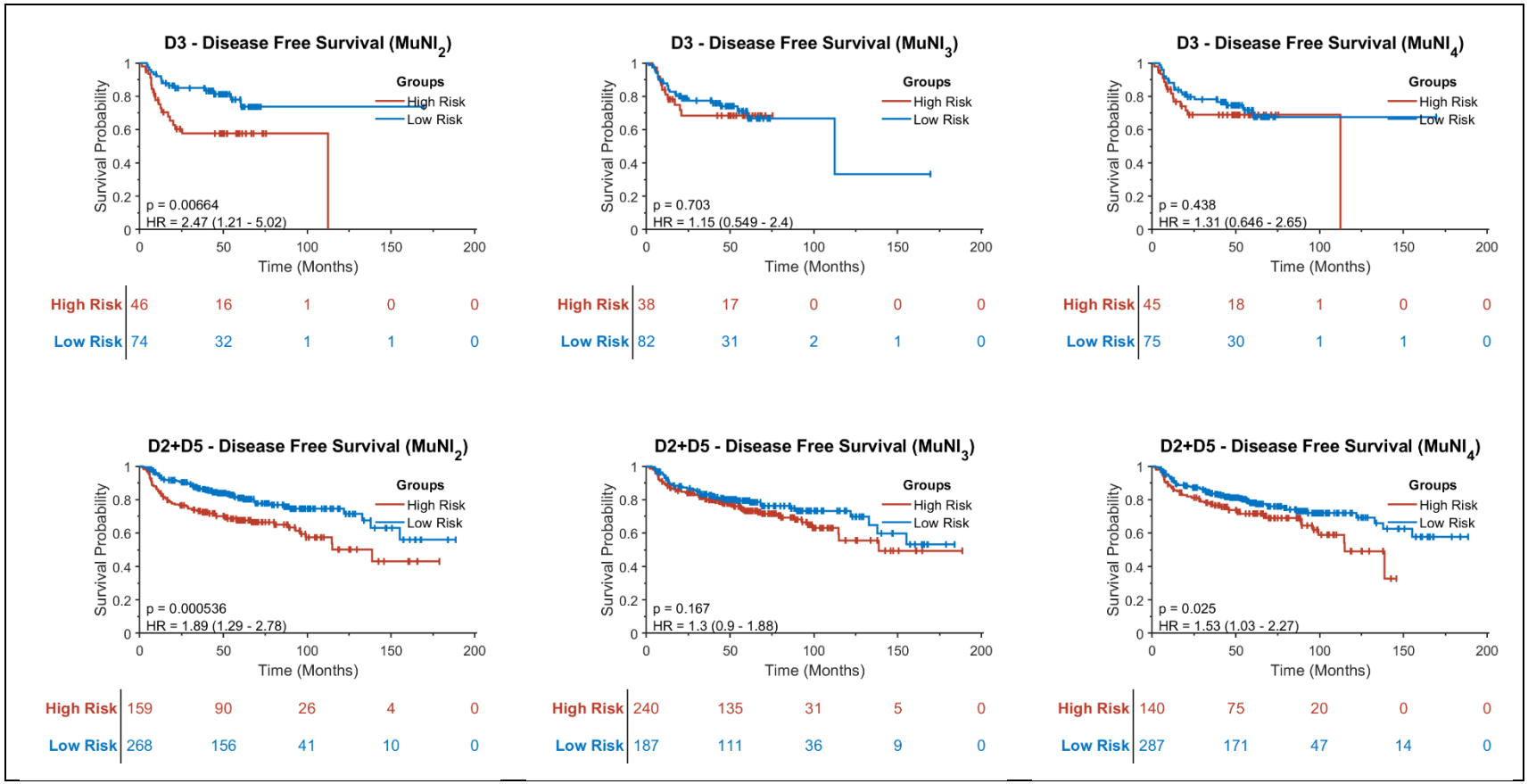
For a WSI, *m* was used to denote the number of tiles extracted from the WSI and M_{MN}^{i} and M_{EP}^{i} corresponded to the number of detected MNs and EP cells in tile *i* extracted from the WSI, respectively. The normalized MuNI for the WSI was then defined as the ratio of total MNs to EP cells. Different variants of the MuNI were also analyzed in terms of their prognostic abilities. One of them, $MuNI_2$, normalizes the number of MNs to the number of total cells, instead of EP cells. In other MuNI variants, we further partitioned tissue compartments into tumor and non-tumor regions utilizing another convolutional neural network (VGG19) (4). Then, two indices were calculated by normalizing the MNs 1) by the number of EP, $MuNI_3$, and 2) by the total number of cells, $MuNI_4$, both measurements were carried out within the tumor regions of the WSI only. The variants of MuNI are defined as follows:

$$MuNI_2 = \frac{\sum_{i=1}^{m} M_{MN}^{i}}{\sum_{i=1}^{m} M_{TC}^{i}}$$

$$MuNI_3 = \frac{\sum_{i=1}^{m} M_{MN_{TUMOR}}^{i}}{\sum_{i=1}^{m} M_{EP_{TUMOR}}^{i}}$$

$$MuNI_4 = \frac{\sum_{i=1}^{m} M_{MN_{TUMOR}}^{i}}{\sum_{i=1}^{m} M_{TC_{TUMOR}}^{i}}$$

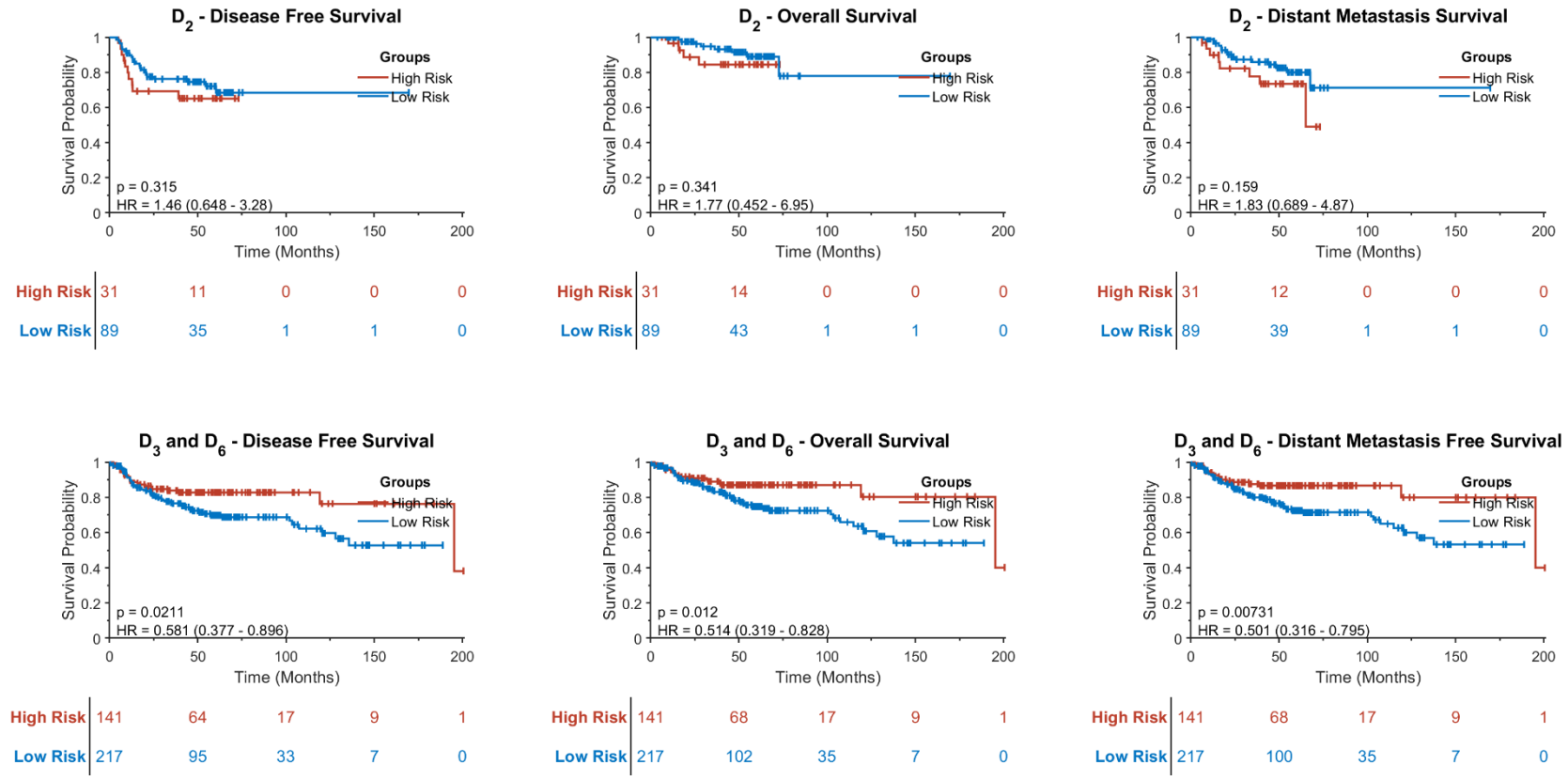
where M_{TC}^{i} is the number of total cells in tile *i*. A cutoff to stratify the patients into high- and low- risk was determined as the mean value of MuNIs within the set D_3, and then applied to the combination of D_2 and D_5 (Suppl. Figure 3).



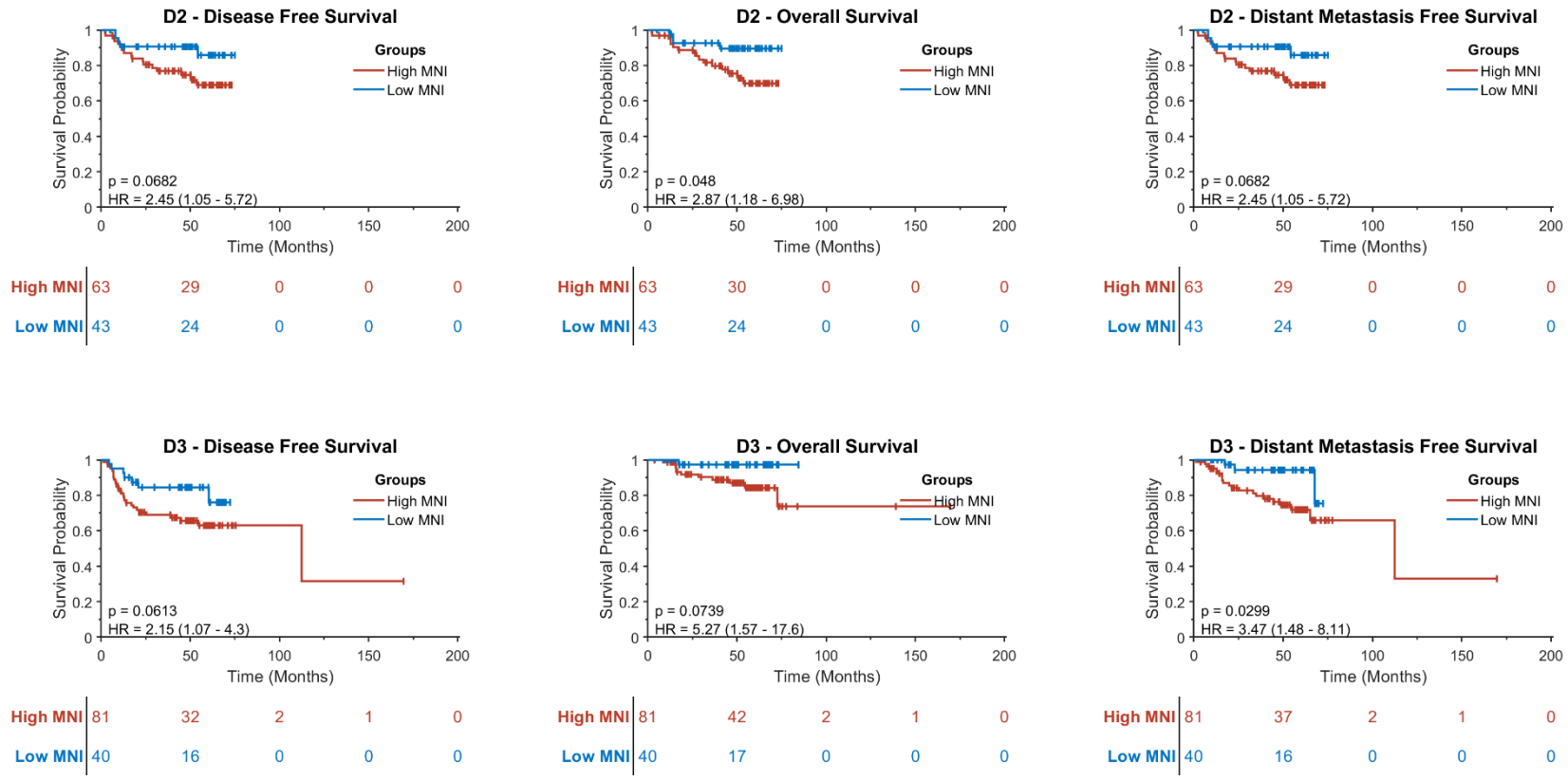
Suppl. Figure 3 – KM- DFS curves as calculated for the different variants of MuNI; $MuNI_2$ (left most column), $MuNI_3$ (center column), and $MuNI_4$ (right most column), for the patients in D_3 (top row) and the combination of D_2 and D_5 (bottom row).

Supplementary Method 3: Pathologist's visual reads of MNs

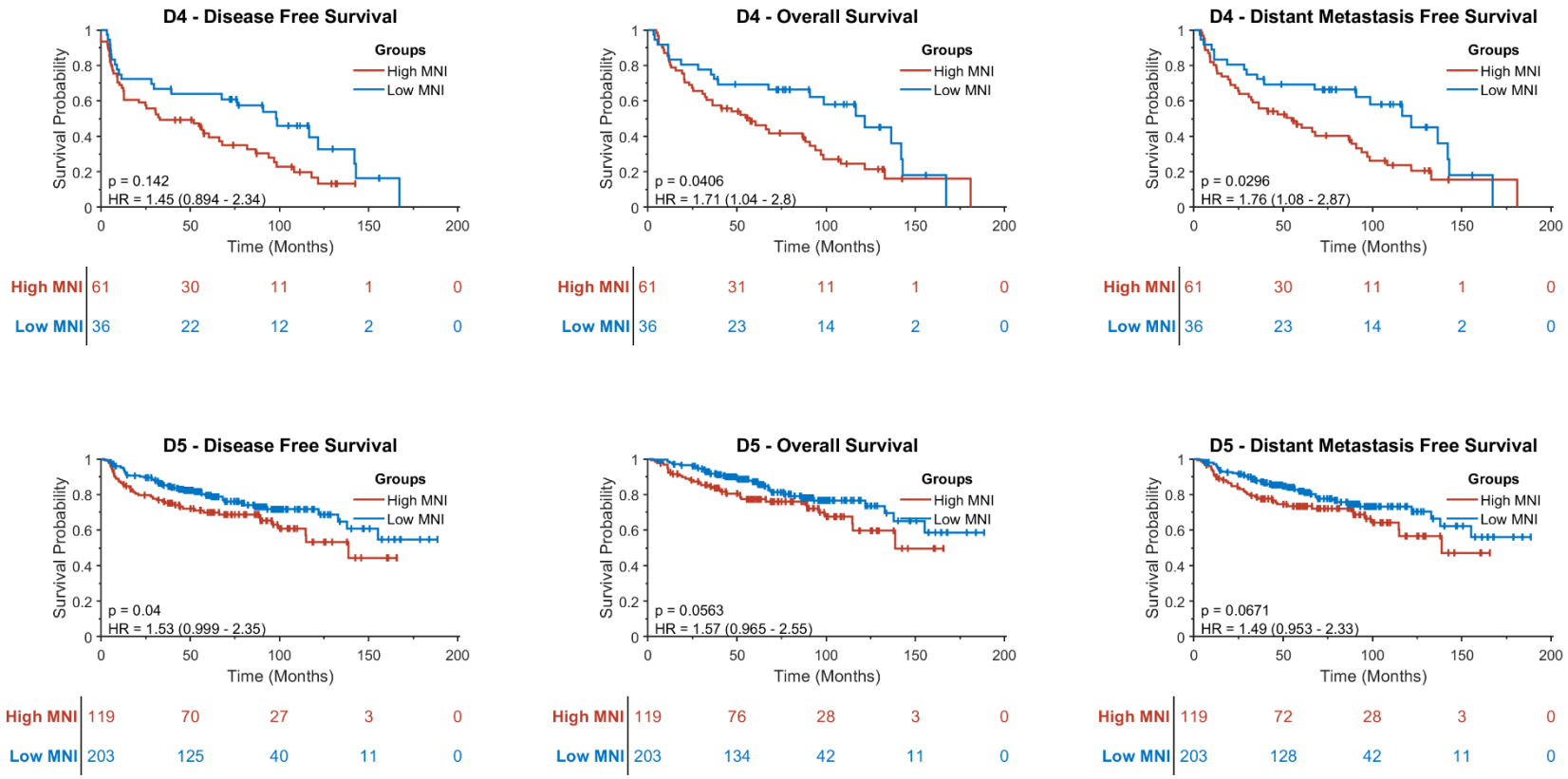
The representative single H&E tumor slides from the D_2, D_3, and D_6 cohorts were reviewed visually by the main study pathologist (JSL) for the presence and semiquantitation of MN without knowledge of patient outcomes. A MN cell per visual analysis was defined as one that clearly had 3 or more nuclei in the same cell. Once a hotspot (highest area of MN) was identified, 10 consecutive high-power fields were counted for MN cells generating a visual MN "index" ($MuNI_{Manual}$) between 0 and maximum number of cells identified. A cutoff to stratify the patients into high- and low- risk was determined as the mean value of $MuNI_{Manual}$ within the set D_2, and then applied to the combination of D_3 and D_6. The KM survival curves show that the visual reads were not prognostic for DFS, OS or DMFS (Suppl. Figure 4).



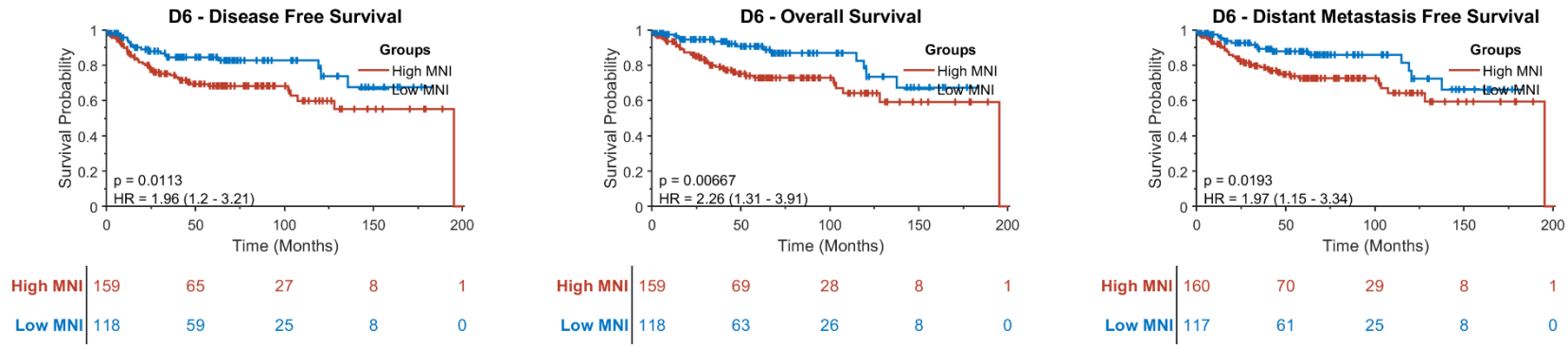
Suppl. Figure 4 – KM- DFS, OS, and DMFS curves for $MuNI_{Manual}$ among the patient in D_2 (top row), and for the combination of D_3 and D_6 (bottom row).



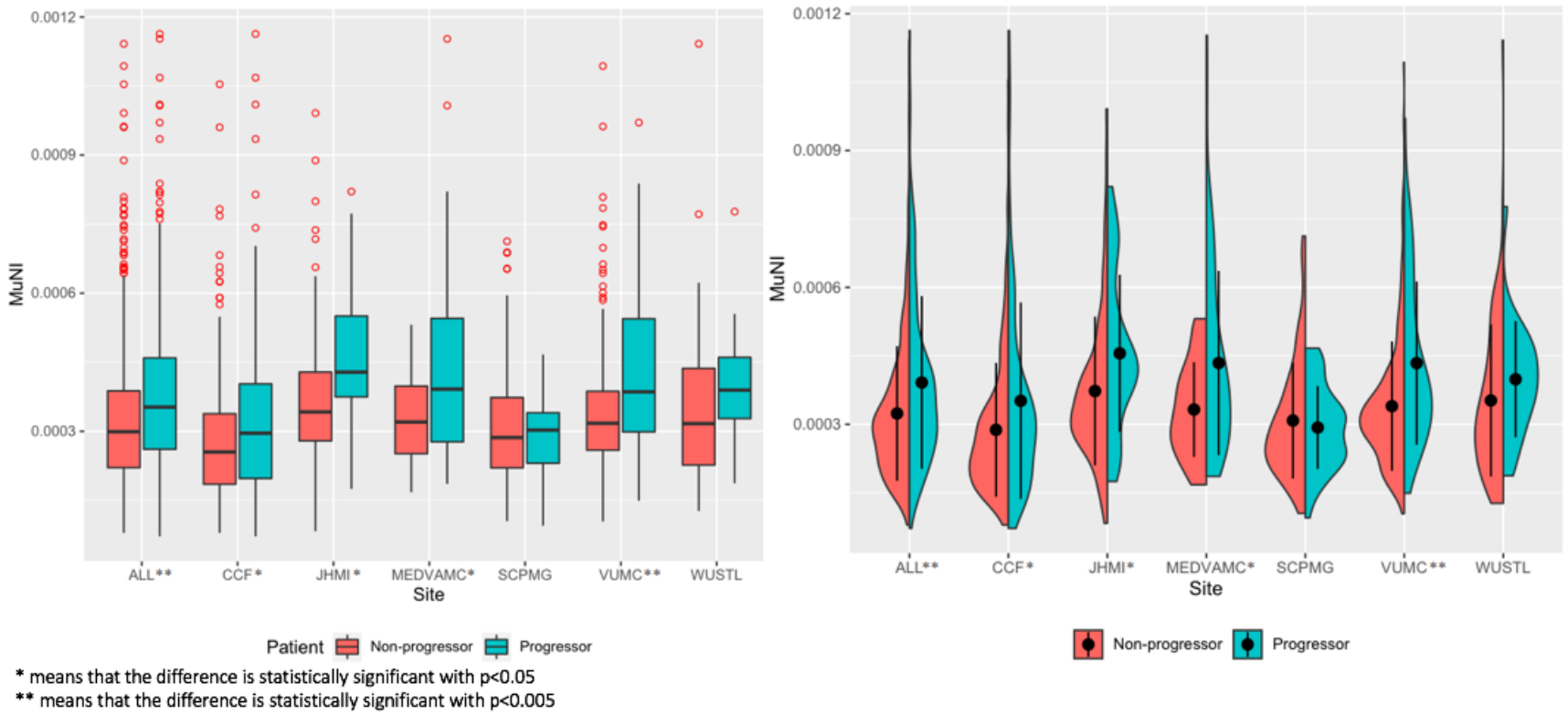
Suppl. Figure 5 – KM- DFS, OS, and DMFS curves for the patients in D_2 (top row) and D_3 (bottom row) stratified into high- and low-risk by MuNI. A cutoff to stratify the patients into high- and low- risk was determined as the mean value of MuNIs within the set D_1.



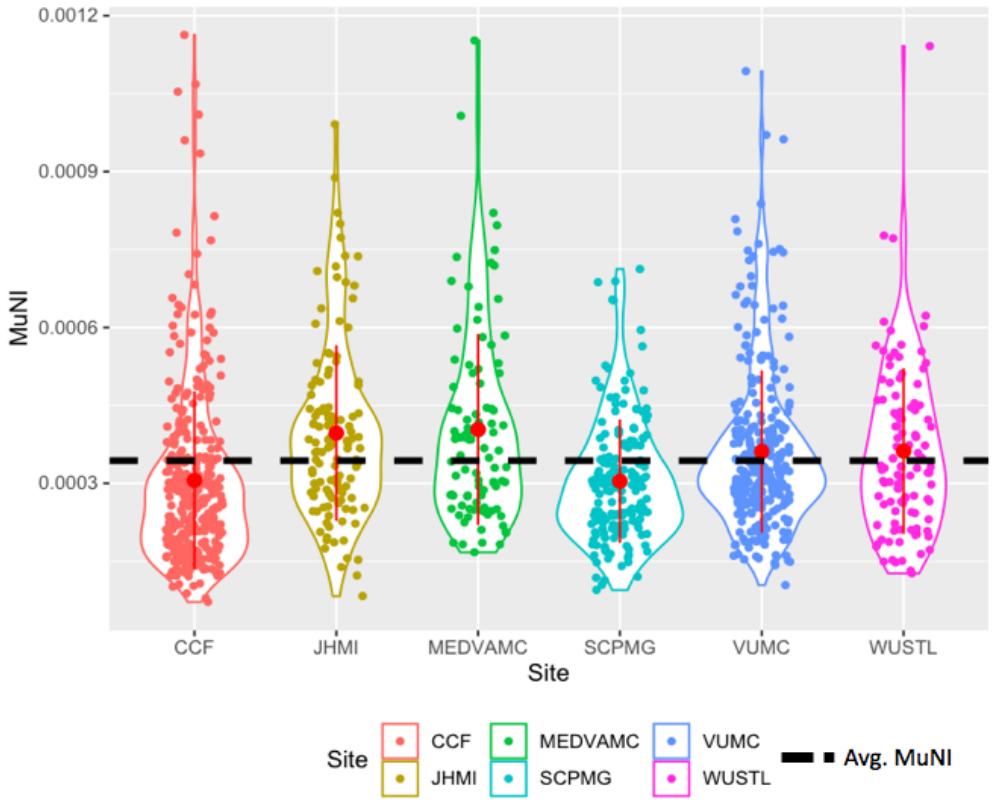
Suppl. Figure 6 – KM- DFS, OS, and DMFS curves for the patients in D_4 (top row) and D_5 (bottom row) stratified into high- and low-risk by MuNI. A cutoff to stratify the patients into high- and low- risk was determined as the mean value of MuNIs within the set D_1.



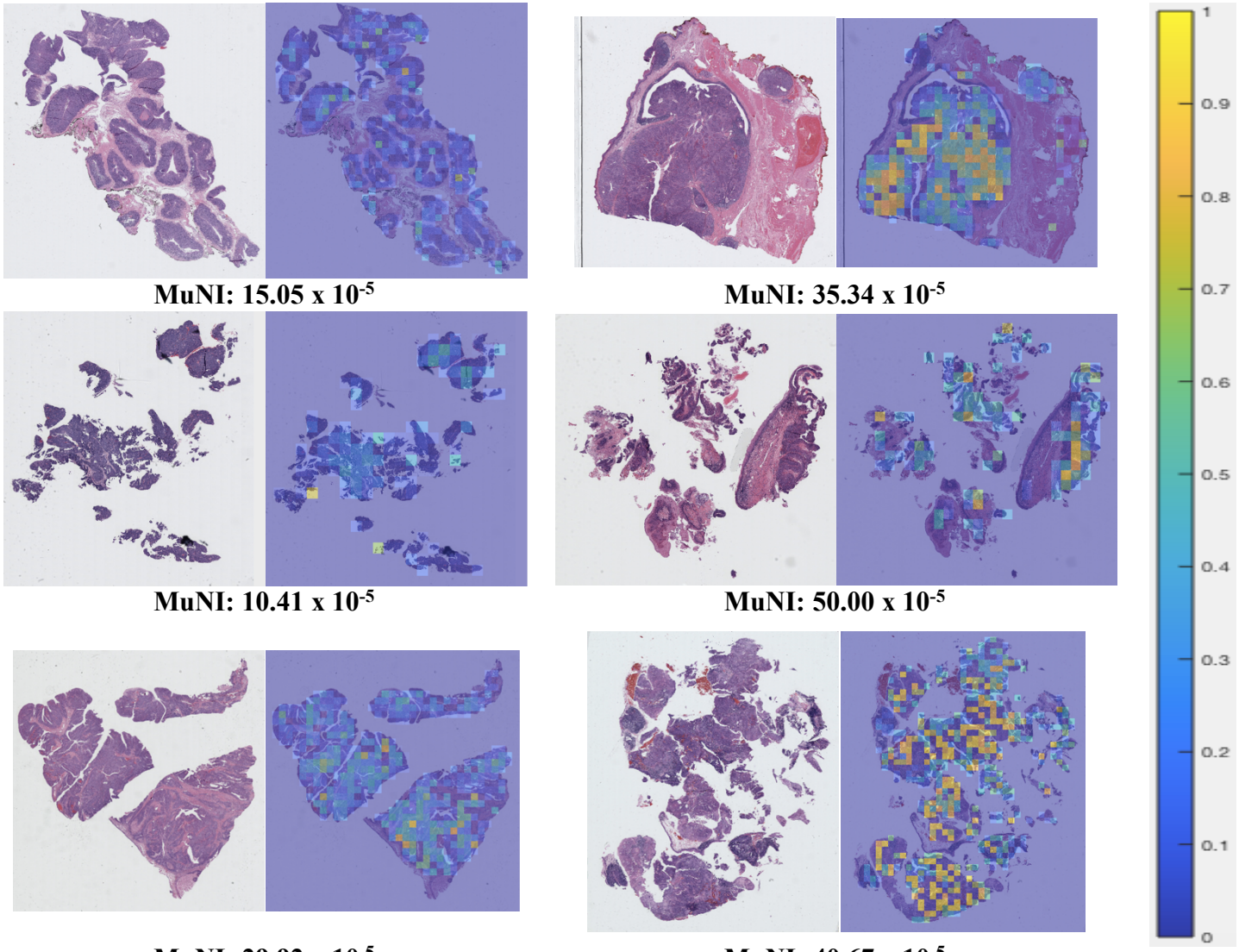
Suppl. Figure 7 – KM- DFS, OS, and DMFS curves for the patients in D_6 stratified into high- and low-risk by MuNI. A cutoff to stratify the patients into high- and low- risk was determined as the mean value of MuNIs within the set D_1.



Suppl. Figure 8 – Comparison of patients who would develop tumor progression versus those who would not in terms of their MuNIs across different sites. ** means that difference between the groups is statistically significant with $p<0.005$. * means that the difference between the groups is statistically significant with $p<0.05$ in Mann-Whitney U test.



Suppl. Figure 9 – Distributions of MuNIs across different sites. Red lines show mean ± one standard deviation range. Black dashed line corresponds to the mean of the calculated MuNIs in the entire dataset.



Suppl. Figure 10 – Visual comparison of three high-risk and low-risk samples. MuNIs were calculated for each tile and overlaid on top of WSIs as heatmaps. Yellow color shows the regions with higher MN density, larger MuNI score, whereas blue shows the opposite.

Suppl. Table 1 – Summary of clinical and pathological features of all six cohorts. ± denotes one standard deviation below/above the mean.

Summary of clinical and pathological features of all six cohorts							
	D_1 N (%)	**D_2 N (%)**	**D_3 N (%)**	**D_4 N (%)**	**D_5 N (%)**	**D_6 N (%)**	**ANOVA (2-sided) P-value***
No. of patients	171	106	121	97	322	277	
Age	57.18 ± 9.7	57.62 ± 9.6	57.2 ± 8.3	60.9 ± 9.1	58.73 ± 9.1	57.7 ± 9.6	**0.01**
Gender							
Male	154 (90.06)	89 (83.96)	111 (91.74)	96 (99.97)	285 (88.5)	253 (91.4)	**0.01**
Female	17 (9.94)	17 (16.04)	10 (8.26)	1 (0.03)	37 (11.49)	24 (8.6)	
Race							
White	157 (91.81)	103 (97.17)	114 (95.00)	78 (80.41)	299 (94.0)	271 (97.8)	
Non-white:							**0.02**
Black	10 (5.85)	3 (2.83)	6 (5.00)	14 (14.32)	19 (5.98)	6 (2.16)	
Asian	4 (2.34)	0 (0)	0 (0)	0(0)	0(0)	0(0)	
Smoke							
Ever	111 (67.46)	70 (66.04)	79 (65.83)	79 (81.44)	214 (66.46)	172 (62.1)	**0.03**
Never	60 (32.54)	36 (33.96)	41 (34.17)	18 (18.56)	108 (33.5)	105 (37.9)	
Treatment							
Surgery w. adjuv. therapy	97 (56.73)	64 (60.38)	24 (19.84)	3 (3.09)	0 (0)	79 (28.52)	**<0.001**
Surgery alone	0 (0)	18 (16.98)	1 (0.83)	0 (0)	0 (0)	38 (13.72)	
Definitive non-operat. treat.	72 (42.11)	24 (22.64)	96 (79.34)	94 (96.91)	322 (100)	100 (36.10)	
Specimen type							
Resection	97 (56.73)	82 (77.36)	25 (20.66)	3 (3.09)	0 (0)	117 (42.24)	**<0.001**
Biopsy	72 (42.11)	24 (22.64)	96 (79.34)	94 (96.91)	322 (100)	100 (36.10)	
T-Stage							
T1/T2	91 (53.22)	73 (68.87)	79 (65.26)	56 (57.73)	203 (63.04)	38 (13.7)	**<0.001**
T3/T4	80 (46.78)	33 (31.13)	42 (34.74)	41 (42.27)	119 (37.0)	197 (71.2)	
N-Stage							
N1/N0	127 (74.27)	77 (72.64)	85 (70.25)	24 (24.74)	209 (64.9)	183 (66.1)	**<0.001**

N2/N3	44 (23.73)	29 (27.36)	36 (29.75)	73 (75.26)	113 (35.1)	52 (18.77)	
Overall stage I/II III	 126 (73.68) 45 (26.32)	 87 (82.08) 19 (17.92)	 92 (76.67) 28 (23.33)	 70 (72.17) 27 (27.83)	 240 (74.5) 82 (25.5)	 226 (81.6) 15 (5.42)	**<0.001**
DFS (months) Event Non-event	74.88 ± 37 44 (25.73) 127 (74.27)	43.1 ± 21.2 23 (21.70) 83 (78.30)	40.14 ± 26.0 35 (28.93) 86 (71.07)	59.1 ± 49.1 68 (70.10) 29 (29.90)	66.8 ± 42.3 92 (28.57) 230 (71.4)	55.8 ± 47.8 64 (23.11) 213 (76.9)	**<0.001**
OS (months) Event Non-event	79.3 ± 33.5 37 (21.64) 134 (78.36)	45.5 ± 19.6 20 (18.87) 86 (81.13)	47.02 ± 24.6 12 (9.92) 109 (90.08)	64.90 ± 47.04 64 (65.98) 33 (34.02)	70.5 ± 39.8 71 (22.05) 251 (78.0)	58.5 ± 46.9 52 (18.77) 225 (81.2)	**<0.001**
DMFS (months) Event Non-event	77.03 ± 35.5 40 (23.39) 131 (76.61)	43.91 ± 20.9 22 (20.76) 84 (79.25)	44.05 ± 24.1 24 (19.84) 97 (80.17)	64.08 ± 47.4 65 (67.01) 32 (32.99)	68.55 ± 41.3 84 (26.09) 238 (73.91)	57.97 ± 47.1 55 (19.86) 222 (80.14)	**<0.001**
MuNI	3.04×10^{-4} ± 1.19×10^{-4}	3.62×10^{-4} ± 1.59×10^{-4}	3.97×10^{-4} ± 1.69×10^{-4}	4.03×10^{-4} ± 1.84×10^{-4}	3.06×10^{-4} ± 1.71×10^{-4}	3.61×10^{-4} ± 1.56×10^{-4}	**<0.001**

References:

1. Radford, Alec, et al. “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint, 2015.

2. Isola, Phillip, et al. “Image-to-image translation with conditional adversarial networks,” Proceedings of the IEEE conference on CVPR, 2017.

3. Salimans, Tim, et al. “Improved techniques for training gans,” Advances in neural information processing systems, 2016.

4. K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition,” In ICLR, 2015.