Supplementary Material

Supplementary Method

Image pre-processing

We annotated multiple sections from 93 GoCAR[S1] PAS stained slides for tissue compartment and mononuclear leukocytes (MNL) prediction using ASAP software (https://computationalpathologygroup.github.io/ASAP/) under the guidance of pathologists. Each section was outlined by a boundary and the area within boundary but apart from annotated objects were defined as interstitium by default. The raw sections were divided into 22,692 fixed-sized image tiles at 20x objective-power and then transformed by data augmentation process[S2] including position shifting, rotating, flipping, perspective transforming, color transferring, contrast transforming, or noise feeding. Every fixed-sized tile image was paired with a ground truth classification image labeling group information at the same size for model evaluation. These two types of images were served as input in model construction process.

Deep-learning model generation

We split 93 slides roughly in the ratio 2:1 into discovery set (n=60) and testing set (n=33). Discovery set were used for model construction in training process and testing set were used for final evaluation and were kept untouched during training process. We performed 10-fold cross-validation within discovery set thus divided the discovery set into 10 equal sized portions. During each model training process, we used 9/10 of samples as training set and the left-out 1/10 of samples as validation set to tune the model. As a result, 10 separate base models were created based on 10 partitions and the final prediction were made by aggregating results from each base model.

Two types of CNN structures were used in this study: a model segmenting images at instance level: Mask Region-based Convolutional Neural Network (Mask R-CNN)[S3] and a model segmenting images at pixel level: U-Net[S4]. The former model Mask R-CNN first extracts feature maps from input images by a convolutional backbone structure and generates region proposal of given objects through Region Proposal Network (RPN). These proposed regions are later passed through another neural network to generate multi-categorical classes, bounding boxes and masks for objects. The U-Net model on the other hand makes predictions for each pixel instead of instance on input images. It has a symmetrical "U" shape architecture consisting of an encoder which extracts features from input images by convolution blocks and a decoder which expends contracted vector back to segmentation map at input size. The number of blocks in encoder step is the same as the number of blocks in decoder step.

We constructed a compartment detection model and a MN leukocyte detection model using Mask R-CNN structure[S5] and a tissue segmentation model using U-net structure[S6]. Specifically, the compartment detection model was trained for 90 epochs with batch size of 10 using 1024x1024 pixels input images; the MN leukocyte detection model was trained for 250 epochs with batch size of 6 using 512x512 pixels input images; the tissue segmentation model was trained for 100 epochs with batch size of 8 using 512x512 pixels input images. The detection model used RestNet-

101 as backbone and was tuned based on pre-trained weights from MS COCO dataset[S7]. We applied GSD optimizer with learning rate of 0.001 and momentum of 0.9. Total loss was calculated as the sum of loss of RPN classifier, RPN bounding box, MRCNN classifier, MRCNN bounding box, and MRCNN mask, where cross-entropy was used for classification problems and smooth L1-loss was used for bounding box refinement. The segmentation model was constructed using 3 down-sampling layers and 3 up-sampling layers and the first layer contained 32 feature maps. ADAM optimizer was chosen for weight updating at learning rate of 0.001 and cross-entropy was used for loss function. Best epoch/model was determined by evaluating loss from each training and validation partition. Finally, we applied the best model to testing data set for unbiased model evaluation. Accuracies were measured by True Positive Rate (TPR) and Positive Predictive Value (PPV) and general F_β score[S8] where β=2.

The GPU machine we used was equipped with 36 Intel(R) Xeon(R) W-2195 CPUs (18 cores), 128 GB Memory, and 4 GPUs of Quadro RTX 8000. All processes ran on Ubuntu 18.04 system.

Whole slide investigation and digital feature definition

We applied above three models to whole slide images (WSI) and assembled results into full prediction masks including all tissue compartment objects and MN leukocytes. Since this study focused on the features in cortex, medullar region and adjacent artery along with imperfectly cut or scanned fragments were excluded in WSI analysis. Due to the instance and pixel level prediction nature of MRCNN and U-Net, we were able to perform object counting as well as area estimation. In general, from WSI prediction, we defined basic features such as the size of a slide, the number of glomeruli and tubules, the percentage of glomeruli, tubules and interstitial area over slide, as well as a series of abnormal features focusing on interstitial space, abnormal tubules and inflammation.

To define slide-wide abnormal features, we first introduced the concept of Region of Interest (ROI) window to identify local abnormal regions with respect to interstitium and inflammation (Figure S1A). Given a whole slide prediction image, we applied a 384x384 pixels unit window sliding over the image with stride of 128 pixels. Within each unit window, we examined three metrics and defined two types of ROI: **interstitial ROI** and **inflammatory ROI**. A unit window was determined as interstitial ROI if it had wide interstitial space but narrow space of background noise in tubule-enriched regions as defined as: $Sparsity(I) > 0.35,$ and $Density(O) < 0.2,$ and $Density(B) < 0.2,$ where

$$Sparsity(I) = \frac{area(Interstitial\ Space)}{area(Interstitial\ \ Space)+area(Tubule)} \text{ per unit window} \qquad (1)$$

$$Density(O) = \frac{area(Glo)+area(Other\ groups)}{384\times384} \text{ per unit window} \qquad (2)$$

$$Density(\mathrm{B}) = \frac{\mathrm{area(backgroud)}}{384\times384} \text{ per unit window} \qquad (3)$$

A unit window was determined inflammatory ROI if it had enriched mononuclear leukocytes (MNL) as $Density(MNL)\ > 43$, where

$$Density(MNL) = N(\mathrm{MNL}) \text{ per unit window} \qquad (4)$$

After one slide is processed with sliding-window scanning, the pipeline generated two types of ROI masks highlighting abnormal interstitium area and MN leukocytes infiltration area, respectively.

Abnormal features were then defined in terms of interstitial space, abnormal tubules and MN leukocytes infiltration at WSI level or ROI level:

- Interstitium features: Feature (5) estimates overall percentage of intestinal space over WSI area.

$$\boldsymbol{Abnormal\ Interstitial\ Area\ Percentage} = \frac{area(Interstitial\ Space\ within\ \boldsymbol{interstitial\ ROI})}{area\ (WSI)} \tag{5}$$

- Abnormal tubules feature: Feature (6) summarizes number of abnormal tubules per 1000x1000 unit area.

$$\boldsymbol{Abnormal\ Tubules\ Density} = \frac{N(Abnormal\ Tubules)}{area(WSI)} \times 10^6 \tag{6}$$

- Inflammation features: Feature (7) estimates proportion of MN leukocyte enriched area over WSI area. Feature (8) summarizes average number of MN leukocyte in inflammatory ROI per 1000x1000 unit area.

$$\boldsymbol{MNL\text{-}enriched\ Area\ Percentage} = \frac{area(\boldsymbol{Inflammatory\ ROI})}{area(WSI)} \tag{7}$$

$$\boldsymbol{MNL\ Density}\ (\boldsymbol{infR}) = Average\ N(MNL)\ weighted\ across\ \boldsymbol{Inflammatory\ ROIs} \tag{8}$$

Basically, the digital features were defined in consideration of two aspects: i) how widespread is given abnormal feature over slide such as Abnormal Interstitial Area Percentage (5) and MNL-enriched Area Percentage (7); ii) how dense is given abnormal object per unit area such as Abnormal Tubules Density (6) and MNL Density (infR) (8).

Moreover, we integrated above individual features into composite scores. Because density features had skewed distribution, we first rescaled them through log2 transformation. By multiply coverage feature (area %) by density feature, the Interstitial and Tubular Abnormality Score (ITAS) and MNL infiltration Score (MLIS) were proposed to approximate relative amount of IFTA (Interstitial Fibrosis and Tubular Atrophy) and MNLs. The max(x,0) function in formula (9) and (10) serves as a gate function to ensure non-negative values. A final Composite Damage Score(CDS) was proposed to integrate abnormality regarding all three aspects. Since our model tends to recognize MNLs within interstitial space, we assume the proposed CDS approximates i-IFTA at certain level.

$$\boldsymbol{MNL\ Infiltration\ \ Score}\ (\boldsymbol{MLIS}) = \max\left(\boldsymbol{MNL\text{-}enriched\ Area\ Percentage} \times log_2\big(\boldsymbol{Density}\ (\boldsymbol{infR})\big),\ 0\right) \tag{9}$$

$$\begin{aligned}&\boldsymbol{Interstitial\ and\ Tubular\ Abnormality\ Score\ (ITAS)}\\&\quad= \max\ (\boldsymbol{Abnormal\ Interstitial\ Area\ Percentage} \times\\&\qquad log_2(10\boldsymbol{Abnormal\ Tubules\ Density}),\ 0)\end{aligned} \tag{10}$$

$$\boldsymbol{Composite\ Damage\ Score\ (CDS)} = \boldsymbol{MLIS} + \boldsymbol{ITAS} \tag{11}$$

Some patients had multiple segments per slide or re-scanned slides at the time of biopsy. Digital features were first estimated within each segment and then weighted by relative size of segment $\frac{area\ of\ segment(i)}{total\ area\ of\ all\ segments}$ and summed across multiple segments (except counting of glomeruli which was simply summed across segments). Therefore, features extracted from large segments had more weight than those from small segments. If a patient has re-scanned slides, we perform similar weighted average (by relative size of slide) method to obtain features at patient level. It is worth noting at we observed similar feature outputs for few re-scanned slides, which suggested the consistency/reproducibility of machine production given the same slide.

In summary, our whole slide feature extraction pipeline generated three types of outputs for one WSI: i) whole slide prediction masks demonstrating kidney tissue compartments within a slide. ii) two ROI masks representing interstitium and inflammation abnormality. iii) a comprehensive data report summarizing individual or composite features.

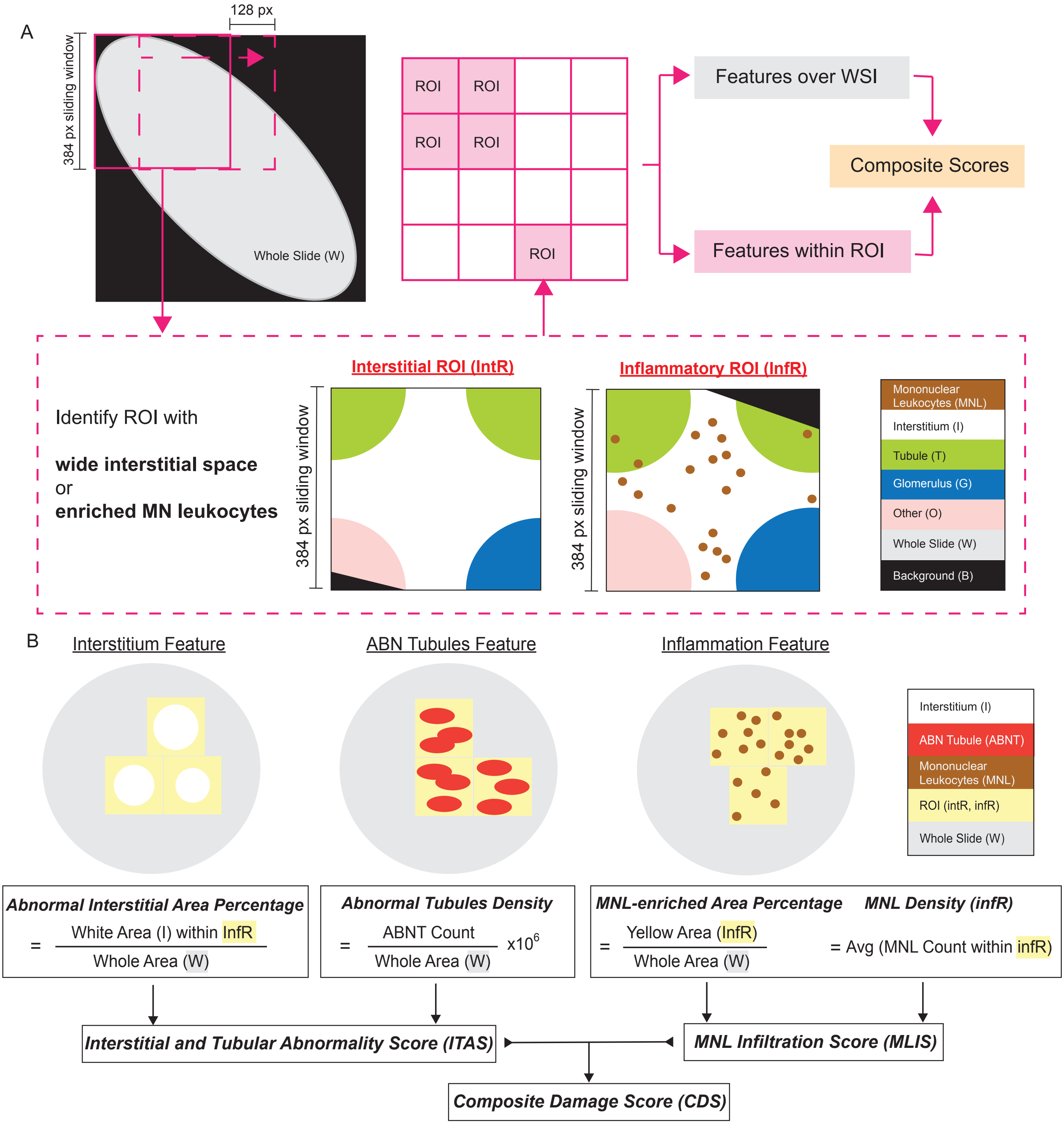


Figure S1. Illustration of slide-wide digital features extraction process and definition. A) Illustration of feature extraction process. We used a 384x384 pixel unit window scanning across WSI at stride of 128-pixel distance. Windows that had wide interstitial space or high amount of MN leukocyte were defined as interstitial regions of interest (intROI, intR) or inflammatory regions of interest (infROI, infR). A series of individual features were defined at ROI or slide level and further integrated into composite scores aiming for overall abnormality estimation. **B)** Illustration of definition (calculation) of individual digital features in interstitium, tubules and MNL infiltration.

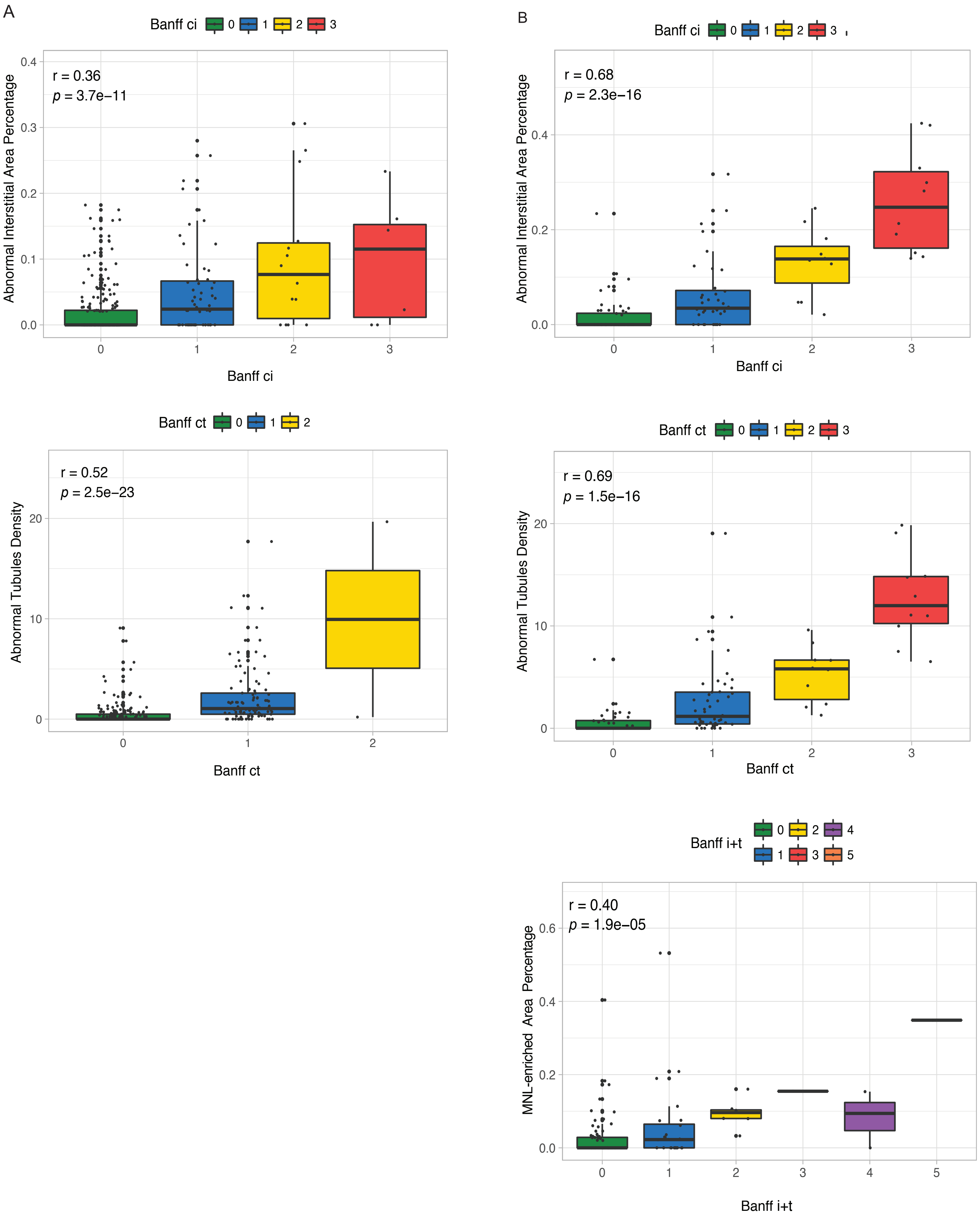


Figure S2. Correlation of digital features with corresponding Banff scores. A) Correlation of Abnormal Interstitial Area Percentage and Banff ci score (upper), Abnormal Tubules Density and Banff ct score (middle) in GoCAR baseline biopsy slides (n=317). **B)** Correlation of Abnormal Interstitial Area Percentage and Banff ci score (top), Abnormal Tubules Density and Banff ct score (middle), MNL-enriched Area Percentage and Banff i+t score (bottom) in AUSCAD 12m post-transplant biopsy slides (n=111). P-values were calculated from Spearman's correlation test.

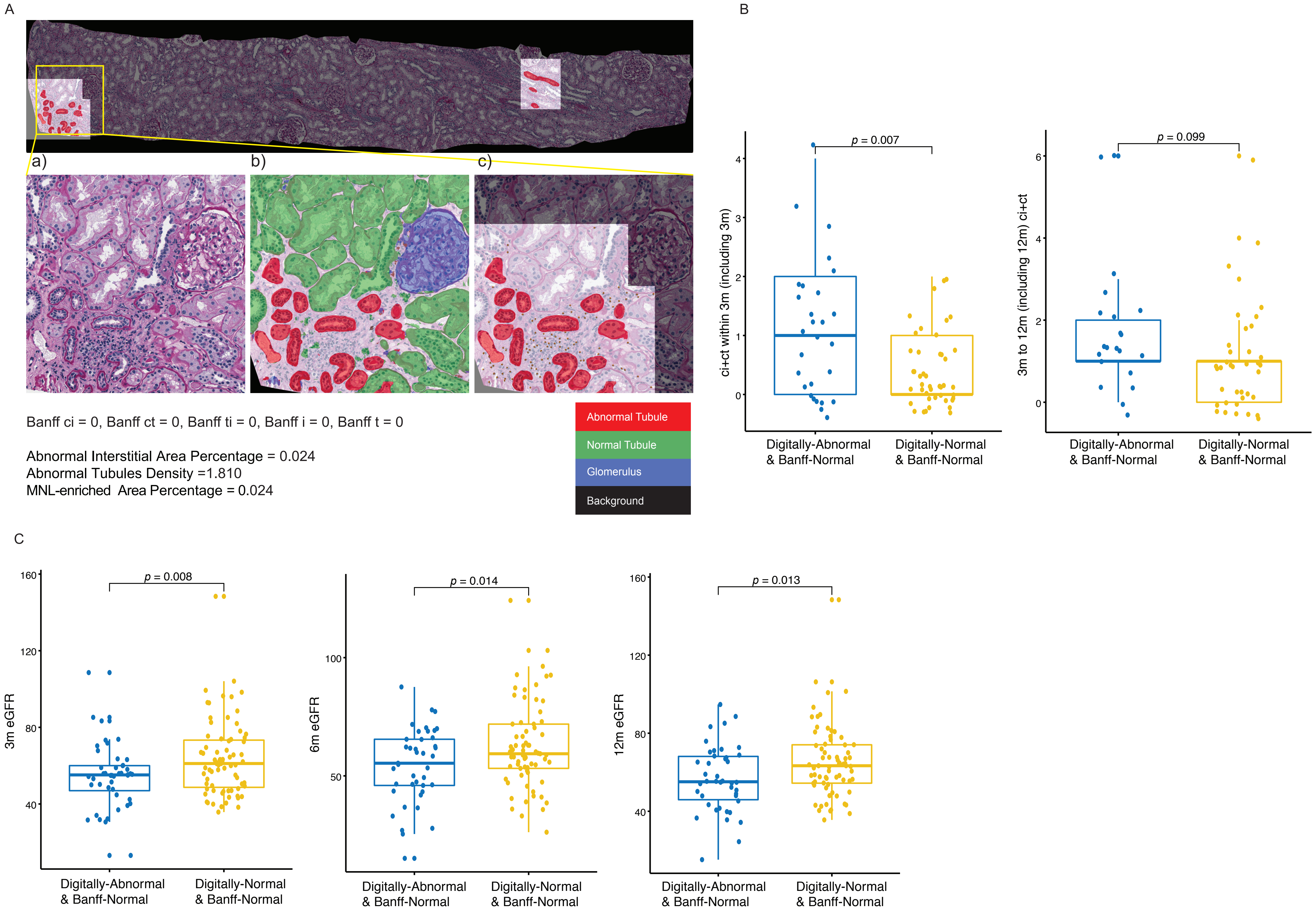


Figure S3. Discrepancy between digital features and Banff scores. A) An example of a WSI determined normal by all Banff scores but abnormal by digital features. Upper-panel shows whole slide image highlighting abnormal interstitium/tubules regions by digital features. Lower-panel shows close-up views of original(a), full compartment prediction(b) and abnormal ROI mask(c) from one abnormal region within yellow box. **B-C)** Comparison of subsequent pathological outcomes **(B)** ci+ct within 3m (including 3m) and ci+ct from 3m to 12m (including 12m) and clinical outcomes **(C)** 3m eGFR, 6m eGFR, 12m eGFR between digitally-abnormal vs. –normal patients who were all determined normal by Banff scores from GoCAR baseline biopsies. P-values are calculated by t-test.

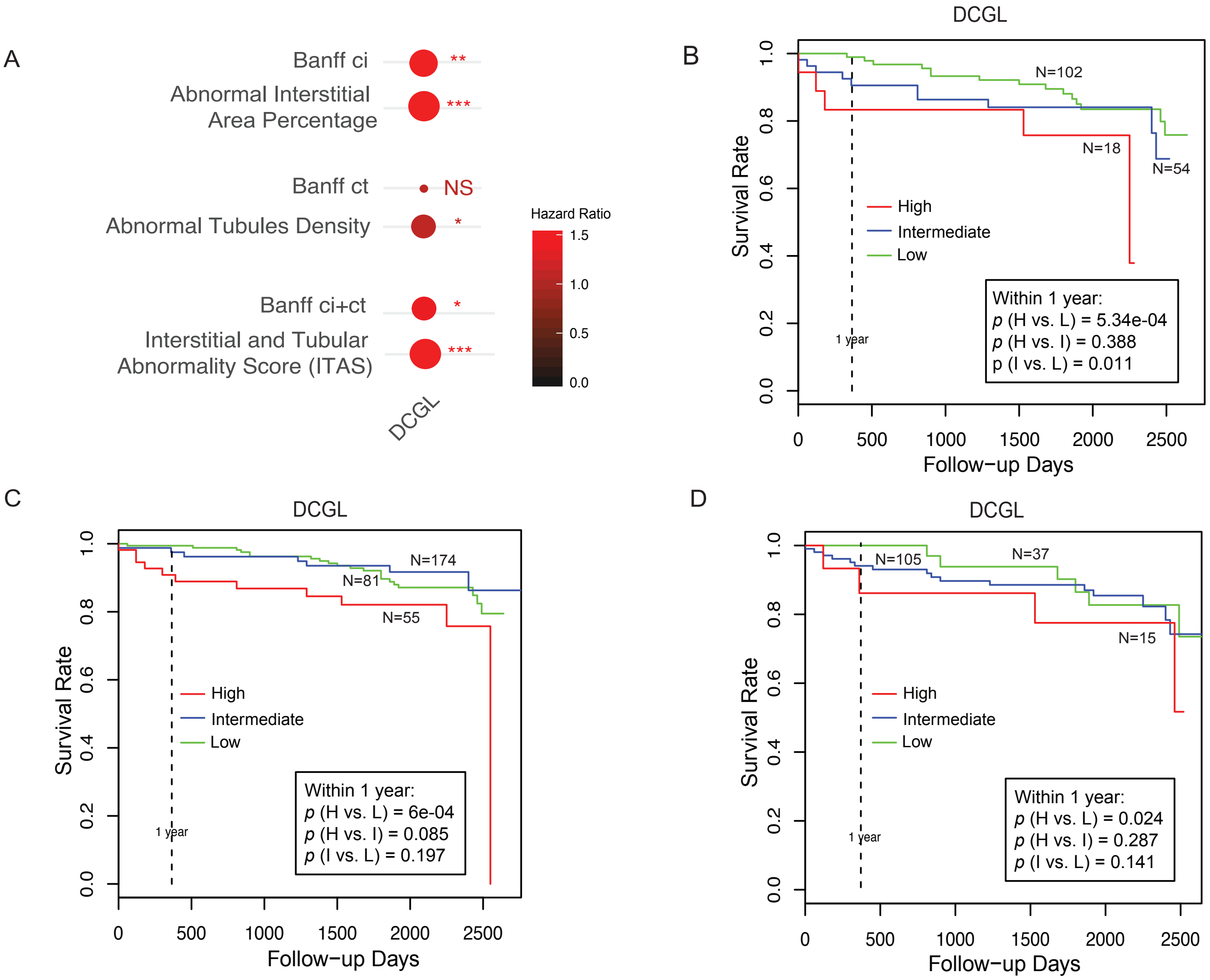


Figure S4. Association of baseline digital features with post-transplant graft outcomes in GoCAR cohort. A) Dot heatmap of association of Banff scores and digital features with post-transplant death-censored graft loss (DCGL) in baseline biopsy slides (n=317). The size of dots and number of asterisks indicate significance level (p-value) of association by Cox proportional hazards regression (NS: p ≥ 0.1; .: 0.05 ≤ p < 0.1; *:0.005 ≤ p < 0.05; **: 5e-04 ≤ p < 0.005; ***: 5e-05 ≤ p < 5e-04; ****:p < 5e-05). Color darkness of dots indicate hazard ratio. **B)** Kaplan-Meier curves of DCGL in ITAS high vs. intermediate vs. low group in deceased donor population in baseline biopsies (n=174). Baseline ITAS groups are defined as: high: ITAS>0.6, intermediate: 0.1≤ITAS≤0.6, low: ITAS<0.1. P-values are calculated by log-rank test. **C)** Kaplan-Meier curves of DCGL in ci+ct high vs. intermediate vs. low group in baseline biopsies. ci+ct groups are defined as: high: ci+ct>1, intermediate: ci+ct=1, low: ci+ct=0. P-values are calculated by log-rank test. **D)** Kaplan-Meier curves of DCGL in KDPI high vs. intermediate vs. low group in deceased donor population in baseline biopsies. KDPI groups are defined in deceased-donor population as: high: KDPI>85%, intermediate: 20%<KDPI≤85%, low: KDPI≤20%. P-values are calculated by log-rank test.

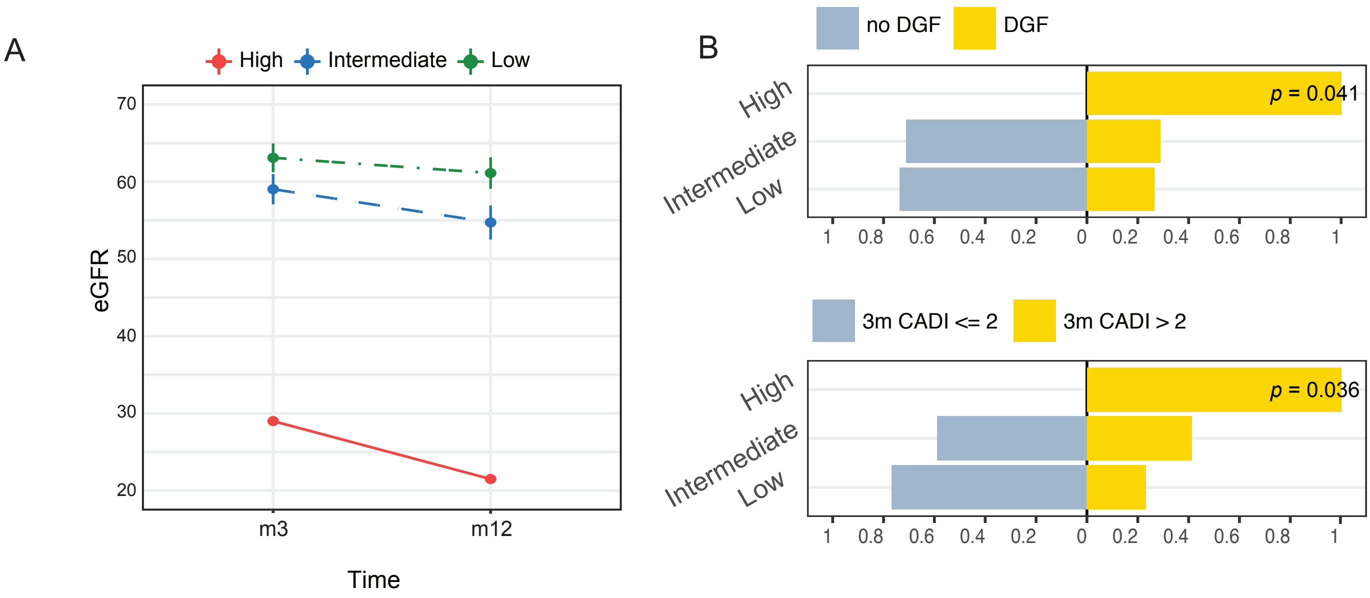


Figure S5. Association of baseline digital features with post-transplant graft outcomes in AUSCAD cohort. A) Average eGFR values over time within 12m post-transplant per baseline ITAS risk group. Error bars represent 0.1x standard deviation from mean values. **B)** Bar charts demonstrating proportions of DGF/no DGF (upper) and 3m post-transplant CADI >2/≤2 (lower) among three baseline ITAS risk groups in whole population. P-values are calculated by Fisher's exact test.

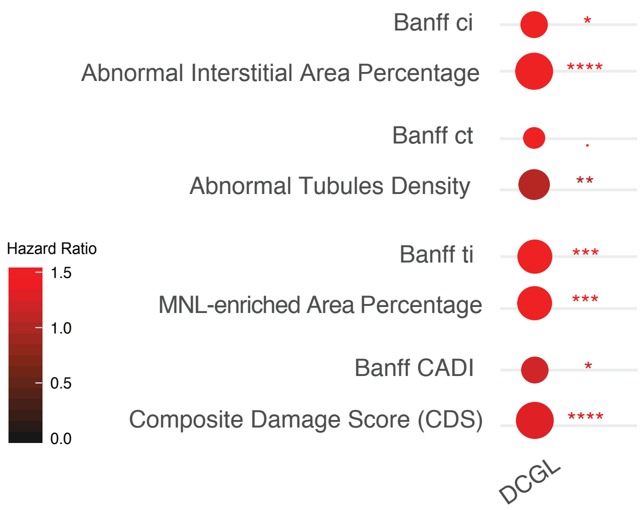
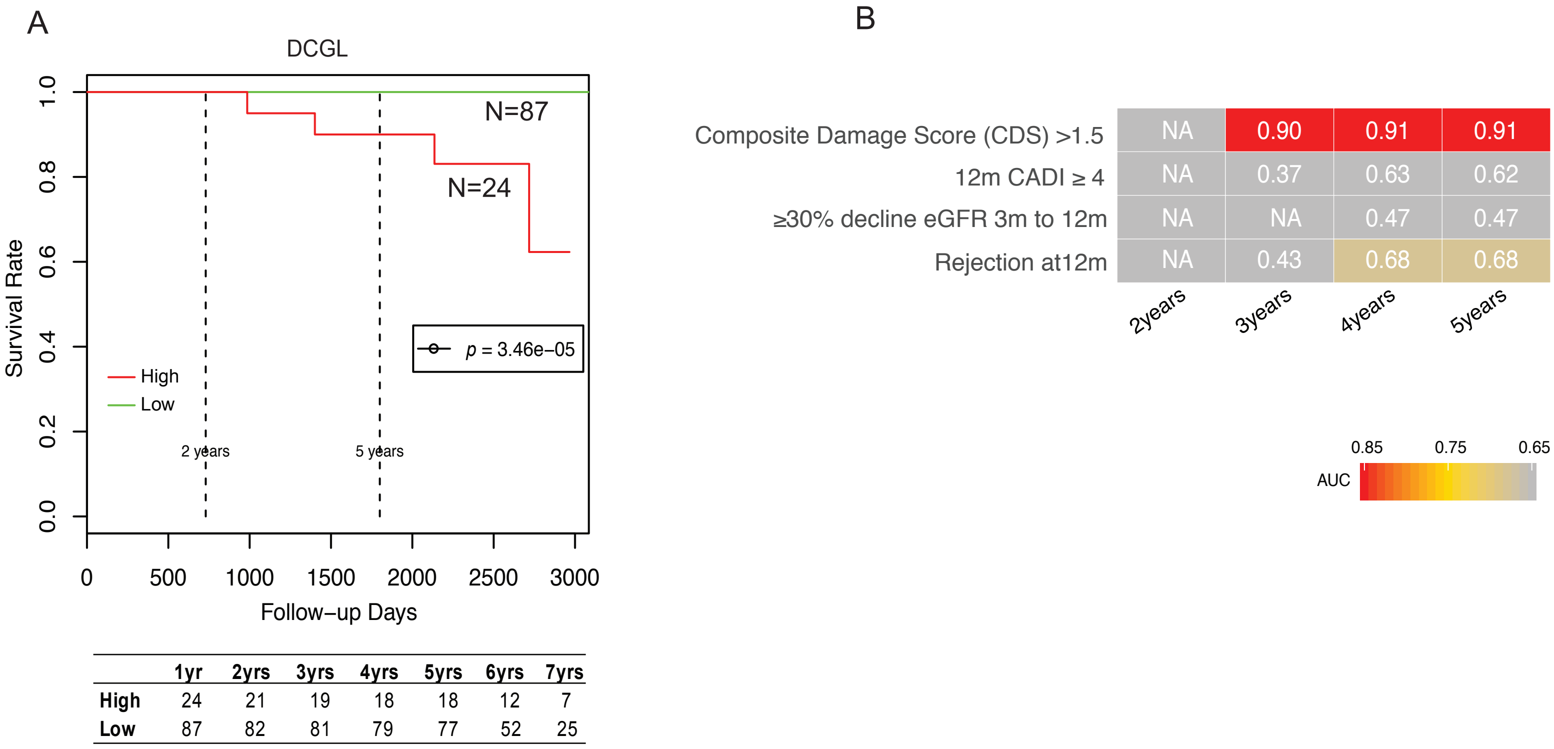


Figure S6. Association of 12m post-transplant digital features with post-transplant graft outcomes in GoCAR cohort. Dot heatmap of association of Banff scores and digital features with post-transplant death-censored graft loss (DCGL) in 12m post-transplant biopsy slides (n=200). The size of dot and number of asterisks indicate significance level (p-value) of association by Cox proportional hazards regression (NS: $p \geq 0.1$; .: $0.05 \leq p < 0.1$; *:$0.005 \leq p < 0.05$; **: $5e-04 \leq p < 0.005$; ***: $5e-05 \leq p < 5e-04$; ****:$p < 5e-05$). Color darkness of dots indicate hazard ratio.



	1yr	2yrs	3yrs	4yrs	5yrs	6yrs	7yrs
High	24	21	19	18	18	12	7
Low	87	82	81	79	77	52	25

	2years	3years	4years	5years
Composite Damage Score (CDS) >1.5	NA	0.90	0.91	0.91
12m CADI ≥ 4	NA	0.37	0.63	0.62
≥30% decline eGFR 3m to 12m	NA	NA	0.47	0.47
Rejection at12m	NA	0.43	0.68	0.68

Figure S7. Association of 12m post-transplant digital features with post-transplant graft outcomes in AUSCAD cohort. A) Kaplan-Meier curves of DCGL in CDS high vs. low group from AUSCAD 12m biopsies (n=111). P-values are calculated by log-rank test. **B)** Heatmap of time-dependent AUCs in predicting DCGL by 12m CDS high/low group and other pathological/clinical factors which were obtained prior to or at 12m. 12m CDS groups are defined as: high: CDS>1.5, low: CDS≤1.5.

Table S1. Accuracy summary of kidney tissue compartment prediction model based on independent testing set.

Group	Area Segmentation			Instance Detection		
	TPR	**PPV**	**F-score**	**TPR**	**PPV**	**F-score**
Interstitium	0.73	0.85	0.75	-	-	-
Glomeruli	0.94	0.87	0.93	0.96	0.97	0.96
All Tubule	0.93	0.84	0.91	0.91	0.85	0.90
Normal Tubule	0.92	0.79	0.89	0.81	0.77	0.80
Abnormal Tubule	0.79	0.78	0.79	0.84	0.76	0.82
Artery	0.84	0.96	0.86	0.75	0.89	0.77
MN Leukocyte	-	-	-	0.77	0.66	0.75
Epithelial cell	-	-	-	0.90	0.67	0.84

Table S2. Association of baseline Banff scores and digital features with graft loss in GoCAR cohort.

a) Association of baseline Banff scores and digital features with death-censored graft loss (DCGL)

Scores	DCGL PH assumption p-value	DCGL p-value	DCGL p-value FDR adjusted	DCGL hazard ratio	DCGL lower CI	DCGL upper CI
Banff ci	0.16	5.0e-04	1.0e-03	1.81	1.29	2.52
Abnormal Interstitial Area Percentage	0.65	5.7e-05	3.4e-04	1.08	1.04	1.13
Banff ct	0.18	8.1e-01	8.1e-01	-	-	-
Abnormal Tubules Density	0.65	1.2e-02	1.8e-02	1.11	1.02	1.20
Banff ci+ct	0.17	2.8e-02	3.4e-02	1.38	1.04	1.84
Interstitial and Tubular Abnormality Score (ITAS)	0.93	1.5e-04	4.5e-04	3.25	1.77	5.97

b) Association of baseline Banff scores and digital features with death-censored graft loss (DCGL) after adjusting for clinical confounders.

Scores	DCGL PH assumption p-value	DCGL p-value	DCGL p-value FDR adjusted	DCGL hazard ratio	DCGL lower CI	DCGL upper CI
Banff ci	0.14	3.2e-02	6.3e-02	1.54	1.04	2.27
Abnormal Interstitial Area Percentage	0.43	8.8e-03	3.7e-02	1.06	1.01	1.10
Banff ct	0.21	8.4e-01	8.4e-01	-	-	-
Abnormal Tubules Density	0.73	7.0e-02	1.1e-01	-	-	-
Banff ci+ct	0.20	2.3e-01	2.7e-01	-	-	-
Interstitial and Tubular Abnormality Score (ITAS)	0.69	1.2e-02	3.7e-02	2.28	1.20	4.35

* Cox p-values are calculated by Wald test from Cox proportional hazards regression. The proportional hazards assumptions are assessed through chi-square goodness of fit test between Schoenfeld residuals and time. Non-significant p-values confirm the assumption. Hazard ratios are not reported if PH assumptions are violated or cox p-values are not significant.

Table S3. Association of baseline digital features with graft loss in AUSCAD cohort.

a) Association of baseline digital features with death-censored graft loss (DCGL).

Scores	DCGL PH Assumption p-value	DCGL p-value	DCGL p-value FDR adjusted	DCGL hazard ratio	DCGL lower CI	DCGL upper CI
Abnormal Interstitial Area Percentage	0.63	3.1e-01	7.0e-01	-	-	-
Abnormal Tubules Density	0.24	7.6e-01	1.0e+00	-	-	-
Interstitial and Tubular Abnormality Score (ITAS)	0.57	3.5e-01	7.0e-01	-	-	-

b) Association of baseline digital features with death-censored graft loss (DCGL) after adjusting for clinical confounders.

Scores	DCGL PH Assumption p-value	DCGL p-value	DCGL p-value FDR adjusted	DCGL hazard ratio	DCGL lower CI	DCGL upper CI
Abnormal Interstitial Area Percentage	0.63	1.8e-01	4.4e-01	-	-	-
Abnormal Tubules Density	0.22	4.6e-01	6.2e-01	-	-	-
Interstitial and Tubular Abnormality Score (ITAS)	0.57	2.2e-01	4.4e-01	-	-	-

* Cox p-values are calculated by Wald test from Cox proportional hazards regression. The proportional hazards assumptions are assessed through chi-square goodness of fit test between Schoenfeld residuals and time. Non-significant p-values confirm the assumption. Hazard ratios are not reported if PH assumptions are violated or cox p-values are not significant.

Table S4. Association of 12m post-transplant Banff scores and digital features with graft loss in GoCAR cohort.

a) Association of 12m Banff scores and digital features with death-censored graft loss (DCGL).

Scores	DCGL PH Assumption p-value	DCGL p-value	DCGL p-value FDR adjusted	DCGL hazard ratio	DCGL lower CI	DCGL upper CI
Banff ci	0.37	1.6e-02	2.1e-02	1.74	1.11	2.73
Abnormal Interstitial Area Percentage	0.31	2.1e-05	8.5e-05	1.05	1.03	1.08
Banff ct	0.48	9.6e-02	9.6e-02	-	-	-
Abnormal Tubules Density	0.60	1.5e-03	2.4e-03	1.06	1.02	1.09
Banff ti	0.19	2.2e-04	4.4e-04	2.10	1.42	3.10
MNL-enriched Area Percentage	0.80	1.6e-04	4.3e-04	1.03	1.02	1.05
Banff CADI	0.81	2.2e-02	2.5e-02	1.22	1.03	1.45
Composite Damage Score (CDS)	0.90	1.8e-05	8.5e-05	1.30	1.15	1.47

b) Association of 12m Banff scores and digital features with death-censored graft loss (DCGL) after adjusting for clinical confounders.

Scores	DCGL PH Assumption p-value	DCGL p-value	DCGL p-value FDR adjusted	DCGL hazard ratio	DCGL lower CI	DCGL upper CI
Banff ci	0.17	7.7e-02	1.0e-01	-	-	-
Abnormal Interstitial Area Percentage	0.17	4.8e-04	3.8e-03	1.05	1.02	1.07
Banff ct	1.00	4.0e-01	4.0e-01	-	-	-
Abnormal Tubules Density	0.26	1.6e-02	2.5e-02	1.05	1.01	1.09
Banff ti	0.07	1.2e-02	2.3e-02	1.76	1.13	2.72
MNL-enriched Area Percentage	0.75	7.6e-03	2.0e-02	1.03	1.01	1.04
Banff CADI	0.60	2.5e-01	2.8e-01	-	-	-
Composite Damage Score (CDS)	0.64	1.5e-03	6.1e-03	1.25	1.09	1.43

* Cox p-values are calculated by Wald test from Cox proportional hazards regression. The proportional hazards assumptions are assessed through chi-square goodness of fit test between Schoenfeld residuals and time. Non-significant p-values confirm the assumption. Hazard ratios are not reported if PH assumptions are violated or cox p-values are not significant.

Table S5. Association of 12m post-transplant Banff scores and digital features with graft loss in AUSCAD cohort.

a) Association of 12m Banff scores and digital features with death-censored graft loss (DCGL).

Scores	DCGL PH Assumption p-value	DCGL p-value	DCGL p-value FDR adjusted	DCGL hazard ratio	DCGL lower CI	DCGL upper CI
Banff ci	0.04	-	-	-	-	-
Abnormal Interstitial Area Percentage	0.46	4.9e-03	1.2e-02	1.08	1.02	1.14
Banff ct	0.03	-	-	-	-	-
Abnormal Tubules Density	0.58	4.2e-02	5.6e-02	1.10	1.00	1.20
Banff i+t	0.79	2.2e-03	1.2e-02	6.32	1.94	20.60
MNL-enriched Area Percentage	0.78	8.8e-03	1.4e-02	1.08	1.02	1.14
Banff CADI	0.03	-	-	-	-	-
Composite Damage Score (CDS)	0.75	6.1e-03	1.2e-02	1.71	1.16	2.50

b) Association of 12m Banff scores and digital features with death-censored graft loss (DCGL) after adjusting for clinical confounders.

Scores	DCGL PH Assumption p-value	DCGL p-value	DCGL p-value FDR adjusted	DCGL hazard ratio	DCGL lower CI	DCGL upper CI
Banff ci	0.03	-	-	-	-	-
Abnormal Interstitial Area Percentage	0.36	1.1e-02	4.2e-02	1.08	1.02	1.14
Banff ct	0.03	-	-	-	-	-
Abnormal Tubules Density	0.36	7.5e-02	8.2e-02	-	-	-
Banff i+t	0.74	3.0e-02	5.1e-02	14.67	1.29	166.94
MNL-enriched Area Percentage	0.78	3.2e-02	5.1e-02	1.07	1.01	1.15
Banff CADI	0.04	-	-	-	-	-
Composite Damage Score (CDS)	0.71	2.2e-02	5.1e-02	1.69	1.08	2.63

* Cox p-values are calculated by Wald test from Cox proportional hazards regression. The proportional hazards assumptions are assessed through chi-square goodness of fit test between Schoenfeld residuals and time. Non-significant p-values confirm the assumption. Hazard ratios are not reported if PH assumptions are violated or cox p-values are not significant.

Supplementary Reference

S1. O'Connell, P.J., et al., *Biopsy transcriptome expression profiling to identify kidney transplants at risk of chronic injury: a multicentre, prospective study.* Lancet, 2016. **388**(10048): p. 983-93.

S2. Shorten, C. and T.M. Khoshgoftaar, *A survey on image data augmentation for deep learning.* Journal of Big Data, 2019. **6**(1): p. 1-48.

S3. He, K., et al. *Mask R-CNN*. 2017. arXiv:1703.06870.

S4. Ronneberger, O., P. Fischer, and T. Brox *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv:1505.04597.

S5. Abdulla, W. *Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow*. GitHub repository 2017; Available from: https://github.com/matterport/Mask_RCNN.

S6. Akeret, J., et al., *Radio frequency interference mitigation using deep convolutional neural networks.* Astronomy and Computing, 2017. **18**: p. 35.

S7. Lin, T.-Y., et al. *Microsoft coco: Common objects in context*. in *European conference on computer vision*. 2014. Springer.

S8. Van Rijsbergen, C.J., *Information Retrieval (2nd ed.).* . 1979: Butterworth-Heinemann.