腎生検病理画像からの糸球体自動検出における 2 施設間の精度比較

An Interhospital Comparison of Glomerular Detection from Whole Slide Images

嶋本 公徳[1※]　中村 一成[2]　山口 亮平[3]　堂本 裕加子[4]
河添 悦昌[1] 塚本 達雄[5] 大江 和彦[6]

Kiminori Shimamoto[1], Issei Nakamura[2], Ryohei Yamaguchi[3], Yukako Domoto[4],
Yoshimasa Kawazoe[1], Tatsuo Tsukamoto[5], and Kazuhiko Ohe[6]

[1] 東京大学大学院 医学系研究科 医療 AI 開発学講座
[1]Artifical Intelligence in Healthcare, Graduate School of Medicine, The University of Tokyo
[2] 株式会社 NTT ドコモサービスイノベーション部
[2] Service Innovation Department, NTT DOCOMO, INC.
[3] 大島記念嬉泉病院
[3] Ohshima Memorial Kisen Hospital
[4] 日本医科大学付属病院 病理診断科
[4] Department of Diagnostic Pathology, Nippon Medical School Hospital
[5] 公益財団法人 田附興風会 医学研究所 北野病院
[5] Department of Nephrology and Dialysis, Tazuke Kofukai Medical Research Institute, Kitano Hospital
[6] 東京大学大学院 医学系研究科 医療情報学分野
[6] Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo

Abstract: Transfer learning is used in where a model trained on one facility is re-used on another facility. If we know how much data is needed to transfer learning and what factors affect the performance on another facility, we will be able to do transfer learning more efficiently. Here we report a case study of transfer learning approach for two Glomerular Detection Models trained on each facility, to show the number of training data required and a factor affecting performance in this case.

背景と目的

腎生検とは、腎臓の組織を採取し、その病理スライドを作成し、顕微鏡でその病態を確認する検査である。腎生検診断において、体内の毒素を排出するのに中心的な役割を果たす糸球体と呼ばれる部分は特に重要な診断対象組織である。その糸球体組織は直径が 200 ㎛程度の球状の組織であり、その直径に比べて広大な病理スライド全体(Whole Slide Image: WSI)の中に 10 個～170 個程存在している。糸球体の診断は、スライドに存在する糸球体を検出し、その状態を診断する負荷の高い作業となる。この作業負

※連絡先：東京大学大学院 医学系研究科
医療 AI 開発学講座
〒113-8655 東京都文京区本郷 7-3-1
中央診療棟 2　22 世紀医療センター8 階
E-mail: kshimamoto@m.u-tokyo.ac.jp

荷を低減するために、機械学習を用いた糸球体検出支援が試みられている[1][2][3]。

一方、機械学習によって作成されるシステム（学習モデル）の振る舞いは学習データに依存することが知られる。病理画像などの医療画像データは、たとえ同一染色方法名の画像であっても、染色方法やデジタル化処理の細部に差異があり、実施施設ごとに画像データの傾向に違いを生じることがある。このため、単一施設データのみを学習した学習モデルを他施設データに適用した場合に、精度を単純に外挿することが出来ない。医療データでは特に、複数施設のデータ利用が容易ではなく、研究開発時に単一施設データしか利用できないこと、加えて、学習モデルを公開する場合に、学習データを公開出来ないことがある。そのような場合に、他施設の学習モデルを自施設に適用する際は、自施設データを追加学習（転移学習）し、自施設データでの精度を確保することになるが、その際に、転移学習データの必要数と、転移時の性能に影響を与える因子が既知であれば、転移の効率化が期待される。そこで今回の研究では、二つの施設の WSI を施設ごとに学習した学習モデルを使い、転移時の性能変化と必要な学習データ数を例示するとともに、転移時の性能に影響を与える因子を例示する。

データ

東京大学医学部附属病院（以下、施設 T）で 2010 年から 2017 年に検体採取された腎生検検体の PAS 染色 WSI 324 枚と、公益財団法人田附興風会医学研究所北野病院（以下、施設 K）で 2005 年から 2017 年に検体採取された腎生検検体の PAS 染色 WSI 354 枚を利用した。これらの WSI から、含まれる糸球体数が多い WSI を施設毎に 300 枚抽出して実験に使用した。すべての WSI には人手により糸球体の領域が四角形でアノテーションされている。これらのデータの取り扱いは、東京大学医学系研究科倫理委員会で承認されたプロトコルに従った(承認番号：11455)。後述のすべての方法は、関連する倫理ガイドラインおよび規制に準拠して実施した。

以下の実験では、各施設 300 枚の WSI を、各 50 枚のデータセットに 6 分割し、50 枚を学習停止判定と確信度の閾値設定に用い、50 枚を性能評価に用い、残りを学習に用いる 6 回の交差検証を行い、F1 スコアによってその性能を評価する。6 分割したデータセットに 1 から 6 までの番号を付与し、各試行の学習用データ（training データ）と、学習停止・閾値設定データ（validation データ）と性能評価データ（test データ）の組み合わせを表 1. に示し、各データセットに含まれる WSI の枚数と糸球体の数を表 2 と表 3 に示す。

表 1　交差検証毎のデータセットの組合わせ

テスト番号	性能評価用 (test)	学習停止・しきい値設定用 (validation)	学習用 (training)
test 1	1	2	3, 4, 5, 6
test 2	2	3	4, 5, 6, 1
test 3	3	4	5, 6, 1, 2
test 4	4	5	6, 1, 2, 3
test 5	5	6	1, 2, 3, 4
test 6	6	1	2, 3, 4, 5

表 2 施設 T のデータセットごとの WSI 数と糸球体数

データセット番号	WSI 数	糸球体数(カッコ内は 1 つの WSI に含まれる糸球体数の最小数と最大数)
1	50	2342(min:17, max:166)
2	50	2342(min:17, max:152)
3	50	2341(min:16, max:150)
4	50	2341(min:16, max:143)
5	50	2341(min:16, max:141)
6	50	2340(min:15, max:140)

表 3　施設 K のデータセットごとの WSI 数と糸球体数

データセット番号	WSI 数	糸球体数(カッコ内は 1 つの WSI に含まれる糸球体数の最小数と最大数)
1	50	1516(min:12, max:153)
2	50	1516(min:11, max:104)
3	50	1516(min:11, max:97)
4	50	1515(min:11, max:96)
5	50	1515(min:11, max:84)
6	50	1515(min:11, max:82)

施設 T に比して、施設 K では、WSI あたりの糸球体数が少ない(約 65%)。

表 4 に WSI の諸元を示す。

表 4 WSI の諸元

	スライドスキャナ	幅 (pixel)	高さ (pixel)	解像度 (㎛/pixel)
施設 T	Hamamatsu NanoZoomer C9600-12	平均: 134,472 min: 36,864 max: 204,800	平均: 32,515 min: 7,680 max: 73,472	x: 0.2277 y: 0.2275
施設 K	Hamamatsu NanoZoomer C9600-12	平均: 75,571 min: 28,672 max: 184.320	平均: 25,087 min: 11,52 max: 54,784	x: 0.2264 y: 0.2263

施設 K の WSI は、平均幅と高さが、施設 T の WSI の 56.2%,と 77.2%と小さく、解像度はわずかに高い。

表 5 に各色階調の平均値を示す。

表 5 色階調平均

	Blue (256 階調)	Green (256 階調)	Red (256 階調)
施設 T	平均:221.08 SD: 9.28 分散:86.41	平均:219.36 SD: 17.21 分散:297.17	平均: 223.88 SD: 12.10 分散:146.90
施設 K	平均:222.86 SD: 10.01 分散:100.54	平均:220.48 SD: 21.77 分散:475.52	平均:227.45 SD: 12.77 分散:163.62
Welch t 値	-2.25	-0.75	-3.46
p 値	0.025	0.45	5.79e-4

赤色と青色で平均に有意の差があり、特に、施設 K の WSI は赤の色調が強いが、その差は 3.57 であり SD の 29%程である。

方法

WSI からの糸球体領域の検出は以下の手順で行った。

(1) 糸球体検出処理：
WSI は最大解像度（対物レンズ 40 倍相当）において 10 万×3 万(pixel)ほどの画素から構成される。このような巨大画像全体を一度に処理することは計算リソースの制限から難しいため、WSI を対物レンズ 5 倍相当の 1/8 の解像度に縮小し、その画像から 2000 ㎛平方のウィンドウ（「窓」）を一定幅で重複させて切り出し、「窓」単位に糸球体を検出した。糸球体の検出には Faster R-CNN 法[4]を用いた。

(2) 重複集約処理:
(1)の処理では、処理単位の「窓」の一部が別の「窓」と重複しているため、重複して検出される糸球体があり、これを集約する。今回の実験では Intersection over Union(IoU)値が 0.35 以上の重複がある糸球体を一つに集約した。

処理単位の「窓」の大きさ 2000 ㎛は、糸球体の直径の概数 200 ㎛の 10 倍の大きさとして設定した。この「窓」の画素数は、対物レンズ 5 倍相当の解像度において 1098×1099(pixel)(施設 T 平均)、1105×1105(pixel)（施設 K 平均）になるが、計算リソースの制限から、それらを 1024×1024(pixel)の画像に変換して学習と検出処理を行った。

学習では、一つの糸球体の中心から上下左右に 1000 ㎛の画像を切り出し、その画像内に含まれる糸球体領域情報を付与し、さらに、学習データを拡張するために、それぞれ 0.5 の確率で上下反転、左右反転した学習用画像を作成し、この画像から含まれる糸球体全てを検出する学習を行った。一つの学習用画像に含まれる糸球体の平均を表 6 に示し、そのヒストグラムを図 1 に示す。

表 6 一つの学習画像に含まれる糸球体数

	一つの学習用画像に含まれる糸球体数
施設 T	平均: 6.21(min: 1, max: 22) SD: 3.20
施設 K	平均:5.84(min: 1, max: 23) SD: 3.08

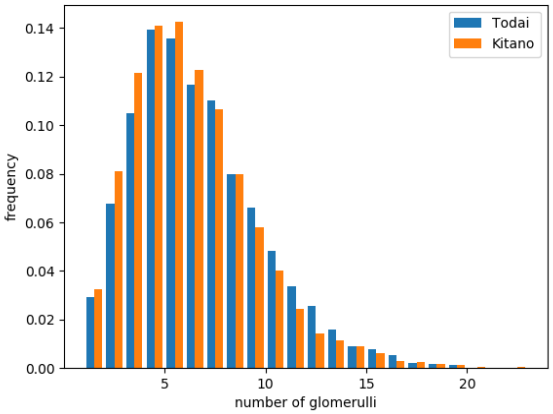

図 1 学習画像に含まれる糸球体数のヒストグラム

一つの学習画像には、最大 23 個、平均約 6 個の糸球体が含まれる。図 1 に示されるように、施設 K の学習データには、一つの学習用データに含まれる糸球体数が施設 T に比して少ないものが多いが、その分布形状は類似している。

一回の学習ステップのミニバッチサイズを 16 とし、5,000 ステップ毎にモデルを保存し、validation データを用いて F1 スコアが最良になる学習回数と確信度のしきい値を求めた。5,000 ステップは 200 枚の WSI を学習する場合に施設 T と施設 K のデータで、それぞれ約 8 epoch と 13 epoch に相当する。

転移学習では、表 1 の学習用データセットを左から順に学習データに追加し、転移学習に用いるデータ数と性能の関係を調べた。つまり、転移学習に用いる WSI を、50 枚、100 枚、150 枚、200 枚とする 4 つのケースを比較した。

結果

6 回の交差検証の F1 スコアの平均値を表 7 に示す。表内の()内の数値は 95%信頼区間、下段太字の数値は「自施設 200」モデルとの差を示す。「自施設 200」は施設 T データに対しては施設 T の WSI を 200 枚学習したモデル、「他施設 200」は施設 K の WSI を 200 枚学習したモデルを表す。施設を区別する場合は、それぞれ「T200」、「K200」と表記する。「他施設 200+自施設 50」は「他施設 200」の学習モデルに自施設の 50 枚の WSI を転移学習したモデルを示す。

表 7 F1 スコアによる評価結果

	モデル	F1 スコア：施設 T	F1 スコア：施設 K
1	自施設 200	0.899(±0.005)	0.898(±0.007)
2	他施設 200	0.862(±0.019)	0.862(±0.020)
		-0.037	**-0.036**
3	他施設 200	0.890(±0.009)	0.896(±0.008)
	+自施設 50	**-0.009**	**-0.002**
4	他施設 200	0.891(±0.010)	0.898(±0.007)
	+自施設 100	**-0.008**	**-0.001**
5	他施設 200	0.892(±0.008)	0.900(±0.006)
	+自施設 150	**-0.007**	**+0.002**
6	他施設 200	0.891(±0.008)	0.897(±0.003)
	+自施設 200	**-0.008**	**-0.001**

「自施設 200」モデルの F1 スコアはそれぞれ 0.899、0.898 であり、この値が転移学習での性能の到達目標となる

表 7 の 2 行目に示す「他施設 200」モデルでは、目標値からそれぞれ 0.037、0.036 低下し、3 行目に示す「他施設 200+自施設 50」モデルでは到達目標と同等の性能が得られた。さらに WSI の数を増やしても有意な性能向上はなかった。

他施設モデル適用時の誤り分析

「自施設 200」と「他施設 200」モデルでの、過検出(FP)糸球体数と検出漏れ(FN)糸球体数を表 8 に示す。表中の「差」は FP、FN それぞれの「他施設 200」から「自施設 200」を引いた値であり、()内の数値は、「差」を「自施設 200」の値で割ったものである。

表 8 他施設モデル適用時の FP と FN の変化

		FP	FN	総糸球体数
施設 T	自施設 200	981	1788	13892
	他施設 200	840	2768	
	差	**-141**	**+980**	
		(-14.4%)	**(54.9%)**	
施設 K	自施設 200	595	1242	9093
	他施設 200	854	1489	
	差	**+259**	**+247**	
		(43.5%)	**(19.9%)**	

施設 T では「他施設 200」で FN が 54.9%増加し、施設 K では FP が 43.5%増加している。そこで、施設 T の FN と施設 K の FP を以下に例示する。

(1) 施設 T の WSI での FN 例

「K200」モデルを用いて施設 T の WSI から糸球体を検出した例を図 2 に示す。図中の黄枠が正解領域、赤枠はモデルが予測した領域である。図 2 において、黄枠のみの、検出できなかった糸球体領域が 4 つ確認できる。

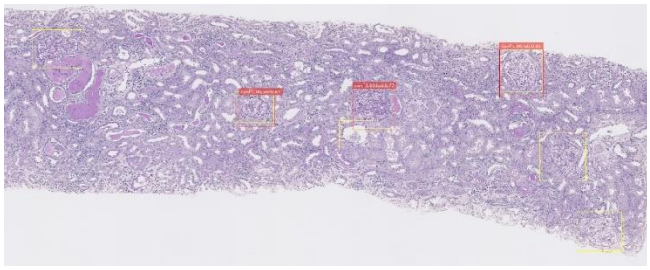

図 2 施設 T の WSI に対する「K200」モデルによる検出結果例（部分）

図 2 と同じ WSI の同じ領域を「T200」モデルで検出した結果を図 3 に示す。図 2 で検出できなかった 4 つの糸球体領域がすべて検出されていることが確認できる。

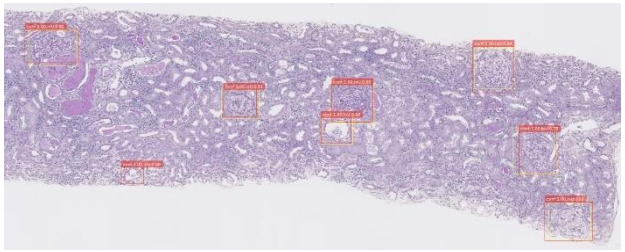

図 3 施設 T の WSI に対する「T200」モデルによる検出結果例（部分）

(2) 施設 K の WSI での FP 例

「T200」モデルを用いて施設 K の WSI から糸球体を検出した場合の FP の例を図 4 に示す。検体組織外の背景の濃淡を糸球体として検出している。

図 4 施設 K の WSI に対する「T200」モデルによる検出結果例（FP 箇所の拡大）

考察

他施設モデルを自施設に適用する場合には、対象画像の染色手法名が同一の場合にも、実施施設ごとに、染色方法やデジタル化処理の細部に差異があり、自施設での性能検証が必要なことを確認した。また、今回の実験では WSI からの糸球体検出処理において、自施設の 50 枚分の WSI を転移学習することで、自施設の WSI を 200 枚学習した場合と同等の性能が得られた。

以下、誤り分析の結果を考察する。図 2 と図 4 に例示した FN と FP の発生理由を推定するために、施設 K の WSI の例を図 5 に示す。

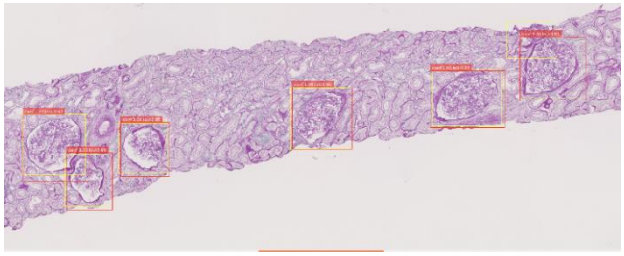

図 5 施設 K の WSI の例（部分）

図 2 と比較して、図 5 の施設 K の WSI の方が糸球体組織の輪郭が鮮鋭に見える。そこで、施設間の画像の鮮鋭性の違いがモデル転移時の性能に影響を与えていると考え、鮮鋭性の違いを評価するために、WSI 毎に Laplacian フィルタで輝度勾配を求め、その分散を比較した[5]。表 9 と図 6 に輝度勾配の分散とそのヒストグラムを示す。施設 T のデータでは分散が 300 未満の WSI が多いが、施設 K のデータでは 250 以上の WSI が多かった。

表 9 輝度勾配の分散の平均

	輝度勾配の分散 (Laplacian フィルタの分散)
施設 T	平均:252.20 (min: 40.68, max:934.84)
施設 K	平均:406.91 (min: 58.96, max: 1283.15)

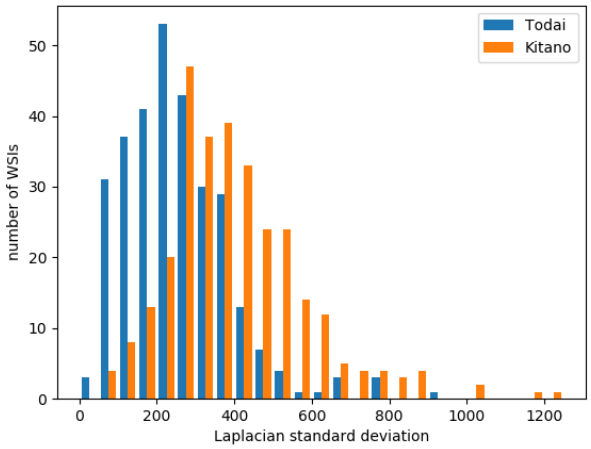

図 6 輝度勾配の分散のヒストグラム

このことから、施設間の鮮鋭性の分布の違いがモデル転移時の性能変化の一因であると考えられた。

WSI の鮮鋭性がデジタル病理画像処理に影響を与えることは既に指摘され、WSI の「ぼけ(blur)」を自動検出する方法も提案されている[6][7]。そのため、今後収集される WSI の鮮鋭性の施設間差異は減少するかもしれないが、すでに蓄積された WSI を利用する際には、鮮鋭性の影響が少ない学習法や利用法が必要となる。そして、既存モデルを転移利用する際の性能向上に、鮮鋭性の差異の情報が利用できる可能性がある。今後の課題として、学習データの鮮鋭性指標を利用して学習モデルの転移時精度向上の可能性を検証するとともに、施設間利用時に有用なデータ要約方法の検討が挙げられる。

結語

今回の実験では、2 つの施設ともに、自施設の 50 枚分の WSI を転移学習することで目標性能に到達した。加えて、施設間の WSI の鮮鋭性の分布の違いが、モデル転移時の性能変化の一因と推測された。

謝辞

本研究は厚生労働科学研究費補助金「臨床研究等 ICT 基盤構築・人工知能実装研究事業(H28-ICT-一般-010)」の助成を受けた

参考文献

[1] M. Temerinac-Ott, et al., Detection of glomeruli in renal pathology by mutual comparison of multiple staining modalities. In Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis, Ljubljana, Slovenia, 18–20 September, (2017)

[2] O. Simon, et al., Multi-radial LBP Features as a Tool for Rapid Glomerular Detection and Assessment in Whole Slide Histopathology Images. Sci. Rep., 8, 2032, (2018)

[3] Y. Kawazoe, et al., Faster R-CNN-Based Glomerular Detection in Multistained Human Whole Slide Images, J.Imaging 4(7), 91, (2018)

[4] Ren, S., et al., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell., 39, 1137–1149, (2017)

[5] R. Bansal, et all., Blur image detection using Laplacian operator and Open-CV, 2016 International Conference System Modeling & Advancement in Research Trends(SMART), pp.63-67, (2016)

[6] X. Lopez, et all., An Automated Blur Detection Method for Histological Whole Slide Imaging, PLOS ONE, Dec. 13, (2013)

[7] G. Campanella, et all., Towards machine learned quality control: A benchmark for sharpness quantification in digital pathology, Computerized Medical Imaging and Graphics, Volume 65, pp.142-151, (2018)