Confidence Calibration for Deep Renal Biopsy Immunofluorescence Image Classification

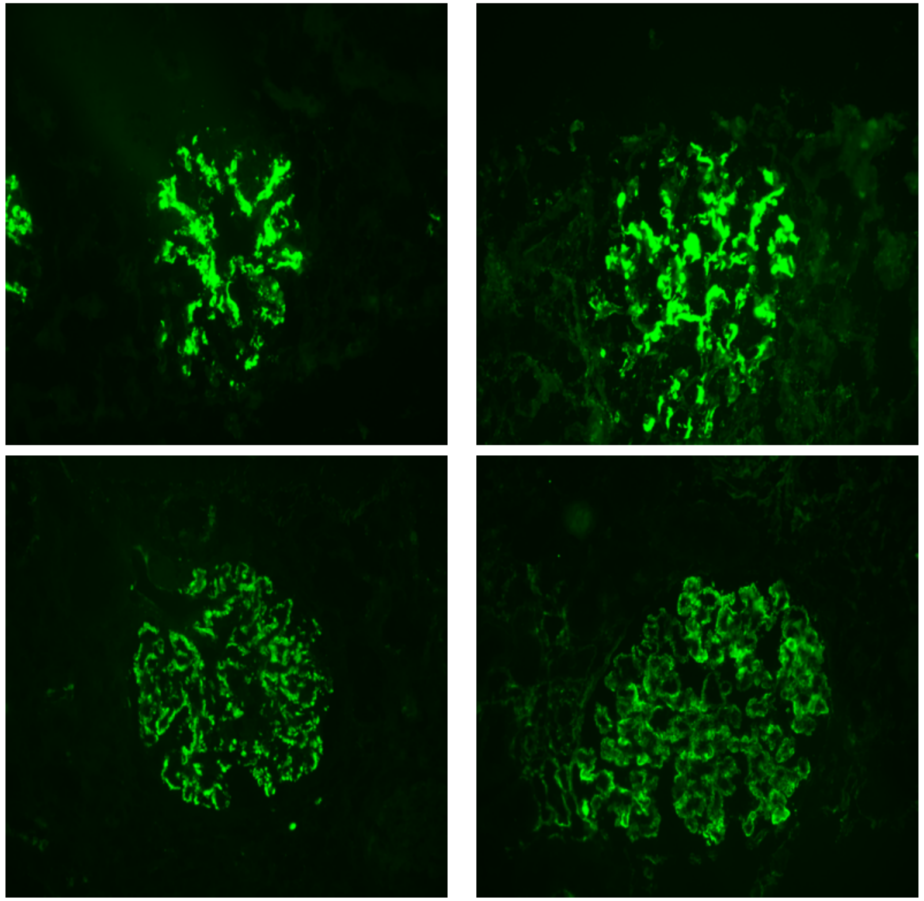
Federico Pollastri[1], Juan Maroñas[2], Federico Bolelli[1], Giulia Ligabue[1], Roberto Paredes[2], Riccardo Magistroni[1], and Costantino Grana[1]

[1]*Università degli Studi di Modena e Reggio Emilia, Italy*
[2]*Universitat Politècnica de València, Spain*

Immunofluorescence in Renal Biopsy

- **Immunofluorescence** is a powerful technique for light microscopy that makes use of fluorescent-labeled antibodies
- It can be used for renal diseases diagnosis
- Pattern of antibody deposits require strong expertise to be analyzed
- This work focuses on using Convolutional Neural Networks (CNNs) for the automatic identification of two deposit patterns:

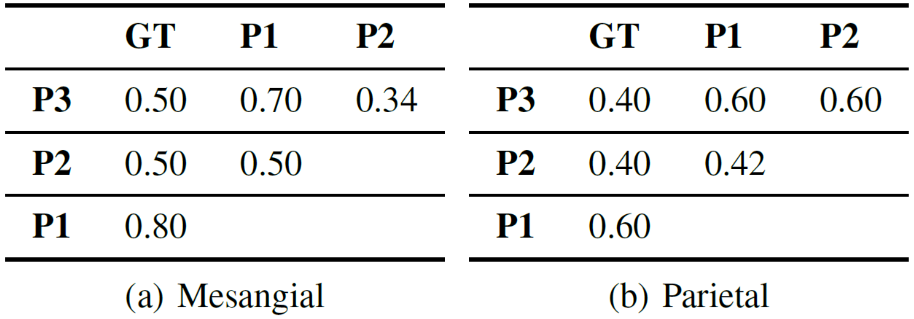
  I. Mesangial - *top row*
  II. Parietal - *bottom row*



Deep Learning in Medical Imaging

- Convolutional Neural Networks have been widely employed in several Medical Imaging tasks such as image classification, detection, segmentation, and others
- Neural Networks are often seen as **black boxes**: this does not suit our task
- Binary predictions are an extremely underwhelming tool for immunofluorescence image analysis
- How can we improve CNNs interpretability?

Inter-rater agreement (Cohen's Kappa) between expert practitioners is very low



	GT	**P1**	**P2**
P3	0.50	0.70	0.34
P2	0.50	0.50	
P1	0.80		

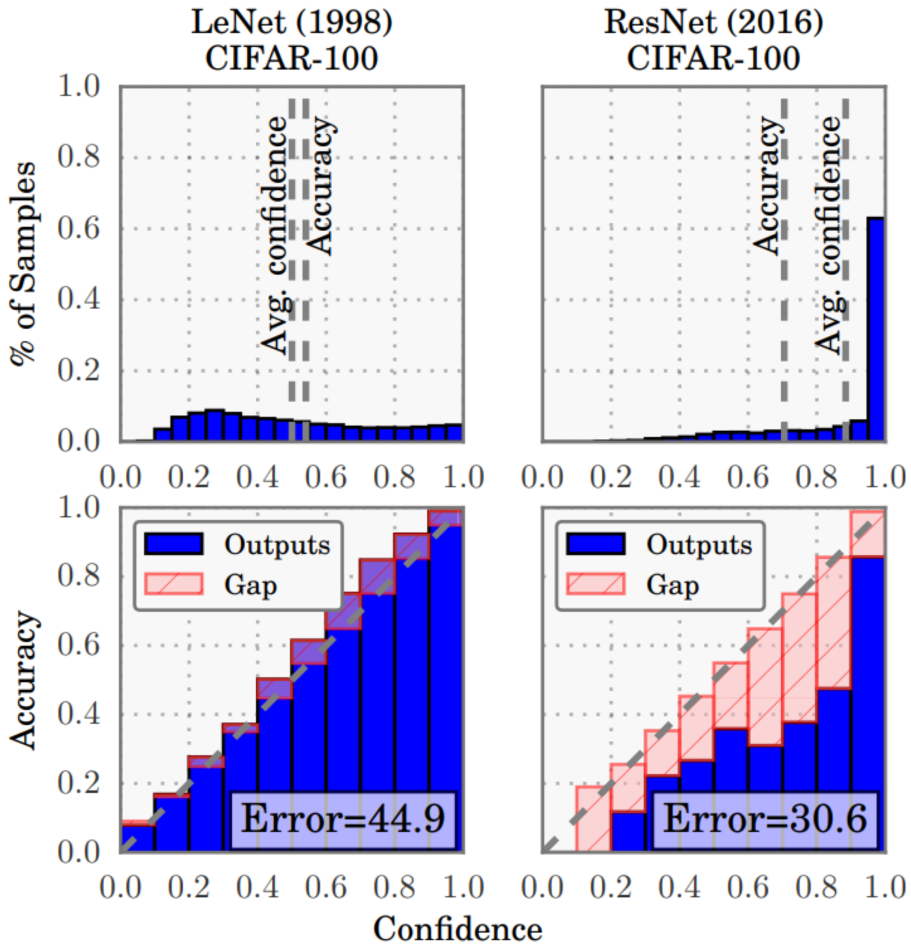
(a) Mesangial

	GT	**P1**	**P2**
P3	0.40	0.60	0.60
P2	0.40	0.42	
P1	0.60		

(b) Parietal

Proposed Method

- **Dataset:**
  - 11k images
  - 3k exhibit parietal pattern
  - 2k exhibit mesangial pattern
  - 1k exhibit both patterns
- **Task** – identification of two mutually non-exclusive patterns (mesangial and parietal)
- **CNNs** – one residual blocks neural network for task
- **Reliable Outputs** – model recalibration
  - Calibrated probabilities – low Expected Calibration Error (ECE) [1]
  - Good discriminative power – high accuracy



[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, *“On Calibration of Modern Neural Networks,” in Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017, pp. 1321–1330.*

Quantitative Results 1/2

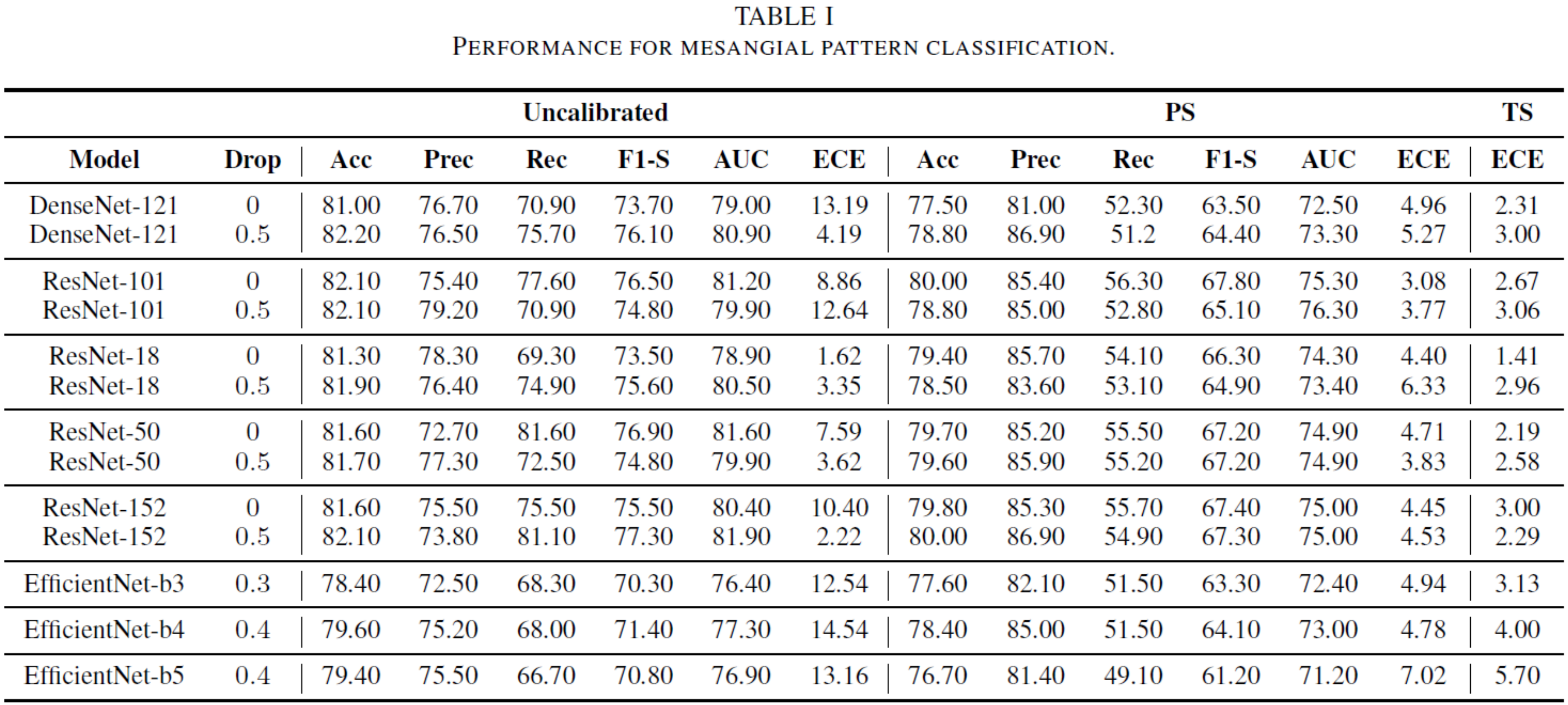


TABLE I

Performance for mesangial pattern classification.

		Uncalibrated						PS						TS
Model	**Drop**	**Acc**	**Prec**	**Rec**	**F1-S**	**AUC**	**ECE**	**Acc**	**Prec**	**Rec**	**F1-S**	**AUC**	**ECE**	**ECE**
DenseNet-121	0	81.00	76.70	70.90	73.70	79.00	13.19	77.50	81.00	52.30	63.50	72.50	4.96	2.31
DenseNet-121	0.5	82.20	76.50	75.70	76.10	80.90	4.19	78.80	86.90	51.2	64.40	73.30	5.27	3.00
ResNet-101	0	82.10	75.40	77.60	76.50	81.20	8.86	80.00	85.40	56.30	67.80	75.30	3.08	2.67
ResNet-101	0.5	82.10	79.20	70.90	74.80	79.90	12.64	78.80	85.00	52.80	65.10	76.30	3.77	3.06
ResNet-18	0	81.30	78.30	69.30	73.50	78.90	1.62	79.40	85.70	54.10	66.30	74.30	4.40	1.41
ResNet-18	0.5	81.90	76.40	74.90	75.60	80.50	3.35	78.50	83.60	53.10	64.90	73.40	6.33	2.96
ResNet-50	0	81.60	72.70	81.60	76.90	81.60	7.59	79.70	85.20	55.50	67.20	74.90	4.71	2.19
ResNet-50	0.5	81.70	77.30	72.50	74.80	79.90	3.62	79.60	85.90	55.20	67.20	74.90	3.83	2.58
ResNet-152	0	81.60	75.50	75.50	75.50	80.40	10.40	79.80	85.30	55.70	67.40	75.00	4.45	3.00
ResNet-152	0.5	82.10	73.80	81.10	77.30	81.90	2.22	80.00	86.90	54.90	67.30	75.00	4.53	2.29
EfficientNet-b3	0.3	78.40	72.50	68.30	70.30	76.40	12.54	77.60	82.10	51.50	63.30	72.40	4.94	3.13
EfficientNet-b4	0.4	79.60	75.20	68.00	71.40	77.30	14.54	78.40	85.00	51.50	64.10	73.00	4.78	4.00
EfficientNet-b5	0.4	79.40	75.50	66.70	70.80	76.90	13.16	76.70	81.40	49.10	61.20	71.20	7.02	5.70

AImage Lab

Quantitative Results 2/2

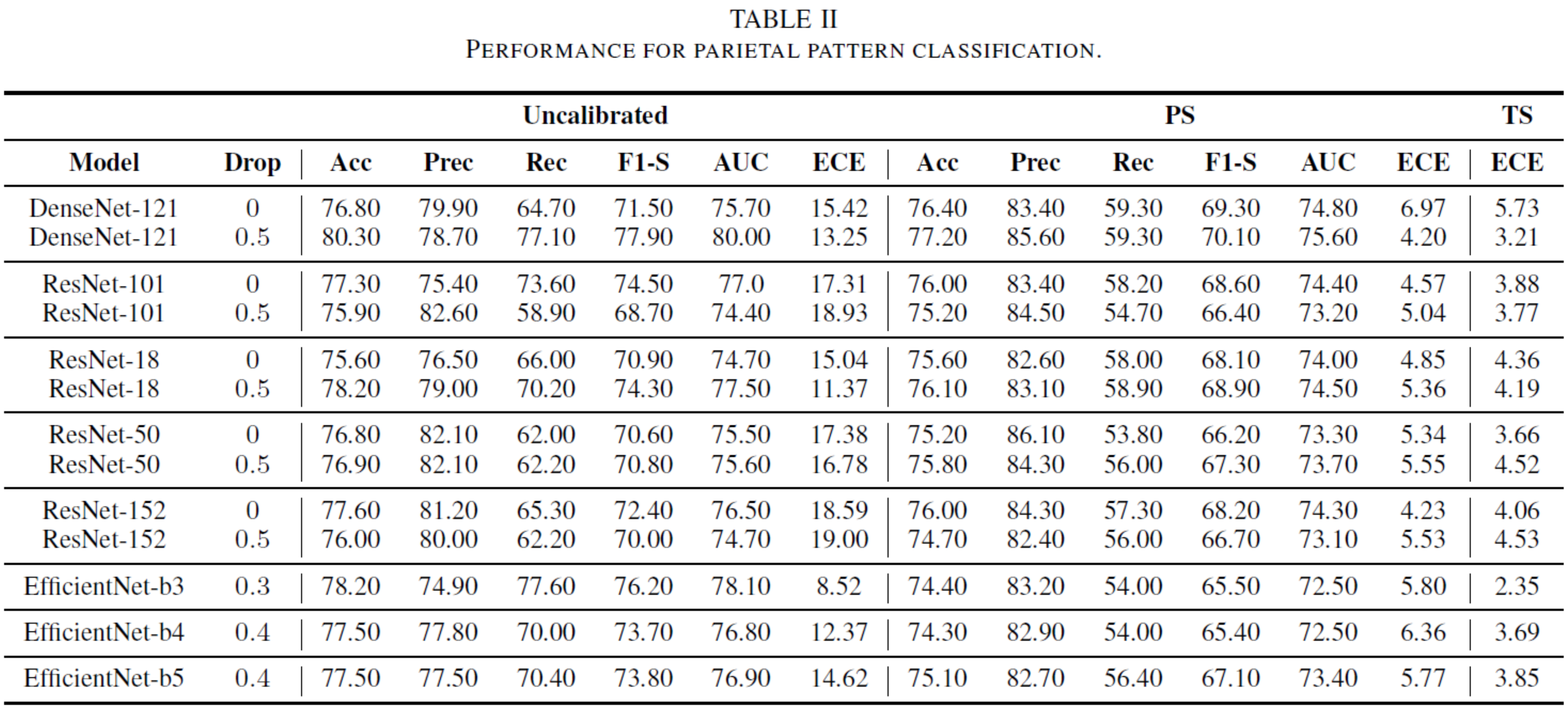


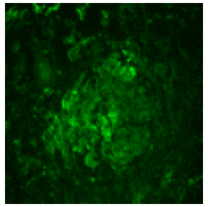
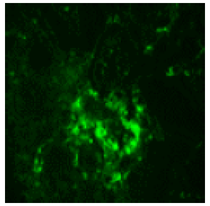
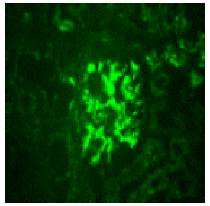
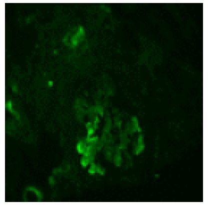
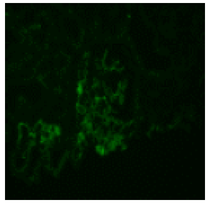
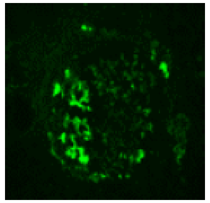
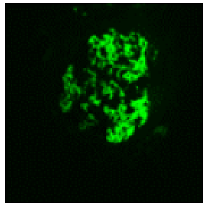
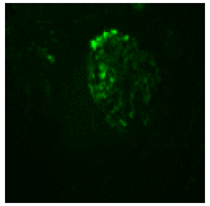
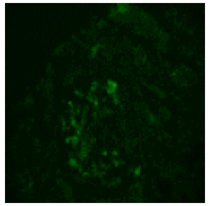
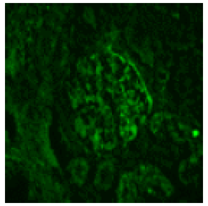
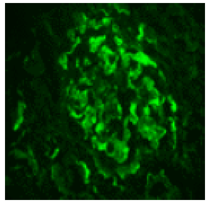
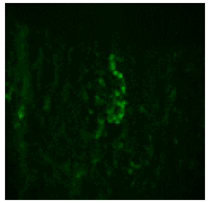
TABLE II
PERFORMANCE FOR PARIETAL PATTERN CLASSIFICATION.

		Uncalibrated						PS						TS
Model	**Drop**	**Acc**	**Prec**	**Rec**	**F1-S**	**AUC**	**ECE**	**Acc**	**Prec**	**Rec**	**F1-S**	**AUC**	**ECE**	**ECE**
DenseNet-121	0	76.80	79.90	64.70	71.50	75.70	15.42	76.40	83.40	59.30	69.30	74.80	6.97	5.73
DenseNet-121	0.5	80.30	78.70	77.10	77.90	80.00	13.25	77.20	85.60	59.30	70.10	75.60	4.20	3.21
ResNet-101	0	77.30	75.40	73.60	74.50	77.0	17.31	76.00	83.40	58.20	68.60	74.40	4.57	3.88
ResNet-101	0.5	75.90	82.60	58.90	68.70	74.40	18.93	75.20	84.50	54.70	66.40	73.20	5.04	3.77
ResNet-18	0	75.60	76.50	66.00	70.90	74.70	15.04	75.60	82.60	58.00	68.10	74.00	4.85	4.36
ResNet-18	0.5	78.20	79.00	70.20	74.30	77.50	11.37	76.10	83.10	58.90	68.90	74.50	5.36	4.19
ResNet-50	0	76.80	82.10	62.00	70.60	75.50	17.38	75.20	86.10	53.80	66.20	73.30	5.34	3.66
ResNet-50	0.5	76.90	82.10	62.20	70.80	75.60	16.78	75.80	84.30	56.00	67.30	73.70	5.55	4.52
ResNet-152	0	77.60	81.20	65.30	72.40	76.50	18.59	76.00	84.30	57.30	68.20	74.30	4.23	4.06
ResNet-152	0.5	76.00	80.00	62.20	70.00	74.70	19.00	74.70	82.40	56.00	66.70	73.10	5.53	4.53
EfficientNet-b3	0.3	78.20	74.90	77.60	76.20	78.10	8.52	74.40	83.20	54.00	65.50	72.50	5.80	2.35
EfficientNet-b4	0.4	77.50	77.80	70.00	73.70	76.80	12.37	74.30	82.90	54.00	65.40	72.50	6.36	3.69
EfficientNet-b5	0.4	77.50	77.50	70.40	73.80	76.90	14.62	75.10	82.70	56.40	67.10	73.40	5.77	3.85

AImage Lab

Qualitative Results

- Expert practitioners provided likelihood scores of the mesangial patterns

GT: yes Pred: no	*Calib* Uncalib	GT: no Pred: yes	*Calib* Uncalib	GT: yes Pred: yes	*Calib* Uncalib *Human*
	*0.830* 0.992		*0.781* 0.980		*0.965* 0.999 *1.000*
	*0.771* 0.977		*0.774* 0.964		*0.771* 0.977 *0.400*
	*0.571* 0.707		*0.572* 0.711		*0.658* 0.883 *0.600*
	*0.562* 0.684		*0.560* 0.679		*0.558* 0.673 *0.400*

- Mitigating CNNs overconfidence is undoubtedly helpful for misclassified samples

- Calibrated probabilities are closer to human-assigned likelihood scores *w.r.t.* uncalibrated outputs

- Re-calibrating the CNNs output reduced the Mean Absolute Error (MAE) by 5%

AImage Lab