npj | precision oncology

Article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-025-00997-4

Explainable, federated deep learning model predicts disease progression risk of cutaneous squamous cell carcinoma

Juan I. Pisula[1,2,12], Doris Helbig[3,12], Lucas Sancéré[1,2], Oana-Diana Persa[4], Corinna Bürger[2,5,6], Anne Fröhlich[7], Carina Lorenz[2,5,6], Sandra Bingmann[3], Dennis Niebel[8], Konstantin Drexler[8], Jennifer Landsberg[7], Roman Thomas[5,9,10], Katarzyna Bozek[1,2,11,13] & Johannes Brägelmann[2,5,6,13] ✉

Predicting cancer patient disease progression is a key step towards personalized medicine and secondary prevention. Risk stratification systems based on clinico-pathological criteria aim to identify high-risk patients, but accurate predictions remain challenging. Deep learning models present new opportunities for patient risk prediction, yet their interpretability has been largely unexplored. We developed a transformer-based approach for predicting progression of cutaneous squamous cell carcinoma (cSCC) patients based on diagnostic histopathology tumor slides. Our initial model showed AUROC = 0.92 on a held-out test set, with average AUROC of 0.65 on external validation cohorts. To further increase generalizability and reduce potential privacy concerns, we trained the model in a federated manner across three clinical centers, reaching AUROC = 0.82 across all cohorts, with image-based risk scores achieving hazard ratios up to 7.42 ($p < 0.01$) in multivariable analyses. Through interpretability analysis, we identified spatial and morphological features predictive of progression, suggesting that tumor boundary information and tissue heterogeneity characterize progressive cSCCs. Trained exclusively on routine diagnostic slides and offering biological insights, our model can improve secondary prevention and understanding of cSCC while enabling deployment across clinical centers without administrative overheads or privacy concerns.

Cutaneous squamous cell carcinoma (cSCC) is the second most prevalent type of non-melanoma skin cancer that is diagnosed in 1 million patients in the USA every year[1]. In the last decades, the incidence of cSCC has risen sharply and is projected to increase further[2]. Even though the majority of cSCCs can be removed by surgical excision, a relevant fraction of patients experiences disease progression by local recurrence or metastases to lymph nodes or other body sites, which is associated with poor prognosis and increased risk of death[3–6]. Due to the high incidence of cSCC, this poses a significant public health concern. Reliable predictors are thus needed to decide which patients will benefit from enhanced secondary prevention e.g. by more frequent follow-up care or additional treatments such as immuno-, chemo- or radiotherapy. Current cSCC staging systems like the American Joint Committee on Cancer (AJCC), the Brigham Women's Hospital (BWH), or the National Comprehensive Cancer Network (NCCN) staging systems aim to provide guidance on risk stratification and clinical management of cSCC patients[7,8]. However, they fall short of reliably identifying patients at high risk of disease progression. Recently, multi-gene expression signatures have been used to predict metastasis risk of cSCCs[9,10]. While these

[1]Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Köln, Germany. [2]Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine and University Hospital Cologne, University of Cologne, Köln, Germany. [3]Department for Dermatology, University Hospital Cologne, Cologne, Germany. [4]Department of Dermatology, Technical University Munich, Munich, Germany. [5]University of Cologne, Faculty of Medicine and University Hospital Cologne, Department of Translational Genomics, Cologne, Germany. [6]University of Cologne, Faculty of Medicine and University Hospital Cologne, Mildred Scheel School of Oncology, Cologne, Germany. [7]Department of Dermatology and Allergology, University Hospital Bonn, Bonn, Germany. [8]Department of Dermatology, University Medical Center Regensburg, Regensburg, Germany. [9]Institute of Pathology, Medical Faculty, University Hospital Cologne, University of Cologne, Cologne, Germany. [10]DKFZ, German Cancer Research Centre, German Cancer Consortium, Heidelberg, Germany. [11]Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Köln, Germany. [12]These authors contributed equally: Juan I. Pisula, Doris Helbig.[13]These authors jointly supervised this work: Katarzyna Bozek, Johannes Brägelmann. ✉e-mail: johannes.braegelmann@uni-koeln.de

signatures help to predict metastasis risk, they have not yet been used to predict local recurrences. In addition, they require measurement of gene expression from patient samples, which limits their potential for translation into clinical routine use.

In addition to clinical parameters such as immunosuppression, several pathological tumor features such as perineural involvement, tumor size, and invasion depth have been associated with increased risk of cSCC progression and are part of existing staging systems like the NCCN risk stratification[4–6,8]. Moreover, specific histological subtypes e.g. desmoplastic cSCC have been linked to higher recurrence and/or metastasis risk[6]. Morphology in histological specimens thus holds information on progression risk. Since deep learning has matched human experts in cancer detection and classification[11], computational pathology methods hold promise to extract information on patient progression from histopathology image data. Building robust models that offer high predictive power across data independent of their source, requires multi-institutional data sets for model training. Obtaining such data sets poses challenges regarding data governance and raises concerns about patient privacy. Federated Learning (FL) is a strategy that limits the logistic overhead and reduces privacy concerns in training a multi-center-based model[12,13]. Moreover, FL simplifies the inclusion of new patients and cohorts for further model training, which in turn facilitates model update, continuous improvement, and clinical applicability.

Here, we present a multiple instance learning transformer-based deep learning model for prediction cSCC progression risk using Hematoxylin-Eosin-(HE-) stained histopathology images acquired during routine care (Fig. 1)[14,15]. Our model, trained in a federated manner on cohorts from three clinical centers, achieved high accuracy in predicting patients at risk of disease progression, which corresponds to significant differences in progression-free survival. We developed explainability methods on our model which provide insights into the tissue areas and cell features associated with increased progression risk. Overall, we present a powerful approach that improves risk-stratification of cSCC patients and offers insights into the underlying cancer biology.

Results

Deep learning on histopathology images predicts cSCC progression risk

Even though specific pathological factors like perineural and lymphatic invasion are established parameters for risk stratification in cSCC, a systematic deep learning approach to comprehensively evaluate histophathological features for progression prediction is currently lacking. Currently, it is not clear if the progression risk of a cSCC can be inferred from a histopathology slide and which elements of the tumor and its microenvironment are decisive of disease progression. To fill this gap, we used a multiple instance learning, transformer-based classifier for the task of progression prediction from Whole Slide Images (WSIs). We trained the model in a federated manner, leveraging data from three different medical centers (Fig. 1)[14,15].

Initially, we trained our model on the Cologne cohort only ($n = 157$ patients, 214 WSIs), achieving cSCC progression status classification accuracy of 0.92 AUROC (95% CI = [0.83–1.00]) in a held-out test set from Cologne (Fig. 2A). In comparison, a multivariable logistic regression model incorporating clinico-pathological parameters associated with risk of disease progression (Suppl. Fig. 1) achieved an AUROC of 0.63 (95% CI = [0.52–0.75]) in the same prediction task and cohort (Fig. 2B). To test the robustness of our deep learning model we assembled two additional cohorts from dermatology departments at the University Hospital Bonn (Bonn cohort, $n = 35$ patients, 133 WSIs) and the Technical University Munich (Munich cohort, $n = 51$ patients, 113 WSIs). While the model trained on the Cologne cohort performed well on the Bonn cohort (AUROC = 0.90, 95% CI = [0.71–0.97]), it failed to generalize to the Munich cohort (AUROC = 0.46, 95% CI = [0.30–0.63]; Fig. 2A). A further analysis

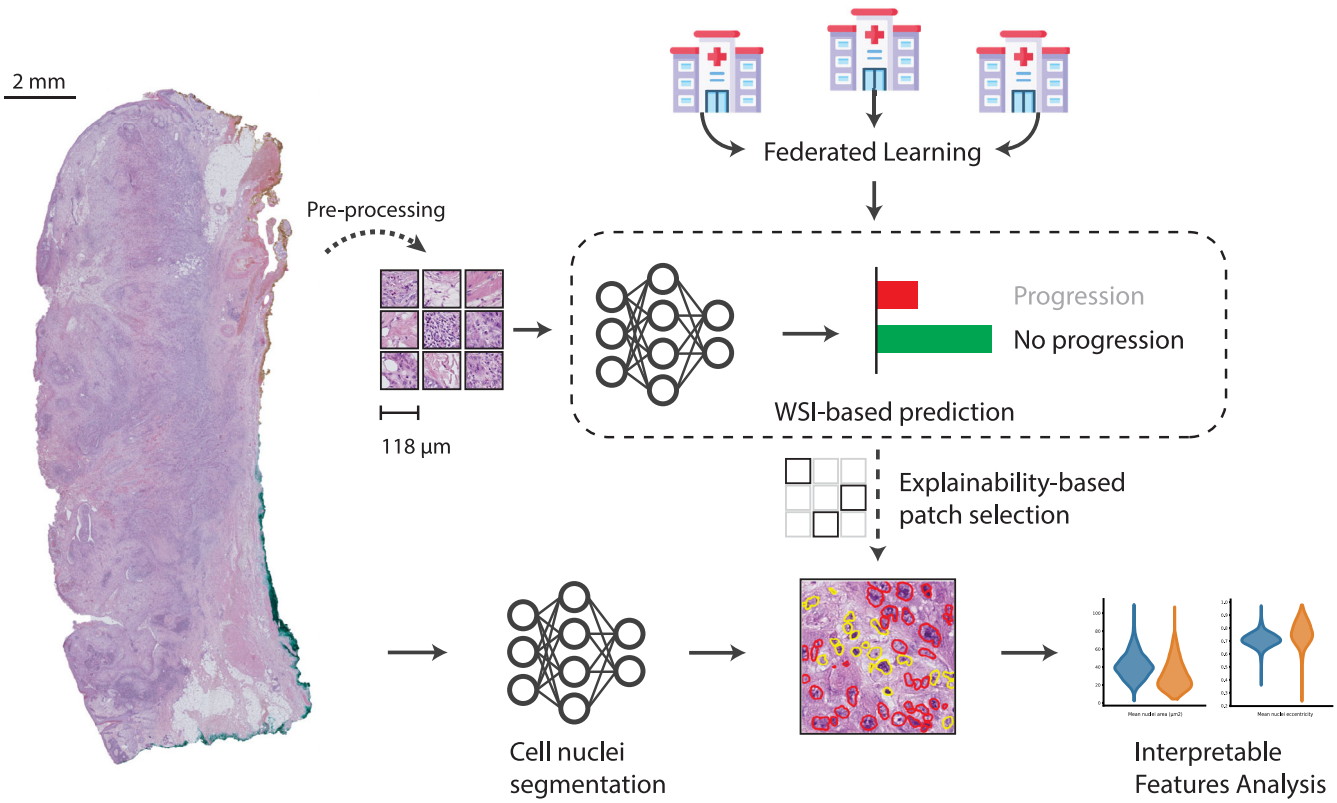

Fig. 1 | We propose a WSI-based cutaneous Squamous Cell Carcinoma (cSCC) progression prediction model, trained on data from three medical centers using Federated Learning. Beyond prediction, we investigate underlying biological features that influence our classifier. We do so by computing cellular-level features with aid of a nuclei segmentation model. We analyze these features in image regions detected as relevant for prediction outcome by Integrated Gradients, an input attribution algorithm for explainable deep neural networks.

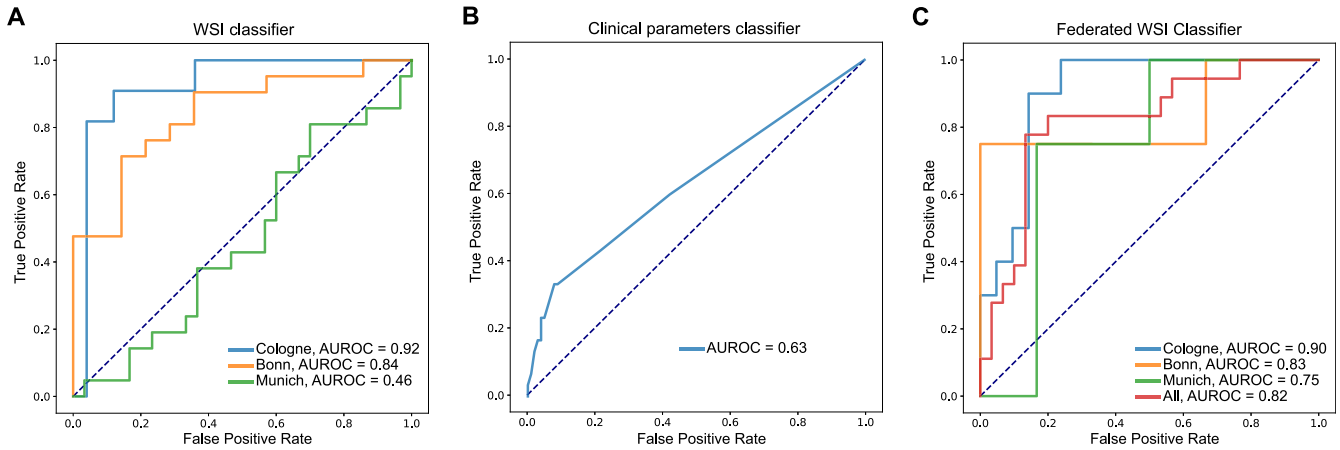

Fig. 2 | ROC curves of the classifiers. A WSI-based classifier trained exclusively on the Cologne cohort and tested on Munich and Bonn cohorts (AUROC = Area under the receiver operator curve). **B** Multivariate logistic regression model based on clinico-pathological parameters associated with progression risk in univariate analysis. Model trained and evaluated on the Cologne cohort. **C** Federated WSI-based classifier.

of this phenomenon indicates that the performance gap is due to systemic differences in image appearance rather than biological differences between the cohorts. The distributions of clinico-pathological variables shown in Table 1 appear comparable, except grading, making it unlikely that these factors contribute to the performance discrepancy. Suppl. Fig. 2 shows UMAP plots of the mean feature vector representations of the slides as computed by the EfficientNet neural network used in our pipeline[16], and by an additional CTransPath model[17]. Both plots indicate distinct visual differences between the two cohorts, potentially arising from differences in tissue processing. This highlights that variation induced by technical procedures, or distribution shift and domain adaptation problems may hamper generalizability of models trained on a single-center cohort.

Federated learning improves generalizability of image-based classification

To improve performance across cohorts, it is crucial to train deep learning models on large and diverse datasets. However transfer of patient data and histological slides across hospitals carries important logistic complexity and poses potential privacy threats. We therefore trained our model in an FL scheme on all three cohorts (Fig. 1)[12]. FL overcomes the data sharing hurdles by reducing the organizational overhead of combining different patient cohorts, since patient data can remain in the respective hospital. Model training is performed locally and only model parameters are shared between the hospitals. Moreover, it enables dynamic patient enrollment and facilitates inclusion of additional centers, which in turn increases its flexibility and the opportunities for clinical deployment. Training on the multi-institutional cohorts using the FL framework did indeed improve model performance. While AUROC on Cologne and Bonn decreased at most by 2%, performance on the Munich cohort increased by 63%, leading to prediction accuracy of AUROC = 0.82 (95% CI = [0.69–0.95]) in the complete dataset (Fig. 2C). This highlights that prediction of disease trajectories is indeed possible for cSCC patients and can be achieved with a deep learning model trained on different cohorts in a federated manner. Such prediction opens possibilities for clinical translation of the model as a tool for the identification of patients at high recurrence risk that may benefit from increased surveillance.

Image-encoded information has higher discriminative power than clinical variables

Several clinico-pathological parameters have been associated with increased risk of disease progression, such as immunosuppression, perineural involvement, tumor size, and invasion depth[4,6]. Similarly, desmoplastic cSCC histology has been linked to higher recurrence and/or metastasis risk[6]. In this experiment, we used the logit output of the deep learning models as a progression risk score, and compared it against clinico-pathological parameters available for the Cologne and Bonn cohorts. Our analyses were done separately on the Cologne cohort, to which we applied a new federated classifier trained solely on the Bonn and Munich cohorts, and on the Bonn cohort, to which we applied the original Cologne model.

From all the variables in the experiment, the image-based models' scores were the most discriminative: risk of progression was 4.2 times higher for high- compared to low-risk patients in the Cologne cohort, and 8.25 times higher in the Bonn cohort, according to univariate Cox proportional hazard models (Fig. 3A, C). These covariates were followed by perineural invasion in Cologne (Suppl. Fig. 3A, hazard ratio = 3.58, $p = 0.004$) and tumor diameter in Bonn (Suppl. Fig. 4A, hazard ratio = 5.35, $p = 0.19$).

Lastly, we joined the informative factors of our clinico-pathological parameters with the deep learning model's output to predict survival using a multivariable model. To this end we combined the deep learning models' predicted risk scores and clinico-pathological parameters that showed a p-value below 0.1 in univariate analyses in multivariable Cox regression models. These combined models showed that the image data carries more information than the clinico-pathological variables (global $p < 0.01$, Fig. 3B, D). In fact, classification of patients in the Cologne cohort as high-risk using the model trained on Bonn and Munich carries a hazard ratio of 5.96 even when adjusting for additional variables (multivariable p-value = 0.001). Similarly, patients classified as high-risk in the Bonn cohort using a model trained on Cologne only have a hazard ratio of 7.42 in a multivariate analysis (multivariable p-value = 0.01). Only the Cologne cohort model exhibits other variables that remain significant: invasion beyond subcutaneous tissue (hazard ratio = 4.53, multivariable p-value = 0.007), and tumor thickness greater than 6 mm (hazard ratio = 2.54, multivariable p-value = 0.026).

Considering that only a fraction of patients shows perineural invasion, vascular invasion, or invasion beyond subcutaneous tissue, and that clinico-pathological information is frequently incomplete, these analyses highlight the potential of our image-based model to reliably identify patients at high risk of disease progression for intensified clinical follow-up.

Explainability analyses highlight factors associated with cSCC progression

In addition to stratifying patients according to their disease progression risk, we assessed which parts of the histological images are predictive of disease progression. In these experiments, we inspected the slides of the independent test set of the federated model. We used Integrated Gradients (IGs) attributions to infer which areas in the WSIs are the most relevant for the federated model to predict the respective patient as

Table 1 | Clinicopathological characteristics of the different cSCC cohorts. *p*-values correspond to chi-squared tests for categorical variables and t-tests for continuous variables

	Cologne	Munich	Bonn	*p*
Patient characteristics				
Patient number	166	51	35	—
Sex				
Male	125 (75.3%)	—	25 (71.4%)	0.632
Female	41 (24.7%)	—	10 (28.6%)	
Age				
>80	80 (48.2%)	21 (41.2%)	17 (48.6%)	0.633
≤80	85 (51.2%)	27 (52.9%)	18 (51.4%)	
Unknown	1 (0.6%)	3 (5.9%)	—	
Age at initial diagnosis (mean ± std)	78 ± 9	79 ± 9	77 ± 9	0.549
Immunosuppression				
Yes	43 (25.9%)	6 (11.8%)	7 (20.0%)	0.137
No	123 (74.1%)	42 (82.4%)	28 (80.0%)	
Unknown		3 (5.9%)		
Tumor characteristics				
Tumor number	219	51	35	
with progress	63 (28.8%)	22 (43.1%)	14 (40.0%)	0.085
without progress	156 (71.2%)	29 (56.9%)	21 (60.0%)	
WSI (total)	219	129	291	—
with progress	66 (30.1%)	70 (52.3%)	214 (73.5%)	
without progress	153 (69.9%)	59 (47.7%)	77 (26.5%)	
WSI (detected tumor)	214	113	133	—
with progress	62 (29.0%)	61 (54.0%)	105 (79.0%)	
without progress	152 (71.0%)	52 (46.0%)	28 (21.0%)	
Invasion depth				
≤6.00 mm	149 (68.0%)	31 (60.8%)	19 (54.3%)	0.083
>6.00 mm	47 (21.5%)	16 (31.4%)	13 (37.1%)	
Unknown	23 (10,5%)	4 (7.8%)	3 (8.6%)	
Grading				
G1	137 (62.6%)	13 (25.5%)	11 (31.4%)	< 0.001
G2	35 (16.0%)	12 (23.5%)	19 (54.3%)	
G3	21 (9.5%)	12 (23.5%)	5 (14.3%)	
G4	12 (5.5%)	8 (15.7%)		
Unknown	14 (6.4%)	6 (11.8%)		
Desmoplasia				
Yes	3 (1.4%)	—	—	1
No	216 (98.6%)	48 (94.1%)	—	
Unknown	—	3 (5.9%)	—	
Perineural Invasion				
Yes	11 (5.0%)	—	—	0.368
No	208 (95.0%)	—	33 (94.3%)	
Unknown	—	—	2 (5.7%)	

progressor/non-progressor[18]. Additionally, we leveraged a separate pipeline we recently established specifically for cSCC, which performs nuclei segmentation and classification of cells into one of six cell types (granulocyte, lymphocyte, plasma, stroma, tumor, and epithelial cell)[19]. We used the cell type detection and classification to analyze the WSI regions with the highest predicted power as attributed by IGs. In the WSI regions with high IGs attribution score we calculated various features of nuclei morphology, cell type composition and spatial distribution (Suppl. Table 1).

We next performed statistical analyses of these features to gain insights into the determinants of cSCC progression. Interestingly, many of the predictive tiles with the highest attribution score for disease progression were outside of the tumor region (Fig. 4A). In fact, attribution scores were low in areas with high tumor cell density, as determined using our cell type classification pipeline (Fig. 4A, bottom left and middle)[19]. Instead, they were high at the tumor border and frequently in areas where the most common cell type was stroma (Fig. 4A, bottom right, Suppl. Fig. 5).

In contrast, for patients without disease progression, the most predictive tiles were located within the tumor and in areas with high tumor cell density. Areas outside the tumor border were, in the case of these patients, not of high value for prediction of non-progression (Fig. 4B). This highlights that different parts of histological sections contain information that distinguishes patients at high vs. low risk of disease progression and that such patient stratification needs to be based not only on the tumor but also its surroundings for adequate predictions.

Additionally, we systematically compared the cell-based features between the tiles that were regarded as most predictive for disease progression or non-progression according to their IGs scores. Numerous parameters with significantly different distributions between the two groups were detected, Fig. 5 shows a subset of the tumor-cell-related features. Non-progressors e.g. showed higher values in Average Nearest Neighbor Ratio (ANNR), indicating a higher uniformity in the way tumor cells were distributed (Fig. 5A, $p < 0.0001$), while progressors had more intermixing of tumor cells with other cell types, i.e. more heterogeneity in tissue composition (Fig. 5C, $p < 0.0001$). Moreover, tumor cells of non-progressors showed differences in their morphology compared to progressors such as larger nucleus size (Fig. 5B, $p < 0.0001$) and lower nuclear eccentricity (Fig. 5D, $p < 0.0001$). In addition, tumors of patients that later experienced disease progression showed higher degrees of nuclear dysmorphia and pleomorphism compared to non-progressors. Tumor cells from non-progressors have larger values of morphological solidity and extent (larger median, negatively-skewed distributions, Suppl. Fig. 6A–D, Suppl. Table 1), while morphological extent has a larger variance in tumor cells from progressors (Suppl. Fig. 6E, Suppl. Table 1).

Further analyses were conducted to corroborate the validity of these results. Calculating features shown in Fig. 5 separately for the tumor border and the inner tumor highlights that both regions exhibit similar feature distributions (Suppl. Fig. 7). Given the performance gap between the Cologne and Munich cohorts in the image-based progression prediction models, we investigated the biological differences between these centers according to our engineered features. Remarkably, the centers display similar feature distributions (Suppl. Fig. 8), and the features exhibit a stronger association with disease progression than with the center of origin (Suppl. Fig. 9).

We next tested whether our cell-based features are sufficient to predict the progression/non-progression of patients based on their respective image tiles using a tree-based classification algorithm XGBoost[20]. Interestingly, using the cell-based features as input resulted in high prediction accuracy (Suppl. Fig. 10, AUROC = 0.98, 95% CI = [0.97–0.99]). This highlights that these features, which we computed using an independent pipeline, do indeed capture relevant biological parameters and variation associated with progression risk of patients. Thus the cellular and morphological features are making explicit the morphological and structural components of the tissues and cells that the deep learning model learned implicitly.

Overall, our explainability analyses indicate that tumor cell-intrinsic properties as well as composition of the microenvironment and growth patterns of the tumor are associated with the difference in prognosis and are captured by our deep learning model to accurately predict progression risk.

Discussion

Deep learning has enabled automation of the analysis of large histopathology images. These digital pathology methods not only provide fast

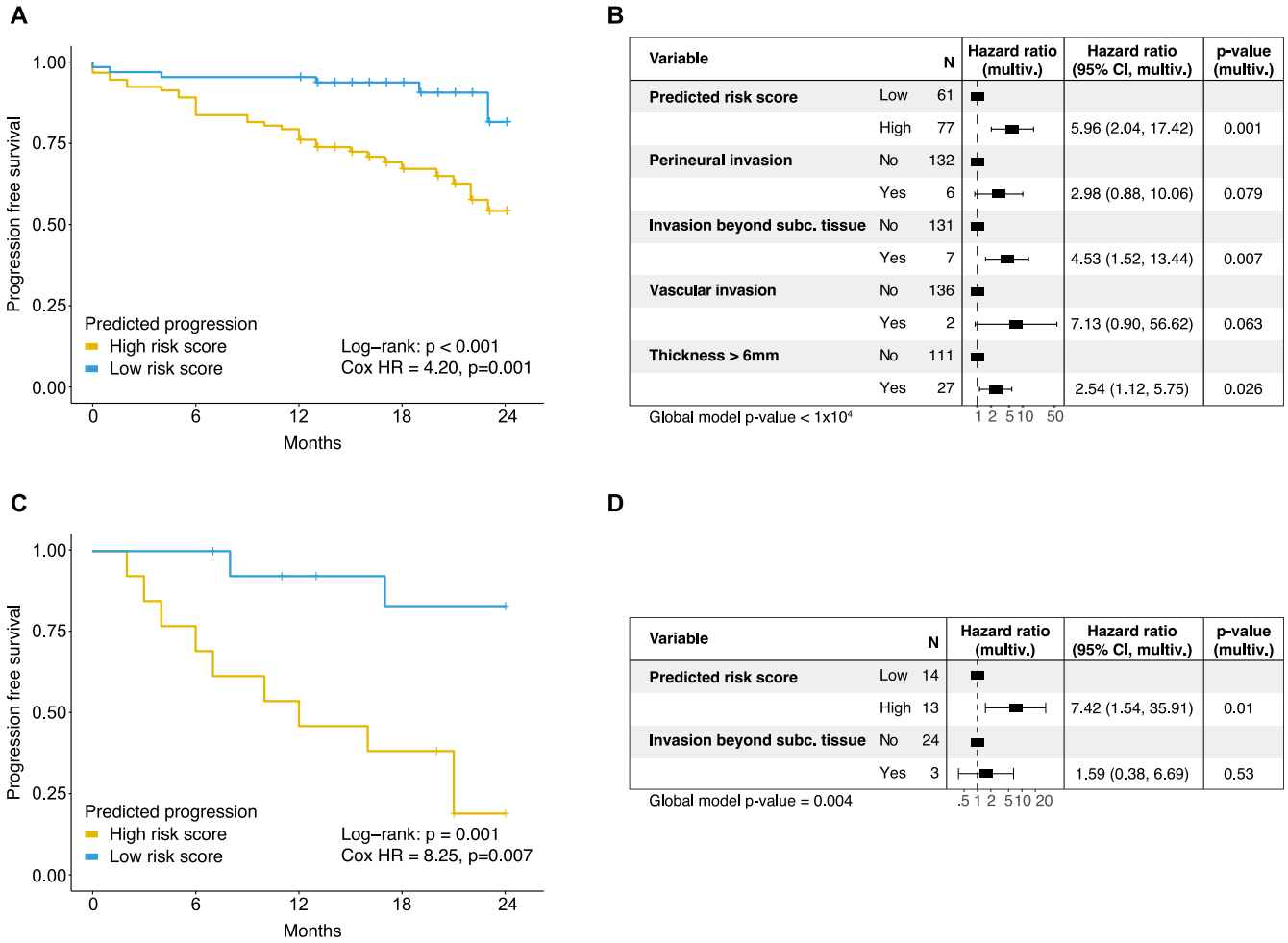

Fig. 3 | Comparison of federated and original deep learning models for survival prediction. A, B Federated model trained on Bonn and Munich cases, applied to Cologne patients. **C, D** Original model trained on Cologne, applied to Bonn patients. **A** Progression-free survival of Cologne patients classified as high vs. low progression risk based on federated deep learning prediction (threshold: Youden index, HR from univariate Cox regression). **B** Multivariable Cox regression for $n = 138$ Cologne patients, integrating federated deep learning risk categories with clinical parameters. **C** Progression-free survival of Bonn patients classified using the original Cologne model (threshold: Youden index, HR from univariate Cox regression). **D** Multivariable Cox regression for $n = 27$ Bonn patients, combining deep learning risk categories with clinical parameters.

and detailed insights into the cellular composition of massive WSIs[21,22], but also allow to identify patterns and anomalies that may be imperceptible to the human eye[23]. Here we present an approach that combines both: a model that detects complex, imperceptible morphological features of a tumor sample that are predictive of patient outcome with an explainability procedure to disentangle what these features are. While patient outcomes might be influenced by multifactorial clinicopathological variables and span variable development trajectories, we demonstrate that, in case of cSCC, prediction of patient progression is possible based on histological images of their tumor samples alone. Via a comprehensive and quantitative analysis of predictive regions of the tumor samples we point to consistent and repetitive patterns in tumor and tumor microenvironment morphology and organization that characterize progression and non-progression patient groups. Our model offers unmatched accuracy compared to the prediction based on clinicopathological features that were the gold standard up till now.

Our analysis combined data from three academic clinical centers: Cologne, Munich, and Bonn. The model trained on a single cohort resulted in an uneven accuracy on the remaining two cohorts, ranging from random predictions to 0.84 AUROC. Our additional analyses (Suppl Fig. 2,8,9) suggest that the performance difference observed between the Cologne and Munich cohorts is more likely due to systemic variations in image appearance, potentially stemming from disparities in tissue processing, such as reagent concentrations or processing times, rather than biological differences between cohorts. Apart from grading, the distribution of clinicopathological variables appears comparable, making it unlikely that these factors contribute to the performance discrepancy. Analysis of features extracted from cell nuclei segmentation further supports this, indicating similar distributions between centers and a stronger association with disease progression than with the center of origin. While digital pathology models require large and multi-center data for better generalization, clinical data sharing carries important administrative hurdles and data protection risks. Here we demonstrate that these difficulties can be overcome by employing an FL training scheme resulting in a model with high accuracy across all cohorts while circumventing cross-center data sharing. Our model development strategy allows for easy incorporation of additional clinical centers in the future which could potentially improve the prediction accuracy further.

Deep learning models have achieved human expert-level accuracy in standard diagnostic tasks such as tumor metastases detection and cancer subtyping[24–26]. These tasks involve detecting patterns that, while sometimes local, subtle, and difficult to notice, are known and described in pathology textbooks. In contrast, while some histopathological parameters may be indicative of disease progression[4–8], predicting patient outcomes based on WSIs remains a complex task. Numerous studies address prediction of cancer progression based on HE-stained samples of tumors across diverse

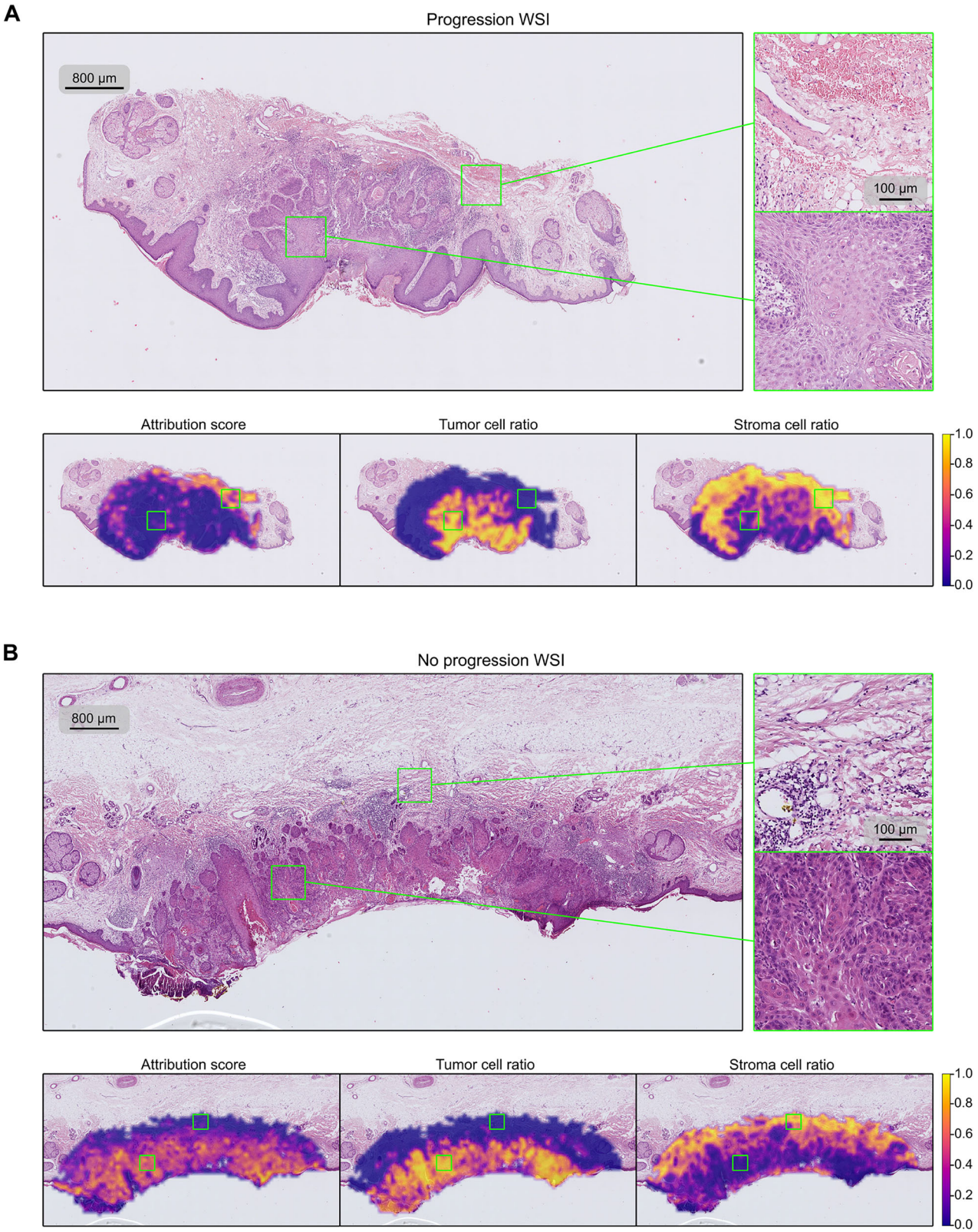

Fig. 4 | Slides and heatmaps of the patches' classifier attribution score, tumor cell ratio, and stroma cell ratio. A Slide of a progression patient, showing that the WSI-based classifier assigns higher importance to the region outside the tumor area (indicated by the tumor cell ratio heatmap). **B** Slide of a non-progression patient, where the high attribution area coincides with the tumor-cell populated areas. Colorbar indicates the slide-normalized heatmap values.

tissue types[27–32], however rarely reaching accuracy > 0.80 AUROC. Notably, combining image with clinical data has improved prediction accuracy in some studies still barely exceeding 0.80 AUROC[33–35]. In cSCC research, the work of Coudray et al. addresses the prediction of disease outcome from WSIs using a bag of visual words classifier, achieving AUROC = 0.689[36]. These examples demonstrate that prediction of patient progression is indeed difficult, and that the accuracy of our model is among the best achieved so far.

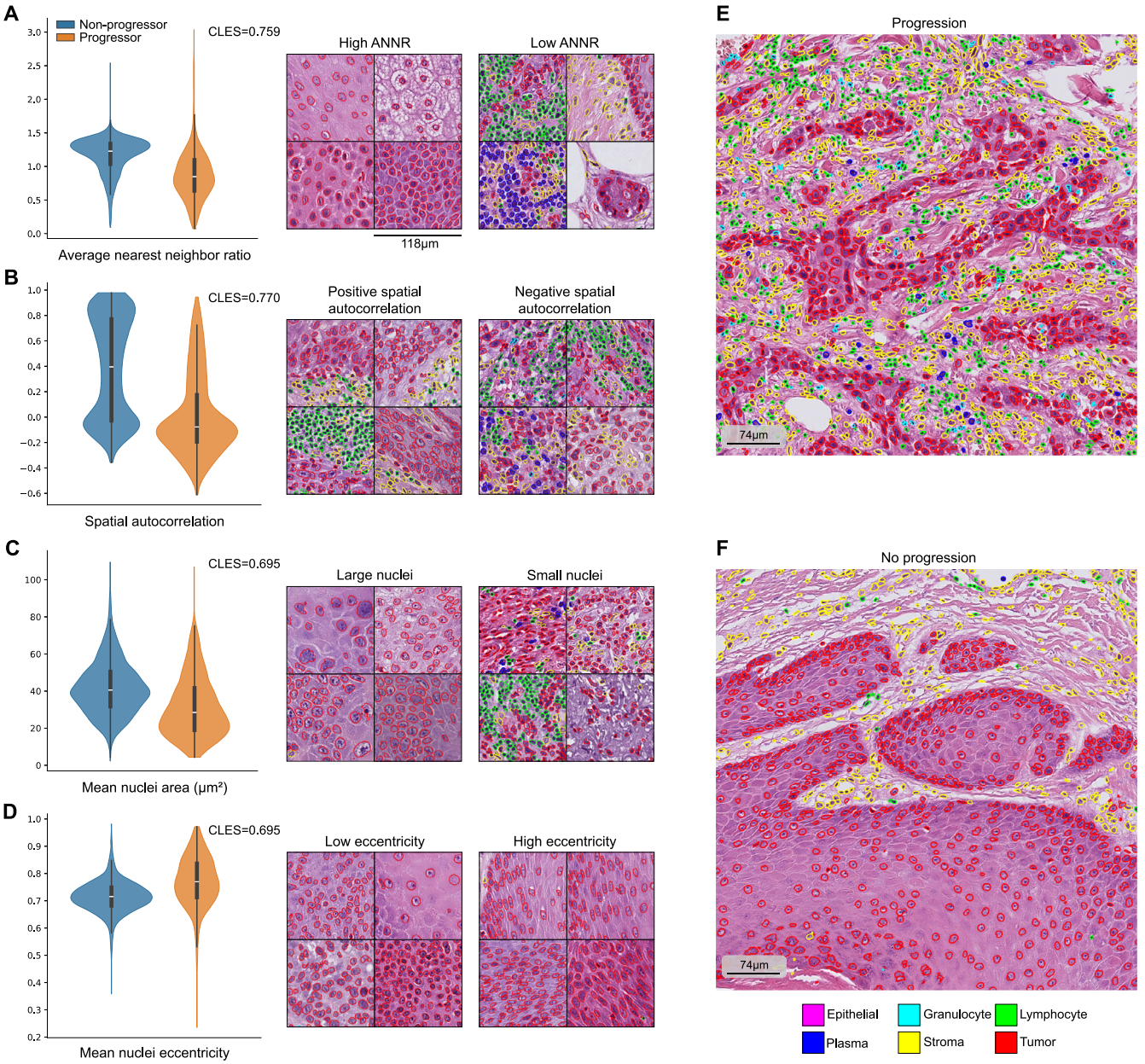

Fig. 5 | Four of the features of the tumor cells used in the analysis. A–D show violin plots and segmented image patches that illustrate these values. In general, progression-associated tumor cells cluster together (**A**), interface with other cell types (**B**), and have smaller (**C**), eccentric nuclei (**D**). These effects are not just local to image patches, but they occur in larger regions, as shown in (**E**, **F**). The displayed CLES (Common Language Effect Size) values are indicated for the group with the largest mean. All features are significantly different in both groups, with p-values < 0.0001 using Mann–Whitney U-test.

Strikingly, progression risk of a patient could be predicted based on histology images alone, exceeding by far the accuracy achieved by a model trained on clinico-pathological features. Unlike clinical parameters[7,8], or gene expression measurements[9,10], which in different clinical centers might follow different standards, be done selectively for some patients only, and come with a high cost, histology is routinely performed in cSCC diagnosis. The fact that tissue slides are available for every patient and that prediction is fast and free of additional costs, considerably increases the facility and potential of our model for clinical use. Moreover, by obviating the need for data sharing, FL greatly facilitates further model training and refinement and its extension to additional centers.

Unlike prediction based on clinical parameters, which are numeric and unambiguous, prediction based on image data is not easy to interpret. Commonly, multiple instance learning models are interpreted using qualitative inspection of image regions with high attention scores[24–26]. Here we adopt a fully quantitative and systematic approach to model interpretation in which we filter predictive patches of each patient group and statistically compare over 524 cell-based features between the two groups. Our features are based on a segmentation model specifically designed for this tumor type and capture a broad range of aspects of sample cell composition, spatial organization of the tissue, as well as nuclei morphology[19]. We point to several noticeable differences in tumor morphology between progressing and non-progressing patients.

Interestingly, the most predictive patches of disease progression were located outside of the tumor region. In contrast, in patients without disease progression, the predictive patches were inside the tumor according to our IGs-based analysis. On the level of cellular morphology and tissue architecture, tumors from patients with disease progression exhibited a higher degree of heterogeneity. Parameters quantifying nuclear morphology showed higher variability and in these patients, cells in the tumor tissues

showed a less uniform distribution. Different areas in and around the cSCC tumor, as well as features of cellular morphology may play distinct roles in the propensity for local recurrence and/or metastatic spread.

Future studies in additional cohorts, ideally together with genomic and transcriptomic experiments will be instrumental to further validate our model and infer cause-and-effect relationships between morphological findings and risk of disease progression. Variability in nuclear shape has been linked to cancer and tumor grading, where higher variability correlates with greater aggressiveness. While mechanisms remain unclear, three key factors contribute to nuclear eccentricity: reduced nuclear envelope proteins (lamin A/B), chromosomal abnormalities, and mechanical forces (e.g., cytoskeletal tension or invasion)[37–42]. These factors likely interact. For example, elongated nuclei are observed in epithelial-mesenchymal transition (EMT), which increases motility and metastasis. In head and neck squamous cell carcinoma, EMT-factor Snail downregulates lamin, enhancing nuclear deformability, elongation, and metastasis propensity[43]. Nuclear deformation during migration through narrow spaces can induce DNA damage, increasing mutational load and tumor aggressiveness[37,38]. Similarly, an AI model linked nuclear area variability to aneuploidy in lung, breast, and colorectal cancer[39]. Nuclear and cellular morphology also influence signaling pathways like MAPK[40]. Multiple studies show that elongated or irregular nuclei correlate with poorer outcomes in epithelial malignancies, including cSCC, underscoring their biological relevance[41,42]. Our findings across multiple cohorts suggest high nuclear eccentricity is an intrinsic tumor trait. Further studies should explore the interplay between biomechanical properties like nuclear shape and biological processes such as EMT in cSCC aggressiveness.

In summary, our study presents an explainable, federated deep learning model that reliably stratifies cSCC patients at high risk of disease progression and identifies their characteristic morphological features. The accuracy, interpretability, and federated implementation of our model hold great promise to better understand the disease and to advance the management of cSCC patients in the future.

Methods

Patient cohorts

For the initial training cohort, all patients with a primary cSCC diagnosed and treated by excision at the Department of Dermatology at the University Hospital Cologne (Cologne cohort) between January 2009 to May 2019 were collected. For these patients we used clinico-pathological parameters based on medical records and pathology reports and performed active follow-up regarding disease progression status. Local recurrence or lymph-node/distant metastasis within 2 years after initial diagnosis was considered a progression event, and was annotated per tumor. Hematoxylin-Eosin (HE) stained slides obtained during routine work-up of surgical samples were available. The final cohort comprised 219 annotated tumors (progress $n = 63$, non-progress $n = 156$) coming from 166 patients.

From the University Hospital Bonn (Bonn cohort) patients diagnosed and treated for cSCC between March 2012 and September 2021 were included. Tumors were excised at the Department of Dermatology or the Department of Oral and Maxillo-facial Surgery and worked up histologically following standard procedures. We identified 23 primary cSCC cases with eventual disease progression (recurrence/metastasis) and randomly selected a group of primary cSCCs without disease progression. Of those, HE slides were available for 35 tumors coming from 35 patients (progress $n = 21$, non-progress $n = 14$).

For the cohort from the Department of Dermatology, Technical University Munich (TU Munich, Munich cohort) we identified patients with a primary cSCC and disease progression and assembled a random cohort of primary cSCCs without disease progression. Of those, HE slides were available for 51 tumors coming from 51 patients (progress $n = 22$, non-progress $n = 29$).

The study was performed in agreement with the Declaration of Helsinki Institutional Review Board of the University Hospital Bonn (vote number 187/16), Ethics committee of the University Hospital Cologne (vote numbers 21–1500, 20–1082 and 22–1330-retro) and institutional review board of the TU Munich (vote number 2024–363-S-CB - 1). Need for informed consent was waived for this retrospective analysis using anonymized data. Clinicopathological parameters of the cohorts are shown in Table 1.

Classification datasets

Whole-slide images (WSIs) were acquired from HE slides using a NanoZoomer Slide Scanner (Hamamatsu) at 40x resolution. In total, we collected 219 WSIs of 166 patients from the University Hospital Cologne, 291 WSIs of 35 patients from the University Hospital Bonn, and 129 WSIs of 51 patients from TU Munich. We filtered out slides without any tumor tissue according to the Segmenter model described by Sancéré et al. The final dataset used for training of the federated deep learning model comprises 214 slides from 157 patients from the University Hospital Cologne, 133 slides from 35 patients from the University Hospital Bonn and 113 slides from 51 patients from TU Munich. From this dataset, 228 slides are from patients showing cSCC progression, and 232 slides are from patients showing no cSCC progression. Data splitting is done in a stratified fashion on patient level, making 65-15-20 splits for training, validation, and testing, respectively.

Pre-processing

Each WSI is tiled into patches of 256×256 pixels at x20 magnification. Patches with less than 50% tissue are discarded, and the remaining patches are processed with an ImageNet pre-trained EfficientNet-v2-L[16], to compute its feature vector representations. The average slide of our dataset produces 11330 feature vectors.

Classification

Each WSI is treated as the sequence of feature vectors corresponding to its non-empty image patches. We use the multiple instance learning classification model described by Pisula and Bozek[44]. Following an approach similar to Lu et al.[45], a transformer model initialized with language-modeling pre-training weights is used for classification. We use a RoBERTa transformer encoder[46], and perform parameter-efficient fine-tuning by only training its normalization layers[45,47]. To reduce compute and memory footprint, we apply multi-head attention pooling at the input to shorten the length of the patch sequence. The embedding vectors from the last layer of the transformer encoder are averaged and fed to a linear layer for the final classification.

Each WSI is classified independently during model training. During inference, in cases where there are multiple slides per patient, we evaluate the model on each one and take the prediction corresponding to the slide with the biggest activation in the positive class output neuron.

We train our model with a Federated Averaging strategy for 50 rounds[12]. Adam is used as the optimizer algorithm, with a learning rate of 1.e-4, weight decay of 5.e-5, and batch size of 4. Model selection is done based on weighted validation AUROC of the three cohorts.

Classification explanation and analysis

Beyond mere disease progression prediction with a deep network classifier, we investigate the biological features that drive our classifier's decision. Our process is threefold: we detect relevant image regions responsible for the model's decision; we compute handcrafted features of the cellular composition of the image regions; and we perform the data analysis itself. This approach is described in detail below.

We use Integrated Gradients (IGs) to identify regions of a WSI that play a role in the classifier's progression prediction[18]. IGs is a deep learning explainability algorithm that attributes the prediction of a deep network to its input features. We apply IGs to our cSCC progression prediction model, to assign a positive score to image patches that contribute to the prediction of the correct class, and a negative score to patches that contribute to the prediction of the opposite outcome. By arranging the IGs attribution scores of the patches in their corresponding spatial locations in the slides, it is possible to visualize these values as heatmaps, as shown in Fig. 4.

We use the HoverNet nuclei segmentation model described by Sancéré et al. on the WSI image patches to identify their cell composition[19,21]. The model detects and classifies cell nuclei into granulocytes, lymphocytes, plasma cells, stroma cells, tumor cells, and non-neoplastic epithelial cells. Once the cells in a patch have been identified, we compute a total of 524 features that summarize the patch into a single feature vector. These features include:

Cell type populations and ratios.

Descriptive statistics (mean, median, variance, skewness, kurtosis, minimum, maximum) of nuclei morphology, such as the mean tumor cells nuclei eccentricity, or the variance in plasma cells nuclei area. These features were computed with the 'skimage.measure' Python package[48].

Descriptive statistics of distances between cell nuclei, such as the median distance between stroma cells and tumor cells.

Average Nearest Neighbor Ratio (ANNR) and Join Count analysis for each cell type.

The features from the last item are used to quantify the spatial arrangement of cells within a patch, and they capture two different aspects of it.

ANNR is used to quantify the observed pattern of distances between cell nuclei in a patch:

$$ANNR = \underline{D_O}/\underline{D_E} \tag{1}$$

where $\underline{D_O}$ is the observed mean distance between each cell and its closest neighbor, and $\underline{D_E}$ is the expected mean distance between each cell and its closest neighbor if the cells were placed randomly:

$$\underline{D_E} = \frac{0.5}{\sqrt{n/A}} \tag{2}$$

where n is the number of cells in a patch, and A is the patch area. An ANNR < 1 indicates clustering (meaning, cells in the patch are closer than a random pattern of cells), and an ANNR > 1 indicates a dispersed or regular pattern of cell nuclei. We compute the ANNR for each cell type in a patch.

Join Count analysis gives a measure of spatial autocorrelation: it describes how the values of a variable at neighboring spatial locations are similar to each other. In our case, the variable of interest is the cell type, where a positive spatial autocorrelation would mean that neighboring cells belong to the same type, and a negative spatial autocorrelation would mean that neighboring cells belong to different classes. Spatial autocorrelation is complementary to ANNR, it quantifies neighboring cell nuclei types disregarding how close or distanced they are.

Our Join Count analysis is computed for each cell type individually, in the following way:

A patch is partitioned into a Voronoi tessellation, using the nuclei centroids as seeds for the regions.

The regions are binary-labeled. Given a cell type, a positive label is assigned to all the cell nuclei belonging to that class, and a negative label is assigned to the remaining regions.

The different types of joins were then counted. Two neighboring cells make a black-black (BB) join if they both are from the positive label (i.e. the cell type being currently analyzed); a black-white (BW) join is formed between two cells of opposite labels; and a white-white (WW) join happens when two cells of the negative label neighbor each other.

This procedure is done for each cell type independently, assigning the positive label (black) to the analyzed cell type and the negative label (white) to all the other cell types. Our measure of spatial autocorrelation is given by:

$$Spatial\ Autocorrelation = (J_{BB} - J_{BW})/J_T \tag{3}$$

where J_{BB}, J_{BW} and J_T are the number of BB joins, the number of BW joins, and the total number of joins, respectively. This equation is positive when the majority of joins in a patch are BB joins, indicating a positive spatial autocorrelation, and is negative when the majority of joins are BW joins, indicating negative spatial autocorrelation.

We apply IGs to all the patients in the test set, and describe their corresponding image patches as previously explained. We use in this analysis the patches coming from tumor regions detected by the Segmenter model described by Sancéré et al.[19,49], plus a surrounding tissue stripe of approximately 800μm of width next to the tumor border. From the totality of patches, we form two groups: A "positive group" of image patches coming from progression patients, which were detected to be explainable of this condition with IGs; and a "negative group" of patches coming from non-progression patients, which were detected to be explainable of this condition with IGs.

To enhance the predictive signal and avoid over-representing patients with bigger tumors, we take a slide's top 10% IGs-scored patches, and limit this quantity to 200 image patches per slide. We compare values of each feature individually between the two groups of patches. We guide our analysis by focusing on features whose values differ between the two groups with an Effect Size bigger than random. We use the Common Language Effect Size (CLES)[50], or probability of superiority, as it has no assumptions about the data distribution, and is straightforward to understand:

$$CLES = P(X > Y) \tag{4}$$

is the probability that a value sampled from group X is bigger than a value sampled from group Y. In our case, the two groups are the positive and the negative groups previously described, and we compute the CLES for each feature with brute force, by exhaustively comparing each value of one group with all the values of the same feature in the other group.

In addition to comparing the feature distributions in both groups, we tested whether the individual patches' feature vectors were sufficient to predict the progression status of their respective patients using an XGBoost classifier[20]. The patches under analysis (coming from the FL model's test set) were split into 80−20 train and test sets, and model selection was done with 3-fold cross-validation on this new train split.

Survival analysis with clinico-pathological variables

Associations of clinico-pathological variables with disease progression and survival were done for all patients with available data. Association with disease progression risk was calculated using logistic regression and reported as odds ratios. Association with survival was done using the Kaplan-Meier method with log-rank test as well as Cox proportional hazard models and reported as hazard ratios with 95% confidence intervals. For multivariable analyses, variables with $p < 0.1$ in univariate analysis were combined. Analyses were done in R statistical environment (v4.3.0).

Data availability

Data is available upon reasonable request to the authors. Code is available at https://github.com/bozeklab/cscc-response.

Received: 13 November 2024; Accepted: 4 June 2025;
Published online: 28 June 2025

References

1. Winge, M. C. G. et al. Advances in cutaneous squamous cell carcinoma. *Nat. Rev. Cancer* **23**, 430–449 (2023).
2. Keim, U. et al. Incidence, mortality and trends of cutaneous squamous cell carcinoma in Germany, the Netherlands, and Scotland. *Eur. J. Cancer Oxf. Engl. 1990* **183**, 60–68 (2023).
3. Brantsch, K. D. et al. Analysis of risk factors determining prognosis of cutaneous squamous-cell carcinoma: a prospective study. *Lancet Oncol.* **9**, 713–720 (2008).
4. Schmults, C. D., Karia, P. S., Carter, J. B., Han, J. & Qureshi, A. A. Factors predictive of recurrence and death from cutaneous squamous

cell carcinoma: a 10-year, single-institution cohort study. *JAMA Dermatol.* **149**, 541–547 (2013).
5. Thompson, A. K., Kelley, B. F., Prokop, L. J., Murad, M. H. & Baum, C. L. Risk factors for cutaneous squamous cell carcinoma recurrence, metastasis, and disease-specific death: a systematic review and meta-analysis. *JAMA Dermatol* **152**, 419–428 (2016).
6. Eigentler, T. K., Dietz, K., Leiter, U., Häfner, H.-M. & Breuninger, H. What causes the death of patients with cutaneous squamous cell carcinoma? a prospective analysis in 1400 patients. *Eur. J. Cancer* **172**, 182–190 (2022).
7. Ruiz, E. S., Karia, P. S., Besaw, R. & Schmults, C. D. Performance of the american joint committee on cancer staging manual, 8th edition vs the brigham and women's hospital tumor classification system for cutaneous squamous cell carcinoma. *JAMA Dermatol.* **155**, 819–825 (2019).
8. Schmults, C. D. et al. NCCN guidelines® insights: squamous cell skin cancer, version 1.2022. *J. Natl. Compr. Cancer Netw. JNCCN* **19**, 1382–1394 (2021).
9. Wysong, A. et al. Validation of a 40-gene expression profile test to predict metastatic risk in localized high-risk cutaneous squamous cell carcinoma. *J. Am. Acad. Dermatol.* **84**, 361–369 (2021).
10. Wang, J. et al. Transcriptomic analysis of cutaneous squamous cell carcinoma reveals a multigene prognostic signature associated with metastasis. *J. Am. Acad. Dermatol.* **89**, 1159–1166 (2023).
11. Haenssle, H. A. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol. J. Eur. Soc. Med Oncol.* **29**, 1836–1842 (2018).
12. McMahan, H. B., Moore E., Ramage, D., Hampson, S. & Arcas, B. A. Y. Communication-efficient learning of deep networks from decentralized data. *arXiv* http://arxiv.org/abs/1602.05629 (2024).
13. Ogier Du Terrail, J. et al. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nat. Med* **29**, 135–146 (2023).
14. Maron, O., Lozano-Pérez, T. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems* (MIT Press, 2024).
15. Vaswani A., et al. Attention is all you need. *arXiv*https://doi.org/10.48550/arXiv.1706.03762.
16. Tan M., Le Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. *ArXiv* https://doi.org/10.48550/arXiv.1905.11946 (2024).
17. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. image Anal.* **81**, 102559 (2022).
18. Sundararajan, M., Taly, A., Yan, Q. Axiomatic attribution for deep networks. *arXiv* http://arxiv.org/abs/1703.01365 (2023).
19. Sancéré, L. Histo-Miner: Tissue features extraction with deep learning from h&e images of squameous cell carcinoma skin cancer. *arXiv* https://doi.org/10.48550/arXiv.2505.04672 (2024).
20. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
21. Graham, S. et al. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).
22. Hörst, F. et al. CellViT: Vision transformers for precise cell segmentation and classification.*arXiv* https://doi.org/10.48550/ARXIV.2306.15350 (2023).
23. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
24. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
25. Shao, Z. et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. *arXiv* https://doi.org/10.48550/arXiv.2106.00908 (2021).
26. Chen R. J. et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. *arXiv*https://doi.org/10.48550/arXiv.2206.02647 (2022)
27. Dietrich, E. et al. Towards explainable end-to-end prostate cancer relapse prediction from h&e images combining self-attention multiple instance learning with a recurrent neural network. In *Proc. Machine Learning for Health* 38–53 (PMLR, 2021).
28. Akram, F. et al. Artificial intelligence-based recurrence prediction outperforms classical histopathological methods in pulmonary adenocarcinoma biopsies. *Lung Cancer* **186**, 107413 (2023).
29. Wu, Z. et al. DeepLRHE: a deep convolutional neural network framework to evaluate the risk of lung cancer recurrence and metastasis from histopathology images. *Front Genet*. **11**, 768 (2020).
30. Xiao, H. et al. Predicting 5 year recurrence risk in colorectal cancer: development and validation of a histology-based deep learning approach. *Br. J. Cancer* **130**, 951–960 (2024).
31. Foersch, S. et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat. Med.* **29**, 430–439 (2023).
32. Shi, Y. et al. Predicting early breast cancer recurrence from histopathological images in the Carolina breast cancer study. *Npj Breast Cancer* **9**, 1–7 (2023).
33. Howard, F. M. et al. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *Npj Breast Cancer* **9**, 1–6 (2023).
34. Yang, J. et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput Struct. Biotechnol. J.* **20**, 333–342 (2022).
35. Lucas, M. et al. Deep learning–based recurrence prediction in patients with non–muscle-invasive bladder cancer. *Eur. Urol. Focus* **8**, 165–172 (2022).
36. Coudray, N. et al. Self-supervised artificial intelligence predicts recurrence, metastasis and disease specific death from primary cutaneous squamous cell carcinoma at diagnosis. *Res. Sq.* **13**, rs.3.rs-3607399 (2023).
37. Shah, P. et al. Nuclear deformation causes DNA damage by increasing replication stress. *Curr. Biol.* **31**, 753–765 (2021).
38. Fan, J.-R. et al. AKT2-mediated nuclear deformation leads to genome instability during epithelial-mesenchymal transition. *Iscience* **26**, 106992 (2023).
39. Abel, J. et al. AI powered quantification of nuclear morphology in cancers enables prediction of genome instability and prognosis. *NPJ Precis. Oncol.* **8**, 134 (2024).
40. Rangamani, P. et al. Decoding information in cell shape. *Cell* **154**, 1356–1369 (2013).
41. Glazer, E. S. et al. Nuclear morphometry identifies a distinct aggressive cellular phenotype in cutaneous squamous cell carcinoma. *Cancer Prev. Res.* **4**, 1770–1777 (2011).
42. Grosser, S. et al. Cell and nucleus shape as an indicator of tissue fluidity in carcinoma. *Phys. Rev. X* **11**, 011033 (2021).
43. Chen, Y.-Q. et al. Snail augments nuclear deformability to promote lymph node metastasis of head and neck squamous cell carcinoma. *Front. Cell Dev. Biol.* **10**, 809738 (2022).
44. Pisula, J. I. & Bozek, K. Efficient WSI classification with sequence reduction and transformers pretrained on text. *Sci. Rep.* **15**, 5612 (2025).
45. Lu, K. Grover, A. Abbeel, P. Mordatch, I. Pretrained transformers as universal computation engines. *arXiv* http://arxiv.org/abs/2103.05247 (2024).
46. Liu, Y. et al. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* https://doi.org/10.48550/arXiv.1907.11692 (2019).

47. Lialin, V., Deshpande, V., Rumshisky, A. Scaling down to scale up: a guide to parameter-efficient fine-tuning. *arXiv* http://arxiv.org/abs/2303.15647 (2024).
48. Walt, S. et al. scikit-image: image processing in python. *PeerJ* **2**, e453 (2014).
49. Strudel R., Garcia R., Laptev I., Schmid C. Segmenter: Transformer for semantic segmentation. *arXiv*https://doi.org/10.48550/arXiv.2105.05633 (2021).
50. McGraw, K. O. & Wong, S. P. A common language effect size statistic. *Psychol. Bull.* **111**, 361–365 (1992).

Acknowledgements

A.F. was partly funded by the Deutsche Krebshilfe through a Mildred Scheel Foundation Grant (grant number 70113307). C.L. was partly funded through the collaborative research center grant on small cell lung cancer (CRC1399, project ID 413326622) by the German Research Foundation (DFG). J.B. receives funding through the collaborative research center grant on small cell lung cancer (CRC1399, project ID 413326622) and predictability in evolution (CRC1310, project ID 325931972) by the German Research Foundation (DFG), a Mildred Scheel Foundation Grant (grant number 70113307) by the Deutsche Krebshilfe and the CANTAR network (NW21-062B) funded through the program "Netzwerke 2021", an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia, Germany. Both K.B. and J.I.P. were hosted by the Center for Molecular Medicine Cologne throughout this research. K.B. and J.I.P. were supported by the BMBF program Junior Group Consortia in Systems Medicine (01ZX1917B) and BMBF program for Female Junior Researchers in Artificial Intelligence (01IS20054).

Author contributions

K.B., J.B. conceived and designed the study. J.I.P., L.S., J.B. developed software and conducted the statistical analyses. D.H., O.P., C.B., A.F., S.B., D.N., K.D., J.L., R.T. provided samples, clinical data, and helped with data curation as well as interpretation of results. J.I.P., D.H., L.S., K.B., J.B. had access to all raw data of the study. J.I.P., D.H., K.B., J.B. drafted the manuscript. All authors read and approved the manuscript and participated in reviewing and editing of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The Authors declare no Competing Non-Financial Interests but the following Competing Financial Interests: D.N. received financial support (speaker's honoraria, advisory boards, travel expense reimbursements or grants) from Abbvie, Almirall, AstraZeneca, Biogen, Boehringer Ingelheim, Bristol-Myers-Squib, GlaxoSmithKline, Incyte, Janssen-Cilag, Kyowa Kirin, LEO Pharma, Lilly, L'Oreal/Cerave, MSD, Novartis, Pfizer, Regeneron and UCB Pharma. J.B. received research funding from Bayer and travel expenses from Merck KG and Bicycle Therapeutics outside the presented work. K.D. received financial support (speaker's honoraria, advisory boards, travel expense reimbursements or grants) from Abbvie, Bristol-Myers-Squib, Novartis, and Pierre-Fabre.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-025-00997-4.

Correspondence and requests for materials should be addressed to Johannes Brägelmann.