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

Lingjie Kong
kong1j@tsinghua.edu.cn
Tsinghua University https://orcid.org/0000-0002-8250-7547
Junhao Liang
Tsinghua University https://orcid.org/0000-0002-6316-661X
Jianghui Yang
Beijing Tsinghua Changgung Hospital
Meilong Wu
Tsinghua University
Weisheng Zhang
Tsinghua University
Qionghai Dai
Tsinghua University https://orcid.org/0000-0001-7043-3061
Hongfang Yin
Beijing Tsinghua Changgung Hospital https://orcid.org/0000-0001-7744-1953
Ying Xiao
Beijing Tsinghua Changgung Hospital https://orcid.org/0000-0003-4316-4825

Article

Keywords:

Posted Date: March 19th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-1452983/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Machine Intelligence on April 3rd, 2023. See the published version at https://doi.org/10.1038/s42256-023-00635-3.

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

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

[1] State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing 100084, China.

[2] Department of Pathology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing 102218, China.

[3] School of Clinical Medicine, Tsinghua University, Beijing 100084, China.

[4] Department of Automation, Tsinghua University, Beijing 102218, China.

[5] IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China.

[6] Current address: Division of Hepatobiliary and Pancreas Surgery, Department of General Surgery, Shenzhen People's Hospital, The Second Clinical Medical College, Jinan University, Shenzhen 518020, Guangdong, China; division of Hepatobiliary and Pancreas Surgery, Department of General Surgery, The First Affiliated Hospital, Southern University of Science and Technology, Shenzhen 518020, Guangdong, China.

[*] To whom correspondence may be addressed. e-mails: konglj@tsinghua.edu.cn, xya02486@btch.edu.cn, yhfa00530@btch.edu.cn, daiqh@tsinghua.edu.cn

Abstract

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 (WSIs) 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 tumor necrosis distribution, for practical prognosis, and verified their potentials in clinical prognosis according to Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK)-derived criteria. 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.

Introduction

Pathological analysis of WSIs is the gold standard for cancer diagnosis and prognosis. Tumor 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 is robust with high sensitivity and specificity[3,4]. Thus there is still a desperate need for identifying additional robust biomarkers to guide tumor diagnosis and prognosis, and to direct the research of tumor mechanism[5–7]. Specifically in cancer prognosis, with the advancement of computational pathology in recent years, 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 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 preclinical discovery[18], the identifying of tissue biomarkers from well-performing prognostic DL models is challenging[8–12]. On one hand, the input multi-dimensional images of WSIs for prognosis are of high information density, compared to 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, Extended Data Fig. 1). However, limited by GPUs (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.

Here we propose an interpretable, human-centric, DL framework, named as 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, based on 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 heatmaps as the macro mode of WSIs (Methods). In order 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, 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 (MLP) 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 based on 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. Based on 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 REMARK-derived criteria[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 1182 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 of largest tumor fraction as the patient's representative WSI, as discussed later. We trained MacroNet in a 10-fold cross-validation on TCGA dataset, tested the generalization of trained model on QHCG dataset. In order 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 tumor patches as inputs, and the latter one is based on both macro mode and micro mode, which attempts to fuse these two modes (Methods, 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, Supplementary Figs. 2, 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, 0.941, and AUC of 0.9980, 0.9984, 0.9974, on QHCG, TCGA, 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 heatmaps 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 10-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 logrank 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 well discrimination between the two risk groups (p-value = 1.25×10^{-7}) compared to M2MNet and clinical staging (Figs. 2d, e, 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 well discrimination between two risk groups (p-value = 7.68×10^{-7}) on QHCG dataset, as M2MNet did (Fig. 2g, 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-value = 0.0057, TCGA dataset; HR: 6.56, 95% CI: 2.01 to 21.36, p-value = 0.0018, QHCG dataset) was independent of other clinicopathological characteristics (Figs. 2f, h, Supplementary Tables 1, 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

In order to interpret why MacroNet can achieve high-performance prognosis and to explore which macro features largely contribute to risk score, we 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 Figs. 7a, c). Then we analyzed the segmentation map of high-risk and low-risk WSIs, and observed that necrosis 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 heatmap, 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 tumor 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, 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 tumor 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 tumor-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 Figs. 7b, d, Methods)[12,29], as an indicator designed based on known clinical experience. The Kaplan-Meier curves and logrank test based p-values showed that NEC and TND can significantly distinguish high-risk and low-risk groups on both TCGA and QHCG datasets (Figs. 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-value = 0.0019, QHCG dataset; HR: 1.80, 95% CI: 1.13 to 2.87, p-value = 0.0133, TCGA dataset) and TND (HR: 6.67, 95% CI: 2.36 to 18.85, p-value = 0.0003, QHCG dataset; HR: 3.00, 95% CI: 1.56 to 5.74, p-value = 0.0009, TCGA dataset) were more significant than most clinical indicators including TILs (Figs. 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-value = 0.0003) and TND (HR: 2.92, 95% CI: 1.52 to 5.60, p-value = 0.0012) can even be used as significant indicators in recurrence prediction (Extended Data Figs. 6c-i, 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 only based on NEC (C-Index: 0.703) or TND (C-Index: 0.691) (Figs. 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, 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 tumor fraction one as the patient's representative WSI. In order 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 (Figs. 5c, d, e). Even for NEC and TND, two objective and universal biomarkers, the results based on largest tumor fraction selection rule were better than 94% of the results based on random selection rule, indicating that the largest tumor 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, 10, Supplementary Note 2, Supplementary Figs. 6, 7). Then we measured the corresponding prognostic performance (*i.e.* 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, 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 few attentions and has not been put into clinical staging guidelines in detail[32–34]. Our findings demonstrate that AI can analyze data more objectively and alert us about missing information. Different from highly diverse tumor tissues, necrosis is easier to be distinguished in both clinics and computer vision, which makes it convenient for clinical prognosis. Meanwhile, the mechanisms between tumor 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, tumor necrosis is postulated to be caused by tumor necrosis factors[35], which have been found significant correlations with TILs[36,37]. However, our results suggest a low correlation between tumor 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, benefiting 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 (MIL) or self-supervised learning (SSL) have achieved well-performance on both WSIs 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 analyzed 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 deep learning. Conventional WSIs pathological annotation methods usually annotate the contour of specific tissues, e.g., tumor 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 can’t be annotated precisely at all (Extended Data Fig. 2e). Moreover, tissue area fractions of different classes in WSIs are quite different, e.g., bile duct reaction tissue may occupy 0.01% of the WSI tissue area while tumor 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 unbalance data brings difficulties to DL training (Extended Data Fig. 2a).

However, when it comes to DL, the desired training set is class balanced, high diversity, and low similarity. And 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 of 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 realize 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 well-performance segmentation.

Purpose of meta annotation. We proposed the meta annotation to close the gap between conventional WSIs 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 tumor, or different subtypes, such as early-stage tumor and late-stage tumor. In our experiment, we pay attention to 7 different types of tissues and empty area (TUM, tumor; Nor, normal; FIB, fibrosis; INF, inflammation; NEC, necrosis; REA, bile duct reaction; STE, steatosis; EMP, empty), and different subtypes of the same tissues (e.g., early-stage tumor vs. late-stage tumor) are considered as intraspecific diversity[9,45,46]. The selected 7 tissue types are common, which basically cover histological features that are easily identified at the resolution level of current WSIs. Based on 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 labeling. For example, for large tumor 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 which only occupy small areas, such as inflammation and bile duct reactions, pathologists use rectangular boxes to enclose their regions as much as possible. After annotating WSIs, nonoverlap 150×150 pixels patches are extracted automatically based on the annotated rectangular boxes. Although the impact of class imbalance has been minimized in the annotating process, there is still a situation that TUM and NOR patches are much more than REA and INF patches. To overcome this problem, during automatically extraction, we specify that TUM and NOR classes are randomly extracted up to 100 patches based on rectangular annotations in one WSI, and all annotated regions of other classes are extracted in full patches. After patches 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 heatmaps 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 convolutional neural network[25] pretrained 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 8 tissue classes (TUM, tumor; Nor, normal; FIB, fibrosis; INF, inflammation; NEC, necrosis; REA, bile duct reaction; STE, steatosis; EMP, empty). Specifically, we used the convolution layers f_{covn} of ResNeXt50 to convert $\mathbf{I}$ into 2048-dimensional feature vector, and modified the last output feature of fully connection layers (f_{fc}) $\boldsymbol{g}$'s dimension to 8:

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

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

After training, the PaSegNet can map the input WSI $\mathbf{W} \in \mathbb{R}^{m\times n\times3}$ to macro mode $\mathbf{M} \in \mathbb{R}^{m'\times n'\times8}$, $m' = \mathrm{int}(m/150)$, $n' = \mathrm{int}(n/150)$:

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

where $\mathbf{M}_{ij} = \boldsymbol{p}(i,j)$, $\mathbf{W}_{ij} = \mathbf{I}(i,j)$, $\mathbf{M}_t \in \mathbb{R}^{m'\times n'\times1}$ is the probability map of class t in 8 tissue classes. The class index $c(i,j)$ of $\mathbf{I}(i,j)$ was selected as :

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

and the segmentation map $\mathbf{S} \in \mathbb{R}^{m'\times n'\times1}$ can be obtained on $\mathbf{M}$ by calculating the class index $c(i,j)$ of each position:

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

where $\mathbf{S}_{ij} = T(i,j)$. Based on the segmentation map, 16 patches of 512×512 RGB images in tumor area were randomly extracted at 20× magnification. For the cases of insufficient tumor area, 16 patches were randomly selected with the highest tumor probability. After color normalizing[49], these patches were combined as the micro mode $\mathbf{C} \in \mathbb{R}^{512\times512\times(3\times16)}$ of the WSI.

Datasets Description

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

Data Source. 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) (Supplementary Fig. 1, 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.

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 1182 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. 1182 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 was split randomly during cross validation.

Deep Learning 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 with that based on tumor cell morphology (micro mode), and whether the combination of tumor 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_{macro_enco}, the feature compression and stabilization module f_{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_{macro_enco} to encode $\mathbf{M}$ into a more compact 2048-dimensional feature space by extracting the information of multi-class spatial distribution and interaction. To further compress the encoded macro feature vector $\boldsymbol{k}_{macro} \in \mathbb{R}^{2048}$ to macro mode representation $\boldsymbol{h}_{macro} \in \mathbb{R}^{32}$ 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_{comp_stab}. Then the final patient-level risk score $\mathbf{RS}_{macro}$ was computed from $\boldsymbol{h}_{macro}$ 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 the equations below:

$$\boldsymbol{k}_{macro} = f_{macro\_enco}(\mathbf{M}) \tag{6}$$

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

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

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_{micro_enco}, the feature compression and stabilization module f_{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_{micro_enco} to encode $\mathbf{C}$ into a more compact 2048-dimensional feature space by extracting the information of micro morphology. Feature compression and stabilization module f_{comp_stab} was used to further compress the encoded micro feature vector $\boldsymbol{k}_{micro} \in \mathbb{R}^{2048}$ to micro mode representation $\boldsymbol{h}_{micro} \in \mathbb{R}^{32}$ and improve the robustness of network. Then the final patient-level risk score $\mathbf{RS}_{micro}$ was computed from $\boldsymbol{h}_{micro}$ using f_{pred}. The whole model is shown the equations below:

$$\boldsymbol{k}_{micro} = f_{micro\_enco}(\mathbf{C}) \tag{9}$$

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

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

M2MNet. To achieve multimodal survival prediction from both macro mode and micro mode, MacroNet and MicroNet were used to extract macro mode representation $\boldsymbol{h}_{macro}$ and micro mode representation $\boldsymbol{h}_{micro}$. Following the unimodal feature representations, multimodal feature representation $\boldsymbol{h}_{fusion} \in \mathbb{R}^{64}$ was obtained by concatenating $\boldsymbol{h}_{macro}$ and $\boldsymbol{h}_{micro}$. In order to integrate the unimodal feature representations more comprehensively, a fusion module f_{fusion} was designed to first use a fully connected layer expand $\boldsymbol{h}_{fusion}$ to a 1024-dimensional fusion feature space and then use feature compression and stabilization module f_{comp_stab} with the prediction module f_{pred} make survival prediction.

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

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

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

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

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 $h(t)$ which 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 proportion hazards model makes semi-parametric approach for estimating the hazard function:

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

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 describe how the hazard varies with features vector $\boldsymbol{X}_i$ of patient i. Based on Cox proportion 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, 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, 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 well-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 of 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. Based on segmentation map $\mathbf{S}$, the tissue fraction of class t in 7 tissue classes (exclude empty) can be written as:

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

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

TIL. Tumor infiltrating lymphocytes (TILs) has been shown to be a key prognostic indicator for a range of cancers[12]. We quantified TILs based on segmentation map **S** and TIL abundance (TILAb) score[29]. Specifically, **S** was divided into m × n equal sized grids, and the grid size was selected as 10 pixels in our work. Then the co-localization score M in terms of the Morisita-Horn index is defined as[53]:

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

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

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

NEC and TND. To characterize and verify necrosis area fraction in WSIs (NEC) and tumor necrosis distribution (TND) were prognostic biomarkers, we built their mathematical models based on **S**. For NEC, we used the tissue fraction model to quantify it:

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

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

TIL quantifies the spatial distribution and the interaction between tumor and inflammation to characterize tumor infiltrating lymphocytes. Whereas, TND is used to quantify the spatial intersection of tumor boundaries and necrosis boundaries, which is essentially the spatial distribution and interaction between tumor 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}:

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

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

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 CPUs (Central Processing Units) and a total of four 3090 GPUs (Graphics Processing Units). Deep learning models were trained with Nvidia softwares CUDA 11.1 and cuDNN 8.0.5. Saliency was implemented using Captum (version 0.2.0)[54]. Statistical analyses such as two-sampled *t*-tests used implementations from SciPy (version 1.4.1), and logrank tests, 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).

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 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.

Code availability

All code was implemented in Python using PyTorch as the primary deep learning 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])

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), the Guoqiang Institute, Tsinghua University (No. 2021GQG1024). Y.X. was supported by the Beijing Tsinghua Changgung Hospital Fund (No. 12021C1009).

Authors 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. analyzed 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.

Figure Legends

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, to obtain the spatial distribution probability heatmaps of 7 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 convolutional neural network 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 modeled based on 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.

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. CI, confidence interval. **b,** C-Index distribution of MacroNet, MicroNet, and M2MNet on TCGA dataset in a 10-fold cross-validation (*n* = 10 independent experiments for MacroNet, MicroNet, and M2Mnet, respectively). 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. **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%-confidence interval of the mean estimate (1000 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 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 and multivariate analyses are described in details in Supplementary Tables 1, 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**, **g**) and multivariable Cox proportional hazards model (**f**, **h**). *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.

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 heatmaps of WSIs segmentation maps and their corresponding visualization results of NEC and TND hypothetical indicators. TUM, tumor; Nor, normal; FIB, fibrosis; INF, inflammation; NEC, necrosis; REA, bile duct reaction; STE, steatosis.

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). **b, d, f, h,** 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; details are shown in Supplementary Tables 7, 8. *P* values according to two-sided log-rank test (**a, c, e, g**) and multivariable Cox proportional hazards model (**b, d, f, h**). *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.

Fig. 5 Exploring the robustness of macro mode indicators. a, Sampling strategy of clinical WSIs. NLP, non-neoplastic liver parenchyma; TC, tumor center; TI, tumor-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 tumor) 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 tumor 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.

References

1. Kather, J. N. & Calderaro, J. Development of AI-based pathology biomarkers in gastrointestinal and liver cancer. *Nat. Rev. Gastroenterol. Hepatol.* (2020) doi:10.1038/s41575-020-0343-3.

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* gutjnl-2020-320930 (2020) doi:10.1136/gutjnl-2020-320930.

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.e6 (2022).

13. Watson, D. S. *et al.* Clinical applications of machine learning algorithms: Beyond the black box. *BMJ* **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. *Nature Medicine* vol. 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. *The Lancet* vol. 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. Explainable Artificial Intelligence (XAI). *Def. Adv. Res. Proj. Agency* 1 (2016).

21. B, W. S. Towards Explainable Artificial. **1**, 5–22 (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. Xie, S., Girshick, R. & Doll, P. Aggregated Residual Transformations for Deep Neural Networks http://arxiv.org/abs/1611.05431v2. *Cvpr* 1492–1500 (2017).

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**, (2021).

41. Liu, Y. *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding

in Tissue. *Cell* **183**, 1665-1681.e18 (2020).

42. Xie, W. *et al.* Prostate cancer risk stratification via non-destructive 3D pathology with deep learning-assisted gland analysis. *Cancer Res.* canres.2843.2021 (2021) doi:10.1158/0008-5472.can-21-2843.

43. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A. & Courtiol, P. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. (2020). Preprint at https://arxiv.org/abs/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. Preprint at https://arxiv.org/abs/2112.06409?context=cs (2021).

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. Smith, P. *et al.* NOBUYUKI OTSU. - 1979 - A Tlreshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **20**, 62–66 (1979).

48. Jia Deng *et al.* ImageNet: A large-scale hierarchical image database. 248–255 (2009) doi:10.1109/cvprw.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. In Proc. International Conference on Machine Learning 2048–2057 (PMLR, 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 https://arxiv.org/abs/2009.07896 (2020).

55. Liang, J & Kong, L. PathFinder. *Zenodo* https://doi.org/10.5281/zenodo.7628549 (2023).

Figures

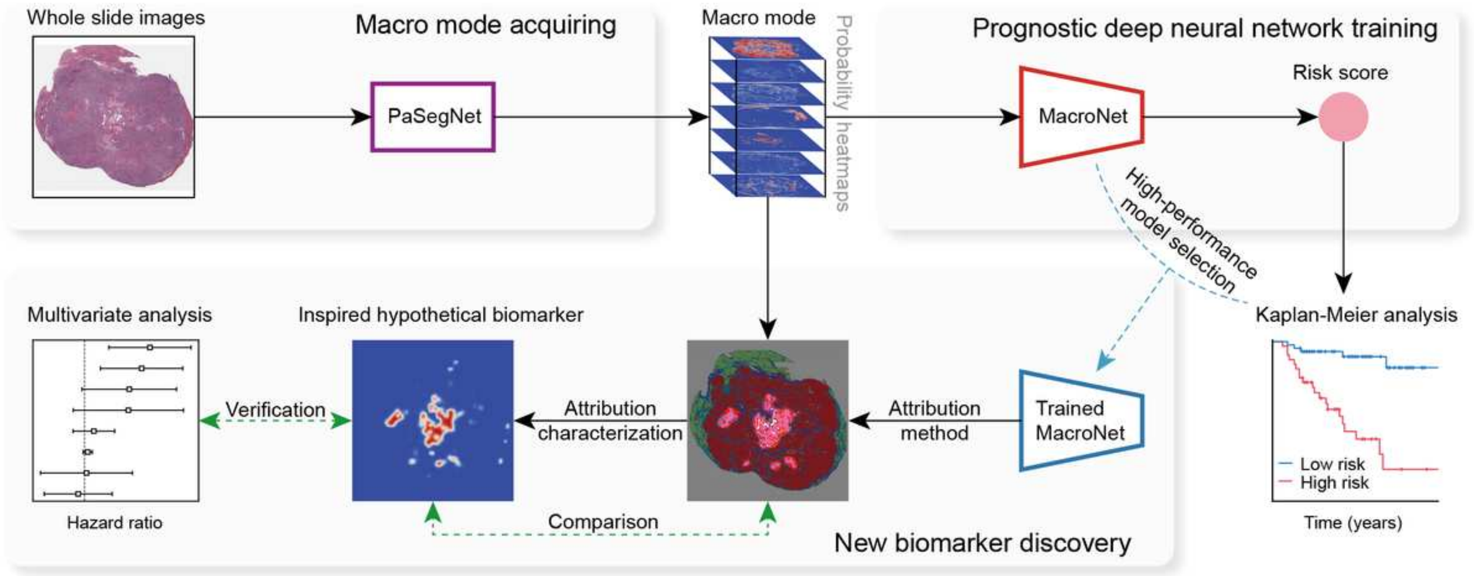

Figure 1

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, to obtain the spatial distribution probability heatmaps of 7 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 convolutional neural network 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 modeled based on 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.

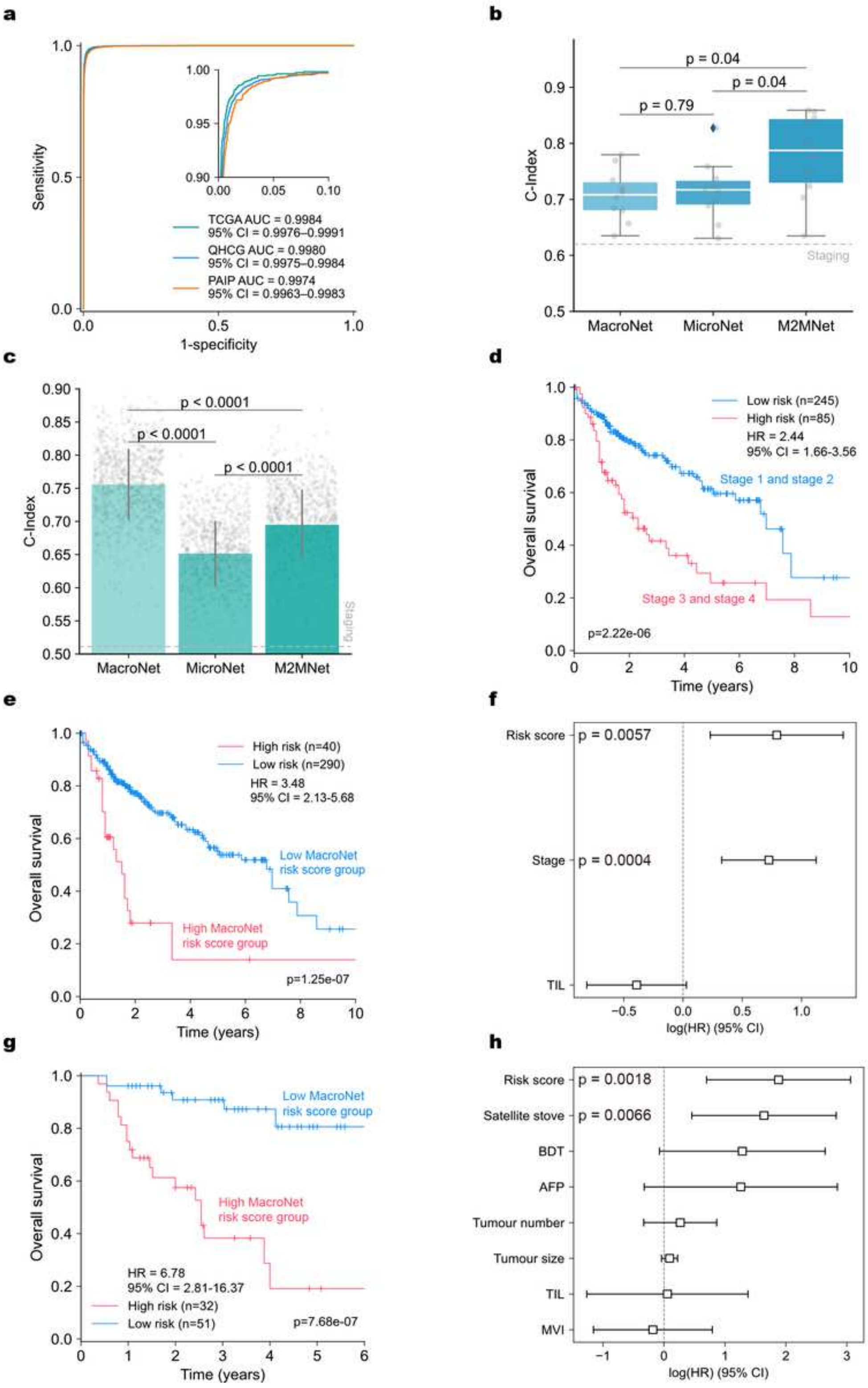

Figure 2

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. CI, confidence interval. **b,** C-Index distribution of MacroNet, MicroNet, and M2MNet on TCGA dataset in a 10-fold cross-validation (*n* = 10 independent experiments for MacroNet, MicroNet, and M2Mnet, respectively). 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. **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%-confidence interval of the mean estimate (1000 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 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 and multivariate analyses are described in details in Supplementary Tables 1, 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**, **g**) and multivariable Cox proportional hazards model (**f**, **h**). *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.

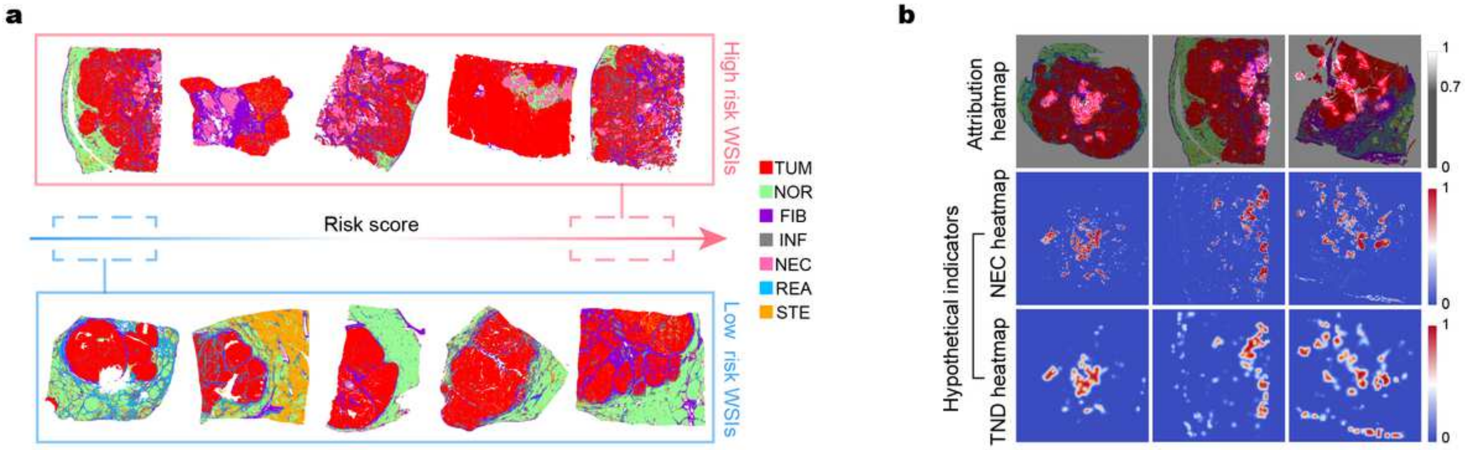

Figure 3

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 heatmaps of WSIs segmentation maps and their corresponding visualization results of NEC and TND hypothetical indicators. TUM, tumor; Nor, normal; FIB, fibrosis; INF, inflammation; NEC, necrosis; REA, bile duct reaction; STE, steatosis.

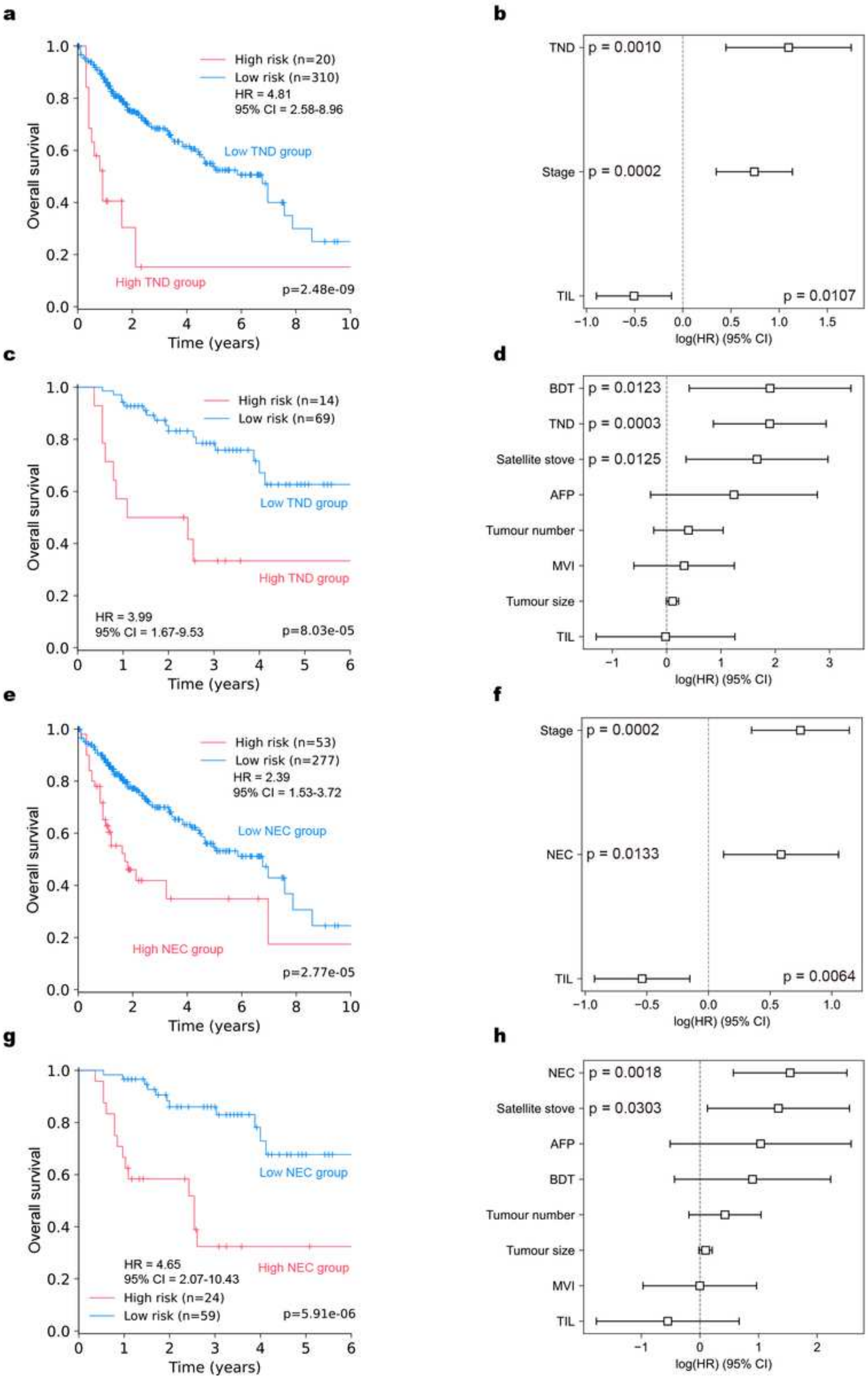

Figure 4

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). **b, d, f, h,** 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; details are shown in Supplementary Tables 7, 8. *P* values according to two-sided log-rank test (**a, c, e, g**) and multivariable Cox proportional hazards model (**b, d, f, h**). *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.

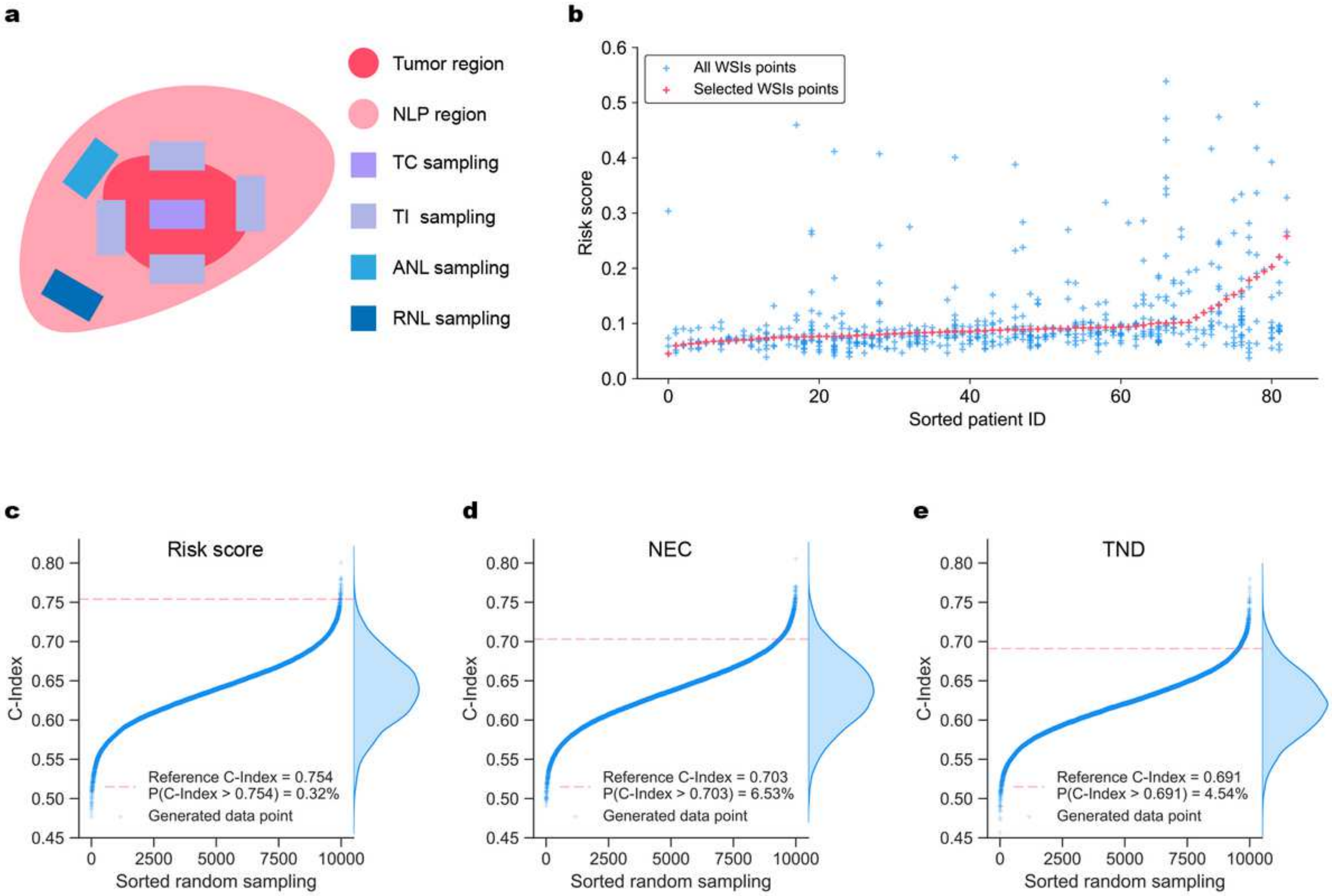

Figure 5

Exploring the robustness of macro mode indicators. a, Sampling strategy of clinical WSIs. NLP, non-neoplastic liver parenchyma; TC, tumor center; TI, tumor-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 tumor) 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 tumor 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 Files

This is a list of supplementary files associated with this preprint. Click to download.

- LiangSupplementaryTables.xlsx
- LiangSI.pdf
- LiangEDFig1.jpg
- LiangEDFig2.jpg
- LiangEDFig3.jpg
- LiangEDFig4.jpg
- LiangEDFig5.jpg
- LiangEDFig6.jpg
- LiangEDFig7.jpg
- LiangEDFig8.jpg
- LiangEDFig9.jpg
- LiangEDFig10.jpg