의학석사 학위논문

인공지능을 사용한 요로상피세포암종의 종양세포분율 측정

Artificial intelligence-based algorithm for estimating the neoplastic cell percentage in urinary tract cancer

울 산 대 학 교 대 학 원

의 학 과

정 진 안

인공지능을 사용한 요로상피세포암종의 종양세포분율 측정

지 도 교 수　　조 영 미

이 논문을 의학석사 학위 논문으로 제출함

2024년　2월

울 산 대 학 교 대 학 원

의　　학　　과

정　진　안

정진안의 의학석사학위 논문을 인준함

심사위원 신 동 명 (인)

심사위원 조 영 미 (인)

심사위원 안 보 경 (인)

울 산 대 학 교 대 학 원

2024 년 2 월

Abstract

Estimating the percentage of neoplastic cells (NCP) is crucial for molecular research. While manual counting by a pathologist is the established method, it is time-consuming and not easily executable. To address this, we gathered scanned images of 34 cases of urinary tract cancer and constructed AI models based on convolutional neural networks to estimate NCP. For external validation, 118 cases were obtained, and multiplexed immunofluorescence (mIF) served as the gold-standard. Each AI model demonstrated strong reliability, displaying high intraclass correlation coefficients (ICC) ranging from 0.82 to 0.88. Moreover, they exhibited consistent results compared to human pathologists, with an ICC value of 0.93. These findings suggest that the algorithm of AI models could help effectively human pathologists in repetitive NCP calculations.

Keywords: Artificial intelligence, immunofluorescence, urologic neoplasms, tumor microenvironment

Table of Contents

List of Tables

List of Figures

Introduction

Bladder cancer is one of the most highly mutated tumors with recurrent mutations and potential therapeutic targets in the majority (69%) of such patients[1]. Recently, personalized treatments based on the genetic alteration such as pan-FGFR inhibitors targeting FGFR3 mutations and FGFR3/2 fusions have been emerging[2]. In addition, several molecular targeting agents are under investigation in clinical trials with promising efficacy reported in HER2-targeting antibody-drug conjugate in HER2-overexpressing bladder cancer[3,4]. For such personalized treatment, accurate molecular diagnosis is prerequisite requirement.

The targeted next-generation sequencing (NGS) is widely used not only to define disease-associated genetic alterations for diagnostic purpose but also to find drug-associated clinically actionable targets for personalized medicine. As an initial step of NGS, an accurate assessment of neoplastic cell percentages (NCP) is essential because solid tumors such as bladder cancer contain variable amount of non-neoplastic cells such as desmoplastic fibroblasts, inflammatory cells, vascular endothelial cells, and smooth muscle cells. Depending on the NCP, a NGS test might be preceded or canceled in specimens, especially those with low tumor content near the cutoff level of the test. NGS test with suboptimal NCP might results in false negative results when the test is preceded in spite of insufficient neoplastic cells and interpreted as negative for a variant. Furthermore, inappropriately assessed NCP produces noise and distorts the relationship between read counts, resulting in inaccurate estimation of copy number variation (CNV) in NGS data[5].

NCP, also referring normal cell contamination, is defined as the fraction of cancer

cells in a tumor. Currently, NCP is determined by visual estimation of tumor sections on H&E-stained slides by pathologists[6,7]. However, the pathologic estimates have been issued for its limited accuracy and wide range of interobserver variation[7,8].

The artificial intelligence (AI) has been emerging as a useful tool for quantitative and qualitative analyses of digital histopathologic images[9]. AI-based quantitative image analyses ware tried to estimate NCP for breast cancer[10] and tumor cellularity for lung cancer[11]. However, urinary tract cancer images are not analyzed yet, although they have diverse morphological appearances[12] and clinical settings they stand.

In an attempt to estimate NCP accurately, a machine learning framework-based AI model using digital images was developed for urinary tract malignancy in our study. The AI-based digital estimates were compared to those of pathologists using multiplex immunofluorescence(mIF)-based NCP as the ground-truth.

Material and Methods

Study Overview

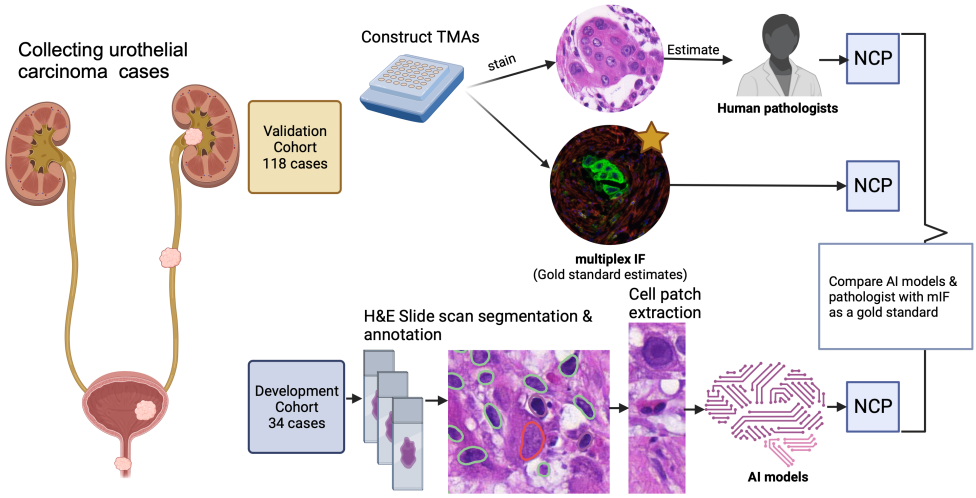

Figure 1. Study overview. AI models were trained using cell patches extracted from the developmental cohort. They estimated NCPs on H&E slides from TMA constructs from another validation cohort. Multiplex IF was also performed and used as a gold standard. NCP estimates by pathologists and AI models were compared for their reliability to multiplex IF. AI, artificial intelligence; NCP, neoplastic cell percentage; H&E, hematoxylin and eosin; TMA, tissue microarray; IF, immunofluorescence.

Patients

This retrospective study was approved by the Institutional Review Board of Asan Medical Center (AMC) (2022-1558) and included patients with pathologically confirmed urinary tract malignancies at Asan Medical Center, Seoul, Republic of Korea and their formalin-fixed, paraffin-embedded (FFPE) tumor tissue blocks were

available for tissue microarray (TMA) construction. Two cohorts were established for this study: a development cohort and a validation cohort.

The development cohort was utilized to create numerous cell patches for training the convolutional neural network models. This cohort was collected between March 2022 and August 2022, encompassing various subtypes and divergent differentiations of urothelial carcinoma, as well as commonly encountered specimen types.

For the validation cohort, NGS cases of the urinary tract were included. This cohort, gathered between May 2019 and February 2022, was designed to simulate a clinical setting using NCP and used to validate the trained machine learning models.

Tissue microarray construction

TMA blocks were generated from the validation cohort and constructed with 1 mm-diameter cores from 10% neutrally buffered FFPE tumor blocks using a tissue microarrayer (Quick-Ray, Unitma Co. Ltd., Seoul, Republic of Korea). Three representative cores from each case were selected in the area with the variable density of viable neoplastic cells, low to high, while trying to avoid necrotic areas[13].

Multiplex IF and multispectral imaging analysis

Four-micron thick tissue sections were cut from the TMA construct; and then transferred onto plus-charged slides for multiplexed fluorescent immunohistochemistry using a Leica Bond Rx™ Automated Stainer (Leica Biosystems) and Opal Polaris 7-Color Automated IHC Detection Kit (AKOYA Biosciences) as previously described[14]. After sequential reactions, tissue sections

were counterstained with DAPI (62248, Thermo Scientific) and mounted with ProLong Gold antifade reagent (P36935, Invitrogen). The primary antibodies used in this study were CD45 (DAKO, Santa Clara, CA, USA,1:200), alpha-smooth muscle actin (αSMA, Zymed, San Francisco, CA, USA, 1:100), pan-cytokeratin (CK, Novus Biologicals, Littleton, CO, 1:300) and the corresponding fluorophores for fluorescence signals were Opal 570, Opal 690 and Opal 780, respectively. The multispectral fluorescence images were acquired on multiplex stained slides, scanned on the Vectra® Polaris Automated Quantitative Pathology Imaging System (Akoya Biosciences), and the images were visualized in the Phenochart Whole Slide Viewer (Akoya Biosciences). Phenotyping of cellular components in the images was performed using inForm image analysis software and phenoptr/phenoptrReports tissue analysis software packages (Akoya Biosciences). Based on the phenotyping, mIF-based NCP was calculated and used as the ground-truth value.

Whole-slide scanning

H&E slides from the development cohort and the validation cohort were scanned with the PANNORAMIC 250 Flash II (3D HISTECH, Budapest, Hungary) at 40x magnification with 0.22 μm/pixel in a single layer.

Deep learning models for NCP estimation

Cell patch generation for AI model training

Scanned slides were assessed using open-source digital pathology software, QuPath v. 3.4.2.[15]. In the development cohort, the most representative regions of interest, at least one per slide, were drawn by the author (J. A. J.). Cell nuclei were

detected with star-convex polygons[16] using the provided QuPath plugin to avoid incomplete segmentation of overlapping nuclei. The detected cells were then manually classified into three classes: tumor cells, stromal cells and immune cells, one at a time. For each detected cell, 100×100 pixel image patches were obtained with their class labels. Extracted cell patches containing 3 classes were divided into 3 sets, training set, validation set, and test set, as 7:2:1 ratio and supplied to convolutional neural networks.

Training AI models

Cell patches in the training and validation sets were transformed to tensors and augmented for generality, handled by Pytorch library version 1.12.1[17]. Cell patches were provided to nine convoluted neural network (CNN) models, basically provided by Pytorch. The models were AlexNet[18], VGG[19], ResNet[20], WideResNet20[21], EfficientNet[22], EfficientNet V2[23], MobileNet V2[24], MobileNet V3[25], and ShuffleNet V2[26]. The Adam optimizer[27] was adopted with default hyperparameters ($\beta1 = 0.9$; $\beta2 = 0.999$; $\varepsilon = 1.0 \times 10 - 8$). The Cross Entropy Loss function and Reduce LR On Plateau function were used as the loss function and learning rate scheduler, respectively. Batch size was set to 128. Learning epoch was set to 80. Models were computed by two GPUs, RTX 3090 (NVIDIA).

Performance metrics of AI models

The performance of the trained models was evaluated in the predetermined test set with four parameters: sensitivity, specificity, precision, accuracy, F1 score and the area under the receiver operating characteristic (AUROC).

Comparison of NCP estimation performance between pathologists and AI models

The NCPs of each core of H&E-stained slides from the TMA constructs of the validation cohort were estimated by seven pathologists with varying levels of expertise and the nine trained CNN models. The pathologists consisted of one uropathologist (YM Cho), two fellows (SE Jeong and BK Ahn), and three residents (KH Kim, HJ Seong, and YI Lee). They were instructed to estimate each individual TMA core by eyeball measurement, not by counting cells individually. They provided NCP estimates on a 5% scale, ranging from 0% to 100%.

To obtain NCP estimates from the CNN models, TMA slides were scanned and cell patches were generated in a similar way as training patches. The trained deep learning models were then applied and classified cell patches into three classes, tumor, stroma, and immune cells, and provided NCP estimates of each TMA core.

The agreement between mIF-based NCP and those of pathologists or AI models was calculated using intraclass correlation coefficients, using ICC (2,1) for comparing between mIF-based NCP and each pathologist or AI model. ICC (2, k) was used for comparing mIF-based NCP and pathologist group or AI model group[28]. The ICC was interpreted as poor (<0.40), fair (0.40-0.59), good (0.60-0.74) and excellent (0.75-1.00) as previously proposed[29,30].

Identifying difficult cases

Mean absolute deviation (MAD) was used for identifying TMA cores with low agreements. MAD is used for measuring of variability of data, implying how much data values are spread out from the mean. The mean of the absolute deviations

around the data's mean is defined as MAD.

$$\text{MAD} = \sum_i \frac{|x_i - \bar{x}|}{n}$$

Results

Patient cohorts

In the development cohort, from 322 cases of invasive high-grade urothelial carcinoma, 34 slides were selected. Data variability and image artifact were also considered not only variable histologic morphology[31] (Table 1).

Twenty cases were transurethral resection and fourteen cases were radical surgical resection specimens, which included radical cystectomy or nephroureterectomy. The cohort consisted of 21 pure invasive urothelial carcinoma and 13 urothelial carcinomas with various divergent differentiations and subtypes, including glandular and squamous differentiation and microcystic, sarcomatoid, poorly differentiated, and micropapillary subtypes (Table 2).

In the validation cohort, 118 specimens were collected, including 85 radical surgical resection specimens and 23 transurethral resection specimens. The validation cohort consisted of 105 cases of invasive high-grade urothelial carcinoma, one case of non-invasive papillary urothelial carcinoma, two cases of collecting duct carcinoma, two cases of urachal adenocarcinoma, and one case of invasive squamous cell carcinoma. Fifty-seven cases of urothelial carcinoma contained various differentiations and histologic subtypes, including squamous, sarcomatoid, glandular, plasmacytoid, micropapillary, giant cell, nested, and microcystic (Table 2).

Table 1. A checklist for findings to include in whole slide images of urinary tract cancer.

Tissue condition
Fixation
1. Well fixed
2. Poorly fixed
Cauterization effect
3. Cauterized
4. Not cauterized
Focus
5. Well focused
6. Partly focused out
Section thickness
7. Thick
8. Thin
Tearing
9. Torn
10. Not torn
Background
11. Clear
12. Hemorrhagic
13. Degenerative
14. Necrotic
Location
15. In tissue
16. Floating
Cellularity
17. High
18. Low
Tumor
Morphology
19. Round
20. Polygonal
21. Irregular
Cohesiveness
22. Cohesive
23. Overlapping
24. Poorly cohesive
Cell membrane
25. Distinct
26. Indistinct
Mitosis
27. Active
28. Inactive
Nucleoli
29. Prominent
30. Inconspicuous

Chromatin
31. Fine
32. Coarse
33. Vesicular
Cytoplasm
34. Eosinophilic
35. Amphophilic
36. Basophilic
37. Clear
38. Scant
Mesenchyme
Location
39. Intra tumoral
40. Peri tumoral
41. Extra tumoral
Structure
42. Adipose
43. Muscularis propria
44. Muscularis mucosae
45. Large vessel
46. Small vessel
47. Papillary core
48. Nerve
49. Granulation tissue
Reaction
50. Desmoplastic
51. Hyalinized
52. Myxoid
53. Edematous
54. Activated
Immune cells
Location
55. Intra tumoral
56. Peri tumoral
57. Extra tumoral
Structure
58. Aggregate
59. Follicle
60. Abscess
Cells
61. Neutrophil
62. Eosinophil
63. Mature lymphocyte
64. Histiocyte

Table 2. Patient characteristics in the development and validation cohort.

Characteristic	Development cohort (n=34)		Validation cohort (n=118)	
	No.	%	No.	%
Sex				
Male	24	70.6	90	76.3
Female	10	29.4	28	23.7
Location				
Urinary Bladder	22	64.7	49	41.5
Renal pelvis	5	14.7	16	13.6
Ureter	4	11.8	27	22.9
Urethra	0	0	1	0.8
Multiple location*	2	5.9	11	9.3
Others†	1	2.9	14	11.9
Procedure				
TURB	20	58.8	23	19.5
Surgical resection‡	14	41.2	85	72.0
Biopsy	0	0	1	0.8
Others§	0	0	9	7.6
Histologic grade				
Low	0	0	1	0.8
High	34	100	105	89.0
Not assessed‖	0	0	12	10.2
Pathologic T staging				
Tx	0	0	11	9.3
T0¶	1	0	2	1.7
T1	14	41.2	7	5.9
T2	9	26.5	29	24.6
T3	10	23.5	53	44.9
T4	0	0	16	13.6
Histologic subtype				
None	21	61.8	56	42.1
Squamous	5	14.7	26	19.5
Micropapillary	2	5.9	23	17.3
Sarcomatoid	2	5.9	8	6.0
Other	6	14.7	15	11.3
Non-urothelial**	0	2.9	5	3.8
LVI				
Not identified	20	58.8	76	64.4
Present	13	38.2	32	26.9
NA	1	2.9	10	8.4
Pathologic LN status				
Nx	25	73.5	52	44.0
Negative (N0)	5	14.7	33	28.0
Positive (N1/N2/N3)	4	11.8	33	28.0
UCIS				
Not identified	18	52.9	36	30.5

Present	15	44.1	64	54.2
N/A	1	2.9	18	15.3
Total	34		118	

* Cases with two or more detected urothelial carcinomas, at same organ or not were counted as multiple location cases.

† Metastectomy cases, such as lymphadenectomy, adrenalectomy, lung wedge resection, were usually assigned.

‡ Radical surgery cases performed to primary urothelial carcinoma for therapeutic purpose were assigned. It contains such as radical cystectomy, nephroureterectomy, distal ureterectomy, and partial cystectomy specimens.

§ Metastectomy cases for metastatic tumor in peritoneum, adrenal gland, and lung were included.

ǁ Metastatic urothelial carcinoma cases not mentioned their histological grade additionally were included.

¶ Pathologic complete remission of primary tumor after neoadjuvant chemotherapy with lymph node metastasis cases were included.

** Collecting duct carcinoma, urachal adenocarcinoma and pure squamous cell carcinoma cases were included.

TURB, transurethral resection of bladder.

Multiplex IF performed as the gold-standard

Three tissue microarrays were constructed with tissue cores in 1mm diameter and used for multiplex IF with 331 cores available for staining (Figure 2). Cell counts of tumor cells, immune cells, and stromal cells were determined by pan-cytokeratin-, CD45-, and SMA-positive cells, respectively (Figure 3). NCPs were calculated by dividing the number of cytokeratin-positive cells by the total number of DAPI-positive cells. They ranged from zero to ninety-nine. Fifteen cores contained no tumor cells. NCP less than 20% were 71 cores (Figure 4). The proportion of cells with a single immunophenotype, positive for only one marker, was 89.4%. The most common double-positive immunophenotype was CD45 and αSMA double positivity, representing 6.9% of total cell counts (Table 3).

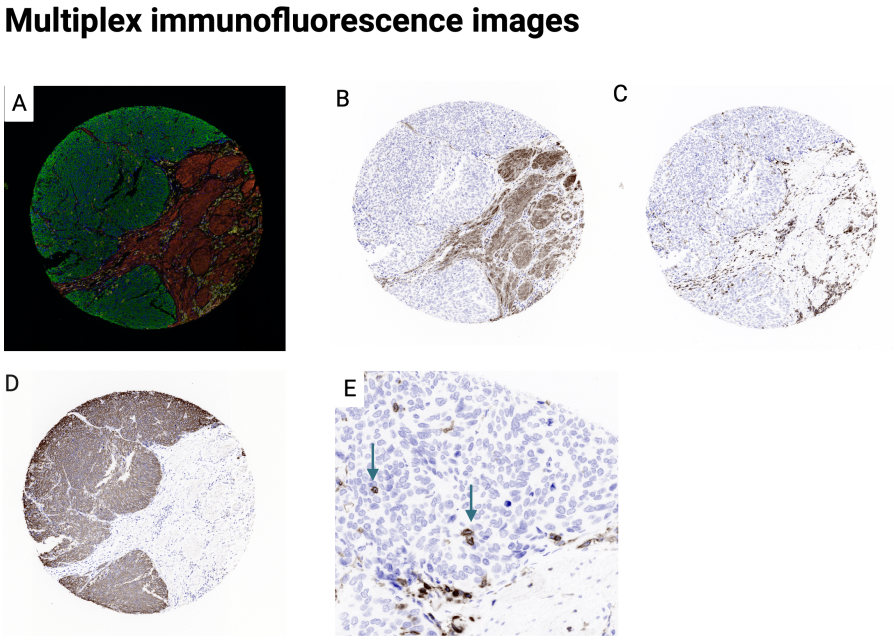

Figure 2. An example of multiplex immunofluorescence image.
(A) Multiplex immunofluorescence images; Blue, DAPI; Green, CK; Red, α SMA; Yellow, CD45. (B, C, D): Images are highlighted by each antibody channel: CK (B), α SMA (C), and CD45(D). (E) CD45 stain highlights intratumoral lymphocytes (blue arrows).

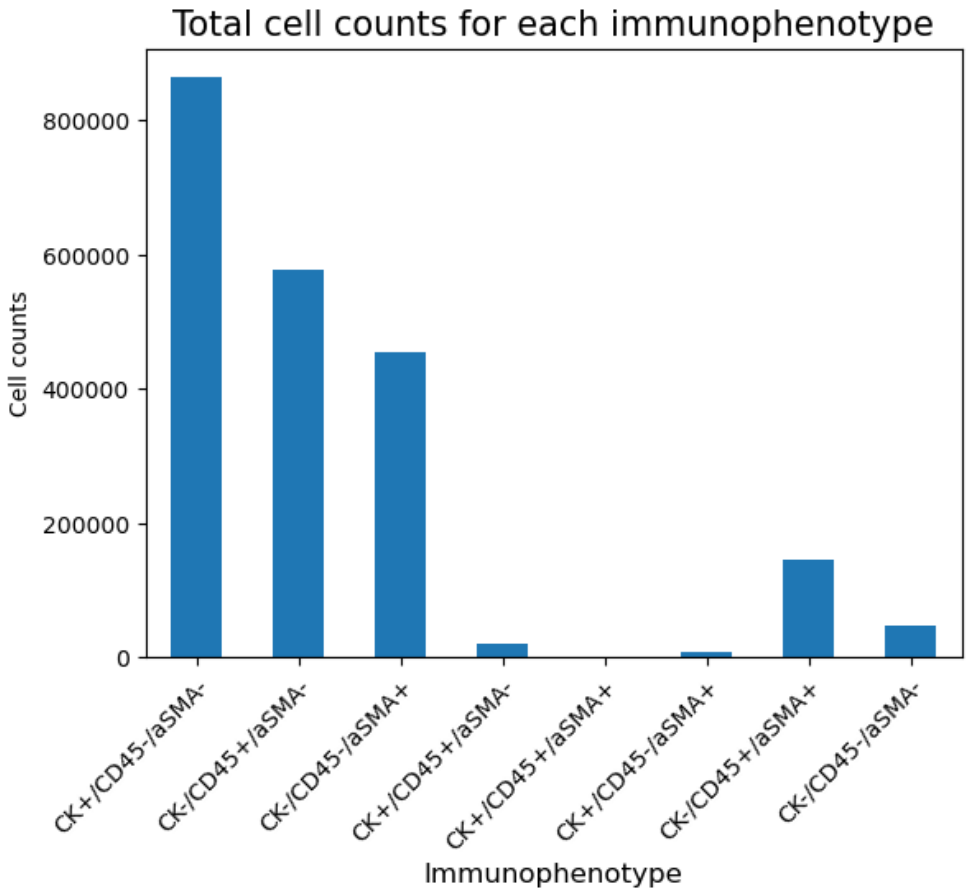

Figure 3. Total cell counts in each TMA core. Immunophenotyped cell counts by mIF for each TMA cores were visualized among eight possible immunophenotypes.

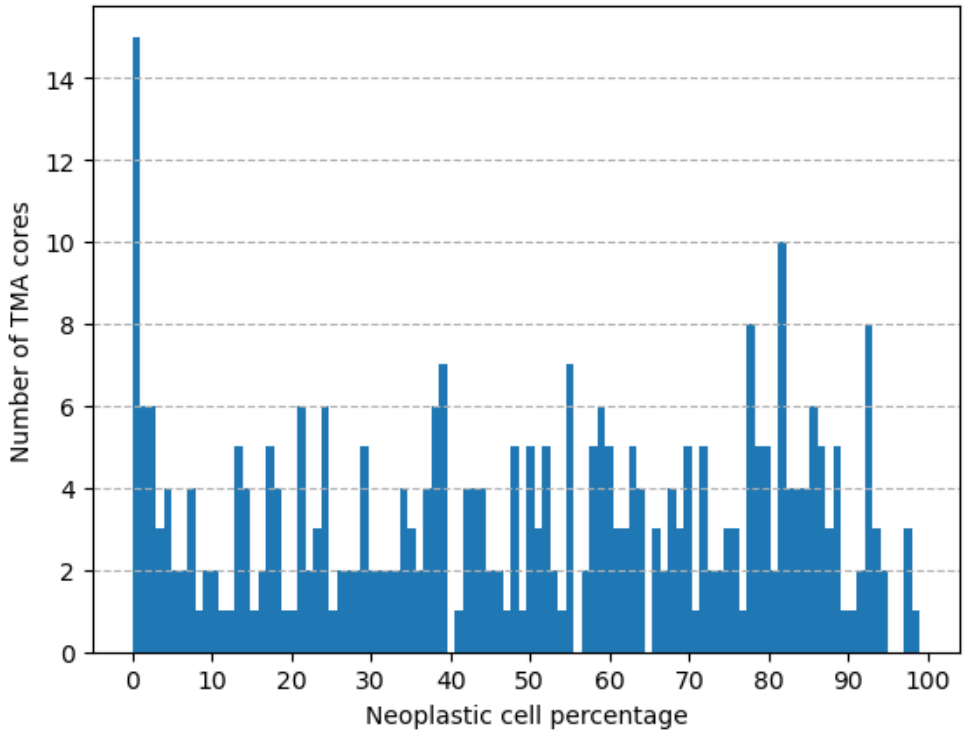

Figure 4. Distribution of mIF-based NCP estimates in the validation cohort.

Table 3. Total immunophenotyped cells by mIF.

Immunophenotype	Cell counts	Percentage (%)
CK+/CD45-/αSMA-	863531	40.7
CK-/CD45+/αSMA-	577724	27.2
CK-/CD45-/αSMA+	455225	21.5
CK+/CD45+/αSMA-	20299	1.0
CK+/CD45+/αSMA+	56	0.0
CK+/CD45-/αSMA+	9035	0.4
CK-/CD45+/αSMA+	145584	6.9
CK-/CD45-/αSMA-	48559	2.3
Total	2120013	100

Optimizing AI models

Nine supervised deep learning models were evaluated to classify provided cell patches into three distinct classes. Within the training cohort, a total of 133,941 cell patches were manually classified, comprising 76,330 tumor cell patches, 24,297 stromal cell patches, and 33,314 immune cell patches (Figure 5). These cell patches were divided into training, validation, and test sets as a 7:2:1 ratio. The image patches in the training set were converted into tensors and augmented to improve generalization using techniques such as random flipping, rotation, padding, and normalization (Figure 6). These augmented patches were then fed into a convolutional neural network model to train it in classifying cell patches into the three specified classes. Model performance was evaluated using the test set (Table 4). Among these models, EfficientNet showed the highest sensitivity (0.94) and accuracy (0.87) while two models (AlexNet and VGG) demonstrated low accuracy and area under the receiver operating characteristic curve (AUROC) scores of 0.55 and 0.5, respectively. (Table 4).

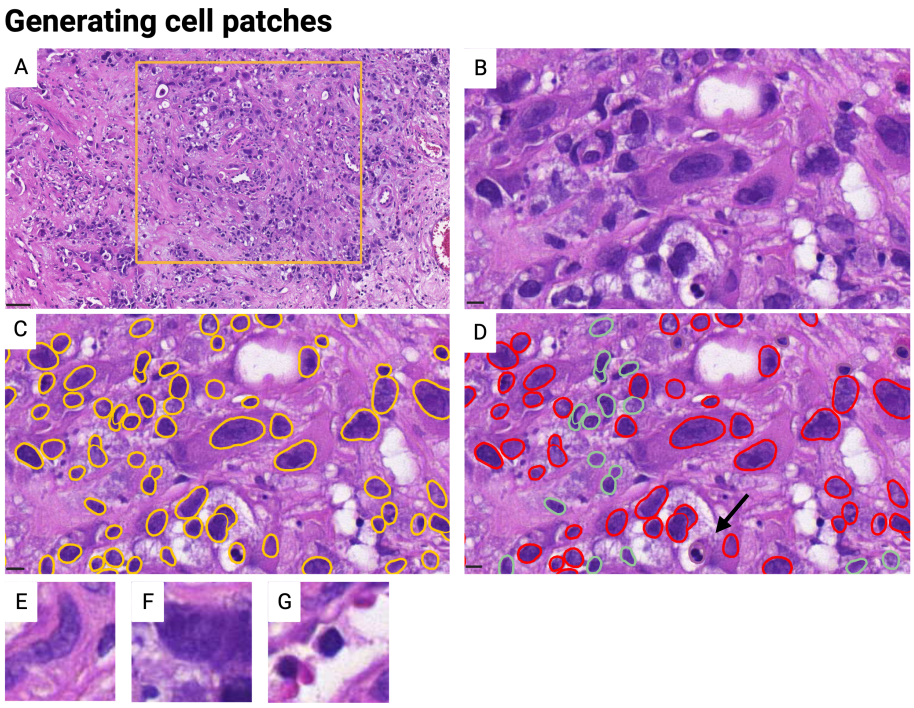

Figure 5. Cell patch generation to train AI models using development cohort. (A) Region of interest is selected for each digitally scanned whole slide images. (B) High magnification view of square area at (A). (C) Segmented nuclei by yellow outline. (D) Manually classified cells (Red: tumor cells; green: stromal cells; purple: immune cells (arrow)). (E, F, G): Examples of extracted 100×100 pixel image patches (E: stromal cells; F: tumor cells; G: immune cells)

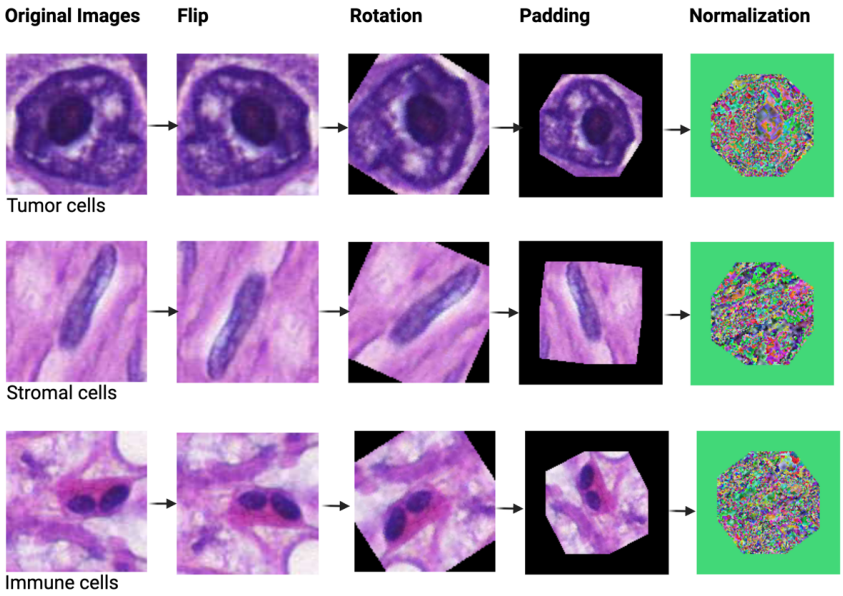

Figure 6. Image augmentation techniques for helping AI training.

Table 4. Model evaluation metrics in the development cohort.

Model	Sensi-tivity	Speci-ficity	Precision	Accuracy	F1-score	AUROC
AlexNet	1.0	0.0	0.55	0.55	0.71	0.5
VGG	1.0	0.0	0.55	0.55	0.71	0.5
EfficientNet	0.94	0.88	0.90	0.87	0.92	0.96
EfficientNet V2	0.93	0.87	0.90	0.86	0.91	0.96
MobileNet V2	0.92	0.86	0.89	0.85	0.90	0.95
MobileNet V3	0.92	0.86	0.89	0.86	0.90	0.95
ResNet	0.92	0.88	0.90	0.86	0.91	0.96
WideResNet	0.92	0.89	0.91	0.86	0.91	0.96
ShuffleNet V2	0.93	0.87	0.89	0.86	0.91	0.96

Comparison of interobserver agreement between AI models and pathologists in NCP assessment

In general, both the pathologist group and the AI model group showed high reliability for mIF-based NCPs. They also showed high reliability for each other (Table 5). The all pathologists estimated NCP with excellent reliability, with ICCs ranging from 0.81 to 0.91, of which the pathologist specializing in uropathology estimated NCP with the highest reliability, 0.91 (95% CI, 0.89-0.93) (Table 6). AI models also estimated NCP with excellent reliability, with ICCs ranging from 0.83 to 0.88. The most reliable model was EfficientNet with an ICC of 0.88 (95% CI, 0.78-0.92). MobileNet V3 was the least reliable model with an ICC of 0.82 (95% CI, 0.7-0.88), but was still more reliable than all but three human pathologists (Figure 7 and Table 7).

To evaluate the accuracy of NCP assessment according to histological variation, we divided the validation cohort into three groups, urothelial carcinoma without divergent differentiation, urothelial carcinoma with divergent differentiation, and non-urothelial carcinoma (e.g., squamous cell carcinoma and urachal adenocarcinoma). ICC values were then measured. AI models demonstrated a robust estimation of NCP in cases of urothelial carcinoma, both with [ICC, 0.94 (95% CI, 0.91-0.95)] and without [ICC, 0.95 (95% CI, 0.94-0.97)] divergent differentiation. However, the AI models demonstrated lower reliability in non-urothelial carcinoma cases, showing lower reliability [ICC, 0.71 (95% CI, 0.07-0.91)] than those of pathologists [ICC, 0.84 (95% CI, 0.41-0.95)] (Table 8 and Table 9).

Table 5. ICC values among each rater group.

Rater group	ICC value	CI 95%

mIF/Pathologists	0.94	[0.91,0.96]
mIF/AI models	0.94	[0.92, 0.95]
Human/AI models	0.93	[0.89, 0.95]

ICC, intraclass correlation coefficient; mIF, multiplex immunofluorescence; CI, confidence interval.

Table 6. ICC values among mIF and each pathologist.

Rater	ICC value	CI 95%
Specialist	0.91	[0.89, 0.93]
Fellow 1	0.80	[0.53, 0.89]
Fellow 2	0.82	[0.78, 0.85]
Resident 1	0.78	[0.72, 0.83]
Resident 2	0.82	[0.78, 0.85]
Resident 3	0.81	[0.73, 0.86]

NCP, neoplastic cell percentage.

Table 7. ICC values among mIF and each AI model.

Model	ICC value	CI 95%
EfficientNet	0.88	[0.78, 0.92]
EfficientNet V2	0.87	[0.84, 0.89]
MobileNet V2	0.87	[0.85, 0.90]
MobileNet V3	0.82	[0.7, 0.88]
ResNet	0.86	[0.76, 0.91]
WideResNet	0.85	[0.8, 0.88]
ShuffleNet V2	0.83	[0.74, 0.88]

Table 8. ICC values among mIF, each pathologist and AI model, divided into three histologic groups.

Rater	Urothelial carcinoma without divergent differentiation (n=56)		Urothelial carcinoma with divergent differentiation (n=57)		Non-urothelial carcinoma (n=5)	
	ICC value	CI 95%	ICC value	CI 95%	ICC value	CI 95%
Specialist	0.91	[0.87, 0.93]	0.94	[0.91, 0.95]	0.75	[0.39, 0.91]
Fellow 1	0.83	[0.59, 0.91]	0.78	[0.48, 0.89]	0.67	[0.13, 0.89]
Fellow 2	0.83	[0.77, 0.87]	0.84	[0.79, 0.88]	0.61	[0.17, 0.85]
Resident 1	0.82	[0.65, 0.89]	0.76	[0.69, 0.82]	0.77	[0.2, 0.93]
Resident 2	0.84	[0.79, 0.88]	0.84	[0.79, 0.88]	0.63	[0.2, 0.86]
Resident 3	0.83	[0.7, 0.89]	0.80	[0.72, 0.85]	0.64	[0.17, 0.87]
EfficientNet	0.89	[0.81, 0.93]	0.89	[0.67, 0.95]	0.62	[0.18, 0.86]
EfficientNetV2	0.88	[0.83, 0.91]	0.90	[0.85, 0.93]	0.60	[0.12, 0.85]
MobileNetV2	0.88	[0.84, 0.91]	0.90	[0.87, 0.93]	0.44	[-0.12, 0.78] *
MobileNetV3	0.81	[0.65, 0.89]	0.87	[0.8, 0.91]	0.33	[-0.18, 0.71] *
ResNet	0.87	[0.8, 0.92]	0.87	[0.65, 0.93]	0.48	[-0.0, 0.79]
WideResNet	0.85	[0.78, 0.9]	0.90	[0.86, 0.92]	0.55	[0.05, 0.83]
ShuffleNetV2	0.83	[0.68, 0.9]	0.89	[0.84, 0.92]	0.54	[0.03, 0.82]

**P* values for F test are greater than 0.05.

Table 9. ICC values among different observer groups divided into three histologic groups.

Rater group	Urothelial carcinoma without divergent differentiation (n=56)		Urothelial carcinoma with divergent differentiation (n=57)		Non-urothelial carcinoma (n=5)	
	ICC value	CI 95%	ICC value	CI 95%	ICC value	CI 95%
mIF/Pathologists	0.94	[0.9, 0.97]	0.94	[0.91, 0.96]	0.84	[0.41, 0.95]
mIF/AI models	0.94	[0.91, 0.95]	0.95	[0.94, 0.97]	0.71	[0.07, 0.91]
Pathologists/AI models	0.93	[0.82, 0.97]	0.94	[0.92, 0.96]	0.84	[0.51, 0.94]

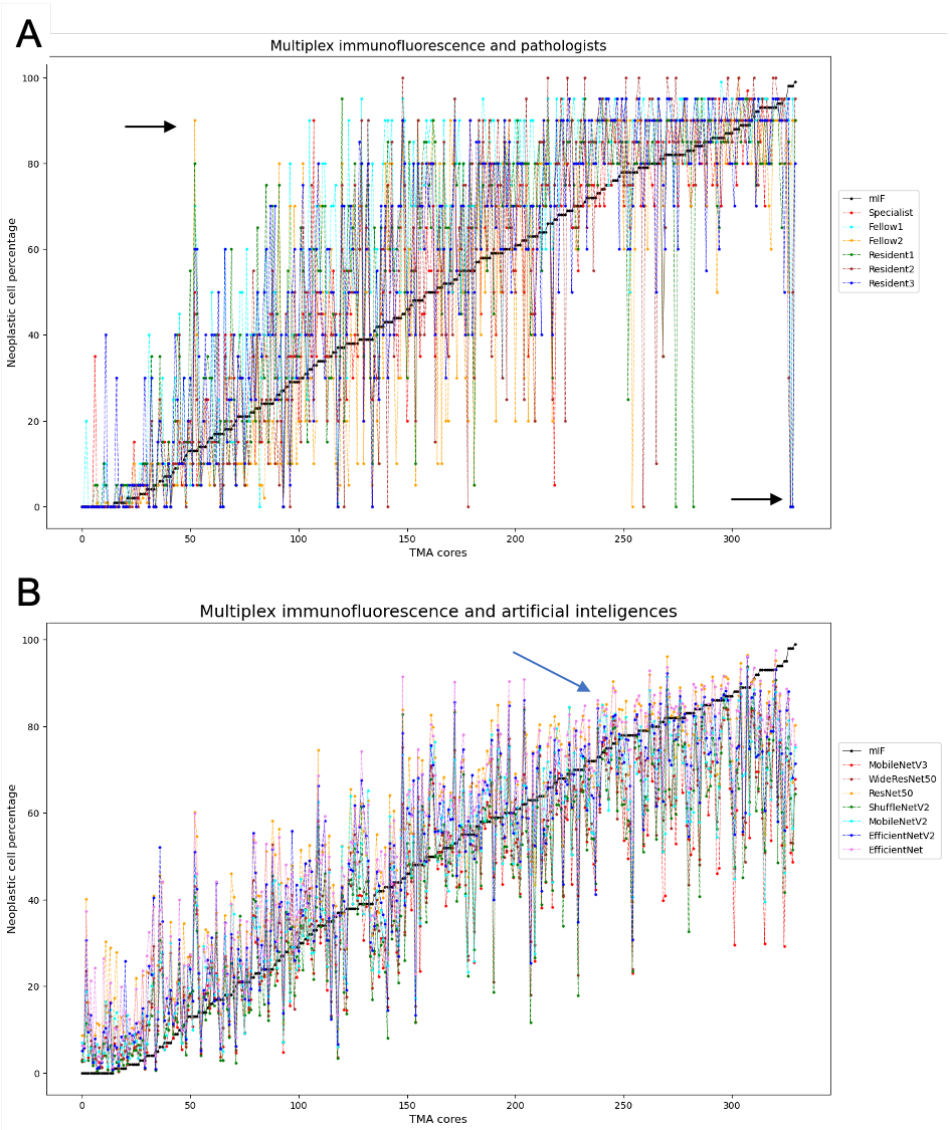

Figure 7. NCP estimations by pathologists (A) and AI models (B) for each TMA core. The values are sorted according to the corresponding mIF values in ascending order from 0 to 99. (A) NCP estimations among pathologists are extremely variable in any range of NCP (black arrows). (B) NCP estimations among AI models are less variable in any range of NCP. However, they have a tendency to underestimate NCP, of which mIF estimates more than 60% (blue arrow).

Difficult cases with high variability

TMA cores with widely spread NCPs (n = 27) were detected by measuring MAD, of which cut-off value was set to 15. Their H&E and mIF slides were reviewed and several features were suspected as possible causes: (1) high cellularity (n = 20); (2) cauterization and/or crush artifact (n = 10); (3) Abundant cytoplasm (n = 6); (4) histologic variant (micropapillary, plasmacytoid and sarcomatoid) (n = 3); (5) learning failure of AI (n = 3); (6) tissue loss by deeper section (n = 1) (Figure 8).

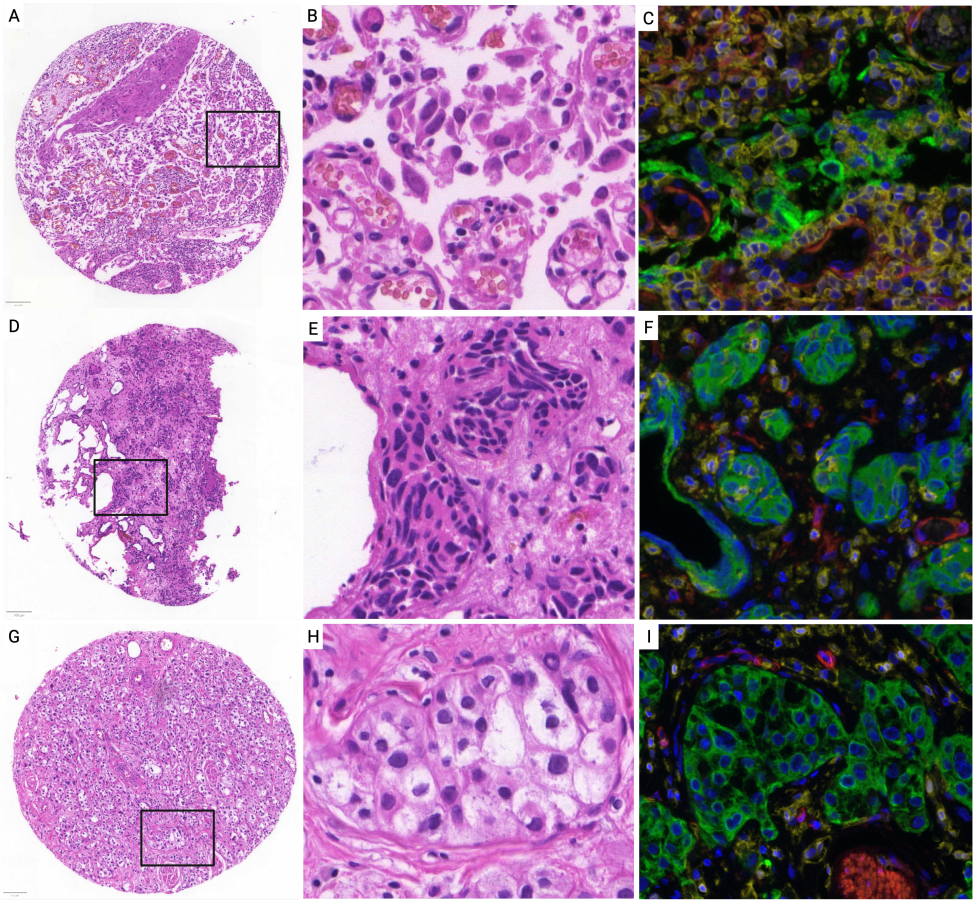

Figure 8. Examples of H&E and mIF images for TMA cores with high mean absolute deviation. Disagreements usually occur in various settings. (A) High cellularity precludes accurate NCP estimations to pathologists (range, 0-80%). Those of AI models (range, 15.5-36.5%) and mIF (29%) are not consistent with them. (B) High magnification view at square area in (A). (C) High magnification view of mIF shows tumor cells (green) are admixed with immune cells (yellow). Surrounding stromal cells were also observed (red). (D) Cauterization artifact makes NCP estimations inaccurate to pathologists (range, 0-70%) and AI models (range, 14.5-45.4%). (E) The artifact restricts an assessment to cytologic details. (F) However, mIF (37%) is less affected by this artifact. (G) Certain cytologic feature, such as abundant cytoplasm makes NCP estimation difficult. (H) Pathologists overestimate NCPs (range, 65-90%) than those of AI models (range, 32.6-48.6%) and (I) mIF (43%).

Discussion

Although pathologists group showed quite excellent reliability compared to group of AI models, their ranges of NCP estimation were quite broad while those of AI models were not (Figure 3). In practice, NCP is estimated by one pathologist. It seems hard to consistently estimate NCPs for human pathologists, especially trainee.

Our study applied mIF as a new method for calculating NCP and used mIF-based estimation as the gold standard. Cytoplasm stained for cytokeratin and CD45 circumferentially surrounded nuclei, so tumor cells and immune cells were easily immunophenotyped, but αSMA staining was not. However, the unphenotyped mesenchymal cells and double-positive cells for CD45 and αSMA did not affect NCP estimation because only cytokeratin-positive tumor cells among the total DAPI-positive nuclei were calculated for NCP estimation.

The AI models provided reliable estimates of NCP initially despite having limited training data. The validation cohort contained approximately ten times as many cases as the training cohort and exhibited a more diverse range of histological appearances. Furthermore, these appearances were not included in their training material, and there were variations in the quality of H&E slides scanned by different WSI scanners. Even though they were not specifically trained for non-urothelial carcinoma cases such as mucinous adenocarcinomas in the development cohort, they showed moderate to good reliability. This suggests that the models have a degree of generalizability across diverse tumor morphologies of urinary cancers.

Although relatively earlier deep learning models such as AlexNet[18] and VGG[19] faced challenges in being trained for the classification of tumor cell patches, other

AI models were considerably trained, achieving an accuracy surpassing 0.85. Nonetheless, there was no overwhelmingly superior model. This observation might indicate that cell classification tasks demand not just substantial computational resources but also sophisticated cognitive capabilities.

Human pathologists experienced fatigue when assessing NCPs for approximately three hundred TMA cores, a challenge not encountered by the models. Many TMA cores were derived from FFPE blocks, which could be considered sufficient in quantity to distinguish the differences between human pathologists and AI models. However, human pathologists had an advantage in examining smaller regions. As the area expands, the task becomes difficult, a challenge usually not encountered by artificial intelligence. Therefore, a one-millimeter diameter TMA core may not accurately represent the conditions of real-world molecular testing, which requires a larger tumor area because it contains a greater number of tumor cells.

References

1. Robertson, A. G., Kim, J., Al-Ahmadie, H., et al. (2018, Aug 9). Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell, 174*(4), 1033.
2. Loriot, Y., Necchi, A., Park, S. H., et al. (2019, Jul 25). Erdafitinib in Locally Advanced or Metastatic Urothelial Carcinoma. *N Engl J Med, 381*(4), 338-348.
3. Patelli, G., Zeppellini, A., Spina, F., et al. (2022, Mar). The evolving panorama of HER2-targeted treatments in metastatic urothelial cancer: A systematic review and future perspectives. *Cancer Treat Rev, 104*, 102351.
4. Sheng, X., Yan, X., Wang, L., et al. (2021, Jan 1). Open-label, Multicenter, Phase II Study of RC48-ADC, a HER2-Targeting Antibody-Drug Conjugate, in Patients with Locally Advanced or Metastatic Urothelial Carcinoma. *Clin Cancer Res, 27*(1), 43-51.
5. Zare, F., Dow, M., Monteleone, N., Hosny, A., & Nabavi, S. (2017, May 31). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics, 18*(1), 286.
6. Dufraing, K., van Krieken, J. H., De Hertogh, G., et al. (2019, Sep). Neoplastic cell percentage estimation in tissue samples for molecular oncology: recommendations from a modified Delphi study. *Histopathology, 75*(3), 312-319.
7. Smits, A. J., Kummer, J. A., de Bruin, P. C., et al. (2014, Feb). The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. *Mod Pathol, 27*(2), 168-174.

8. Haider, S., Tyekucheva, S., Prandi, D., et al. (2020). Systematic Assessment of Tumor Purity and Its Clinical Implications. *JCO Precis Oncol, 4*.
9. Baxi, V., Edwards, R., Montalto, M., & Saha, S. (2022). Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol, 35(1), 23-32.
10. Azimi, V., Chang, Y. H., Thibault, G., Smith, J., Tsujikawa, T., Kukull, B., Jensen, B., Corless, C., Margolin, A., & Gray, J. W. (2017). Breast Cancer Histopathology Image Analysis Pipeline for Tumor Purity Estimation. Proc IEEE Int Symp Biomed Imaging, 2017, 1137-1140.
11. Lin, S., Samsoondar, J. P., Bandari, E., Keow, S., Bikash, B., Tan, D., Martinez-Acevedo, J., Loggie, J., Pham, M., Wu, N. J., Misra, T., Lam, V. H. K., Sansano, I., & Cecchini, M. J. (2023). Digital Quantification of Tumor Cellularity as a Novel Prognostic Feature of Non-Small Cell Lung Carcinoma. Mod Pathol, 36(3), 100055.
12. Amin, M. B. (2009). Histological variants of urothelial carcinoma: diagnostic, therapeutic and prognostic implications. Mod Pathol, 22 Suppl 2, S96-S118.
13. Dufraing, K., van Krieken, J. H., De Hertogh, G., et al. (2019, Sep). Neoplastic cell percentage estimation in tissue samples for molecular oncology: recommendations from a modified Delphi study. *Histopathology, 75*(3), 312-319.
14. Ahn, J., Jin, M., Song, E., Ryu, Y. M., et al. (2021, Jan). Immune Profiling of Advanced Thyroid Cancers Using Fluorescent Multiplex Immunohistochemistry. Thyroid, 31(1), 61-67.

15. Bankhead, P., Loughrey, M. B., Fernandez, J. A., et al. (2017, Dec 4). QuPath: Open source software for digital pathology image analysis. *Sci Rep, 7*(1), 16878.
16. Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (2018). Cell Detection with Star-Convex Polygons. *Medical Image Computing and Computer Assisted Intervention - Miccai 2018, Pt Ii, 11071*, 265-273.
17. Paszke, A., Gross, S., Massa, F., Lerer, A., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems, 32*.
18. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012b). *ImageNet classification with deep convolutional neural networks* Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Lake Tahoe, Nevada.
19. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
20. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.
21. Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.
22. Tan, M. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* Proceedings of the 36th International Conference on Machine Learning.
23. Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster

training. International conference on machine learning.

24. Sandler, M., Howard, A., Zhu, M., et al. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition.
25. Howard, A., Sandler, M., Chu, G., et al. (2019). Searching for mobilenetv3. Proceedings of the IEEE/CVF international conference on computer vision
26. Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. Proceedings of the European conference on computer vision (ECCV).
27. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
28. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, *86*(2), 420-428.
29. Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment, 6*, 284-290.
30. Robert, M. E., Ruschoff, J., Jasani, B., et al. (2023, May). High Interobserver Variability Among Pathologists Using Combined Positive Score to Evaluate PD-L1 Expression in Gastric, Gastroesophageal Junction, and Esophageal Adenocarcinoma. *Mod Pathol, 36*(5), 100154.
31. Homeyer, A., Geissler, C., Schwen, L. O., Zakrzewski, F., Evans, T., Strohmenger, K., Westphal, M., Bulow, R. D., Kargl, M., Karjauv, A., Munne-Bertran, I., Retzlaff, C. O., Romero-Lopez, A., Soltysinski, T., Plass, M., Carvalho, R., Steinbach, P., Lan, Y. C., Bouteldja, N., . . . Zerbe,

N. (2022). Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology. Mod Pathol, 35(12), 1759-1769.

국문요약

종양세포비율(NCP)을 추정하는 것은 분자 연구에 있어 매우 중요하다. 현재는 병리학자가 수작업으로 종양세포비율을 세고 있지만, 시간이 많이 걸리고 번거로운 작업으로 수행이 쉽지 않다. 이 문제를 해결하기 위해 34건의 요로암 스캔 이미지를 수집하여 컨볼루션 신경망 기반의 AI 모델들을 구축하였다. 외부 검증을 위해 118건의 추가 사례를 확보했으며, 다중 면역 형광(mIF) 검사를 통해 취득한 NCP를 기준값으로 사용하였다. 각 AI 모델은 0.82-0.88의 높은 급내 상관관계 계수(ICC)를 나타내며 강력한 신뢰성을 입증하였다. 병리의사와 비교했을 때 0.93의 ICC 값으로 일관된 결과를 나타냈다. 이러한 결과는 AI 모델의 알고리즘이 NCP 계산에서 병리의사를 효과적으로 보조할 수 있음을 시사한다.