Domain Adaptation-Based Deep Learning for Automated Tumor Cell (TC) Scoring and Survival Analysis on PD-L1 Stained Tissue Images

Ansh Kapil, Armin Meier, Keith Steele, Marlon Rebelatto, Katharina Nekolla, Alexander Haragan, Abraham Silva, Aleksandra Zuraw, Craig Barker, Marietta L. Scott, Tobias Wiestler, Simon Lanzmich, Günter Schmidt, and Nicolas Brieu

***Abstract*— We report the ability of two deep learning-based decision systems to stratify non-small cell lung cancer (NSCLC) patients treated with checkpoint inhibitor therapy into two distinct survival groups. Both systems analyze functional and morphological properties of epithelial regions in digital histopathology whole slide images stained with the SP263 PD-L1 antibody. The first system learns to replicate the pathologist assessment of the Tumor Cell (TC) score with a cut-point for positivity at 25% for patient stratification. The second system is free from assumptions related to TC scoring and directly learns patient stratification from the overall survival time and event information. Both systems are built on a novel unpaired domain adaptation deep learning solution for epithelial region segmentation. This approach significantly reduces the need for large pixel-precise manually annotated datasets while superseding serial sectioning or re-staining of slides to obtain ground truth by cytokeratin staining. The capacity of the first system to replicate the TC scoring by pathologists is evaluated on 703 unseen cases, with an addition of 97 cases from an independent cohort. Our results show Lin's concordance values of 0.93 and 0.96 against pathologist scoring, respectively. The ability of the first and second system to stratify anti-PD-L1 treated patients is evaluated on 151 clinical samples. Both systems show similar stratification powers (first system: HR = 0.539, p = 0.004 and second system: HR = 0.525, p = 0.003) compared to TC scoring by pathologists (HR = 0.574, p = 0.01).**

***Index Terms*— Deep learning, digital pathology, domain adaptation, oncology, PD-L1 biomarker.**

Manuscript received February 28, 2021; revised May 4, 2021; accepted May 9, 2021. Date of publication May 18, 2021; date of current version August 31, 2021. *(Corresponding author: Ansh Kapil.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Institutional Review Boards under Application No. NCT01693562, and performed in line with the Declaration of Helsinki and good clinical practice guidelines.

Ansh Kapil, Armin Meier, Katharina Nekolla, Tobias Wiestler, Simon Lanzmich, and Günter Schmidt are with Astrazeneca Computational Pathology GmbH, 80636 Munich, Germany (e-mail: ansh.kapil@astrazeneca.com; armin.meier@astrazeneca.com; katharina.nekolla@astrazeneca.com; tobias.wiestler@astrazeneca.com; simon.lanzmich@astrazeneca.com; guenter.schmidt@astrazeneca.com).

Keith Steele was with the Early Oncology Translational Medicine, Oncology Research and Development, AstraZeneca, Gaithersburg, MD 20878 USA (e-mail: steeleke@comcast.net).

Marlon Rebelatto was with the Precision Medicine and Biosamples, Oncology Research and Development, AstraZeneca, Gaithersburg, MD 20878 USA. He is now with the Early Oncology Translational Medicine, Oncology Research and Development, AstraZeneca, Gaithersburg, MD 20878 USA (e-mail: marlon.rebelatto@astrazeneca.com).

Alexander Haragan is with the Department of Molecular and Clinical Cancer Medicine, Royal Liverpool University Hospital, Liverpool L7 8XP, U.K. (e-mail: alex.haragan@nhs.net).

Abraham Silva, Aleksandra Zuraw, and Nicolas Brieu were with Astrazeneca Computational Pathology GmbH/Definiens GmbH, 80636 Munich, Germany (e-mail: abrahamsilvac@gmail.com; olkazuraw@gmail.com; nicolas.brieu@gmail.com).

Craig Barker and Marietta L. Scott are with the Precision Medicine and Biosamples, Oncology Research and Development, AstraZeneca, Cambridge CB2 0AA, U.K. (e-mail: craig.barker@astrazeneca.com; marietta.scott@astrazeneca.com).

Digital Object Identifier 10.1109/TMI.2021.3081396

I. Introduction

THE introduction of checkpoint inhibitor therapies in immuno-oncology has significantly improved the life expectancy of cancer patients. In particular in melanoma, non-small cell lung cancer (NSCLC) and bladder cancer, the survival benefit provided by therapies targeting the cell death protein (PD-1) or the programmed death ligand 1 (PD-L1) has led to a number of drug approvals which have changed the clinical routine. The human immunoglobulin G1 kappa monoclonal antibody durvalumab targets the immune escape of cancer by blocking the PD-L1 protein on tumor cells, and thereby promoting T-cell mediated tumor killing [1], [2]. Based on this mechanism of action it was hypothesized that the percentage of PD-L1 positive epithelial cells in tumor tissue samples is associated with response to therapy and overall survival. That hypothesis was confirmed in several studies and led to the development of a predictive histopathological scoring system, the Tumor Cell (TC) scoring, using PD-L1 stained tissue sections [3]. The definition of single cell positivity is determined by pathologist assessment of PD-L1 staining intensity on the membrane of the tumor cell. On slide level, the negative or positive PD-L1 status is determined by comparing the TC score to an assay specific cut-off value [4]; a TC score above the cut-off being indicative of cancers that are more likely to respond to therapy [5]. The TC scoring is typically performed by a pathologist viewing the PD-L1 stained histological tissue under a microscope. However, scoring of PD-L1 stained tissue is a challenging task [6]. The first challenge is that PD-L1 does not solely

stain the membrane of neoplastic epithelial cells. Immune cells (e.g. macrophages and lymphocytes), necrotic and stromal regions may show positive PD-L1 staining, but should not be included in the scoring. Also, regions showing tumor epithelial cells with cytoplasmic, but no membrane staining should be counted as negative. Another challenge associated with visual assessment is the difficulty for the human observers to estimate heterogeneous distributions of cell populations with spatially intermixed positive and negative tumor regions [7]. In some instances, these challenges make the TC scoring subject to variability among pathologists [7], which can lead to subjectivity in the therapeutic decision process.

The first contribution of this work is the introduction of an automated image analysis (IA) based TC scoring system, based on convolutional neural networks (CNN), that accurately replicates the pathologist scoring. Our results show high concordance between the IA algorithm and the pathologist scoring with respect to (i) the continuous TC scoring, (ii) the binary decision on the PD-L1 status at the 25% cut-point and (iii) the identification of a patient subgroup with improved NSCLC survival in the NCT01693562 clinical trial. We formulate the IA based TC scoring task as a three class segmentation problem - the three classes being (1) *TC(+)*: PD-L1 positive tumor epithelial regions ; (2) *TC(−)*: PD-L1 negative tumor epithelial regions; (3) *Other*: non-epithelial regions, which are not considered in the TC scoring such as immune, stromal and necrotic regions. This is followed by the calculation of the TC score as the ratio of the pixel counts (i.e. surface area) of the first regions to the union of the first and second regions: $TC_{score} = |TC(+)| / |TC(+) \cup TC(-)|$. Designed to reproduce the pathologist scoring, the proposed IA based TC scoring algorithm builds on extensive prior hypotheses such as a) the definition of cells that should be counted positive (or negative, respectively), b) the definition of the TC score from the segmented epithelial regions, and c) the cut-off value used for determining the PD-L1 status to perform patient stratification. While this system could assist the pathologists in taking more robust therapeutic decisions, it does not enable the discovery of novel stratification rules. This leads us towards the second contribution of this work, that is the introduction of a CNN-based survival analysis system to predict time-to-event outcome from PD-L1 stained histology images. Following the recent work by Mobadersany *et al.* [8], the proposed system automatically identifies visual patterns as well as the regions associated with low and high probability of being 'at risk'. Using hard attention (please refer to section II-B for more details on hard attention) on automatically segmented epithelial regions and patient-based stratified sampling, we are able to bypass some of the constraints associated with previous works on weakly supervised learning applied to histopathology images [8]–[11]. We show that the modifications enable the replication of TC score-based survival analysis without the need of either a large tissue dataset [9], [11] or restriction of the analysis to manually selected regions of interest (ROIs) [8] or to preselected small tissue samples such as tissue micro arrays (TMAs) [10]. Our approach enables, for the first time, the retrospective analysis by deep survival learning of a small clinical trial dataset consisting of needle biopsies and resected tissue samples.

The two above contributions build on the automatic segmentation of epithelial regions. For TC score replication, the score can be derived from the three-class segmentation problem. For survival analysis, it enables the attention to be focused on regions a-priori known to be information-rich. The aforementioned challenges associated with PD-L1 staining and TC scoring, together with the demonstrated performance of deep learning methods in digital pathology IA [9], [11]–[16] leads us towards this set of methods, and more particularly towards deep learning based semantic segmentation networks [17], [18]. Semantic segmentation networks have been successfully applied to a variety of histopathological image analysis tasks, though mostly in the H&E domain. For instance, Liu *et al.* [19] used Inception V3 network [20] to perform a patch based classification on whole slide images to produce course segmentation maps to detect cancer metastasis. To obtain dense segmentation results, Scheurer *et al.* [21] used a UNet network [18] with an efficient net B7 backbone [22] for classification of cutaneous lymphoma and eczema. To incorporate more contextual information from large histopathological images into the networks, Graham *et al.* [23] used multi-scale network to perform instance-based nuclei segmentation. Another idea to incorporate more contextual information was proposed by Rijthoven *et al.* [24], where they proposed multi-resolution networks to segment and classify various tissue regions in non-small cell lung cancer and breast cancer samples. While these methods provide accurate results, training of these networks requires, however, the use of large datasets with pixel-precise labels. In practice, the generation of such datasets is a manual process, and is therefore hampered by the associated high costs of data procurement as well as by the requirement of specialized expert annotators. To lower the need for manual training annotations, Mahmood *et al.* [25] proposed a method to synthesize H&E patches from segmentation masks using adversarial methods followed by training networks on them for nuclei segmentation. Chan *et al.* [26] proposed a weakly supervised method for semantic segmentation of histological tissue subtype.

Keeping the requirement of lowering the demand for manual annotations, we propose to automatically generate artificial training annotations using slides stained with the epithelial marker Pan-Cytokeratin (PanCK). PanCK is a pan-epithelial marker and has a good specificity of staining on different types of epithelium. Availability of PanCK stain enables easy and fast segmentation of epithelial regions using computer-based heuristics (Color deconvolution, Otsu thresholding followed by morphological operations). Human input is minimal on the PanCK sections, it being limited to a rough removal of macroscopic regions showing staining not specific to epithelial regions. The time consuming task of pixel precise labeling of the epithelium regions solely relies on computer-based heuristics, and thus large amounts of precise epithelial labels can be collected with relatively low effort. Therefore usage of PanCK as a helper stain for epithelial segmentation in this study was a natural choice. While the idea of using a

helper stain to generate training labels has been previously described [27]–[29], the methods relied on a training cohort designed such that the target stain (PD-L1) and the helper stain (PanCK) are both available on the same cases, either through the registration of consecutive slides [28] or the consecutive staining of the same slide [27], [29]. To bypass the need for either serial sections or subsequent staining, we exploit recent advances in deep generative adversarial networks (GANs), especially in unpaired image-to-image translation using CycleGAN [25], [30], [31]. We introduce an end-to-end trainable network (cf. Fig. 2c) named DASGAN (Domain Adaptation and Segmentation Generative Adversarial Network) that (i) jointly performs unpaired image-to-image translation and semantic segmentation and (ii) can leverage training annotations simultaneously from both the PanCK and PD-L1 stain domains despite a conflicting number of classes. The latter is a key characteristic of the proposed method: while PanCK staining provides unpaired cues for the two-class problem of epithelial segmentation, the automatic TC scoring algorithm additionally requires the classification of epithelial regions into PD-L1 positive or PD-L1 negative, yielding a three-class segmentation task.

II. Materials and Methods

This work introduces two novel deep learning-based image analysis (IA) methodologies. First the DASGAN network, which is an extension of the CycleGAN architecture [30] towards an end-to-end network for joint domain adaptation and segmentation. Second, an extension of the deep survival learning methodology [8] with hard-attention on epithelial regions.

A. The DASGAN Network

The DASGAN model builds on the existing CycleGAN model [30], which we recall here for completeness purposes. Two generators $G_{BA} : \mathcal{X}_B \rightarrow \mathcal{X}'_A$ and $G_{AB} : \mathcal{X}_A \rightarrow \mathcal{X}'_B$ are trained to synthesize samples in domain A (PD-L1) from real samples in domain B (PanCK) and vice versa. Two discriminators D_A and D_B are trained in opposition to identify synthetic from real samples in the two domains. The parameters of the two discriminator and two generator networks are learned in an adversarial manner following a minimax game on the two adversarial losses $\mathcal{L}^{AB}_{GAN}$ and $\mathcal{L}^{BA}_{GAN}$:

$$\min_{G_{AB}} \max_{D_B} \mathcal{L}^{AB}_{GAN} := \mathbb{E}_{x_B \sim \mathcal{X}_B} \log(D_B(x_B)) + \mathbb{E}_{x_A \sim \mathcal{X}_A} \log(1 - D_B(G_{AB}(x_A))) \quad (1)$$

$$\min_{G_{BA}} \max_{D_A} \mathcal{L}^{BA}_{GAN} := \mathbb{E}_{x_A \sim \mathcal{X}_A} \log(D_A(x_A)) + \mathbb{E}_{x_B \sim \mathcal{X}_B} \log(1 - D_A(G_{BA}(x_B))) \quad (2)$$

The necessity of having image pairs for image translation between A and B is bypassed using a cycle consistent loss $\mathcal{L}_{cycle}$ [30]. The cycle loss is defined to prevent mode collapse of the two GAN models and to constrain the invertibility of the translated domains, based on the translation of the synthesized samples $x'_B = G_{AB}(x_A)$ and $x'_A = G_{BA}(x_B)$ back to their original domains A and B:

$$\mathcal{L}_{cycle} := \mathbb{E}_{x_A \sim \mathcal{X}_A} \left\| x_A - G_{BA}(x'_B) \right\|_1 + \mathbb{E}_{x_B \sim \mathcal{X}_B} \left\| x_B - G_{AB}(x'_A) \right\|_1 \quad (3)$$

Following the auxiliary classifier generative adversarial network (AC-GAN) [32], we extend the CycleGAN model [30] to obtain segmentation maps as auxiliary from the two discriminator networks D_A and D_B operating on the domain A (PD-L1) and the domain B (PanCK). The proposed network is illustrated in Fig. 2. We condition the input images of the two generator networks G_{AB} and G_{BA}, which transform real PD-L1 images into translated PanCK images and real PanCK images into translated PD-L1 images, respectively, with the respective ground truth segmentation masks. To this end, the segmentation mask is concatenated with the original RGB image layers across the input image channel axis. The respective concatenated volumes go through a series of transformations by generators G_{AB} and G_{BA} to produce synthetic images in the respective target stain domains B and A. As a second extension, the two discriminator networks D_A and D_B are extended to predict pixel-wise class probability maps in addition to predicting the correct source of image. To this end, and to propagate the class specific information to the generator, a segmentation loss is introduced to the discriminator in addition to the original adversarial loss:

$$\mathcal{L}_{seg} := \mathcal{L}_{CE}(y_A^{true}, y_A^{pred}) + \mathcal{L}_{CE}(y_B^{true}, y_B^{pred}) \quad (4)$$

where $\mathcal{L}_{CE}(y^{true}, y^{pred}) = -\sum y^{true} \log(y^{pred})$ denotes the categorical cross-entropy loss and y^{true} and y^{pred} correspond to the ground truth and the predicted label maps, respectively. This results in the following loss for training the proposed DASGAN network:

$$\mathcal{L} := \mathcal{L}^{AB}_{GAN} + \mathcal{L}^{BA}_{GAN} + \lambda_1 \mathcal{L}_{cycle} + \lambda_2 \mathcal{L}_{seg} \quad (5)$$

with $\mathcal{L}^{AB}_{GAN}$, $\mathcal{L}^{BA}_{GAN}$ and $\mathcal{L}_{cycle}$ denoting the two adversarial and the cycle consistency losses [30] and $\lambda_1 = 10$, $\lambda_2 = 1$ weighting the losses associated with the cycle constraint and the segmentation auxiliary task, respectively. Only the discriminator D_A is employed at time of prediction but the use of a symmetric discriminator D_B ensures the balancing of the two counter-playing GAN networks.

The architectures of the two generators in the proposed DASGAN are similar to that in the original CycleGAN paper [30]. We included the following modifications. First, the input images and the segmentation mask are concatenated. Second, for the two discriminators, weights between the prediction of the source distribution and of the semantic segmentation posterior maps are shared in the first three convolutional layers and the branch for semantic segmentation extended to include three resnet blocks and three deconvolutional layers. Spectral normalization [33] and self-attention blocks [34] are added in the discriminators and generators to increase training stability and to model long structural dependencies respectively.

The resulting DASGAN network makes it possible to combine, at training time, annotations from any two stain domains

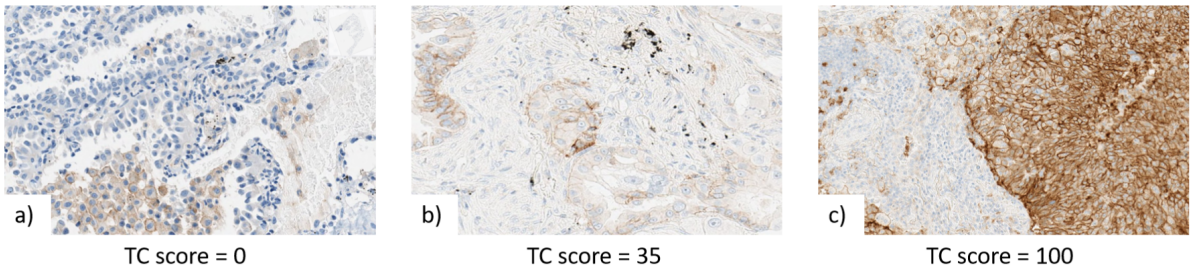


Fig. 1. Three examples of regions with different PD-L1 TC scores according to pathologist assessment. a) shows no PD-L1 staining in tumor cells, staining is only seen in macrophages which are not counted in the TC scoring b) shows a heterogeneous tumor staining in ≈35% of tumor cells and c) shows staining in ≈100% of tumor cells. In this work, we automate the scoring pipeline using novel domain-adaptation based deep learning method.

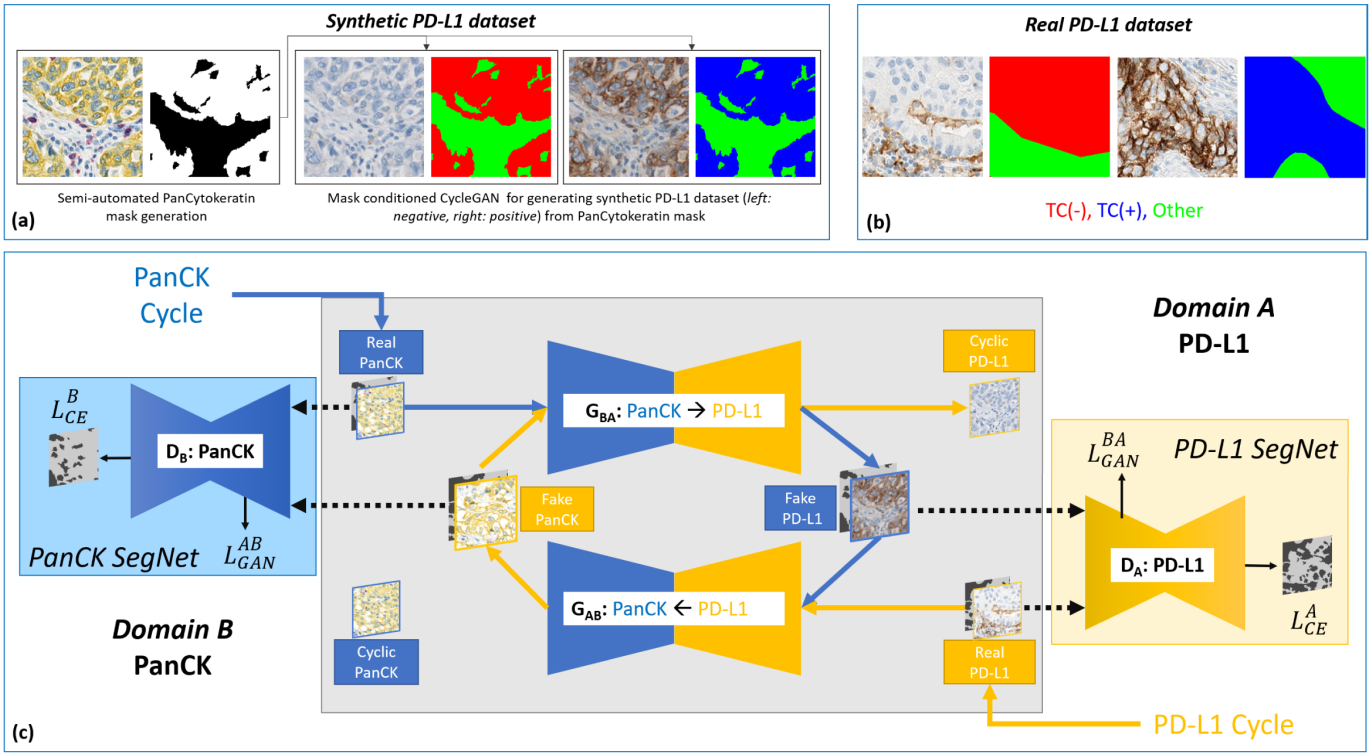


Fig. 2. Synthetic (a) and real (b) PD-L1 datasets generated from the semi-automated segmentation of PanCK images and manual annotations, respectively. (c) DASGAN model for joint domain adaptation and semantic segmentation. NB: the two cycle losses between the real and the cyclic images are not displayed for clarity purposes.

and independent cohorts. Given a dataset of PanCK stained images, large amounts of dense and pixel-precise ground truth data are created with a limited cost and burden of manual annotations. The above DASGAN network enables the binary segmentation of epithelial vs. other regions on PD-L1 images. To further differentiate between PD-L1 positive and PD-L1 negative epithelial regions, a three-class pixel-wise mask conditioning is introduced to DASGAN. Each PanCK binary segmentation mask is transformed into two examples of (i) a PD-L1 positive and (ii) a PD-L1 negative epithelial masks. After domain adaptation through G_{BA}, this results in two synthetic PD-L1 image versions of the same PanCK image (c.f. Fig. 2(a)): the first stained with PD-L1, the second not. Given a PanCK binary segmentation mask, a PD-L1 negative epithelial mask is built by giving the labels 0 and 1 to the non-epithelial and the epithelial regions, respectively. This conditions G_{BA} to yield a PD-L1 negative image. Similarly, a PD-L1 positive epithelial mask is generated from the same PanCK mask by giving the respective labels 0 and 2 instead, respectively. This conditions G_{BA} to yield a PD-L1 positive image from the same PanCK image.

To show the effectiveness of the proposed method, the segmentation network of DASGAN i.e. the discriminator's segmentation head is also trained as a standalone network (called "Segnet" is Fig.4 and Fig.5) and used as a baseline in results. The architectures have been kept exactly the same for fair comparison.

1) Dataset Description: The complete dataset was divided into three parts: 1) The training set used for model training,

2) the validation dataset used for model selection and 3) the unseen test dataset used to report performance.

The **training set** consists of $N_{PanCK} = 56$ PanCK (AE1/AE3 clone, Ventana [35]) stained WSIs of NSCLC samples and of $N_{PD-L1} = 69$ WSIs of the same indication and stained with the SP263 PD-L1 clone. The PanCK images and the PD-L1 images are unpaired and come from two independent patient cohorts. The good specificity of PanCK staining makes it possible to obtain a reliable semi-automatic segmentation of epithelial regions in PanCK images. More precisely this is done by using heuristics: color deconvolution, Otsu thresholding, and closing morphological operations (cf. Fig. 2a). Some human supervision is used in addition to discard macroscopic regions that are non-specifically stained (e.g. necrosis) or that are outside of the tumor regions (e.g. benign epithelium), thereby ensuring the purity of PanCK based training samples for epithelial regions. 194k patches were uniformly sampled from the PanCK images. ROIs on PD-L1 sections were manually annotated for the three classes of interest by pathologists - class 0: Non epithelial regions (including stroma, macrophages, immune cell clusters, necrotic regions etc.), class 1: PD-L1 unstained epithelial regions ($TC(-)$) and class 2: PD-L1 positively stained epithelium regions ($TC(+)$). For two class epithelial segmentation, class 1 ($TC(-)$) and class 2 ($TC(+)$) in the aforementioned ROI labels were combined. To study the impact of N_{PD-L1} on the segmentation accuracy, we report results with three different configurations for the training and validation sets: (i) 44K patches from $N_{PD-L1} = 22$ slides, (ii) 103K patches from $N_{PD-L1} = 49$ slides and (iii) 149K patches from $N_{PD-L1} = 69$ slides, all patches from (i) being included in (ii) and those of (ii) in (iii). All patches were 128×128 pixels sampled at $10\times$ resolution ($1\mu m$/px). The PanCK-based training set, as well as PD-L1 validation and test set remain unchanged in these experiments.

The **validation set** is similarly generated from another $N_{PD-L1} = 28$ partially annotated PD-L1 stained WSIs. The samples in the validation set are not used for model training, instead solely used for tuning the training hyper-parameters (e.g. learning rate, batch size) and for selection of the best model. The latter more precisely works as follows: (1) At the end of each training epoch, the model is recorded together with the corresponding accuracy metric (segmentation F1 score) computed on the validation set; (2) this leads, at the end of training, to a collection of model candidates and associated metrics; (3) the best model is selected among this collection of model candidates as the one maximizing the recorded metric. The selected best model is then applied on the unseen test set to report the final segmentation performance.

The **test set** for epithelial segmentation consisted of 106 regions of interest (ROI) of $500 \times 500 \mu m$ selected from 25 test whole slide images (WSIs) stained for PD-L1 with Ventana SP263 antibody. Similar to the training and validation sets, these ROIs were manually annotated for the three classes of interest by pathologists. The ROIs were selected to cover a high variability of different NSCLC sub-types (adeno, squamous), growth patterns (acinar, papillary and solid) and sample types (needle biopsies and resectates). The test WSIs were exclusively employed for evaluation purposes and were used neither for training the network nor selecting the best model hyper-parameters. The segmentation accuracy was measured for each class of interest on the unseen test set with the aggregated f1 score, which is defined as the harmonic mean between the aggregated precision and recall.

2) Network Training: Training and inference were performed using the TensorFlow library. All models were trained on a single NVIDIA V100 GPU with 32GB of memory and Adam optimization performed for both the generators (lr = 1e-4, beta1 = 0.5) and the discriminators (lr = 5e-4, beta1 = 0.5) for 150k iterations. Because the same architecture D_A is used by all networks for segmentation, the prediction time is the same for all networks: 0.08 sec for 512×512 pixels is measured on NVIDIA K80 GPU.

B. Deep Survival Learning

Deep survival learning is a paradigm of learning 'at risk' visual patterns in images directly from the time and event survival information using a CNN. To enable training of a neural network by back-propagation and handle survival data, Faraggi *et al.* [36] proposed a neural network based method to implement the Cox model in a loss function. Mobadersany *et al.* [8] recently extended on the idea proposed by Faraggi *et al.* [36] and use CNNs to compute the negative partial log likelihood associated with Cox model directly from images. Our work builds on this work, which we recall here for completeness. A patch based CNN is trained to predict 'at risk' value $\beta^T X$ derived from a Cox model, where X denote the realized values of the different covariates and β is a vector of linear coefficients. The CNN is trained by back-propagation to minimize the following negative partial log likelihood:

$$L(\beta, X) = \sum_{i \in U} \left(\beta^T X_i - \log \sum_{j \in \Omega_i} \exp(\beta^T X_j) \right) \quad (6)$$

where U is the set of samples with event and Ω_i is the set of samples with overall survival time higher than of the i^{th} sample. The loss is computed over all ROIs sampled from all cases [10], thereby the time and event information being propagated from the patient to each of its constituting ROI.

Survival learning in its original form, just like other deep learning applications, requires large datasets to learn meaningful 'at risk' patterns that can lead to robust predictions of high or low survivor groups. In case of smaller datasets, the problem of overfitting the model parameters to the data is often encountered. Also, since it is unlikely that every region in an image encodes information about the patient being 'at risk', propagating time and event indifferently to all possible ROIs yields training patches with low signal-to-noise ratio. To avoid the manual selection of ROIs, we employ a region-focused survival analysis, where the training ROIs are automatically sampled from the previously segmented regions of interest. This process of focusing the survival learning only on certain pre-selected regions, in this case epithelial regions, we refer to as "hard attention". To avoid a disproportionately large number of patches coming from tissue resections versus smaller needle biopsies during training, we introduce a stratified sampling

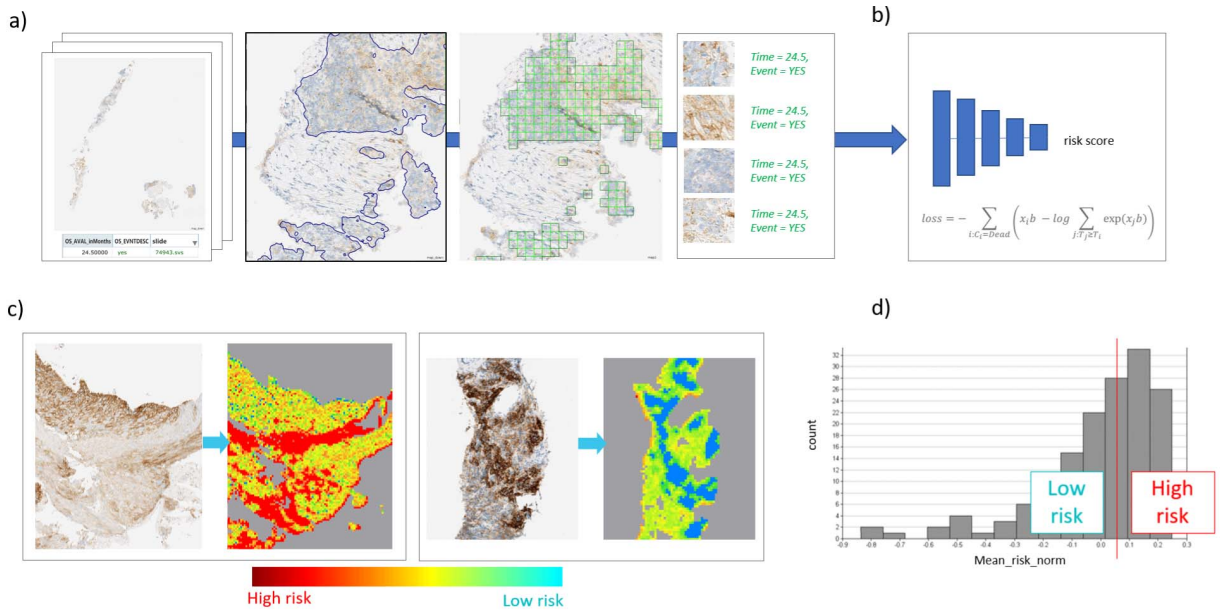


Fig. 3. The overview of region focused survival CNN workflow. a) Densely sample ROIs from DASGAN-detected epithelial regions and label each sample with patient overall survival information $\{time + event\}$. b) Train a CNN using Cox survival loss [8]. c) Generate a risk map for each patient by predicting with the trained network using a sliding window approach and z-score standardization of the resulting risk values. d) For each patient, aggregate the normalized risk map into a single risk score by averaging the individual normalized risk values of all ROIs in epithelial regions. Stratify the patients by the median of their aggregated risk scores.

scheme where a random subset of up to 10k patches are sampled per image. The training procedure remains otherwise similar to that of the original survival model. The resolution of the input patches is standardized to a magnification of $10\times$ i.e. $1\mu/px$. The details of the approach are shown in Fig. 3.

The network used in this work is relatively shallow and consists of three convolutional blocks. Each block is built of a 3 convolutional layer (with 2, 4, 8 filters, respectively) interleaved with a batch normalization layer, a ReLu activation and a 2×2 maximum pooling layer. These are followed by a fully connected layer of 8 nodes, a dropout layer ($r = 0.4$) and a linear activation layer. Input patches are 58×58 pixels and are augmented for brightness, saturation and gamma correction. The system was applied in a two-fold procedure, the first model trained the first fold being applied on the second fold and the second model trained on the second fold being applied on the first fold. For both folds, training was performed using the adam optimizer ($lr = 1e{-}5$, $\beta_1 = 0.5$, $\beta_2 = 0.9$) for 200 epochs. The split between the first and the second folds was provided externally to this analysis. The network was applied on the respective other fold using a sliding window approach, thereby generating low resolution risk score maps for each patient.

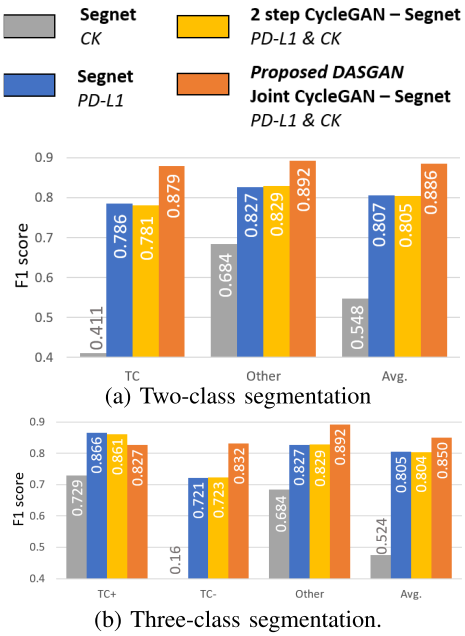
For each fold independently, the risk score values were z-score standardized to zero mean and unit standard deviation. This standardization allowed us to standardize survival risk values between the two analysis folds. For each patient, the obtained standardized risk score values within the detected epithelium regions were finally averaged, yielding one risk score per patient. Patients are finally split based on the cohort median of their aggregated risk scores into low and high risk subgroups, respectively. For the baseline comparison, the same steps are repeated without considering the hard attention on epithelial regions. Cox regression analysis is performed on the resulting stratification.

III. Results

A. Epithelial Segmentation

The segmentation accuracy was studied as a function of the availability of manually annotated PD-L1 images for training. With a limited number of manual annotations (cf. Fig. 4a), the proposed DASGAN model outperformed all the baseline models, i.e. (i) the semantic segmentation model trained solely on real and manually annotated PD-L1 images; (ii) the semantic segmentation model trained solely on the PD-L1-translated and automatically annotated PanCK images; and (iii) the two-step domain adaptation and semantic segmentation model trained on the real PD-L1 images and on the PD-L1-translated PanCK images. The DASGAN and the three baseline models are detailed in the Methods (section II) sections.

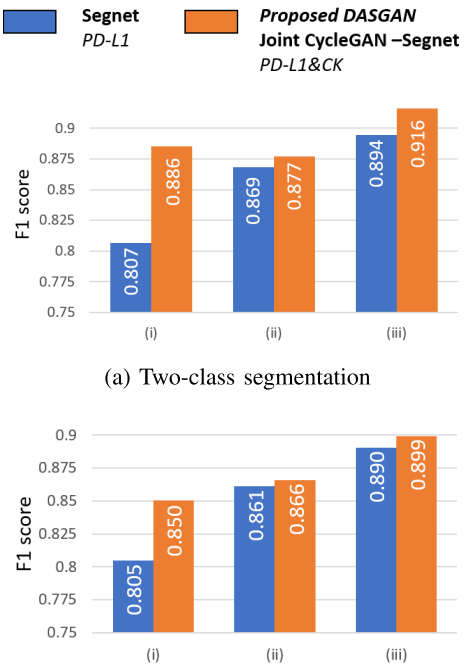
1) Two Class - Epithelial vs Non-Epithelial Segmentation: Mean f1 score of $f_1 = 0.886$ is reported for the proposed network. The two-step model did not improve the segmentation results ($f_1 = 0.805$) compared to the training on sole real PD-L1 images ($f_1 = 0.807$). Training on PD-L1-translated PanCK images only did not yield accurate segmentation results ($f_1 = 0.548$). As shown in Fig. 5a, the more manual annotations were available for training, the more the difference between the DASGAN and the best of the three baseline models decreases. In case of highest data availability, f1 scores of $f_1 = 0.894$ and $f_1 = 0.916$ were achieved by the baseline and the DASGAN models, respectively.



(a) Two-class segmentation

(b) Three-class segmentation.

Fig. 4. Segmentation accuracy on the unseen test ROIs (N = 106), for the three baseline models (in gray, blue and yellow) and for the proposed DASGAN (in orange) under the condition (i) of low availability of manual annotations on real PD-L1 images. F1 scores are reported for each class of interest - epithelial (TC), epithelial positive (TC(+)), epithelial negative (TC(−)) and non-epithelial regions (Other), together with their average score Avg. in both scenarios of (a) epithelial detection and (b) replication of TC score.



(a) Two-class segmentation

(b) Three-class segmentation.

Fig. 5. Segmentation accuracy (avg. f1-score) on the unseen test ROIs (N=106), for the best of the baseline models trained only on real PD-L1 samples (blue) and the proposed DASGAN model trained on both real and PD-L1-translated PanCK samples (orange), for increasing availability (i)-(ii)-(iii) of manual annotations on real PD-L1 images. Both scenarios of (a) epithelial detection and (b) replication of TC score are reported.

2) Three-Class Segmentation: In this three-class configuration, DASGAN enables the segmentation of PD-L1 positive and PD-L1 negative epithelial regions. We used the same test ROIs as mentioned above, with the $TC(+)$ and $TC(-)$ regions given different classification labels. In case of relative shortage of manual annotations, DASGAN yielded a segmentation f1 score of $f_1 = 0.850$ averaged over the three classes of interest, i.e. here (i) PD-L1 positive epithelial (TC(+)) vs. (ii) PD-L1 negative epithelial (TC(-)) vs. (iii) non-epithelial regions. F1 scores of $f_1 = 0.805$ and $f_1 = 0.807$ were reported for the two-step and the PD-L1 baseline methods, respectively (cf. Fig. 4b). As above, DASGAN systematically outperformed the best of the baseline approaches across increasing availability of labeled PD-L1 training images, reaching a maximum of $f_1 = 0.899$ (cf. Fig. 5b). Fig. 6 provides a qualitative example of epithelial segmentation produced by DASGAN.

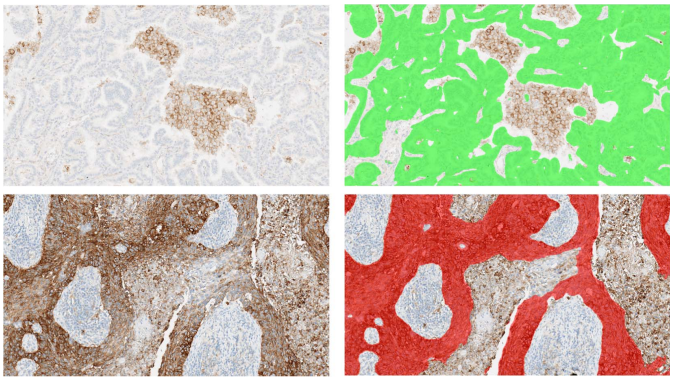


Fig. 6. Example of PD-L1 negative (green) and PD-L1 positive (red) epithelial regions segmented by the proposed DASGAN model. Notice the PD-L1 positively stained macrophage clusters (top row) and the PD-L1 positively stained necrotic regions (bottom row) are excluded from analysis automatically by the DASGAN model, and hence excluded from TC scoring.

B. Tumor Cell Scoring

In addition to the above analytical study, we quantitatively assessed the clinical relevance of the proposed IA based TC scoring methodology on a set of 703 PD-L1 stained WSIs that, similarly to the above test WSIs, were neither used for training nor model selection nor hyper-parameter optimization. The n = 703 slides comprises of 3 patient cohorts, two of them (with n = 434 and n = 118 unseen samples, respectively) consists of patients that received standard of care (SOC) treatment and one cohort (with n = 151 unseen samples) consists of Durvalumab treated patients. Because this clinical test set originated from the same three patient cohorts as the PD-L1 WSIs used for training and hyper-parameter optimization, an independent clinical validation cohort with 97 PD-L1 stained slides was acquired after freezing of the developed TC scoring algorithm. This set consisted of mostly resectates obtained from patients that received standard of care treatment (chemo/radio therapy followed by surgery) treatment. Results on this independent cohort provided an unbiased performance estimate of the proposed image-based TC scoring algorithm. Fig. 7 shows the bar plot of mean and standard deviation of the image analysis TC scores against

TABLE I
VALIDATION OF AUTOMATED PD-L1 TC SCORES AGAINST PATHOLOGIST TC SCORES

Cohort	Lin's CC	Pearson CC	Mean absolute error	OPA	PPA	NPA
Development cohort (N=703 unseen slides)	0.93	0.94	7.3	0.92	0.91	0.93
Independent validation cohort (N=97 unseen slides)	0.96	0.96	6.24	0.95	0.88	1.0

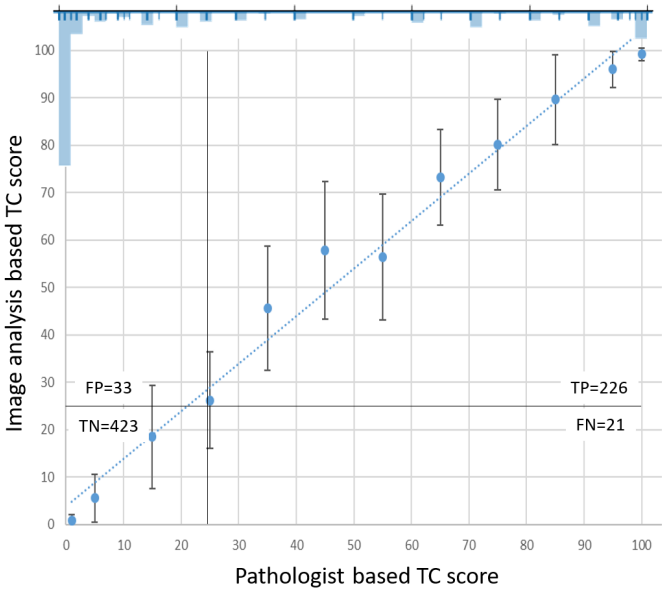


Fig. 7. Bar plot showing, on the first validation set (N = 703), the mean and standard deviation of the DASGAN based TC scores. The inverted histogram at the top shows the relative distribution of cases with respect to the pathologist based TC score. The dotted blue line is the regression line showing the overall trend of image analysis based TC score to the pathologist based TC score. The vertical and horizontal lines illustrate the 25% cut-off deciding on the PD-L1 positive or PD-L1 negative status. Resulting true and false positive classifications as well as true and false negative classifications are reported between the image based algorithm and the pathologist based TC scores.

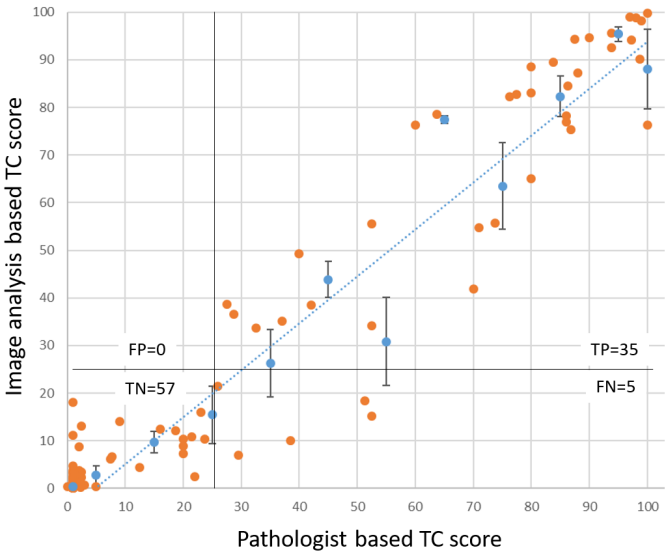


Fig. 8. Scatter plot (orange dots) between the pathologist based and the DASGAN based TC scores on a second and independent dataset (N = 97). The dotted blue line is the regression line showing the overall trend of image analysis based TC score to the pathologist based TC score. The vertical and horizontal lines illustrate the 25% cut-off deciding on the PD-L1 positive or PD-L1 negative status. The bars (with blue centers) show the mean and standard deviation of the DASGAN based TC scores. Resulting true and false positive classifications as well as true and false negative classifications are reported between the image based algorithm and the pathologist based TC scores.

the pathologist-based TC scores on the first clinical image set. Lin's concordance coefficient of $Lcc = 0.93$, Pearson correlation coefficient of $Pcc = 0.94$ and mean absolute error of $MAE = 7.30$ are reported between the estimated and the true TC score values. Applying the cut-off value of 25% on the respective TC scores, we obtained an Overall Predictive Agreement of $OPA = 0.92$, a Positive Predictive Agreement of $PPA = 0.91$ and a Negative Predictive Agreement of $NPA = 0.93$. Fig. 8 displays the scatter plot of the IA based and the pathologist based TC scores on the independent clinical validation cohort. We found a similarly high concordance between the pathologist and the algorithm, with a Lin's concordance coefficient of $Lcc = 0.96$, a Pearson correlation coefficient of $Pcc = 0.96$, a mean absolute error value of $MAE = 6.24$, and the following agreement values on the PD-L1 status: $PPA = 0.88$, $NPA = 1.0$ and $OPA = 0.95$. Table I summarizes these results.

C. Survival Analysis

Rebelatto *et al.* [5] showed that a cut-off of 25% of tumor cells with PD-L1 membrane staining of any intensity best discriminated responders from non-responders. We apply the same criteria for determining the PD-L1 status with automated PD-L1 scoring by DASGAN. Survival information for Overall Survival (OS) was available for one of the three initial patient cohorts, more precisely on the set of (N = 163) core needle biopsies and tissue resections from the NCT01693562 clinical trial (NSCLC). Survival analysis was performed solely on the subset of patients, whose tissue samples remained unseen to the training of DASGAN model (N = 151). Patients were stratified using the 25% cut-off value on the pathologist based TC score and the DASGAN based TC score. Performing a Cox regression analysis resulted in hazard ratios and associated p-values of $HR = 0.574$, $p = 0.01$ and $HR = 0.539$, $p = 0.004$, respectively. Here, the hazard ratios significantly smaller than 1 relate to a reduced risk of seeing an event (Event = Death in case of Overall Survival) in the biomarker positive cases as compared to the biomarker negative ones. Applying the proposed end-to-end survival learning methodology without hard attention on epithelial regions resulted in a hazard ratio of $HR = 0.762$ and p-value of $p = 0.199$. Applying the same end-to-end survival learning methodology with hard attention on the DASGAN-segmented epithelial regions resulted into $HR = 0.525$ and $p = 0.003$. These values are summarized in Table II. Associated Kaplan-Meier curves are displayed in Fig. 9. Figures 9 e),f) show comparison of automated TC scores from DASGAN and automated risk

TABLE II

SURVIVAL ANALYSIS FOR OS RESULTS USING THE TC SCORE BASED METHODS (MANUAL PATHOLOGY BASED AND AUTOMATED) AND DEEP SURVIVAL LEARNING BASED (NON-FOCUSED AND FOCUSED). ** SPECIFIES SIGNIFICANCE (HR P-VALUE ≤ 0.01)

Method	Criteria for grouping	Hazard ratio	HR P-value
Manual TC scores (median by 3 pathologists)	TC score ≥ 25	0.574	0.01**
Automated TC scores from DASGAN	TC score ≥ 25	0.539	0.004**
Deep survival learning (non focused)	median risk score	0.762	0.199
Deep survival learning (focused)	median risk score	0.525	0.003**

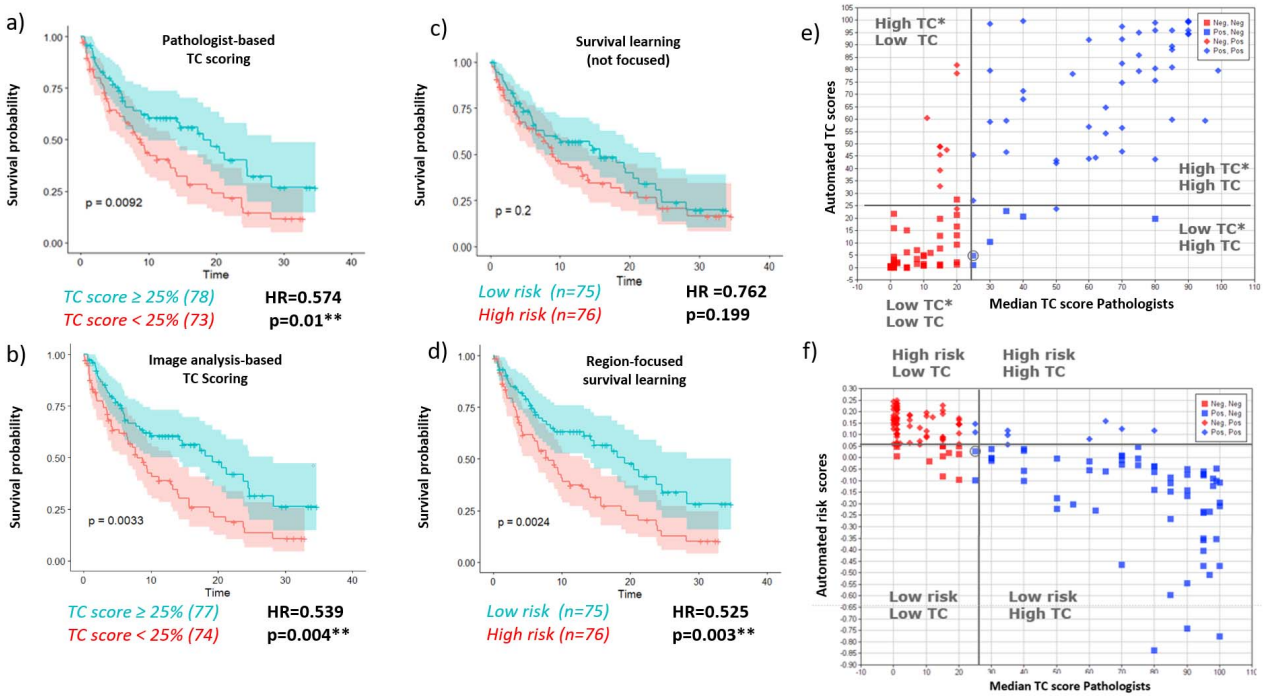


Fig. 9. Kaplan-Meier curves with low and high surviving groups for OS obtained by a) pathologist-based TC scoring b) automated TC scoring c) non-focused survival learning and d) newly proposed region-focused survival learning. We show that the IA based survival analysis, both automated TC score based as well using region-focused survival analysis, gives patient stratification as good as obtained from assessment by pathologists. e), f) show comparison of automated TC scores from DASGAN and automated risk scores from region-focused survival net in grouping patients suitable for anti PD-L1 treatment.

scores from region-focused survival net in grouping patients suitable for anti PD-L1 treatment. We can clearly see that low TC score corresponds to high risk scores and vice versa. For instance, in e) the group **Low TC*Low TC** corresponds to the group where both the automated TC scores as well as the median pathologist scores suggests that the patient belongs to the group showing low PD-L1 expression. A similar group can be seen in f) with high risk and low median TC scores by the pathologists. Both groups correspond to the group of patients that are less likely to respond to anti PD-L1 therapy. Objective metrics like OPA ($OPA_{TC} = 0.8874$, $OPA_{Risk} = 0.8675$), PPA ($PPA_{TC} = 0.9041$, $PPA_{Risk} = 0.8767$) and NPA ($NPA_{TC} = 0.8717$, $PPA_{Risk} = 0.8589$) shows the similarity in the groups obtained by both TC score based and risk score based methods, respectively.

IV. DISCUSSION

We confirm that the IA based PD-L1 TC scoring system matches the predictive ability of the pathologist TC scoring. The concordance was shown on a large number of slides (Table I) that suggests the robustness of the proposed system. We also show that similar predictive ability can be achieved by training a shallow CNN in an end-to-end manner directly on the overall survival information with the condition that the attention of this CNN is focused on epithelial regions only. While this latter finding should still be confirmed on an independent cohort, the current results show the ability of the proposed hard-attention methodology to perform further hypothesis-free exploratory analysis of small clinical datasets, with the potential of discovering novel predictive biomarkers.

A natural extension to the region focused survival CNNs would be to perform a study to find region-focused hypothesis-free biomarkers from other tissue areas e.g. in tumor-associated stroma. This would give us more insights about the role different immune cell populations have on survival which would be complementary to the features that we derive from epithelial regions. The applicability of this method is not limited to lung cancer, but can be applied to various cancer types and stains which allow for the creation of new biological insights as well as suggestions for therapeutic targets. In addition to the image data, other clinical variables like age,

smoking status and other omics data could be integrated into the CNN.

There are certain limitations to the proposed DASGAN method in this study which we would like to address in the future. In the current setup, the concordance of automated TC score is shown against only one pathologist for both development and validation cohorts. An interesting direction of work will be to compare the inter-observer variability of the PD-L1 TC scoring in a multi-pathologist setting where multiple pathologists will score the same slides. This will help us understand in which samples pathologists are more likely to agree/disagree on TC scoring. As also mentioned in the introduction and motivation of this work, the heterogeneous distributions of cell populations with spatially intermixed positively and negatively stained tumor regions are often difficult to score and lead subjectivity in the scoring [7]. A slight trend for this observation can be seen in the mid TC score range (25%-75%) where automated scoring has the most standard deviation against the pathologist scoring. These cases are more likely to be heterogeneous in terms of spatial distribution of PD-L1 stained tumor cells. Another limitation of the study is that the DASGAN model for epithelium segmentation segments all kinds of epithelium i.e. there is no differentiation between malignant and benign epithelium compartments. Due to this limitation, imprecision in the reported TC scores might occur, especially in the cases where the benign epithelium shows different staining than malignant epithelium. This discordance is more likely to happen in resections since there are chances of having more benign epithelium than biopsies where the amount of benign epithelium is relatively low ($<10\%$). This limitation can either be solved by manually delineating malignant cancerous "tumor core" regions or by extending the model to automatically separate the benign epithelium from the malignant ones. The latter can be done using helper stain for benign epithelium in conjunction with PanCK.

In this study, the DASGAN as well as the region-focused survival net were developed and validated on PD-L1 SP263 assay. We would expect that the proposed method could be applied to other PD-L1 assays as well, like the 22C3 and 28-8 Dako PD-L1 assays. Currently we do not have any data to support this claim. We expect so because, first, the estimation of the PD-L1 status in these clones also depends solely on estimating the PD-L1 expression on tumor cells and second, because these two assays appear relatively similar in staining pattern with the SP263 assay [37]. However, since each PD-L1 clone is associated with a different clinically relevant cut-off value, the determination of the patient status would have to be adapted according to their respective guidelines. For the Ventana SP142 clone, the methodology in the current form would not be directly applicable since i) staining of tumor cells with the Ventana SP142 assay has been shown to be less concordant [37] (e.g. to stain fewer tumor cells) and ii) the decision on the PD-L1 status is not solely based on TC score but also on the percentage of tumor area occupied by PD-L1 expressing tumor-infiltrating immune cells [38].

The applicability of DASGAN is not limited to PanCK helper stain. The same methodology can be used with other helper stains CD3/CD8/CD20 (for lymphocyte cell populations), CD68 (for macrophage population). Furthermore, DASGAN can be used with different histopathological image modalities (e.g. H&E and multiplexed immunofluorescene).

The advantage of using such a system is that the results are always reproducible as compared to a human rating which might be subjective. Additionally, the segmentation results might be used as an assistance to the pathologists to make a more informed diagnosis. With further validation of the proposed DASGAN for PD-L1 TC scoring and the future extensions that takes into account automatic removal of benign epithelium from TC scoring, this system can potentially become a companion diagnostic (CDx) to prospectively select patients that might benefit from PD-L1 checkpoint inhibitor therapy.

Author Contribution Statement

N.B., A.K., A.M. and G.S. designed the study. A.K., N.B, A.M. and K.N. developed the two image analysis-based systems for automated TC scoring and end-to-end survival learning. T.W. and S.L. developed the heuristic-based analysis of the PanCK-stained images. M.L.S and C.B. provided PD-L1 stained images. A.Z. provided manual region annotation for training the DASGAN. A.S. provided TC scores, in particular these on the independent patient cohort used for validating the image-based TC scoring system. K.S. and M.R. provided the PD-L1 stained images used for survival analysis. A.H. provided the independent patient cohort used for evaluating the image-based TC scoring algorithm. N.B. and A.K wrote the manuscript. All authors reviewed the manuscript.

Data and Code Availability

The code can be made available on reasonable request for all deep learning models mentioned in this work. The data used in this work is sensitive patient data which is confidential and hence cannot be made available.

References

[1] W. Zou, J. D. Wolchok, and L. Chen, "PD-L1 (B7-H1) and PD-1 pathway blockade for cancer therapy: Mechanisms, response biomarkers, and combinations," *Sci. Transl. Med.*, vol. 8, no. 328, p. 328rv4, 2016.

[2] C. Grigg and N. A. Rizvi, "PD-L1 biomarker testing for non-small cell lung cancer: Truth or fiction?" *J. ImmunoTherapy Cancer*, vol. 4, no. 1, p. 48, Dec. 2016.

[3] M. Udall *et al.*, "PD-L1 diagnostic tests: A systematic literature review of scoring algorithms and test-validation metrics," *Diagnostic Pathol.*, vol. 13, no. 1, p. 12, Dec. 2018.

[4] H. Kim, H. J. Kwon, S. Y. Park, E. Park, and J.-H. Chung, "PD-L1 immunohistochemical assays for assessment of therapeutic strategies involving immune checkpoint inhibitors in non-small cell lung cancer: A comparative study," *Oncotarget*, vol. 8, no. 58, p. 98524, 2017.

[5] M. C. Rebelatto *et al.*, "Development of a programmed cell death ligand-1 immunohistochemical assay validated for analysis of non-small cell lung cancer and head and neck squamous cell carcinoma," *Diagnostic Pathol.*, vol. 11, no. 1, p. 95, Dec. 2016.

[6] R. D. Ventana Medical System Inc. (2016). *Ventana PD-L1 (SP263) Assay Staining of Non-Small Cell Lung Cancer—Interpretation Guide*. [Online]. Available: http://www.ventana.com

[7] M. S. Tsao *et al.*, "PD-L1 immunohistochemistry comparability study in real-life clinical samples: Results of blueprint phase 2 project," *J. Thoracic Oncol.*, vol. 13, no. 9, pp. 1302–1311, Sep. 2018.

[8] P. Mobadersany *et al.*, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018.

[9] P. Courtiol *et al.*, "Deep learning-based classification of mesothelioma improves prediction of patient outcome," *Nature Med.*, vol. 25, pp. 1519—1525, Oct. 2019.

[10] A. Meier *et al.*, "End-to-end learning to predict survival in patients with gastric cancer using convolutional neural networks," in *Proc. Ann. Oncol., ESMO Congr.*, 2018, pp. 814–857.

[11] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019.

[12] G. Litjens *et al.*, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Sci. Rep.*, vol. 6, no. 1, p. 26286, Sep. 2016.

[13] A. Cruz-Roa *et al.*, "Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent," *Sci. Rep.*, vol. 7, no. 1, p. 46450, Jun. 2017.

[14] A. Kapil *et al.*, "Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.

[15] N. Brieu *et al.*, "Domain adaptation-based augmentation for weakly supervised nuclei detection," in *Proc. Workshop Comput. Pathol. (MICCAI)*, 2019, [Online]. Available: http://arxiv.org/abs/1907.04681

[16] A. Kapil *et al.*, "DASGAN—Joint domain adaptation and segmentation for the analysis of epithelial regions in histopathology PD-L1 images," in *Proc. Workshop Comput. Pathol. (MICCAI)*, 2019. [Online]. Available: http://arxiv.org/abs/1906.11118

[17] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[18] O. Ronneberger *et al.*, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Springer, 2015, pp. 234–241.

[19] Y. Liu *et al.*, "Detecting cancer metastases on gigapixel pathology images," 2017, *arXiv:1703.02442*. [Online]. Available: http://arxiv.org/abs/1703.02442

[20] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[21] J. Scheurer, C. Ferrari, L. B. T. Bom, M. Beer, W. Kempf, and L. Haug, "Semantic segmentation of histopathological slides for the classification of cutaneous lymphoma and eczema," in *Proc. Annu. Conf. Med. Image Understand. Anal.* Springer, 2020, pp. 26–42.

[22] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[23] S. Graham and N. M. Rajpoot, "SAMS-NET: Stain-aware multi-scale network for instance-based nuclei segmentation in histology images," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 590–594.

[24] M. van Rijthoven, M. Balkenhol, K. Siliņa, J. van der Laak, and F. Ciompi, "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101890.

[25] F. Mahmood *et al.*, "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3257–3267, Nov. 2020.

[26] L. Chan, M. Hosseini, C. Rowsell, K. Plataniotis, and S. Damaskinos, "HistoSegNet: Semantic segmentation of histological tissue type in whole slide images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10662–10671.

[27] W. Bulten *et al.*, "Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.

[28] N. Harder *et al.*, "Segmentation of prostate glands based on H&E or IHC counterstain with minimal manual annotation in prostate cancer," ISBI, 2019, Paper TuP2O-05.6. [Online]. Available: https://embs.papercept.net/conferences/conferences/ISBI19/program/ISBI19_ContentListWeb_2.html

[29] D. Tellez *et al.*, "Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2126–2136, Sep. 2018.

[30] J.-Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017, *arXiv:1703.10593*. https://arxiv.org/abs/1703.10593

[31] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "StainGAN: Stain style transfer for digital histological images," 2018, *arXiv:1804.01601*. [Online]. Available: http://arxiv.org/abs/1804.01601

[32] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," 2016, *arXiv:1610.09585*. [Online]. Available: http://arxiv.org/abs/1610.09585

[33] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: http://arxiv.org/abs/1802.05957

[34] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318

[35] R. Diagnostics. *Anti-Pan Keratin (AE1/AE3/PCK26) Primary Antibody*. [Online]. Available: http://ww.ventanadiscovery.com/product/69?type=64

[36] D. Faraggi and R. Simon, "A neural network model for survival data," *Statist. Med.*, vol. 14, no. 1, pp. 73–82, Jan. 1995.

[37] M. Zajac *et al.*, "Concordance among four commercially available, validated programmed cell death ligand-1 assays in urothelial carcinoma," *Diagnostic Pathol.*, vol. 14, no. 1, pp. 1–10, Dec. 2019.

[38] R. Diagnostics. *Ventana PD-l1 (SP142) Assay*. [Online]. Available: https://diagnostics.roche.com/content/dam/diagnostics/us/en/resource-center/PD-L1-SP142-Brochure-LUNG-US.pdf