ORIGINAL RESEARCH

Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning

Jie-Yi Shi,[1] Xiaodong Wang,[2] Guang-Yu Ding,[1] Zhou Dong,[3] Jing Han,[4] Zehui Guan,[3] Li-Jie Ma,[5] Yuxuan Zheng,[2] Lei Zhang,[2] Guan-Zhen Yu,[6] Xiao-Ying Wang,[1] Zhen-Bin Ding,[1] Ai-Wu Ke,[1] Haoqing Yang,[2] Liming Wang,[2] Lirong Ai,[3] Ya Cao,[7] Jian Zhou,[1,8] Jia Fan ,[1,8] Xiyang Liu,[2] Qiang Gao [1,8,9]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2020-320930).

For numbered affiliations see end of article.

Correspondence to
Dr Qiang Gao, Liver Cancer Institute, Zhongshan Hospital Fudan University, Shanghai 200032, China;
gao.qiang@zs-hospital.sh.cn
Professor Xiyang Liu;
xyliu@xidian.edu.cn

J-YS, XW, G-YD and ZD are joint first authors.

Received 19 February 2020
Revised 2 July 2020
Accepted 20 July 2020

To cite: Shi J-Y, Wang X, Ding G-Y, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2020-320930

ABSTRACT

Objective Tumour pathology contains rich information, including tissue structure and cell morphology, that reflects disease progression and patient survival. However, phenotypic information is subtle and complex, making the discovery of prognostic indicators from pathological images challenging.

Design An interpretable, weakly supervised deep learning framework incorporating prior knowledge was proposed to analyse hepatocellular carcinoma (HCC) and explore new prognostic phenotypes on pathological whole-slide images (WSIs) from the Zhongshan cohort of 1125 HCC patients (2451 WSIs) and TCGA cohort of 320 HCC patients (320 WSIs). A 'tumour risk score (TRS)' was established to evaluate patient outcomes, and then risk activation mapping (RAM) was applied to visualise the pathological phenotypes of TRS. The multi-omics data of The Cancer Genome Atlas(TCGA) HCC were used to assess the potential pathogenesis underlying TRS.

Results Survival analysis revealed that TRS was an independent prognosticator in both the Zhongshan cohort ($p<0.0001$) and TCGA cohort ($p=0.0003$). The predictive ability of TRS was superior to and independent of clinical staging systems, and TRS could evenly stratify patients into up to five groups with significantly different prognoses. Notably, sinusoidal capillarisation, prominent nucleoli and karyotheca, the nucleus/cytoplasm ratio and infiltrating inflammatory cells were identified as the main underlying features of TRS. The multi-omics data of TCGA HCC hint at the relevance of TRS to tumour immune infiltration and genetic alterations such as the *FAT3* and *RYR2* mutations.

Conclusion Our deep learning framework is an effective and labour-saving method for decoding pathological images, providing a valuable means for HCC risk stratification and precise patient treatment.

Significance of this study

What is already known on this subject?

- Pathological images of hepatocellular carcinoma (HCC) contain rich phenotypic and molecular information that is essential for tumour diagnosis, prognosis and precision management.
- Deep learning can adaptively extract and interpret tumour image features, representing a promising avenue for automatic diagnosis and predicting patient outcomes.

What are the new findings?

- A deep learning method that includes tissue type discrimination, automatic sampling and prognostication is constructed to output a risk score, identifying the tumour risk score (TRS) as a prognostic factor superior to clinical staging systems in HCC.
- The histological features associated with a high TRS are discerned as sinusoidal capillarisation, prominent nucleoli and karyotheca, a high nucleus/cytoplasm ratio and the absence of tumour-infiltrating immune cells.
- The integration of multi-omics data and histological features indicates the relevance of TRS to the tumour-infiltrating immune cells and gene mutations such as *FAT3* and *RYR2*.

How might it impact on clinical practice in the foreseeable future?

- Our prognostic network based on weakly supervised deep learning is an effective and labour-saving method to improve patient stratification and clinical management in HCC.
- TRS, featured by tumour angiogenesis, cell morphology and immune infiltration, may serve as an innovative determinant of the response to combinational immune, cellular function and anti-angiogenic therapy in HCC.

INTRODUCTION

Hepatocellular carcinoma (HCC), accounting for 75% to 85% of primary liver cancer, is the sixth most common and fourth deadly malignancy globally.[1] HCC is highly heterogeneous at the histological, molecular and genetic levels, making its prognostic stratification and personalised management challenging. During the past decades, high-throughput techniques have been endeavoured to stratify HCC by genomic, transcriptomic and proteomic profiles,[2 3] but the links between those molecular traits and clinical decision-making have not been fully unveiled.

Pathological analysis of HCC specimens is essential for patient prognostic classification beyond diagnosis. The phenotypic information present in pathological tissue reflects the overall effect of the tumour microenvironment on the behaviour of cancer cells. In addition to conventional pathological features such as microvascular invasion and tumour differentiation, novel prognostic histopathological features, including intratumour tertiary lymphoid structures,[4] vessels encapsulating tumour clusters[5] and the macrotrabecular-massive subtype,[6] have been reported in HCC recently. However, histopathological images that contain numerous phenotypic descriptions of the molecular processes underlying HCC initiation and progression remain largely unknown to subjective judgement.[7] Further in-depth analysis is needed to translate the histopathological features of HCC into prognostic and predictive algorithms to improve patient stratification and clinical management.

Deep learning can adaptively extract image features based on learning objectives, emerging as a novel approach for tumour diagnosis and prognosis based on histopathological images.[8–10] Deep learning, as a 'black box' model, is difficult to deconstruct and interpret the features it extracts, making its clinical application impracticable. A method that integrates the patient prognostication and visualisation of prognostic pathological phenotypes may facilitate oncologists to explore new pathological features. In this regard, network activations as saliency masks for visualisations have become a popular method to explain deep learning;[11 12] however, their application in pathological images needs further improvement. The existing methods can only visualise local coarse features rather than accurately decode cell-level pathological features, such as nuclear atypia, mitotic activity, cellular density and tissue architecture. Additionally, a pathological image contains enormous subtle and complex information, rendering pixel-level annotations a huge challenge. Weakly supervised learning that only requires slide-level labels for training partially addresses the above considerations. However, the existing weakly supervised learning usually requires tens of thousands of samples to converge, making it not widely used.

Herein, we present a deep learning framework for predicting HCC patient prognosis based on pathological images from the Zhongshan cohort and the Cancer Genome Atlas (TCGA) HCC cohort (figure 1). We first used manually labelled haematoxylin and eosin (H&E) images as learning materials to train tissue

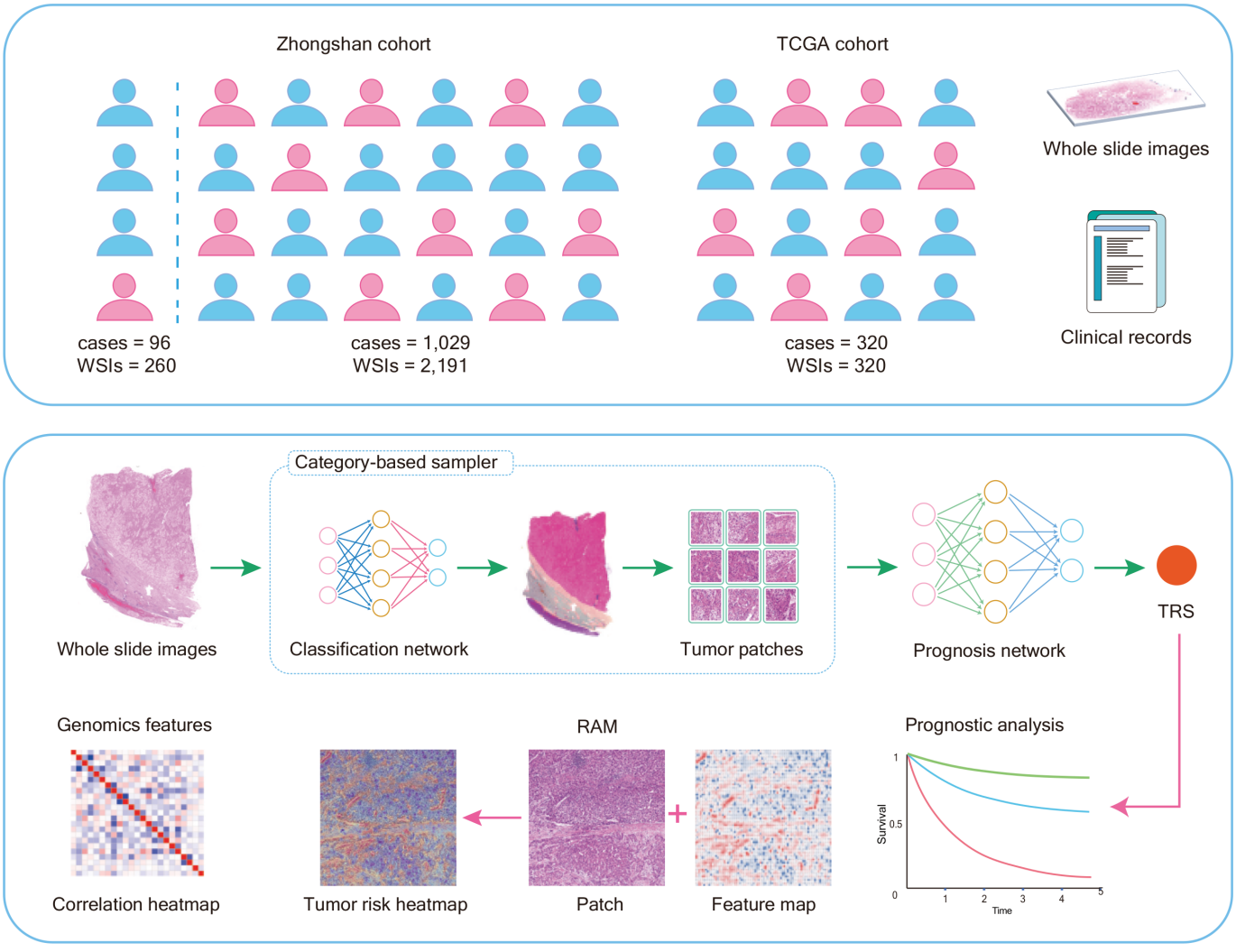

Figure 1 Data and workflow for the prognostic analysis of liver cancer with deep learning. We first developed the classification network using 260 whole-slide images (WSIs) as the category-based sampling. The network was then used to analyse the remaining WSIs and generate the segmentation maps. We randomly sampled tiles from each type of tissue based on these segmentation maps. Next, we trained the prognostic network and calculated a tumour risk score (TRS) for each patient. Finally, we used TRS to predict patient prognosis, and integrate transcriptomics, genomics and neural network heatmaps to identify interpretable features. TCGA, The Cancer Genome Atlas.

category-based local sampling as prior knowledge to prune the feature space. Next, the 'tumour risk score (TRS)' based on weakly supervised deep learning was established, using sampled tiles from tumour tissue as inputs and patient outcomes as labels. Finally, we explored the TRS-related features by risk activation mapping and the relevance of TRS to tumour immune infiltration and gene mutations using The Cancer Genome Atlas (TCGA) HCC data.

METHODS

Patients and follow-up

In our initial cohort, 1125 HCC patients with 2550 pathological slides were randomly selected from patients who had received curative hepatectomy for HCC without distant metastasis or any prior anticancer treatments between 2009 and 2013 at Zhongshan Hospital of Fudan University. The postoperative surveillance, adjuvant treatments and post-recurrent management were performed according to our uniform guidelines.[13] Tumour differentiation was determined by the Edmondson grading system, and liver cirrhosis was assessed by the Scheuer's system. Tumours were staged according to the International Union Against Cancer Tumour-Node-Metastasis (TNM) Classification (eighth edition) and the Barcelona Clinic Liver Cancer (BCLC) staging system. Among these slides, 99 were excluded for poor section quality and the remaining 2451 slides were captured and scanned for whole-slide images (WSIs). In addition to 260 WSIs of 96 patients randomly selected for manual annotation, all the 2191 WSIs of 1029 patients were randomly divided into a training set and a validation set (online supplementary table S1). Overall survival (OS) was defined as the time between surgery and death, cancer-specific survival (CSS) was defined as the time from surgery to death caused by HCC, and recurrence-free survival (RFS) was defined as the time between surgery and recurrence. Patients without recurrence or death were censored at the last follow-up. In the Zhongshan validation cohort, 184 patients died, 75.5% of whom had tumour recurrence; among the 344 living patients, 32.3% were diagnosed with tumour recurrence. The median duration of follow-up was 56 months (range, 14∼98 months). The 1-year, 3-year and 5-year overall survival rates were 89.7%, 74.1% and 68.2%, respectively, and the 1-year, 3-year and 5-year recurrence rates were 23.0%, 45.6% and 58.8%, respectively. Informed consent forms were signed by each patient.

We collected 376 HCC patients from the TCGA database via the Genomic Data Commons (https://gdc.cancer.gov/) with prognostic information and qualified pathological images as an independent cohort. We excluded cases with poor image quality (such as pen marks or poor staining) and whose OS was less than 1 month. Finally, this data set included 320 HCC patients for validation (online supplementary table S2). However, the patients of this cohort lacked recurrence information; thus, OS was the major index in survival analysis.

Slice preparation

All the specimens were fixed with 4% neutral formaldehyde, embedded in paraffin, sectioned at 4 μm thickness and stained with H&E. For CD34 immunostaining, the primary antibody was mouse anti-human CD34 (1:100 dilution; clone QBEND-10; Abcam) using the protocol as described previously.[14]

Sampling methods

We applied two sampling methods—global random sampling and tissue category-based local sampling. The background area was excluded based on the 40/256×tissue mask obtained by Otsu.[15] For global random sampling, we randomly selected 256×256 tiles in the tissue area of WSIs. For tissue category-based local sampling, we trained a tissue classification network, and then locally sampled tiles for different tissues based on the output of the network (online supplementary figure S1A, details in the online supplementary methods).

Classification network

We mainly aimed to identify tumour tissue, adjacent normal liver tissue, tumour-associated stroma and haemorrhage and necrosis regions. We modified the neural conditional random field (NCRF)[16] network for multi-classification, based on ResNet-18[17] and the conditional random field (CRF) structure (online supplementary figure S1A). ResNet-18 was used to extract features of tiles, and CRF was used to model the spatial correlation of tiles (details in the online supplementary methods).

Prognostic network

We took each patient as an example, with the survival time as the label and tiles sampled from the WSIs as the input. We separately trained the prognostic model with samples at different magnification scales as the input (online supplementary figure S1B, details in the online supplementary methods). After training, we sampled the tiles at different scales and inputted them into these prognostic networks to generate each patient's TRS. We then averaged the prediction results from different scale networks to obtain the final TRS.

Standardisation of WSIs in the TCGA HCC cohort

Due to the variability introduced during slide preparation and scanning, we could not directly apply the model trained by the Zhongshan cohort to the TCGA cohort to obtain satisfactory results. We performed enhanced CycleGAN[18] to convert the staining pattern of WSIs from the TCGA cohort to the Zhongshan cohort to standardise the staining conditions between the two data sets (details in the online supplementary methods).

Visualisation of prognosis-related features

To identify features related to prognosis, we modified the class activation mapping (CAM)[12] to the risk activation mapping (RAM) to display the heatmap on the tiles. On this basis, we carried out t-SNE and K-means algorithms to visualise and cluster the featured regions and quantified these potential features using a stacked histogram (details in the online supplementary methods).

Correlation between tumour immune infiltration and TRS

To systematically evaluate the correlation between TRS and immune infiltration in the TCGA HCC cohort, we applied the CIBERSORT algorithm and LM22 gene signature[19] for specific discrimination of 22 human immune phenotypes based on transcriptome data, including T cells, B cells, NK (natural killer) cells, macrophages, monocytes, dendritic cells and myeloid-derived suppressor cells. The CIBERSORT score was available in The Cancer Immunome Atlas (https://tcia.at/). In addition, T-Effector signature (CD8A, EOMES, PRF1, IFNG, and CD274) was computed using package GSVA.[20]

Statistical analysis

We used 0.5 as the classification threshold to calculate the accuracy rate. We calculated the C-index by risk scores and patient OS to verify performance. Cox regression analyses were

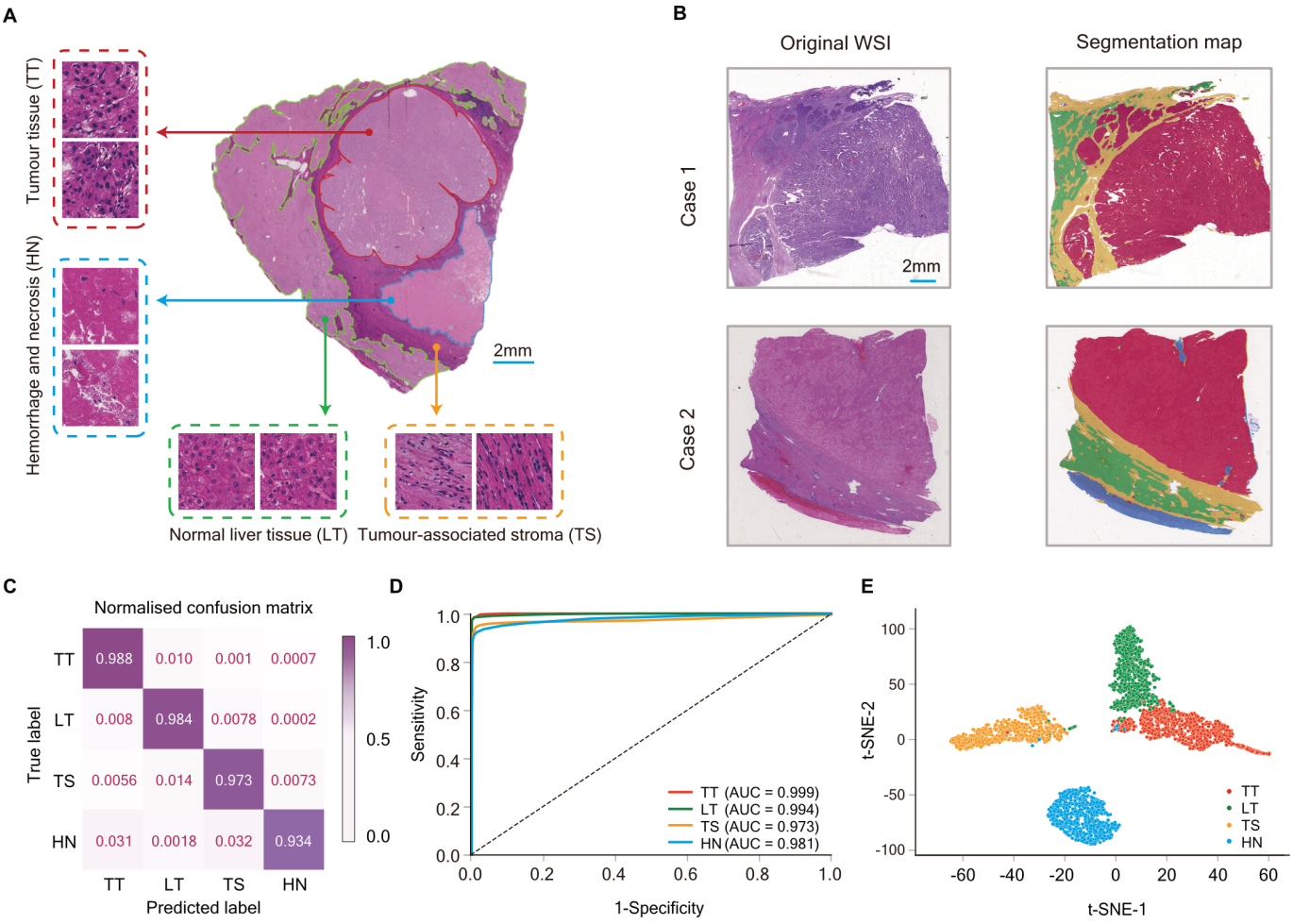

Figure 2 Visualisation of WSI classification for category-based sampling in the Zhongshan cohort. (A) Labelled WSI. We manually labelled WSIs to train the classification network (red for tumour tissue, green for normal liver tissue, yellow for tumour-associated stroma and blue for haemorrhage and necrosis) for category-based sampling. (B) Original WSI and segmentation maps of the classification network output. Left: original WSI images; Right: corresponding segmentation maps. (C) Confusion matrices of the classification results. (D) AUC of each tissue category of the classification network on the validation set. (E) Visualisation of the classification results using the t-SNE algorithm.

performed using the predicted risk scores for the validation set. The Mantel-Cox log-rank test was used for survival analysis. To compare two groups, the variables were evaluated by unpaired Student's t-test. Analysis of multivariate data was performed using multivariate analysis of variance with semiparametric designs. Correlation coefficients were computed by Spearman and distance correlation analyses. All the correlation heatmaps were generated using the pheatmap function (https://github.com/raivokolde/pheatmap). A two-sided p value less than 0.05 was considered statistically significant. Scikit-learn was used to calculate the area under the curve (AUC). SPSS 25.0 was used for survival analysis.

RESULTS

Training of tissue category-based local sampling within HCC tissue

We first built the category-based sampling by training a classification network of HCC tissue. Generally, HCC tissue comprises various tissue components, including tumour cells, adjacent normal liver tissues, tumour-associated stroma and haemorrhage and necrosis regions. Two experienced experts manually annotated the four types of tissue regions on 260 WSIs using polygons of different colours to draw the outline (figure 2A), and this step was independently checked by another pathologist.

Typical examples of the output of the classification network are shown in figure 2B, with different colours representing distinct tissue components. The number of tiles in each category to train the network is shown in online supplementary table S3. After full training, we tested the performance of the classification network (figure 2C), revealing an accuracy of 0.982 (table 1). The AUC of each tissue category of our classification network exceeded 0.973, with the highest in tumour tissue classification (AUC=0.999) (figure 2D). We then evaluated the discrimination of our classification network on different tissues using the

Table 1 Performance of the classification network in the Zhongshan cohort

Category	Sensitivity	Specificity	PPV	NPV	Accuracy
Normal liver tissue	0.984	0.989	0.982	0.990	0.982
Tumour tissue	0.988	0.992	0.988	0.992	
Tumour-associated stroma	0.973	0.995	0.981	0.993	
Haemorrhage and necrosis	0.934	0.998	0.862	0.999	

PPV, positive predictive value; NPV, negative predictive value;

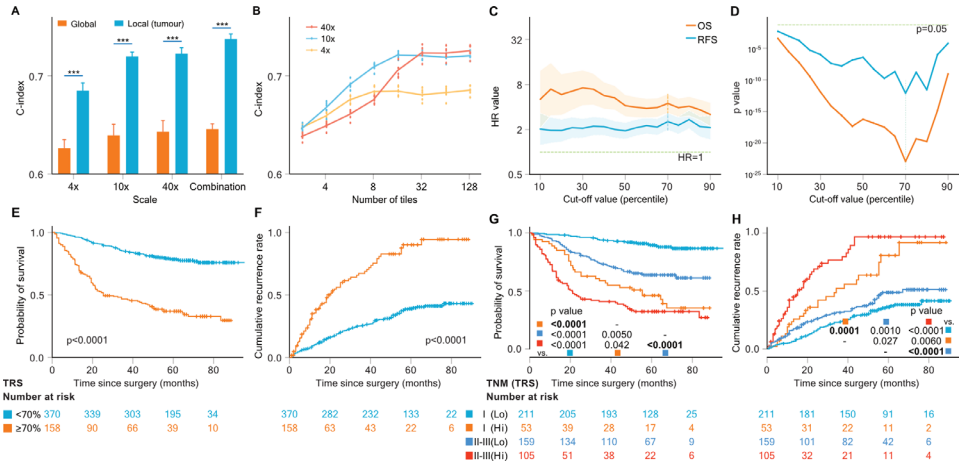

Figure 3 Evaluation of the prognostic performance between patients in the validation set and the Zhongshan cohort. (A) C-indices via risk scores according to different sampling methods under different magnification scales. (B) C-indices via risk scores according to different sampling counts under different magnification scales. (C and D) HR value and p value for different cut-off values of TRS. (E and F) Kaplan-Meier curves of survival and recurrence for high (Hi) TRS (>70% TRS) and low (Lo) TRS (≤70% TRS). (G and H) Kaplan-Meier curves for survival and recurrence of TRS (70% as cut-off) at different TNM stages. n.s., p>0.05; *, 0.05>p >0.01; **, 0.01>p >0.001; ***, p<0.001. OS, overall survival; RFS, recurrence-free survival; TNM, tumour-node-metastasis; TRS, tumour risk score.

t-SNE algorithm and found an obvious separation of each tissue component (figure 2E).

Next, we analysed the remaining 2191 WSIs using the classification network and confirmed the output by two experienced pathologists. Only approximately 2.51% (55/2191) of WSIs required manual re-annotation due to contamination, depigmentation and overlapping, indicating the comparable performance of our classification network to professional pathologists.

Prognostic network predicting patient outcome in HCC

To extract prognostic information from WSIs, we constructed a prognostic network with ResNet-50[17] and linear Cox regression. The patients in the Zhongshan cohort were randomly split into the training set (501 patients with 1221 WSIs) and validation set (528 patients with 970 WSIs).

We applied the sampled tiles as the inputs and the patient prognosis as the labels to train the prognostic network (online supplementary table S4). The output of the network that we called 'risk score' was a relative value to assess the prognostic risk of each patient. We first compared the performance of the global random sampling method (whole slide and irrespective of tissue category) with that of the tissue category-based local sampling method, in which 32 tiles were sampled for each patient as input. As shown in figure 3A, we computed the C-index via risk scores with different sampling methods under different magnification scales. The performance of local tumour sampling was significantly higher than that of global random sampling at every magnification scale (p<0.0001). The performance of the combination of the three magnification scales was always better than that of a single magnification scale (online supplementary figure S2A-C) because of the capture of various features at different magnifications. The median C-index of risk scores from multi-scale tumour tissues even reached 0.731. This result demonstrated our hypothesis that local tumour sampling as prior knowledge can significantly reduce the complexity of prognostic network optimisation.

Next, we tested the effect of different sample counts taken as input on network performance (figure 3B). As the number of samples increased, the network performance of different magnification scales gradually increased until a platform period appeared. As the magnification increased, more tiles were required to saturate the network performance. Specifically, at least 32 tiles are required for the 40×magnification scale, 16 tiles for the 10×magnification scale and 8 tiles for the 4×magnification scale. Only with a sufficient number of samples can the network extract the overall features of the tumour tissue possibly because of the well-known intratumour heterogeneity of HCC.[21]

Transcriptomic signatures of adjacent liver tissues and features of tumour-associated stroma in HCC contain valuable prognostic information.[22 23] Thus, we also sampled tiles from adjacent normal liver tissues and tumour-associated stroma regions at multi-scale as the network's input using the category-based local sampling method. The performance of tumour tissue was better than that of the other two types of tissue at the multi-scale combination (online supplementary figure S2D). The prognostic ability of the TRS alone was highly comparable to the combination of the three tissue components (p=0.405), implying that the features captured from tumour tissue played a dominant role in risk scoring.

We further performed survival analysis using TRS obtained from the prognostic network in the validation set by separating the patients into high-risk and low-risk groups. Strikingly, significant cut-offs showed a broad spectrum, ranging from 10% to 90% in terms OS and RFS, and the corresponding HR values were all >1 (figure 3C). Because the performance was the best at 70% (figure 3D), we used 70% as the cut-off for subsequent patient dichotomisation. The survival curves showed that patients with high TRS showed significantly worse survival (HR: 4.47, 95% CI: 3.34 to 5.99, p<0.0001 for OS in figure 3E; HR: 4.98, 95% CI: 3.66 to 6.77, p<0.0001 for CSS in online supplementary figure S3) and higher recurrence (HR: 2.55, 95% CI: 1.98 to 3.30, p<0.0001) than those with low TRS (figure 3F).

Based on TRS, HCC patients can be equally divided into three, four and even five groups with significantly different prognoses (online supplementary figure S4A-F), demonstrating robust and quantitative stratification of our prognostic risk network for predicting clinical outcome. Moreover, stratified analyses based on postoperative treatments showed that

Table 2 Univariable and multivariable analyses of factors associated with overall survival and recurrence-free survival for the Zhongshan cohort

	Overall survival					Recurrence-free survival				
	Univariable	Multivariable				Univariable	Multivariable			
Variable	P value	HR	95% CI		P value	P value	HR	95% CI		P value
Age (>60 vs ≤60 years)	0.35	–	–	–	–	0.99	–	–	–	–
Gender (female vs male)	0.15	–	–	–	–	0.035	1.36	0.9	2.02	0.15
HBV (positive vs negative)	0.53	–	–	–	–	0.24				
HCV (positive vs negative)	0.87	–	–	–	–	0.10	–	–	–	–
Alpha fetoprotein (>20 vs ≤20 ng/mL)	<0.0001	1.11	0.79	1.57	0.54	0.033	1.06	0.81	1.39	0.68
Liver cirrhosis (S4e-4 vs S0-3)	0.31	–	–	–	–	0.13	–	–	–	–
Tumour number (multiple vs single)	<0.0001	2.44	1.63	3.65	<0.0001	<0.0001	2.14	1.44	3.17	<0.0001
Tumour size (>5 vs ≤5)	<0.0001	1.70	1.21	2.38	0.002	<0.0001	1.36	1.03	1.79	0.032
Tumour encapsulation (no vs yes)	0.001	1.54	1.14	2.06	0.004	0.12	–	–	–	–
Differentiation (III-IV vs I-II)	0.0002	1.21	0.89	1.64	0.23	0.13	–	–	–	–
Vascular invasion (yes vs no)	<0.0001	2.53	1.46	4.40	0.001	<0.0001	1.70	1.–6	2.71	0.027
TNM (II-III vs I)	<0.0001	1.59	0.86	2.96	0.14	<0.0001	0.78	0.46	1.34	0.37
BCLC (B-C vs 0-A)	<0.0001	0.35	0.22	0.57	<0.0001	<0.0001	1.08	0.70	1.73	0.76
TRS (high vs low)	<0.0001	3.32	2.36	4.69	<0.0001	<0.0001	2.05	1.54	2.73	<0.0001

HBV, hepatitis B virus; HCV, hepatitis C virus; TNM, tumour-node-metastasis; BCLC, Barcelona clinic liver cancer; TRS, tumour risk score.;

high TRS impaired patient OS (online supplementary figure S5A) and RFS (online supplementary figure S5B), regardless of different postoperative management strategies. Multivariable analysis revealed that the association of high TRS with unfavourable OS (HR: 3.32, 95% CI: 2.36 to 4.69, p<0.0001) or RFS (HR: 2.05, 95% CI: 1.54 to 2.73, p<0.0001) was independent of other clinicopathological characteristics (table 2). Notably, TRS remained a predictor of patient survival in all subgroups of each patient characteristic (figure 4). For example, after stratifying patients according to TNM stage (figure 3G,H) or BCLC stage (online supplementary figure S5C,D), TRS remained strongly correlated with patient prognosis irrespective of early or advanced stages, supporting the superior predictive ability of our prognostic risk network.

To compare the performance of TRS to a combination of all clinicopathological variables, we integrated the clinicopathological characteristics (p<0.05 in univariate analysis) into a combined score. The prognostic power of TRS was stronger for OS (p<0.001, online supplementary figure S6A) but suboptimal for RFS (p=0.043, online supplementary figure S6B) compared with the combined clinical score. As expected, the integration of TRS with the combined clinical score had the strongest prognostic power for both survival (p<0.001) and recurrence (p<0.001, online supplementary figure S6A,B).

Prognostic performance in the TCGA HCC cohort

We next validated our proposed framework on the TCGA HCC data set (320 patients with 320 WSIs). We first standardised the TCGA data set through enhanced CycleGAN (online supplementary figure S7) and then tested the networks without fine-tuning. We randomly selected 29 WSIs for annotation to test the performance of the classification network. Most of the pathological haematoxylin and eosin (H&E) images from TCGA only contained tumour tissue; thus, we only annotated tumours. The accuracy of the classification network reached 0.921 (figure 5A), and the median C-index of the prognostic network reached 0.713 (figure 5B). We also tested the TCGA cohort without WSI standardisation and found that the accuracy rate of the classification network was 0.782 and the C-index of the prognosis network was 0.602 (figure 5B). These results showed that the variation caused by staining and the scanner in different data sets could be alleviated by WSI standardisation.

The patients in the high-risk group showed significantly inferior survival compared with those in the low-risk group (HR: 3.49, 95% CI: 2.41 to 5.06, p<0.0001; Figure 5C), using 70% as the cut-off. Multivariable analysis also revealed that TRS was an independent prognosticator for patient survival (HR: 2.64, 95% CI: 1.49 to 5.12, p=0.0003; online supplementary table S5). TRS was still correlated with patient survival when stratified by tumour stage (figure 5D). These results confirmed that our proposed deep learning framework is robust for HCC prognostic stratification and superior to conventional clinical staging systems.

Additionally, exploiting the TCGA HCC data, we investigated the relevance of TRS to the tumour immune microenvironment and gene mutations (figure 5E). We found that TRS was positively correlated with the infiltration of neutrophils (p=0.0086, r=0.148), M0 macrophages (p=0.024, r=0.128) and regulatory T cells (p=0.029, r=0.123), but negatively correlated with the infiltration of CD4+ memory T cells (p=0.025, r=−0.126) and T-Effector signature (p<0.0001, r=−0.226). Additionally, TRS showed positive correlations with *FAT3* (p=0.0007, OR=0.193), *RYR2* (p=0.0092, OR=0.148) and *TTN* (p=0.031, OR=0.123) mutations.

Visualisation of histological features associated with prognosis risk

To open the black box of deep learning approaches, we used RAM to display the histological features associated with patient outcomes on WSIs. To accurately describe fine-grained features, we removed the last three residual blocks of the feature extraction module (ResNet-50). Finally, heatmaps were obtained by weighting the activation feature maps of the last layer generated in the inference process with featured weights, where redder and bluer indicated higher and lower risks of death (or recurrence), respectively. Our visualisation method requires only one inference to obtain a heatmap without calculating gradients, indicating higher efficiency and universality.

	Event/patients (%)		HR (95% CI)	log-rank p value
	Low-risk	High-risk		
Alpha fetoprotein (ng/mL)				
≤20	30/173 (17.3%)	23/35 (65.7%)	7.10 (4.10 to 12.30)	<0.0001
>20	52/196 (26.5%)	79/124 (63.7%)	3.40 (2.39 to 4.83)	<0.0001
Tumour number				
Single	51/298 (17.1%)	73/122 (59.8%)	5.19 (3.62 to 7.44)	<0.0001
Multiple	31/72 (43.1%)	29/36 (80.6%)	3.27 (1.95 to 5.48)	<0.0001
Tumour size (cm)				
≤ 5	47/265 (17.7%)	31/52 (59.6%)	4.96 (3.14 to 7.83)	<0.0001
> 5	35/105 (33.3%)	71/106 (70.0%)	2.93 (1.95 to 4.40)	<0.0001
Tumour encapsulation				
No	37/140 (26.4%)	55/78 (70.5%)	4.32 (2.83 to 6.59)	<0.0001
Yes	45/230 (19.6%)	47/80 (58.8%)	4.28 (2.84 to 6.45)	<0.0001
Tumour Differentiation				
I-II	45/239 (18.8%)	45/72 (62.5%)	4.86 (3.20 to 7.37)	<0.0001
III-IV	37/131 (28.2%)	57/86 (66.3%)	3.66 (2.42 to 5.56)	<0.0001
Vascular invasion				
No	40/249 (16.1%)	36/61 (63.9%)	5.31 (3.38 to 8.35)	<0.0001
Yes	42/121 (34.7%)	66/97 (68.0%)	2.97 (2.01 to 4.39)	<0.0001
TNM				
I	24/211 (11.4%)	31/53 (58.5%)	7.47 (4.37 to 12.77)	<0.0001
II-III	58/159 (36.5%)	71/105 (67.6%)	2.82 (1.99 to 4.00)	<0.0001
BCLC				
0-A	40/241 (16.6%)	41/62 (66.1%)	6.26 (4.04 to 9.71)	<0.0001
B-C	42/129 (23.3%)	61/96 (63.5%)	2.80 (1.89 to 4.16)	<0.0001

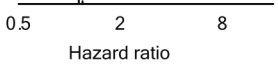

Figure 4 Forest plot of tumour risk score for the Zhongshan cohort in overall survival analysis.

By visualising and clustering the featured regions in heatmaps, we detected four separated features at the 40×scale (figure 6A,B), and two features at the 10×scale and 4×scale (online supplementary figure S8A-C). Next, our pathologists reviewed and proposed the possible annotations on the featured regions. In the heatmaps at the 40×scale, some sinusoidal areas were labelled in red, and immunostaining revealed that these sinusoids were outlined by CD34+ cells (figure 6C), suggesting sinusoidal capillarisation in these areas. Likewise, prominent nucleoli and karyotheca cells and a high nucleus/cytoplasm (N/C) ratio were also displayed in red. Thus, the above three features were probably associated with the increased risks of recurrence and death. By contrast, the heatmaps specifically highlighted the regions of inflammatory cells in blue, indicating a low risk of recurrence and death. These inflammatory cells tend to be smaller and rounder with dark and homogeneous staining compared with tumour cells. Additionally, sinusoidal capillarisation and inflammatory cells were identified at the 10×scale and 4×scale. Importantly, most (>60%) sinusoidal capillarisation, prominent nucleoli and karyotheca cells and high N/C ratio cells appeared in patients with high TRS, while most (>60%) infiltrated inflammatory cells occurred in patients with low TRS ($p<0.001$) (figure 6D and online supplemental figure 8D). The C-index for the combination of the four features reached 0.724 (figure 6E), indicating the deconstruction of the TRS predicted by our model.

Together, the heatmaps implied that the histological features of tumour angiogenesis, cell morphology and inflammatory microenvironment might correlate with patient outcome, which could be captured by our prognostic network.

DISCUSSION

Although genetic, transcriptomic and proteomic profiles are increasingly attempted to predict tumour aggressiveness, tumour histology remains essential in prognostic stratification. Several recent studies have preliminarily demonstrated that deep learning-based architecture could contribute to predicting patient prognosis from digital pathological images.[24–30] Herein, we developed an interpretable prognosis network based on weakly supervised learning incorporating prior knowledge. It can visualise the prognostic-related features at a fine-grained level and has surprising accuracy in prognostic prediction.

Some authors[9 30] have reported the method of only using clinical diagnostic records as labels for training neural networks to avoid expensive and time-consuming pixel-wise manual annotations. However, they need at least 10 000 cases to obtain good performance, a requirement that is time-consuming and costly.

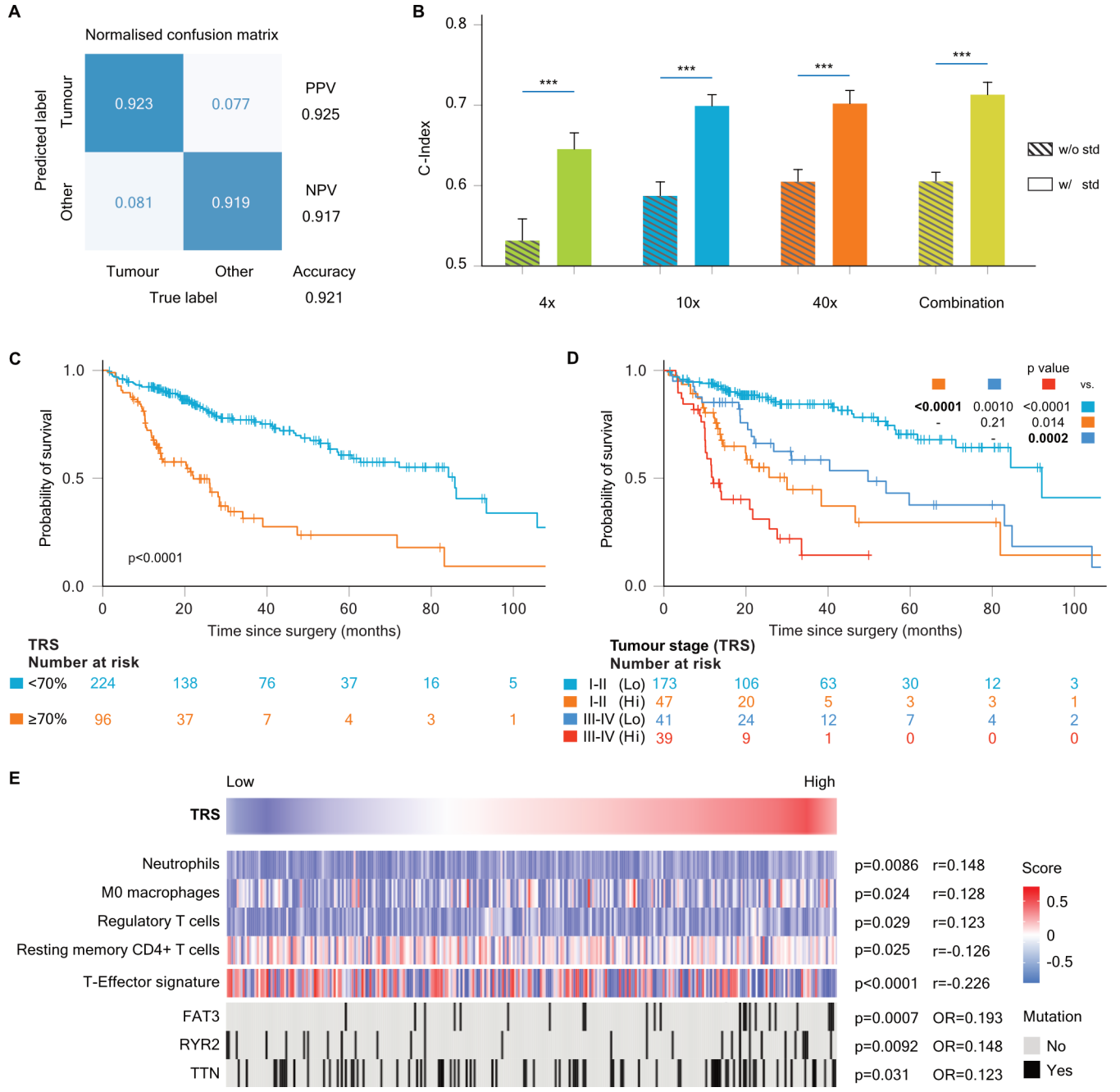

Figure 5 Evaluation of the prognostic results for patients in the validation set from the The Cancer Genome Atlas (TCGA) cohort. (A) Confusion matrices of the classification results (PPV, positive predictive value; NPV, negative predictive value). (B) C-indices via TRS under different magnification scales with (w/) or without (w/o) WSI standardisation (std). (C) Kaplan-Meier curves of overall survival for high TRS (>70% TRS) and low TRS (≤70% TRS) (D) Kaplan-Meier curves for overall survival of TRS (70% as cut-off) at different tumour stages. (E) Relevance of TRS to tumour immune infiltration and gene mutations. Hi, high; Lo, low; TRS, tumour risk score.

By contrast, we successfully applied weakly supervised learning on smaller data sets effectively. Training a deep neural network in image analysis is an optimisation of all possible solutions to tremendous image information. By implicitly constraining the solution space with prior knowledge, the complexity of the problem can be effectively reduced to justify the need for less data. In our study, an easily available prior knowledge for pathological images is the tissue category. Thus, we trained tissue category-based local sampling using minor manual annotation, thereby obtaining a large number of more meaningful tissue tiles on HCC slices as inputs. Using prior knowledge can effectively reduce the complexity of network optimisation and improve the performance of weakly supervised learning, even on small data sets.

Recently, Saillard *et al* has reported a method of predicting postoperative survival of HCC patients using deep-learning on histological slides.[29] Compared with their study, we have advantages on multi-scale sampling, multi-omics analysis and fine-grained heatmap visualisation. Using tissue decomposition, we performed multi-scale automatic sampling from specific tissue regions to calibrate prognostic risk scores. Compared with non-tumour tissue regions, the TRS had the greatest prognostic

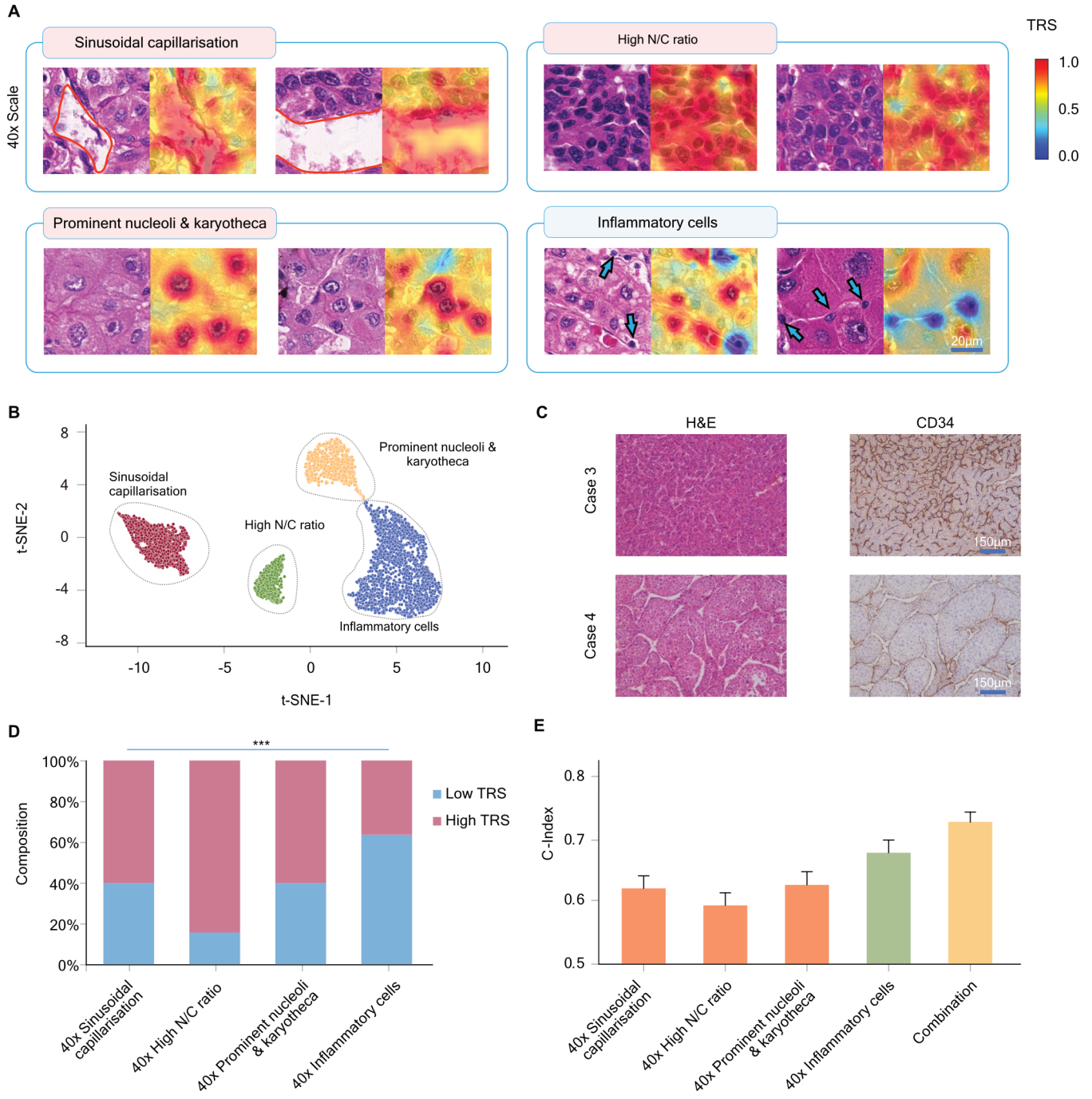

Figure 6 Visualisation and clustering of the risk heatmaps. (A) Visualisation of the heatmaps of high-risk (sinusoidal capillarisation, high N/C ratio and prominent nucleoli and karyotheca) and low-risk features (inflammatory cells) related to TRS at the 40×scale. Pathologists reviewed and defined each feature. On the left of each panel is the original tile of WSIs; on the right is the heatmap. The red area indicates potentially increased risk, and the blue area indicates reduced risk. Red curve circled indicates sinusoidal capillarisation and blue arrows indicate inflammatory cells. (B) The four potential features extracted from heatmaps at the 40×scale were separated by the t-SNE algorithm. (C) Immunostaining revealed that the sinusoids in the tumour were outlined by CD34+ cells, suggesting sinusoidal capillarisation in these areas. (D) The distribution of each potential feature was quantified using a stacked histogram in patients grouped by TRS. The data were compared using multivariate analysis of variance. (E) C-index for each potential feature and the combination of these features. ***, $p < 0.001$. N/C,nucleus/cytoplasm; TRS, tumour risk score; WSIs, whole-slide images.

power in HCC (C-index, 0.731) with an extended range of significant cut-offs (10%~90%) and independently correlated with patient outcome. Clinically, our prognostic risk network could be used to automatically detect HCC tumour regions and calculate TRS for each patient. Notably, TRS predicted the survival of HCC patients stratified for all the conventional clinicopathological characteristics, indicating its value in identifying HCC patients at higher risk of rapid progression, even in an early-stage subpopulation.

Using TCGA HCC data, we validated our prognostic risk model and further revealed the relevance of TRS to gene mutations. In particular, TRS was correlated with mutations of *FAT3* and *RYR2*, which are associated with tumour progression.[31 32] Nevertheless, the findings need to be prospectively verified in a

larger cohort. Previous literature[24 33] has reported that combining genomics and pathological imaging might further improve prognostic accuracy. In the future, the integration of genomic, immunological and histological data into a deep learning-based prediction framework would be a promising direction to provide a more objective, multidimensional and functionally relevant prognostic output.

It is challenging for doctors to extract new prognostic features directly from pathological images. Herein, we propose a quantitative and interpretable method to simplify this process. Combined with visualising and clustering of the fine-grained heatmaps, the doctors just need to define each potential feature discovered by our method. We found that the sinusoidal capillarisation, prominent nucleoli and karyotheca, high N/C ratio and immune response were powerful pathological features affecting survival. In addition to tumour cell intrinsic features, evidence has been accumulating that the composition, localisation and nature of immune cells in HCC could alter tumour biological behaviour and affect disease progression.[14 34] Indeed, we observed an anti-tumour effect of these infiltrated immune cells, which may be T cells, such as CD4+ memory T cells or effector T cells, based on the TCGA data. Additionally, angiogenesis gradually occurs by the microvascularisation of endothelial cells, which is associated with HCC invasiveness[35] and consistent with our findings. Similarly, Saillard *et al* identified macrotrabecular architectural pattern, vascular spaces, cytological atypia and nuclear hyperchomasia as high-risk features, while immune cell infiltration as low-risk feature.[29] The consistency of the two studies perfectly illustrated that tissue structure, cellular morphology and immune response were the three essential features affecting patient prognosis. Presently, combinational antiangiogenics and immunotherapies represent the main avenues in the treatment of advanced HCC;[36] thus, TRS may be an important marker to predict and evaluate the effect of these combination therapies.

In conclusion, this study proposed a weakly supervised deep-learning framework to facilitate patient prognosis in HCC. The prognostic features discovered by our deep-learning framework were formularised by TRS and visualised by heatmaps. We confirmed that tumour immune infiltration might favour and microvascularisation may compromise HCC patient survival. Our work indicates that weakly supervised deep learning is an effective and labour-saving method for predicting patient clinical outcomes and warrants further study and extensive application.

Author affiliations
[1]Department of Liver Surgery and Transplantation, Liver Cancer Institute, Zhongshan Hospital, and Key Laboratory of Carcinogenesis and Cancer Invasion (Ministry of Education), Fudan University, Shanghai, P. R. China
[2]School of Computer Science and Technology, Xidian University, Xi'an, P. R. China
[3]School of Computer Science, Northwestern Polytechnical University, Xi'an, P. R. China
[4]Department of Pathology, Zhongshan Hospital Fudan University, Shanghai, P. R. China
[5]Department of General Surgery, Zhongshan Hospital (South), Public Health Clinical Centre, Fudan University, Shanghai, P. R. China
[6]Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, P. R. China
[7]Cancer Research Institute, Xiangya School of Medicine, Central South University, Hunan, P. R. China
[8]Institute of Biomedical Sciences, Fudan University, Shanghai, P. R. China
[9]State Key Laboratory of Genetic Engineering at Fudan University, Shanghai, P. R. China

Contributors QG and XL conceived and directed the project. JS and GD collected the original slides and manually labelled whole-slide images, and JH and GY checked the annotation. XW led the deep learning algorithm development and evaluation. ZD, ZG, YZ, LZ and HY wrote the code for different tasks. LM collected and analysed the TCGA data. GY helped to scan all the hepatocellular carcinoma slides. XW, ZD and AK provided clinical guidance. HY, LW and LA contributed to the analysis of the data. JZ, YC and JF provided strategic guidance. JS and XW wrote the manuscript with the assistance and feedback of the other co-authors.

Funding This work was supported by the National Natural Science Foundation of China (No. 91859105, 8196112802, 81872321 and 81802302), National Key R&D Program of China (2018YFC0116500), Basic Research Project from Technology Commission of Shanghai Municipality (No. 17JC1402200), Shanghai Municipal Key Clinical Specialty and Program of Shanghai Academic Research Leader (19XD1420700).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval Ethical approval was obtained from the Zhongshan Hospital Research Ethics Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement The data from Zhongshan Hospital that support the findings of this study are available upon reasonable request from the corresponding author (QG). The data from Zhongshan Hospital are not publicly available, because they contain protected patient privacy information. The external validation of TCGA data set is publicly available at the TCGA portal (https://portal.gdc.cancer.gov). We provide a manifest linking to the sample IDs considered in the study (at https://github.com/wangxiaodong1021/HCC_Prognostic). We also provided annotated files of TCGA tumour regions (at https://github.com/wangxiaodong1021/HCC_Prognostic). Code availability: All code related to this method was written in Python. Custom code related to the image extraction, preprocessing pipeline, deep-learning model builder, data provider and experimenter driver were available (at https://github.com/wangxiaodong1021/HCC_Prognostic).

ORCID iDs
Jia Fan http://orcid.org/0000-0001-6386-1068
Qiang Gao http://orcid.org/0000-0002-9947-7268

REFERENCES

1 Bray F, Ferlay J, Soerjomataram I, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68:394–424.
2 Jiang Y, Sun A, Zhao Y, *et al*. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* 2019;567:257–61.
3 Gao Q, Zhu H, Dong L, *et al*. Integrated Proteogenomic characterization of HBV-related hepatocellular carcinoma. *Cell* 2019;179:561–77.
4 Calderaro J, Petitprez F, Becht E, *et al*. Intra-tumoral tertiary lymphoid structures are associated with a low risk of early recurrence of hepatocellular carcinoma. *J Hepatol* 2019;70:58–65.
5 Renne SL, Woo HY, Allegra S, *et al*. Vessels encapsulating tumor clusters (VETC) is a powerful predictor of aggressive hepatocellular carcinoma. *Hepatology* 2020;71:183–95.
6 Calderaro J, Couchy G, Imbeaud S, *et al*. Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification. *J Hepatol* 2017;67:727–38.
7 Calderaro J, Ziol M, Paradis V, *et al*. Molecular and histological correlations in liver cancer. *J Hepatol* 2019;71:616–30.
8 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199–210.
9 Campanella G, Hanna MG, Geneslaw L, *et al*. Clinical-Grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019;25:1301–9.
10 Kather JN, Krisam J, Charoentong P, *et al*. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med* 2019;16:e1002730.
11 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-cam: visual explanations from deep networks via gradient-based localization. *Proc IEEE Inter Conf Comput Vis* 2017.
12 Zhou B, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization. *Proc IEEE Conf Comput Vis Pattern Recognit* 2016.
13 Zhou J, Sun H-C, Wang Z, *et al*. Guidelines for diagnosis and treatment of primary liver cancer in China (2017 edition). *Liver Cancer* 2018;7:235–60.
14 Shi J-Y, Gao Q, Wang Z-C, *et al*. Margin-infiltrating CD20(+) B cells display an atypical memory phenotype and correlate with favorable prognosis in hepatocellular carcinoma. *Clin Cancer Res* 2013;19:5994–6005.
15 Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 1979;9:62–6.

16 Li Y, Ping W. Cancer metastasis detection with neural conditional random field. *arXiv* 2018.
17 He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis Pattern Recognit* 2016.
18 Zhou N, Cai D, Han X, *et al*. Enhanced Cycle-Consistent Generative Adversarial Network for Color Normalization of H&E Stained Images. *Med Image Comput Comput Assist Interv* 2019;11764:694–702.
19 Newman AM, Liu CL, Green MR, *et al*. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;12:453–7.
20 McDermott DF, Huseni MA, Atkins MB, *et al*. Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nat Med* 2018;24:749–57.
21 Gao Q, Wang Z-C, Duan M, *et al*. Cell Culture System for Analysis of Genetic Heterogeneity Within Hepatocellular Carcinomas and Response to Pharmacologic Agents. *Gastroenterology* 2017;152:232–42.
22 Hoshida Y, Villanueva A, Kobayashi M, *et al*. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *N Engl J Med* 2008;359:1995–2004.
23 Gao Q, Wang X-Y, Qiu S-J, *et al*. Tumor stroma reaction-related gene signature predicts clinical outcome in human hepatocellular carcinoma. *Cancer Sci* 2011;102:1522–31.
24 Mobadersany P, Yousefi S, Amgad M, *et al*. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci U S A* 2018;115:E2970–9.
25 N Kalimuthu S, Wilson GW, Grant RC, *et al*. Morphological classification of pancreatic ductal adenocarcinoma that predicts molecular subtypes and correlates with clinical outcome. *Gut* 2020;69:317–28.
26 Kleppe A, Albregtsen F, Vlatkovic L, *et al*. Chromatin organisation and cancer prognosis: a pan-cancer study. *Lancet Oncol* 2018;19:356–69.
27 Reichling C, Taieb J, Derangere V, *et al*. Artificial intelligence-guided tissue analysis combined with immune infiltrate assessment predicts stage III colon cancer outcomes in PETACC08 study. *Gut* 2020;69:681–90.
28 Liao H, Xiong T, Peng J, *et al*. Classification and prognosis prediction from histopathological images of hepatocellular carcinoma by a fully automated pipeline based on machine learning. *Ann Surg Oncol* 2020;27:2359–69.
29 Saillard C, Schmauch B, Laifa O, *et al*. Predicting survival after hepatocellular carcinoma resection using deep-learning on histological slides. *Hepatology* 2020. doi:10.1002/hep.31207. [Epub ahead of print: 28 Feb 2020].
30 Courtiol P, Maussion C, Moarii M, *et al*. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat Med* 2019;25:1519–25.
31 Katoh M. Function and cancer genomics of fat family genes (review). *Int J Oncol* 2012;41:1913–8.
32 Liang JQ, Teoh N, Xu L, *et al*. Dietary cholesterol promotes steatohepatitis related hepatocellular carcinoma through dysregulated metabolism and calcium signaling. *Nat Commun* 2018;9:4490.
33 Chaudhary K, Poirion OB, Lu L, *et al*. Deep Learning-Based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;24:1248–59.
34 Ringelhan M, Pfister D, O'Connor T, *et al*. The immunology of hepatocellular carcinoma. *Nat Immunol* 2018;19:222–32.
35 Dong Z-R, Sun D, Yang Y-F, *et al*. Tmprss4 drives angiogenesis in hepatocellular carcinoma by promoting HB-EGF expression and proteolytic cleavage. *Hepatology* 2019. doi:10.1002/hep.31076. [Epub ahead of print: 22 Dec 2019].
36 Finn RS, Qin S, Ikeda M, *et al*. Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma. *N Engl J Med* 2020;382:1894–905.