nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00635-3

Deep learning supported discovery of biomarkers for clinical prognosis of liver cancer

Received: 6 April 2022

Accepted: 27 February 2023

Published online: 3 April 2023

Check for updates

Junhao Liang[1], Weisheng Zhang[1], Jianghui Yang[2], Meilong Wu[3,6,7], Qionghai Dai[4] ✉, Hongfang Yin[2] ✉, Ying Xiao[2] ✉ & Lingjie Kong[1,5] ✉

Tissue biomarkers are crucial for cancer diagnosis, prognosis assessment and treatment planning. However, there are few known biomarkers that are robust enough to show true analytical and clinical value. Deep learning (DL)-based computational pathology can be used as a strategy to predict survival, but the limited interpretability and generalizability prevent acceptance in clinical practice. Here we present an interpretable human-centric DL-guided framework called PathFinder (Pathological-biomarker-finder) that can help pathologists to discover new tissue biomarkers from well-performing DL models. By combining sparse multi-class tissue spatial distribution information of whole slide images with attribution methods, PathFinder can achieve localization, characterization and verification of potential biomarkers, while guaranteeing state-of-the-art prognostic performance. Using PathFinder, we discovered that spatial distribution of necrosis in liver cancer, a long-neglected factor, has a strong relationship with patient prognosis. We therefore proposed two clinically independent indicators, including necrosis area fraction and tumour necrosis distribution, for practical prognosis, and verified their potential in clinical prognosis according to criteria derived from the Reporting Recommendations for Tumor Marker Prognostic Studies. Our work demonstrates a successful example of introducing DL into clinical practice in a knowledge discovery way, and the approach may be adopted in identifying biomarkers in various cancer types and modalities.

Pathological analysis of whole slide images (WSIs) is the gold standard for cancer diagnosis and prognosis. Tumour classification, staging and prognosis are assessed according to tissue biomarkers on WSIs[1,2]. Unfortunately, even though various tissue biomarkers have been proposed, few of them are robust with high sensitivity and specificity[3,4]. Thus there is still a desperate need for identifying additional robust biomarkers to guide tumour diagnosis and prognosis, and to direct the research of tumour mechanism[5–7]. Specifically in cancer prognosis, with the

[1]State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing, China. [2]Department of Pathology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, China. [3]School of Clinical Medicine, Tsinghua University, Beijing, China. [4]Department of Automation, Tsinghua University, Beijing, China. [5]IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China. [6]Present address: Division of Hepatobiliary and Pancreas Surgery, Department of General Surgery, Shenzhen People's Hospital, The Second Clinical Medical College, Jinan University, Shenzhen, China. [7]Present address: Division of Hepatobiliary and Pancreas Surgery, Department of General Surgery, The First Affiliated Hospital, Southern University of Science and Technology, Shenzhen, China. ✉e-mail: daiqh@tsinghua.edu.cn; yhfa00530@btch.edu.cn; xya02486@btch.edu.cn; konglj@tsinghua.edu.cn

https://doi.org/10.1038/s42256-023-00635-3

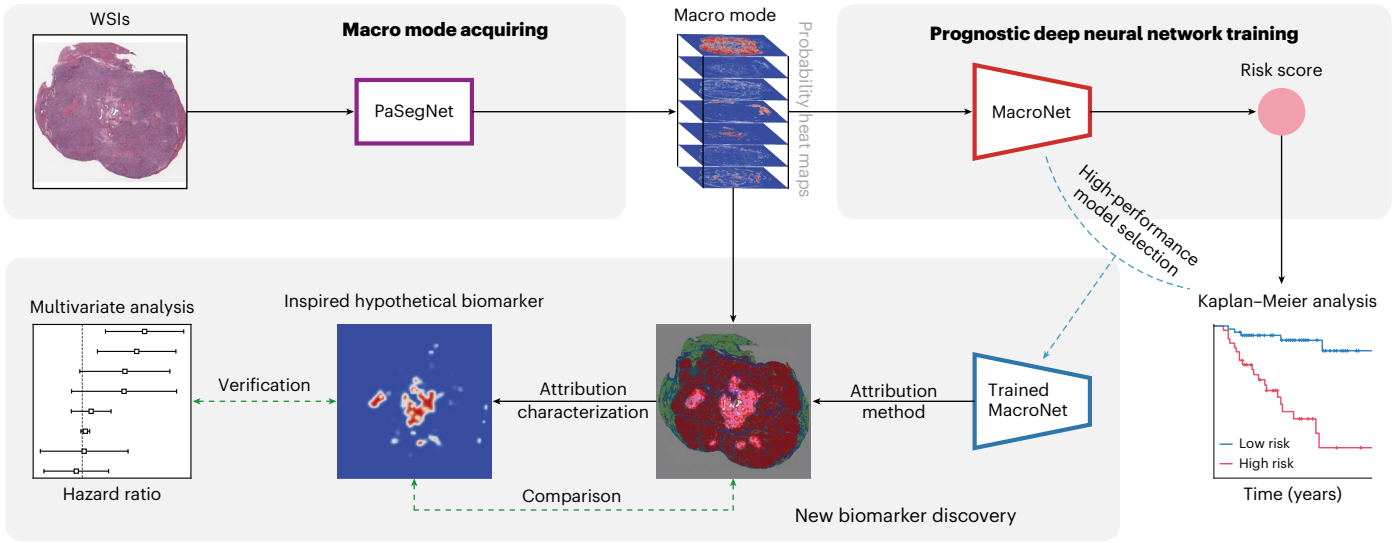


Fig. 1 | The workflow of PathFinder. Digitized high-resolution histology slides of patients serve as the input into the framework. The WSI is first processed with PaSegNet, a convolutional neural network (CNN), to obtain the spatial distribution probability heat maps of seven common liver tissues. The achieved macro mode and the corresponding survival time are used as the image–label pair to train the MacroNet, a prognostic CNN with the output of corresponding risk score for guiding the patient's prognosis. Then one can apply the attribution method to the trained, well-performing MacroNet to explore the model's spatial focus area, from which to get the inspiration of potential prognostic biomarkers. Following that, these hypothetical biomarkers are modelled on the basis of the macro mode to achieve quantification and characterization, in which the ones similar to the attribution map after visualization are selected as candidate biomarkers and used as indicators for multivariate analysis. After testing with clinical dataset, the significantly independent prognostic indicators can be identified.

advancement of computational pathology in recent years, deep learning (DL) models based on end-to-end training can predict a risk score that outperforms current clinical staging, showing the potential of learning knowledge from current medical data[8–12]. However, due to limited interpretability and generalizability, DL-based risk score is still difficult to be accepted as a useful biomarker for clinical prognosis[6,13,14].

Considering that clinicians are likely to keep playing the central role in patient care, it is essential to focus the development and evaluation of artificial intelligence (AI)-based clinical algorithms on their potential to augment rather than replace human intelligence[15–17]. Although some studies have attempted to use established biomarkers and attribution methods to verify the credibility of abstract risk scores[8–11], this strategy fails in generating new knowledge for clinical prognosis. Knowledge discovery based on AI, especially the discovery of new or dominant prognostic biomarkers of clear pathological significance and explicit mathematical model, will open up new direction of human-centric AI for cancer prognosis.

Different from that in the fields of genetics where biologically informed sparse DL models combined with attribution methods has been used to guide pre-clinical discovery[18], identifying tissue biomarkers from well-performing prognostic DL models is challenging[8–12]. On the one hand, the input multi-dimensional images of WSIs for prognosis are of high information density compared with molecular data inputs, which are usually one-dimensional vector and have specific labels or descriptions[18]. Thus it is difficult to build a sparse network while guaranteeing the prognostic performance[19]. On the other hand, current attribution methods usually achieve a two-dimensional attribution map for spatial attribution positioning[13,14], which is far from locating specific high-attribution features in high-information-density input. These two problems lead to insufficient interpretation, as low-dimensional attribution knowledge is used to interpret abstract results based on high-dimensional inputs. Even worse, it makes one use pre-existing knowledge in explanation, which contradicts the aim of discovering new biomarkers[19–21].

Histologically, gigapixel WSIs can be regarded as self-multimodal information sources with both slide-level macro mode and region-level micro mode[14]. The former contains multi-class tissue spatial distribution and interaction information, while the latter contains cell texture and structure information (Methods and Extended Data Fig. 1). However, limited by graphics processing units memory and deep neural network architecture, WSIs are generally cut into patches and only the micro mode information is paid attention to in most DL-based studies[9,10,22–24]. Moreover, in clinics, due to the lack of precise quantification of WSIs, the relationship between tissue spatial distribution and patients' prognostic result is still not clear.

In this Article, we propose an interpretable, human-centric, DL framework, named PathFinder, that uses the sparse multi-class tissue spatial distribution information of WSIs for assessing prognosis and discovering new biomarkers. Using the macro mode of WSIs, which is of low information density that perfectly matches current spatial-positioning attribution methods, PathFinder can achieve state-of-the-art prognostic performance. Inspired by the exact and intuitive attribution maps of PathFinder, we found spatial distribution of necrosis in liver, a common but overlooked pathological morphology, has a strong relationship with patients' prognosis, on the basis of which we characterized two significant indicators for clinical prognosis.

Interpretable AI-based framework for biomarker discovery

Figure 1 shows the workflow of PathFinder. It consists of three parts: macro mode acquiring, prognostic deep neural network training and new biomarker discovery. We first trained the multi-class tissue segmentation network PaSegNet to obtain the multi-class tissue probability heat maps as the macro mode of WSIs (Methods). To acquire high-quality macro mode, we proposed meta-annotation, a data-centric annotation method that combined with pathological priors to bridge the gap between current pathological annotation methods and DL training requirements, and achieved efficient, high diversity and low similarity class-balanced training dataset (Methods and Extended Data Fig. 2). With the macro mode of WSIs, we built MacroNet for high-precision prognosis, which is composed of a convolution feature

extractor and a multilayer perceptron with a batch normalize layer[25] (Methods). Using only time-to-event patient death information as the input mode label and Cox proportional likelihood loss as the network loss, the MacroNet can learn to predict the patients' risk score on the basis of macro mode only. Then we used attribution methods on the trained MacroNet to acquire the attribution map of input image[26], and overlapped the attribution map on the corresponding multi-class segmentation map. The generated two-dimensional attribution map shows the spatial areas that MacroNet focuses on, which matches well with the sparse multi-class tissue spatial distribution information, making the interpretation more direct and objective. On the basis of integrative analysis of macro mode and attribution map, pathologists can propose the hypothesis of the biomarkers that the model is concerned with, followed by quantitatively characterization. The new biomarkers, whose visualizations are similar with the corresponding attribution map, were used as indicators to perform multivariate analysis according to criteria derived from the Reporting Recommendations for Tumor Marker Prognostic Studies[27]. After testing with clinical dataset, new biomarkers of significantly independent prognostic effect were discovered.

With PathFinder, we performed the discovery of new tissue biomarkers for clinical prognosis of hepatocellular carcinoma (HCC), which is the fourth leading cause of cancer-related death worldwide[28]. In this study, we collected 342 WSIs from 330 patient samples in The Cancer Genome Atlas Liver Hepatocellular Carcinoma dataset (TCGA dataset) and 1,182 WSIs from 83 patient samples in Beijing Tsinghua Changgung Hospital dataset (QHCG dataset) (Extended Data Fig. 3 and Supplementary Fig. 1). As for the case that there are multiple WSIs for a patient, we selected the one with the largest tumour fraction as the patient's representative WSI, as discussed later. We trained MacroNet in a ten-fold cross-validation on TCGA dataset, and tested the generalization of the trained model on the QHCG dataset. To better compare the prognostic performance of MacroNet, we also designed and trained MicroNet and M2MNet for prognosis task. The former one is based on micro mode, which takes high-resolution tumour patches as inputs, and the latter one is based on both macro mode and micro mode, which attempts to fuse these two modes (Methods and Extended Data Fig. 4).

Evaluation of model performance

We first evaluated the multi-class classification performance of PaSegNet on the internal test set of QHCG dataset and external independent test sets including TCGA dataset and Pathology AI Platform 2019 challenge dataset (PAIP dataset). Confusion matrices and receiver operating characteristic (ROC) curves are used to demonstrate classification results (Fig. 2a and Supplementary Figs. 2 and 3). The macro-average accuracy and area under the curve (AUC) are selected to evaluate model performance. Across all test sets, PaSegNet achieved accuracy of 0.948, 0.956 and 0.941, and AUC of 0.9980, 0.9984 and 0.9974, on QHCG, TCGA and PAIP test set, respectively. The results show that the PaSegNet trained on the meta-annotated dataset can achieve accurate multi-class tissue classification. To evaluate the segmentation performance of WSIs, we further visualized the multi-class tissue probability heat maps and segmentation maps obtained by PaSegNet, both of which demonstrate that the model can accurately and smoothly segment WSIs and identify small key lesion areas (Extended Data Fig. 5). In general, PaSegNet trained on the meta-annotation dataset can efficiently quantify WSIs' macro mode and ensure the following prognostic network training.

We next evaluated the prognostic capability of MacroNet, MicroNet and M2MNet, by using ten-fold cross-validation on TCGA dataset. To compare the performance of prognostic networks, we used the median of cross-validated concordance index (C-Index) to measure the predictive accuracy of each model, Kaplan–Meier curves to visualize the quality of patient stratification between predicted high-risk and low-risk patients, and the log-rank test to test the statistical difference between high-risk and low-risk groups (Supplementary Note 1). MacroNet achieved a C-Index of 0.708, similar to the C-Index 0.717 using MicroNet and lower than the C-Index 0.787 using M2MNet (Fig. 2b). In visualizing the Kaplan–Meier survival curves of predicted high-risk and low-risk patient groups, MacroNet also showed good discrimination between the two risk groups ($P = 1.25 \times 10^{-7}$) compared with M2MNet and clinical staging (Fig. 2d,e and Extended Data Fig. 6a). In addition, we also reported dynamic area under the curve (AUC; termed as Survival AUC) to measure the prognostic performance of the networks. Similar conclusion can be achieved as MacroNet achieved the Survival AUC of 0.732, similar to the Survival AUC 0.729 using MicroNet and lower than the Survival AUC 0.832 using M2MNet (Supplementary Fig. 4a).

We further evaluated the models' generalization capability by training the models on TCGA dataset and testing them on QHCG dataset. MacroNet achieved a C-Index of 0.754, whereas M2MNet and MicroNet achieved C-Indices of 0.695 and 0.652, respectively (Fig. 2c). On Survival AUC, we observed similar model performances, with MacroNet reaching an AUC of 0.796 compared with 0.733 in M2MNet and 0.666 in MicroNet (Supplementary Fig. 4b). These results demonstrated that MacroNet has stronger generalization ability in prognosis. In addition, the Kaplan–Meier survival curves of MacroNet showed good discrimination between two risk groups ($P = 7.68 \times 10^{-7}$) on QHCG dataset, as M2MNet did (Fig. 2g and Extended Data Fig. 6b). Furthermore, the multivariable analysis revealed that the risk score predicted by MacroNet (hazard ratio (HR) 2.21, 95% confidence interval (CI) 1.26 to 3.86, $P = 0.0057$, TCGA dataset; HR 6.56, 95% CI 2.01 to 21.36, $P = 0.0018$, QHCG dataset) was independent of other clinicopathological characteristics (Fig. 2f,h and Supplementary Tables 1 and 2), and the risk scores generated by MicroNet and M2MNet were also independent of other clinicopathological characteristics (Supplementary Tables 3–6). These results indicate that MacroNet can achieve state-of-the-art prognostic performance using only macro mode of WSIs and has potential in finding useful prognostic biomarkers.

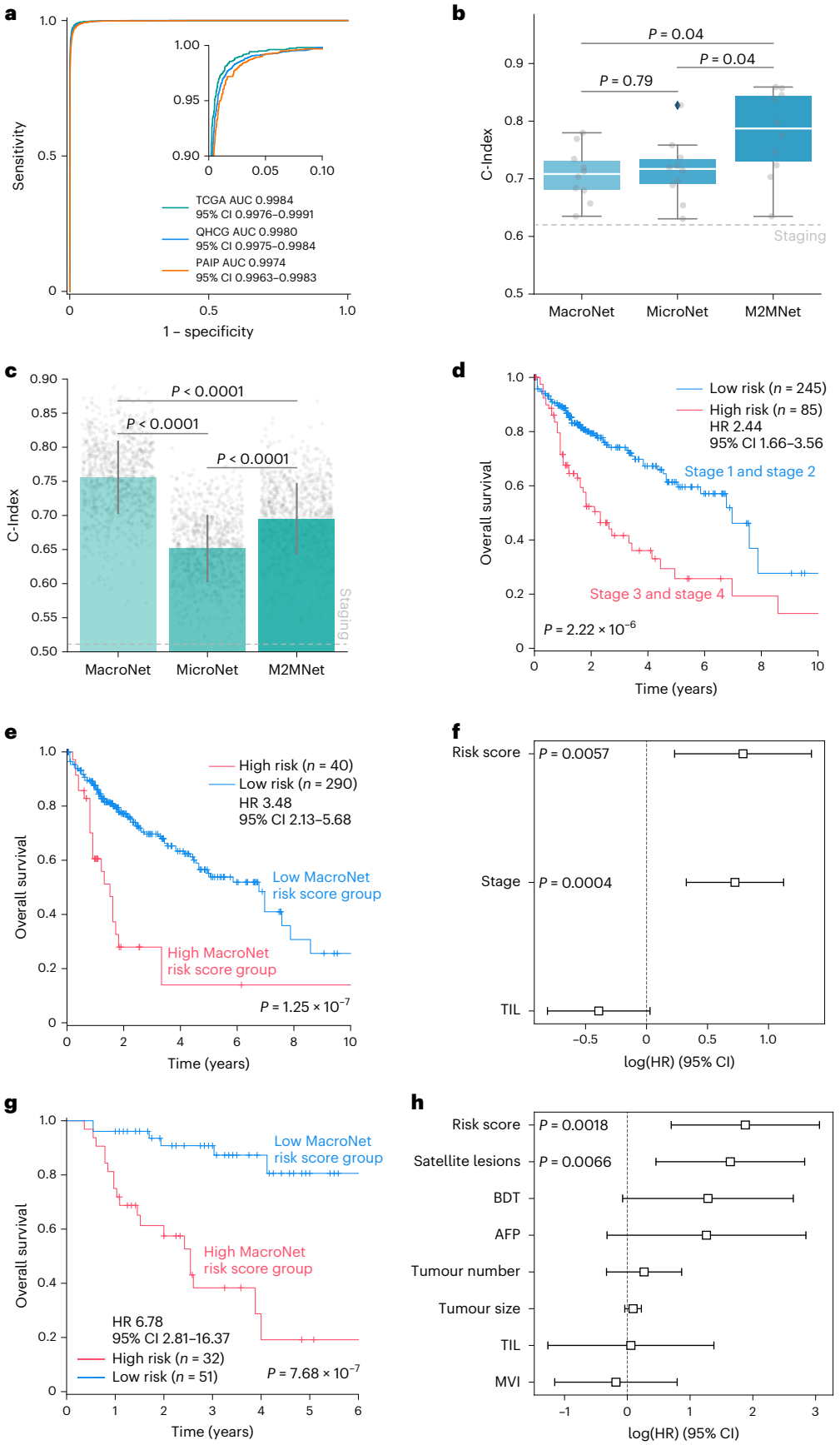
Discovery, characterization and verification of biomarkers

To interpret why MacroNet can achieve high-performance prognosis and to explore which macro features largely contribute to risk score, we

Fig. 2 | Performance of PathFinder in the discovery of new tissue biomarkers for clinical prognosis of HCC. a, ROC curves for the multi-class tissue classification, evaluated on the internal test set (QHCG) and external independent test sets (TCGA, PAIP). The central measure of the CIs is the median. **b**, C-Index distribution of MacroNet, MicroNet and M2MNet on TCGA dataset in a ten-fold cross-validation ($n = 10$ independent experiments for MacroNet, MicroNet and M2Mnet, respectively). Box plot whiskers extend to the smallest and largest value within 1.5 times the interquartile ranges of hinges, and box centre and hinges indicate median and first and third quartiles, respectively. **c**, C-Index performance of MacroNet, MicroNet and M2MNet on QHCG test set ($n = 83$ patients). The data are presented as mean values, and the error bars show the 95% CI of the mean estimate (1,000 bootstrapping samples). **d**, Kaplan–Meier analysis of patient stratification of clinical staging patients on TCGA dataset. **e**,**g**, Kaplan–Meier analysis of patient stratification of low- and high-risk patients via MacroNet on TCGA dataset (**e**) and QHCG dataset (**g**), respectively. **f**,**h**, Multivariable analysis of factors associated with overall survival and MacroNet risk score on TCGA dataset ($n = 330$ patients) (**f**) and QHCG dataset ($n = 83$ patients) (**h**), respectively; the data are presented as HR estimates (squares), and the error bars show the 95% CI of the HR estimate, according to multivariable Cox proportional hazards model; the results of univariate and multivariate analyses are described in details in Supplementary Tables 1 and 2. *P* values according to two-sided Mann–Whitney–Wilcoxon test (**b**), two-sided two-sample *t*-test (**c**), two-sided log-rank test (**d**, **e**, and **g**) and multivariable Cox proportional hazards model (**f** and **h**). *n*, sample size; Stage, AJCC staging; TIL, tumour infiltrating lymphocytes digital score; BDT, bile duct thrombosis; AFP, alpha-fetoprotein; MVI, microvascular invasion.

https://doi.org/10.1038/s42256-023-00635-3

conducted an integrated analysis from both global and individual perspectives. We counted the difference in the tissue fractions in patients of high-risk scores and low-risk scores from a global perspective, and found that the necrosis fraction is significantly higher in the high-risk score group (Extended Data Fig. 7a,c). Then we analysed the segmentation map of high-risk and low-risk WSIs, and observed that necrosis



https://doi.org/10.1038/s42256-023-00635-3

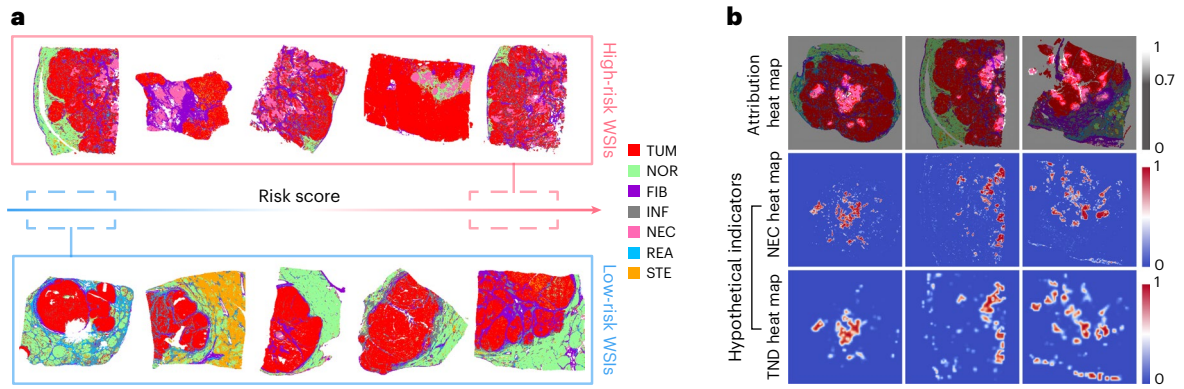


Fig. 3 | Discovery and characterization of new tissue biomarkers. a, Segmentation maps of low- and high-risk WSIs predicted by MacroNet on TCGA dataset and QHCG dataset. **b**, Attribution heat maps of WSI segmentation maps and their corresponding visualization results of NEC and TND hypothetical indicators.

occurred in every high-risk WSI, but not in all low-risk WSIs (Fig. 3a). From an individual perspective, we used the attribute method to locate the areas where MacroNet focused on in the form of a two-dimensional heat map, and overlapped the result with the segmentation map for better visualization (Fig. 1). We discovered that the areas of high contribution are almost the junctions of necrosis and other tissues (Fig. 3b), which is consistent with our former conclusions obtained from the global perspective. All the discoveries inspired us that spatial distribution of necrosis may have a strong relationship with HCC prognosis.

To make the DL-based MacroNet acceptable in clinical practice, we proposed two hypotheses of new biomarkers, namely necrosis area fraction in WSIs (NEC) and tumour necrosis distribution (TND), based on above integrated analyses and inspirations of MacroNet. We first established mathematical models of these two indicators to characterize them, and achieved their quantification based on the existing macro mode (Methods). By visualizing these two indicators and comparing them with the corresponding attribution map, we found that these two hypothetical indicators can well characterize the features that MacroNet pays attention to (Fig. 3b and Extended Data Fig. 8), indicating that these two clinically available indicators are of great potential to affect the prognosis of the risk score given by MacroNet. It also should be noted that these biomarkers are objective and universal pathological features, considering that NEC is a common and inherent attribute of WSIs, and TND is a newly designed indicator that takes into account the spatial distribution and interaction between tumour and necrosis.

To verify whether NEC and TND are independent prognostic indicators, we investigated the prognostic significance of these two indicators on both TCGA and QHCG datasets using Kaplan–Meier curves and Cox hazard analysis by conducting univariate and multivariate analyses of clinicopathological parameters. Additionally, to compare the performance with new clinical indicators inspired by AI, we quantified tumour-infiltrating lymphocytes (TILs), which is already known as a prognostic factor and is significantly different between high-risk group and low-risk group (Extended Data Fig. 7b,d and Methods)[12,29], as an indicator designed on the basis of known clinical experience. The Kaplan–Meier curves and *P* values based on log-rank test showed that NEC and TND can significantly distinguish high-risk and low-risk groups on both TCGA and QHCG datasets (Fig. 4a,c,e,g). The univariate and multivariable analyses revealed that the dependences of overall survival on NEC (HR 4.66, 95% CI 1.77 to 12.28, $P = 0.0019$, QHCG dataset; HR 1.80, 95% CI 1.13 to 2.87, $P = 0.0133$, TCGA dataset) and TND (HR 6.67, 95% CI 2.36 to 18.85, $P = 0.0003$, QHCG dataset; HR 3.00, 95% CI 1.56 to 5.74, $P = 0.0009$, TCGA dataset) were more significant than most clinical indicators including TILs (Fig. 4b,d,f,h). This suggests that the two indicators are independent of other clinicopathological characteristics. In addition, NEC (HR 3.31, 95% CI 1.73 to 6.30, $P = 0.0003$) and TND (HR 2.92, 95% CI 1.52 to 5.60, $P = 0.0012$) can even be used as significant indicators in recurrence prediction (Extended Data Figs. 6c–i and Supplementary Table 9). It is worth noting that the Cox's proportional hazard model was able to achieve a C-Index 0.7 without utilizing additionally clinical variables or risk score predicted by DL methods, as it makes predictions based on only NEC (C-Index 0.703) or TND (C-Index 0.691) (Fig. 5d,e). In addition, taking other clinical factors together into consideration, the C-Indices of NEC and TND can be further improved to 0.831 and 0.845, indicating the value of these two indicators in clinical prognosis (Supplementary Fig. 5).

Overall, the above results verified spatial distribution of necrosis as a new biomarker for prognosis. We demonstrated that the prognostic performance of the AI inspired indicators based on WSIs macro mode is comparable to the performances of various DL models based on WSIs micro mode, genomics and multimodality[9–12].

Robustness of macro mode indicators

In clinical practice, there are generally many WSIs with different sampling positions from a patient (Fig. 5a). As the micro mode is not greatly affected by the sampling locations, the prognostic DL models trained on the micro mode rarely discuss the situation where a patient has multiple WSIs[8]. However, different sampling positions will cause huge differences in the macro mode, which will lead to deviations in the risk scores predicted by MacroNet (Fig. 5b and Extended Data Figs. 7e,f). Exploring how to select representative WSI from multiple WSIs of a patient becomes an unavoidable problem in applying macro indicators in clinical prognosis.

In our former study, we selected the largest tumour fraction one as the patient's representative WSI. To explore the robustness and effectiveness of this selection rule in clinical prognosis, we calculated the risk score, TND and NEC of all WSIs, and randomly selected one from the multiple WSIs of a patient as the representative WSI, with C-Index being used to measure the accuracy of prognosis under this random sampling standard. After 10,000 simulations under random selection strategy, the prognostic performance of our former selection rule is better than most random selections (Fig. 5c,d,e). Even for NEC and TND, two objective and universal biomarkers, the results based on largest tumour fraction selection rule were better than 94% of the results based on random selection rule, indicating that the largest tumour fraction selection rule can be adopted with NEC and TND biomarkers for clinical prognosis.

Besides, it is important to verify the prognostic robustness of these two indicators calculated from segmentation maps with different accuracies. We first calculated the TND and NEC scores corresponding to the segmentation maps generated by 11 commonly used convolutional neural networks (CNNs) (Extended Data Figs. 9 and 10,

https://doi.org/10.1038/s42256-023-00635-3

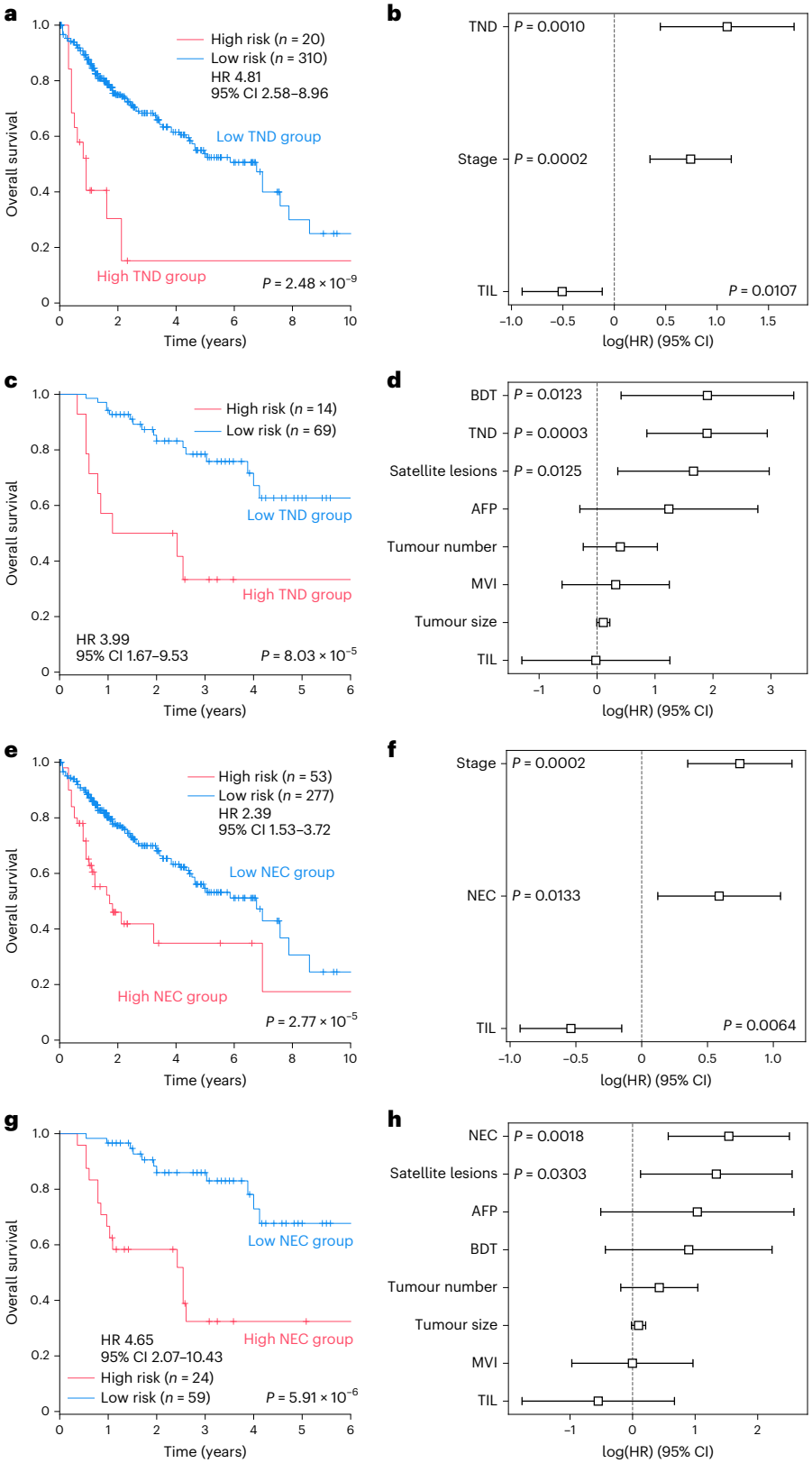


Fig. 4 | Verification of new tissue biomarkers. a,**c**, Kaplan–Meier analysis of patient stratification of low (low TND score) and high-risk (high TND score) patients on TCGA dataset (**a**) and QHCG dataset (**c**). **b**,**d**, Multivariable analyses of TND and other factors associated with overall survival on TCGA dataset (**b**) ($n = 330$ patients) and QHCG dataset (**d**) ($n = 83$ patients). **e**,**g**, Kaplan–Meier analysis of patient stratification of low (low NEC score) and high-risk (high NEC score) patients on TCGA dataset (**e**) and QHCG dataset (**g**). **f**,**h**, Multivariable analyses of NEC and other factors associated with overall survival on TCGA dataset (**f**) ($n = 330$ patients) and QHCG dataset (**h**) ($n = 83$ patients). In **b**, **d**, **f** and **h**, the data are presented as HR estimates (squares) and the error bars show the 95% CI of the HR estimate, according to multivariable Cox proportional hazards model; details are presented in Supplementary Tables 7 and 8. *P* values according to two-sided log-rank test (**a**, **c**, **e** and **g**) and multivariable Cox proportional hazards model (**b**, **d**, **f** and **h**). *n*, sample size; Stage, AJCC staging; TIL, tumour infiltrating lymphocytes digital score; BDT, bile duct thrombosis; AFP, alpha-fetoprotein; MVI, microvascular invasion.

https://doi.org/10.1038/s42256-023-00635-3

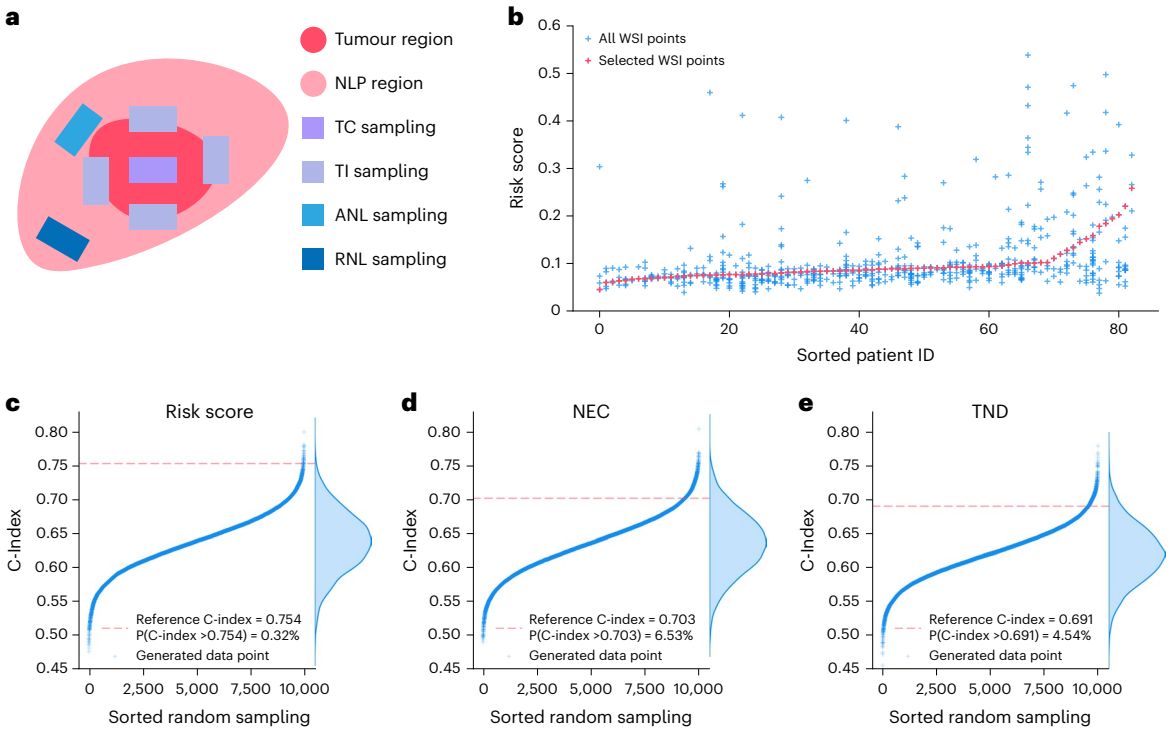


Fig. 5 | Exploring the robustness of macro mode indicators. a, Sampling strategy of clinical WSIs. NLP, non-neoplastic liver parenchyma; TC, tumour centre; TI, tumour–liver interface; ANL, adjacent non-neoplastic liver; RNL, remote non-neoplastic liver. **b**, Deviations in the risk scores predicted by MacroNet from different WSIs of a patient. The risk scores of all WSIs (excluded WSIs without tumour) of 83 patients are ranked in ascending order based on the selected WSI points. Each patient has more than one WSIs points (blue points on a specific abscissa), in which the selected WSIs to characterize the patient's final risk score is labelled as red points. **c**–**e**, Random selection strategy simulations of MacroNet risk score (**c**), NEC (**d**), and TND (**e**), respectively. The red dotted lines represent C-Indices of MacroNet risk score, NEC, and TND under the largest tumour fraction selection rule. Each blue point represents the C-Index of one random selection simulation, and all the blue points are ranked in ascending order based on their C-Indices. The distribution of these points with respect to the C-Index is shown on the right side of the image.

Supplementary Note 2 and Supplementary Figs. 6 and 7). Then we measured the corresponding prognostic performance (that is, C-Index) of NEC and TND scores. No major difference was found in the TND and NEC scores calculated from segmentation results generated by different CNNs of the same patient, and the overall trend of score ranking remains relatively consistent across all patients (Extended Data Figs. 9a and 10a). More specifically, except for AlexNet that has poor classification performance, the C-Indices of TND (Extended Data Figs. 9b,c) and NEC (Extended Data Figs. 10b,c) obtained from segmentation maps generated by other CNNs are close. These indicate the robustness of these two indicators on prognosis, which further illustrates their generalization ability and usability in clinical practice.

Discussion

We present PathFinder as a complete framework of AI-inspired discovery of clinically acceptable biomarkers. Instead of using DL to predict a risk score from WSIs[8–11,24], we focus on proposing human-centric workflows for inspiring pathologists to discover new clinically acceptable biomarkers from well-performing black boxes. We show a method of bridging AI and clinical prognosis, and prove the potential of AI in learning and exploring new prognostic biomarkers based on large datasets and objective survival information.

To overcome the limited interpretability and generalizability of DL-based risk scores, we proposed to simplify the input of DL models and explored the relationship between multi-class tissue spatial distribution and prognosis. Different from utilizing pre-trained networks to compress WSIs[8,10,24,30], our input is more sparse and has explicit medical meaning, which enables the attribution method to characterize the biomarkers that the model focuses on more accurately. Our results show that the prognostic performance of DL is still good even when the input is reduced from WSIs of several gigabytes to macro mode of several megabytes. This indicates that the multi-class tissue spatial distribution of WSIs has prognostic information and the conventional inputs of prognostic DL models are redundant.

In this study, we did not target AI as a substitute for pathologists, but as a tool for pathologists to mine dominate biomarkers. Just as AI guides mathematical intuition[31], pathologists can formulate specific hypotheses based on their clinical experience, and then use PathFinder to deeply mine the connection between hypotheses-relevant information and prognosis. Inspired by PathFinder, we defined two necrosis-related clinical prognostic indicators, NEC and TND, and demonstrated their feasibility in HCC prognosis. Even as a common pathological morphology in liver cancer, spatial distribution of necrosis has caught little attention and has not been put into clinical staging guidelines in detail[32–34]. Our findings demonstrate that AI can analyse data more objectively and alert us about missing information. Different from highly diverse tumour tissues, necrosis is easier to be distinguished in both clinics and computer vision, which makes it convenient for clinical prognosis. Meanwhile, the mechanisms between tumour and necrosis are still unclear. The significant effect of TND and NEC on prognosis may suggest that the spatial distribution of tissue is worth considering in researches of necrosis mechanisms. Additionally, tumour necrosis is postulated to be caused by tumour necrosis factors[35], which have been found significant correlations with TILs[36,37]. However, our results suggest a low correlation between tumour necrosis and TILs (Extended Data Figs. 6j,k), indicating that HCC necrosis may have its own specific causes and mechanisms.

As products of knowledge discovery, TND and NEC have clear pathological significance and explicit mathematical model. The strong generalizability of these new biomarkers is evaluated on TCGA and QHCG datasets, suggesting the great advantages of human-centric AI for knowledge discovery and clinical prognosis.

Same as all commonly used DL models, the focusing features of PathFinder would be affected by training data and hyperparameters. In addition, the intra-individual variability of the macro mode cannot be ignored. However, we explored the robustness of macro mode and gave a feasible selection rule for macro mode variability problem.

In PathFinder, the macro mode can achieve state-of-the-art prognostic performance as micro mode does. Considering that numerous studies have achieved multi-class tissue segmentation across various cancer types[38,39], further exploration of the impact of these ready-made segmentation maps on prognosis may lead to new discoveries. Moreover, benefitting from its simple and easy-to-use features, PathFinder can be easily migrated to similar tasks such as spatial multi-omics and three-dimensional pathological prognosis to discover new biomarkers in different modalities[40–42]. We expect PathFinder as a fundamental mechanism to better integrate the two fields of clinical prognosis and AI, and inspire more meaningful discoveries.

Methods

Meta-annotation

The acquisition of annotated data is a major challenge for deep-learning-based computational pathology. Recently, annotation-free methods such as multi-instance learning or self-supervised learning have achieved good performance on both WSI segmentation, diagnosis and prognosis[22,43]. However, these annotation-free approaches usually require a large amount of data and computing power to make up the cost for the lack of existing pathology priors during training. Improving annotation method and/or dataset quality without changing existing supervised learning method may be another means to solve the dilemma[44]. Here we analysed the gap between pathological annotation and DL, and proposed the meta-annotation based on existing pathological priors and training requirements of DL models to achieve efficient and high-quality pathological image annotation and dataset generation.

The gap between pathological annotation and DL. Conventional WSIs pathological annotation methods usually annotate the contour of specific tissues, for example, tumour boundaries (Extended Data Fig. 2b). However, annotating WSIs is time consuming and laborious due to the complex boundaries and large scale. Furthermore, the tissue boundaries always contain other tissues which are difficult to exclude by annotating (Extended Data Fig. 2d), which would introduce noise label data into the DL training set (Extended Data Fig. 2a). Some of the WSIs regions are completely mixed by multiple tissue types that cannot be annotated precisely (Extended Data Fig. 2e). Moreover, tissue area fractions of different classes in WSIs are quite different, for example, bile duct reaction tissue may occupy 0.01% of the WSI tissue area while tumour tissue occupies 60%. In addition, a tissue type with a large area in one WSI is always similar in content, which is redundant (Extended Data Fig. 2f). Such unbalanced data bring difficulties to DL training (Extended Data Fig. 2a).

However, when it comes to DL, the desired training set is class balanced, has high diversity and has low similarity. Even a small dataset can achieve a high performance if it has such features (Extended Data Fig. 2a). Most segmentation tasks first classify patches and then stitch them together according to their spatial distribution, to acquire the segmentation map of WSI. However, it is difficult to annotate the junction of tissues and give a specific label to the segmented patches from tissue boundary. Meanwhile, according to the pathological priors, most specific tissues on a WSI are actually similar (Extended Data Fig. 2f), and using all specific tissues as the training set will cause serious problems of data imbalance. Therefore, for the segmentation methods based on patches classification, it is not advisable to perform the complete annotation of outer contours to improve the training performance. Designing new annotation methods based on the requirements of DL and the properties of WSI may enable efficient data annotation and good-performance segmentation.

Purpose of meta-annotation. We proposed the meta-annotation to close the gap between conventional WSI pathological annotation and DL training requirements. Meta-annotation method aims to ensure the diversity of annotated tissues while reducing redundant annotation between similar tissues based on WSIs prior and pathologists' experience. The basic purpose of pathological annotation is to label different classes of tissues, where the classes can be different types of tissues, such as fibrosis and tumour, or different subtypes, such as early-stage tumour and late-stage tumour. In our experiment, we pay attention to seven different types of tissue and empty area (TUM, tumour; Nor, normal; FIB, fibrosis; INF, inflammation; NEC, necrosis; REA, bile duct reaction; STE, steatosis; EMP, empty), and different subtypes of the same tissue (for example, early-stage tumour versus late-stage tumour) are considered as intra-specific diversity[9,45,46]. The selected seven tissue types are common, which basically cover histological features that are easily identified at the resolution level of current WSIs. On the basis of such classification, we can study macro spatial distributions of multi-class tissue.

Details of meta-annotation. The process of meta-annotation and the acquisition of PaSegNet dataset for segmentation is shown in Extended Data Fig. 2g. For the WSI that needs to be annotated, pathologists use rectangular boxes to annotate typical areas to reduce the difficulty of labelling. For example, for large tumour or normal regions, pathologists only annotate a small region of inside areas, and perform sampling in multiple spatial regions to ensure high diversity and low similarity of the data. For tissue types that occupy only small areas, such as inflammation and bile duct reactions, pathologists use rectangular boxes to enclose their regions as much as possible. After annotating WSIs, non-overlap 150 × 150 pixels patches are extracted automatically on the basis of the annotated rectangular boxes. Although the impact of class imbalance has been minimized in the annotating process, TUM and NOR patches are still much more frequent than REA and INF patches. To overcome this problem, during automatic extraction, we specify that TUM and NOR classes undergo random extraction of up to 100 patches based on rectangular annotations in one WSI, and all annotated regions of other classes are extracted in full patches. After patch extraction, resampling is applied to the extracted dataset to achieve better class balance, which leads to the final meta-annotation training set.

WSI decoupling and sparsification

To overcome the problem of the high information density of WSIs and make prognostic DL model more suitable for current attribution methods, we decoupled the input WSI into macro mode and micro mode. In our study, we selected the multi-class tissue probability heat maps as the macro mode and the morphology of tissue patches as the micro mode of WSIs (Extended Data Fig. 1). We first used OTSU method to remove background[47], divided the non-background area into 150 × 150 RGB image patches at 20× magnification, and recorded the locations of all patches. Then we proposed PaSegNet f_{seg}, a ResNeXt50-based multi-class classification CNN[25] pre-trained on ImageNet[48], to encode the input patch $\mathbf{I}(i,j) \in \mathbb{R}^{150\times150\times3}$ into probability vector $\boldsymbol{p}(i,j) \in \mathbb{R}^{8}$, where (i,j) is the location of patch $\mathbf{I}$, $\boldsymbol{p}_t$ is the probability of $\mathbf{I}$ belonging to class t in eight tissue classes. Specifically, we used the convolution layers f_{covn} of ResNeXt50 to convert $\mathbf{I}$ into 2,048-dimensional feature vector, and modified the last output feature of fully connected layers (f_{fc}) $\boldsymbol{g}$'s dimension to 8:

$$\boldsymbol{p}(i,j) = \mathrm{softmax}\left(f_{\mathrm{fc}}\left(f_{\mathrm{covn}}\left(\mathbf{I}(i,j)\right)\right)\right) = \mathrm{softmax}\left(\boldsymbol{g}\right) = f_{\mathrm{seg}}\left(\mathbf{I}(i,j)\right) \quad (1)$$

$$\boldsymbol{p}_t(i,j) = \frac{\exp(\boldsymbol{g}_t)}{\sum_{j=1}^{8}\exp(\boldsymbol{g}_j)} \quad (2)$$

After training, the PaSegNet can map the input WSI $\mathbf{W} \in \mathbb{R}^{m\times n\times 3}$ to macro mode $\mathbf{M} \in \mathbb{R}^{m'\times n'\times 8}$, $m' = \mathrm{int}(m/150)$, $n' = \mathrm{int}(n/150)$:

$$\mathbf{M} = f_{\mathrm{seg}}\left(\mathbf{W}\right) \quad (3)$$

where $\mathbf{M}_{ij} = \boldsymbol{p}(i,j)$, $\mathbf{W}_{ij} = \mathbf{I}(i,j)$, $\mathbf{M}_t \in \mathbb{R}^{m'\times n'\times 1}$ is the probability map of class t in eight tissue classes. The class index $c(i,j)$ of $\mathbf{I}(i,j)$ was selected as:

$$c(i,j) = \underset{t}{\mathrm{argmax}}\left(\boldsymbol{p}(i,j)\right) \quad (4)$$

and the segmentation map $\mathbf{S} \in \mathbb{R}^{m'\times n'\times 1}$ can be obtained on $\mathbf{M}$ by calculating the class index $c(i,j)$ of each position:

$$\mathbf{S} = \underset{t}{\mathrm{argmax}}\left(\mathbf{M}\right) \quad (5)$$

where $\mathbf{S}_{ij} = T(i,j)$, $T(i,j)$ represents the tissue class t of $\mathbf{I}(i,j)$. On the basis of the segmentation map, 16 patches of 512 × 512 RGB images in tumour area were randomly extracted at 20× magnification. For the cases of insufficient tumour area, 16 patches were randomly selected with the highest tumour probability. After colour normalizing[49], these patches were combined as the micro mode $\mathbf{C} \in \mathbb{R}^{512\times 512\times(3\times 16)}$ of the WSI.

Datasets description

A summary of the selection and study design of the data used in this work is shown in Supplementary Fig. 1 and Extended Data Fig. 3.

Data source. The data used in this work come from two publicly available datasets, TCGA dataset and PAIP dataset, and the in-house dataset of QHCG dataset (Supplementary Fig. 1 and Extended Data Fig. 3a). In the TCGA dataset, there are 342 WSIs of 330 patients, and each WSI has the clinical information correspondingly. In the PAIP dataset, there are 100 WSIs, but no clinical or survival information available. In the QHCG dataset, there are 1,182 WSIs of 83 patients with clinical information and 151 external WSIs without clinical information. In this study, all WSIs were processed at 20× magnification.

Datasets for WSI segmentation. The training set for segmentation was obtained by meta-annotation on the 151 WSIs with no clinical information of QHCG dataset. The extracted training set had 40,000 patches for each class. The test sets were composed of an internal test set and an external test set to characterize the classification performance and generalization ability of the trained model. The internal test set was randomly annotated by pathologists in QHCG's 1,182 WSIs that were not included in the training set and were not from a same patient, and each class had 550 patches. The external test sets contained TCGA test set and PAIP test set, from which 200 patches per class were randomly extracted, separately.

Datasets for prognosis. A total of 1,182 WSIs from 83 clinically informative patients in QHCG dataset and 342 WSIs from 330 patients in TCGA dataset were used to train and test the prognostic network. The macro mode obtained by WSI decoupling and the patients' survival information constituted the MacroNet prognosis dataset; the micro mode obtained by WSI decoupling and the patients' survival information constituted the MicroNet prognosis dataset. Macro mode, micro mode and patients' survival information constituted the multimodal M2MNet prognostic dataset. The data were split randomly during cross-validation.

DL network architecture

Considering that the macro mode on prognosis has not been explored, while it may have advantages in being easy to interpret with attribution methods, we designed MicroNet, MacroNet and M2MNet, to test whether the performance of macro mode on prognosis can be comparable to that based on tumour cell morphology (micro mode), and whether the combination of tumour cell morphology and spatial distribution information is helpful for prognosis. A summary of network architectures is shown in Extended Data Fig. 4.

MacroNet. To perform survival prediction from macro mode of WSIs, we extended ResNeXt50 to learn the representation feature vector of macro mode and give corresponding risk score by receiving multi-channel sparse macro mode and making survival regression. The MacroNet f_{macro} can be described by three components, the macro mode encoding module $f_{\mathrm{macro\_enco}}$, the feature compression and stabilization module $f_{\mathrm{comp\_stab}}$, and the prediction module f_{pred}. Specifically, we modified the input channel number of ResNeXt50 to 8 to match channel number of sparse macro mode $\mathbf{M}$. The modified convolution layers were selected as macro mode encoding module $f_{\mathrm{macro\_enco}}$ to encode $\mathbf{M}$ into a more compact 2,048-dimensional feature space by extracting the information of multi-class spatial distribution and interaction. To further compress the encoded macro feature vector $\boldsymbol{k}_{\mathrm{macro}} \in \mathbb{R}^{2048}$ to macro mode representation $\boldsymbol{h}_{\mathrm{macro}} \in \mathbb{R}^{32}$ and improve the robustness of network, a fully connected layer (FC) followed by batch normalization (BN) and rectified linear unit (ReLU) constructed feature compression and stabilization module $f_{\mathrm{comp\_stab}}$. Then the final patient-level risk score $\mathbf{RS}_{\mathrm{macro}}$ was computed from $\boldsymbol{h}_{\mathrm{macro}}$ using f_{pred}, a fully connected layer with weights $\mathbf{V} \in \mathbb{R}^{1\times 32}$ and survival loss function (described in detail in 'Loss function'). The whole model is shown in the equations below:

$$\boldsymbol{k}_{\mathrm{macro}} = f_{\mathrm{macro\_enco}}\left(\mathbf{M}\right) \quad (6)$$

$$\boldsymbol{h}_{\mathrm{macro}} = f_{\mathrm{comp\_stab}}\left(\boldsymbol{k}_{\mathrm{macro}}\right) = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{FC}\left(\boldsymbol{k}_{\mathrm{macro}}\right)\right)\right) \quad (7)$$

$$\mathbf{RS}_{\mathrm{macro}} = f_{\mathrm{pred}}\left(\boldsymbol{h}_{\mathrm{macro}}\right) = \mathbf{V}\boldsymbol{h}_{\mathrm{macro}}^{\mathrm{T}} \quad (8)$$

MicroNet. To perform survival prediction from micro mode of WSIs, we extended ResNeXt50 to learn the representation feature vector of micro mode and give corresponding risk score by receiving multi-channel micro mode and making survival regression. The MicroNet f_{micro} can be described by three components, the micro mode encoding module $f_{\mathrm{micro\_enco}}$, the feature compression and stabilization module $f_{\mathrm{comp\_stab}}$, and the prediction module f_{pred}. Specifically, we modified the input channel number of ResNeXt50 to 48 to match channel number of micro mode $\mathbf{C}$. The modified convolution layers were selected as macro mode encoding module $f_{\mathrm{micro\_enco}}$ to encode $\mathbf{C}$ into a more compact 2,048-dimensional feature space by extracting the information of micro morphology. Feature compression and stabilization module $f_{\mathrm{comp\_stab}}$ was used to further compress the encoded micro feature vector $\boldsymbol{k}_{\mathrm{micro}} \in \mathbb{R}^{2048}$ to micro mode representation $\boldsymbol{h}_{\mathrm{micro}} \in \mathbb{R}^{32}$ and improve the robustness of network. Then the final patient-level risk score $\mathbf{RS}_{\mathrm{micro}}$ was computed from $\boldsymbol{h}_{\mathrm{micro}}$ using f_{pred}. The whole model is shown in the equations below:

$$\boldsymbol{k}_{\mathrm{micro}} = f_{\mathrm{micro\_enco}}\left(\mathbf{C}\right) \quad (9)$$

$$\boldsymbol{h}_{\mathrm{micro}} = f_{\mathrm{comp\_stab}}\left(\boldsymbol{k}_{\mathrm{micro}}\right) = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{FC}\left(\boldsymbol{k}_{\mathrm{micro}}\right)\right)\right) \quad (10)$$

$$\mathbf{RS}_{\mathrm{micro}} = f_{\mathrm{pred}}\left(\boldsymbol{h}_{\mathrm{micro}}\right) = \mathbf{V}\boldsymbol{h}_{\mathrm{micro}}^{\mathrm{T}} \quad (11)$$

M2MNet. To achieve multimodal survival prediction from both macro mode and micro mode, MacroNet and MicroNet were used to extract

macro mode representation $\boldsymbol{h}_{\text{macro}}$ and micro mode representation $\boldsymbol{h}_{\text{micro}}$. Following the unimodal feature representations, multimodal feature representation $\boldsymbol{h}_{\text{fusion}} \in \mathbb{R}^{64}$ was obtained by concatenating $\boldsymbol{h}_{\text{macro}}$ and $\boldsymbol{h}_{\text{micro}}$. To integrate the unimodal feature representations more comprehensively, a fusion module f_{fusion} was designed to first use a fully connected layer expand $\boldsymbol{h}_{\text{fusion}}$ to a 1,024-dimensional fusion feature space and then use feature compression and stabilization module $f_{\text{comp\_stab}}$ with the prediction module f_{pred} make survival prediction.

$$\boldsymbol{h}_{\text{macro}} = f_{\text{comp\_stab}}\left(f_{\text{macro\_enco}}\left(\mathbf{M}\right)\right) \tag{12}$$

$$\boldsymbol{h}_{\text{micro}} = f_{\text{comp\_stab}}\left(f_{\text{micro\_enco}}\left(\mathbf{C}\right)\right) \tag{13}$$

$$\boldsymbol{h}_{\text{fusion}} = \boldsymbol{h}_{\text{macro}} \oplus \boldsymbol{h}_{\text{micro}} \tag{14}$$

$$\mathbf{RS}_{\text{M2M}} = f_{\text{fusion}}\left(\boldsymbol{h}_{\text{fusion}}\right) \tag{15}$$

Loss function. To perform survival prediction for both unimodal and multimodal networks, we selected the negative Cox partial log-likelihood as the loss function[50]. Let the survival function $S(t) = P(T \geq t_0)$ be the probability of a patient surviving longer than time t_0, where T is a continuous random variable that represents patient survival time, the hazard function $h(t)$ that describes probability that an event occurs instantaneously at a time t (after t_0) can be written as:

$$h(t) = \lim_{\partial t \to 0} \frac{P(t \leq T \leq t + \partial t \mid T \geq t)}{\partial t} \tag{16}$$

and the survival function $S(t)$ is the integration of the hazard function $h(t)$ over the time between t and t_0:

$$S(t) = \exp\left(-\int_0^t h(x)\,\partial x\right) \tag{17}$$

Assuming that the hazard function can be parameterized as an exponential linear function, Cox proportional hazards model makes a semi-parametric approach for estimating the hazard function:

$$h(t|\boldsymbol{X}_i) = b_0(t)\, e^{\boldsymbol{X}_i^T \cdot \boldsymbol{\beta}} \tag{18}$$

where $b_0(t)$ is the baseline hazard that describes how the risk of an event changes over time, $\boldsymbol{\beta}$ is model parameters vector that describes how the hazard varies with features vector $\boldsymbol{X}_i$ of patient i. On the basis of the Cox proportional hazards model, the negative Cox partial log-likelihood is as follows:

$$l(\boldsymbol{\beta}) = -\sum_{i \in U}\left(\boldsymbol{X}_i^T \cdot \boldsymbol{\beta} - \log \sum_{j \in R_i} e^{\boldsymbol{X}_j^T \cdot \boldsymbol{\beta}}\right) \tag{19}$$

where U is the set of uncensored patients, $R_i = \{j \mid Y_j \geq Y_i\}$ is the set of patients whose time of death or last follow-up Y_j is later than patient i. In this loss function, $\boldsymbol{X}_i^T \cdot \boldsymbol{\beta}$ can be regarded as the risk score given by f_{pred}, where $\boldsymbol{\beta}$ is the weights of f_{pred} and $\boldsymbol{X}_i$ is the feature vector of patient i input into f_{pred}. To train MacroNet, MicroNet and M2MNet for survival prediction, we used the negative Cox partial log-likelihood combined with deep networks as loss function, with the derivative of the loss function used as error during back-propagation.

Training details. MacroNet and MicroNet were trained end-to-end with a mini-batch size of 64, using Adam optimization with a learning rate of 5×10^{-3}, b_1 coefficient of 0.9, b_2 coefficient of 0.999 and L_2 weight decay of 4×10^{-4}. M2MNet was trained end-to-end with a mini-batch size of 32, using Adam optimization with a learning rate of 1×10^{-3}, b_1 coefficient of 0.9, b_2 coefficient of 0.999 and L_2 weight decay of 4×10^{-4}. To mitigate model overfitting during training, we also added a L_1 regularization term with weight 3×10^{-4} to the loss function and used dropout layers with $P = 0.25$ during M2MNet training.

Attribution methods

To explore the good-performance prognostic model $\hat{f}$, we used attribution techniques to find features or structures that are relevant to the prediction made by $\hat{f}$, which may guide us to discover new biomarkers. There are many attribution techniques to achieve such work, including gradient-based methods[51], feature occlusion and attention weights methods[52]. However, most current attribution techniques can only give attribution maps to achieve two-dimensional contribution spatial location, which may be insufficient to interpret the high-information-density input.

To overcome this problem and explore the relationship between macro mode and prognosis, we decoupled input WSIs into sparse macro mode and trained high-performance MacroNet. The macro mode, which only has tissue spatial distribution and interaction information, matches well with the attribution maps produced by current attribution techniques, and the extremely sparse and explicit information of macro mode makes the interpretation more objective and accurate. In this work, we used saliency maps, which were generated by calculating the gradient of the loss function for risk score with respect to the input pixels[26], combined with segmentation maps of WSIs to achieve interpretation. For better visualization, we made the transparency corresponding to the first 30% of the values in the generated saliency map increasing linearly, and overlapped the saliency map with corresponding segmentation map. The discovered features can then be useful for guiding hypotheses for new biomarkers.

Quantification of WSI macro mode

Tissue fraction. On the basis of the segmentation map $\mathbf{S}$, the tissue fraction of class t in seven tissue classes (exclude empty) can be written as:

$$\text{Fraction}_t = \frac{N_t}{N - N_{\text{empty}}} \tag{20}$$

where N_t is the number of pixels belong to class t in $\mathbf{S}$, N_{empty} is the number of empty pixels in $\mathbf{S}$ and N is the number of all pixels in $\mathbf{S}$.

TIL. TILs have been shown to be a key prognostic indicator for a range of cancers[12]. We quantified TILs on the basis of segmentation map $\mathbf{S}$ and TIL abundance (TILAb) score[29]. Specifically, $\mathbf{S}$ was divided into $m \times n$ equal sized grids, and the grid size was selected as ten pixels in our work. Then the co-localization score M in terms of the Morisita–Horn index is defined as[53]

$$M = \frac{2\sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\text{INF}} \times p_{ij}^{\text{TUM}}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\text{INF}}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\text{TUM}}\right)^2} \tag{21}$$

where p_{ij}^{INF} and p_{ij}^{TUM} represent the percentage of inflammation and tumour regions in the (i,j)th grid-cell, respectively. Considering the inflammatory proliferation in tumour as a good prognostic indicator for patient survival, the quantified TILs can be written as:

$$\text{TIL} = \begin{cases} \frac{M}{2} \times \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\text{INF}}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\text{TUM}}\right)}, \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\text{TUM}}\right) > 0 \\ 1, \quad \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\text{TUM}}\right) \leq 0 \end{cases} \tag{22}$$

NEC and TND. To characterize and verify that NEC and TND were prognostic biomarkers, we built their mathematical models based on $\mathbf{S}$. For NEC, we used the tissue fraction model to quantify it:

$$\mathrm{NEC} = \mathrm{Fraction}_{\mathrm{NEC}} = \frac{N_{\mathrm{NEC}}}{N - N_{\mathrm{empty}}} \tag{23}$$

where N_{NEC} is the number of pixels belong to necrosis in **S**.

TIL quantifies the spatial distribution and the interaction between tumour and inflammation to characterize TILs, whereas TND is used to quantify the spatial intersection of tumour boundaries and necrosis boundaries, which is essentially the spatial distribution and interaction between tumour and necrosis, to characterize high attribution areas for MacroNet prognosis.

Therefore, we modified TIL into TND by changing p_{ij}^{INF} into the percentage of necrosis regions in the (i,j)th grid-cell p_{ij}^{NEC}:

$$M' = \frac{2\sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\mathrm{NEC}} \times p_{ij}^{\mathrm{TUM}}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\mathrm{NEC}}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\mathrm{TUM}}\right)^2} \tag{24}$$

$$\mathrm{TND} = \begin{cases} \frac{M'}{2} \times \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\mathrm{NEC}}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\mathrm{TUM}}\right)}, \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\mathrm{TUM}}\right) > 0 \\ 1, \sum_{i=1}^{m}\sum_{j=1}^{n}\left(p_{ij}^{\mathrm{TUM}}\right) \le 0 \end{cases} \tag{25}$$

Computational hardware and software

Python (version 3.7.9) packages used by the project include PyTorch (version 1.8.0), Lifelines (version 0.25.11), NumPy (version 1.19.2), Pandas (version 1.2.2), Albumentations (version 0.5.2), OpenCV (version 4.5.1), Pillow (version 7.2.0) and OpenSlide (version 1.1.2). All WSIs were processed on Intel Xeon multi-core central processing units and a total of four Nvidia 3090 graphics processing units. DL models were trained with Nvidia software CUDA 11.1 and cuDNN 8.0.5. Saliency was implemented using Captum (version 0.2.0) (ref. 54). Statistical analyses such as two-sample *t*-tests used implementations from SciPy (version 1.4.1), and log-rank tests and univariable and multivariable analyses used implementations from Lifelines (version 0.25.11). Plotting and visualization packages were generated using Seaborn (version 0.9.0) and Matplotlib (version 3.1.1).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The TCGA diagnostic whole-slide data and corresponding clinical information are available from NIH genomic data commons (https://portal.gdc.cancer.gov/projects/TCGA-LIHC). The PAIP histology data and corresponding annotations are available from the Pathology AI Platform 2019 challenge (https://paip2019.grand-challenge.org/Dataset/). Restrictions apply to the availability of the QHCG data, including WSIs and generated PaSegNet dataset, which were used with institutional permission through institutional review board approval for the current study, and are thus not publicly available. Please email all requests for academic use of raw and processed data to the corresponding author. All requests will be evaluated on the basis of institutional and departmental policies to determine whether the data requested are subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic purposes and will require a formal material transfer agreement. Source data are provided with this paper.

Code availability

All code was implemented in Python using PyTorch as the primary DL package. All code and scripts to reproduce the experiments of this paper are available at https://github.com/Biooptics2021/PathFinder. The code is also available at https://zenodo.org/record/7628549 (ref. 55).

References

1. Kather, J. N. & Calderaro, J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. *Nat. Rev. Gastroenterol. Hepatol.* https://doi.org/10.1038/s41575-020-0343-3 (2020).
2. Ludwig, J. A. & Weinstein, J. N. Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* **5**, 845–856 (2005).
3. Bosman, F. T. & True, L. D. Prognostic biomarkers: an introduction. *Virchows Arch.* **464**, 253–256 (2014).
4. Mandalà, M. & Massi, D. Tissue prognostic biomarkers in primary cutaneous melanoma. *Virchows Arch.* **464**, 265–281 (2014).
5. Hamilton, P. W. et al. Digital pathology and image analysis in tissue biomarker research. *Methods* **70**, 59–73 (2014).
6. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
7. Kleppe, A. et al. Designing deep learning studies in cancer diagnostics. *Nat. Rev. Cancer* **21**, 199–211 (2021).
8. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
9. Shi, J.-Y. et al. Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning. *Gut* https://doi.org/10.1136/gutjnl-2020-320930 (2020).
10. Saillard, C. et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology* **72**, 2000–2013 (2020).
11. Chen, R. J. et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **0062**, 1–1 (2020).
12. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878 (2022).
13. Watson, D. S. et al. Clinical applications of machine learning algorithms: beyond the black box. *Br. Med. J.* **364**, 10–13 (2019).
14. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126 (2022).
15. Vasey, B. et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.* **27**, 186–187 (2021).
16. Kundu, S. AI in medicine must be explainable. *Nat. Med.* **27**, 1328 (2021).
17. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* **393**, 1577–1579 (2019).
18. Elmarakeby, H. A. et al. Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).
19. Barredo Arrieta, A. et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
20. Gunning, D. et al. XAI—Explainable artificial intelligence. *Sci. Robot.* **4**, eaay7120 (2019).
21. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
22. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
23. Bulten, W. et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* **21**, 233–241 (2020).
24. Skrede, O. J. et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet* **395**, 350–360 (2020).

25. S. Xie, R. Girshick, P. Dollár, Z. Tu, & K. He. Aggregated Residual Transformations for Deep Neural Networks. in 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5987–5995 (2017). doi: 10.1109/CVPR.2017.634
26. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. *2nd Int. Conf. Learn. Represent. ICLR 2014 - Work. Track Proc.* 1–8 (2014).
27. McShane, L. M. et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). *Br. J. Cancer* **93**, 387–391 (2005).
28. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
29. Shaban, M. et al. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. *Sci. Rep.* **9**, 1–13 (2019).
30. Tellez, D., Litjens, G., Van Der Laak, J. & Ciompi, F. Neural image compression for gigapixel histopathology image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 567–578 (2021).
31. Davies, A. et al. Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74 (2021).
32. Bijelic, L. & Rubio, E. R. Tumor necrosis in hepatocellular carcinoma—unfairly overlooked? *Ann. Surg. Oncol.* **28**, 600–601 (2021).
33. Wei, T. et al. Tumor necrosis impacts prognosis of patients undergoing curative-intent hepatocellular carcinoma. *Ann. Surg. Oncol.* **28**, 797–805 (2021).
34. Ling, Y. H. et al. Tumor necrosis as a poor prognostic predictor on postoperative survival of patients with solitary small hepatocellular carcinoma. *BMC Cancer* **20**, 1–9 (2020).
35. Vakkila, J. & Lotze, M. T. Inflammation and necrosis promote tumour growth. *Nat. Rev. Immunol.* **4**, 641–648 (2004).
36. Minervini, A. et al. Prognostic role of histological necrosis for nonmetastatic clear cell renal cell carcinoma: correlation with pathological features and molecular markers. *J. Urol.* **180**, 1284–1289 (2008).
37. Trentin, L. et al. Tumour-infiltrating lymphocytes bear the 75 kDa tumour necrosis factor receptor. *Br. J. Cancer* **71**, 240–245 (1995).
38. Mercan, E. et al. Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. *JAMA Netw. Open* **2**, 1–11 (2019).
39. Javed, S., Mahmood, A., Werghi, N., Benes, K. & Rajpoot, N. Multiplex cellular communities in multi-gigapixel colorectal cancer histology images for tissue phenotyping. *IEEE Trans. Image Process.* **29**, 1–1 (2020).
40. Wu, R. et al. Comprehensive analysis of spatial architecture in primary liver cancer. *Sci. Adv.* **7**, eabg3750 (2021).
41. Liu, Y. et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell* **183**, 1665–1681 (2020).
42. Xie, W. et al. Prostate cancer risk stratification via non-destructive 3D pathology with deep learning-assisted gland analysis. *Cancer Res.* https://doi.org/10.1158/0008-5472.can-21-2843 (2021).
43. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A. & Courtiol, P. Self-supervision closes the gap between weak and strong supervision in histology. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2012.03583 (2020).
44. Whang, S. E., Roh, Y., Song, H. & Lee, J.-G. Data collection and quality challenges in deep learning: a data-centric AI perspective. *VLDB J.* (2023). doi: 10.1007/s00778-022-00775-9
45. Yamashita, R. et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol.* **22**, 132–141 (2021).
46. Kather, J. N. et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med.* **16**, 1–22 (2019).
47. Otsu, N. A threshold selection method from gray level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
48. J. Deng et al. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). doi: 10.1109/CVPR.2009.5206848
49. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**, 1962–1971 (2016).
50. Verweij, P. J. M. & Van Houwelingen, H. C. Penalized likelihood in Cox regression. *Stat. Med.* **13**, 2427–2436 (1994).
51. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. *34th Int. Conf. Mach. Learn. ICML 2017* **7**, 5109–5118 (2017).
52. Xu, K. et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *32th Int. Conf. Mach. Learn. ICML 2015* **37**, 2048–2057 (2015).
53. Horn, H. S. Measurement of 'overlap' in comparative ecological studies. *Am. Nat.* **100**, 419–424 (1966).
54. Kokhlikyan, N. et al. Captum: a unified and generic model interpretability library for PyTorch An Overview of the Algorithms. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2009.07896 (2020).
55. Liang, J & Kong, L. PathFinder. *Zenodo* https://doi.org/10.5281/zenodo.7628549 (2023).

Acknowledgements

We thank Y. Gao, S. Yang and X. Chen for helpful comments on the manuscript. The study by L.K. and J.L. was partially supported by the STI2030-Major Projects (no. 2022ZD0212000), National Natural Science Foundation of China (NSFC) (nos. 61831014, and 32021002), Tsinghua-Foshan Innovation Special Fund (TFISF) (no. 2021THFS0207) and the Guoqiang Institute, Tsinghua University (no. 2021GQG1024). Y.X. was supported by the Beijing Tsinghua Changgung Hospital Fund (no. 12021C1009).

Author contributions

L.K. and J.L. conceived the idea. L.K. supervised the project. J.L. and Y.X. performed the experiments. Y.X., Y.J. and W.M. curated the QHCG dataset. J.L., Y.X. and W.Z. analysed the results. Q.D. and H.Y. provided helpful discussions on the project design. J.L. and L.K. prepared the manuscript with inputs from all co-authors.

Competing interests

The authors declare that they have no competing financial interests.

Additional information

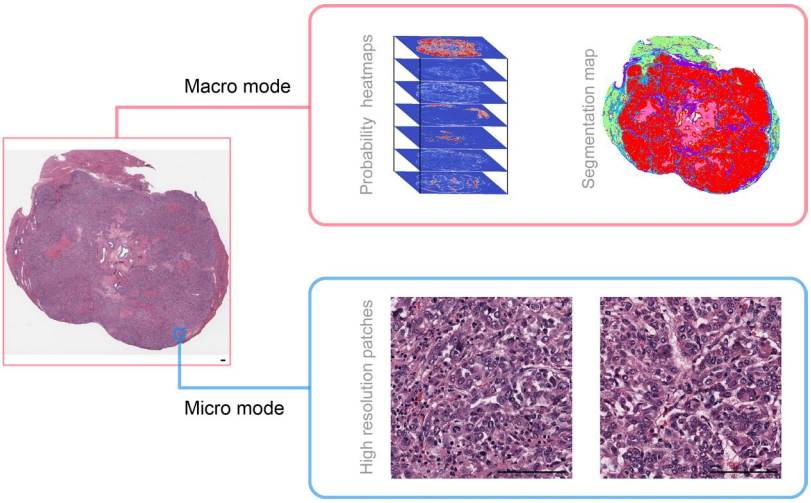
Extended data is available for this paper at https://doi.org/10.1038/s42256-023-00635-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00635-3.

Correspondence and requests for materials should be addressed to Qionghai Dai, Hongfang Yin, Ying Xiao or Lingjie Kong.

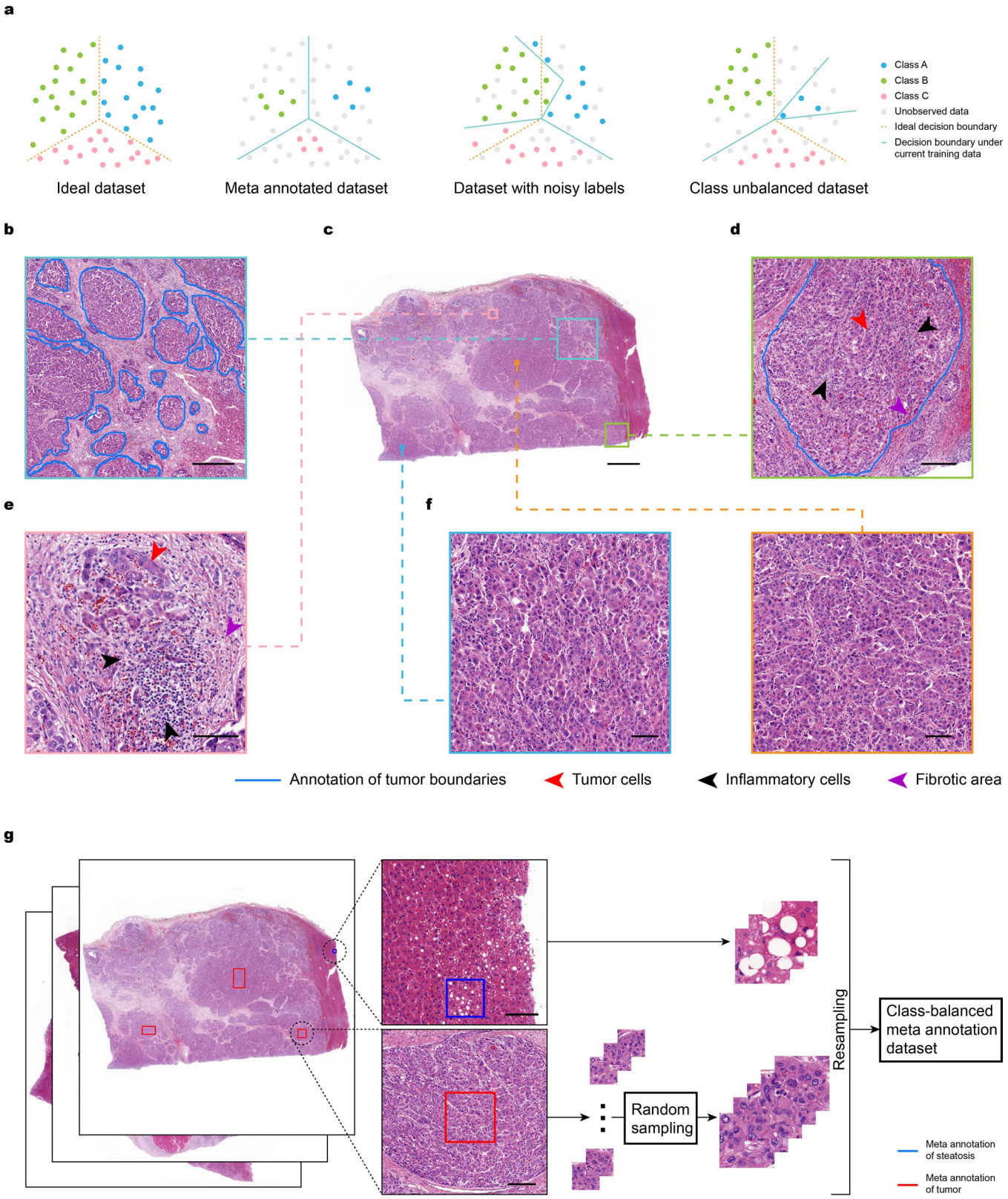
Peer review information *Nature Machine Intelligence* thanks Jiguang Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-023-00635-3



Extended Data Fig. 1 | The macro mode and micro mode in our model. Macro mode mainly focuses on the global information at low WSIs resolution. In this case, the spatial distribution information of different tissue types is included, while the high-resolution cell morphology information is discarded. On the contrary, micro mode mainly focuses on the region-level information at high spatial resolution. In this case, the high-resolution cell morphology information is included, while the tissue spatial distribution and contextual information are ignored. Scale bar: 100 μm.

https://doi.org/10.1038/s42256-023-00635-3

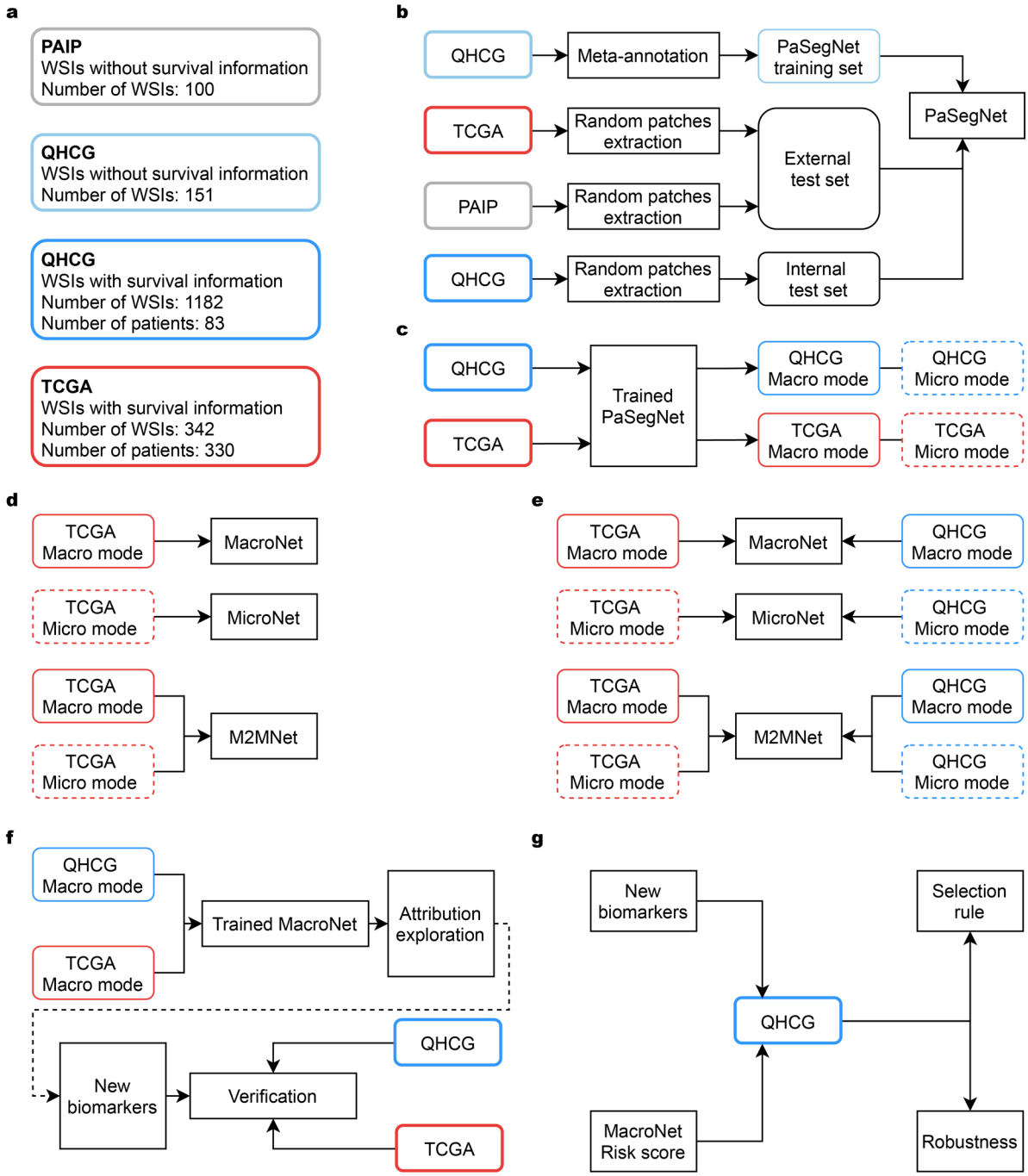


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s42256-023-00635-3

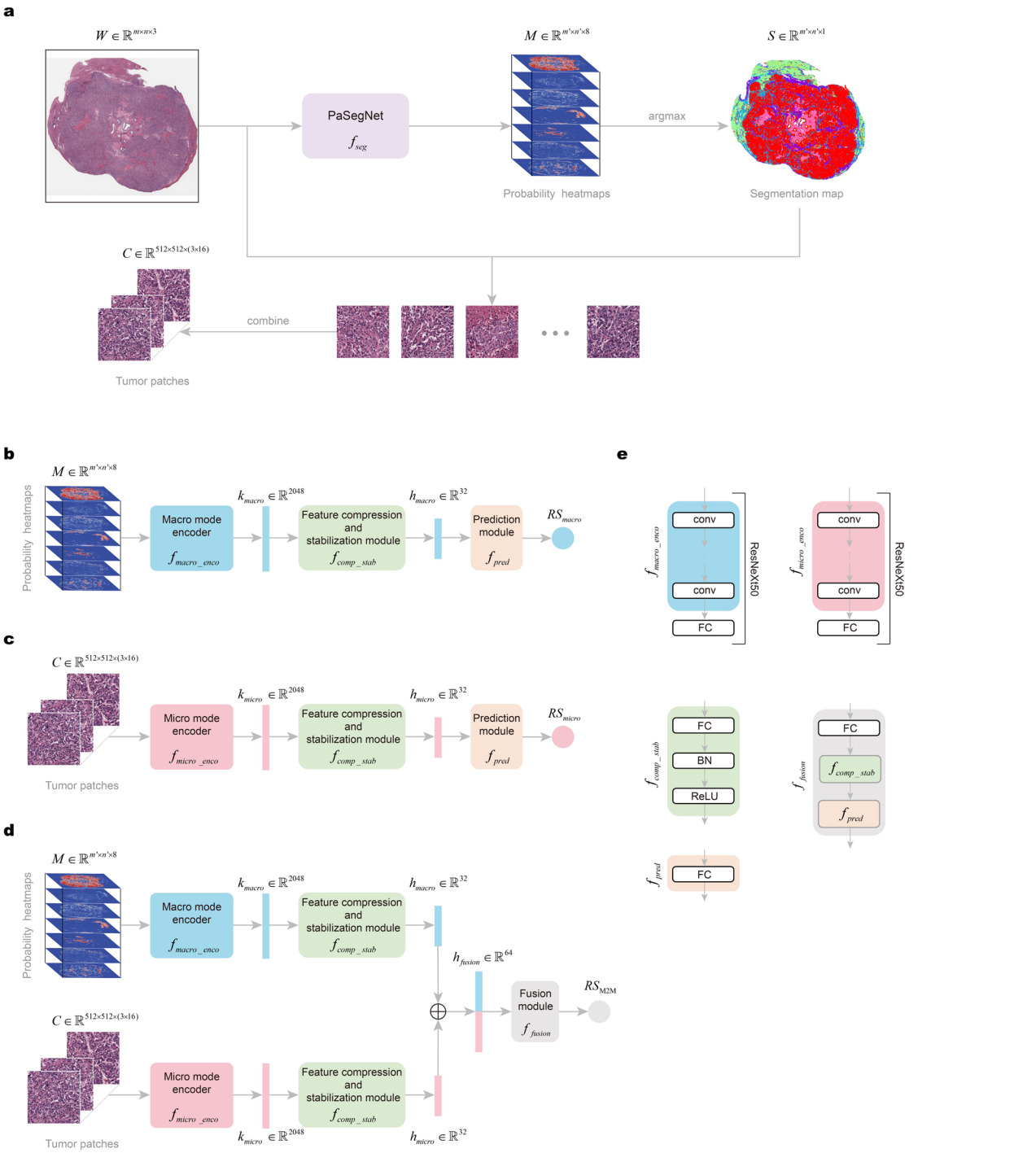
Extended Data Fig. 2 | The gap between pathological annotation and deep learning, and the pipeline of meta annotation. a, The distributions of data points and decision boundaries in latent feature space of different situations. In an ideal situation, DL can learn an ideal decision boundary based on enough and class-balanced data. However, the actual data distribution is often not clear. The dataset we observed usually has noisy labels near the decision boundary, which makes the decision boundary learned by the model jitter in the ideal boundary area, or is class-unbalanced, which makes the decision boundary deviate from the ideal boundary. Meta annotated dataset, collecting a small number of representative data points in each class, is still possible to make the decision boundary close to the ideal boundary. **b**, Conventional pathological annotation method. It usually takes a long time to complete pixel-level annotation of complex tissues. Scale bar, 500 μm. **c**, WSI example. Scale bar, 2000 μm. **d**, The borders or interiors of tumor regions annotated by conventional methods still contain other types of tissue. Scale bar, 200 μm. **e**, An example of annotating regions with great difficulties. Multiple classes of tissue are mixed together. Scale bar, 100 μm. **f**, The tumor morphologies at different spatial locations of the WSI are similar. Scale bar, 100 μm. **g**, The pipeline of our proposed meta annotation. Scale bar: 100 μm (above), 200 μm (below).

https://doi.org/10.1038/s42256-023-00635-3



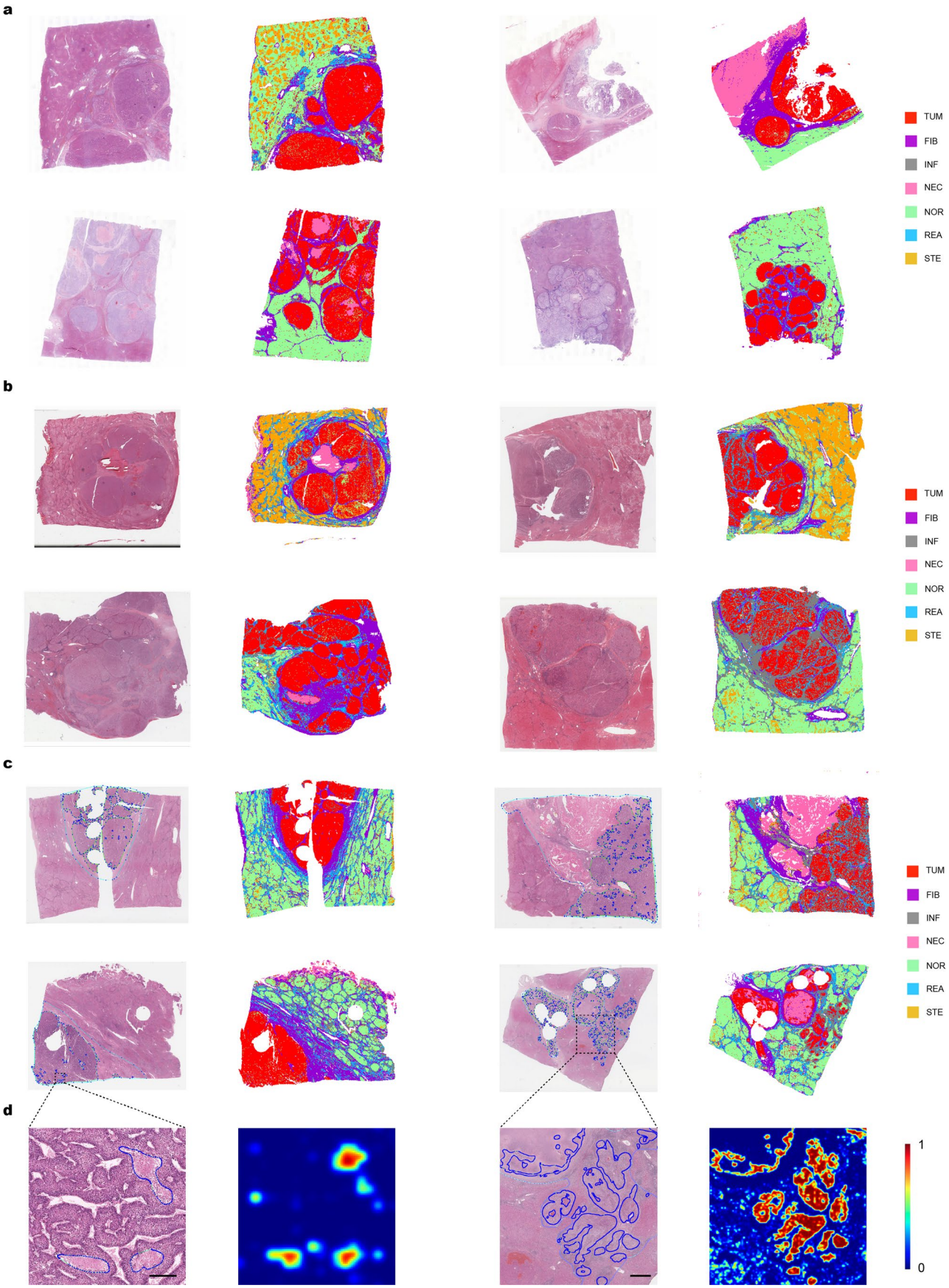
Extended Data Fig. 3 | Summary of study design and data usage. **a**, Information of datasets. **b**, Training and validation of PaSegNet. **c**, Acquiring macro mode and micro mode by WSI decoupling and sparsification. **d**, 10-fold cross-validations of prognosis networks on TCGA dataset. **e**, Generalization ability test. The prognosis networks were first trained on TCGA dataset and then tested on QHCG dataset. **f**, Discovery, characterization, and verification of new biomarkers. **g**, Exploration of macro mode robustness and multiple WSIs selection rule.

https://doi.org/10.1038/s42256-023-00635-3



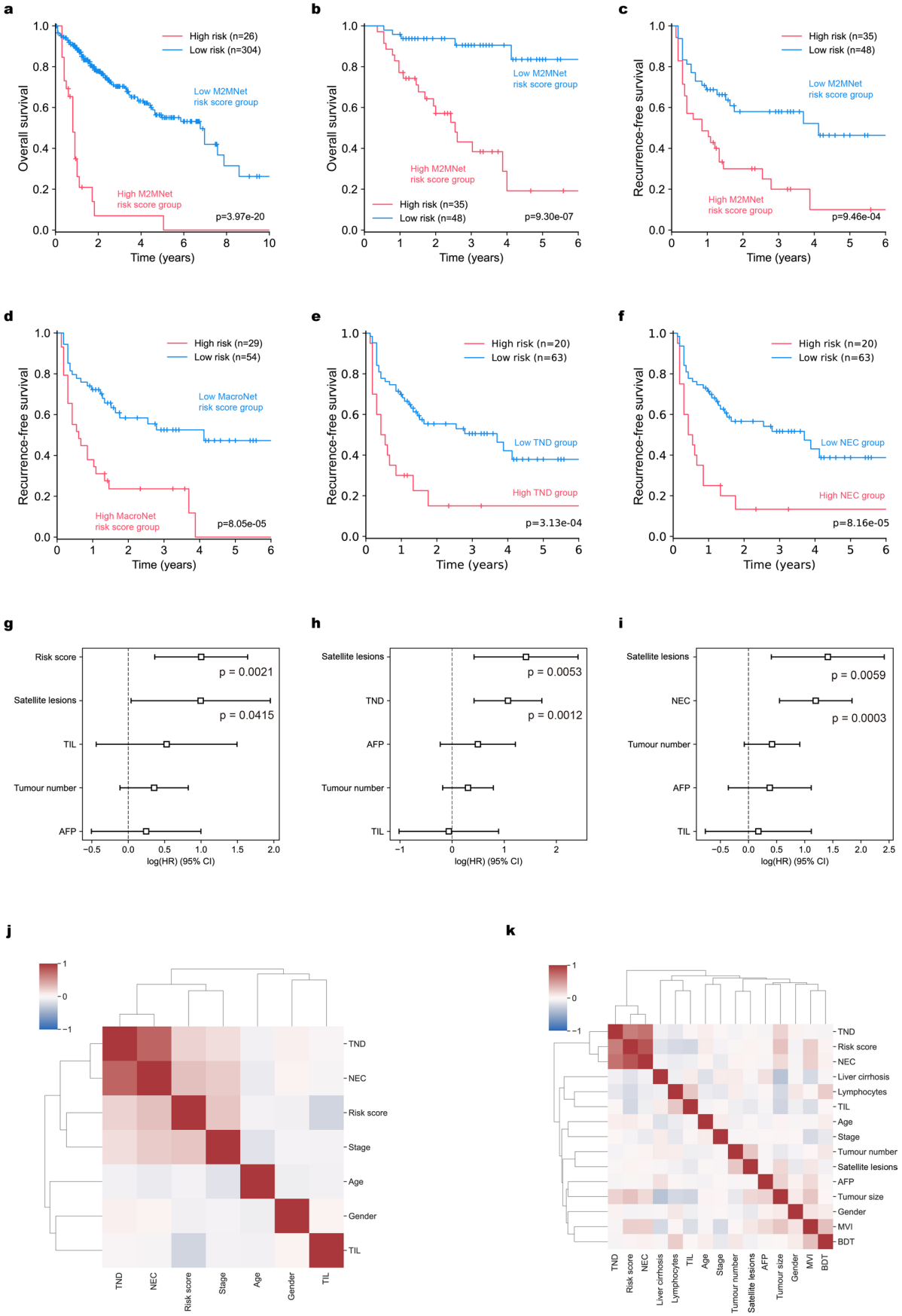
Extended Data Fig. 4 | Neural network architectures and detailed processes of various modes. a, The process of obtaining probability heatmaps, segmentation maps, and tumor patches based on PaSegNet. **b, c, d**, Neural network architectures and detailed processes of MacroNet (**b**), MicroNet (**c**) and M2MNet (**d**), respectively. **e**, The detailed architecture of each neural network module in the model.

https://doi.org/10.1038/s42256-023-00635-3



Extended Data Fig. 5 | Segmentation results. a, Segmentation results of QHCG WSIs. **b**, Segmentation results of TCGA WSIs. **c**, Segmentation results of PAIP WSIs. **d**, Segmentation results of small key lesion regions. Left, necrosis regions and corresponding probability heatmap. Scale bar, 250 μm. Right, tumor regions and corresponding probability heatmap. Scale bar, 1 mm. TUM, tumor; Nor, normal; FIB, fibrosis; INF, inflammation; NEC, necrosis; REA, bile duct reaction; STE, steatosis.

https://doi.org/10.1038/s42256-023-00635-3

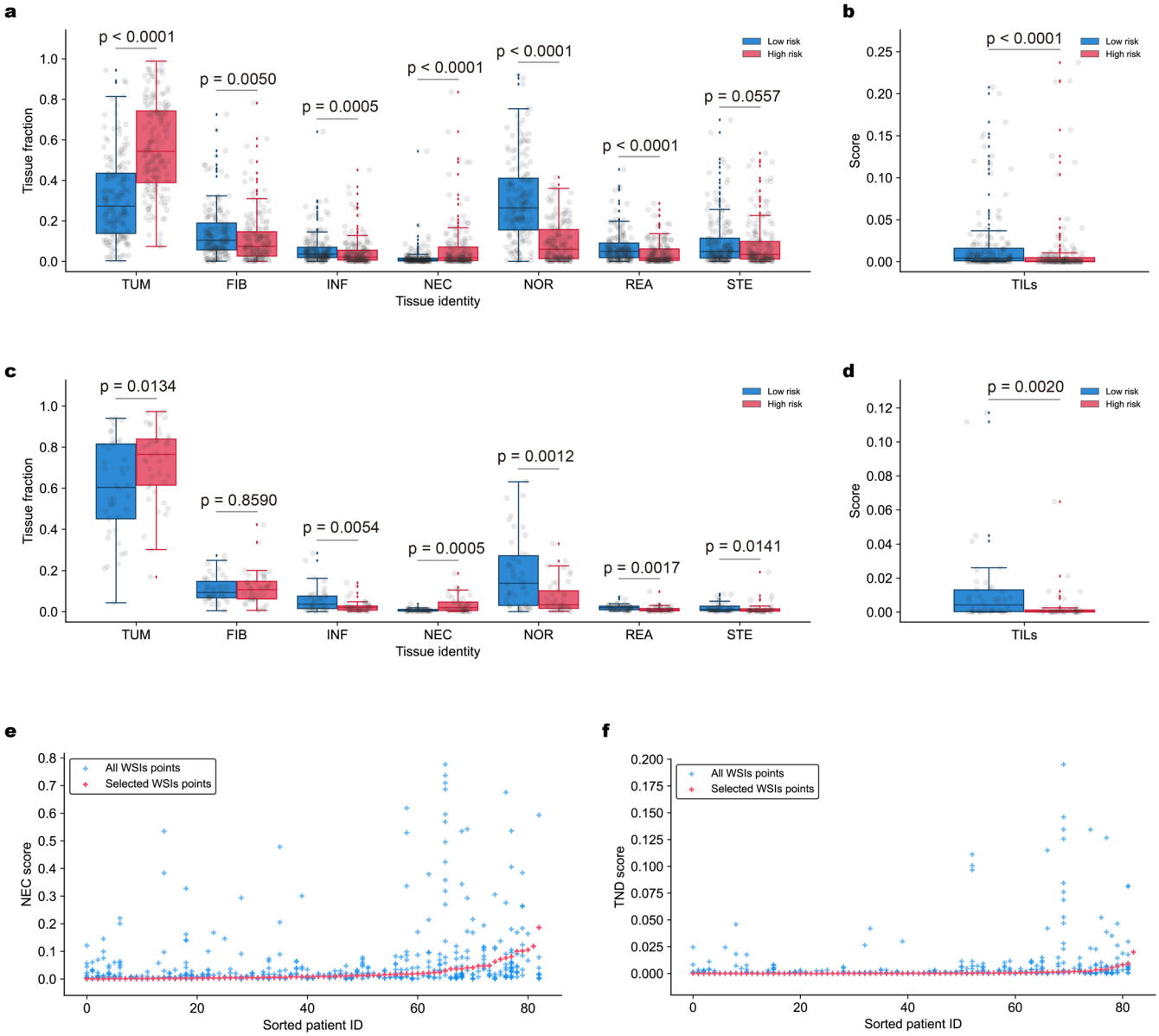


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s42256-023-00635-3

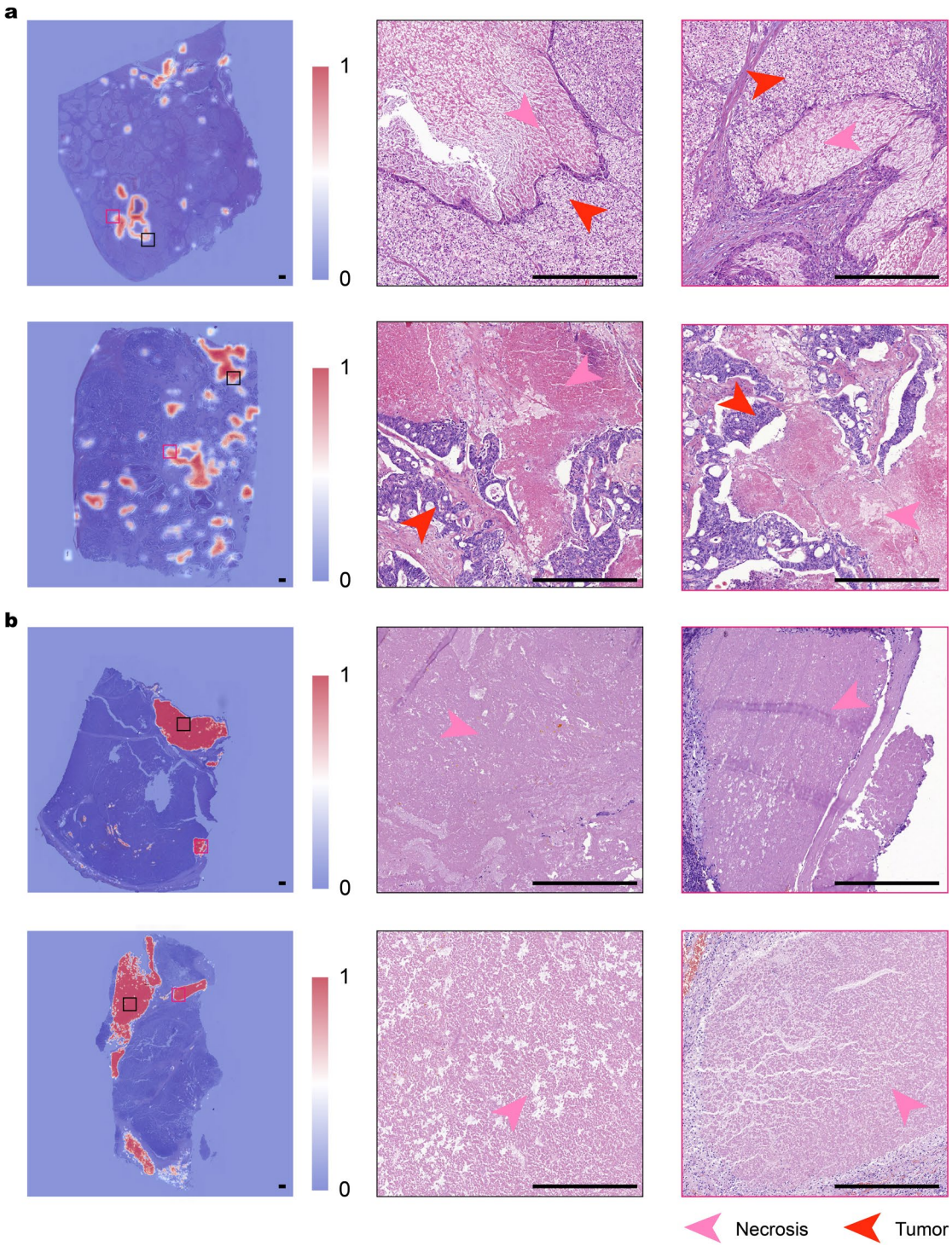
Extended Data Fig. 6 | Survival and recurrence analyses on TCGA and QHCG dataset, and the correlation maps of clinical parameters. a, b, Kaplan-Meier analyses of patient stratification of low and high death risk patients via M2MNet on TCGA dataset (**a**) and QHCG dataset (**b**). **c-f**, Kaplan-Meier analyses of patient stratification of low and high recurrence risk patients via M2MNet (**c**), MacroNet (**d**), TND (**e**), and NEC (**f**) on QHCG dataset. **g-i**, Multivariable analyses of factors associated with recurrence and MacroNet (**g**), TND (**h**), and NEC (**i**) on QHCG dataset (n = 83 patients); the data are presented as hazard ratio estimates (squares) and the error bars show the 95%-confidence interval of the hazard ratio estimate, according to multivariable Cox proportional hazards model. The results of univariate, multivariate analyses, and the abbreviations of each variable are detailed in Supplementary Table 3. **j, k**, Correlation maps of clinical parameters on TCGA dataset (**j**) and QHCG dataset (**k**). P values according to two-sided log-rank test (**a-f**) and multivariable Cox proportional hazards model (**g-i**). n, sample size; HR, hazard ratio; Stage, AJCC staging; TIL, tumor infiltrating lymphocytes digital score; BDT, bile duct thrombosis; AFP, alpha-fetoprotein; MVI, microvascular invasion.

https://doi.org/10.1038/s42256-023-00635-3



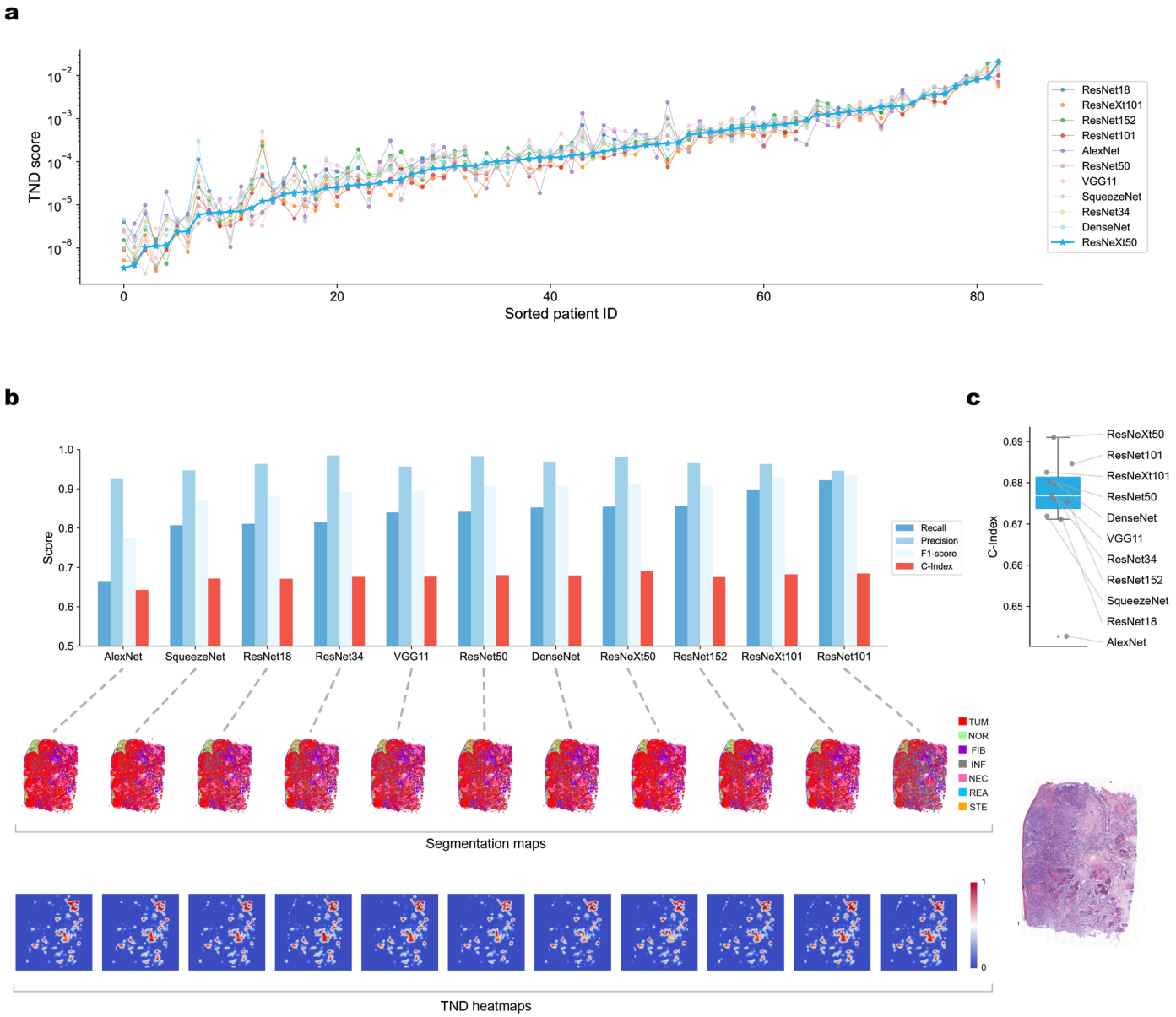
Extended Data Fig. 7 | Quantification analysis of macro mode, and the indicator distributions among all WSIs. a, Quantification of tissue fraction on TCGA dataset ($n = 330$ patients). **b**, Quantification of TIL on TCGA dataset ($n = 330$ patients). **c**, Quantification of tissue fraction on QHCG dataset ($n = 83$ patients). **d**, Quantification of TIL on QHCG dataset ($n = 83$ patients). **e**, Distribution of NEC score from different WSIs of a same patient. **f**, Distribution of TND score from different WSIs of a same patient. **a**-**d**, The median risk score value is taken as the cutoff value of high risk group and low risk group; the significance level shown is determined using a two-sided Mann-Whitney-Wilcoxon test; boxplot whiskers extend to the smallest and largest value within 1.5 times the interquartile ranges of hinges, and box centre and hinges indicate median and first and third quartiles, respectively. TIL, tumor infiltrating lymphocytes digital score; TUM, tumor; Nor, normal; FIB, fibrosis; INF, inflammation; NEC, necrosis; REA, bile duct reaction; STE, steatosis.

https://doi.org/10.1038/s42256-023-00635-3



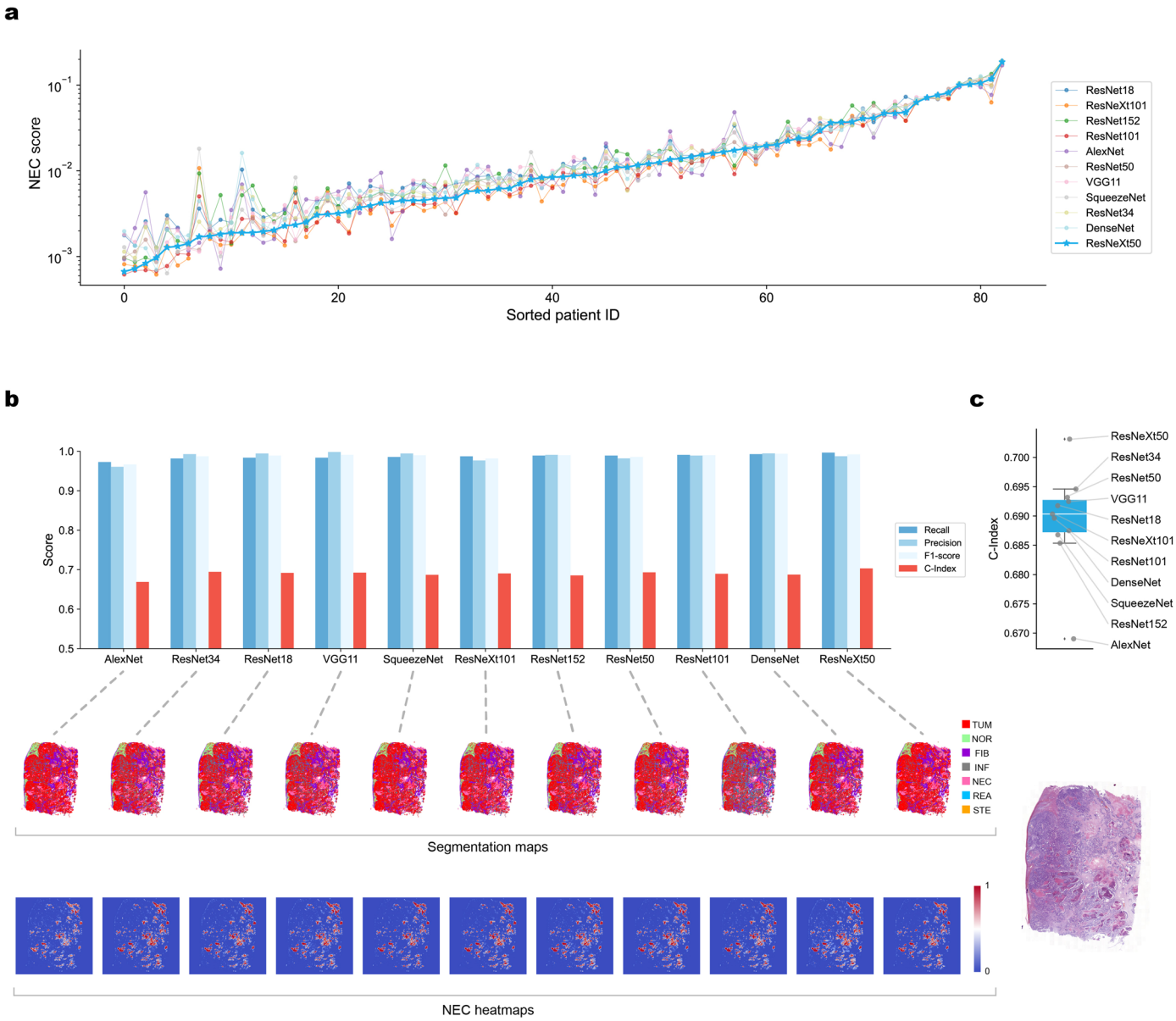
Extended Data Fig. 8 | The localization results of corresponding pathological features of TND and NEC. a, TND heatmaps and pathological features of its localization. **b**, NEC heatmaps and pathological features of its localization. The zoom-in views of pathological slides are from the heatmaps labelled in black and red boxes. Scale bar: 500 μm.

https://doi.org/10.1038/s42256-023-00635-3



Extended Data Fig. 9 | Robustness of TND under different segmentation accuracies. a, TND scores calculated for each patient based on segmentation results generated by 11 CNNs. The TND scores corresponding to ResNeXt50 (the CNN used in this study) are marked with an opaque blue asterisk. Patients are ranked based on TND scores corresponding to ResNeXt50. **b**, Classification performance, segmentation results, TND heatmaps, and prognostic performance of different CNNs. Histograms include recall, precision, and F1-score for each CNN's 'tumor' category tested on QHCG test set, as well as TND prognostic performance (C-Index) based on segmentation maps generated by each CNN. **c**, Prognostic performance distributions of different CNNs ($n = 11$ networks). Boxplot whiskers extend to the smallest and largest value within 1.5 times the interquartile ranges of hinges, and box centre and hinges indicate median and first and third quartiles, respectively.

https://doi.org/10.1038/s42256-023-00635-3



Extended Data Fig. 10 | Robustness of NEC under different segmentation accuracies. a, NEC scores calculated for each patient based on segmentation results generated by 11 CNNs. The NEC scores corresponding to ResNeXt50 (the CNN used in this study) are marked with an opaque blue asterisk. Patients are ranked based on NEC scores corresponding to ResNeXt50. **b**, Classification performance, segmentation results, NEC heatmaps, and prognostic performance of different CNNs. Histograms include recall, precision, and F1-score for each CNN's 'necrosis' category tested on QHCG test set, as well as NEC prognostic performance (C-Index) based on segmentation maps generated by each CNN. **c**, Prognostic performance distributions of different CNNs ($n = 11$ networks). Boxplot whiskers extend to the smallest and largest value within 1.5 times the interquartile ranges of hinges, and box centre and hinges indicate median and first and third quartiles, respectively.

nature portfolio

Corresponding author(s): Lingjie Kong

Last updated by author(s): Feb 14, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All code and scripts to reproduce the experiments of this paper are available at https://github.com/Biooptics2021/PathFinder.
Data analysis	OpenSlide (version 1.1.2) was used for reading whole slide images. Analysis code was written in Python (version 3.7.9). These Python libraries were used: PyTorch (version 1.8.0), Lifelines (version 0.25.11), NumPy (version 1.19.2), Pandas (version 1.2.2), Albumentations (version 0.5.2), OpenCV (version 4.5.1), Pillow (version 7.2.0), Captum (version 0.2.0), SciPy (version 1.4.1), Seaborn (version 0.9.0) and Matplotlib (version 3.1.1). Deep learning models were trained with Nvidia softwares CUDA 11.1 and cuDNN 8.0.5.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The TCGA diagnostic whole-slide data and corresponding clinical information are available from NIH genomic data commons (https://portal.gdc.cancer.gov/

projects/TCGA-LIHC). The PAIP histology data and corresponding annotations are available from the Pathology AI Platform 2019 challenge (https://paip2019.grand-challenge.org/Dataset/). Restrictions apply to the availability of the QHCG data, including whole slide images and generated PaSegNet dataset, which were used with institutional permission through IRB approval for the current study, and are thus not publicly available. Please email all requests for academic use of raw and processed data to the corresponding author. All requests will be evaluated based on institutional and departmental policies to determine whether the data requested is subject to intellectual property or patient privacy obligations. Data can only be shared for non-commercial academic purposes and will require a formal material transfer agreement.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Detailed gender distributions for the datasets used in this study are provided in Extended Data Figs. 3c, d.
Population characteristics	Public Data: TCGA and PAIP contain data from a diverse population representing multiple hospitals. QHCG Data: Patient demographics are consist with demographics of all patients who undergo pathology diagnosis at the hospital. Population characteristics including prognosis information are summarized in the Datasets Description section, Extended Data Fig. 2 presents clinical parameters and characteristics distribution of patients, Extended Data Fig. 3 presents the study design and data usage.
Recruitment	Patients were not directly involved or recruited for the study. This study involved retrospective analysis of pathology slides from patients obtained during standard clinical care.
Ethics oversight	This study was approved by the Beijing Tsinghua Changgung Hospital institutional review board. Only retrospective data was used for research, without any active involvement of patients.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The data used in this work comes from two publicly available datasets, including The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA dataset) and Pathology AI Platform 2019 challenge (PAIP dataset), and the in-house dataset of Beijing Tsinghua Changgung Hospital (QHCG dataset) (Extended Data Fig. 2, Extended Data Fig. 3a). In the TCGA dataset, there are 342 WSIs of 330 patients, and each WSI has the clinical information correspondingly. In the PAIP dataset, there are 100 WSIs, but no clinical or survival information available. In the QHCG dataset, there are 1182 WSIs of 83 patients with clinical information and 151 external WSIs without clinical information. In this study, all WSIs were processed at 20× magnification.
Data exclusions	Pre-established exclusion criteria included: 1. Cases that did not have available prognostic information. 2. Slides that were corrupted or did not clear for prognosis. 3. Slides that did not contain tissue spatial distribution information (like needle biopsy). No other cases or slides were excluded.
Replication	All training and validation experiments were performed using 10-fold cross validation. Machine Learning results have been generated with versioned Python-code for reproducibility.
Randomization	It was a Machine Learning study on retrospective data. For each 10-fold cross validation on TCGA dataset, patients were randomly assigned to either the training or the validation set. QHCG dataset was used as an external test set in this study.
Blinding	No blinding was used in our experiments. As it was a retrospective study, all relevant data had been collected prior to training, validation, and testing of the model, so there was no need/way to perform this in a blinded fashion. No subjective evaluation which required blinding was performed in our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging