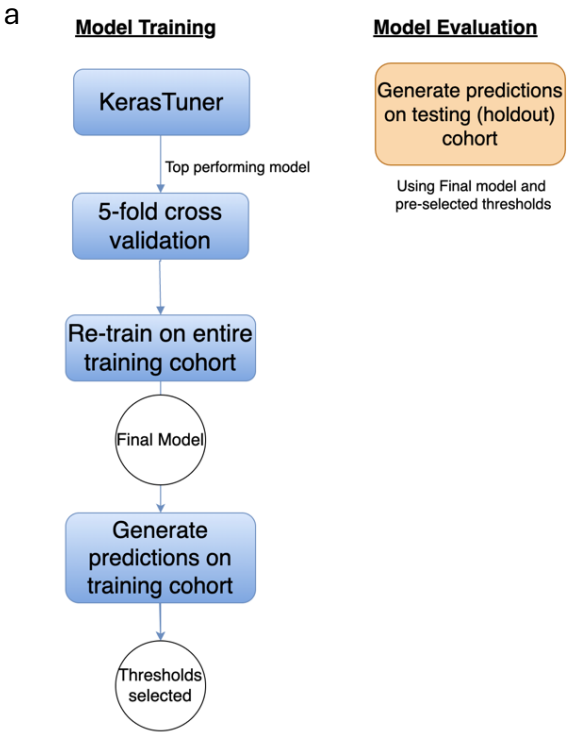
Supplementary Material

Supplementary Figure 1. Model training and evaluation

a



b

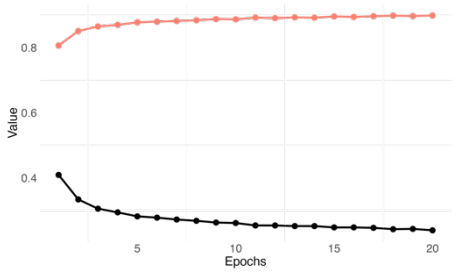
Parameters compared using KerasTuner:

- Network: ResNet50, MobileNetV2, VGG16, InceptionV3
- Dropout: 0.1, 0.2, 0.3
- Initial learning rate: 1e-4, 1e-5, 1e-6
- Fixed parameters: batch size 48, epochs 20, sample cap 500 tiles/tumour, pretrained on Imagenet, tile size 512x512 pixels, decay=0.01

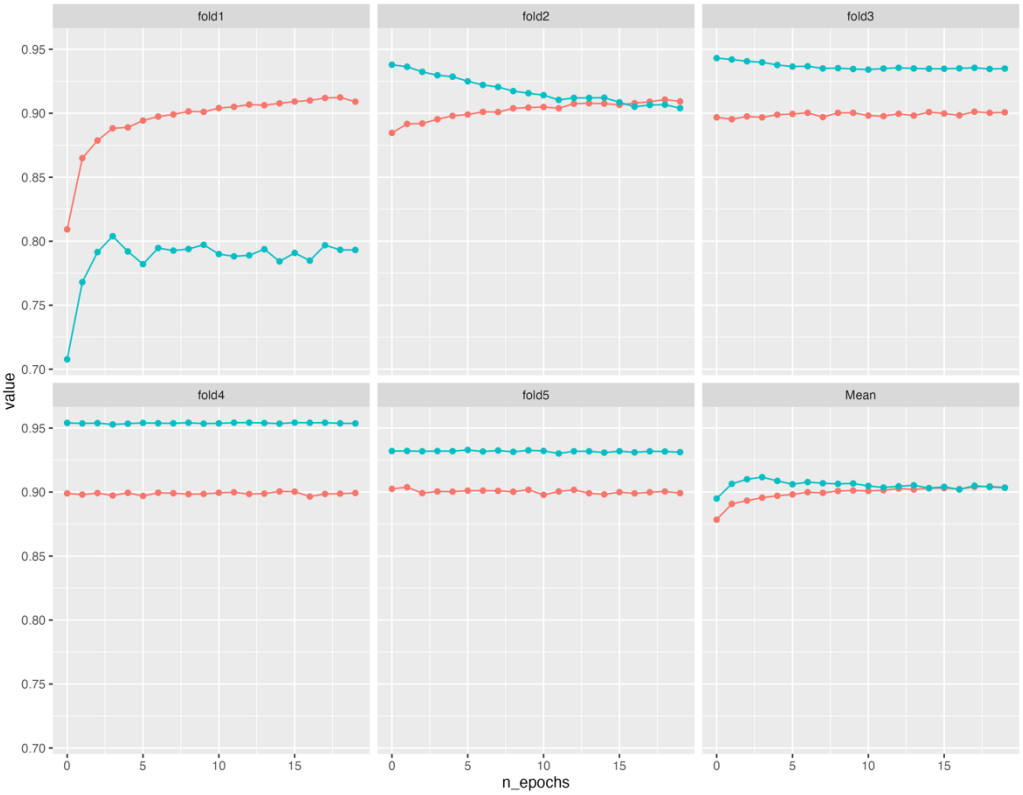
Additional parameters evaluated for Model 2:

- Networks: ResNetV2, Resnet101, InceptionResNetV2
- Batch size: 16, 24
- Initial learning rate: 1e-2, 1e-3
- Dropout: 0.5
- Not pretrained on ImageNet
- Trained on entire WSI (vs limiting to Region of interest)
- Tile size: 256x256 pixels
- Without colour normalisation
- Sample cap: 1000 tiles/tumour

d



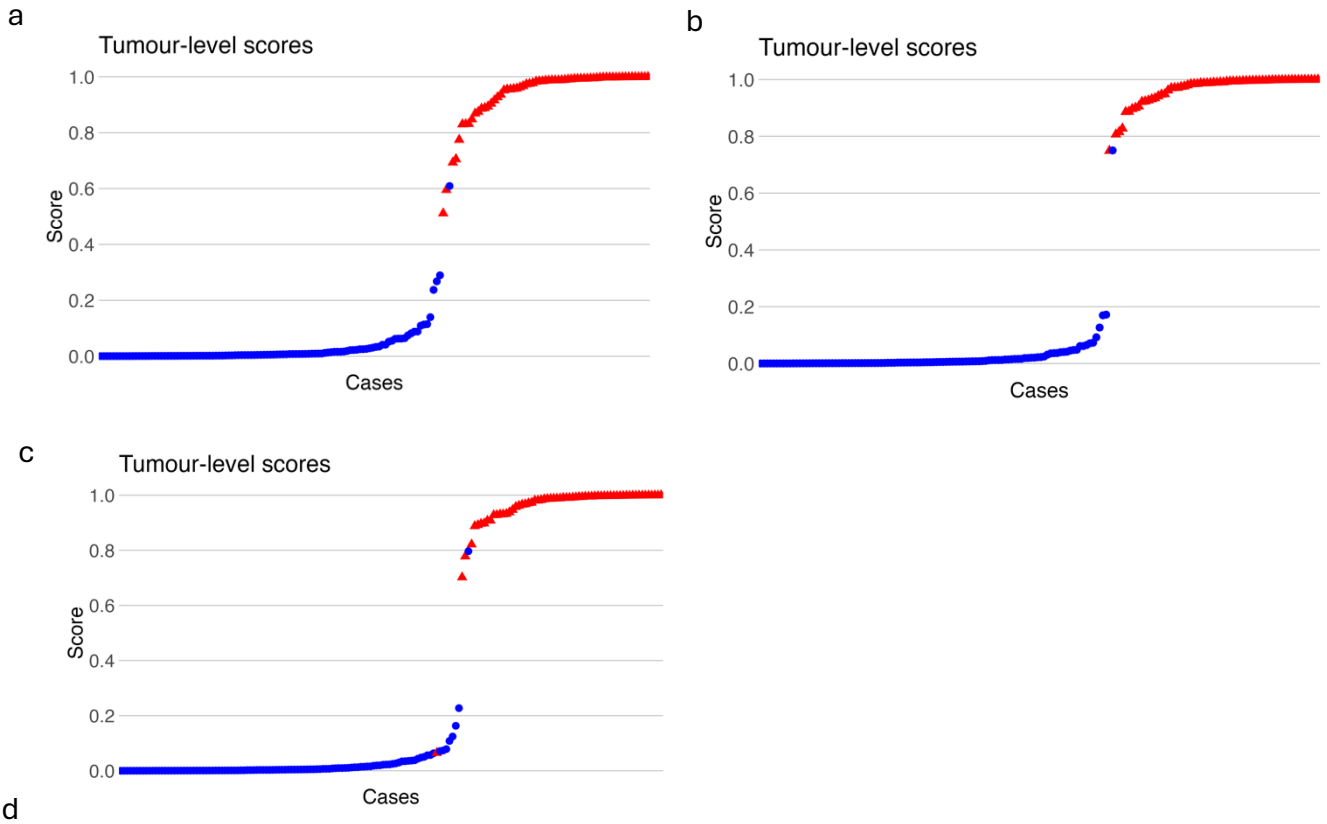
c



a Diagram summarising the model training and evaluation pipeline followed for both models. Firstly, hypertuning using the KerasTuner was used for selection of the ideal model architecture and hyperparameters, through systematic comparison of widely-used convolutional neural network backbones and parameters. The top performing model was then evaluated using 5-fold cross-validation. A final model was then re-trained on the entire

training cohort. To select a threshold for binary classification, the final model was used to generate predictions on the training cohort. Model evaluation was performed by generating predictions on the hold-out testing cohort and comparing model predictions to ground truth. **b** List of parameters compared using the KerasTuner and additional important parameters evaluated individually. **c** Five-fold cross validation curves for Model 2, with training accuracy in red and validation accuracy in blue. **d** Model 2 final training curve, with training accuracy in red and training loss in black. After 20 epochs, the model reached accuracy 0.90 and loss 0.24.

Supplementary Figure 2. Model 2 threshold selection

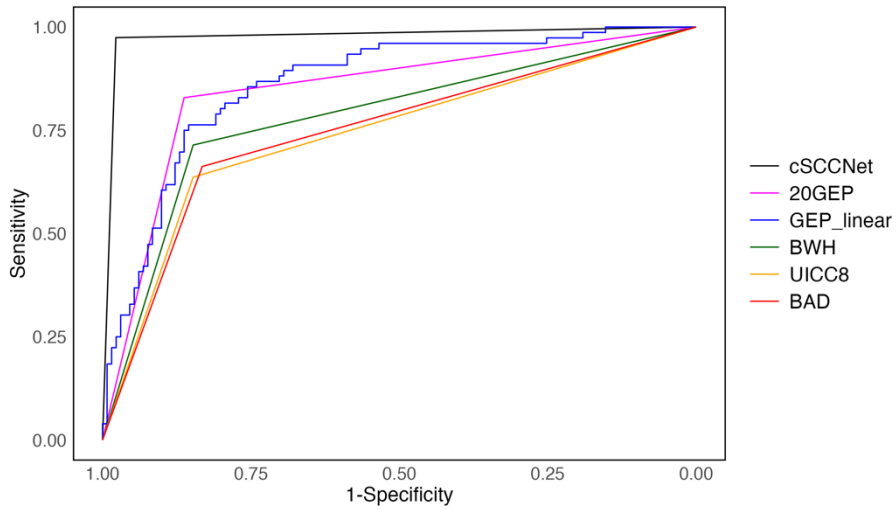


Threshold	TP	FP	FN	TN	accuracy	sensitivity	specificity	PPV	NPV
>50% ≥ 0.5	63	1	1	107	99%	98%	99%	98%	99%
>30% ≥ 0.7	63	2	1	106	98%	98%	98%	97%	99%
>20% ≥ 0.7	64	4	0	104	98%	100%	96%	94%	100%
>10% ≥ 0.9	64	2	0	106	99%	100%	98%	97%	100%
Median >0.8	61	0	3	108	98%	95%	100%	100%	97%
Median >0.2	63	2	1	106	98%	98%	98%	97%	99%

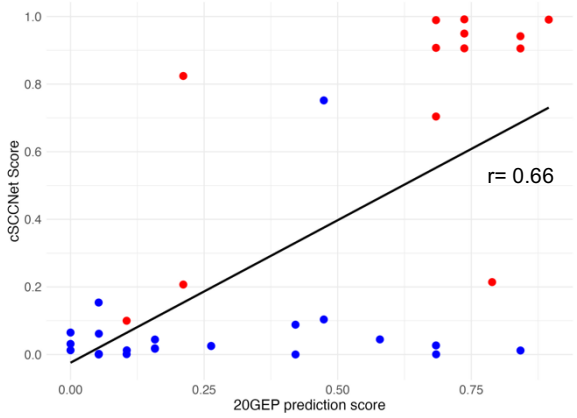
Results given for 172 samples in the training cohort. **a** Scatterplot showing Model 2 median tile scores for metastasising cases as red triangles and non-metastasising cases as blue circles. All tiles within the pathologist-annotated regions of interest (ROI) were included. **b** To improve the separation between low- and high-risk tumours, borderline tiles (with scores 0.3-0.7) were excluded. Scatterplot showing Model 2 median tile scores, for tiles within pathologist ROI and after excluding tiles with borderline scores. A median score >0.2 was selected as the definition for 'high-risk' cSCC, based on graphical representations and accuracy statistics. **c** Both models were then used in series. Scatterplot showing Model 2 median tile scores, for tiles selected by Model 1 and after excluding borderline tiles. **d** Comparison of various Model 2 thresholds in predicting risk of cSCC metastasis when both models were used in series and after excluding borderline tiles. Thresholds: '>50% ≥ 0.5' (>50% of tiles have score ≥0.5), '>30% ≥ 0.7' (>30% of tiles have score ≥ 0.7), '>20% ≥ 0.7' (>20% of tiles have score ≥ 0.7), '>10% ≥ 0.9' (>10% of tiles have score ≥ 0.9), median tile score > 0.8 and median tile score > 0.2. FN: false negatives; FP: false positives; NPV: negative predictive value; PPV: positive predictive value; TN: true negatives; TP: true positives.

Supplementary Figure 3. Model 2 evaluation

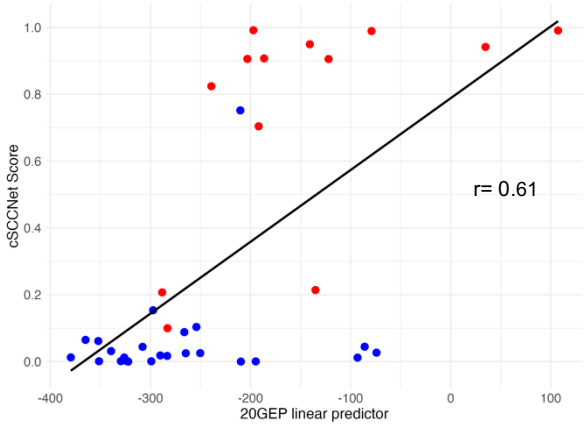
a



b



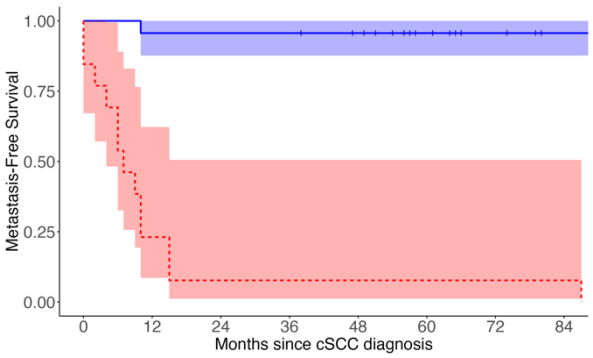
c



d

Variable	Hazard Ratio	95% confidence interval	p-value
cSCCNet	53.4	6.8-417.1	1e-4
20-GEP	7.3	2.0-27.1	0.003
UICC8	4.4	1.5-12.9	0.006
BWH	4.7	1.5-14.5	0.008
BAD	4.4	1.5-12.9	0.006
Diameter	1.1	1.03-1.1	1.4e-5
Differentiation	14.4	1.9-110	0.01
Thickness	1.13	1.0-1.3	0.04
LVI	6.0	1.6-22.9	0.009

e



f

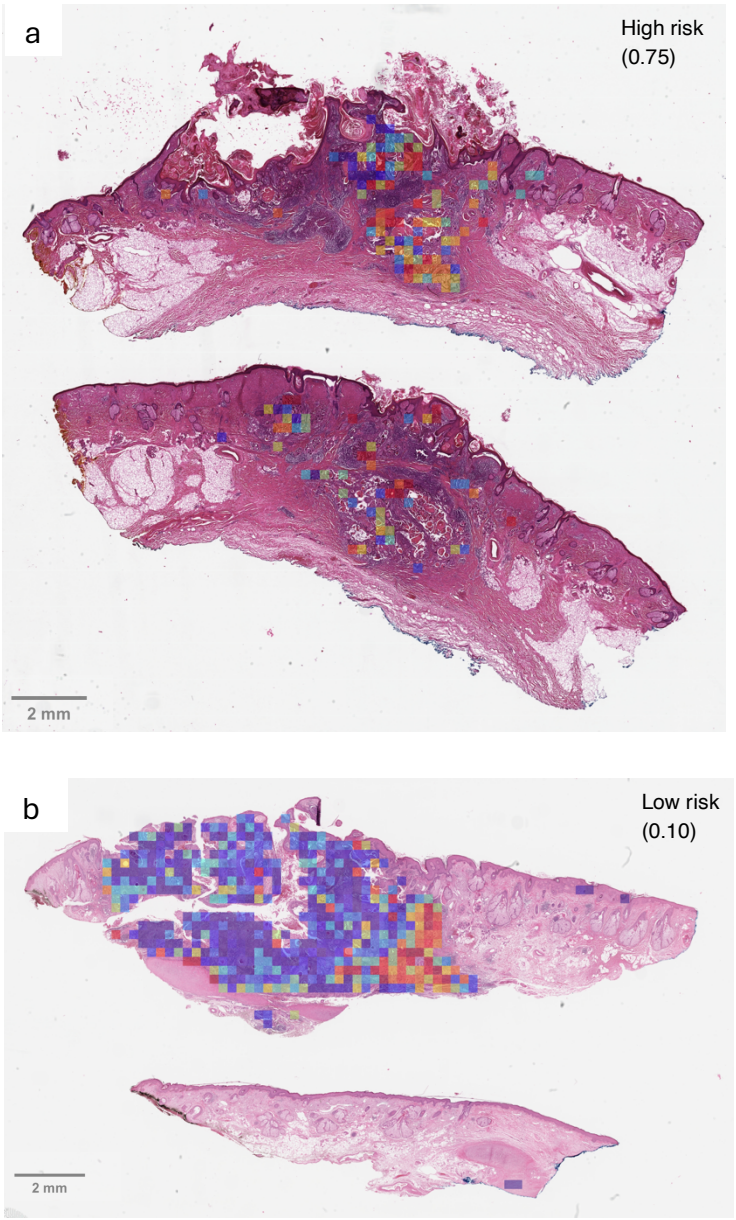
Variable	Hazard Ratio	95% confidence interval	p-value
Age	0.94	0.88-1	0.05
Sex	2.9	0.3-29.3	0.37
Differentiation	0.6	0.05-8.3	0.74
UICC8	2.8	0.8-9.5	0.10
cSCCNet	74.9	4.8-1170.6	0.002

g

Variable	Hazard Ratio	95% confidence interval	p-value
Age	0.93	0.88-0.99	0.02
Sex	2.2	0.2-21.7	0.49
BWH	2.2	0.7-7.4	0.20
cSCCNet	68.4	6.0-785.3	6.9e-4

a Receiver operating characteristic (ROC) curves are shown for cSCCNet (black), the 20-gene expression profile (20-GEP) prediction score based on k-nearest neighbours analysis (20GEP, purple) and linear predictor (GEP_linear, blue), Brigham and Women's Hospital classification (BWH, green), 8th edition Union for International Cancer Control classification (UICC8, yellow), and British Association of Dermatologists' cSCC guidelines (BAD, red) on the entire cohort (n = 212). Of note, this includes the training cases for the cSCCNet and 20-GEP models, which might bias their performance. The area under the ROC curves (AUC) and their 95% confidence intervals were: 0.98 (0.95-1) for cSCCNet, 0.85 (0.79-0.90) for 20-GEP prediction score, 0.86 (0.81-0.91) for the 20-GEP linear predictor, 0.74 (0.68-0.80) for UICC8, 0.78 (0.72-0.84) BWH, and 0.74 (0.69-0.81) for BAD. For 37 cases in the testing cohort where complete data were available, the Pearson correlation between the cSCCNet score and 20-GEP test showed a moderate positive correlation: **b** 0.66 (p = 9e-6) for the 20-GEP prediction score and **c** 0.61 (p = 7e-5) for the 20-GEP linear predictor score. **d** On univariate analysis, features predictive of metastasis (Wald test, $p < 0.05$) in the testing cohort included the cSCCNet classification, 20-GEP, UICC8, BWH, BAD Very High risk grade, tumour diameter, differentiation, thickness, and presence of lymphovascular invasion (LVI). Age, sex, site of primary cSCC and presence of perineural invasion were not statistically significant in the testing cohort; however, all were significant ($p < 0.05$) when assessed in the entire cohort (n=212), suggesting an impact of sample size. Margin status was not a significant predictor of outcome. **e** Kaplan-Meier curve showing metastasis-free survival after cSCC diagnosis, stratified by cSCCNet prediction, with high-risk cases in red and low-risk cases in blue. **f** On multivariate analysis for predicting the risk of metastasis in the testing cohort (n=35), cSCCNet was a significant predictor, independently of age, sex, tumour differentiation, or UICC8. **g** cSCCNet was also a significant predictor when multivariate analysis was repeated with BWH. As differentiation is already included within BWH, it was not included as a separate variable.

Supplementary Figure 4. Review of incorrect cases



Heatmaps for the testing cohort cases misclassified by cSCCNet (n=2/40), with Model 2 tile scores converted to colour using a blue to red scale, for scores 0-1 (low to high-risk). The tumour-level aggregate scores are on the top right corner of each case, with scores >0.20 representing 'high-risk' tumours. **a** Non-metastasising scalp cSCC classified as high-risk by the cSCCNet model and by staging criteria (8th edition Union for International Cancer Control classification, UICC8, and Brigham and Women's Hospital classification, BWH). It is poorly differentiated and invades beyond subcutis. Heatmaps show that Model 1 had failed to select >60% of the ROI; the small number of tiles passed to Model 1 contained poorly differentiated carcinoma. **b** A metastasising pinna cSCC with incomplete excision margins was classified as low-risk by the cSCCNet model. The majority of the tumour was moderately-differentiated with good keratinisation; however, there was extension beyond cartilage. It was high grade on staging criteria (AJCC8 T3/BWH T2b). A small area of poorly differentiated carcinoma was present and was classified as 'high-risk' by the model.

Supplementary Table 1. Baseline clinicopathological characteristics

	Training cohort n = 172[a]	Testing cohort n=40
Metastasising cases Non-metastasising cases	64 (37) 108 (63)	14 (35) 26 (65)
Age, years	80 (71-84)	82 (75-86)
Sex Male Female	 114 (67) 57 (33)	 29 (73) 11 (28)
Site Head and Neck	 112 (65)	 24 (60)
Tumour diameter Median, mm >=20 mm	 15 (10-23) 58/167 (35)	 15 (10-25) 15 (38)
Differentiation Poorly differentiated Moderately differentiated Well differentiated	 87 (51) 63 (37) 22 (13)	 21 (53) 13 (33) 6 (15)
Thickness, mm Median >6mm	 3 24/170 (14)	 3 3/36 (8)
Invasion to Dermis Subcutis Beyond subcutaneous fat	 89 (53) 45 (27) 35 (21)	 24 (60) 9 (23) 7 (18)
Perineural invasion	22/170 (13)	3/39 (8)
Lymphovascular invasion	11/167 (7)	3/38 (8)
Follow-up, months	64 (44-80)	65 (54-80)
UICC8 pT1 pT2 pT3	 96 (57) 16 (9) 57 (34)	 22 (56) 5 (13) 12 (31)
BWH T1 T2a T2b T3	 61 (36) 45 (27) 58 (34) 5 (3)	 12 (31) 15 (38) 10 (26) 2 (5)
BAD LR HR VHR	 45 (27) 64 (38) 60 (36)	 11 (28) 15 (38) 13 (33)

This table does not include the 15 cases used to train Model 1 only. Numbers in brackets are percentages or interquartile range, as appropriate. There were no statistically significant differences ($p > 0.05$) between the training and testing cohorts, using the Mann-Whitney U test for continuous variables and Fisher's exact test for categorical variables. Although all cSCC were treated with wide local excision, 30/210 tumours (21/172 training and 9/40 testing) had incomplete excision margins, and all had either re-excision or adjuvant radiotherapy, except one case in the training cohort which had already metastasized at the time of presentation of the primary lesion. [a]172 tumours from 171 patients. BAD: British Association of Dermatologists' cSCC guidelines; BWH: Brigham and Women's Hospital classification; UICC8: the 8th edition Union for International Cancer Control classification.