A deep-learning workflow to predict upper tract urothelial cancer subtypes
supporting the prioritization of patients for molecular testing

Miriam Angeloni1,2,3, Thomas van Doeveren4, Sebastian Lindner1,2,3, Patrick
Volland1,2,3, Jorina Schmelmer1,2,3, Sebastian Foersch6, Christian Matek1,2,3, Robert
Stoehr1,2,3, Carol I. Geppert1,2,3, Hendrik Heers7, Sven Wach2,3,8, Helge Taubert2,3,8,
Danijel Sikic2,3,8, Bernd Wullich2,3,8, Geert J. L. H. van Leenders9, Vasily
Zaburdaev10,11, Markus Eckstein1,2,3, Arndt Hartmann1,2,3, Joost L. Boormans4, Fulvia
Ferrazzi1,2,3,5,†,*, Veronika Bahlinger1,2,3,†,*

1Institute of Pathology, University Hospital Erlangen-Nürnberg, Friedrich-Alexander-
Universität Erlangen-Nürnberg, Erlangen, Germany.
2Comprehensive Cancer Center Erlangen-EMN (CCC ER-EMN), Erlangen, Germany.
3Bavarian Cancer Research Center (BZKF), Erlangen, Germany.
4Department of Urology, Erasmus MC Urothelial Cancer Research Group,
Rotterdam, The Netherlands.
5Department of Nephropathology, Institute of Pathology, University Hospital
Erlangen-Nürnberg, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen,
Germany.
6Institute of Pathology, University Medical Center Mainz, Mainz, Germany.
7Department of Urology, Philipps-Universität Marburg, Marburg, Germany.
8Department of Urology and Pediatric Urology, University Hospital Erlangen,
Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Department of Pathology, Erasmus MC Cancer Institute, University Medical Centre, Rotterdam, the Netherlands.

Department of Biology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany.

Max-Planck-Zentrum für Physik und Medizin, Erlangen, Germany.

†These authors share co-senior authorship.

*Corresponding authors:

Fulvia Ferrazzi, PhD
Department of Nephropathology, Institute of Pathology
Friedrich-Alexander-Universität Erlangen- Nürnberg
Krankenhausstr. 8-10, 91054 Erlangen, Germany
E-mail: fulvia.ferrazzi@uk-erlangen.de; Tel. +49-9131-8543677

Veronika Bahlinger, MD
Institute of Pathology
Friedrich-Alexander-Universität Erlangen- Nürnberg
Krankenhausstr. 8-10, 91054 Erlangen, Germany
E-mail: veronika.bahlinger@uk-erlangen.de; Tel. +49-9131-8547791

Running title: Deep-learning to predict UTUC subtypes

Keywords: Deep-learning; Digital pathology; Immunohistochemistry; Molecular subtypes; Targeted therapy; Upper tract urothelial carcinoma; Whole-slide images
Abstract

**Background:** Urothelial carcinoma of the bladder (UBC) comprises several molecular subtypes, which are associated with different targetable therapeutic options. However, if and how these associations extend to the rare upper tract urothelial carcinoma (UTUC) remains unclear.

**Objective:** Identifying UTUC protein-based subtypes and developing a deep-learning (DL) workflow to predict these subtypes directly from histopathological H&E slides.

**Design, Setting, and Participants:** Subtypes in a retrospective cohort of 163 invasive samples were assigned on the basis of the immunohistochemical expression of three luminal (FOXA1, GATA3, CK20) and three basal (CD44, CK5, CK14) markers. DL model building relied on a transfer-learning approach.

**Outcome Measurements and Statistical Analysis:** Classification performance was measured via repeated cross-validation, including assessment of the area under the receiver operating characteristic (AUROC). The association of the predicted subtypes with histological features, PD-L1 status, and **FGFR3** mutation was investigated.

**Results and Limitations:** Distinctive luminal and basal subtypes were identified and could be successfully predicted by the DL (AUROC 95th CI: 0.62-0.99). Predictions showed morphology as well as presence of **FGFR3**-mutations and PD-L1 positivity that were consistent with the predicted subtype. Testing of the DL model on an independent cohort highlighted the importance to accommodate histological subtypes.
Conclusions: Our DL workflow is able to predict protein-based UTUC subtypes directly from H&E slides. Furthermore, the predicted subtypes associate with the presence of targetable genetic alterations.

Patient Summary: UTUC is an aggressive, yet understudied, disease. Here, we present an artificial intelligence algorithm that can predict UTUC subtypes directly from routine histopathological slides and support the identification of patients that may benefit from targeted therapy.
Introduction

Urothelial carcinomas (UCs) are malignant epithelial neoplasms arising from the urothelial lining of the urinary tract [1]. The rare upper tract UC (UTUC) represents 5-10% of all UCs, while the remaining are urothelial bladder cancer (UBC). UTUC is frequently associated with poor prognosis, with two-thirds of patients being diagnosed at an invasive tumor stage [2]. Owing to the histological similarity between UTUC and UBC [3, 4], and the preponderance of the latter, UTUC is an understudied disease.

Transcriptome-based subtyping allowed the stratification of muscle-invasive BC (MIBC) into different molecular subtypes [5-10], and in 2020 a consensus classification identified luminal and basal as the two distinctive subtypes [1, 11]. These subtypes offer valuable support for guiding targeted therapy options. Indeed, the luminal subtype appears associated with higher responsiveness to Fibroblast growth factor receptor 3 (FGFR3)-targeted therapies, and the basal subtype to immunotherapies such as programmed cell death ligand 1 (PD-L1) and programmed cell death protein 1 (PD-1) inhibitors [11, 12]. For UTUC, only few studies have so far investigated its genomic and transcriptomics landscape [13-15]. Additionally, no consensus subtypes have been identified yet [3, 16].

Assessment of molecular subtypes via high-throughput sequencing is neither ubiquitously available, nor cost-effective. Thus, alternative subtyping approaches should be considered. In MIBC, studies showed a substantial overlap between molecular and protein-based subtypes [17, 18], identifiable via more widespread, routinely applicable immunohistochemistry (IHC) analyses. Moreover, AI-based approaches have recently emerged as novel research tools able to provide automated and accurate pathological
diagnoses, leveraging the information residing into whole-slide images (WSIs) [19].

Here, we comprehensively characterize UTUC protein-based subtypes and propose a deep learning (DL) approach that can predict them relying on routine digitized H&E slides.

**Patients and methods**

**Patient cohorts**

The German cohort comprised formalin-fixed paraffin embedded material of N = 163 retrospectively analyzed UTUC patients diagnosed between 1995 and 2012 at University Hospital Erlangen-Nürnberg (Erlangen, Germany) and University Hospital Gießen and Marburg (Marburg, Germany). The Dutch cohort comprised N = 55 UTUC samples diagnosed between 2017 and 2020 as part of a multicenter, phase II, prospective trial conducted at University Medical Center Rotterdam (Rotterdam, The Netherlands) [20]. For both cohorts one whole slide image (WSI) per patient was used and all samples were invasive (i.e., with tumor stage ≥ pT1). All cases were systematically reviewed by two uropathologists (V.B., A.H.) according to the tumor, node, metastasis (TNM) classification (2017) and the classification of genitourinary tumors [21]. Clinico-pathological characteristics of the two cohorts are summarized in Table 1. The uropathologists also evaluated slides in terms of histological subtype, infiltration type, tumor type, and stroma content. Approval for this study was obtained from the Ethics Committee of the Friedrich-Alexander University Erlangen-Nürnberg (No. 329_16B) and the Erasmus Medical Centre Rotterdam (METC 2017-227 NL60919.078.17). All patients gave informed consent. The study was carried out in
accordance with the Declaration of Helsinki and to the principles of Good Clinical Practice (GCP).

**Clustering-based subtype assignment and statistical analyses.**

For protein-based subtyping, four tumor cores were punched for each patient and stained with three luminal (CK20, FOXA1, and GATA3) and three basal (CD44, CK14, and CK5) markers at the Institute of Pathology, University Hospital Erlangen, using tissue microarrays (TMAs). Expression was quantified using the histoscore (H-Score).

The same marker panel was used for IHC evaluations at the whole-slide level on selected cases. PD-L1 expression was assessed on TMAs using both immunoscore (IC) and combined positive score (CPS) [22], while FGFR3 mutational status was assessed using the SNaPshot method [23].

Hierarchical clustering and statistical analyses were performed within the R environment v.4.0.3 [24]. Association between categorical variables was assessed using Fisher’s exact test. To compare the distribution of continuous variables, the Wilcoxon rank-sum test for independent samples (two groups), the Kruskal-Wallis test (more than two), or the one-tailed Wilcoxon signed-rank test (paired samples) were used. Analyses of overall survival (OS) and disease-specific survival (DSS) were performed using the Kaplan-Meier (K-M) estimator and statistical differences assessed through the log-rank test. P-values < 0.05 were considered statistically significant.

**WSIs annotation and pre-processing**

Slides belonging to the two cohorts were digitized in the respective pathology centers using a Panoramic P250 scanner (3DHistech, Budapest, Hungary) at different resolution levels. For each WSI, tumor tissue was manually annotated in QuPath.
(v.0.2.3) by a trained observer (M.A.) under expert supervision (V.B.) (Supplementary Figure 1). An automated Python-based pipeline was implemented to tessellate the identified tumor areas into non-overlapping tiles of 512x512 pixel edge length, perform quality filtering, and stain-normalization (Supplementary Figure 2).

**Deep-learning algorithm and its validation**

Our DL framework relied on a transfer-learning approach by fine-tuning a ResNet50 [25] initialized with weights pre-trained on the ImageNet database [26]. To account for class imbalance, the number of tiles belonging to each class within the training set was equalized. A repeated three-fold cross-validation was used to estimate the model’s generalization accuracy and error. Here, to ensure independence between training and validation folds, random splitting was performed at patient level (Supplementary Figure 3). DL model’s prediction for single tiles of the validation set were averaged to obtain a WSI-level subtype prediction. For each repetition, the area under the receiver operating curve (AUROC), accuracy, precision, recall, and F1-Score were assessed as mean and 95% confidence interval (CI) across the three validation folds relying on Student’s t-distribution. Confusion matrices for a given repetition were obtained by concatenating the predictions for the three validation folds. The predicted class in the independent test cohort was taken equal to the one with the highest average prediction value across those obtained using the three trained models.

Further details for the methods are in the Supplementary Material.
Results

Hierarchical clustering of protein marker expression identifies luminal, basal, and indifferent UTUC subtypes

To characterize protein-based UTUC subtypes, the expression of three basal (CK5, CK14, CD44) and three luminal (CK20, FOXA1, GATA3) markers utilized in published UBC studies [17, 18] was assessed in a cohort of 163 invasive samples, referred to as ‘German cohort’. Hierarchical clustering of marker expression identified: a ‘luminal’ cluster (80 samples, 49%), a ‘basal’ cluster (42 samples, 26%), and an ‘indifferent’ cluster (41 samples, 25%) with low expression of both basal and luminal markers (Figure 1A). The basal samples exhibited shorter overall survival (OS) and disease-specific survival (DSS) than the luminal and indifferent samples (Figure 1B). Tumor stages differed across the three subtypes ($p = 0.01$), with almost half of pT4 samples in the basal group (Supplementary Table 1). Both infiltration ($p = 0.02$) and tumor type ($p = 0.02$) differed between basal and indifferent cases (Figure 1 C-D). In contrast, no significant differences were found when comparing luminal and indifferent subtypes. Indeed, basal samples showed a clear prevalence of diffusely infiltrative and non-papillary tumors, while the other two subtypes showed a higher proportion of pushing and papillary tumors. Also, luminal and indifferent subtypes were characterized by a comparable proportion of fibrotic and leiomyomatous samples, while the basal subtype had a prevalence of fibrotic stroma (Figure 1E).

Collectively, hierarchical clustering based on the IHC expression of six well-characterized UBC markers identified three distinctive protein-based UTUC subtypes, with high histological similarity of the indifferent to the luminal subtype.
Deep-learning successfully predicts luminal and basal subtypes and identifies candidate heterogeneous slides

We hypothesized that an AI approach could predict the identified subtypes from routinely available H&E slides. To test this hypothesis, 100,178 luminal, 66,770 basal and 57,874 indifferent tumor patches from the 163 annotated tumor areas of the German cohort were used to train and validate a DL-based classifier in a three-time repeated three-fold cross-validation setting (Supplementary Table 2; Supplementary Figure 4). While the model correctly predicted most basal (AUROC = 0.77; 95% CI, 0.67-0.86; repetition three) and luminal (AUROC = 0.71; 95% CI, 0.44-0.99; repetition two) samples, more than 55% of indifferent samples were predicted luminal. This difficulty of the DL model in predicting the indifferent subtype was consistent with the observed histomorphological similarity between luminal and indifferent subtypes.

Therefore, we decided to train the model focusing on the most distinctive luminal and basal subtypes, again relying on a repeated cross-validation setting (Figure 2A). Our classifier achieved a mean AUROC of at least 0.8 (AUROC = 0.83; 95% CI, 0.67-0.99; repetition one) and mean accuracy of ≥ 0.75 (mean accuracy = 0.79; 95% CI, 0.75-0.84; repetition two) (Supplementary Table 3). For further analyses, we focused on the results of repetition 2, as it showed the best accuracy and most consistent performance metrics across the three hold-out folds (Supplementary Table 3). First, the so called “high-confidence” slides were identified, i.e., those slides predicted luminal or basal with a score of at least 0.7, irrespective of their true subtype (Supplementary Table 4). The true positive rates in the high-confidence luminal and basal slides were respectively 86% (50/58) and 87.56% (14/16). Visual inspection of tile-level prediction
maps of the top high-confidence slides confirmed the pathological description of these
subtypes at histological level, i.e., dense nuclei with small stroma bridges as distinctive
feature of the luminal subtype (Figure 3A) and dense stroma and keratinization for the
basal subtype (Figure 3B) [1]. In the top scoring luminal slide, 99.9% of tiles were
predicted luminal. In the top basal, 90% of tiles were predicted as basal. These
predictions were confirmed by whole-slide level staining with the six luminal and basal
markers (Supplementary Figure 5 A-B), thus suggesting that our DL model can
successfully distinguish between luminal and basal subtypes.

Next, we focused on 22 ‘low-confidence’ predictions with prediction scores
between 0.4 and 0.6 (Supplementary Table 5). These slides showed no significant
difference in the distribution of luminal and basal marker expression (p = 0.43). Tile-
level prediction maps allowed categorizing these slides into: ‘heterogeneous slides’ with
and slides without distinguishable clusters of predicted luminal and basal tiles. Visual
inspection of a selected candidate heterogeneous slide supported the predictions, with
basal and luminal tiles showing the characteristic histological features (Figure 3C).
Furthermore, whole-slide IHC validation showed positivity of the entire tumor area for
the three luminal markers and the basal marker CK14 (Supplementary Figure 5C).
Notably, in the luminal-predicted area, the CK14 basal marker appeared expressed only
in the outer cellular layer, while in the basal-predicted area it appeared in all cell layers
(Figure 3C). Taken together, the DL model was able to identify candidate heterogeneous
slides showing co-presence of luminal and basal areas.
High-confidence predicted slides show the expected histology features and significantly associate with PD-L1 and FGFR3 status

To further characterize the high-confidence predictions, we examined also for them marker expression and morphological characteristics. High-confidence predicted luminal and basal slides showed higher expression of luminal \((p = 6.62\times10^{-9})\) or basal markers \((p = 0.002)\), respectively (Figure 4A).

High-confidence luminal predictions were mainly characterized by papillary tumors, with not otherwise specified (NOS) or histological subtype, and with pushing infiltration type. Instead, high-confidence basal predictions were mainly non-papillary tumors either squamous or with a histological subtype component and diffusely infiltrating (Figure 4B). Of note, five out of the eight wrongly predicted luminal slides were pushing papillary tumors with NOS histology, whereas both wrongly predicted basal slides were diffusely infiltrating, non-papillary tumors with variant histology. These results show that high-confidence predictions showed morphologic features consistent with the predicted subtype.

Next, we investigated the association of high-confidence slides with clinically-relevant biomarkers (Figure 4C). The proportion of PD-L1-positive samples was higher in high-confidence predicted basal slides, both using IC \((p = 0.01)\) and CPS \((p < 0.001)\). Vice versa, the proportion of FGFR3-mutated samples was higher in high-confidence predicted luminal slides \((p\text{-value}=0.002)\). Interestingly, one wrongly predicted basal slide was actually PD-L1-positive, whereas four wrongly predicted luminal slides were FGFR3 mutated.
External validation of the deep-learning model highlights the importance of subtype histology

To test the generalization ability of the DL model, an external cohort of 55 invasive UTUC patients, referred to as the ‘Dutch cohort’, was used. Hierarchical clustering identified also in this cohort luminal (31 samples, 56%), basal (4 samples, 7%), and indifferent (20 samples, 36%) subtypes, with expression profiles matching those in the German cohort (Supplementary Figure 6A).

The DL model correctly classified all luminal samples, with an average prediction score of 0.89, but not the four basal samples (Supplementary Figure 6B). Three of these samples showed histological subtypes with glandular, squamous and sarcomatoid features; one sample was instead predominantly papillary with NOS histology. However, in this fourth sample, basal features could be observed at the invasion front, where two out of four TMA cores were punched. Interestingly, the tile-level prediction map highlighted a small tumor area predicted basal in correspondence to the invasion front, whereas the remaining papillary area was predicted luminal (Supplementary Figure 6C). Thus, the difficulties in predicting the basal samples might be due to histological subtype.

Discussion

We identified three protein-based UTUC subtypes, which partially resembled UBC protein-based subtypes [17, 18]. The indifferent subtype, similarly to UBC double negative, presented low expression of both luminal and basal IHC markers, and would require further validation using more comprehensive methods such as RNA sequencing.
Our characterization in terms of infiltration type, tumor type, and stroma content, shed light into the histological similarity between the indifferent and luminal subtypes. This similarity was reflected in the performance of a three-class DL model, which predicted a large number of indifferent samples as luminal. Instead, we showed that a DL model could successfully discriminate between the most distinctive luminal and basal subtypes relying on routinely available H&E slides. Furthermore, the DL model identified candidate heterogeneous slides for which whole-slide IHC validation confirmed the co-presence of luminal and basal areas closely matching the tile-level predictions.

At histopathology level, invasive UCs present different morphological appearances [1]. As observed in Weyerer et al., MIBC histological subtypes are strong indicators of mRNA-/protein-based subtypes. Notably, squamous differentiation, characterized by intercellular bridges or keratinization, only occurs in the basal subtype, whereas conventional UCs (NOS) are mainly associated with the luminal subtype [17]. The high-confidence predictions by our model, even including those slides where the predicted subtype did not match the expression-based subtype, showed histological features consistent with the predicted subtype. Moreover, high-confidence predicted luminal and basal slides were significantly associated with FGFR3 and PD-L1, respectively. Because of the implementation of anti-PD-L1 and PD-1 therapies, and specific FGFR3 inhibitors [21, 27] in UCs, histopathology labs have been facing increased requests for PD-L1 assessment and FGFR3 genetic alteration testing. Our DL model thus offers a valuable support to pathologists for the prioritization of FGFR3/PD-L1 testing in UTUC patients. This would also contribute to extending
patients’ access to targeted therapies and improve their management and care in the clinical practice.

Several challenges were encountered during our study. Hierarchical clustering in the independent Dutch cohort highlighted the same biological tendency observed in the German cohort, which is even more remarkable considering the prospective nature of the former. This strongly supported the existence of three distinct UTUC subtypes. However, the cluster membership of single samples might vary with varying samples being clustered. This uncertainty in the training sample labels might negatively affect the DL model performance. An additional level of uncertainty in training labels was due to the assessment of marker expression in selected TMA cores. It is common practice to stain several tumors at once, while also taking into account tumor heterogeneity. However, expression in TMA cores might not be representative of whole-slide level expression, as clearly seen for the predominantly papillary case in the Dutch cohort with a basal-like morphology at the invasion front. RNA-sequencing analyses might offer a more robust subtype assignment, thanks to genome-wide profiling, yet might still fail to correctly characterize heterogeneous samples. Another challenge was linked to histological subtypes. Given the rarity of UTUC, we had decided not to exclude them, as instead done in previous UBC studies [18]. However, as the results on the Dutch cohort showed, this might have impaired model performance. This issue could be mitigated by the future availability of an even more extensive UTUC cohort.

Collectively, our results show that the most distinctive protein-based UTUC subtypes can be predicted directly from H&E slides and are associated with the presence of targetable genetic alterations. Thus, our study lays the foundation for an AI-
based tool to support UTUC diagnosis and extend patients’ access to targeted treatments.
References


Figure Legends

Figure 1 – Protein-based UTUC subtypes and their histopathological characterization. (A) Heatmap visualization of the hierarchical clustering analysis performed on the expression of the three basal (CD44, CK5, CK14) and three luminal (FOXA1, GATA3, CK20) markers in our UTUC German cohort (N = 163 invasive samples). Heatmap colors represent markers expression (standardized H-Score; red: high expression, blue: low expression). The color ribbon at the top of the heatmap indicates the three protein-based subtypes: luminal (red), indifferent (green), basal (blue). (B) Kaplan-Meier overall survival (top) and disease-specific survival (bottom) curves for the identified subtypes. P-values from log-rank test. (C-E) Characterization of the identified subtypes in terms of infiltration type (C), tumor type (D) and stroma content (E). On the left representative histology images and on the right bar plot distributions. The histology images are framed in the respective color used in the bar plot to the right. P-values from Fisher’s exact tests (p-values < 0.05 in italic bold).

Figure 2 – Deep-learning based prediction of luminal and basal protein-based subtypes. (A) Steps of the deep learning framework: (1) Starting from 80 luminal and 42 basal WSIs, a library made up of 100,178 luminal (red) and 66,770 basal (blue) stain-normalized tiles is generated using an automated, custom-developed, pre-processing pipeline. (2) The tiles library is used for training the network using three-fold cross validation (grey: training folds, yellow: validation fold). Tiles of the trained set are balanced between the two classes. The cross-validation is repeated three times. (3) For each training/validation set combination, a deep-learning model is trained using a
transfer-learning approach. For each tile, the model outputs a prediction value for the
luminal ($p_{\text{luminal}}$) and for the basal ($p_{\text{basal}}$) class (4). WSI-level predictions for the luminal
($p_{\text{WSI\ luminal}}$) and basal ($p_{\text{WSI\ basal}}$) subtypes are calculated by averaging the tile-level
predictions for each class. The subtype associated with the highest prediction is
assigned to the entire slide (4). In the schematization color intensity is proportional to
the prediction score. (B) AUROC for the three repetitions (with basal type as positive
class). The mean AUROC ± standard deviation across folds is reported for each
repetition. AUROC: area under the receiver operating characteristic; $N_L$: number of
luminal slides; $N_B$: number of basal slides.

Figure 3 – Whole-slide IHC-validation of deep-learning predictions. (A) Slide
predicted with the highest $p_{\text{WSI\ luminal}}$; (B) Slide with the highest $p_{\text{WSI\ basal}}$; (C) Candidate
heterogeneous slide. For all three slides, from left to right: digitized whole-slide with
annotated tumor areas (red); tile-level prediction map (red: luminal, blue: basal; intensity
dependent on prediction score); selected areas and corresponding whole-slide IHC-
validation using CK20 as representative luminal marker and CK14 as representative
basal marker. Whole-slide IHC-validation with all the six luminal/basal markers is
provided in Supplementary Figure 5.

Figure 4 – Characterization of the high-confidence predicted luminal and basal
slides. (A) Boxplot distributions of the mean expression values for the basal (CK14,
CK5, CD44) and luminal (CK20, FOXA1, GATA3) markers in the high-confidence
predicted luminal (left) and basal (right) slides. Shape represents the true slide label
(triangle: basal; square: luminal); color represents the prediction accuracy (grey: correct predictions; violet: incorrect predictions). P-values from one-tailed Wilcoxon signed-rank test. (B) Characterization of the high-confidence predicted slides in terms of tumor type (left), histology (middle), and infiltration type (right). An overview of the morphological features for the wrongly predicted slides is provided through the colored symbols above the bars. (C) Characterization of the high-confidence predicted slides in terms of clinical-relevant biomarkers: PD-L1 status (measured as CPS = combined positive score; IC = immunoscore) and FGFR3 mutational status. An overview of PD-L1 and FGFR3 status of the wrongly predicted slides is provided through the colored symbols above the bars. n: number of samples; p-values from one-tailed Fisher’s exact test.
**Tables**

**Table 1 - Clinico-pathological variables characterizing the German and the Dutch cohorts.** Estimates are given as median (minimum, maximum) or frequency (percentage) with respect to the total number of analyzed samples (N = 163 for the German cohort and N = 55 for the Dutch cohort). A hyphen (-) is used to indicate information/variable’s value not available within a given cohort.

<table>
<thead>
<tr>
<th>Clinico-pathological variable</th>
<th>German cohort</th>
<th>Dutch cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age at diagnosis (yrs), median (min-max)</strong></td>
<td>73 (47-94)</td>
<td>71 (52-85)</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>52 (31.9)</td>
<td>21 (38.2)</td>
</tr>
<tr>
<td>Male</td>
<td>111 (68.1)</td>
<td>34 (61.8)</td>
</tr>
<tr>
<td><strong>Histological Grade (WHO 1973), n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>G2</td>
<td>52 (31.9)</td>
<td>12 (21.8)</td>
</tr>
<tr>
<td>G3</td>
<td>111 (68.1)</td>
<td>36 (65.5)</td>
</tr>
<tr>
<td>Missing</td>
<td>0 (0%)</td>
<td>7 (12.7)</td>
</tr>
<tr>
<td><strong>Histological Grade (WHO 2004), n (%)</strong></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>high</td>
<td></td>
<td>54 (98.2)</td>
</tr>
<tr>
<td>low</td>
<td></td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Missing</td>
<td></td>
<td>1 (1.8)</td>
</tr>
<tr>
<td><strong>Histological Grade (WHO 2016), n (%)</strong></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>high</td>
<td>142 (87.1)</td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>21 (12.9)</td>
<td></td>
</tr>
<tr>
<td><strong>Primary Tumor, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pT1</td>
<td>33 (20.2)</td>
<td>17 (30.9)</td>
</tr>
<tr>
<td>pT2</td>
<td>28 (17.2)</td>
<td>13 (23.6)</td>
</tr>
<tr>
<td>pT3</td>
<td>81 (49.7)</td>
<td>25 (45.5)</td>
</tr>
<tr>
<td>pT4</td>
<td>21 (12.9)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Regional Lymph Nodes, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pN0</td>
<td>54 (33.1)</td>
<td>22 (40%)</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>pN1</td>
<td>16 (9.8)</td>
<td>3 (5.5%)</td>
</tr>
<tr>
<td>pN2</td>
<td>14 (8.6)</td>
<td>2 (3.6%)</td>
</tr>
<tr>
<td>Missing</td>
<td>79 (48.5)</td>
<td>28 (50.9%)</td>
</tr>
</tbody>
</table>

**Distant Metastasis, n (%)**

<table>
<thead>
<tr>
<th>cM0</th>
<th>79 (48.5)</th>
<th>54 (98.2%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cM1</td>
<td>8 (4.9)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Missing</td>
<td>76 (46.6)</td>
<td>1 (1.8%)</td>
</tr>
</tbody>
</table>
Supplementary Figure Legends

**Supplementary Figure 1 – Example of tumor annotation.** (left) Overview of a digitized whole slide image (WSI) with tumor tissue annotation borders in red; (right) zoomed-in tumor area representative of the employed annotation criteria: necrosis (1), bleeding/blood vessels (2), peri-tumoral lymphocytes (3) and scanning artifacts such as blurring and poor fixation (4) were excluded from the annotated area.

**Supplementary Figure 2 – Overview of the WSIs pre-processing workflow.** Graphical representation of the main steps performed by the automated Python-based pre-processing pipeline. A QuPath project (Project.qpproj) or a set of WSIs to process (provided via a csv file, or specified individually as a list of one or more WSI names) can be used as input for the pipeline. The pipeline performs the following steps: (1) The annotated tumor area (red) within each slide is tessellated into smaller tiles; (2) that the tiles undergo a quality-filtering step based on lower/upper thresholds identified on the log10-transformed median intensity values distribution; (3) tiles are stain-normalized according to the Macenko method.

**Supplementary Figure 3 – Overview of the training set generation procedure in a three-fold cross-validation setting.** In the example, each icon represents a patient (in our case corresponding to a slide, as only one slide per patient was used) of the analyzed cohort and the three colors (red, blue and green) represent three different classes of patients. The patients’ cohort is randomly split in three stratified partitions, i.e., with the same class distribution as in the whole cohort. Two partitions out of three
are used, in turn, to train the deep-learning model while the correspondent hold-out
folds are used for validation. Tiles belonging to the two training partitions are pooled
together and class. In the example tiles are represented by the square symbol and
inherit the same color (i.e., class) of the parent patient.

Supplementary Figure 4 – Deep-learning model performance in the prediction of
the luminal, basal, and indifferent subtypes. AUROC curves and confusion matrices
for (A) repetition 1, (B) repetition 2 and (C) repetition 3 of the three-fold cross-validation.
AUROC curves are shown for each subtype (blue: basal, green: indifferent, red:
luminal). The mean AUROC ± standard deviation (sd) is reported for each repetition.
Confusion matrices are normalized over the true class (row). AUROC: area under the
receiver operating characteristic.

Supplementary Figure 5 – Whole-slide IHC-validation with the entire markers
panel of the samples shown in Figure 3. Whole-slide IHC-validation with the three
basal (CK14, CK5, CD44) and three luminal (CK20, FOXA1, GATA3) markers for (A)
the top high-confidence predicted luminal slide, (B) the top high-confidence predicted
basal slide, and (C) a candidate heterogeneous slide.

Supplementary Figure 6 – Validation of the deep-learning model on the Dutch
cohort. (A) Heatmap visualization of the hierarchical clustering analysis performed on
the expression of the three basal (CD44, CK5, CK14) and three luminal (FOXA1,
GATA3, CK20) markers in the independent UTUC cohort from the Netherlands (N = 55
invasive samples). Heatmap colors represent markers expression (standardized H-Score; red: high expression, blue: low expression). The color ribbon at the top of the heatmap indicates the three protein-based subtypes: luminal (red), indifferent (green), basal (blue). (B) Confusion matrix with a summary of model’s generalization accuracy. (C) Selected basal slide predicted luminal by the model. From left to right: digitized whole-slide with annotated tumor areas (red); tile-level prediction map (red: luminal, blue: basal; intensity dependent on prediction score); zoomed area for a better visualization of the basal morphological features of the invasion front.
Supplementary Tables

Supplementary Table 1. Association analysis of the identified subtypes with the main clinico-pathological variables. Estimates are given as median (minimum, maximum) or frequency (percentage) calculated on the total available samples. F: female; M: male. Kruskal-Wallis and Fisher’s exact tests were used respectively for continuous and categorical variables. Statistically significant p-values are highlighted in italic bold.

<table>
<thead>
<tr>
<th>Clinico-pathological variable</th>
<th>Luminal</th>
<th>Indifferent</th>
<th>Basal</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Age at diagnosis (yrs), median (min-max)</em></td>
<td>72 (47-94)</td>
<td>70 (48-85)</td>
<td>77 (51-87)</td>
<td><em>0.02</em></td>
</tr>
<tr>
<td><em>Gender, n (%)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>24 (30)</td>
<td>9 (22)</td>
<td>19 (45.2)</td>
<td>0.07</td>
</tr>
<tr>
<td>M</td>
<td>56 (70)</td>
<td>32 (78)</td>
<td>23 (54.8)</td>
<td></td>
</tr>
<tr>
<td><em>Histological Grade (WHO 2016), n (%)</em></td>
<td></td>
<td></td>
<td></td>
<td>0.18</td>
</tr>
<tr>
<td>low</td>
<td>12 (15)</td>
<td>7 (17.1)</td>
<td>2 (4.8)</td>
<td></td>
</tr>
<tr>
<td>high</td>
<td>68 (85)</td>
<td>34 (82.9)</td>
<td>40 (95.2)</td>
<td></td>
</tr>
<tr>
<td><em>Primary Tumor, n (%)</em></td>
<td></td>
<td></td>
<td></td>
<td><em>0.01</em></td>
</tr>
<tr>
<td>pT1</td>
<td>20 (25)</td>
<td>10 (24.4)</td>
<td>3 (7.1)</td>
<td></td>
</tr>
<tr>
<td>pT2</td>
<td>19 (23.8)</td>
<td>4 (9.8)</td>
<td>5 (11.9)</td>
<td></td>
</tr>
<tr>
<td>pT3</td>
<td>36 (45)</td>
<td>21 (51.2)</td>
<td>24 (57.1)</td>
<td></td>
</tr>
<tr>
<td>pT4</td>
<td>5 (6.2)</td>
<td>6 (14.6)</td>
<td>10 (23.8)</td>
<td></td>
</tr>
<tr>
<td><em>Regional Lymph Nodes, n (%)</em></td>
<td></td>
<td></td>
<td></td>
<td>0.11</td>
</tr>
<tr>
<td>N0</td>
<td>31 (77.5)</td>
<td>11 (47.8)</td>
<td>12 (57.1)</td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>4 (10)</td>
<td>6 (26.1)</td>
<td>6 (28.6)</td>
<td></td>
</tr>
<tr>
<td>N2</td>
<td>5 (12.5)</td>
<td>6 (26.1)</td>
<td>3 (14.3)</td>
<td></td>
</tr>
<tr>
<td><em>Distant Metastasis, n (%)</em></td>
<td></td>
<td></td>
<td></td>
<td>0.57</td>
</tr>
<tr>
<td>M0</td>
<td>36 (87.8)</td>
<td>23 (95.8)</td>
<td>20 (90.9)</td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>5 (12.2)</td>
<td>1 (4.2)</td>
<td>2 (9.1)</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 2 – Performance metrics of the deep-learning classifier in the prediction of the luminal, basal, and indifferent protein-based subtypes for the three repetitions. For each repetition a measure of precision, recall, F1-Score, and AUROC is reported separately for the luminal, basal, and indifferent subtype together with the overall accuracy across the three groups. All performance metrics are indicated as mean [95% CI] across the hold-out folds. **AUROC**: area under the receiver operating characteristic; CI: confidence interval.

<table>
<thead>
<tr>
<th>Repetition</th>
<th>Basal</th>
<th>Luminal</th>
<th>Indifferent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1-Score</td>
</tr>
<tr>
<td>Repetition 1</td>
<td>0.43 [0.28-0.58]</td>
<td>0.59 [0.43-0.75]</td>
<td>0.11 [0.0-0.43]</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.43 [0.12-0.74]</td>
<td>0.74 [0.56-0.91]</td>
</tr>
<tr>
<td></td>
<td>F1-Score</td>
<td>0.43 [0.2-0.66]</td>
<td>0.65 [0.5-0.81]</td>
</tr>
<tr>
<td></td>
<td>AUROC</td>
<td>0.7 [0.5-0.9]</td>
<td>0.66 [0.58-0.74]</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repetition 2</td>
<td>0.51[0.38-0.64]</td>
<td>0.63 [0.44-0.82]</td>
<td>0.35 [0.15-0.56]</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.57 [0.1-1]</td>
<td>0.69 [0.57-0.8]</td>
</tr>
<tr>
<td></td>
<td>F1-Score</td>
<td>0.54 [0.24-0.83]</td>
<td>0.65 [0.59-0.71]</td>
</tr>
<tr>
<td></td>
<td>AUROC</td>
<td>0.75 [0.56-0.94]</td>
<td>0.71 [0.44-0.99]</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repetition 3</td>
<td>0.51 [0.37-0.64]</td>
<td>0.55 [0.5-0.6]</td>
<td>0.33 [0.0-0.7]</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.55 [0.45-0.65]</td>
<td>0.59 [0.02-1]</td>
</tr>
<tr>
<td></td>
<td>F1-Score</td>
<td>0.52 [0.46-0.59]</td>
<td>0.55 [0.23-0.88]</td>
</tr>
<tr>
<td></td>
<td>AUROC</td>
<td>0.77 [0.67-0.86]</td>
<td>0.66 [0.62-0.71]</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 3 – Performance metrics of the deep-learning classifier in the prediction of the luminal and basal protein-based subtypes for the three repetitions. For each repetition a measure of accuracy, AUC, sensitivity, specificity, precision, and F1-Score is reported for the basal subtype (here taken as the “positive” class) as mean [95% CI] across the hold-out folds. AUROC: area under the receiver operating characteristic; CI: confidence interval.

<table>
<thead>
<tr>
<th></th>
<th>Repetition 1</th>
<th>Repetition 2</th>
<th>Repetition 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.75 [0.52-0.99]</td>
<td>0.79 [0.75-0.84]</td>
<td>0.75 [0.48-1]</td>
</tr>
<tr>
<td>AUROC</td>
<td>0.83 [0.67-0.99]</td>
<td>0.8 [0.62-0.99]</td>
<td>0.81 [0.65-0.96]</td>
</tr>
<tr>
<td>Sensitivity (Recall)</td>
<td>0.67 [0.46-0.87]</td>
<td>0.6 [0.49-0.7]</td>
<td>0.6 [0.32-0.87]</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.8 [0.42-1]</td>
<td>0.9 [0.84-0.96]</td>
<td>0.83 [0.55-1]</td>
</tr>
<tr>
<td>Precision</td>
<td>0.67 [0.25-1]</td>
<td>0.76 [0.67-0.85]</td>
<td>0.66 [0.2-1]</td>
</tr>
<tr>
<td>F1-Score</td>
<td>0.66 [0.42-0.9]</td>
<td>0.67 [0.6-0.73]</td>
<td>0.62 [0.27-0.98]</td>
</tr>
</tbody>
</table>

Supplementary Table 4 – High-confidence predicted luminal and basal slides. The table lists all slides predicted luminal (sheet 1) or basal (sheet 2) with a prediction score higher or equal than 0.7. For each slide the true label (column “Subtype”), i.e., the label assigned by hierarchical clustering analysis, the prediction score for the basal (column “PredScore_Basal”) and luminal (column “PredScore_Luminal”) classes, the histological characterization (columns E-I), as well as the H-Score associated with the six markers panel, are reported. High-confidence predicted luminal and basal slides are sorted by descending values of PredScore_Luminal and PredScore_Basal respectively.

Supplementary Table 5 – Low-confidence predicted slides. The table lists all slides for which the prediction score for the basal class was in the range [0.4-0.6]. For each
slide the true label (column “Subtype”), i.e., the label assigned by hierarchical clustering analysis, the prediction score for the basal (column “PredScore_Basal”) and luminal (column “PredScore_Luminal”) classes, the histological characterization (columns E-I), as well as the H-Score associated with the six markers panel, are reported.
Author contributions

Conception and design: Angeloni, Ferrazzi, Bahlinger

Acquisition of data: Angeloni, van Doeveren, Volland, Schmelmer, Stoehr, Geppert, Heers, Wach, Taubert, Sikic, Wullich, van Leenders, Eckstein, Boormans, Ferrazzi, Bahlinger

Analysis and interpretation of data: Angeloni, Lindner, Ferrazzi, Bahlinger

Drafting of the manuscript: Angeloni, Ferrazzi, Bahlinger

Critical revision of the manuscript for important intellectual content: All authors

Statistical analysis: Angeloni, Ferrazzi

Obtaining funding: Eckstein, Hartmann, Boormans, Ferrazzi

Administrative, technical, or material support: None

Supervision: Hartmann, Ferrazzi, Bahlinger

Other: Foersch (resources), Matek (resources), Zaburdaev (resources)

Financial disclosures

Miriam Angeloni: No conflicts of interest

Thomas van Doeveren: No conflicts of interest

Sebastian Lindner: No conflicts of interest

Patrick Volland: No conflicts of interest

Jorina Schmelmer: No conflicts of interest

Sebastian Foersch: No conflicts of interest

Christian Matek: No conflicts of interest

Robert Stoehr: No conflicts of interest

Carol I. Geppert: No conflicts of interest
Hendrik Heers: No conflicts of interest

Sven Wach: No conflicts of interest

Helge Taubert: No conflicts of interest

Danijel Sikic: No conflicts of interest

Bernd Wullich: Speaker's honoraria from MSD and Janssen-Cilag

Geert J. L. H. van Leenders: No conflicts of interest

Vasily Zaburdaev: No conflicts of interest

Markus Eckstein: Personal fees, travel costs and speaker’s honoraria from MSD, AstraZeneca, Janssen-Cilag, Cepheid, Roche, Astellas, Diaceutics; research funding from AstraZeneca, Janssen-Cilag, STRATIFYER, Cepheid, Roche, Gilead; advisory roles for Diaceutics, MSD, AstraZeneca, Janssen-Cilag, GenomicHealth and Owkin.

Arndt Hartmann: honoraria for lectures for Abbvie, AstraZeneca, Biontech, BMS, Boehringer Ingelheim, Cepheid, Diaceutics, Ipsen, Janssen, Lilly, MSD, Nanostring, Novartis, Roche, and 3DHistotech; advisory role for Abbvie, AstraZeneca, Biontech, BMS, Boehringer Ingelheim, Cepheid, Diaceutics, Gilead, Illumina, Ipsen, Janssen, Lilly, MSD, Nanostring, Novartis, Qiagen, QUIP GmbH, Roche, and 3DHistotech; and research funding from AstraZeneca, Biontech, Cepheid, Gilead, Illumina, Janssen, Nanostring, Qiagen, QUIP GmbH, Roche, and STRATIFYER.

Joost L. Boormans: Consultancy (all paid to Erasmus MC): Janssen, Merck, MSD, BMS, AstraZeneca; Research support: Merck/Pfizer, Janssen, MSD.

Fulvia Ferrazzi: No conflicts of interest

Veronika Bahlinger: Personal fees, travel costs and speaker's honoraria from MSD; research funding from Gilead.
Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; SFB TRR 305 – Z01 to FF and Z02 to AH; SFB TRR 374 – INF1 to FF), by the Federal Ministry of Education and Research (BMBF) and the Dutch Cancer Society (KWF kankerbestrijding) in the framework of the ERA-NET TRANSCAN-2 initiative (to AH and JB), by the IZKF FAU Erlangen-Nürnberg (step 2 clinician scientist program to ME), by the Else Kröner-Fresenius-Stiftung (grant number: 2020_EKEA.129 to ME), by the Bavarian Center of Cancer Research (BZKF; Young Clinical Scientist Fellowship to ME), and the Federal Ministry of Education and Research (BMBF) 01KD2211B (ID: 100577017 to ME).

Acknowledgments

We are grateful to Natascha Leicht, Christa Winkelmann, and Verena Popp for excellent technical assistance in the laboratory.
Angeloni et al., Figure 1

A

Expression

Subtype

Luminal
Indifferent
Basal

B

Overall survival

Disease-specific survival

Probability

p=0.08
p=0.32

C

Infiltration type

Diffusely infiltrative
Infiltrative
Pushing

100 μm

p=0.02
p=0.87

D

Tumor type

Papillary
Non-papillary

1 mm

p=0.02
p=0.25

E

Stroma content

Myxoid
Leyomiomatous
Fibrotic

100 μm

p=0.11
p=0.15

medRxiv preprint doi: https://doi.org/10.1101/2023.06.14.23291350; this version posted June 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.
Angeloni et al., Figure 2

A

1) $N = 80$  
$N_B = 42$  
100,178 tiles  
67,770 tiles  

2) 3-fold CV

3) Training set balancing

4) $p_{luminal} = \sum_{i=1}^{N_B} p_{luminal, i} / N_B$
$p_{basal} = \sum_{i=1}^{N_T} p_{basal, i} / N_T$

if $p_{luminal} > p_{basal}$ then Basal, else Luminal

$N_T =$ number of tiles

B

Repetition 1  
Repetition 2  
Repetition 3

True positive rate

AUROC = 0.83 ± 0.05  
AUROC = 0.80 ± 0.06  
AUROC = 0.81 ± 0.05

False positive rate

False positive rate

False positive rate
Angeloni et al., Supp. Figure 1

2 mm

250 μm
Stratified random partitioning at patient level

Hold-out validation folds

Training folds

Training set balancing at tile level
Angeloni et al., Supp. Figure 4

A

Repetition 1

B

Repetition 2

C

Repetition 3

True positive rate

False positive rate

True subtype

Indifferent

Basal

Luminal

Predicted subtype

Basal

Indifferent

Luminal

0.43 0.26 0.31

0.27 0.049 0.69

0.15 0.11 0.74

0.17 0.14 0.69

Basal (AUROC = 0.70±0.07)

Luminal (AUROC = 0.66±0.03)

Indifferent (AUROC = 0.43±0.06)

Basal (AUROC = 0.75±0.06)

Luminal (AUROC = 0.71±0.09)

Indifferent (AUROC = 0.58±0.10)

Basal (AUROC = 0.77±0.03)

Luminal (AUROC = 0.71±0.09)

Indifferent (AUROC = 0.55±0.03)

Basal (AUROC = 0.75±0.06)

Luminal (AUROC = 0.71±0.09)

Indifferent (AUROC = 0.55±0.03)

CC-BY-NC-ND 4.0 International license
It is made available under a CC-BY-NC-ND 4.0 International license.
Angeloni et al., Supp. Figure 5

A

CK14  CK5  CD44  CK20  FOXA1  GATA3

5 mm

B

CK14  CK5  CD44  CK20  FOXA1  GATA3

5 mm

C

CK14  CK5  CD44  CK20  FOXA1  GATA3

5 mm
Angeloni et al., Supp. Figure 6

medRxiv preprint doi: https://doi.org/10.1101/2023.06.14.23291350; this version posted June 22, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.