Table of Contents

Supplementary Materials

Study Data

WSIs were collected into the Sheffield cohort based on the following criteria: a histological diagnosis of oral epithelial dysplasia (OED), sufficient availability of tissue (i.e. excluding tangentially cut sections, tissue with artefacts, verrucous lesions) and at least five years of follow-up data from the initial diagnosis (including treatment and transformation information). To ensure diagnostic consistency, all cases were independently evaluated by at least two certified/consultant pathologists (PMS, PMF, DJB, KDH), who provided an initial diagnosis based on the WHO grading system (2008-2016). To confirm the WHO (2017) grade and assign binary grades, the cases were blindly re-evaluated by an Oral & Maxillofacial Pathologist (SAK) and an Oral Surgeon with a specialist interest and expertise in OED analysis (HM). Cases with disagreement were resolved through discussion within the team. A total of 193 unique OED cases (with 270 slides) are used in the analysis for the internal Sheffield cohort. The external Birmingham-Belfast cohort consisted of 89 unique OED cases (with 89 slides). A summary of the internal/external cohorts are provided in Table 1. We also provide a CONSORT diagram in Figure 1 showing case collection.

Table 1
Demographic, characteristic and outcome data for the OED cases from Sheffield, Birmingham and Belfast cohorts.

	Sheffield	Belfast	Birmingham
n	270	42	47
Median Age (IQR)	64 (55 – 74)	62 (52 – 71)	61 (51 – 70)
Sex, *n* (%)			
Female	124 (46)	21 (50)	24 (51)
Male	145 (54)	21 (50)	23 (49)
Site, (%)			
Buccal Mucosa	35 (13)	0 (0)	6 (13)
Tongue	118 (44)	29 (69)	30 (64)
Floor of Mouth	57 (21)	9 (21)	3 (6)
Other	60 (22)	4 (10)	8 (17)
WHO grade, *n* (%)			
Mild	91 (34)	6 (14)	24 (51)
Moderate	104 (39)	25 (60)	18 (38)
Severe	75 (28)	11 (26)	5 (11)
Binary grade, *n* (%)			
Low-risk	169 (63)	7 (17)	28 (60)
High-risk	101 (37)	35 (83)	19 (40)
Transformation, *n* (%)	57 (21)	30 (71)	10 (21)
Median Follow-up Time, *months* (IQR)	95 (57 – 108)	42 (23 – 74)	47 (32 – 66)
Scanner, *n* (%)			
Aperio CS2	173 (64)	0 (0)	0 (0)
Hamamatsu NanoZoomer 360	97 (36)	0 (0)	0 (0)
Aperio AT2	0 (0)	42 (100)	0 (0)
Pannoramic 250	0 (0)	0 (0)	47 (100)

Note. Malignancy transformation is used as the survival endpoint.

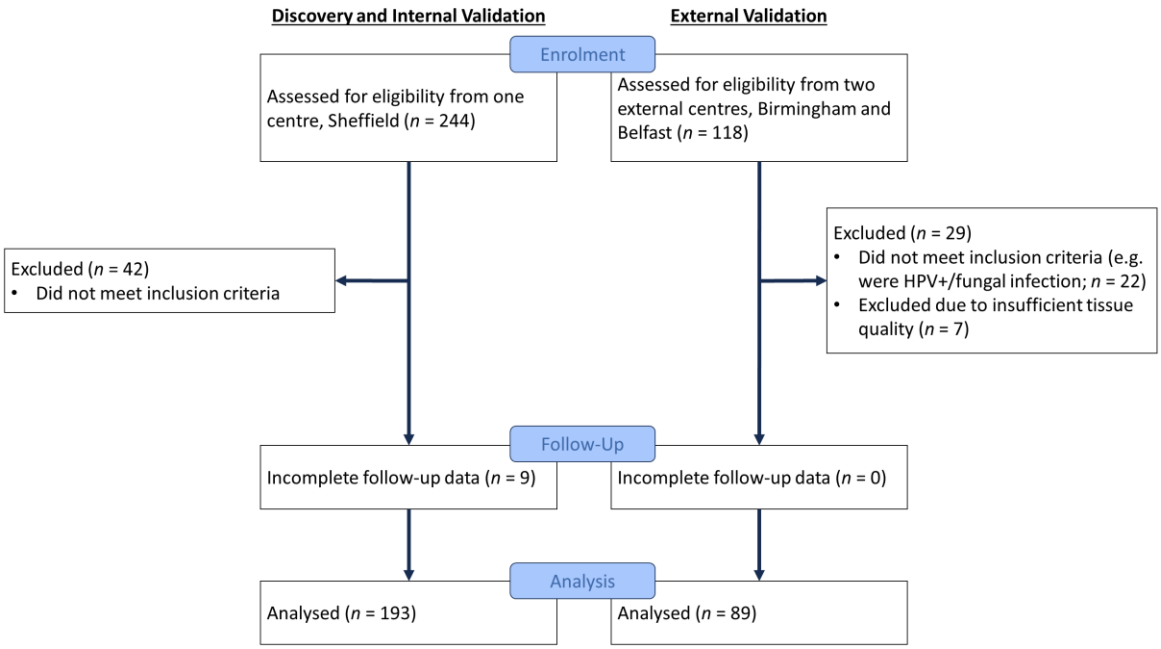

Figure 1
CONSORT diagram for the internal and external cases used in this study.

Ground Truth Labels

For training our segmentation models, an expert oral and maxillofacial pathologist (SAK) exhaustively manually delineated the three layers of the epithelium (basal layer, epithelial layer, superior keratin layer) in 59 OED cases, in addition to nine controls (collected with the Aperio scanner subject to the above protocols), using our inhouse WASABI software (a customised version of HistomicsTK). We then generated tissue masks for each of the segmented WSIs via Otsu thresholding and the further removal of small holes and objects. A layer mask was then generated for each WSI by combining the layer segmentations with the tissue mask.

The manual segmentation of individual nuclei within WSIs is laborious and subject to inter/intra-rater variability. Thus nuclear instance masks were only generated for a subset of cases. 30 regions of interest (ROIs) were created (one per case), and the pathologist (SAK) placed a point annotation on each nucleus within each of the ROIs. Each nucleus was labelled as either epithelial nuclei or 'other' nuclei. The point annotations were then used within the NuClick framework to generate nuclear boundaries (1,2). NuClick is a deep learning framework that takes a raw image and a guiding signal 'click' as an input and then produces a nuclear instance boundary as an output. This method has been found previously to be superior to fully-automated approaches for generating nuclear instance segmentations, particularly in the cases of touching/overlapping nuclei (1). Following the generation of nuclear instance masks in all 30 ROIs, a quality control step was added to ensure that all nuclei segmentations were of a high quality. The masks were then manually refined when found to be visibly incorrect. This typically occurred to nuclei in the granular layer of the epithelium. In total, this method resulted in 71,757 labelled nuclei segmentations across the 30 ROIs. These annotations were then used to train our deep learning segmentation models.

We aimed to make HoVer-Net+ as generalisable as possible; thus, in the training of our model we further split up our Sheffield dataset into smaller cohorts for testing the model generalisability. Thus of the 68 cases/controls that we had layer (and some nuclear) annotations for, we further split it up into two cohorts, cohort 1 (C1) and cohort 2 (C2; see Table 2 for breakdown), so that each cohort had images from two different scanners.

Table 2
Sheffield cohorts used for training HoVer-Net+.

	Layers		Nuclei	
	Cohort 1	Cohort 2	Cohort 1	Cohort 2
Cases, *n*				
Aperio CS2	38	0	20	0
Hamamatsu NanoZoomer 360	0	21	0	10
Controls, *n*				
Aperio CS2	5	4	0	0
Hamamatsu NanoZoomer 360	0	0	0	0

Supplementary Methods

Layer and Nuclear Segmentation

HoVer-Net+ is multi-task learning method, thus, it is assumed that the training of the encoder via different datasets/tasks will only increase its generalisability and performance (3). To train our HoVer-Net+ model, we have taken a multi-stage, multi-cohort approach. At 20× magnification (0.50 mpp), it is logical that upon patch extraction, we will produce many more patches with layer segmentations (68 WSIs) than nuclear annotations (30 ROIs). Thus, the primary step of our training schemes will be to train the encoder and LS decoder for layer segmentation. Following this, we could then train the nuclear segmentation/classification decoders (NP/HoVer/NC branches). We tessellated our WSIs, layer masks and nuclear instance maps into smaller patches of size 256 × 256 (no overlap) at 20× magnification (0.50 mpp). For C1, this resulted in 1,139 patches with nuclear instance maps (and corresponding layer masks), and 24,532 patches with layer masks. For C2, this resulted in 866 patches with nuclear instance maps, and over 7,731 patches with layer masks. For training purposes, an 80/20 train/test split was used for all models. Our training scheme consists of three stages:

1. We first aim to train HoVer-Net+ for layer segmentation alone, based on all the layer segmentation patches (model: HoVer-Net+$_{LS}$).
2. We then aim to train the entire network based on patches with both layer and nuclear annotations (model: HoVer-Net+$_{LS,All}$).
3. Finally, we freeze the encoder and train the LS decoder only on the segmentation patches again (model: HoVer-Net+$_{LS,All,LS}$)

During the second stage of training the LS branch was trained based on less layer annotations than in stage one, and thus will be overfit to the smaller sample used. Thus, the third stage aims to retrain the LS decoder again with all the layer annotations available.

We trained the HoVer-Net+ models in stages one and two over two phases. In phase one, only the decoder branches were trained for 20 epochs. In phase two, all branches (including the encoder) were trained for 30 epochs. A batch size of 8 and 4 on each GPU were used across these two phases respectively. In stage three, only the second phase of training was used as the encoder was kept frozen. For all training, the Adam optimiser was used with a learning rate that decayed initially from 10^{-4} to 10^{-5} after 10 epochs in each phase. During training, we applied the following random data augmentations: flip, rotation, Gaussian blur, median blur, and colour perturbation. We additionally, tested the effect of stain augmentation during our experiments, using the TIAToolbox (4) implementation, a method that has been shown to effectively counter scanner-induced domain-shifts, in order to create more generalisable models (5–7).

Comparative experiments were performed to test HoVer-Net+ model against other state-of-the-art models for nuclear segmentation and classification. For layer segmentation, we compared HoVer-Net+ to U-Net (8), a state-of-the-art model for semantic segmentation, following stage one of layer segmentation training. We additionally tested the performance of HoVer-Net+ for layer segmentation following further training on nuclear segmentation/classification. For nuclear segmentation/classification we further compared HoVer-Net+ to HoVer-Net (9) and U-Net (8). These models were trained based on their default parameters. For these comparisons, HoVer-Net and HoVer-Net+ were pretrained on PanNuke (10), and U-Net was pretrained on ImageNet. We further aimed to determine the generalizability of our HoVer-Net+ model to new datasets and different domain-invariance techniques.

Slide-level Transformation Prediction

We generated tile-level features for use in our MLP model for transformation prediction (OMTscoring). For each tile, we calculated 104 morphological and 64 spatial features. The morphological features were obtained

from 13 shape features for each nucleus in a tile (eccentricity, convex area, contour area, extent, perimeter, solidity, orientation, radius, major/minor axis, equivalent diameter, bounding box area/aspect ratio) with four tile-level statistics (mean, minimum, maximum, standard deviation) per nuclear type (epithelial and other). This resulted in 104 morphological features per tile. We computed the number of different nuclear types within a small radius of a nuclear instance, resulting in four counts per tile (number of epithelial nuclei around another nucleus, number of epithelial nuclei around epithelial nuclei, number of other nuclei around epithelial nuclei, and finally the number of other nuclei around other nuclei) over four varying radii (100, 200, 300 and 400 pixel radii). Finally, we took tile-level summary statistics (mean, minimum, maximum, standard deviation) across these 16 features, resulting in 64 spatial features per tile.

Supplementary Results

Layer and Nuclear Segmentation

We aimed to test the generalisability of our proposed HoVer-Net+ model to new datasets in the presence of a domain shift (e.g. by scanner/site), when trained for layer segmentation alone. The data in C1 were collected on an Aperio scanner, whilst the data in C2 were predominantly collected on a Hamamatsu scanner. Thus, we expect a certain level of domain-shift between these cohorts. Our initial experiments therefore aimed to test our method, when trained/tested on combinations of the two Sheffield cohorts with annotations (C1 and C2). We also tested the effect of using different pretrained models (e.g. PanNuke (10) or ImageNet), and stain augmentation. The results for these experiments are displayed in Table 3. Training on both cohorts generally improved performance across both datasets, whereas stain augmentation was seen to not necessarily be beneficial. Overall, trial 5 gained the best overall F1-score, when pretrained on PanNuke and trained on both C1 and C2, without any stain augmentation. We suggest that when both datasets are present within both the train and test set, the model learns enough variation over the different scanners used, and thus stain augmentation is not necessarily useful.

Following optimisation, our HoVer-Net+ model (trained for layer segmentation) was compared with U-Net (no stain augmentation, trained on C1 and C2). The results for these comparative experiments can also be seen in Table 4. Here, multiple trials of HoVer-Net+ are displayed. First, HoVer-Net+$_{LS}$ describes the model trained for layer segmentation alone. Unsurprisingly, this method achieved the best F1-score of 0.819, whilst U-Net (ResNet-50 encoder) achieved the worst results for layer segmentation. We additionally make further comparisons, showing how the layer segmentation performance changes with the additional training of the HoVer-Net+ model with nuclear information. HoVer-Net+$_{LS,All}$ describes the model in its second stage of training. Following on from the first stage, where the model was trained for layer segmentation alone, the second stage describes the model that is then trained to perform simultaneous nuclear instance segmentation/classification and layer segmentation. Intuitively, there is a slight drop in the model performance, with incorporation of nuclear information, as there are less patches with both nuclear and layer information, and thus the model will be slightly overfitted to this smaller sample size. However, we combated this by introducing a third stage of training, where we froze the encoder, and the nuclear instance segmentation/classification decoder branches, and further tuned the LS decoder branch on the entire cohort of layer patches (HoVer-Net+$_{LS,All,LS}$). This acted to improve layer segmentation only slightly when compared to HoVer-Net+$_{LS,All}$.

The HoVer-Net+ results for nuclear instance segmentation and classification are included in Table 5; where we have additionally compared the performance against HoVer-Net (9) and U-Net (8). U-Net has expectedly gained the worst performance, owing to its inability to differentiate touching nuclei well. This is best illustrated in Figure 2, where touching basal layer epithelial nuclei are poorly discriminated. Within this table, HoVer-Net+$_{LS,All}$ was initially trained on all layer segmentation data, and then further trained on all nuclei/layer data. Following the training of HoVer-Net+ on nuclei and layer information (e.g. HoVer-Net+$_{LS,All}$), unsurprisingly there is a decrease in performance for layer segmentation (see Table 5). The mean F1-score for HoVer-Net+ decreases from 0.819 to 0.812 owing to there being much fewer patches with both layer and nuclear annotations. We therefore opted to further finetune the HoVer-Net+ models (LS decoder only) on the entire cohort of layer patches (stage three of training). Following finetuning, the mean F1-score of HoVer-Net+ (HoVer-Net+$_{LS,All,LS}$ in Table 5) only increased marginally from 0.812 to 0.813. Since, this stage was done with the LS decoders alone, this should have little effect on the performance of HoVer-Net+ on nuclear segmentation/classification, only really effecting the nuclear classification performance. As can be seen from Table 5 this didn't affect the segmentation/detection quality of HoVer-Net+$_{LS,All,LS}$; however, did act to improve with the classification of nuclei. This demonstrates the benefits of using HoVer-Net+ in a multi-task learning framework. When we compare both of our HoVer-Net+ models to HoVer-Net we see an improved performance across both datasets,

showing the superiority of the method over both cohorts C1 and C2. Thus, this model is used for inference on all WSIs in both the internal and external datasets seen in the next step of our analysis pipeline.

The output layer and nuclear segmentations of HoVer-Net+ on a ROI from the test set are displayed in Figure 2. The top row shows the GT layer segmentation on the left followed by HoVer-Net+ (HoVer-Net+$_{LS,All,LS}$) and U-Net. Here, we can see that HoVer-Net+ appears to have the best quality segmentation. The blue boxes show areas where HoVer-Net+ has segmented regions that the GT segmentation has misclassified. In general, U-Net's segmentations appear more spurious. The bottom row shows the nuclear segmentations for the GT, HoVer-Net+, HoVer-Net and U-Net (from left to right). Both HoVer-Net and HoVer-Net+ have incorrectly classified some epithelial nuclei as keratin nuclei, however, HoVer-Net appears to have misclassified many more of these cases, demonstrating the benefits of including the LS decoder branch.

Added Value of HoVer-Net+

In the first stage of our experiment, we tested the performance and generalisability of our HoVer-Net+ method to new cohorts and domain invariance techniques. HoVer-Net+, when trained for layer segmentation, achieved the best results when pretrained on PanNuke over ImageNet. This is perhaps unsurprising, demonstrating the merits of pretraining on domain-specific images. We additionally tested how well HoVer-Net+ performed when trained/tested on different datasets. Overall, the mean F1-score across both C1 and C2 (used in training HoVer-Net+) was highest when trained over both datasets. Counterintuitively, we also found stain augmentation to have mixed effects on model performance, even when testing on new data. Since, we wished to generate a model that generalised as much as possible to new datasets we used the model trained on both datasets going forward, without any stain augmentation.

We performed multiple experiments testing the order of training HoVer-Net+ for each of the different tasks. We first trained the model on just layer segmentation data, as there were many more layer annotations than nuclear annotations. Next, we further trained the HoVer-Net+ model for nuclear segmentation/classification and layer segmentation (NP, HoVer, NC decoders pretrained on PanNuke, LS decoder trained in previous stage). Following this training (labelled HoVer-Net+$_{LS,All}$), there was an expected reduction in layer segmentation performance, owing to the encoder and LS decoder being further trained on a much smaller subset of data, and therefore being more susceptible to overfitting. We therefore circumvented this issue, by freezing the encoder and the nuclear decoder branches (NP, HoVer, NC), and finetuning the LS decoder branch on the entire layer segmentation data. This then resulted in an improved performance for layer segmentation (labelled HoVer-Net+$_{LS,All,LS}$) when compared to the HoVer-Net+$_{LS,All}$. Since the nuclear decoder branches were not altered here, this did not affect the nuclear segmentation performance of the model; but did improve the performance for nuclear sub-classification of epithelial nuclear types (e.g. as basal, epithelial, keratin), which is dependent on the output of the LS branch. This demonstrates the merits of our multi-task learning approach, where training over multiple tasks can improve the representation of the images learned by the encoder, to benefit both tasks.

Following the optimisation of HoVer-Net+ we compared its performance to U-Net for layer segmentation. Here, we found HoVer-Net+ to get the best performance, with U-Net being worse. We additionally compared HoVer-Net+ to HoVer-Net and U-Net for nuclear segmentation/classification, where we found HoVer-Net+ to gain the best results. Overall, these results express the benefits of our multi-task learning HoVer-Net+ method, which we expect to generalise well to new data.

Table 3

Experiments for layer segmentation on both dataset 1 and 2, based on F1-scores.

Trial	Train Set	Pretrain.	Stain. Aug.	Test C1						Test C2						Overall Mean
				Bkgd.	Other	Basal	Epith.	Keratin	Mean	Bkgd.	Other	Basal	Epith.	Keratin	Mean	
1	C2	PanNuke	N	0.872	0.765	0.563	0.802	0.644	0.729	0.910	**0.853**	**0.771**	0.888	0.791	0.843	0.786
2	C2	PanNuke	Y	0.822	0.713	0.557	0.739	0.547	0.675	0.909	0.828	0.714	0.861	0.774	0.817	0.746
3	C1	PanNuke	N	0.842	0.807	0.717	0.819	0.660	0.769	0.844	0.813	0.720	0.851	0.780	0.802	0.785
4	C1	PanNuke	Y	0.849	0.819	0.700	0.838	0.682	0.778	0.876	0.835	0.727	0.865	0.772	0.815	0.796
5	**C1,C2**	**PanNuke**	**N**	0.874	**0.849**	**0.721**	**0.858**	**0.717**	**0.804**	0.904	0.849	0.753	0.880	0.789	0.835	**0.819**
6	C1,C2	PanNuke	Y	0.850	0.821	0.706	0.835	0.674	0.777	0.901	0.851	0.739	0.884	**0.801**	0.835	0.806
7	C2	ImageNet	N	0.869	0.774	0.594	0.797	0.627	0.732	**0.911**	0.851	0.760	**0.891**	0.798	**0.842**	0.787
8	C2	ImageNet	Y	0.843	0.742	0.544	0.781	0.603	0.703	0.910	0.842	0.695	0.868	0.776	0.818	0.760
9	C1	ImageNet	N	**0.878**	**0.849**	0.715	0.852	0.698	0.798	0.864	0.473	0.533	0.446	0.251	0.513	0.656
10	C1	ImageNet	Y	0.847	0.807	0.711	0.826	0.664	0.771	0.822	0.806	0.722	0.846	0.783	0.796	0.783
11	C1,C2	ImageNet	N	0.864	0.808	0.701	0.815	0.654	0.768	0.905	0.777	0.714	0.809	0.578	0.756	0.762
12	C1,C2	ImageNet	Y	0.817	0.755	0.672	0.780	0.577	0.720	0.888	0.834	0.713	0.870	0.784	0.818	0.769

Table 4

Comparative experiments for layer segmentation on both dataset 1 and 2, based on F1-scores.

Model	Pretrain.	Test C1						Test C2						Overall Mean
		Bkgd.	Other	Basal	Epith.	Keratin	Mean	Bkgd.	Other	Basal	Epith.	Keratin	Mean	
U-Net	ImageNet	**0.880**	**0.849**	**0.739**	0.853	0.708	**0.806**	**0.905**	0.777	0.714	0.809	0.578	0.756	0.781
HoVer-Net+$_{LS}$	PanNuke	0.874	**0.849**	0.721	**0.858**	**0.717**	0.804	0.904	**0.849**	0.753	0.880	0.789	0.835	**0.819**
HoVer-Net+$_{LS,All}$	PanNuke	0.862	0.837	0.722	0.846	0.707	0.795	0.902	0.834	0.759	0.872	0.780	0.829	0.812
HoVer-Net+$_{LS,All,LS}$	PanNuke	0.853	0.827	0.716	0.851	0.700	0.789	0.898	**0.849**	**0.764**	**0.885**	**0.792**	**0.837**	0.813

Note. HoVer-Net+$_{LS}$ was trained on layer data only. HoVer-Net+$_{LS,All}$ was trained on layer data only, then patches with both nuclei and layer data only. HoVer-Net+$_{LS,All,LS}$ was trained on layer data only, then patches with both nuclei and layer data only, finally, the LS decoder alone was finetuned on all layer data.

Table 5

Comparative experiments for nuclear instance segmentation and classification.

Model	Test C1									Test C2								
	Dice	AJI	DQ	SQ	PQ	F^d	F_c^o	F_c^b	F_c^e	Dice	AJI	DQ	SQ	PQ	F^d	F_c^o	F_c^b	F_c^e
U-Net	0.605	0.378	0.473	0.606	0.311	0.627	0.585	0.245	0.529	0.582	0.354	0.444	0.581	0.302	0.652	0.541	0.275	0.522
HoVer-Net	0.713	0.589	0.721	**0.730**	**0.536**	**0.813**	0.680	0.542	0.590	0.660	0.591	0.708	**0.641**	0.473	0.820	0.621	0.494	0.619
HoVer-Net+$_{LS,All}$	**0.716**	**0.598**	**0.724**	**0.730**	0.533	0.810	**0.719**	**0.599**	0.633	**0.665**	**0.635**	**0.753**	**0.641**	**0.478**	**0.833**	0.715	0.590	0.646
HoVer-Net+$_{LS,All,LS}$	**0.716**	**0.598**	**0.724**	**0.730**	0.533	0.810	0.717	0.593	**0.646**	**0.665**	**0.635**	**0.753**	**0.641**	**0.478**	**0.833**	**0.727**	**0.621**	**0.668**

Note. HoVer-Net+$_{LS,All}$ was trained on layer data only, then patches with both nuclei and layer data only. HoVer-Net+$_{LS,All,LS}$ was trained on layer data only, then patches with both nuclei and layer data only, finally, the LS decoder alone was finetuned on all layer data.

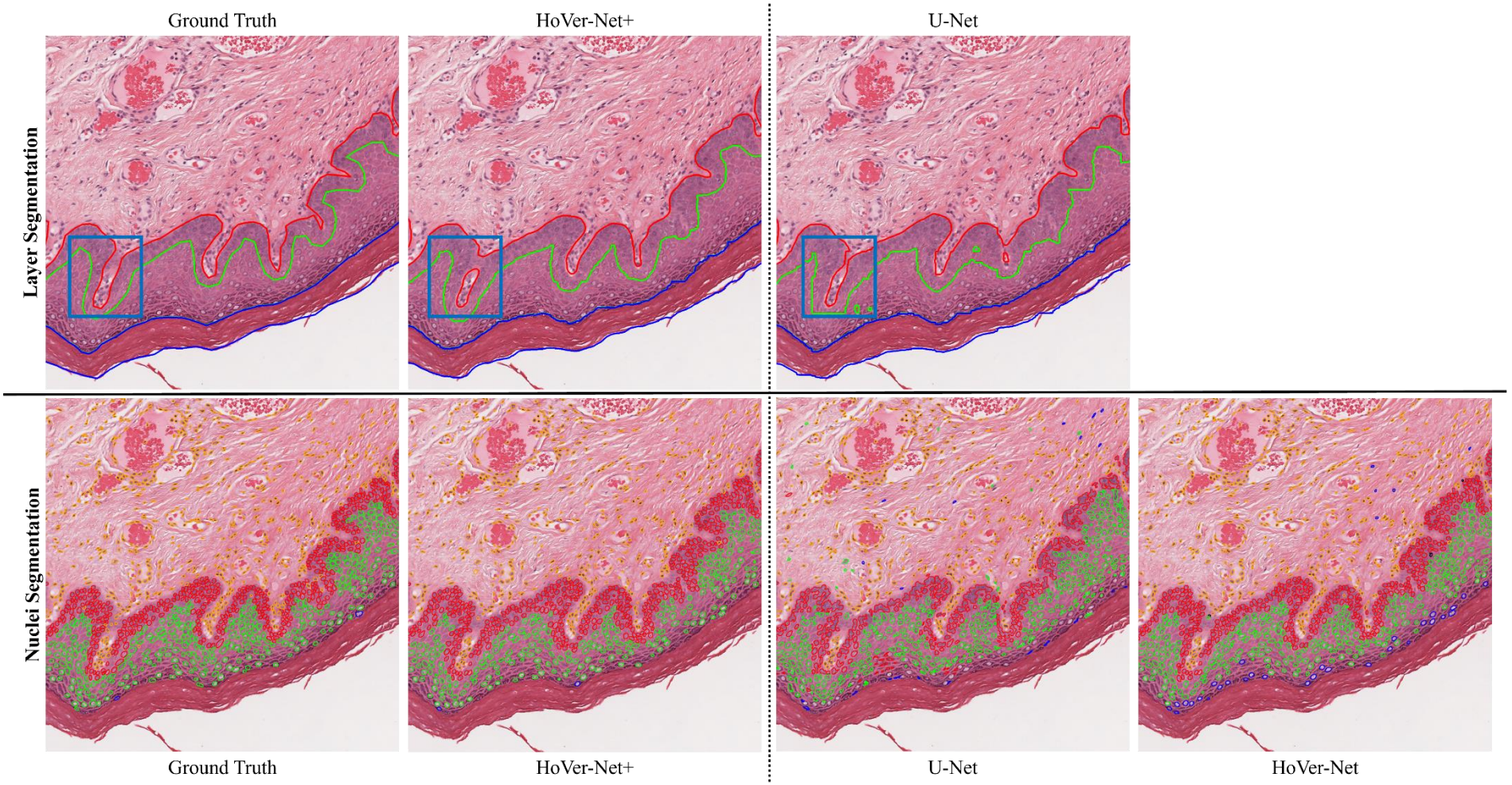

Figure 2
Top row: A comparison of the GT layer segmentations vs. HoVer-Net+$_{LS,All,LS}$ and U-Net. In the images, the red line represents the basal layer, the green epithelium, and the blue keratin. For clarity, other tissue region segmentations have been excluded from these results. Bottom row: A comparison of the GT nuclear segmentations vs. HoVer-Net+$_{LS,All,LS}$, HoVer-Net and U-Net. In the images, the red line represents the basal nuclei, green epithelial nuclei, blue keratin nuclei, orange other nuclei and finally, black represents unlabelled nuclei. Note, the dividing line between the annotations on the left and right is to display that the HoVer-Net+ segmentations for nuclei and layers are from the same model. On the right, specifically with U-Net, these are from two separately trained models.

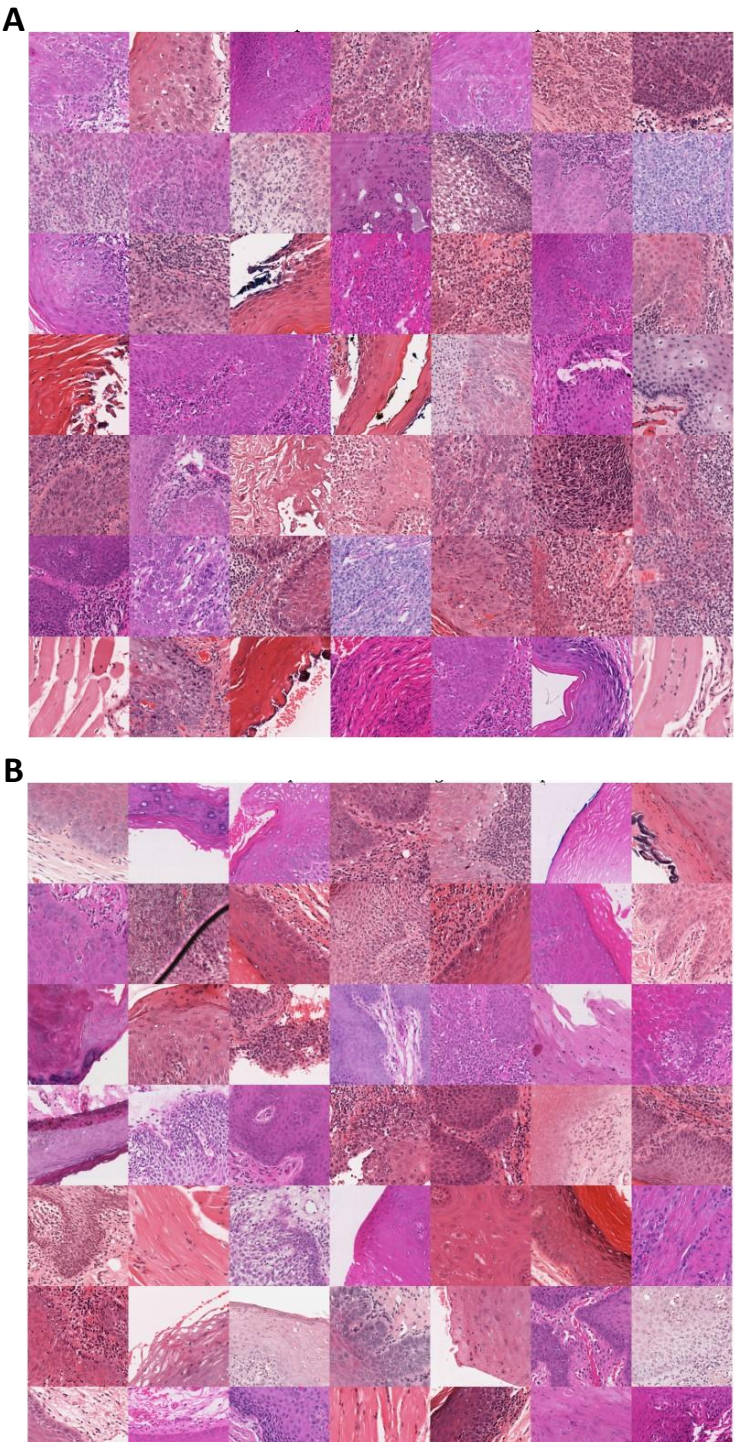

Figure 3
Random patch montage of A) top predicted patches in true positives and B) bottom predicted patches in true negatives from our MLP.

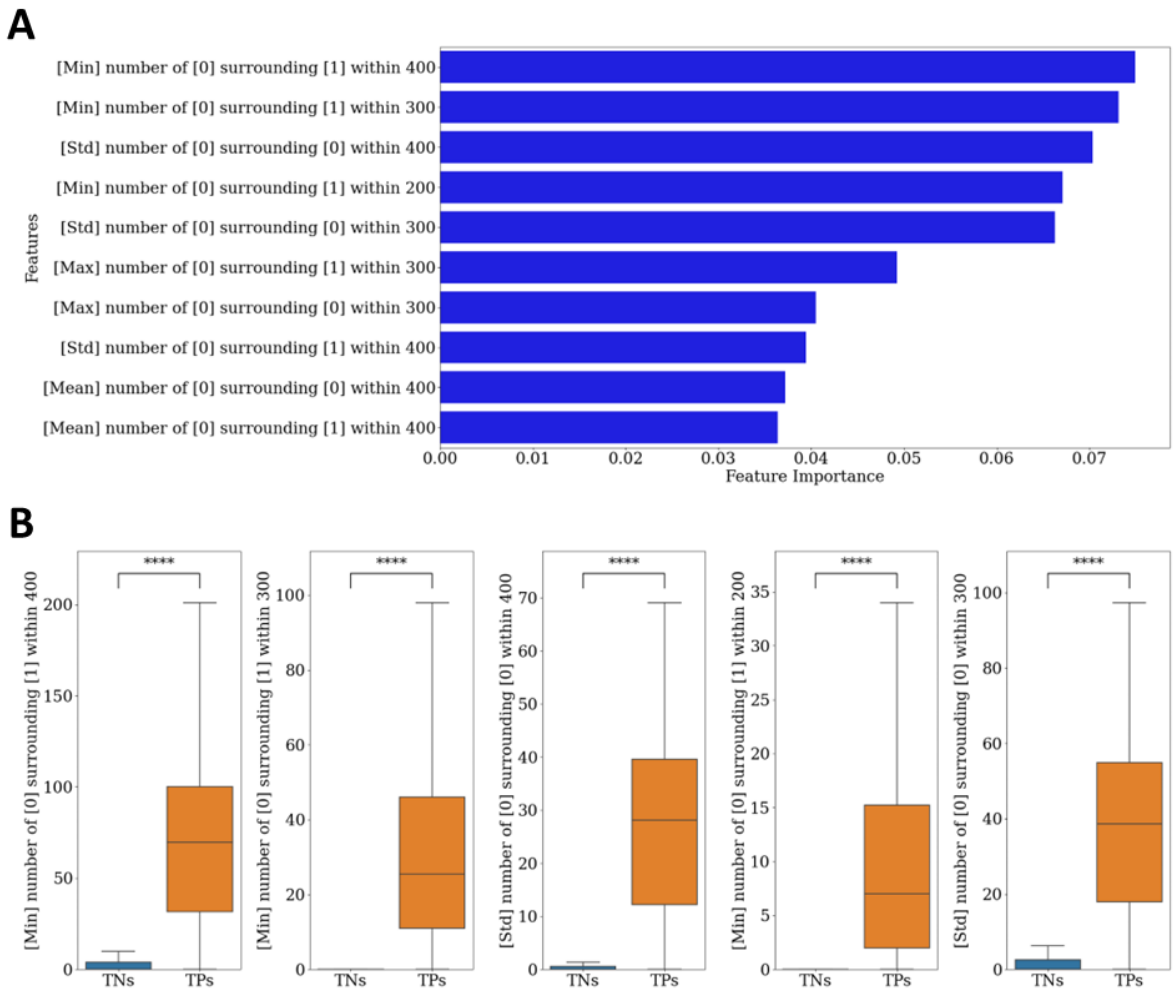

Figure 4
A) The top ten most important morphological/spatial features found via a Random Forest, and B) bosplots showing the distribution of the top five of these features. [0] are "other" nuclei, whilst [1] are epithelial nuclei. Distance is measure in microns.

Slide-level Transformation Prediction with our MLP

Survival Analysis

For slide-level prediction, we used a multi-layer perceptron (MLP), trained with the iterative draw-and-rank method introduced by Bilal *et al.* (2021) (11) with our tile-level morphological/spatial features (our OMTscoring model). Results from the Cox proportional hazard model (see Table 6) showed that both the binary grade ($p < 0.01$, HR = 3.96 [1.45, 11.10]) and *OMTscore* ($p < 0.0001$, HR = 8.48 [3.87, 21.30]) were statistically significant, with the *OMTscore* exhibiting the highest hazard ratio (HR), indicating better prognostic utility. No other clinical variables were found to be significant.

For external validation, the multivariate Cox PH models (see Table 6) showed no variables to be statistically significant. However, both the *OMTscore* ($p = 0.32$, HR = 3.01 [0.71, 20.62]), and the binary grade ($p = 0.14$, HR = 2.64 [0.70, 8.84]) showed high hazard ratios, highlighting their prognostic utility over the other clinical variables and the WHO grade ($p = 0.96$, HR = 1.27 [0.64, 2.50]).

Coldspot/hotspot Patches

For slide-level prediction, we used a multi-layer perceptron (MLP), trained with the iterative draw-and-rank method introduced by Bilal *et al.* (2021) (11) with our tile-level morphological/spatial features (our OMTscoring model). We display a randomly generated selection of the top five most predictive patches for TP

cases and bottom five patches from TN cases on internal testing in Figure 3. Visually, it is clear that there appear to be a larger number of immune cells both within the epithelium and the connective tissue within the TP cases.

Hotspot-Coldspot Feature Importance

We additionally aimed to determine which of the 168 nuclear features used to train our MLP were most important for making the final prediction (on internal testing). Thus, we trained a Random Forest classifier based on the 168 nuclear features in the top five/bottom five predicted patches by our MLP model. We then ranked the feature importance (mean decrease in impurity, MDI), selected the top ten features and performed two-tailed t-tests (with FDR correction), to determine statistical significance.

We present the top ten features found from this analysis (using MDI importance) in Figure 4 A. Spatial features were generally estimated to be most important and made up the top ten features. We further display boxplots showing the distribution of the top five of these features in in Figure 4 B and provide results from a two-tail t-test (following FDR correction). The minimum number of "other" nuclei surrounding an epithelial nucleus within 400 microns had the highest importance, although all of the top ten features had large effect sizes ($d >$ 2.00) and were shown to be significant ($p < 0.0001$). These top features again suggest the importance of lymphocytes (e.g. "other" nuclei) and their proximity to epithelial nuclei (e.g. epithelial-infiltrating/peri-epithelial lymphocytes).

Hotspot-Coldspot Feature Exploration

We did not incorporate any mitotic features into our models, however, we additionally aimed to determine whether our top/bottom predicted patches could further inform us of their importance. We therefore, ran the MIDOG 2022 winning code (available online: adamshephard/TIA-mitosis (github.com)) on the top/bottom predicted patches from our analyses. Since mitotic figures are generally sparse in dysplasia, we generated a mitotic count per slide over all of the five top/bottom predicted patches by the model (in TP vs TN cases, respectively). Only slides that included one or more mitotic figures were included in this analysis. A two-tailed t-test was then used to determine statistical significance.

The MIDOG 22 winning model appeared to work well on our dataset (by visual inspection). However, we found no difference between the number of mitotic figures discovered in TPs vs TNs ($d = 0.06$, $p = 0.82$). This was perhaps an unsurprising finding, as mitotic figures were not included as a feature in our MLP models.

External Testing

We additionally aimed to test the performance of our MLP model, on external validation, with the exclusion of cases that HoVer-Net+'s nuclear/region segmentation results were deemed insufficient quality (n = 13). The results of this analyses are shown in Table 7. We see a clear increase in performance compared to without the quality control (QC). This shows the importance of having a robust segmentation pipeline for this analysis. With training on more diverse data, we expect HoVer-Net+ to generalise better to unseen datasets.

Comparison to IDaRS for Transformation Prediction

We additionally compared our MLP method for transformation prediction to another state-of-the-art method. Thus, we compared it to IDaRS (11), trained with ResNet-34 deep features, in its original published form. We suggest that one of the main advantages of our MLP trained with morphological/spatial features, is that these features are in theory "domain-agnostic". Thus, providing the deep learning models used to generate the nuclear segmentations generalise, then these features should generalise well to new, unseen datasets. In contrast, we suggest that deep models such as IDaRS may not generalise well to unseen data. Thus, we have additionally employed domain adaption techniques to make the IDaRS model robust to new data, in order to make a fair-comparison of the IDaRS method to our MLP. Thus, we additionally include the IDaRS model, trained using a) Macenko stain augmentation, b) domain adversarial training, and c) a combination of these two techniques. Domain adversarial training was introduced by Ganin *et al.* (2016) (12), and aims to force the encoder part of the CNN to learn domain-agnostic features. Within this work, an additional classifier head is added to the IDaRS model that aims to predict the domain from which the provided features are from (in this case, the scanner used). The addition of a gradient reversal layer aims to maximise the error in this classification task, thus incentivising the classification model to learn domain-agnostic features.

In all models presented in this work, the patches used for feature generation are based on HoVer-Net+'s segmentation of the epithelium. All models were additionally trained with a symmetric cross-entropy loss function, and the Adam optimiser. We chose the parameters $k = 5$ top predictive patches and $r = 45$ random

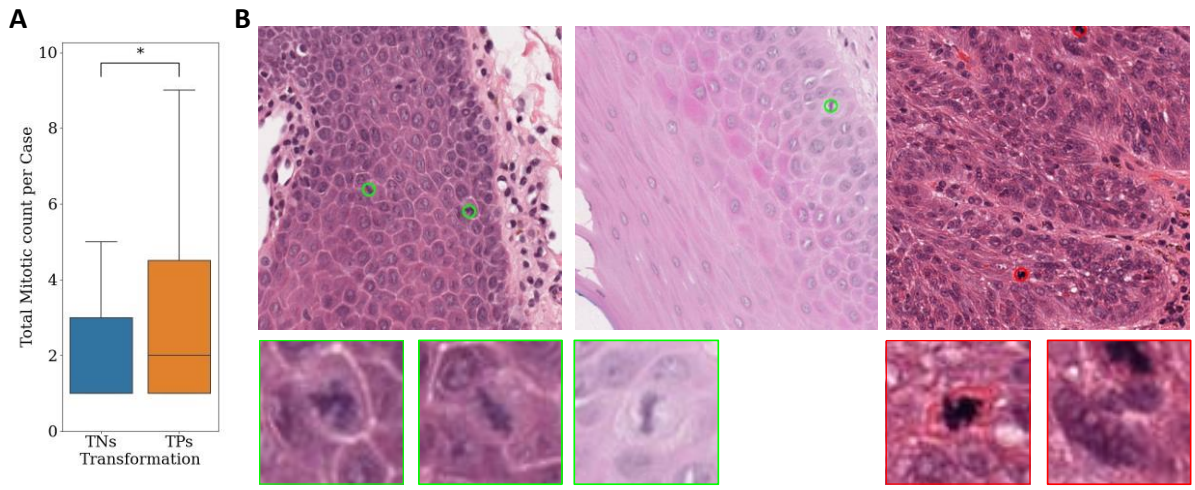

Figure 5
A) Mitotic figure counts (MIDOG 2022 winning algorithm), in the top five predicted patches from true positive cases against the bottom five predicted patches from true negative cases. B) Detections with algorithm correctly detected mitoses (green) and incorrectly missed mitoses (red).

patches in the iterative draw-and-rank method, with a batch size of 256. All models were trained until convergence with a minimum of 30 and a maximum of 100 epochs.

In these experiments (see Table 8), we have compared the performance of our MLP trained with patch-level morphological/spatial features to the IDaRS model (based on deep features from a trained ResNet-34 model). One of the main advantages of our MLP trained with morphological features, is that these features are in theory "domain-agnostic", thus, we additionally compare our model to IDaRS trained with a) stain augmentation (IDaRS-SA), b) domain adversarial training (IDaRS-DA), and c) IDaRS trained with both stain augmentation and domain adversarial training (IDaRS-SA-DA).

Overall, the results tell an interesting story, with relation to domain adaption. The raw IDaRS model that obtained high results on internal cross-validation (AUROC = 0.78), has not generalised well to the new unseen data (AUROC = 0.51. The various domain adaption techniques, such as domain adversarial training and stain augmentation, did not help to improve model performance on internal testing; but, helped improve the generalisability of the IDaRS model on external testing. However, ultimately, the model that was trained with domain-agnostic features (e.g. our MLP) achieved the highest results on external testing (AUROC = 0.75).

Feature Discovery

IDaRS generated predictions based on deep features learned through training (here, ResNet-34). This model differs from the MLP pipeline used primarily in this study, in that the features that are used in the MLP are deliberately generated, with the purpose of being interpretable and important. However, the IDaRS framework does generate good predictions for internal validation. Thus, we also explored the hot/coldspots found by the IDaRS model (on internal testing, IDaRS-DA-SA) to determine what were the important features used for prediction. Interestingly, we found a significantly higher number of mitotic figures in the TP patches than the TN patches (d = 0.63, p = 0.01; see Figure 5). This suggests that the IDaRS model has independently learned to account for mitosis as a deep feature.

Table 6

Multivariate Cox Proportional Hazard Model output for malignant transformation based on the OMTscore, and other clinical variables.

	Internal Validation – Sheffield (*n* = 270)				External Validation – Combined (*n* = 89)			
	p	HR	Lower 95% HR	Upper 95% HR	*p*	HR	Lower 95% HR	Upper 95% HR
OMTscore	**< 0.0001**	**8.48**	**3.87**	**21.30**	0.32	3.01	0.71	20.62
Binary Grade	**< 0.001**	**3.96**	**1.45**	**11.10**	0.14	2.64	0.70	8.84
WHO Grade	1.00	1.06	0.57	2.04	0.96	1.27	0.64	2.50
Age	0.54	1.01	0.98	1.03	1.00	1.00	0.97	1.02
Sex	0.60	1.34	0.71	2.51	0.81	1.29	0.61	2.62
Site	0.36	1.19	0.85	1.67	0.07	1.59	1.03	2.55

HR = Hazard Ratio.

Table 7

Slide-level mean (standard deviation) results for transformation prediction when trained on the Sheffield cohorts and tested on the external Birmingham-Belfast data, with QC of HoVer-Net+ output.

		Birmingham		Belfast		Combined	
Model	Features	F1-score	AUROC	F1-score	AUROC	F1-score	AUROC
OMTscore	**Morph./Spatial**	0.47 (0.03)	**0.80 (0.01)**	**0.84 (0.01)**	**0.71 (0.05)**	0.73 (0.02)	**0.78 (0.01)**
Binary Grade	Low- vs High-Risk	**0.60**	**0.80**	0.80	0.56	**0.75**	0.73
WHO Grade	Mild/Mod. vs Severe	0.44	0.64	0.39	0.51	0.40	0.60
WHO Grade	Mild vs Mod./Severe	0.57	0.78	0.79	0.52	0.74	0.70

Table 8
Slide-level mean (standard deviation) results for transformation prediction.

		Internal Testing		External Testing					
		Sheffield		Birmingham		Belfast		Combined	
Model	Features	F1-score	AUROC	F1-score	AUROC	F1-score	AUROC	F1-score	AUROC
IDaRS	ResNet-34	0.59 (0.11)	0.78 (0.11)	0.41 (0.03)	0.66 (0.04)	0.37 (0.32)	0.58 (0.01)	0.40 (0.19)	0.51 (0.09)
IDaRS-SA	ResNet-34	**0.59 (0.07)**	**0.80 (0.07)**	0.21 (0.19)	0.66 (0.04)	0.56 (0.24)	0.52 (0.07)	0.48 (0.23)	0.74 (0.02)
IDaRS-DA	ResNet-34	0.55 (0.10)	0.73 (0.11)	0.17 (0.19)	0.68 (0.05)	0.59 (0.16)	0.49 (0.04)	0.50 (0.17)	0.72 (0.06)
IDaRS-SA-DA	ResNet-34	0.53 (0.08)	0.72 (0.11)	0.00 (0.00)	0.70 (0.07)	0.09 (0.11)	0.51 (0.05)	0.07 (0.09)	**0.76 (0.01)**
OMTscore	**Morph./Spatial**	0.57 (0.08)	0.77 (0.08)	0.44 (0.01)	0.73 (0.01)	**0.84 (0.02)**	**0.71 (0.03)**	0.69 (0.01)	0.75 (0.01)
Binary Grade	Low- vs High-Risk	0.51 (0.08)	0.71 (0.06)	0.55	0.75	0.80	0.56	**0.72**	0.72
WHO Grade	Mild/Mod. vs Severe	0.34 (0.16)	0.58 (0.11)	0.40	0.63	0.39	0.51	0.39	0.69
WHO Grade	Mild vs Mod./Severe	0.46 (0.08)	0.68 (0.05)	**0.55**	**0.76**	0.79	0.52	0.71	0.69

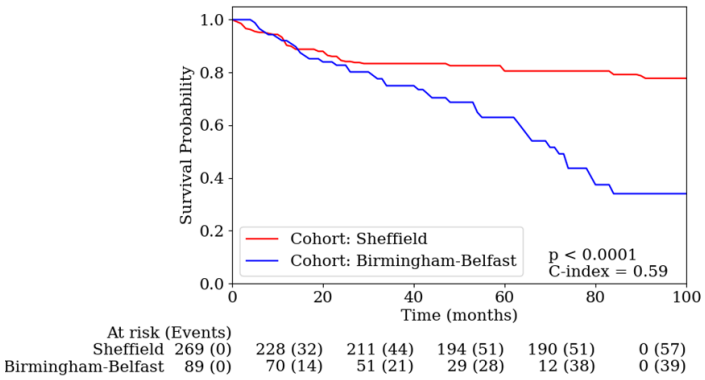

Figure 6
Kaplan-Meier transformation-free survival curves for the Sheffield and Birmingham-Belfast datasets.

Supplementary Discussion

The Difficulty of the Task

In this work, we aimed to generate a model that would better predict transformation (our *OMTscore*), using interpretable features. We also wished to validate this model on external cohorts. However, some comments must be included on the difficulty of this task. OED data can be relatively sparse, and we therefore found it difficult to collate large external cohorts. Even with the inclusion of the Birmingham and Belfast data, we found that the survival information of these cohorts varied drastically, adding to the difficulty of the task. We have therefore included the independent survival curves of our cohorts in Figure 6. Here, we can see how the survival curves for these cohorts vary drastically, with low concordance and a significant difference via a log-rank test. This confounding variable is the clinical reality, and thus adds to the strength of our model's performance.

Domain Generalisation and Future Work

In this work, we aimed to generate a model that would better predict transformation, using interpretable features. Ultimately, these morphological/spatial features were based on shape/size differences (e.g. the differences in the mean radius of nuclei in a patch) in different nuclear types (e.g. epithelial vs other nuclei). Overall, we found our MLP model to gain the best results for predicting transformation on external testing. Interestingly, despite the success of IDaRS with deep features in the past, we have found that by instead using morphological/spatial features one can achieve better performance. We also argue that the incorporation of such features in favour of deep features, makes our models more interpretable. IDaRS, which initially gained the highest AUROC values on internal testing, failed to generalise to the new, unseen data without any form of domain alignment/augmentation techniques. As was expected, the addition of stain augmentation and domain adversarial training, improved the performance of IDaRS on new data (from AUROC = 0.51 to AUROC = 0.76). However, ultimately our MLP model, trained on interpretable morphological/spatial features gained the highest performance in terms of both AUROC and F1-score (AUROC = 0.75; F1 = 0.69). The F1-scores of the IDaRS-based models were generally low, suggesting a lack of model convergence. Our MLP model features are essentially domain-agnostic (when segmented accurately), and we suggest that this is the reason for their superior performance.

Despite this, we still explored the hotspot/coldspot regions produced by IDaRS, and found that mitoses potentially contributed to the model performance. Future work should look at incorporating tile-level (or even slide-level) mitotic scoring into the MLP model, to further improve the performance.

Supplementary References

1. Alemi Koohbanani N, Jahanifar M, Zamani Tajadin N, Rajpoot N. NuClick: A deep learning framework for interactive segmentation of microscopic images. Med Image Anal. 2020;65.

2. Jahanifar M, Koohbanani NA, Rajpoot N. NuClick: From Clicks in the Nuclei to Nuclear Boundaries. arXiv [Internet]. 2019;(Dl). Available from: http://arxiv.org/abs/1909.03253

3. Graham S, Vu QD, Jahanifar M, Raza SEA, Minhas F, Snead D, et al. One model is all you need: Multi-task learning enables simultaneous histology image segmentation and classification. Med Image Anal. 2023;83(October 2022).

4. Pocock J, Graham S, Vu QD, Jahanifar M, Deshpande S, Hadjigeorghiou G, et al. TIAToolbox as an end-to-end library for advanced tissue image analytics. Commun Med [Internet]. 2022;2(1):120. Available from: https://www.nature.com/articles/s43856-022-00186-5

5. Aubreville M, Stathonikos N, Bertram CA, Klopfleisch R, ter Hoeve N, Ciompi F, et al. Mitosis domain generalization in histopathology images — The MIDOG challenge [Internet]. Vol. 84, Medical Image Analysis. Elsevier; 2023 [cited 2023 Jan 6]. p. 102699. Available from: http://arxiv.org/abs/2204.03742

6. Jahanifar M, Shephard A, Tajeddin NZ, Bashir RMS, Bilal M, Khurram SA, et al. Stain-Robust Mitotic Figure Detection for the Mitosis Domain Generalization Challenge. arXiv [Internet]. 2021;3–5. Available from: http://arxiv.org/abs/2109.00853

7. Jahanifar M, Shephard A, Zamanitajeddin N, Raza SEA, Rajpoot N. Stain-Robust Mitotic Figure Detection for MIDOG 2022 Challenge. 2022;1–3. Available from: http://arxiv.org/abs/2208.12587

8. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2015;9351:234–41.

9. Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal [Internet]. 2019;58:101563. Available from: https://doi.org/10.1016/j.media.2019.101563

10. Gamper J, Koohbanani NA, Benes K, Graham S, Jahanifar M, Khurram SA, et al. PanNuke Dataset Extension, Insights and Baselines. 2020;1–12. Available from: http://arxiv.org/abs/2003.10778

11. Bilal M, Raza SEA, Azam A, Graham S, Ilyas M, Cree IA, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. Lancet Digit Heal [Internet]. 2021;3(12):e763–72. Available from: http://dx.doi.org/10.1016/S2589-7500(21)00180-1

12. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. In: Advances in Computer Vision and Pattern Recognition. 2017. p. 189–209.