Method of Tumor Pathological Micronecrosis Quantification Via Deep Learning From Label Fuzzy Proportions

Qiancheng Ye, Qi Zhang, Yu Tian, Tianshu Zhou, Hongbin Ge, Jiajun Wu, Na Lu, Xueli Bai, Tingbo Liang, and Jingsong Li

***Abstract*—The presence of necrosis is associated with tumor progression and patient outcomes in many cancers, but existing analyses rarely adopt quantitative methods because the manual quantification of histopathological features is too expensive. We aim to accurately identify necrotic regions on hematoxylin and eosin (HE)–stained slides and to calculate the ratio of necrosis with minimal annotations on the images. An adaptive method named Learning from Label Fuzzy Proportions (LLFP) was introduced to histopathological image analysis. Two datasets of liver cancer HE slides were collected to verify the feasibility of the method by training on the internal set using cross validation and performing validation on the external set, along with ensemble learning to improve performance. The models from cross validation performed relatively stably in identifying necrosis, with a Concordance Index of the Slide Necrosis Score (CISNS) of 0.9165±0.0089 in the internal test set. The integration model improved the CISNS to 0.9341 and achieved a CISNS of 0.8278 on the external set. There were significant differences in survival ($p = 0.0060$) between the three groups divided according to the calculated necrosis ratio. The proposed method can build an integration model good at distinguishing necrosis and capable of clinical assistance as an automatic tool to stratify patients with different risks or as a cluster tool for the quantification of histopathological features. We presented a method effective for identifying histopathological features and suggested that the extent of necrosis, especially micronecrosis, in liver cancer is related to patient outcomes.**

***Index Terms*—Deep learning, ensemble learning, histopathological image analysis, learning from label fuzzy proportions, tumor micronecrosis quantification.**

Manuscript received September 27, 2020; revised January 25, 2021 and March 20, 2021; accepted March 30, 2021. Date of publication April 6, 2021; date of current version September 3, 2021. This work was financially supported in part by the Major Scientific Project of Zhejiang Lab under Grant 2020ND8AD01, in part by the National Natural Science Foundation of China under Grants 81771936, 81801796, 81830089, 81871320, in part by the National Key Research and Development Program of China under Grant 2018YFC0116901, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR20H160002, and in part by the Key Research Development Program of Zhejiang Province under Grant 2020C03117. *(Qiancheng Ye, Qi Zhang, and Yu Tian contributed equally to this work.) (Corresponding authors: Jingsong Li; Tingbo Liang and Xueli Bai.)*

Qiancheng Ye, Yu Tian, and Tianshu Zhou are with the Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, China (e-mail: yqc_5823595@126.com; ty.1987823@163.com; zts@zju.edu.cn).

Qi Zhang, Hongbin Ge, Jiajun Wu, Na Lu, Xueli Bai, and Tingbo Liang are with the Department of Hepatobiliary and Pancreatic Surgery, the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China, and with the Key Laboratory of Pancreatic Disease of Zhejiang Province, Hangzhou 310003, China, and also with the Innovation Center for the Study of Pancreatic Diseases of Zhejiang Province, Hangzhou 310003, China, and also with the Zhejiang Clinical Research Center of Hepatobiliary and Pancreatic Diseases, Hangzhou 310003, China (e-mail: qi.zhang@zju.edu.cn; 3120101751@zju.edu.cn; 21918349@zju.edu.cn; 3150103137@zju.edu.cn; shirleybai@zju.edu.cn; liangtingbo@zju.edu.cn).

Jingsong Li is with the Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, China, and also with Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou 311100, China (e-mail: ljs@zju.edu.cn).

Digital Object Identifier 10.1109/JBHI.2021.3071276

I. Introduction

WHEN necrosis is observed in surgical samples from cancer patients without any preoperative therapy, it is usually caused by the lack of blood flow (containing oxygen and nutrition within) in the center of the neoplasm, indicating the rapid progression of the tumor. It has been reported in many cancers, such as colorectal carcinoma [1]–[3], pancreatic ductal carcinoma [4], non-small-cell lung cancer [5], renal cell carcinoma [6]–[11], urothelial carcinoma [12], glioblastoma multiforme [13], soft tissue sarcomas [14], breast cancer [15], epithelioid pleural mesothelioma [16], and medullary thyroid carcinoma [17], that primary pathological necrosis can be a good prognostic indicator, whereas in other types, such as liver cancer, the relation between necrosis and prognosis is controversial [18], [19]. Existing studies were qualitative or semiquantitative, resulting in inconsistent conclusions [9] about the relation between necrosis and prognosis. Most of these studies focused on massive necrosis and neglected micronecrosis, which appears as small necrotic foci and is thus easy to overlook and hard to quantify. In addition, as part of the tumor microenvironment, necrosis is considered to be associated with tumor progression [20], [21], metastases [22], immune response [23], chemoresistance [24] and gene expression levels [25] in certain cancers. It can also be used to evaluate the effect of neoadjuvant therapy [26].

Based on the abovementioned possible clinical applications and biological significance, the evaluation of necrosis should be included in routine pathological examinations [1], [4], [7].

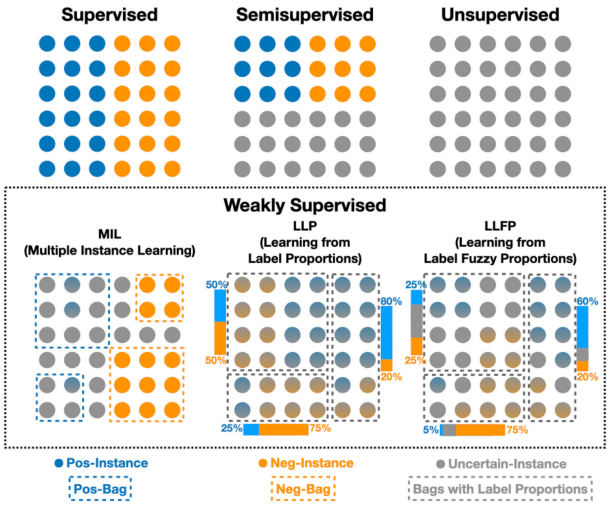

Fig. 1. Different learning frameworks in binary settings. Dots of blue, orange, or gray denote the positive, negative, and unknown samples, respectively. Rectangles of blue, orange, or gray denote the bags with global labels of positive, negative, or proportional. Dots of gray shades in blue or orange are the potential positive or negative instances in the bag.

However, in reality, due to the heavy workload of pathologists, pathological reports hardly cover necrosis and contain either miss-reported or vague descriptions. To assess the degree of necrosis in tumor samples, manual scoring is required. As the size and information presented on histopathological images are enormous, the task is time consuming, and it is difficult to pin down a precise ratio of necrosis on one slide. Additionally, intraindividual variance and interindividual variance are almost inevitable in the manual assessment of histopathological slides, leading to poor reproducibility [27]–[32]. It is important to identify necrosis accurately and calculate its proportion by automatic means.

In recent years, digital and computational pathology has rapidly progressed with the help of scanning equipment, computing servers, and deep learning algorithms. Studies on whole-slide images (WSIs) regarding cancer diagnosis [33]–[36], subtype classification [37], [38], and prognosis stratification [39]–[43] have obtained great results and have the potential to assist clinicians in providing better health care. By using supervised machine learning methods such as random forest, support vector machine, and convolutional neural networks (CNN), human tumor necrosis can be identified with an accuracy of 80% to 94.67% at a scale of dozens of slides [44]–[47], but none of these studies provided external validation results, leaving the problem of overfitting unchecked. Acute hepatic necrosis over microscopic images could be classified by a CNN with an accuracy of 99.33% [48]. Nevertheless, there is no such investigation that both quantifies micronecrosis and explores the relation between necrosis and patient outcomes in a quantitative manner.

The calculation of the necrosis ratio can be considered as a matter of image identification in the area of computer vision. It usually exploits deep neural networks (DNNs) within a supervised learning (SL) framework (as in Fig. 1) to accomplish feature extraction and image classification. The shortcoming of SL is that it requires a large amount of labeled images, which is quite a burden to experts in medicine. Especially in a histopathological scenario with many billion-pixel WSIs, it is too expensive to provide enough labeled image patches to accomplish excellent performance of DNNs, let alone pixel-level annotations such as the delineation of different tissues. Compared with SL, weakly supervised learning (WSL) can relatively reduce the labeling cost and maintain or even improve performance and robustness at the same time [34], [49]. A basic WSL framework called multiple instance learning (MIL) takes a bag of instances with only bag labels but no instance labels to accomplish the task. In terms of WSIs, only a global label of the image (bag) is required, and the definition of a region or the class of a certain patch (instance) is unclear for the model. In a binary setting, an image in which none of the patches are positive can be assigned to the negative group, and if at least one patch in the image is positive, it will be defined as a positive image. With the help of negative bags, the features of negative instances in them can be captured by DNNs, and negative instances in positive bags can be eliminated, leaving positive instances and completing the learning process. However, because of the common presence of necrosis, more or less, on a tumor slide, it is not easy to obtain absolutely negative slides, namely, no necrosis within, leading to wrongful classifications. As another WSL framework, learning from label proportions (LLP) [50]–[52], requires the ratios of instances belonging to each class in a bag, which adds restrictions to the learning convergence. By realizing the proportions fitting of the classes in bags, LLP accomplishes the mission of image classification. Unfortunately, a precise proportion label on a billion-pixel WSI may require as much labor as SL, thus being almost unrealistic. Fuzzy learning can exploit the fuzzy feature representation to reduce data uncertainty in the task of brain MRI segmentation [53] or assign fuzzy labels to the training sample and has shown good performance in person reidentification [54], but it is still unexplored in the field of computational pathology.

In this paper, we proposed a deep-learning-based method to identify and quantify histopathological features such as necrosis on WSIs and proved the feasibility of the method. The main contributions of this work are summarized as follows:

1) A novel method named Deep Learning from Label Fuzzy Proportions (DLLFP) is proposed to accomplish the mission of histopathological quantification at a minimal cost for image labeling.
2) A large-scale cohort of primary liver cancer patients was recruited from the First Affiliated Hospital, Zhejiang University School of Medicine (FAH-ZJUMS). This cohort consisted of 1070 patients, and their hematoxylin and eosin (HE)–stained slides after surgical resection were used for cross validation to verify the validity of the proposed method in identifying necrosis in liver cancer. The proposed method was adapted to the existing necrosis scoring rule in FAH-ZJUMS, so no additional labeling needed to be done.
3) An integration method that mimics the consultations of pathologists and is based on the ideology of ensemble learning was adopted as postprocessing to determine the final ratio of necrosis on the slide by the agreement of

multiple models, realizing a more stable identification of tumor necrosis.

4) An external dataset from The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) data collection was used to validate the robustness of the proposed method and the models trained on the private FAH-ZJUMS dataset. Clinical follow-up data from TCGA-LIHC were used to assess the relation between necrosis and prognosis in liver cancer.

II. Materials & Methods

A. Methodology

1) Learning from Label Fuzzy Proportions (LLFP): In brief, LLFP allows the label proportion to be a range of ratios rather than a precise ratio as in LLP, reducing the requirement of label precision and the workload of labeling while maintaining constraint conditions to a certain extent.

Consider that the whole dataset consists of n bags

$$B_i = \left\{ x_i^1,\ \ldots,\ x_i^{N_i} \right\},\ i \in \{1,\ 2,\ \ldots,\ n\}. \tag{1}$$

The proportion labels of each class in the ith bag compose a vector $\boldsymbol{R_i}$ in the form of

$$\boldsymbol{R_i} = \left[R_i^1,\ \ldots,\ R_i^k \ldots, R_i^C \right]^T, \tag{2}$$

where C is the class number of the task. Since R_i^k is a fuzzy proportion, which means range, $R_{i\,L}^k$ represents the minimal ratio of an instance belonging to the *k*th class in B_i, then $\sum_{k=1}^{C} R_{i\,L}^k \leq 1$. The maximal ratio shall be $R_{i\,H}^k$ and $R_{i\,H}^k \leq 1 - \sum_{c \neq k} R_{i\,L}^c$.

The selection of the training sets should be

$$S_i^k = \left\{ r_1^k, \ldots, r_j^k, \ldots, r_{R_{i\,L}^k * N_i}^k \right\}, \tag{3}$$

in which r_j^k stands for the ranking instance of B_i by the *k*th class probability, with the highest instance as r_1^k and the lowest one as $r_{N_i}^k$.

Cross-entropy loss was used as the loss function of training, defined as

$$-\frac{1}{S_i} \sum_{j=1}^{S_i} \sum_{k=1}^{C} \omega^k l_j^k \log\left(p_j^k\right). \tag{4}$$

S_i is the number of labeled samples in the ith bag. ω^k is the weight assigned to the *k*th class. $p = \{p^k \mid p^k \in [01],\ k = 1, \ldots, C,\ \sum_{k=1}^{C} p_j^k = 1\}$ denotes the output probability of instance r_j^k. $l = \{l^k \mid l^k \in \{01\},\ k = 1, \ldots, C,\ \sum_{k=1}^{C} l_j^k = 1\}$ denotes the true label of the corresponding instance.

With minimization of the loss, the proportion of the *k*th class in B_i will be inside the proportion range, that is,

$$\frac{1}{N_i} \sum_{j=1}^{N_i} p_j^k \in \left[R_{i\,L}^k, R_{i\,H}^k \right]. \tag{5}$$

In the binary-class classification settings, given that the label proportion of positive instances in a bag of instance size N ranges from R_{L} to R_{H}, the minimal number of positive instances in the bag should be N $\times$ R_{L}, and the minimal number of negative instances in the bag should be N $\times$ (1 - R_{H}), leaving N $\times$ (R_{H} - R_{L}) instances in the bag to be of uncertain class and thus out of the learning process. In the multiple-class classification settings, with the label proportions of each class, the minimal number of instances of every class can be calculated and utilized similar to that with the binary settings. When there are hierarchical relations between the classes, the minimal number of instances belonging to the classes inside the hierarchy can be derived based on the multiplicative product of the proportions by series. Let R_{FL} and R_{FH} be the fuzzy proportions of the father concept class F, whereas P_{SL} and P_{SH} are the fuzzy proportions of the son concept class S. The minimal number of class S would be N $\times$ R_{FL} $\times$ R_{SL}, and the minimal number of class F would be N $\times$ R_{FL} $\times$ (1 - R_{SH}). After the minimal instance number of a certain class is calculated, a corresponding number of instances can be selected by ranking the probability during the training process. Unlike MIL, which takes the most likely tiles to train, or the fully supervised method, which trains tiles equally, the proposed fuzzy proportional learning could focus more on the easily misclassified tiles.

2) Preprocessing (Tissue Segmentation and Tiling): Using the difference between the RGB channels of the slide thumbnail image, Otsu's method [55] can clarify the threshold to discard blank background and remove edge artifacts and pen marks of different colors.

After foreground tissue is segmented, sliding windows are used to cut WSIs into tiles with a size of 224 pixels $\times$ 224 pixels at different magnification levels (5x, 10x, and 20x). To ensure that models trained at different magnification levels have the number of training tiles at the same magnitude, the overlap ratio of the sliding window, defined as the overlap area ratio of two directly adjacent tiles, at each magnification is different: 0 at 20x, 50% at 10x and 75% at 5x.

3) Postprocessing (Model Ensemble): For a histopathological proportion fitting problem, a model ensemble can be achieved at two levels, one at the tile level and one at the slide level. The tile-level ensemble takes the predicted probabilities of the models on one tile and makes the final prediction of the tile with a certain operator, then aggregates the tile predictions on the slides into target class proportions. The slide-level ensemble directly makes target class proportions by integrating the slide proportion calculated by each model with a certain operator. The operator mentioned above can be as simple as average-pooling or as complicated as random forest, recurrent neural networks, etc.

B. Datasets

1) FAH-ZJUMS: The Research Ethics Committee of the First Affiliated Hospital of Zhejiang University School of Medicine approved this study (No. 2018-115). A total of 1070 primary liver cancer patients and their corresponding 5181 slides (referred to as the Full Set hereinafter) were finally included in this study, with an average slide number per case of 4.82. The majority of the patients were diagnosed with hepatocellular carcinoma (HCC), while there were few patients with intrahepatic

cholangiocarcinoma and mixed cell carcinoma. The slides were scanned at 20x magnification. Except for 3 slides with failed focus that were excluded, there were no other slide curation criteria. After preprocessing, a total of 96 million tiles of liver tissue were created, reducing the calculative cost by approximately 80% from 470 million tiles without foreground tissue segmentation.

2) TCGA-LIHC: From the TCGA portal, the public official download tool was used to collect data from the LIHC project [56], including diagnostic histopathological slides, corresponding pathology reports, clinical data including follow-ups, etc. A total of 364 HCC cases with 372 slides (referred to as the External Set hereinafter) were used as an external validation dataset and did not participate in any of the training processes. These cases originated from 34 medical centers around the USA. A total of 335 patients had survival information.

3) Manual Scoring Label: Slide-level necrosis scoring was performed on both the Full Set and External Set by two pathologists independently. All slides were assessed based on the ratio of the tumor necrosis area to the tumor area and then classified into 4 grades, corresponding to 0-5%, 5-20%, 20-50%, and 50-100%. Since adjacent normal tissues were not included in the calculation of the necrosis score, part of the Full Set (1566 slides from 308 cases, referred to as the Label Complete Set hereinafter) was given a slide-level tumor score to help DNNs distinguish nontumor areas from tumor areas. The tumor score was defined as the ratio of the tumor area to the whole foreground tissue area, and it also has 4 grades, corresponding to 0%, 0-33%, 33-67%, and 67-100%. The final score was based on the agreement of the group, and the slide score was reviewed if the score did not match. The average scoring time for a slide is around 2 min.

C. Experimental Design

The overall experimental workflow is shown in Fig. 2.

1) Hardware and Software: All experiments were conducted on a computing server containing one Tesla V100 DGXS 32G graphics processing unit (GPU), an Intel Xeon E5-2683 v4 central processing unit (CPU), 512 GB memory and 12 TB SSD storage.

All codes were written in Python (version 3.7.3), with Openslide [57] (version 1.1.1) to preprocess the WSIs, PyTorch [58] (version 1.3.1) to build the models, scikit-learn [59] (version 0.21.2) to complete the evaluation metrics, scipy (version 1.5.2) to conduct statistical tests, lifelines (version 0.23.4) to perform survival analysis, and matplotlib [60] (version 3.1.3) and seaborn (version 0.9.0) to generate plots.

ResNet [61] was selected for the image classification task based on previous investigations in the field of computational pathology because it has the best performance and a reasonable training cost [34], [62]. Transfer learning was adopted using the pretrained ResNet-34 on ImageNet to build the classifier. Cross-entropy was used as the loss function, and minimization of the loss was achieved via stochastic gradient descent (SGD) using the Adam optimizer and a learning rate of 0.0001. The batch size was set to 256. Early stopping was used to avoid overfitting. ω^k in Equation (4) were all set as 1.

2) Experimental Goal: The purpose of the experiment was to calculate the proportion of tumor necrosis, defined as the ratio of the tumor necrosis area to the tumor area, on HE slides. Generally, resected tumors are accompanied by noncancerous tissue adjacent to cancerous tissue, and along with hepatectomy, other organs near the liver, such as the pancreas, gallbladder, and spleen, might be partly resected to check carcinoma invasion. These two types of tissues, defined as cancer-adjacent and liver-adjacent, shall be distinguished and left out of the calculation of the necrosis proportion. Hence, the models to build will address a 4-class classification problem: (C_0) liver-adjacent; (C_1) cancer-adjacent; (C_2) liver cancer; and (C_3) liver necrosis.

If a WSI mainly contains liver-adjacent tissue, 80% of the tiles from the image will be used as training tiles for C_0. For the remaining classes, the number of training tiles for each class is derived from the necrosis score and tumor score of the slide. For example, when a WSI has a tumor score of 2 (33-67% of the foreground tissue is tumor tissue) and a necrosis score of 1 (5-20% necrosis in tumor tissue), the image should contain at least 33%, 26.4% (33% $\times$ 80%), and 1.65% (33% $\times$ 5%) of tiles belonging to C_1, C_2 and C_3, respectively. For slides, outside of the Label Complete Set, without the tumor score, the ratio of the tumor is calculated by the classifier and determined adaptively in a semisupervised way.

With the tiles in the slide classified into each class, the necrosis proportion can be computed as $\frac{N_{C3}}{N_{C2}+N_{C3}}$. For some slides with small areas of liver cancer, the highest slide-level necrosis proportion was limited to 5%, and the maximal necrosis score was adjusted to 1. The definitions of small areas of liver cancer on slides were as follows: 1) area of cancerous tissue in the slide is smaller than 25 mm^2 (5mm$*$5mm, $N_{C2} + N_{C3} < 2000$); 2) area of foreground tissue in the slide is smaller than 63 mm^2 ($N_{tiles} < 5000$); 3) slide ruled as cancer-adjacent using the threshold determined by the training set as the smallest area of cancerous tissue among all cancerous slides; and 4) slide ruled as liver-adjacent using the threshold determined by the training set as the smallest area of liver-adjacent tissue among all liver-adjacent slides.

3) Pretrain Test: To determine the optimal magnification for DNNs to distinguish different kinds of tissues, three 5-fold cross validation experiments on the Label Complete Set were performed at the levels of 5x, 10x, and 20x, the best of which was chosen as the magnification for the formal model building later.

Another 5-fold cross validation at the optimal magnification was conducted, resulting in 10 premodels. All tiles can be grouped into 11 clusters by the times of being defined as necrosis by these models, with 10 being most likely to be necrosis and 0 being the least. Approximately 3000 tiles from the slides in the Label Complete Set were sampled for necrosis labeling, and whether the tile region is necrotic was determined to the extent of 5 grades, corresponding to None (0%), Few (0-50%), Some (50-100%), All (100%) and Suspect (cannot be labeled based on only the field of view around the tile). The average labeling time for a tile is 4 s.

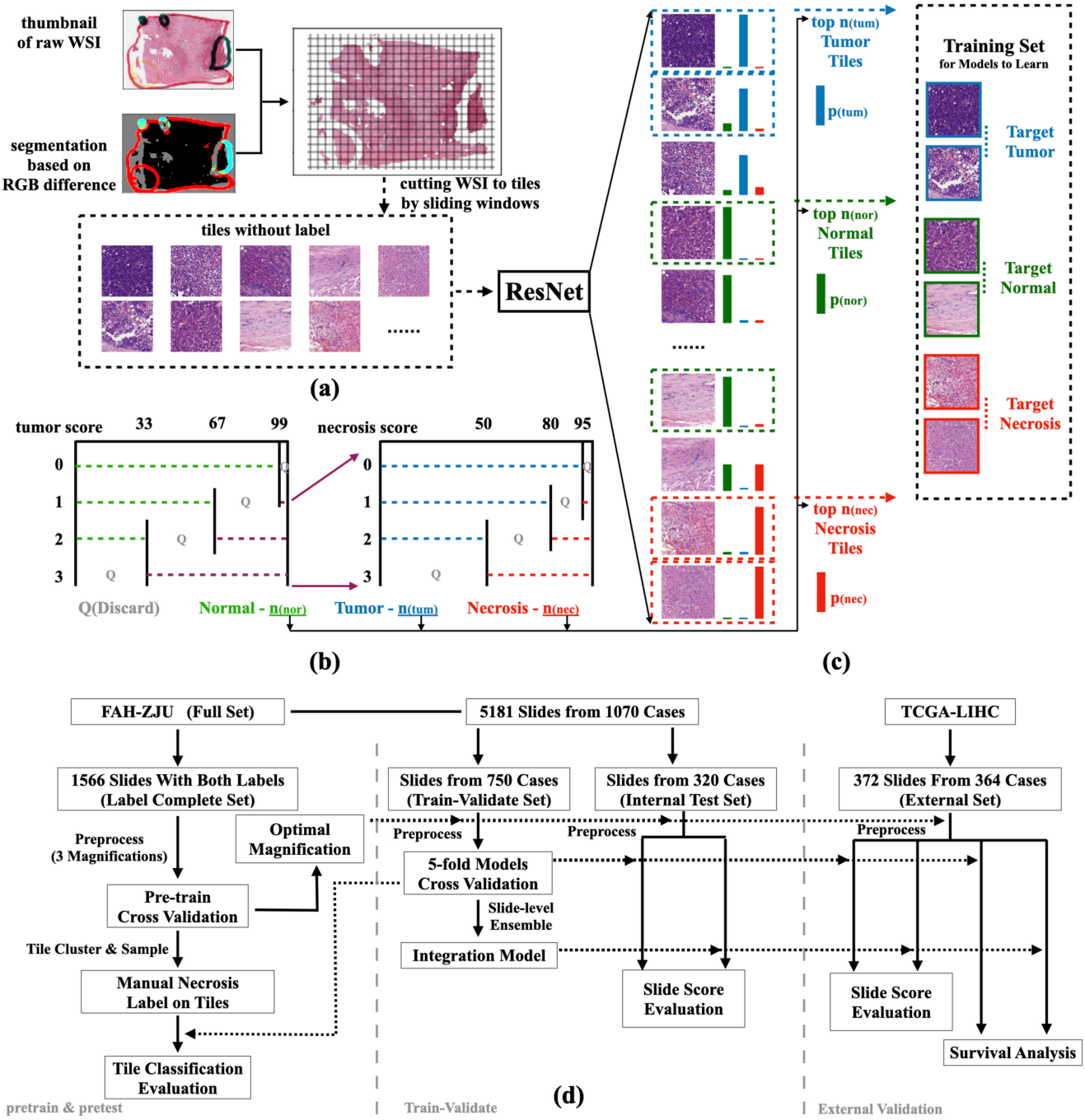

Fig. 2. Schematic diagram of the experiments. (a) Preprocessing of the experiment on HE slides with RGB segmentation and window sliding to produce tiles for ResNet inference. (b) Calculation of the minimal number of tiles belonging to each class based on the manual tumor score and necrosis score with corresponding proportions. (c) Selection of training targets in tiles from (a) according to number of tiles in each class calculated from (b) and ranking probabilities. (d) Workflow of the experiment, mainly composed of three parts: pretraining and pretesting to obtain the optimal magnification and prepare the tile classification evaluation (left of the dividing line), training and testing the integration model (middle between the dividing lines), and externally validating the model and exploring the relation between necrosis and survival (right of the dividing line).

4) Model Building and Integration: Formal model building divided the Full Set into a train-validate set consisting of 750 cases and an internal test set consisting of 320 cases. 5-fold cross validation was performed on the training-validation set, and the 5 best models identifying and quantifying necrosis in the validation process were tested on the internal test set and the External Set. In the cross validations, since some of the WSIs had no tumor score labels, we adopted a semisupervised approach to train the models. First, slides with both a tumor score label and necrosis score label were utilized to pretrain the model for 12 epochs, after which the loss function became stable. Thereafter, the tumor ratios of the WSIs with no tumor score labels could be calculated during every training iteration before the tile ranking and the selection of the training set. The

threshold of the probability of a tile belonging to a class was 0.5 in the process of determining the tumor ratio.

The integration of the 5 models from cross validation adopted the slide-level ensemble, which removed the highest and the lowest necrosis ratios calculated by the 5 models and made the average of the remaining three the final proportion.

5) Evaluation Metrics: There are three evaluation metrics at the slide level. The first one is called Match, meaning the matching ratio of the 4 grades of manual necrosis scoring and the necrosis ratio calculated by the model binned into 4 grades using 5%, 20%, and 50% as thresholds. However, the manual necrosis score is not a precise quantitative calculation but a vague estimation of the degree of necrosis, so the second metric, Concordance Index of the Slide Necrosis Score (CISNS), is introduced to indicate the consistency between the calculated ratio and the manual score. CISNS represents the proportion of correctly calculated ratio ranking pairs in all pairs with comparable scores; that is, when the necrosis score of a slide is higher than the other, the calculated necrosis ratio of the former by the model should be higher as well. In addition, in practical applications, clinicians pay more attention to whether the model can distinguish between rarely having micronecrosis (manual score of 0, ratio of necrosis less than 5%) and definitely having necrosis (manual score from 1 to 3, ratio of necrosis more than 5%). Therefore, the area under the receiver operating characteristic curve (AUROC) of the binary classification of the slide necrosis score is introduced as the third metric.

The evaluation metric at the tile level is classification accuracy. However, due to the massive number of tiles at a million level, it is completely impractical to gain full labels to calculate the accuracy. Therefore, label sampling is used to estimate the tile-level accuracy. Using the above pretrained models to cluster the sampled label set as a test set, each cluster in the set is upsampled to the number of the original cluster to obtain the estimated actual label distribution, and then 11 clusters of distributions are summed to obtain the overall label estimate. With the overall estimated tile labels, the corresponding test results can be inferred by the model, and the estimated accuracy can be calculated.

6) Baseline Settings: For comparison with the proposed method, several experiments were conducted as baselines. The basic MIL (bMIL) utilized binary labels of tumors (positive or negative) and necrosis (greater than 5% or otherwise). The proportional variation in MIL (pMIL) replaced the binary label of tumors in bMIL with the proportional label of tumors. The supervised variation in bMIL (sMIL) replaced the binary label of necrosis in bMIL with the supervised label of necrosis obtained in the process of necrosis tile labeling. The supervised variation in pMIL (spMIL) took advantage of the proportional label of tumors and the supervised label of necrosis.

III. Results

A. Pretrain with Cross Validation

The models from cross validation conducted at 5x achieved the best validation performance compared with the other two magnifications, with a mean CISNS of 0.9010±0.0151 (shown in Table I). Hence, 5x was defined as the optimal magnification to identify necrosis.

TABLE I
Comparison of Model Performance at Different Magnifications

CISNS	5x	10x	20x
k1	0.8986	0.8457	0.8311
k2	0.8773	0.8126	0.7905
k3	0.9129	0.9124	0.8929
k4	0.9010	0.8445	0.8386
k5	0.9150	0.8638	0.8486
mean	**0.9010 (±0.0151)**	**0.8558 (±0.0366)**	**0.8403 (±0.0368)**

k1 to k5 denote the 1st to 5th model derived from the 5-fold cross validation

The overall trend between the tile-level necrosis label and the times ruled as necrosis was relatively consistent (shown in Table II), as all of the tiles sampled from Times-0 are labeled as None and the majority of the tiles sampled from Times-8 to Times-10 are not labeled as None. However, the false discovery rate is quite high considering that a certain amount of tiles labeled as None were identified as necrotic by more than half of the models.

B. Slide-Level Necrosis Score Proportion Fitting Performance

In general, all 5 models from cross validation performed well in identifying necrosis and calculating the ratio in the internal test set, with a CISNS of 0.9165±0.0089, a Match of 0.7573±0.0264, and an AUROC of 0.9159±0.0096 (shown in Table III).

With the slide-level ensemble, the integration model achieved a CISNS of 0.9341, a Match of 0.7560, and an AUROC of 0.9336; the consistency with manual scoring and binary classification performance were better than those of the models from each fold. When the cutoff value of the necrosis ratio was set at 0.0606 to divide the slides according to a binary classification of a manual score of 0 and higher than 0, the accuracy was the highest, with a sensitivity of 0.8406 and a specificity of 0.8928.

From the distribution of the calculated necrosis ratio by the integration model (shown in Fig. 3a), we can see some outliers of higher calculation for Score-0 and Score-1 and of lower calculation for Score-1 and Score-2. Except for the defect of the model's recognition ability itself, the causes of deviation included the following: 1) inaccuracy of manual scoring, especially rather high overall scoring leading to a high false positive rate; 2) slides calculated too high on the necrosis ratio are more likely to be cholangiocarcinoma or mixed cell carcinoma with larger areas of fibrous regions, which could be easily mistaken as necrotic.

With the help of the integration model, some wrongly scored slides by manual evaluation can be identified and reviewed again to be more precise. Some examples of slide necrosis score amendments are shown in Fig. 3(c-f). Fig. 3(c) shows a slide with scattering micronecrosis, whereas Fig. 3(d) shows a slide with

TABLE II
DISTRIBUTIONS OF THE TILE LABELS ACROSS CLUSTERS

Times	# of Tiles	# of Samples	None	Suspect	Few	Some	All
0	1214732	196	196	0	0	0	0
1	110670	237	230	0	6	1	0
2	33662	311	289	2	15	3	2
3	16681	359	328	7	15	6	3
4	9968	296	266	3	20	6	1
5	6587	227	178	9	25	9	6
6	5291	225	169	4	36	9	7
7	4711	195	113	5	41	20	16
8	5011	182	77	8	42	31	24
9	6257	189	85	4	35	38	27
10	19272	692	160	25	76	211	220
Total	1432842	3109	2091	67	311	334	306

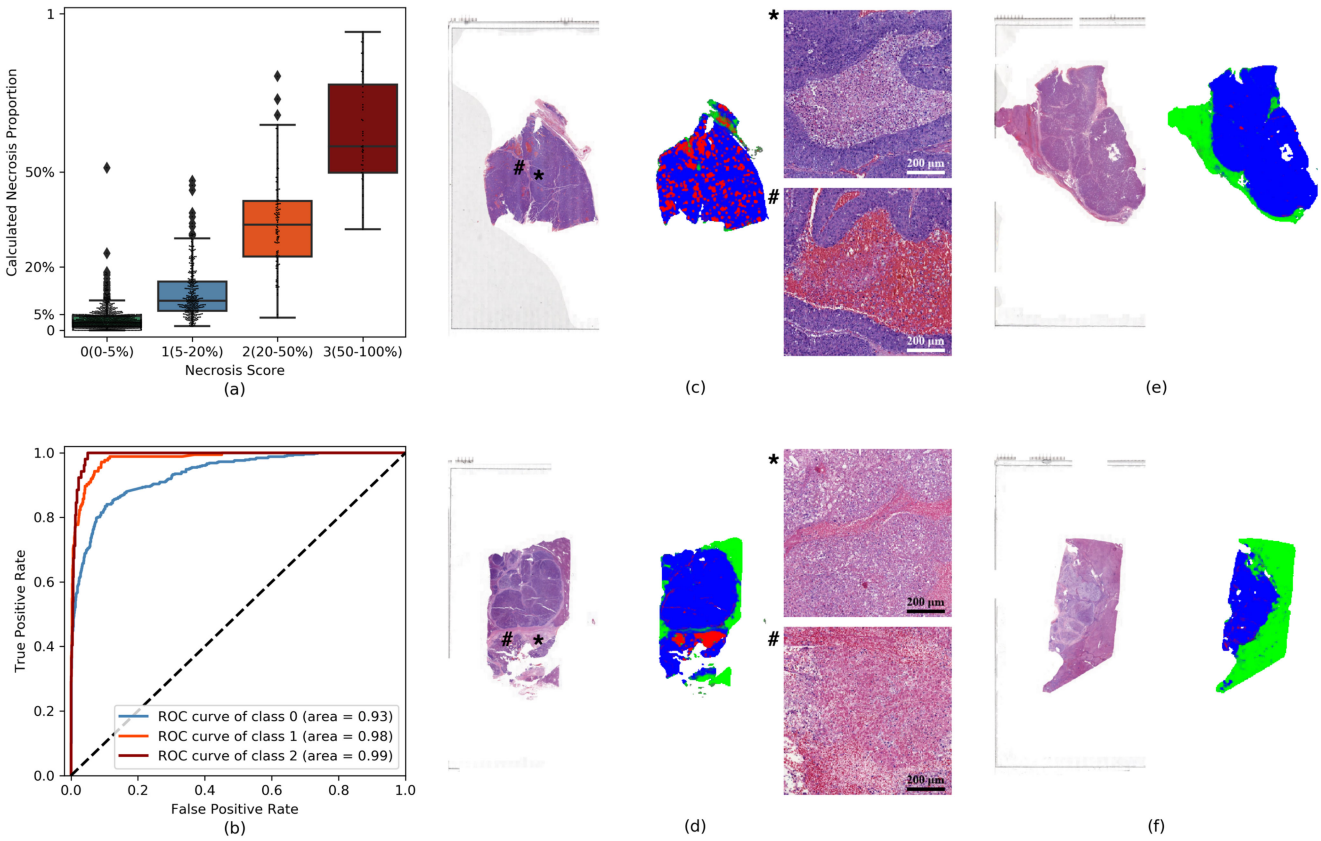

Fig. 3. Test results of the integration model in the internal set. (a) Distribution of the necrosis ratio with manual scoring shows high consistency. (b) ROC curves for three binary classification tasks show great performance in distinguishing slides with various degrees of necrosis. (c-f) Examples of amendment to manual necrosis evaluation. Thumbnails are on the left with inference heatmaps adjacent on the right, on which green represents normal tissue, blue represents cancer tissue, and red represents necrotic tissue. Slides with an original necrosis score of 0 but a calculated necrosis ratio higher than 5% (c, d) were reviewed to have micronecrosis and amended as a necrosis score of 1. Possible missed necrotic regions are marked with an asterisk or hash mark on the left and presented on the right of (c, d). Slides with an original necrosis score of 1 but a calculated necrosis ratio lower than 5% (e, f) were reviewed and amended as a necrosis score of 0.

necrosis adjacent to the tumor capsule, both overlooked by manual evaluation but captured by the machine. Fig. 3(e, f) shows two slides with little necrosis that were labeled as a necrosis score of 1.

C. Tile-Level Classification Accuracy

The most ideal fitting results can be achieved when the recalls of the None-class, Few-class, Some-class, and All-class tiles are 0, some value between 0 and 0.5, some value between 0.5 and 1, and 1, respectively. Since the goal is to calculate the ratio of

TABLE III
SLIDE-LEVEL NECROSIS PROPORTION FITTING PERFORMANCE

kth-fold	CISNS	Match	AUROC
1	0.9298	0.7802	0.9296
2	0.9110	0.7581	0.9114
3	0.9215	0.7655	0.9216
4	0.9113	0.7124	0.9111
5	0.9089	0.7702	0.9058
mean	**0.9165 (±0.0089)**	**0.7573 (±0.0264)**	**0.9159 (±0.0096)**
Integration	**0.9341**	**0.7560**	**0.9336**

the area, the accuracies of None-class (absolutely negative) and All-class (absolutely positive) are more important than the other two. It is acceptable that models produce an average estimated false positive rate of 0.0270±0.0093 for None-class tiles and an average estimated false negative rate of 0.0573±0.0127 for All-class tiles (shown in Table IV), indicating that the excellent performance of slide-level necrosis score fitting came from the accurate identification of necrosis rather than coincidence. It is normal for the recalls of the Few-class and Some-class tiles to be lower than those of the All-class tiles. If tiles in the Few-class and Some-class are recalled at the level of All-class (approximately 0.95), the calculated necrosis proportion will become too large, causing the fitting performance of the slide score to decrease. False negative tiles were relatively untypical, and false positive tiles were mainly summarized as three types: hemorrhage, fibrosis, and tumor secretions (shown in Fig. 4).

D. External Test on TCGA

The average CISNS of the models tested on the external set was 0.7615±0.0603, and the integration model achieved a CISNS of 0.8278 (shown in Table V). Although the performance decreased compared to the performance on the internal set, the integration model shows great robustness in computing the necrosis ratio, considering that the data were strongly heterogeneous from multiple medical centers due to different protocols of sample preparation and scanning equipment.

Based on the calculated necrosis ratio, a data-driven approach was used to determine groups of patients with various degrees of survival risks. Three of the five models from cross validation can divide patients into groups with significant overall survival (OS) differences ($p<0.05$, shown in Table V). When patients were divided into three groups by the integration model using necrosis ratio thresholds of 0.01 and 0.54, a significant difference ($p = 0.0060$, Fig. 5c) in OS was observed. However, there were no survival differences between the groups differentiated by manual scoring ($p = 0.4$, Fig. 5a) or by the calculated necrosis ratio with manual thresholds of 0.05 and 0.2 ($p = 0.6$, Fig. 5b).

E. Comparisons with Relating Methods

Some comparative methods were conducted, and the results are shown in Fig. 6 and Table VI, along with the estimated time costs of labeling. With the near-minimal time costs of annotation,

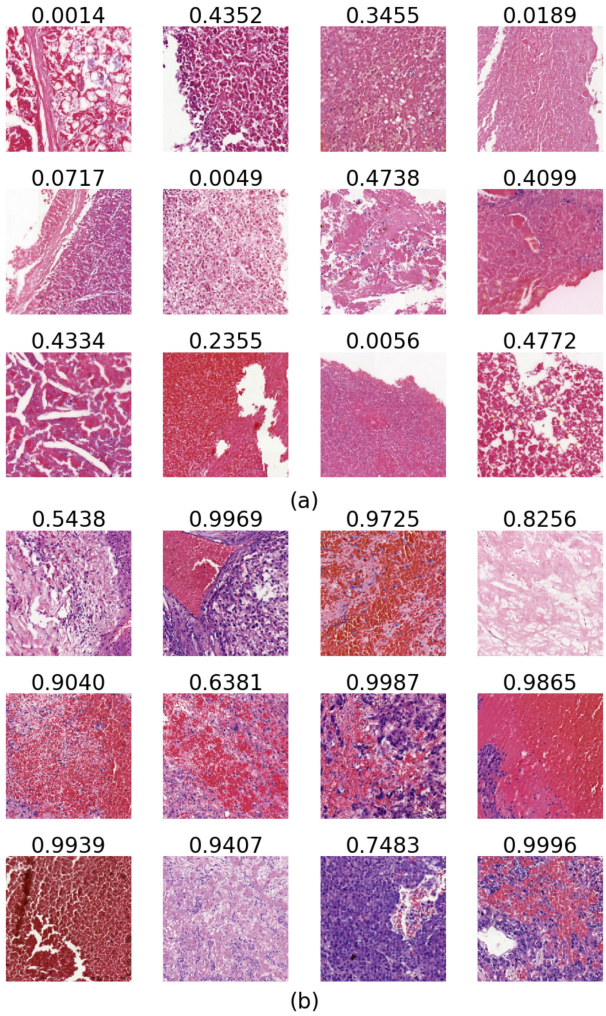

Fig. 4. Examples of wrongly classified tiles. (a) Sampled false negative tiles, untypical necrosis with poor quality. (b) Sampled false positive tiles, mainly hemorrhage, fibrosis, or secretion. All tiles are at 5x magnification with 224 pixels on the width and height, and the probabilities of being necrosis calculated are marked above.

the proposed LLFP achieved comparable or superior accuracy to existing methods in both the internal and external test sets. Moreover, the methods containing the part of proportional learning (LLFP, spMIL and pMIL) showed significant advantages in external validation (Fig. 6(d)).

IV. DISCUSSION

The method proposed in this paper called DLLFP excels in identifying histopathological features on HE slides. The results of the experiment show that DLLFP can realize the accurate identification of different tissues and the calculation of the necrosis proportion in liver cancer at minimal labeling costs while maintaining the constraint of the learning process. Furthermore, the model ensemble was used to avoid identification errors produced by partial models in cross validation and improve performance. Regarding the tasks of identification and the quantitative calculation of other pathological features (fibrosis, steatosis) that can distinguish subtypes of clinical manifestations, the training

TABLE IV
ESTIMATED TILE-LEVEL RECALLS

kth-fold	None (FP)	Suspect	Few	Some	All (TP)
1	0.0257	0.7198	0.7158	0.8514	0.9350
2	0.0285	0.6203	0.5777	0.8607	0.9466
3	0.0167	0.7411	0.6393	0.8278	0.9578
4	0.0417	0.6104	0.7341	0.8418	0.9252
5	0.0223	0.7144	0.5272	0.8298	0.9488
mean	**0.0270 (±0. 0093)**	**0.6812 (±0. 0610)**	**0.6388 (±0. 0883)**	**0.8423 (±0. 0140)**	**0.9427 (±0.0127)**

TABLE V
RESULTS OF EXTERNAL TESTING ON TCGA

kth-fold	CISNS	Thres-L	Thres-M	Thres-H
1	0.8328	NS	NS	NS
2	0.7098	0.01-0.02	0.15-0.16	0.51-0.61
3	0.6884	NS	NS	NS
4	0.7900	0.01-0.03	0.1-0.28/0.33-0.4	0.6-0.7
5	0.7866	0.01/0.03-0.07	0.19-0.44	0.47-0.49/0.51-0.76
mean	**0.7615 (±0.0603)**			
Integration	**0.8278**	0.01	NS	0.54-0.6

Values in Thres-L/M/H represent necrosis thresholds to divide patients into two groups with significant survival differences ($p<0.05$). NS, not significant.

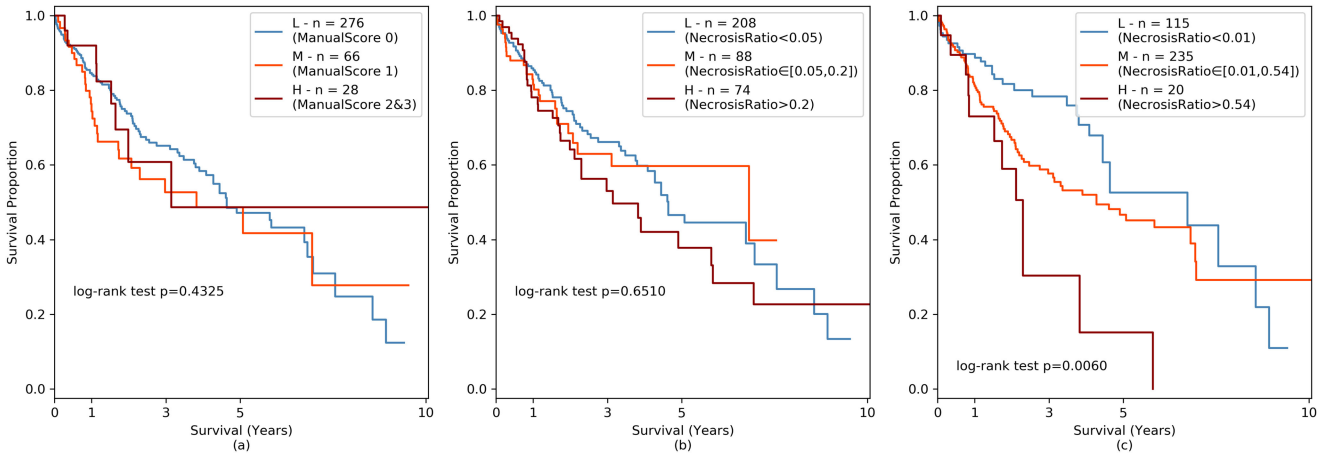

Fig. 5. Comparison between manual scoring and data-driven approach in distinguishing patients with different outcomes based on histopathological necrosis. (a) Overall survival curves by the manual scoring of necrosis. (b) Overall survival curves by the calculated necrosis ratio with the same threshold as the manual score. (c) Overall survival curves by the calculated necrosis ratio via the optimal threshold, with significant differences observed. L, low-risk group. M, median-risk group. H, high-risk group.

strategy and workflow proposed in this paper can be used as a research paradigm.

The results of external validation on TCGA-LIHC show the robustness of the model and its potential to be applied across multiple centers to identify necrosis. With clinical follow-up data and the calculated necrosis ratio, we found that the greater the presence of necrosis in liver cancer was, the worse the OS of the patient. Through the quantitative calculation of necrosis and novel patient stratification, the correlation between necrosis and prognosis in liver cancer can be established, whereas traditional semiquantitative analysis, namely, manual scoring, cannot do the same, showing the advantage of a data-driven approach in survival analysis. The major operation resulting in a significant survival difference in Fig. 5c compared with Fig. 5b is the selection of thresholds to differentiate patients. By lowering the threshold between the low-risk group and the median-risk group to 0.01, the patients with necrosis ratios from 0.01 to 0.05 were moved to the median group, increasing the survival rate of the low-risk group, particularly from the 1st year to the 5th year. Although the extent of necrosis is not very severe, the prognosis of these patients with micronecrosis is poorer than that of patients without necrosis (necrosis ratio lower than

TABLE VI
METRICS OF THE INTEGRATION MODELS BY EXISTING METHODS

methods	int-CISNS	int-Match	int-AUROC	ext-CISNS	ext-Survival	Estimated label cost (time)
bMIL	0.8963	0.7473	0.9017	0.7533	NS	baseline (10000 min)
sMIL	0.9274	**0.7816**	0.9306	0.7947	NS	baseline + 3000 min
pMIL	0.9275	0.7440	0.9267	0.8073	NS	baseline*
spMIL	0.9221	0.7587	0.9196	0.8273	p<0.05	baseline* + 3000 min
LLFP	**0.9341**	0.7560	**0.9336**	**0.8278**	**p<0.01**	baseline*

The prefix int- before the metrics stands for internal test, whereas the prefix ext- stands for external test. The column of ext-Survival records the log rank test of the survival difference between the three groups divided by the calculated necrosis ratios. NS, not significant. The asterisk (∗) in the column of estimated label cost means slightly higher than the time before the ∗ but can be ignored.

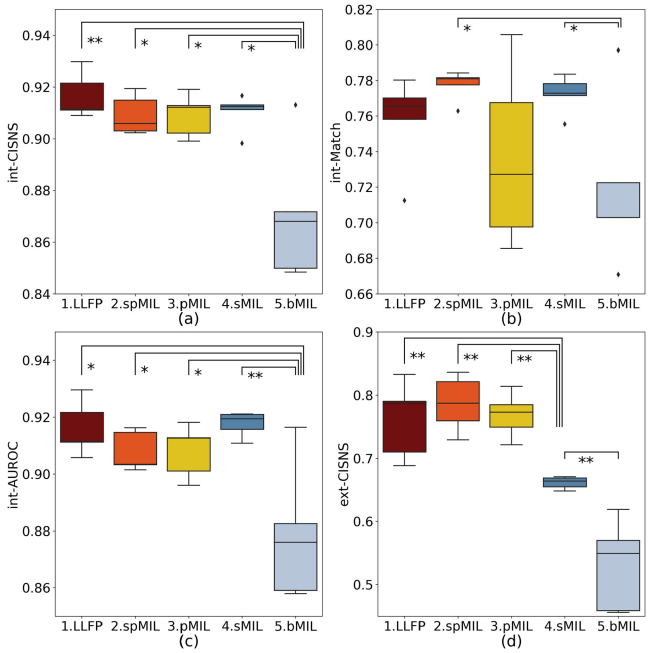

Fig. 6. Comparisons of the cross validation results of the baseline methods. The prefix int- before the metrics stands for internal test, whereas the prefix ext- stands for external test. The single asterisk (∗) denotes p<0.05 and double asterisks (∗∗) denote p<0.01 by pairwise t-test.

0.01). On WSIs, it is possible for manual evaluation to overlook micronecrosis, which might influence patient outcomes, but machines would capture the existence of micronecrosis even in a small proportion. In addition, since the thresholds used in manual scoring were subjectively estimated by clinicians to determine the degree of necrosis, it is more reasonable to reselect thresholds rather than transfer these assumed thresholds directly to quantitative analysis, as shown in Fig. 5b. The selection of optimal thresholds to make finer patient stratification was made possible as a result of quantitative analysis, though these optimal thresholds may need to be further validated by a larger dataset with follow-up data.

The ultimate goal for machine participation in medical care is not to replace clinicians but to provide assistance to clinicians to achieve better medical service; that is, the performance of a clinician with the help of a machine exceeds the performance of either the clinician or the machine alone. In this study, the model could identify some inaccurate labels as shown in Fig. 3(c-f) with the confirmation of experts, representing a good example of computer-aided decision support. In the setting of clinical application, to reduce the workload and improve precision, the trained model can be used as a cluster tool to assist with necrosis quantification. If the classification accuracy of the tile within each class basically meets the requirements of the clinician, the result of the model calculation will be used as the quantitative necrosis ratio; if the accuracy of the tiles fails to meet the requirements, the necrosis ratio will be amended according to the error rate or completely manually evaluated. Meanwhile, this part of the wrongly classified tiles can be fed back as a training bag to the model for parameter adjustment, known as online learning. In addition, to address the problem of inaccurate manual scoring and improve the classification performance, some adaptive learning methods [63] could be introduced in the training-validation process to identify and correct inaccurate labels. Furthermore, larger datasets, better if from multiple centers, might help with model performance to increase the diversity of the data labeling and control the bias of the manual errors.

Since the scale of the datasets in this study was quite large, we did not use preprocessing solutions such as color normalization or data augmentation, which can be considered to improve robustness. In addition, to reduce errors, manual guiding by introducing tile labels to the converging process, combining WSL and SL, may accelerate the convergence and increase accuracy [64]. When implemented at multiple centers, due to the substantial differences in sample preparation and image scanning equipment across different medical centers, the model should be initially calibrated with a small number of sample slides from the local center to prevent tremendous deviation.

Along with pathological features such as necrosis, other clinical factors relating to the prognosis of cancer patients can be integrated to build a more comprehensive prognostic model, since the quantification of pathological features is practically attainable by the proposed method.

V. CONCLUSION

This study proposed a novel method for the quantification of histopathological features on WSIs and collected the largest histopathological image dataset of liver cancer ever known.

Deep learning was exploited in the fuzzy proportion label scenario to maintain relatively strict conditions for convergence at an acceptable cost of labeling. The feasibility of the method for quantifying micronecrosis in liver cancer was demonstrated by utilizing two large-scale datasets from FAH-ZJUMS and TCGA, on which, through ensemble learning, the integration model achieved consistency with manual scoring at 0.9341 and 0.8278, respectively. The integration model shows good robustness on heterogeneous datasets and has the potential to assist clinicians with a faster and more accurate quantification of micronecrosis. Following the research paradigm conducted in this article, other pathological changes active in the tumor microenvironment, such as fibrosis and angiogenesis, on HE slides could be learned and identified automatically as long as their manually evaluated scores (proportion labels) are provided.

As a preliminary exploratory study, this article has obtained some positive results for the quantitative calculation of liver cancer micronecrosis and the differentiation of patients' survival risks based on the extent of necrosis. The proposed method will be implemented as a clinical assistant tool for larger-scale prospective research to validate its applicability in the real world with the intention to actually improve the outcome of patients by providing appropriate risk warnings.

Acknowledgment

The authors would like to thank the TCGA Research Network. The results shown here regarding external test set are based on data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

[1] M. J. Pollheimer *et al.*, "Tumor necrosis is a new promising prognostic factor in colorectal cancer," *Hum. Pathol.,* vol. 41, no. 12, pp. 1749–1757, Dec. 2010.
[2] C. H. Richards *et al.*, "Prognostic value of tumour necrosis and host inflammatory responses in colorectal cancer," *Brit. J. Surg.,* vol. 99, no. 2, pp. 287–294, Feb. 2012.
[3] S. A. Vayrynen *et al.*, "Clinical impact and network of determinants of tumour necrosis in colorectal cancer," *Brit. J. Cancer,* vol. 114, no. 12, pp. 1334–1342, Jun. 2016.
[4] N. Hiraoka *et al.*, "Tumour necrosis is a postoperative prognostic marker for pancreatic cancer patients with a high interobserver reproducibility in histological evaluation," *Brit. J. Cancer,* vol. 103, no. 7, pp. 1057–1065, Sep. 2010.
[5] D. E. Swinson *et al.*, "Tumour necrosis is an independent prognostic marker in non-small cell lung cancer: Correlation with biological variables," *Lung Cancer,* vol. 37, no. 3, pp. 235–240, 2002.
[6] J. C. Cheville *et al.*, "Sarcomatoid renal cell carcinoma: An examination of underlying histologic subtype and an analysis of associations with patient outcome," *Amer. J. Surg. Pathol.,* vol. 28, no. 4, pp. 435–441, 2004.
[7] S. Sengupta *et al.*, "Histologic coagulative tumor necrosis as a prognostic indicator of renal cell carcinoma aggressiveness," *Cancer,* vol. 104, no. 3, pp. 511–520, Aug. 2005.
[8] A. Minervini *et al.*, "Prognostic role of histological necrosis for non-metastatic clear cell renal cell carcinoma: Correlation with pathological features and molecular markers," *J. Urol.,* vol. 180, no. 4, pp. 1284–1289, Oct. 2008.
[9] T. Klatte *et al.*, "Presence of tumor necrosis is not a significant predictor of survival in clear cell renal cell carcinoma: Higher prognostic accuracy of extent based rather than presence/absence classification," *J. Urol.,* vol. 181, no. 4, pp. 1558–1564, Apr. 2009.
[10] M. D. Katz *et al.*, "Percent microscopic tumor necrosis and survival after curative surgery for renal cell carcinoma," *J. Urol.,* vol. 183, no. 3, pp. 909–914, Mar. 2010.
[11] L.-Y. Khor *et al.*, "Tumor necrosis adds prognostically significant information to grade in clear cell renal cell carcinoma," *Amer. J. Surg. Pathol.,* vol. 40, no. 9, pp. 1224–1231, 2016.
[12] R. Zigeuner *et al.*, "Tumour necrosis is an indicator of aggressive biology in patients with urothelial carcinoma of the upper urinary tract," *Eur. Urol.,* vol. 57, no. 4, pp. 575–581, Apr. 2010.
[13] F. G. Barker *et al.*, "Necrosis as a prognostic factor in glioblastoma multiforme," *Cancer,* vol. 77, no. 6, pp. 1161–1166, 1996.
[14] H. Hashimoto *et al.*, "Prognostic significance of histologic parameters of soft tissue sarcomas," *Cancer,* vol. 70, no. 12, pp. 2816–2822, 1992.
[15] R. Leek *et al.*, "Necrosis correlates with high vascular density and focal macrophage infiltration in invasive carcinoma of the breast," *Brit. J. Cancer,* vol. 79, no. 5, pp. 991–995, 1999.
[16] L. E. Rosen *et al.*, "Nuclear grade and necrosis predict prognosis in malignant epithelioid pleural mesothelioma: A multi-institutional study," *Mod. Pathol.,* vol. 31, no. 4, pp. 598–606, Apr. 2018.
[17] B. Alzumaili *et al.*, "Grading of medullary thyroid carcinoma on the basis of tumor necrosis and high mitotic rate is an independent predictor of poor outcome," *Mod. Pathol.*, Apr. 2020.
[18] J. Haratake *et al.*, "Predictable factors for estimating prognosis of patients after resection of hepatocellular carcinoma," *Cancer,* vol. 72, no. 4, pp. 1178–1183, 1993.
[19] Y. Soini *et al.*, "Hepatocellular carcinomas with a high proliferation index and a low degree of apoptosis and necrosis are associated with a shortened survival," *Brit. J. Cancer,* vol. 73, no. 9, pp. 1025–1030, 1996.
[20] J. Vakkila and M. T. Lotze, "Inflammation and necrosis promote tumour growth," *Nature Rev. Immunol.,* vol. 4, no. 8, pp. 641–648, 2004.
[21] A. Karsch-Bluman *et al.*, "Tissue necrosis and its role in cancer progression," *Oncogene,* vol. 38, no. 11, pp. 1920–1935, Mar. 2019.
[22] D. Jiao *et al.*, "Necroptosis of tumor cells leads to tumor necrosis and promotes tumor metastasis," *Cell Res,* vol. 28, no. 8, pp. 868–870, Aug. 2018.
[23] M. T. Lotze and K. J. Tracey, "High-mobility group box 1 protein (HMGB1): Nuclear weapon in the immune arsenal," *Nature Rev. Immunol.,* vol. 5, no. 4, pp. 331–342, Apr. 2005.
[24] O. Tredan *et al.*, "Drug resistance and the solid tumor microenvironment," *J. Nat. Cancer Inst.,* vol. 99, no. 19, pp. 1441–1454, Oct. 2007.
[25] L. A. Cooper *et al.*, "The tumor microenvironment strongly impacts master transcriptional regulators and gene expression class of glioblastoma," *Amer. J. Pathol.,* vol. 180, no. 5, pp. 2108–2119, May 2012.
[26] S. Salah *et al.*, "Tumor necrosis and clinical outcomes following neoadjuvant therapy in soft tissue sarcoma: A systematic review and meta-analysis," *Cancer Treat. Rev.,* vol. 69, pp. 1–10, Sep. 2018.
[27] F. M. C. S. Group and P. Bedossa, "Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C," *Hepatology,* vol. 20, no. 1, pp. 15–20, 1994.
[28] W. C. Allsbrook, Jr. *et al.*, "Interobserver reproducibility of gleason grading of prostatic carcinoma: General pathologist," *Hum. Pathol.,* vol. 32, no. 1, pp. 81–88, Jan. 2001.
[29] M. H. Stoler and M. Schiffman, "Interobserver reproducibility of cervical cytologic and histologic interpretations: Realistic estimates from the ASCUS-LSIL triage study," *JAMA,* vol. 285, no. 11, pp. 1500–1505, 2001.
[30] A. M. El-Badry *et al.*, "Assessment of hepatic steatosis by expert pathologists: The end of a gold standard," *Ann. Surg.,* vol. 250, no. 5, pp. 691–697, Nov. 2009.
[31] R. K. Jain *et al.*, "Atypical ductal hyperplasia: Interobserver and intraobserver variability," *Mod. Pathol.,* vol. 24, no. 7, pp. 917–923, Jul. 2011.
[32] L. E. Franzen *et al.*, "Semiquantitative evaluation overestimates the degree of steatosis in liver biopsies: A comparison to stereological point counting," *Mod. Pathol.,* vol. 18, no. 7, pp. 912–916, Jul. 2005.
[33] C. Mercan *et al.*, "Multi-Instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE Trans. Med. Imag.,* vol. 37, no. 1, pp. 316–325, Jan. 2018.
[34] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.,* vol. 25, no. 8, pp. 1301–1309, Aug. 2019.
[35] D. Karimi *et al.*, "Deep learning-based gleason grading of prostate cancer from histopathology images-role of multiscale decision aggregation and data augmentation," *IEEE J. Biomed. Health Inform.,* vol. 24, no. 5, pp. 1413–1426, May 2020.
[36] C. Sun *et al.*, "Deep learning-based classification of liver cancer histopathology images using only global labels," *IEEE J. Biomed. Health Inform.,* vol. 24, no. 6, pp. 1643–1651, Jun. 2020.
[37] N. Coudray *et al.*, "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *Nature Med.,* vol. 24, no. 10, pp. 1559–1567, Oct. 2018.

[38] K. H. Yu *et al.*, "Classifying non-small cell lung cancer types and transcriptomic subtypes using convolutional neural networks," *J. Amer. Med. Inform. Assoc.,* vol. 27, no. 5, pp. 757–769, May 2020.
[39] J. N. Kather *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature Med.,* vol. 25, no. 7, pp. 1054–1056, Jul. 2019.
[40] J. N. Kather *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS Med.,* vol. 16, no. 1, p. e1002730, Jan. 2019.
[41] P. Courtiol *et al.*, "Deep learning-based classification of mesothelioma improves prediction of patient outcome," *Nature Med.,* vol. 25, no. 10, pp. 1519–1525, Oct. 2019.
[42] Y. Yamamoto *et al.*, "Automated acquisition of explainable knowledge from unannotated histopathology images," *Nature Commun.,* vol. 10, no. 1, pp. 5642, Dec. 2019.
[43] F. Xing *et al.*, "Pixel-to-pixel learning with weak supervision for single-stage nucleus recognition in Ki67 images," *IEEE Trans. Biomed. Eng.,* vol. 66, no. 11, pp. 3088–3097, Nov. 2019.
[44] A. Homeyer *et al.*, "Practical quantification of necrosis in histological whole-slide images," *Comput. Med. Imag. Graph.,* vol. 37, no. 4, pp. 313–322, Jun. 2013.
[45] G. Carneiro *et al.*, "Automatic detection of necrosis, normoxia and hypoxia in tumors from multimodal cytological images," *IEEE Int. Conf. Image Process. (ICIP)*, 2015, pp. 2429–2433.
[46] H. Sharma *et al.*, "Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology," *Comput. Med. Imag. Graph.,* vol. 61, pp. 2–13, Nov. 2017.
[47] H. B. Arunachalam *et al.*, "Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models," *PLoS One,* vol. 14, no. 4, pp. e0210706, 2019.
[48] N. D. Roy *et al.*, "Detection of necrosis in mice liver tissue using deep convolutional neural network," *Int. Conf. Patter Recognit. Mach. Intell.*, pp. 32–40, 2019.
[49] G. Quellec *et al.*, "Multiple-Instance learning for medical image and video analysis," *IEEE Rev. Biomed. Eng.,* vol. 10, pp. 213–234, 2017.
[50] N. Quadrianto *et al.*, "Estimating labels from label proportions," *J. Mach. Learn. Res.,* vol. 10, no. 10, 2009.
[51] K. Fan *et al.*, "Learning a generative classifier from label proportions," *Neurocomputing,* vol. 139, pp. 47–55, 2014.
[52] G. Bortsova *et al.*, "Deep learning from label proportions for emphysema quantification," *Med. Image Comput. Comput. Assisted Intervention – MICCAI* 2018, pp. 768–776, 2018.
[53] Y. Deng *et al.*, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Trans. Fuzzy Syst.,* vol. 25, no. 4, pp. 1006–1012, 2017.
[54] Z. Zhang *et al.*, "Fuzzy multilayer clustering and fuzzy label regularization for unsupervised person Re-identification," *IEEE Trans. Fuzzy Syst.*, pp. 1–1, 2019.
[55] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.,* vol. 9, no. 1, pp. 62–66, 1979.
[56] T. C. G. A. R. Network, "Comprehensive and integrative genomic characterization of hepatocellular carcinoma," *Cell,* vol. 169, no. 7, pp. 1327–1341 e23, Jun. 2017.
[57] A. Goode *et al.*, "OpenSlide: A vendor-neutral software foundation for digital pathology," *J. Pathol. Inform.,* vol. 4, 2013.
[58] A. Paszke *et al.*, "Automatic differentiation in pytorch," 2017.
[59] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.,* vol. 12, pp. 2825–2830, 2011.
[60] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.,* vol. 9, no. 3, pp. 90–95, 2007.
[61] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016, pp. 770–778.
[62] R. Mormont *et al.*, "Comparison of deep transfer learning strategies for digital pathology," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2262–2271.
[63] D. Hao *et al.*, "Inaccurate labels in weakly-supervised deep learning: Automatic identification and correction and their impact on classification performance," *IEEE J. Biomed. Health Inform.,* vol. 24, no. 9, pp. 2701–2710, Sep. 2020.
[64] H. Yang *et al.*, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE Trans. Med. Imag.,* vol. 39, no. 5, pp. 1306–1315, May 2020.