Research Topics

Computational pathology for precision diagnosis, treatment, and prognosis of cancer

Edited by
Jun Cheng, Kun Huang and Jun Xu

Published in
Frontiers in Medicine

ISSN 1664-8714
ISBN 978-2-8325-4488-4
DOI 10.3389/978-2-8325-4488-4

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Computational pathology for precision diagnosis, treatment, and prognosis of cancer

Topic editors

Jun Cheng — Shenzhen University, China

Kun Huang — Indiana University Bloomington, United States

Jun Xu — Nanjing University of Information Science and Technology, China

Citation

Cheng, J., Huang, K., Xu, J., eds. (2024). *Computational pathology for precision diagnosis, treatment, and prognosis of cancer*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4488-4

Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 06 June 2023
DOI 10.3389/fmed.2023.1209666

OPEN ACCESS

EDITED AND REVIEWED BY
Luigi M. Terracciano,
University of Basel, Switzerland

*CORRESPONDENCE
Jun Cheng
✉ juncheng@szu.edu.cn

RECEIVED 21 April 2023
ACCEPTED 23 May 2023
PUBLISHED 06 June 2023

CITATION
Cheng J, Huang K and Xu J (2023) Editorial: Computational pathology for precision diagnosis, treatment, and prognosis of cancer. *Front. Med.* 10:1209666.
doi: 10.3389/fmed.2023.1209666

Editorial: Computational pathology for precision diagnosis, treatment, and prognosis of cancer

Jun Cheng[1,2,3]*, Kun Huang[4,5] and Jun Xu[6]

[1]National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China, [2]Medical Ultrasound Image Computing (MUSIC) Laboratory, Shenzhen University, Shenzhen, China, [3]Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, China, [4]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States, [5]Regenstrief Institute, Indianapolis, IN, United States, [6]Institute for AI in Medicine, School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China

KEYWORDS

computational pathology, digital pathology, machine learning, deep learning, medical image analysis, biomedical data analysis

Editorial on the Research Topic
Computational pathology for precision diagnosis, treatment, and prognosis of cancer

Histopathology is considered the gold standard in determining the presence and nature of tumors. Technological advances in automated high-speed and high-resolution whole-slide imaging have laid the foundation for a digital revolution in microscopy. Digital histopathological images can be analyzed efficiently with image analysis and machine learning techniques. These techniques have shown great potential for extracting sub-visual, quantitative, and valuable features from whole-slide images to characterize tumors and support clinical decision (1, 2). Besides histopathological images, other data modalities, such as radiological images and multi-omics data, are also used to assist the decision-making process for cancer diagnosis, treatment, and prognosis (3, 4). At present, it is not clear how these macroscopic, microscopic, and molecular features are related. Exploring the association between different data modalities can give new insights into diseases.

This Research Topic is to highlight some latest developments in computational pathology that use either classical image analysis or state-of-the-art deep learning solutions for improved clinical decision making. A brief summary of the articles in this Research Topic is provided below.

Segmentation of regions of interest is usually an important step in the workflow of computer-aided diagnosis. Shi et al. collected a new enteroscope biopsy histopathological image dataset for image segmentation tasks and submitted it to a public data repository. This dataset contains 2,228 colorectal tissue images and their corresponding ground-truth annotations with the size of 224 × 224 pixels. To cover the transition process from normal to cancerous tissue, this dataset includes six tumor differentiation stages: normal, polyp, low-grade intraepithelial neoplasia, high-grade intraepithelial neoplasia, serrated adenoma, and adenocarcinoma. In this work, they compared the segmentation performance of five classical machine learning methods and three deep learning methods. Generally, the deep learning methods outperformed the classical machine learning methods by a large margin in all six tissue types. This study and the released dataset can serve as a good benchmark for colorectal histopathological image segmentation. Zhao et al. proposed a deep segmentation network to distinguish cancerous and intestinal metaplasia regions from normal gastric tissue. The segmentation results of multiple whole slide images from a specimen were mapped to the macroscopic image of the specimen. For a convenient use, they developed a software to automate the construction of mucosal recovery maps, which can expedite the learning process of early gastric cancer diagnosis.

Annotating pathological images requires professional knowledge and is time-consuming and costly. The annotations of most existing public datasets focus on the ground truth labels about what the diseases and lesions are, rather than why and how they are discovered and decided. Therefore, these datasets are not directly applicable for clinical use. To address this issue, Zhang et al. proposed a new annotation form, PathNarrative, which includes a hierarchical decision-to-reason data structure, a narrative annotation process, and a multimodal interactive annotation tool. PathNarrative can help collect both decision-to-reason labels and multimodal information on vision, language, voice, and behavioral trajectories. To verify the efficacy of this new annotation tool for human-AI collaborative diagnosis, they experimented on a colorectal pathological dataset with classification and captioning tasks. The experimental results show that the classification and captioning tasks achieve better performance with refined annotations, provide explainable details for doctors to make clinical decisions, and thus enhance doctors' trustworthiness and confidence to collaborate with artificial intelligence models.

Hu et al. performed a comparative study of gastric pathological image classification. They used a publicly available dataset, GasHisSDB, which contains three sub-datasets with different image sizes (80 × 80, 120 × 120, and 160 × 160 pixels). Seven classical machine learning classifiers and four deep learning classifiers were tested. For the classical machine learning classifiers, five feature extraction methods were used, including color histogram, luminance histogram, histogram of oriented gradient, local binary patterns, and gray-level co-occurrence matrix. Overall, the deep learning classifiers achieved much higher accuracy than the classical machine learning classifiers, no matter what kinds of features were used. In addition, they found that the deep learning classifiers misclassified different samples, implying that it is possible to use ensemble learning to obtain better predictive performance. Fully supervised methods require a sufficient quantity of images with annotations. However, in medical field it is difficult to collect and label data, which needs to be performed by experts. Wang et al. proposed a self-supervised learning method to classify malignant and non-malignant pathological images in eyelid melanoma. This method took advantage of a relatively abundant quantity of unlabeled data and a limited quantity of labeled data to learn features. In the self-supervised setting with a subset of images labeled, the proposed method achieved the best performance compared with five fully supervised methods.

Another popular research interest in computational pathology is to predict cancer survival based on quantitative image features and associate these features with molecular data. In a study by Couetil et al., interpretable histopathological features were extracted from whole slide images to predict 5-years survival and 5-years metastasis of melanoma. They used the morphological feature set described in a previous study (2) and introduced additional features to describe lymphocytes and other small, hyperchromatic cells. In total, 135 morphological features were extracted. Four classical machine learning models were implemented, including random forest, support vector machine, k-nearest neighbors, and logistic regression. This approach yielded a maximum F1 score of 0.72 and 0.73 for predicting survival and metastasis, respectively. Tumor-stroma reaction (TSR) is a critical feature in many solid tumors. Jiang et al. trained a serial of deep learning models to identify tumor vs. stroma regions and predict three types of TSR scores (fibrosis, stromal cellularity, and orientation of stromal cells) in ovarian carcinoma. Within the tumor-stroma interface region, they found that the TSR fibrosis scores were strongly associated with patient survival. Correlating the TRS fibrosis scores with gene expression data, they further found that the positively correlated genes were enriched in 14 KEGG pathways that are mostly associated with cancer signaling aberrations. This genotype-phenotype association analysis enables discovering the molecular basis of tissue morphological changes.

Author contributions

Editorial writing: JC. Review and editing: JC, KH, and JX. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Wang X, Barrera C, Bera K, Viswanathan VS, Azarianpour-Esfahani S, Koyuncu C, et al. Spatial interplay patterns of cancer nuclei and tumor-infiltrating lymphocytes (TILs) predict clinical benefit for immune checkpoint inhibitors. *Sci Adv.* (2022) 8:3966. doi: 10.1126/sciadv.abn3966

2. Cheng J, Han Z, Mehra R, Shao W, Cheng M, Feng Q, et al. Computational analysis of pathological images enables a better diagnosis of TFE3 Xp11.2 translocation renal cell carcinoma. *Nat Commun.* (2020) 11:1–9. doi: 10.1038/s41467-020-15671-5

3. Chen RJ, Lu MY, Williamson DFK, Chen TY, Lipkova J, Noor Z, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell.* (2022) 40:865–78.e6. doi: 10.1016/j.ccell.2022.07.004

4. Shao W, Han Z, Cheng J, Cheng L, Wang T, Sun L, et al. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. *IEEE Trans Med Imaging.* (2020) 39:99–110. doi: 10.1109/TMI.2019.2920608

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 07 September 2022
DOI 10.3389/fmed.2022.994467

Computational tumor stroma reaction evaluation led to novel prognosis-associated fibrosis and molecular signature discoveries in high-grade serous ovarian carcinoma

Jun Jiang[1†], Burak Tekin[2†], Lin Yuan[3†], Sebastian Armasu[1], Stacey J. Winham[1], Ellen L. Goode[1], Hongfang Liu[4*‡], Yajue Huang[2*‡], Ruifeng Guo[2*‡] and Chen Wang[1*‡]

[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, United States, [2]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States, [3]Pathology Center, Shanghai General Hospital, Shanghai, China, [4]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, United States

Background: As one of the key criteria to differentiate benign vs. malignant tumors in ovarian and other solid cancers, tumor-stroma reaction (TSR) is long observed by pathologists and has been found correlated with patient prognosis. However, paucity of study aims to overcome subjective bias or automate TSR evaluation for enabling association analysis to a large cohort.

Materials and methods: Serving as positive and negative sets of TSR studies, H&E slides of primary tumors of high-grade serous ovarian carcinoma (HGSOC) (n = 291) and serous borderline ovarian tumor (SBOT) (n = 15) were digitally scanned. Three pathologist-defined quantification criteria were used to characterize the extents of TSR. Scores for each criterion were annotated (0/1/2 as none-low/intermediate/high) in the training set consisting of 18,265 H&E patches. Serial of deep learning (DL) models were trained to identify tumor vs. stroma regions and predict TSR scores. After cross-validation and independent validations, the trained models were generalized to the entire HGSOC cohort and correlated with clinical characteristics. In a subset of cases tumor transcriptomes were available, gene- and pathway-level association studies were conducted with TSR scores.

Results: The trained models accurately identified the tumor stroma tissue regions and predicted TSR scores. Within tumor stroma interface region, TSR fibrosis scores were strongly associated with patient prognosis. Cancer signaling aberrations associated 14 KEGG pathways were also found positively correlated with TSR-fibrosis score.

Check for updates

OPEN ACCESS

EDITED BY
Jun Cheng,
Shenzhen University, China

REVIEWED BY
Shidan Wang,
University of Texas Southwestern Medical Center, United States
Yao Wu,
Children's National Hospital, United States

*CORRESPONDENCE
Chen Wang
wang.chen@mayo.edu
Hongfang Liu
liu.hongfang@mayo.edu
Yajue Huang
huang.yajue@mayo.edu
Ruifeng Guo
guo.ruifeng@mayo.edu

[†]These authors have contributed equally to this work and share first authorship

[‡]These authors share last authorship

SPECIALTY SECTION
This article was submitted to Pathology, a section of the journal Frontiers in Medicine

RECEIVED 14 July 2022
ACCEPTED 15 August 2022
PUBLISHED 07 September 2022

CITATION
Jiang J, Tekin B, Yuan L, Armasu S, Winham SJ, Goode EL, Liu H, Huang Y, Guo R and Wang C (2022) Computational tumor stroma reaction evaluation led to novel prognosis-associated fibrosis and molecular signature discoveries in high-grade serous ovarian carcinoma. *Front. Med.* 9:994467. doi: 10.3389/fmed.2022.994467

Conclusion: With the aid of DL, TSR evaluation could be generalized to large cohort to enable prognostic association analysis and facilitate discovering novel gene and pathways associated with disease progress.

KEYWORDS

tumor-stroma reaction, high-grade serous ovarian carcinoma, digital pathology, prognostic fibrosis, molecular signature

Introduction

Ovarian cancer (OC) is one of the leading causes of mortality among cancers in women. Pathologically, ovarian cancer is divided into high-grade and low-grade carcinomas, and the high-grade carcinomas can be further classified into various histological subtypes, most commonly serous, endometrioid and clear cell. Among them, high-grade serous ovarian carcinoma (HGSOC) is the prevalent histotype and accounts for vast majority of ovarian cancer associated mortality (1). These patients often present with rapid clinical progression, disseminated peritoneal metastasis, distant metastasis, and resistance to treatments. On the other hand, low-grade ovarian carcinomas usually present with slow-progressing diseases and are associated with much lower mortality but protracted clinical courses. Diagnostically, low-grade ovarian carcinomas can be difficult to distinguish from ovarian borderline tumors with much more indolent clinical behavior, while sometimes may share overlapping histological features with aggressive high-grade carcinomas. While all malignant OCs regardless histology types are treated similarly using platinum-based front-line chemotherapies, different surgical resections and chemotherapy treatments options could be applied to different histologic subtypes. In clinical diagnosis, recognizable histological features play a critical role in differentiating these subtypes.

Although histological diagnosis of HGSOC has been well-established, many studies have shown highly heterogenous clinical courses in these patients (2–4). Interestingly, pathologists have long observed the high variability of tumor associated stroma reaction in HGSOCs in daily practice (4, 5). Similar to a process of normal wound healing, the tumor-stroma reaction (TSR) in cancer has been associated with increased extracellular matrix and production of growth factors to facilitate recovery growth of injured tissues (6). In ovarian tumors, histopathological examination of tumor-stroma reaction is critical to differentiate low-grade serous carcinoma from serous borderline serous tumor (SBOT), with the latter lacking tumor triggered stroma reaction. More importantly, tumor-stroma reaction has been reported to facilitate tumorigenesis and associated with prognostic differences in many solid cancers such as cholangiocarcinoma, pancreatic cancer, melanoma, and OC (7–10). Though numerous studies have demonstrated that the interactions between tumor cells and stroma play a critical role in cancer progression and metastasis across multiple cancer types (11–13), the association of histological feature of stromal reaction with molecular mechanism is still underexplored. One of major reasons responsible for this gap is the lack of quantitative evaluation of TSR in solid tumors. In daily pathology practice and many research studies, TSR were examined by manually reviewing H&E-stained slides by individual pathologist, which is highly subjective and labor-intensive. Interobserver variability remains a main challenge thus limiting large-scale investigation of TSR. More importantly, evaluation of TSR by pathologist relies on personal experience, while an unbiased quantification becomes unrealistic which may cause heterogenous quality of the TSR scoring data with poorly reproducibility.

With the advancement of digital pathology, there has been substantial interest in exploring the role of quantitative attributes computationally extracted from H&E-stained whole slide images (WSI). Li et al. (14) introduced a digital pathology-based pipeline to early-stage estrogen receptor-positive invasive breast cancers for association analysis. Their results suggested that the orientation disorder of collagen fiber is prognostic for early-stage breast cancer (14). Geessink et al. (15) trained a deep learning model to segment relevant tissue types in rectal cancer histology and subsequently calculate tumor-stroma ratio for intra-tumoral stroma. Their results showed that tumor-stroma ratio is an independent prognosticator in rectal cancer when assessed automatically in user-provided stroma hot-spots (15). Failmezger et al. (16) introduced topological features extraction method to quantify stromal recruitment for immunosuppression in melanoma histology using graph based spatial model. This research revealed that tumors with high stromal clustering and barrier had reduced expression of pathways involved in naïve CD4 signaling, MAPK, and PI3K signaling, and indicated that computational histology-based stromal phenotypes within the tumor microenvironment are significantly associated with prognosis and immune exclusion in melanoma (16).

In this paper, we present a digital pathology-based pipeline that is able to prognosticate patient survival by estimating degree of TSR directly from multiple aspects of digitized H&E images. Specifically, the automated pipeline consists of image

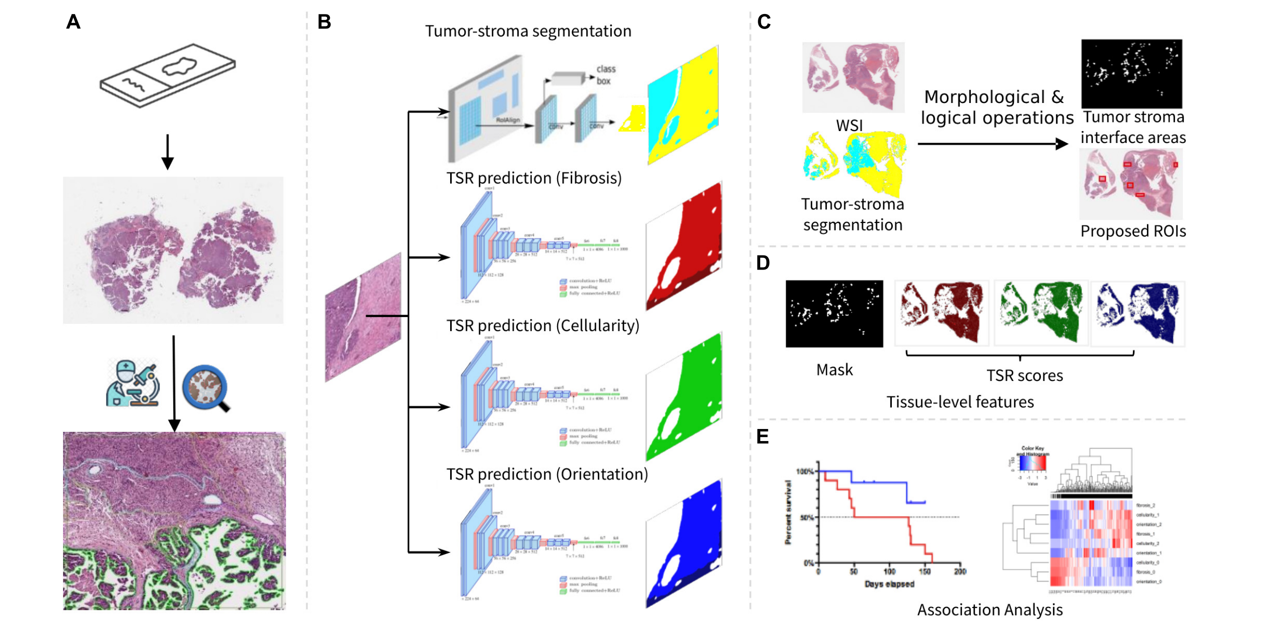


FIGURE 1
Overview of our research workflow. **(A)** Slide scanning and annotation. **(B)** Tumor-stroma segmentation and TSR score estimation. **(C)** Tumor stroma interface area identification. **(D)** Tissue-level feature summarization. **(E)** Association analysis.

processing techniques combined with several machine learning models trained from pathologists' annotations. Serving as the cores of the pipeline, the trained models were used to identify tumor-associated stroma regions, from which we subsequently predicted TSR scores with H&E images as inputs. As shown in **Figure 1**, the developed pipeline was applied to our research cohort to establish associations between tissue-level features, prognosis, and molecular pathways of HGSOC.

Materials and methods

Cohort summary

Our research cohort consists of 291 HGSOC and 15 serous borderline ovarian tumor (SBOT) cases ascertained at the Mayo Clinic between 1994 and 2009. The SBOT cases should not have significant TSR by definition, therefore served as the negative control in both TSR score prediction and evaluation. As the cohort selection criteria, all the cases had retrievable clinical, molecular, and tissue blocks. Survival data were obtained from the Mayo Clinic Tumor Registry, electronic medical records. Gene expression profiles and histological information, including tissue sites (primary or metastasis) and tumor stage, were collected from the EHR system in Mayo Clinic. The cohort characteristics were summarized in **Table 1**. All the slides were scanned in Pathology Research Core at Mayo Clinic with a digital whole-slide scanner (Aperio Scanscope XT). To preserve cell and tissue details, the slides were scanned with 40× resolution (pixel size: 0.25 um). Imaging quality was manually checked by histology technicians when the slides were scanned. All patients provided informed consent for use of their tissues and data in research; all protocols were approved by the Mayo Clinic Institutional Review Board.

Pathologist-guided image annotation

According to the consensus of three pathologists, three types of histopathological evaluation metrics (sub-scores) were defined as criteria to characterize the extent of TSR: (i) increased fibrosis, characterized by collagen deposition, (ii) increased stromal cellularity due to fibroblastic and/or myofibroblastic proliferation, and (iii) orientation of stromal cells (14). Based on histopathologic examination of H&E-stained slide areas, the sub-TSR scores were assessed as 0 (none/weak), 1 (intermediate) or 2 (strong) (**Supplementary Figure 1**). The criteria were chosen to reflect the histopathologic changes commonly observed and evaluated in the clinical practice. With regard to the criterion of stromal cellularity, a minimal density of (myo)fibroblasts was assigned a score of 0, while a score of 2 was given if the area occupied by (myo)fibroblasts exceeded the area occupied by the acellular stroma in a given field. As for the fibrosis criterion, minimal deposition of fine collagen fibers with significant fiber spacing was assigned a score of 0, whereas dense collagen deposition with sclerosis was classified as 2. In terms of orientation of stromal cells, a relatively

TABLE 1 Research cohort statistics.

	Overall (*N* = 291)
Histology*	
High grade serous	291 (100.0%)
Age at diagnosis	
Mean (SD)	63.337 (11.231)
Median	64.000
Q1, Q3	56.000, 71.000
Range	24.000 - 89.000
Age at diagnosis (group)	
[20,50] (premenopausal)	32 (11.0%)
(50,90] (postmenopausal)	259 (89.0%)
Stage	
3	217 (74.6%)
4	74 (25.4%)
Grade	
2	1 (0.3%)
3	290 (99.7%)
Vital status at Last Follow-up	
Alive	34 (11.7%)
Deceased	257 (88.3%)
Months from Diagnosis to Enrollment	
Mean (SD)	0.989 (8.425)
Median	0.000
Q1, Q3	0.000, 0.082
Range	0.000 - 107.664
Months from Diagnosis to Last Follow-up	
Mean (SD)	50.358 (43.148)
Median	37.072
Q1, Q3	17.763, 70.197
Range	0.263 - 196.711
Median Time to Last Follow-up (months)	
Events	257
Median Survival	37.434
Debulking Status	
Missing	1
Optimal	220 (75.9%)
Suboptimal	70 (24.1%)
Suboptimal	70 (24.1%)

*Since the SBOT cases were only included in training deep learning models for providing negative controls, the characteristics of SBOT cases were not included in this table.

linear, unidirectional orientation was assigned a score of 2, while a haphazard orientation without appreciable directionality was scored as 0.

To create an annotated dataset for model training, five HGSOC slides were randomly selected. Within each slide, five most representative regions of interest (ROIs) were circled by pathologists in tumor-stroma interface regions, which were areas where the borders of the tumor islands came into close proximity with the surrounding stromal areas and the stroma exhibited morphologic characteristics different than the non-neoplastic ovarian stroma (Figure 1A). Two experienced pathologists were invited to annotate three TSR scores using an interactive tool named QuPath (17). In each slide, five most representative ROIs were circled by pathologists in tumor-stroma interface regions. The size of each ROI was at least 1,024*1,024 pixels. Within each ROI, polygons were used to annotate homogeneous regions with the same TSR scores. Sub-regions with the same TSR scores were labeled to the same category (Supplementary Figure 1B).

Dataset preprocessing

Using a framework developed in our previous work (18), TSR annotations were parsed using Groovy script within QuPath and converted into a pair of image and annotation mask for each annotated ROI. For the convenience of visualization, annotation masks were encoded from dark to light R/G/B colors for each TSR scoring metric (Supplementary Figures 1D–F). To extract regular size of images and annotation masks for model training and evaluation, a 256*256 pixel sliding window was applied to the ROIs. Taking full capacity of pathologists' annotations, the sampling stride was set to 128 for the aim of creating augmented/enlarged dataset. In total, 11,240 image patches with TSR annotations were obtained from HGSOC cases.

Considering that SBOT confirmed cases were free of significant TSR, we proposed to train the TSR prediction model with image samples from SBOT WSIs as negative controls. Thus, image patches from SBOT cases were also prepared and labeled to TSR score zero. We randomly selected five slides from our previous research (18), from which 1,405 image patches from stroma regions were randomly extracted and added into "annotated" dataset. In total, 7,025 image patches were obtained from SBOT cases.

To identify tumor-stroma interface areas and quantify TSR inside these regions, we repurposed TSR annotations for tumor-stroma segmentation modeling. The stroma region is defined as all the tissue region except the tumor region. Within each annotated ROI, overall stroma regions were obtained by merging all three different TSR score regions, while the tumor regions were defined as the remaining tissue regions inside the ROIs. The same sliding-window sampling strategy was used to extract image patches for segmentation modeling.

Tumor-stroma tissue segmentation

In order to identify tumor-stroma interface areas where TSR occurs, tumor and stroma regions were segmented using a deep learning neural network named Mask-RCNN (19), as shown in Figure 1B. Mask -RCNN was preferred in this study as it has been used in many histological image processing tasks (20, 21) and was reported to be more robust than the U-Net for image segmentation (22, 23). The hyperparameters of Mask-RCNN, such as the dimension of convolutional layers (input dimension = 256 × 256), learning rate (lr = 0.01) and RPN anchor scales (RPN = [8, 16, 32, 64, 128]) were modified to adapt to our image segmentation task. Taking annotated dataset described in Data Preprocessing section, images from HGSOC cases were shuffled and divided into training, validation, and testing groups (3,317:1,105:1,110). From the training subset, the input layers of Mask-RCNN took both original images and tumor-stroma multilabel masks as training samples. In the training process, tumor and stroma areas were iteratively proposed by a sub-structure of Mask-RCNN named Region Proposal Network (RPN). Fully connected network layers were concatenated to the forehead layers to identify differences between ground truth (annotations) and proposed segmentation masks. By minimizing the differences, Mask-RCNN was trained to segment tumor and stroma within WSIs. To facilitate training convergence, weights from the model pretrained with Coco dataset (24) were loaded to our model as initial settings and fine-tuned on our training dataset. Specifically, the model trained at the 315th epoch reached the lowest loss in the validation set.

In the testing phase, tumor stroma regions were segmented and saved as multilabel masks for each image patch from the hold-out testing dataset using trained tumor-stroma segmentation models. Three commonly used evaluation metrics for image segmentation tasks (25, 26) were calculated to measure concordance between model prediction and ground truth, including DSC (Dice similarity coefficient), IoU (intersection of union) and AP (averaged precision).

Tumor-stroma reaction scoring modeling

In our work, TSR score estimation was formulated into an image classification problem. In other words, different TSR scores corresponded to different image categories. We employed a commonly used DL network architecture named VGG16 (27) as our image classification model, as this model also achieved an encouraging performance on identifying tumor infiltrating lymphocytes (TIL) (28). To estimate three TSR scores, three VGG16 models (**Figure 1B**) were trained with the combinational dataset (from both HGSOC and SBOT cases) established in data preprocessing steps. The annotated dataset was divided into training, validation, and testing subsets (12,743:2,247:2,249, 6,000 training samples were from SBOT as negative control). During the training phase, the training dataset was divided into batches (32 samples per batch) to meet the computational resource limitations. The maximum training epoch was set to 30. At the end of each training epoch, training loss was calculated on the validation dataset. To avoid overfitting, the training process was set to stop when the loss variation is less than 10^{-3} within four epochs. To increase generalizability and avoid bias from different H&E-staining conditions, training image dataset was augmented using linear image transformation, such as rotation and flipping. With the same training strategy, three VGG16 image classification models were trained independently to estimate fibrosis, cellularity, and orientation TSR scores, respectively (**Figure 1B**). During the testing phase, three TSR scores were estimated for each input image from the hold-out testing dataset using the three trained VGG16 models. The performances of the three models were evaluated by comparing the concordance between model estimation and human annotation.

Extrinsic model evaluation

Before applying our models to the entire research cohort, it was essential to evaluate model performances on an independent dataset (extrinsic) as our DL models were trained and evaluated using annotated images from ROIs (intrinsic). To this end, we developed an interactive evaluation tool (**Supplementary Figure 2**) for the assessment on a dataset that was independent of annotated ROIs and WSIs. For the sake of pathologists' convenience, the original image as well as corresponding tumor-stroma segmentation and TSR scores predicted by the trained models were loaded into this evaluation tool. With the original image in the center, eight neighborhood image patches were also shown in the user interface as additional references for pathologists to make an accurate judgment. The concordance (average precision) between pathologists' evaluations and predictions were recorded as pathologists proceeded reviewing by clicking buttons and checking boxes.

To create an independent dataset for extrinsic evaluation, image patch-level TSR score distribution was calculated for each slide. Slides with ultra-high (TSR = 2) or low (TSR = 1) TSR score ratio were selected to epitomize the performances of our trained models. Unannotated cases were selected based on TSR score ratio distribution, only the upper 5% quantile and low 5% quantile were included for manual evaluation. To limit the number of images to be evaluated, we randomly selected at least 10 but less than 30 images from each slide within tumor stroma interfacing regions.

Applying deep learning models to research cohort

Trained DL models were generalized to the research cohort according to the following procedures.

Tissue detection and patch extraction

To cut down the computational cost, only the tissue regions (foreground of WSIs) were included in the testing phase. Tissue regions were detected within a down-sampled whole slide image (down-sample rate = 128), and then image patches were extracted from tissue regions accordingly. To detect foreground within down-sampled whole slide images, the color space was converted from RGB to LAB, and then threshold method was applied to L channel for tissue detections. As a commonly used foreground detection method for bright field whole slide images, feasibility of this method has been proved in our previous research (18, 29). By mapping the coordinates of pixels within detected tissue regions back to original WSI, image patches (256*256 pixels) were extracted in a tiling manner. With the same threshold-based method, foreground within the original resolution of image patches was detected. If more than 50% of the patch is background, it was excluded from our study. In the end, the extracted images were fed into our trained models for patch-level predictions.

Tumor-stroma interface area identification

To investigate TSR in invasive tumor front, tumor-stroma interface areas were identified based on tumor-stroma segmentation results. By down-sampling (r = 1/128) and stitching patch level results back to their original locations, slide level tissue segmentation (tumor vs. stroma) were reconstructed. Within slide level tumor-stroma segmentation results, tumor-stroma interface areas were identified using serials of image morphological and logical operations (**Figure 1C** and **Supplementary Figure 5**). The calculation process can be formulated as follows:

$$Tumor_{core} = C\left(I_T, S\right)$$

$$Stroma_{core} = C\left(I_S, S\right)$$

$$ROI_{interface} = and\left(xor\left[D\left(Stroma_{core}, S\right), E\left(Stroma_{core}, S\right)\right],\right.$$

$$\left. Stroma_{core}, D\left(Tumor_{core}, S\right)\right),$$

in which, I_T and I_S denotes the tumor and stroma multilabel image from slide level segmentation, respectively. S denotes the structural elements for morphological operations, including closing C(I, S), erosion E(I, S) and dilation D(I, S) (30).

To evaluate the accuracy of this automatic tumor-stroma localization method, the multilabel masks of interface areas and the counterpart WSIs were shown side-by-side and reviewed by our pathologists. Specifically, top five largest connected components were detected as the representative sub-regions for detailed reviews (31). The misidentifications were recorded for quantitative assessment metrics calculation.

Slide level feature summarization

To enable association analysis, TSRs were summarized to abstract slide level descriptors (**Figures 1D,E**). Since the ROI (tumor-stroma interface area) size varies from case to case, we used mean and standard deviation to denote slide level characteristics. Normalized distributions of TSR scores were calculated by counting TSR scores of each image patch within tumor-stroma interface regions. The entire assembled workflow was generalized to all the slides in our cohort. The summarized features were prepared for association analysis.

Downstream association analysis

The summarized TSR characteristics were associated with clinical and molecular information. In our work, only HGSOCs were included for downstream analysis.

Clinical associations

In HGSOC cases, for each TSR score (Fibrosis, Cellularity, and Orientation), median split was used to divide patients into two groups (i.e., score-high and low) to facilitate categorical comparisons. For univariate and multivariable [adjusted for age, FIGO stage (IV vs. III), and residual tumor after primary debulking surgery] survival analysis, a Cox proportional hazards regression model was used, and hazard ratios (HRs) and associated 95% confidence intervals (CIs) were estimated. All statistical tests were two-sided, and a *P*-value of less than 0.05 was considered statistically significant.

Molecular associations

Tumor gene expression profiles were measured using Agilent Whole Human Genome 4x44K Expression Arrays and processed as previously described (2, 4). For gene-level association analysis, normalized expression levels of each gene were correlated with each TSR score from the same tumors using Spearman Rank correlation. For over-representation pathway analysis purposes, genes with positive and negative correlations with each TSR score (nominal p-value < 0.05) were analyzed using DAVID bioinformatics tool (32, 33), to reveal pathways statistically enriched in correlated gene sets. False discovery rates were computed to correct for multiple hypothesis testing.

Results

Tumor-stroma segmentation

The developed tumor stroma segmentation model identified the tumor vs. stroma region within both HGSOC and

SBOT WSIs (**Figure 2A**). **Figure 2B** demonstrates the whole slide level segmentation results by stitching patch-level segmentation results together according to patch locations. Different tissue types were colored according to predicted categories. More examples of slide-level segmentations were shown in **Supplementary Figure 3**.

In the hold-out testing subset (N = 1,110) reserved for segmentation accuracy assessment, DSC, IoU and AP achieved 93.5, 88.65, and 95.34%, respectively (**Figure 3A**). Moreover, we observed that IoU and DSC dramatically decreased if there was a tissue type misdetection (**Figure 3B**). Our evaluation also suggested that IoU and DSC were highly correlated to each other (**Figure 3C**), and AP could be a more suitable metric for measuring our segmentation accuracy. Since the hold-out testing set is from annotated ROIs, our evaluation results suggested that the trained tumor-stroma segmentation model performed well in intrinsic cases.

Based on the criteria mentioned in the methods, 15 unannotated slides were identified, from which 615 image patches were sampled for independent tumor-stroma segmentation and TSR score evaluation. By analyzing the review records from the extrinsic evaluation tool (**Supplementary Figure 2**), our model achieved 90.6% accuracy, indicating that pathologists generally agreed with our tumor-stroma segmentation performance within independent WSIs. It is noteworthy that our trained model can be applied to the entire research cohort to generate tumor-stroma segmentation across the whole slide for the downstream analytical steps.

Based on our tumor-stroma segmentation results, some WSIs with low stroma tissue areas were identified. By checking the original images of these cases within QuPath, the pathologists confirmed that our tumor-stroma segmentation results were accurate, as the tumor islands occupied the majority of these slides, while the stromal areas consisted of a significant number of adipocytes, necrosis, and/or hemorrhage, with minimal collagenous and cellular stroma [**Supplementary Figure 4** case (4)].

Tumor-stroma reaction scoring and evaluations

Using our trained models, three TSR scores were predicted for each patch inside the stroma regions based on the tumor stoma segmentation results. The spatial overview of slide-level TSR was reconstructed by mapping three TSR scores to different colors with ranked saturation (**Figure 2C**). More examples of slide level segmentations are shown in **Supplementary Figure 4**. More zoomed in results can be found in our GitHub repository.

Our qualitative results suggested that the predicted TSR scores were not even in stroma regions of all HGSOC. Heterogeneity between regions and cases were high, as shown in **Figure 2C** and **Supplementary Figure 4**. Confusion matrices were calculated to quantitatively evaluate model performances on the hold-out testing dataset (N = 2,249). The results indicated that our model achieved accurate TSR score estimation, especially in predicting fibrosis (>90%), as shown in **Figure 4A**. In the extrinsic evaluation, with an average precision over 82.8%, the results suggested that the trained model can be generalized to our research cohort for objective TSR scoring.

We also observed false positives (TSR scores > 0) in some region from three SBOT cases, as illustrated **Supplementary Figure 4** case (3). By mapping the TSR scores back to the WSIs and observing with high resolution, we identified that these false-positive predictions were presumably due to these slides having mostly non-neoplastic ovarian stroma, which inherently has a relatively cellular and fibrotic composition. Since our model is mainly trained on annotated HGSOC regions, the trained model did not capture texture patterns within normal SBOT regions.

Violin plots illustrated the distributions of TSR score ratios per case. **Figure 4B** indicates that the majority of image patches had low TSR scores (TSR = 0), regardless of the diagnosis being SBOT or HGSOC. However, compared to SBOT cases, HGSOC cases were more likely to have higher TSR scores (>1), especially for fibrosis score. We observed a significant proportion of image patches from HGSOC cases had fibrosis TSR scores of 1. After checking the training dataset, we confirmed that half (3,339 out of 6,743) of the annotated images of HGSOC cases had moderate fibrosis TSR scores, indicating that the ambiguity of fibrosis scores could be high, and our TSR scoring model was trained to match pathologists' interpretations.

Identification of tumor-stroma interface regions

Our tumor-stroma interface region identification strategy identified five regions within each testing slide (**Supplementary Figure 5**). The proposed interface regions were localized and overlayed to the WSIs. According to pathologists' manual review, 83.4% (207 out of 250) proposed tumor-stroma regions were confirmed to be tumor-stroma interface area. After checking the falsely proposed tumor-stroma interface regions, we found flaws of tumor-stroma segmentation in those regions to be responsible for the failure, indicating that the tumor-stroma interface identification relies heavily on tumor-stroma segmentation results. To facilitate replication of our work, all the code for this paper is public available on GitHub.[1] The pretrained models for tumor-stroma segmentation

[1] https://github.com/smujiang/TumorStromaReaction

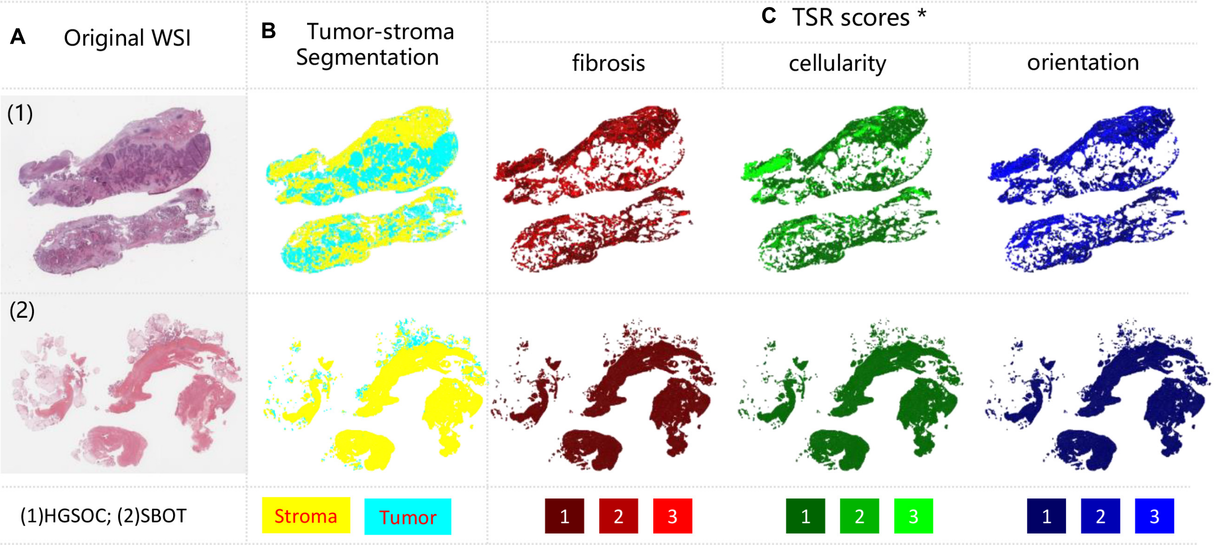


FIGURE 2
Examples of tumor-stroma segmentation and TSR scoring results. **(A)** Original WSIs, with HGSOC and SBOT each; **(B)** tumor-stroma segmentation, tumor and stroma were encoded with cyan and yellow; **(C)** TSR scores measured from three metrics, including fibrosis (Red), cellularity (Green) and orientation (Blue). Each metric was encoded from dark to light color, denoting TSR score from low to high. *For better visualization, TSR scores within all stroma regions were shown, but only the tumor-stroma interface regions were included for analysis.

and tumor-stroma reaction prediction can be obtained *via* contacting authors.

Tumor-stroma reaction clinical and molecular associations

All three TSR scores were significantly elevated in HGSOC cases vs. SBOT cases ($p < 0.001$, Figure 5A). Moreover, in HGSOC cases, higher fibrosis score (>median) was significantly associated with worse survival ($p = 0.02$; Figure 5B), and the prognostic association remained significant ($p = 0.04$; Figure 5C) after multivariate adjusting for other established prognostic factors (age at diagnosis, stage, and residual tumor after surgical debulking). In order to gain further insight into possible molecular mechanisms associated with each TSR score, gene-level correlations were computed between mRNA level of each gene and TSR score from the same tumors; and significant associations were found in two correlations: (1) correlation between fibrosis and molecular findings, and (2) correlation between cellularity and molecular findings (Figure 5D). Further genetic analysis suggested different molecular bases between the three TSR scores. Through pathway enrichment analysis, genes positively correlated with TSR-fibrosis score were found to be enriched in 14 KEGG pathways [FDR (false discovery rate) < 5%], which are mostly associated with cancer signaling aberrations. On the other hand, genes positively correlated with the TSR-orientation score were enriched in 79 KEGG pathways, with leading significant pathways implicated with immune response (Supplementary Table 1). In contrast, genes having positive correlations with TSR-cellularity score were only significantly enriched with one KEGG pathway (hsa01100: Metabolic pathways; FDR = 0.04). Detailed molecular association results including gene- and pathway-level results were shown in Supplementary Table 1.

Discussion

In this study focusing on digital analysis of reactions between tumor and stroma, combined with a critical pathology review using subjective scoring systems, we demonstrated highly concordant computational prediction based on VGG16 DL structure with training annotations and independent validations by multiple pathologists in HGSOC. Conventionally, TSR is often lumped as an overall subjective assessment by pathologists. Herein we further dissected it into three essential aspects of TSR and examined their individual and combined significance by digital detection. With a series of digital detection and quantification procedures including tumor-stroma interface area detections, the trained DL model has been successfully generalized to a large OC cohort from a single institution consisting of nearly 300 patients with long-term clinical follow-up and tumor transcriptome data. Interestingly, among the three aspects of TSR, the data revealed significant prognosis association only with fibrosis

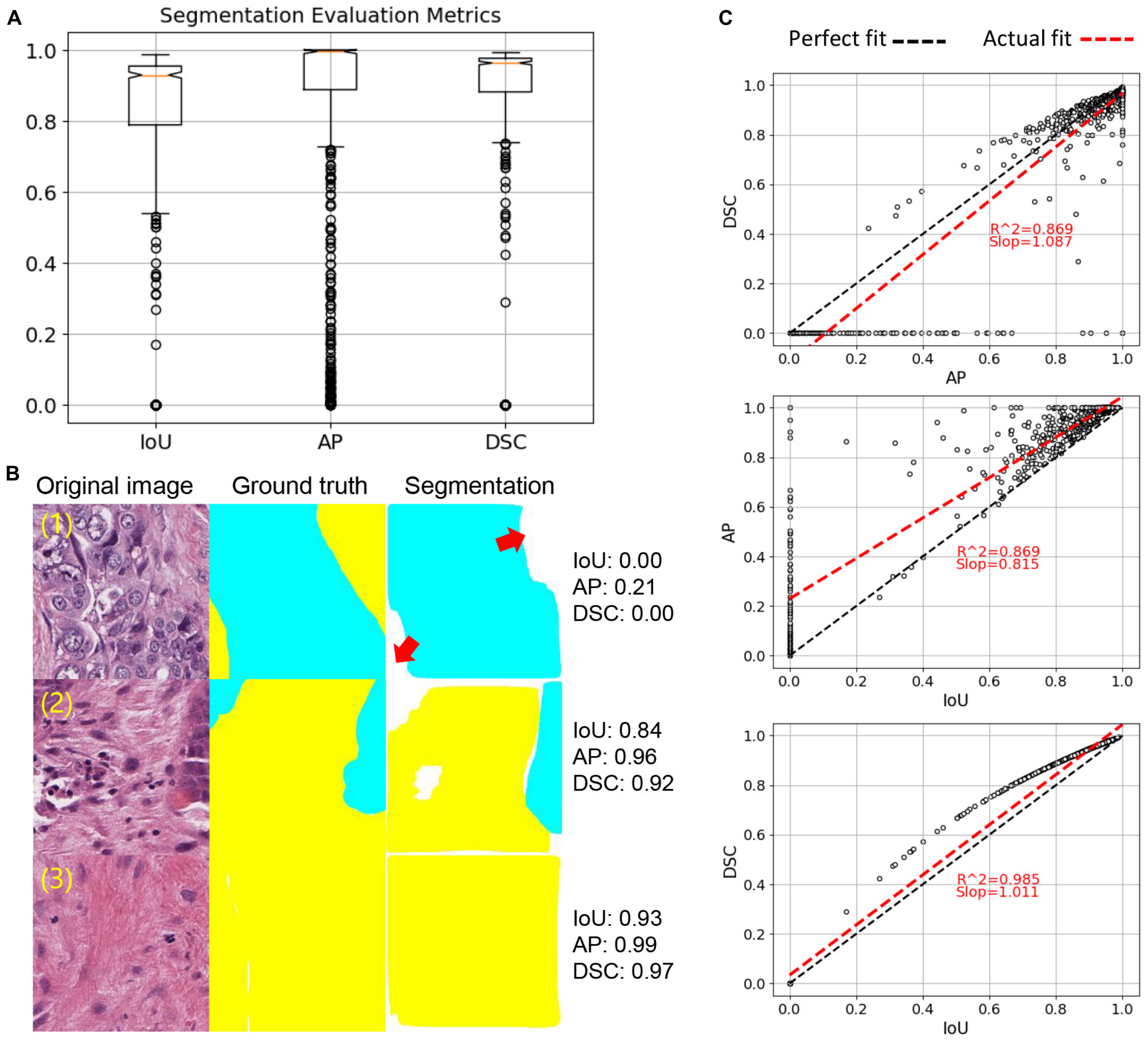


FIGURE 3
Tumor-stroma segmentation evaluation. **(A)** Boxplot of three evaluation metrics, including IoU, AP and DSC. **(B)** Examples of segmentation. Red arrows point to missed targets in segmentation. **(C)** Correlation of three evaluation metrics. Each dot represents an image sample. Linear regression was used to calculate correlation.

score. This is the first study demonstrating the outstanding significance of fibrosis over other TSR pathological features in HGSOC, indicating these features did not carry the equal weight regarding clinical significance. At the design phase of this study, the orientation score was selected as one of the three quantification criteria, mainly to further characterize the fibroblastic and/or myofibroblastic proliferation. In fact, collagen fiber organization has been associated with prognosis in breast cancer in the literature and DL approaches have been employed to quantify this as a histomorphometric feature (14, 34). However, our team observed this criterion to be a highly subjective one. In addition, the orientation score did not reveal any significant prognosis associations in our transcriptome association analysis, calling the usefulness of this criterion with regard to ovarian cancer TSR assessment into question. These observations warrant further study on individual components of the TSR, as well as aberrant gene- and pathway-level activities associated with different digital TSR scores.

Of note, the design of the TSR scoring in our study was not specific or limited to HGSOC, and the same histological principle can be applied to in majority types of solid cancers. Therefore, the digital platform developed by our study can be potentially generalized and applied to various tumor types. These findings highlight potentials of powerful DL approaches to generalize digital pathology-based predictions for large-scale translational research and enable molecular discoveries to better understand tumorigenesis and cancer progression.

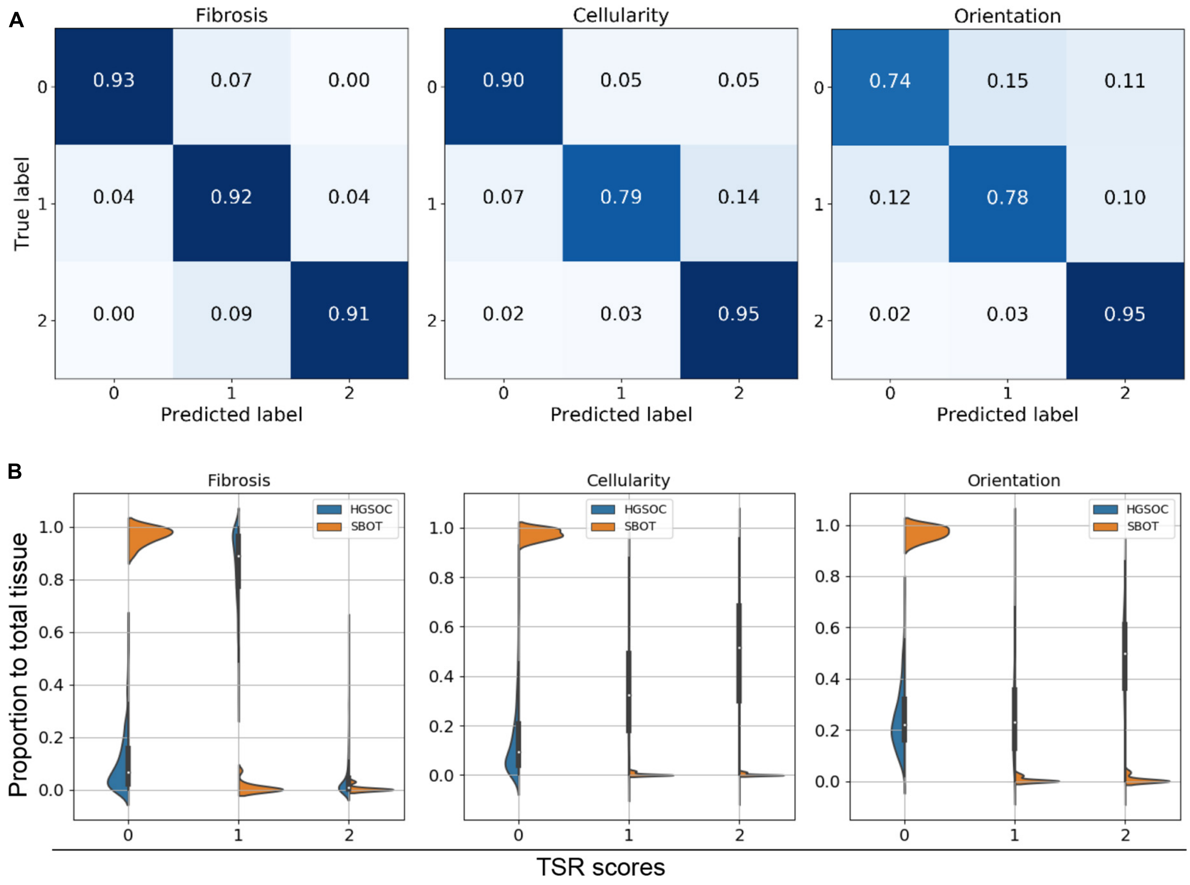
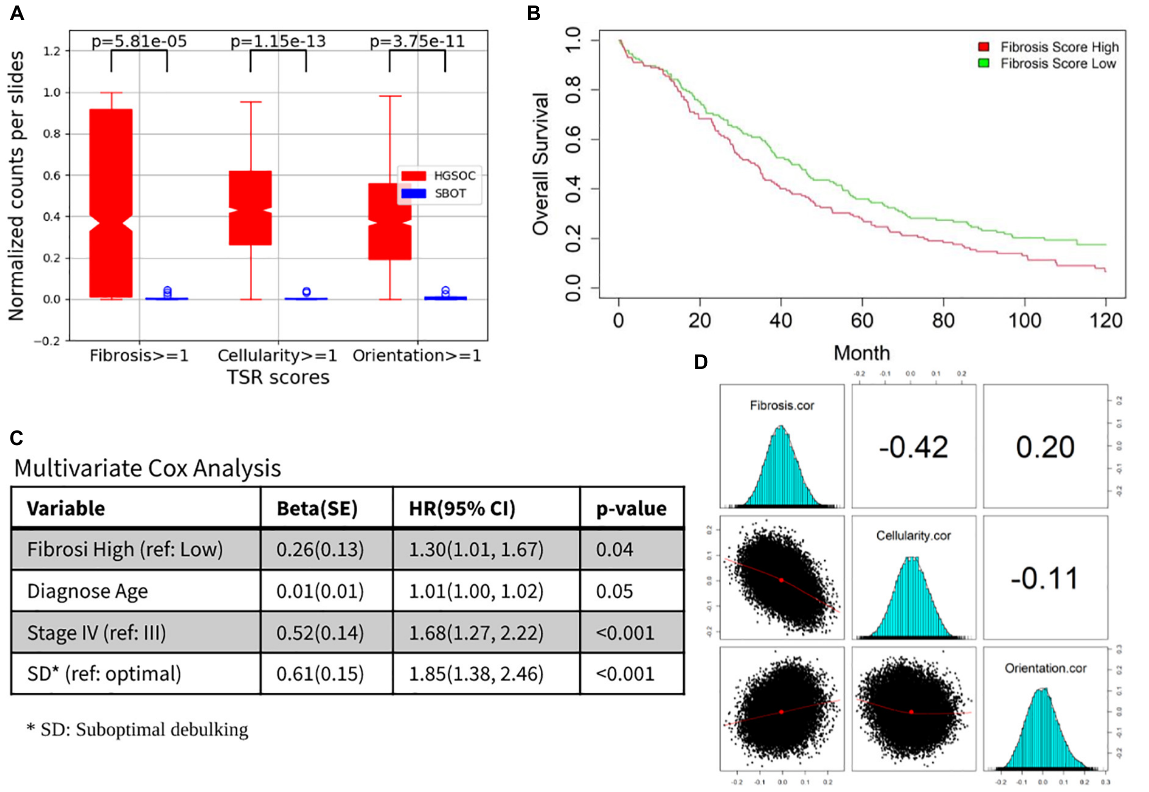


FIGURE 4
Tumor-stroma reaction (TSR) scoring evaluation. **(A)** Confusion matrix of intrinsic evaluation. **(B)** Violine plot for three TSR metrics within HGSOC vs. SBOT. Majority of SBOT images have low TSR score, no matter in which metric.

To consolidate our discoveries, we considered including other pre-existing cases into our research cohort. However, we found that images scanned at different times could be dramatically different in hue even if they were from the same patient, same institution (Mayo Clinic) and shared the same staining and image acquisition protocol. This inconsistency may be due to multiple technical variables, including scanner settings and/or age of the H&E-stained slides. Since these batch effects in pathology image data could be hidden variables in deep learning digital pathology that compromise the accuracy of classification systems (35, 36), we opted for not including our previous HGSOC data into this research. Many previous studies introduced color normalization methods to minimize staining inconsistencies (29, 37). Though it is hard to measure the preservation of diagnostic information after image transformation, many integrative studies investigating cancer subtype classification and prognosis association achieved optimistic performances by introducing image normalization (38, 39). From this point of view, color normalization could be beneficial for assorted research cohort from miscellaneous data sources, especially from multiple institutions. Meanwhile, we also noticed that some investigators improved their model performance by synthesizing images using generative adversarial network (GAN) (40, 41), which could be another potential way to enhance the generalizability of our TSR estimation model.

Although our tumor-stroma interface region detection relies on patch-level tumor-stroma segmentation, the strategy we introduced (**Supplementary Figure 5**) could partially offset this limitation. To generate a slide-level overview of tissue context (tumor vs. stroma) for image morphological manipulations, patch-level tumor-stroma segmentation results were down-sampled and stitched back to their original locations. In this process, the tissue type (tumor or stroma) in slide-level was determined by the dominated component of patch-level segmentation. In other words, for tumor-stroma interface region detection task, tumor-stroma segmentation results were not required to achieve pixel-level accuracy.



Multivariate Cox Analysis

Variable	Beta(SE)	HR(95% CI)	p-value
Fibrosi High (ref: Low)	0.26(0.13)	1.30(1.01, 1.67)	0.04
Diagnose Age	0.01(0.01)	1.01(1.00, 1.02)	0.05
Stage IV (ref: III)	0.52(0.14)	1.68(1.27, 2.22)	<0.001
SD* (ref: optimal)	0.61(0.15)	1.85(1.38, 2.46)	<0.001

* SD: Suboptimal debulking

FIGURE 5

Prognosis and molecular associations of fibrosis score. **(A)** Tumor-stroma reaction (TSR) score boxplots for HGSOC and SBOT groups. **(B)** Overall survival differences between fibrosis high vs. low. **(C)** The prognostic association for other established prognostic factors (age at diagnosis, stage and residual tumor after surgical debulking). **(D)** Correlation between fibrosis/cellularity/orientation and molecular findings.

We observed that our TSR scoring models highlighted some regions with a potentially high TSR in five SBOT cases, for example, case (3) in **Supplementary Figure 4**. The main reason contributing to this flaw is that our models were not trained to differentiate normal vs. abnormal ovarian stroma. We anticipate that our TSR scoring pipeline can achieve a better estimation if models can be trained using extra normal vs. abnormal ovarian stroma annotations. Meanwhile, we acknowledge that using TSR as the sole measurement is not enough to describe the complex tumor micro-environment (TME). It has been reported that TILs can also be assessed with the aid of digital pathology in advanced-stage, HPV-negative head and neck tumors (42). We will introduce more interpretable measurements for pathology image metadata summarization, which will bring more opportunities for novel discoveries. To integrate cellular level features into large cohort analysis, we also plan to introduce more advanced cell segmentation modules to our workflow for better cell level representations (43). As reported in our previous work (44), over- and under-segmentations lead to inaccurate downstream analysis impute to erroneous features calculated based on them.

Another imperfection of our study is way we summarize predicted TSR score to slide level for association analysis. For the sake of simplicity, we employed simple statistics and assigned equal weights to patch level TSR predictions. However, the relative importance of tissue regions contributing to the diagnosis could be dramatically different depending on tissue context. In this study, we introduced tumor-stroma interface area identification methods that were aimed at mimicking pathologists' diagnoses. This strategy is simple and works well in most cases; however, it highly depends on the tumor-stroma segmentation accuracy. We noticed that some studies proposed to introduce multi-resolution analysis for capturing subtle tissue features within different WSI scales (45, 46). Attention based deep neural networks (47) are also tangible options for locating diagnostically relevant regions and assign those regions with higher weights in automatic analysis. We will consider these solutions to fill the gaps in our current work.

Further limitations of the current study include that all the samples were from a single institution, requiring further validations in external cohorts. H&E imaging-based digital pathology studies as such may be also affected by paraffin block preservation protocols and digital scanning parameter settings, which could lead to model differences when generalizing to WSI samples collected and scanned following different protocols. Computational developments and evaluations will be made to address these challenges.

Conclusion

Our developed system achieved encouraging performances in tissue segmentation and TSR score predictions and generalized successfully to a large single-institution OC cohort, resulting in novel discoveries of clinical prognosis associations and molecular findings implicated in different TSR scores.

Data availability statement

The original contributions presented in this study are publicly available. The source code can be found here: https://github.com/smujiang/TumorStromaReaction. The pre-trained model can be obtained by contacting authors.

Ethics statement

The studies involving human participants were reviewed and approved by Mayo Clinic Institutional Review Board. The patients/participants provided their written informed consent to participate in this study. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

RG and YH conceived of the presented histological concept. BT, LY, and RG defined data annotation criteria, performed annotation, and validated model performance. JJ and CW conceived and performed majority of experiments and manuscript drafting. SA conducted cohort characteristic statistics. SW contributed to the final version of the manuscript. HL provided funding resources and supervised the deep learning parts of this project. EG also supervised the funding of this work. All authors discussed the results and contributed to the final manuscript.

Funding

This research was partially supported by the Mayo Clinic Ovarian Cancer SPORE (P50 CA136393), National Institutes of Health's National Institute (R01EB019403), Mayo Clinic Center for Individualized Medicine, and Mayo Clinic Department of Laboratory Medicine and Pathology, Division of Anatomic Pathology Intramural Research Fund.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.994467/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Tumor-stroma reaction (TSR) annotation. **(A)** Original image within ROI selected for annotation. **(B)** Annotated ROIs. Polygons were used to label regions with different TSR scores. **(C)** Legend of three TSR score measurements. **(D–F)** Parsed annotations. TSR scores were encoded into R/G/B colors to represent three measurements (fibrosis, cellularity, and orientation), respectively; Panels **(G,H)** are two zoom in examples. Panel **(G)** was annotated as Fibrosis = 2, Cellularity = 1, Orientation = 1; Panel **(H)** was annotated as Fibrosis = 1, Cellularity = 2, Orientation = 2.

SUPPLEMENTARY FIGURE 2
Interactive tool for extrinsic evaluation. Source code available in our GitHub. Buttons and checkboxes on the right are clickable, pathologists' interactions were recorded for extrinsic evaluation.

SUPPLEMENTARY FIGURE 3
Extra examples of tumor-stroma segmentation results, including five HGSOCs, five SBOTs and their tumor stroma segmentation results.

SUPPLEMENTARY FIGURE 4
Extra examples (two HGSOC and two SBOT) of TSR scoring results. TSR scores measured with fibrosis (Red), cellularity (Green), and orientation (Blue). From dark to light, TSR scores were encoded into R/G/B colors. *For better visualization, TSR scores within all stroma regions were shown, but only the tumor-stroma interface regions were included for analysis.

SUPPLEMENTARY FIGURE 5
Tumor-stroma interface area identification. **(A)** Original WSI and tumor-stroma segmentation results. **(B)** Morphological and logical operations were conducted on tumor-stroma segmentation for localizing tumor-stroma interface regions. **(C)** Proposed ROIs (red rectangles) for TSR score summarization.

SUPPLEMENTARY TABLE 1
Detailed molecular association results, including top 10 genes positively/negatively associated with TSR-Fibrosis score, and pathways positively associated with Fibrosis and Orientation scores, respectively.

References

1. Lisio M-A, Fu L, Goyeneche A, Gao Z-H, Telleria C. High-grade serous ovarian cancer: basic sciences, clinical and therapeutic standpoints. *Int J Mol Sci.* (2019) 20:952. doi: 10.3390/ijms20040952

2. Wang C, Armasu SM, Kalli KR, Maurer MJ, Heinzen EP, Keeney GL, et al. Pooled clustering of high-grade serous ovarian cancer gene expression leads to novel consensus subtypes associated with survival and surgical outcomes. *Clin Cancer Res.* (2017) 23:4077–85. doi: 10.1158/1078-0432.CCR-17-0246

3. Talhouk A, George J, Wang C, Budden T, Tan TZ, Chiu DS, et al. Development and validation of the gene expression predictor of high-grade serous ovarian carcinoma molecular subTYPE (PrOTYPE). *Clin Cancer Res.* (2020) 26:5411–23. doi: 10.1158/1557-3265.OVCA19-A03

4. Konecny GE, Wang C, Hamidi H, Winterhoff B, Kalli KR, Dering J, et al. Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *J Natl Cancer Inst.* (2014) 106:dju249. doi: 10.1093/jnci/dju249

5. Murakami R, Matsumura N, Mandai M, Yoshihara K, Tanabe H, Nakai H, et al. Establishment of a novel histopathological classification of high-grade serous ovarian carcinoma correlated with prognostically distinct gene expression subtypes. *Am J Pathol.* (2016) 186:1103–13. doi: 10.1016/j.ajpath.2015.12.029

6. Ohtani H. Stromal reaction in cancer tissue: pathophysiologic significance of the expression of matrix-degrading enzymes in relation to matrix turnover and immune/inflammatory reactions. *Pathol Int.* (1998) 48:1–9. doi: 10.1111/j.1440-1827.1998.tb03820.x

7. Sirica AE, Gores GJ. Desmoplastic stroma and cholangiocarcinoma: clinical implications and therapeutic targeting. *Hepatology (Baltimore, Md).* (2014) 59:2397. doi: 10.1002/hep.26762

8. Wang LM, Silva MA, D'Costa Z, Bockelmann R, Soonawalla Z, Liu S, et al. The prognostic role of desmoplastic stroma in pancreatic ductal adenocarcinoma. *Oncotarget.* (2016) 7:4183. doi: 10.18632/oncotarget.6770

9. Busam KJ, Mujumdar U, Hummer AJ, Nobrega J, Hawkins WG, Coit DG, et al. Cutaneous desmoplastic melanoma: reappraisal of morphologic heterogeneity and prognostic factors. *Am J Surg Pathol.* (2004) 28:1518–25. doi: 10.1097/01.pas.0000141391.91677.a4

10. Davidson B, Trope CG, Reich R. The role of the tumor stroma in ovarian cancer. *Front Oncol.* (2014) 4:104. doi: 10.3389/fonc.2014.00104

11. Bremnes RM, Dønnem T, Al-Saad S, Al-Shibli K, Andersen S, Sirera R, et al. The role of tumor stroma in cancer progression and prognosis: emphasis on carcinoma-associated fibroblasts and non-small cell lung cancer. *J Thorac Oncol.* (2011) 6:209–17. doi: 10.1097/JTO.0b013e3181f8a1bd

12. Freeman MR, Li Q, Chung LW. Can stroma reaction predict cancer lethality? *Clin Cancer Res.* (2013) 19:4905–7. doi: 10.1158/1078-0432.CCR-13-1694

13. Ueno H, Ishiguro M, Nakatani E, Ishikawa T, Uetake H, Murotani K, et al. Prognostic value of desmoplastic reaction characterisation in stage II colon cancer: prospective validation in a Phase 3 study (SACURA Trial). *Br J Cancer.* (2021) 124:1088–97. doi: 10.1038/s41416-020-01222-8

14. Li HJ, Bera K, Toro P, Fu PF, Zhang ZL, Lu C, et al. Collagen fiber orientation disorder from H&E images is prognostic for early stage breast cancer: clinical trial validation. *Npj Breast Cancer.* (2021) 7:104. doi: 10.1038/s41523-021-00310-z

15. Geessink OGF, Baidoshvili A, Klaase JM, Ehteshami Bejnordi B, Litjens GJS, van Pelt GW, et al. Computer aided quantification of intratumoral stroma yields an independent prognosticator in rectal cancer. *Cell Oncol.* (2019) 42:331–41. doi: 10.1007/s13402-019-00429-z

16. Failmezger H, Muralidhar S, Rullan A, de Andrea CE, Sahai E, Yuan Y. Topological tumor graphs: a graph-based spatial model to infer stromal recruitment for immunosuppression in melanoma histology. *Cancer Res.* (2020) 80:1199–209. doi: 10.1158/0008-5472.CAN-19-2268

17. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep.* (2017) 7:1–7. doi: 10.1038/s41598-017-17204-5

18. Jiang J, Tekin B, Guo R, Liu H, Huang Y, Wang C. Digital pathology-based study of cell-and tissue-level morphologic features in serous borderline ovarian tumor and high-grade serous ovarian cancer. *J Pathol Inform.* (2021) 12:24. doi: 10.4103/jpi.jpi_76_20

19. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: (2017). p. 2961–9. doi: 10.1109/ICCV.2017.322

20. Wang S, Rong R, Yang DM, Fujimoto J, Yan S, Cai L, et al. Computational staining of pathology images to study the tumor microenvironment in lung cancer. *Cancer Res.* (2020) 80:2056–66. doi: 10.1158/0008-5472.CAN-19-1629

21. Mulay S, Ram K, Sivaprakasam M, Vinekar A. Early detection of retinopathy of prematurity stage using deep learning approach. *Proceedings of the SPIE 10950, Medical Imaging 2019: Computer-Aided Diagnosis*. San Diego, CA: SPIE (2019). p. 758–64. doi: 10.1117/12.2512719

22. Quoc TTP, Linh TT, Minh TNT. Comparing U-Net convolutional network with mask R-CNN in agricultural area segmentation on satellite images. *Proceedings of the 2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*. Ho Chi Minh: IEEE (2020). p. 124–9. doi: 10.1109/NICS51282.2020.9335856

23. Durkee MS, Abraham R, Ai J, Fuhrman JD, Clark MR, Giger ML. Comparing Mask R-CNN and U-Net architectures for robust automatic segmentation of immune cells in immunofluorescence images of Lupus Nephritis biopsies. *Proceedings of the Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissue*. Bellingham, WA: SPIE (2021). p. 109–15. doi: 10.1117/12.2577785

24. Abdulla W. *Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow*. San Francisco, CA: GitHub repository (2017).

25. Jiang J, Wu Y, Huang M, Yang W, Chen W, Feng Q. 3D brain tumor segmentation in multimodal MR images based on learning population-and patient-specific feature sets. *Comput Med Imaging Graph.* (2013) 37:512–21. doi: 10.1016/j.compmedimag.2013.05.007

26. Zhou D, Fang J, Song X, Guan C, Yin J, Dai Y, et al. Iou loss for 2d/3d object detection. *Proceedings of the 2019 International Conference on 3D Vision (3DV)*. Quebec City, QC: IEEE (2019). p. 85–94. doi: 10.1109/3DV.2019.00019

27. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint] (2014):doi: 10.48550/arXiv.1409.1556

28. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* (2018) 23:181.–193.

29. Jiang J, Prodduturi N, Chen D, Gu Q, Flotte T, Feng Q, et al. Image-to-image translation for automatic ink removal in whole slide images. *J Med Imaging.* (2020) 7:057502. doi: 10.1117/1.JMI.7.5.057502

30. Gil JY, Kimmel R. Efficient dilation, erosion, opening, and closing algorithms. *IEEE Trans Pattern Analy Mach Intellig.* (2002) 24:1606–17. doi: 10.1109/TPAMI.2002.1114852

31. Dundar MM, Badve S, Bilgin G, Raykar V, Jain R, Sertel O, et al. Computerized classification of intraductal breast lesions using histopathological images. *IEEE Trans Biomed Eng.* (2011) 58:1977–84. doi: 10.1109/TBME.2011.2110648

32. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protocols.* (2009) 4:44–57. doi: 10.1038/nprot.2008.211

33. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* (2009) 37:1–13. doi: 10.1093/nar/gkn923

34. Bredfeldt JS, Liu Y, Conklin MW, Keely PJ, Mackie TR, Eliceiri KW. Automated quantification of aligned collagen for human breast carcinoma prognosis. *J Pathol Inform.* (2014) 5:28. doi: 10.4103/2153-3539.139707

35. Fei T, Zhang T, Shi W, Yu T. Mitigating the adverse impact of batch effects in sample pattern detection. *Bioinformatics.* (2018) 34:2634–41. doi: 10.1093/bioinformatics/bty117

36. Schmitt M, Maron RC, Hekler A, Stenzinger A, Hauschild A, Weichenthal M, et al. Hidden variables in deep learning digital pathology and their potential to cause batch effects: prediction model study. *J Med Intern Res.* (2021) 23:e23436. doi: 10.2196/23436

37. Zheng Y, Jiang Z, Zhang H, Xie F, Shi J, Xue C. Adaptive color deconvolution for histological WSI normalization. *Comput Methods Prog Biomed.* (2019) 170:107–20. doi: 10.1016/j.cmpb.2019.01.008

38. Boschman J, Farahani H, Darbandsari A, Ahmadvand P, Van Spankeren A, Farnell D, et al. The utility of color normalization for AI-based diagnosis of hematoxylin and eosin-stained pathology images. *J Pathol.* (2022) 256:15–24. doi: 10.1002/path.5797

39. Van Eycke Y-R, Allard J, Salmon I, Debeir O, Decaestecker C. Image processing in digital pathology: an opportunity to solve inter-batch variability of immunohistochemical staining. *Sci Rep.* (2017) 7:42964. doi: 10.1038/srep42964

40. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol.* (2019) 20:e253–61. doi: 10.1016/S1470-2045(19)30154-8

41. Hou L, Agarwal A, Samaras D, Kurc TM, Gupta RR, Saltz JH. Robust histopathology image analysis: to label or to synthesize? *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA: (2019). p. 8533–42. doi: 10.1109/CVPR.2019.00873

42. de Ruiter EJ, de Roest RH, Brakenhoff RH, Leemans CR, de Bree R, Terhaard CH, et al. Digital pathology-aided assessment of tumor-infiltrating T lymphocytes in advanced stage, HPV-negative head and neck tumors. *Cancer Immunol Immunother.* (2020) 69:581–91. doi: 10.1007/s00262-020-02481-3

43. Weigert M, Schmidt U, Haase R, Sugawara K, Myers G. "Star-convex polyhedra for 3d object detection and segmentation in microscopy," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Los Alamitos, CA: IEEE Computer Society) (2020). 3666–73. doi: 10.1109/WACV45572.2020.9093435

44. Jiang J, Tekin B, Guo R, Liu H, Huang Y, Wang C. Digital pathology-based study of cell-and tissue-level morphologic features in serous borderline ovarian tumor and high-grade serous ovarian cancer. *arXiv* [Preprint] (2020). doi: 10.48550/arXiv.2008.12479

45. Van Rijthoven M, Balkenhol M, Siliòa K, Van Der Laak J, Ciompi F. HookNet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med Image Analy.* (2021) 68:101890. doi: 10.1016/j.media.2020.101890

46. Marini N, Otálora S, Ciompi F, Silvello G, Marchesin S, Vatrano S, et al. Multi-Scale task multiple instance learning for the classification of digital pathology images with global annotations. *Proc Mach Learn Res.* (2021) 156:170–81.

47. Sornapudi S, Stanley RJ, Stoecker WV, Long R, Xue Z, Zuna R, et al. DeepCIN: attention-based cervical histology image classification with sequential feature modeling for pathologist-level accuracy. *J Pathol Inform.* (2020) 11:40. doi: 10.4103/jpi.jpi_50_20

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fmed.2022.965429

Association of adenosine signaling gene signature with estrogen receptor-positive breast and prostate cancer bone metastasis

Daniel Brian Shropshire[1†], Francisca M. Acosta[1†], Kun Fang[2], Jaime Benavides[1,3], Lu-Zhe Sun[4], Victor X. Jin[2] and Jean X. Jiang[1*]

[1]Department of Biochemistry and Structural Biology, University of Texas Health Science Center, San Antonio, TX, United States, [2]Division of Biostatistics and MCW Cancer Center, Medical College of Wisconsin, Milwaukee, WI, United States, [3]Department of Biomedical Engineering and Chemical Engineering, The University of Texas at San Antonio, San Antonio, TX, United States, [4]Department of Cell Systems and Anatomy, University of Texas Health Science Center, San Antonio, TX, United States

Check for updates

OPEN ACCESS

EDITED BY
Jun Cheng,
Shenzhen University, China

REVIEWED BY
Xiaohui Zhan,
Chongqing Medical University, China
Fuhai Li,
Washington University in St. Louis,
United States

*CORRESPONDENCE
Jean X. Jiang
jiangj@uthscsa.edu

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Pathology, a section of the journal Frontiers in Medicine

RECEIVED 09 June 2022
ACCEPTED 23 August 2022
PUBLISHED 15 September 2022

CITATION
Shropshire DB, Acosta FM, Fang K, Benavides J, Sun L-Z, Jin VX and Jiang JX (2022) Association of adenosine signaling gene signature with estrogen receptor-positive breast and prostate cancer bone metastasis. *Front. Med.* 9:965429.
doi: 10.3389/fmed.2022.965429

Bone metastasis is a common and devastating consequence of several major cancer types, including breast and prostate. Osteocytes are the predominant bone cell, and through connexin (Cx) 43 hemichannels release ATP to the bone microenvironment that can be hydrolyzed to adenosine. Here, we investigated how genes related to ATP paracrine signaling are involved in two common bone-metastasizing malignancies, estrogen receptor positive (ER^+) breast and prostate cancers. Compared to other sites, bone metastases of both cancer types expressed higher levels of ENTPD1 and NT5E, which encode CD39 and CD73, respectively, and hydrolyze ATP to adenosine. ADORA3, encoding the adenosine A3 receptor, had a similar expression pattern. In primary ER^+ breast cancer, high levels of the triplet ENTPD1/NT5E/ADORA3 expression signature was correlated with lower overall, distant metastasis-free, and progression-free survival. In ER^+ bone metastasis biopsies, this expression signature is associated with lower survival. This expression signature was also higher in bone-metastasizing primary prostate cancers than in those that caused other tumor events or did not lead to progressive disease. In 3D culture, a non-hydrolyzable ATP analog inhibited the growth of breast and prostate cancer cell lines more than ATP did. A3 inhibition also reduced spheroid growth. Large-scale screens by the Drug Repurposing Hub found ER^+ breast cancer cell lines were uniquely sensitive to adenosine receptor antagonists. Together, these data suggest a vital role for extracellular ATP degradation and adenosine receptor signaling in cancer bone metastasis, and this study provides potential diagnostic means for bone metastasis and specific targets for treatment and prevention.

KEYWORDS

metastasis, bone, breast, prostate, purinergic, osteocyte

Introduction

Bone is the most common site for distant metastasis by breast and prostate cancers and has devastating impacts on patients (1, 2). Complications include severe pain, pathologic fractures, life-threatening hypercalcemia, and spinal cord compression (3, 4). Furthermore, patients with bone metastases have poor overall prognosis and lower life expectancies (5–7). Understanding the process that permits breast and prostate cancer bone metastasis and knowing how to derail it is critical for improving patient outcomes for the second-leading cause of cancer deaths in women and men, respectively. The microenvironment of distant organs plays a vital role in the process of metastasis to that site (8). Despite this, few drugs specifically target metastatic sites. Bisphosphonates induce osteoclast apoptosis, promote osteocyte Cx43 hemichannel activity (9, 10), and are used to treat bone metastases of various types, including prostate and breast (11, 12). More uniquely, they were clinically validated to prevent breast cancer metastasis to bone in postmenopausal women (13).

Osteocytes comprise roughly 90% of bone cells and are dominant regulators of the local microenvironment (14). In normal bone physiology, they coordinate the actions of bone-building osteoblasts and bone-degrading osteoclasts (14). Osteocytes are rich in Cx43 hemichannels, through which small paracrine signaling molecules such as prostaglandins and ATP are released and influence both normal bone cells and metastatic cancer cells (15, 16). Our previous study found osteocytes expressing Cx43 with impaired hemichannel and gap junction activity promoted the growth of triple-negative breast cancer in bone, while osteocytes with impaired Cx43 gap junction but retained hemichannel function had no such effect (17). Further investigation showed that a stable extracellular ATP (eATP) analog decreased triple-negative breast cancer cell migration, while extracellular adenosine (eADO) increased it, and thus preventing eATP degradation to eADO can enhance the inhibitory effect of eATP on cancer cell migration (16).

A recent surge of interest in purinergic signaling in cancer is primarily on its role in immunology. In tumors, eATP is elevated and generally stimulates the immune system (18). This eATP can be hydrolyzed to AMP by CD39, encoded by the gene ENTPD1, and further degraded to adenosine by CD73, encoded by the gene NT5E (18). The immunosuppressive function of eADO is in part mediated by binding to T cell adenosine 2A receptors (A2ARs) (18). This rationale has led to interest in inhibiting eADO production in tumors as a way of improving outcomes alone or combined with PD1-PDL1 inhibition (19, 20). However, much less attention has been given to the non-immunologic functions of eATP and eADO in cancer development and progression. Our studies on ATP release by osteocytic hemichannels in bone and the effects of eATP and eADO signaling on triple-negative breast cancer led us to investigate whether tumor cells increase eATP hydrolysis to promote bone metastasis. We focused on estrogen receptor-positive (ER^{+}) breast cancer, which accounts for 77% of breast cancer bone metastases (1), and prostate cancer, which also primarily metastasizes to bone (2, 7).

Materials and methods

Materials

Spheroid culture plates were purchased from Corning (Corning, NY, United States; cat. 4515). ATP was purchased from Sigma Aldrich (St. Louis, MO, United States; cat. A2383). ATPγS was purchased from Fisher (Hampton, NH, United States; cat. 40-801-0). Both were dissolved in Dulbecco's phosphate buffered saline (Gibco cat. 14190). MRS-1220 (cat. 12-175) was purchased from Fisher and dissolved in dimethyl sulfoxide (Fisher cat. 67-68-5). The rest of the reagents were purchased either from Fisher or Sigma.

Cell culture, 3D culture, and quantification

MCF-7 cells were a gift from Dr. Michael Brattain maintained in Dulbecco's Modification of Eagle's Medium (DMEM) with 10% fetal bovine serum (FBS). 22Rv1 cells were a gift from Dr. Tim Huang at University of Texas Health Science Center at San Antonio and were maintained in RPMI-1640 with 10% FBS. Cells were kept in a 5% CO_2 incubator.

For 3D culture, 2,000 cells per well were seeded in ultra-low adherent U-bottom 96-well plates with drug or vehicle in DMEM with 2.5% FBS (MCF-7) or RPMI-1640 with 2.5% FBS (22Rv1). Photos were taken using a Keyence BZ-X710 microscope (Keyence, Osaka, Japan) using a 20X phase contrast objective (Nikon, Tokyo, Japan). Sphere cross-sectional area was measured using ImageJ, (21) which was used to determine volume. Statistical comparisons were made using t-test or two-way ANOVA with the Geisser-Greenhouse correction and Tukey's post-test. EC50 values were calculated in Graphpad Prism v9 using a four-parameter logistical model.

Ribonucleic acid expression in metastases, and comparison with primary tumor

Microarray datasets GSE74685, GSE14020, GSE32269, and GSE47561 were downloaded from the Gene Expression Omnibus. GSE14020 raw fluorescence CEL files were processed using BART (22). Datasets were chosen based on clinical characteristics (**Supplementary Table 1**), using workflow as shown in **Supplementary Figure 1**. Differential gene expression analysis for **Supplementary Table 2** was performed using the limma bioinformatics package (23). We compared log2-transformed data in metastatic locations containing at least 5 samples using one-way ANOVA and Dunnett's multiple comparisons test. Expression between primary and metastatic tumors was compared using a t-test. For breast cancer, the Robust Microchip Array (RMA) function in Bioconductor was used to process primary and metastatic data.

Survival analysis

Distant metastasis-free survival analysis in ER^+ breast cancers was performed on microarray data using KMPlot (24, 25). We used ER^+ patients because the first distant metastasis in these patients is usually located in bone (1). Expression data was used to predict ER status when not histologically determined. Patients were separated into high- and low-expressing tumors by median, as evenly as possible. Overall and disease-specific survival were performed using data from the TCGA BRCA (26) cohort accessed through Xena browser (27) and analyzed through KMPlot (24). Signatures were calculated by the average expression [$\log_2$(norm_count + 1)] of the three genes when noted in **Figures 1B**, **2**. Survival analysis for samples taken from established bone metastases used GSE124647 and cohorts were separated by median expression. Significance was determined by $p < 0.05$.

Gene signature and Gleason score correlation

Signature correlation was performed on ER^+ tumors in the TCGA BRCA cohort using a previously published gene set (26, 28). TCGA PRAD (29) data (counts) were downloaded through Xena browser (27). DKFZ data (counts) were downloaded from cBio Cancer Genetics Portal (30, 31). Expression signatures were the average expression of the three genes in each sample. Pearson method was used for correlation analysis. One-way ANOVA with a test for linear trend was used to find increasing averages with increasing Gleason scores and Kruskall-Wallis test with multiple comparisons for comparing signature expression between primary tumors with or without bone metastases and other events. Significance was determined by $p < 0.05$.

Drug sensitivity determination

Drug screen was performed using PRISM technique (32) by the Drug Repurposing Hub, as reported (33). Analyses were performed on the 19Q3 screen. Data were analyzed on DepMap (34) portal, which uses the Limma R statistical package (23). Significance was determined by $p < 0.0005$.

Statistics

Statistical analyses were performed on Graphpad Prism v9 unless otherwise noted. Graphs reflect mean ± SD, except DepMap screen in which boxes represent median ± 1 interquartile range and whiskers represent 5th and 95th percentiles. $^*p < 0.05$; $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$, except for DepMap screen where $p < 0.0005$ is significant.

Results

ENTPD1, NT5E, and ADORA3 show higher expression in bone metastases than in other sites of metastasis or in primary tumors

We previously identified ATP released by active hemichannels as a potential inhibitor of triple-negative breast cancer growth in bone (16, 17). Because hemichannels are rare in most tissues but are well established in bone, we hypothesized that downregulating one or more ATP receptors would enable bone metastasis, and this receptor would have lower expression levels in bone metastases compared to metastases at other locations. We investigated this in microarray gene expression data from patients with metastatic breast cancer

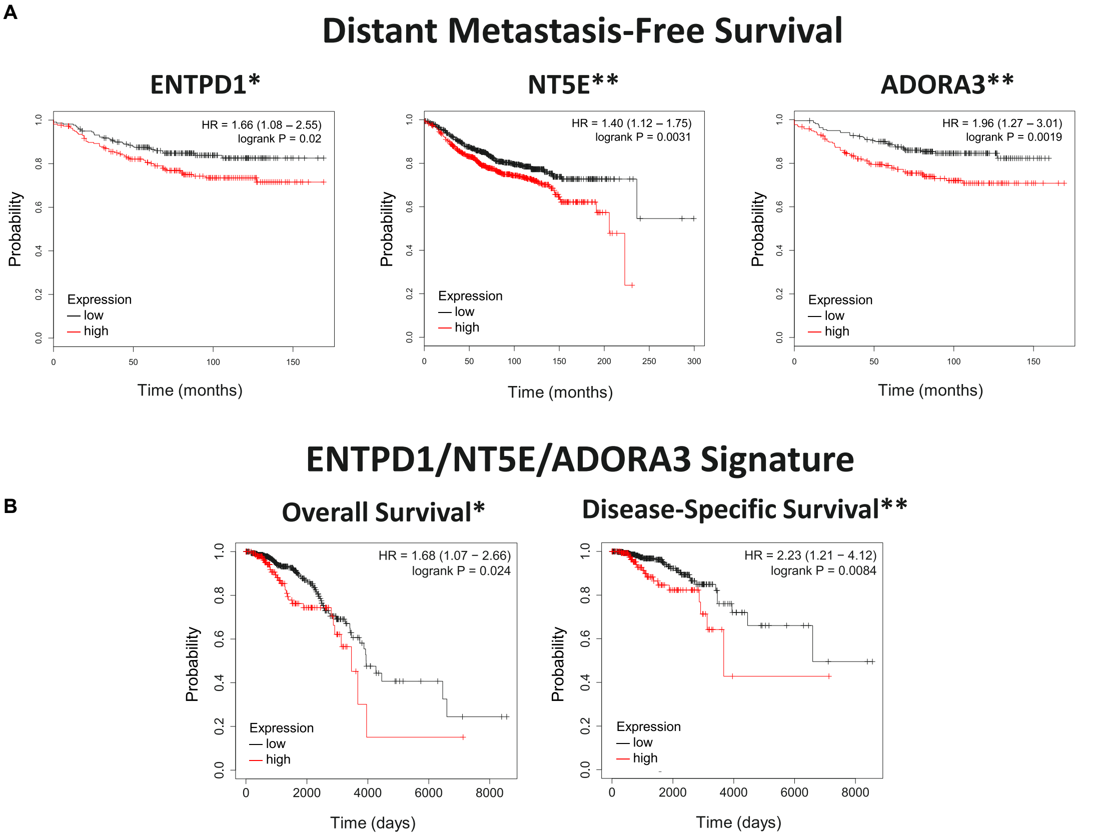


FIGURE 1
ENTPD1, NT5E, and ADORA3 expression in primary ER$^+$ breast cancer correlates with poor outcomes. **(A)** Kaplan Meier plots of distant metastasis-free survival in ER$^+$ breast cancer. The high-expression groups for ENTPD1 (HR = 1.66), NT5E (HR = 1.4), and ADORA3 (HR = 1.96) all have a significantly greater chance of distant metastasis or death. Analysis was made using KMPlot **(24)**. **(B)** Kaplan Meier plots of overall and disease-specific survival of ER$^+$ breast cancer patients in the TCGA BRCA cohort based on ENTPD1/NT5E/ADORA3 signature expression (calculated by the average expression [log2(norm_count + 1)] of the three genes, and separated by median). In this separate cohort than **(A)**, the high-expression group had significantly lower overall and disease-specific survival (HR = 1.68 and 2.23, respectively). *$p < 0.05$; **$p < 0.01$.

(GSE14020). Surprisingly, none of the ATP receptors was differentially expressed between bone and other metastatic sites (Supplementary Table 2). However, ENTPD1 and NT5E, which encode genes that degrade eATP to eADO, were more highly expressed in bone metastases than in metastases to other sites (Figure 3A),top. We next investigated which receptors are activated by the excess eADO formed by eATP hydrolysis and found increased expression of ADORA3, encoding A3R, in bone metastases (Figure 3A),top. We also analyzed gene expression in metastatic prostate cancer (GSE74685) and found a similar expression pattern, with ENTPD1, NT5E, and ADORA3 upregulation in bone metastases than in other metastases (Figure 3A), bottom.

After determining that these three genes are more highly expressed in bone metastases than in other metastases, we further compared their expression between bone metastases and primary tumors. ENTPD1, NT5E, and ADORA3 showed higher expression in bone metastases than in primary breast cancers (Figure 3B),top, GSE47561. Similarly, castrate-resistant bone metastases had higher expression of these three genes than did primary prostate cancer (Figure 3B), bottom, GSE32269. Taken together, we demonstrated that the expression of two genes that hydrolyze eATP to eADO and the eADO receptor ADORA3 are more highly expressed in bone metastases than in other metastases or in primary tumors.

High expression of ENTPD1, NT5E, and ADORA3 in primary ER$^+$ breast cancer is correlated with lower distant metastasis-free survival, overall survival, and disease-specific survival

Next, we investigated whether primary tumors with higher expression of these genes are more likely to metastasize to bone. Since bone is the site of first metastasis for the majority of patients with ER$^+$ breast cancer, (35) distant metastasis-free survival in these patients should largely reflect bone

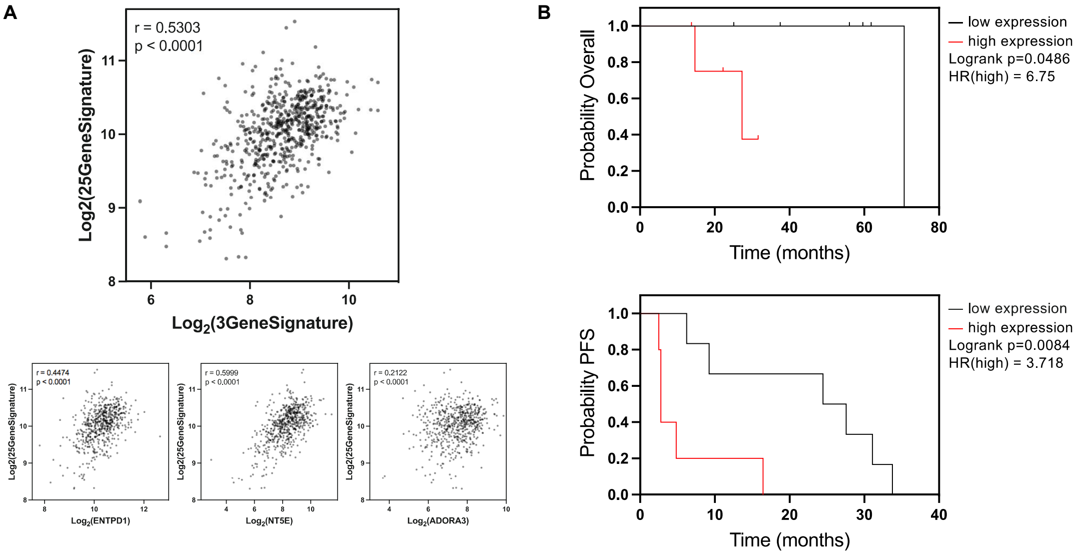


FIGURE 2
ENTPD1/NT5E/ADORA3 signature is correlated with breast cancer osteotropic signature and lower survival in bone patients with established bone metastases. **(A)** A previous study (28) found 25 genes to be upregulated in circulating breast cancer cells from patients with bone metastases compared to patients with extraskeletal metastases, forming a putative bone metastasis-specific signature. We compared ENTPD1/NT5E/ADORA3 expression signature to the 25-gene osteotropic signature in the ER^+ TCGA BRCA cohort. A strong Pearson correlation (r = 0.5303) was observed. Individually, ENTPD1, NT5E, and ADORA3 were each correlated $r > 0.2$. These data imply that the 3-gene signature is not an overall metastasis marker in ER^+ breast cancer and is specifically associated with bone metastasis. **(B)** Kaplan Meier plots displaying overall and progression-free survival in patients with ER^+ breast cancer bone metastasis. In GSE124647, gene expression was measured in bone biopsy samples. We split this cohort into high- and low-expressing ENTPD1/NT5E/ADORA3 signature (calculated by the average expression [log2(norm_count + 1)] of the three genes, and separated by median). In this n = 13 cohort, high 3-gene signature expression was associated with lower overall (HR = 6.75) and progression-free (HR = 3.718) survival, further suggesting ectonucleotidase and ADORA3 expression enables breast cancer growth in the bone microenvironment. *$p < 0.05$; **$p < 0.01$; ****$p < 0.0001$.

metastasis. Primary breast cancer microarray expression studies that reported this outcome were normalized and pooled by KMPlot (24). We found the high-expression cohort for each of ENTPD1, NT5E, and ADORA3 had significantly lower distant metastasis-free survival (**Figure 1A**). None of the other adenosine receptors was significantly correlated with distant metastasis in patients with ER^+ breast cancer (**Supplementary Figure 2A**). To further explore how gene expression in primary tumors might be related to prognosis, we analyzed overall and disease-specific survival among those with ER^+ tumors among the TCGA BRCA cohort. The top half of the 3-gene expression signature (calculated by the average expression [log2(norm_count + 1)] of the three genes, and separated by median) fared more poorly in both outcomes, with an especially strong relationship with disease-specific survival (**Figure 1B**). Additionally, there was generally a stronger relationship with the signature than each individual gene (**Supplementary Figure 2B**). The ENTPD1/NT5E/ADORA3 expression signature was not correlated with either outcome in ER^-, HER2-enriched, or basal breast cancers, which do not share the same metastatic behavior (**Supplementary Figure 3A**), nor were signatures combining expression of the ectonucleotidases with any of the other aADO receptors in ER^+ tumors (**Supplementary Figure 3B**).

Because these outcomes do not measure bone metastasis specifically, we compared the 3-gene signature to an osteotropic breast cancer gene signature (28). To determine this signature, targeted RNA-Seq was performed on circulating cancer cells of patients with metastatic breast cancer. There were 25 genes upregulated in patients with bone metastases compared to patients with extraskeletal metastases. The expression signature combining these 25 genes exhibited a strong correlation with the 3-gene ENTPD1/NT5E/ADORA3 signature in ER^+ breast cancers in the TCGA BRCA cohort, and this relationship is also observed with each of the three genes individually (**Figure 2A**). Because these genes are associated with metastasis to bone, but not to other locations, this suggests that our data are specifically reflective of bone metastasis and not of the overall metastatic ability or aggressiveness.

ENTPD1/NT5E/ADORA3 gene signature in breast cancer bone metastases can predict poor prognosis

If the higher expression of these genes facilitates breast cancer growth in bone, then their elevated expression in already established bone metastases may promote further tumor progression. We compared overall survival and progression-free survival in ER$^+$ breast cancer bone metastases based on the median expression of the three-gene signature. Patients whose tumors were above the median expression level had a significantly lower overall survival and progression-free survival than patients below the median expression (Figure 2B). Notably, expression of none of these genes was individually correlated with overall survival (Supplementary Figure 4, top) and only ADORA3 was significantly correlated with progression-free survival (Supplementary Figure 4, bottom).

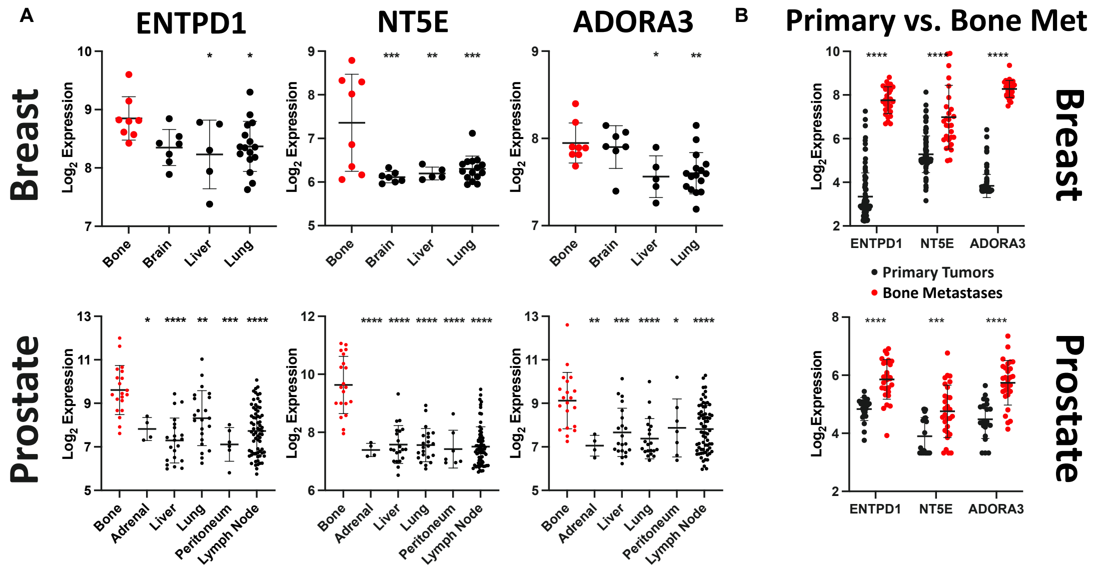


FIGURE 3
ENTPD1, NT5E, and ADORA3 expression are much higher in bone metastases than in other metastases or primary tumors. **(A)** Relative expression of ENTPD1, NT5E, and ADORA3 in metastatic breast and prostate tumors in various organs in GEO datasets GSE14020 and GSE74685. We found that ENTPD1, NT5E, and ADORA3 tend to be more highly expressed in breast and prostate cancer bone metastases (red) than in metastases to other sites (black). **(B)** We found significantly higher expression of ENTPD1, NT5E, and ADORA3 in bone metastases (red) than in primary breast (black, GSE47561) or prostate (black, GSE32269) cancers. One-way ANOVA with Dunnett's post-test was used in **(A)** and unpaired Student's *t*-test was used in **(B)**. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$.

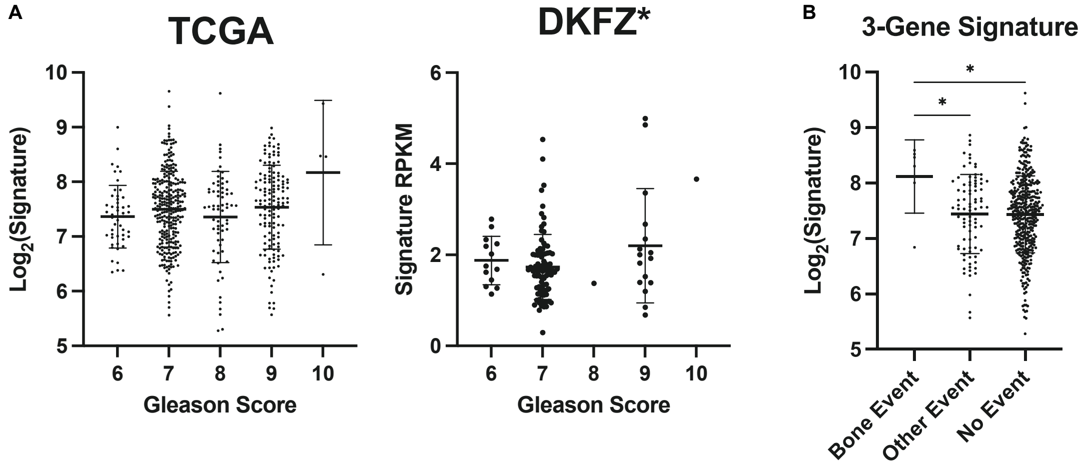


FIGURE 4
ENTPD1/NT5E/ADORA3 expression level is higher in primary prostate tumors that metastasize to bone. Gleason scores reflect the undifferentiation of prostate tumors. Tumors are given two scores that are often added together such as in TCGA PRAD dataset. Tumors with higher scores are more likely to metastasize to bone and other poor outcomes (36). **(A)** An increasing Gleason score is associated with higher ENTPD1/NT5E/ADORA3 expression signature in the DKFZ but not the TCGA PRAD cohort. **(B)** 3-gene expression signature is higher in tumors that form bone metastases, than in tumors that do not progress or cause other events. **(A)** One-way ANOVA with test for trend, **(B)** Kruskal–Wallis test with multiple comparisons. *$p < 0.05$.

This analysis provides further support for the hypothesis that the eATP to eADO hydrolysis might be directly linked to the process of breast cancer bone metastases and overall survival.

Higher expression of ENTPD1/NT5E/ADORA3 gene signature in primary prostate cancer is associated with bone metastasis, but not other progression

We first compared their expression levels across Gleason scores, a measure of tumoral undifferentiation where tumors are given two scores for the dominant and non-dominant phenotype that are often combined into one score. Higher Gleason scores are associated with a worse prognosis and a greater likelihood of recurrence, bone metastasis, and mortality (36–38). There was a significant trend of increasing signature expression with increasing Gleason scores in the German Cancer Research Center cohort (Deutsches Krebsforschungszentrum, DKFZ), (39) but not TCGA PRAD cohort (**Figure 4A**). Thus, the ENTPD1/NT5E/ADORA3 signature does not have a strong relationship with undifferentiation. Notably, most patients who present with localized disease and high Gleason scores do not suffer from bone metastasis in the next 15 years (40). Further, Gleason score and the National Comprehensive Cancer Network combined clinicopathologic score are outperformed by the FDA approved Decipher® Genomic Classifier (41, 42). In the TCGA PRAD cohort, bone-event-causing primary tumors had higher 3-gene signature expression than did those that did not progress, and those that caused other new tumor events, which in this cohort comprise biochemical recurrence, new primary tumor, locoregional metastasis, and distant metastasis to other locations (**Figure 4B**). This suggests that the ENTPD1/NT5E/ADORA3 expression signature is specific for bone metastasis and not of other disease progressions.

Extracellular ATP and A3R antagonist MRS-1220 inhibit breast and prostate cancer cell growth in 3D culture, and non-hydrolyzable ATP analog ATPγS causes stronger reduction

We next used relevant *in vitro* models to determine whether these data reflect a confounding variable or if higher expression of these genes may facilitate bone metastasis. We first compared the effects of ATP and its non-hydrolyzable analog ATPγS on MCF-7, (43) an ER$^+$ breast cancer cell line and 22Rv1, (44) a prostate cancer cell line that originated from the primary tumor of a patient with bone metastasis (45) and that generates mixed osteoblastic and osteolytic tumors in bone (46). In 3D culture conditions ATPγS strongly inhibited growth of MCF-7

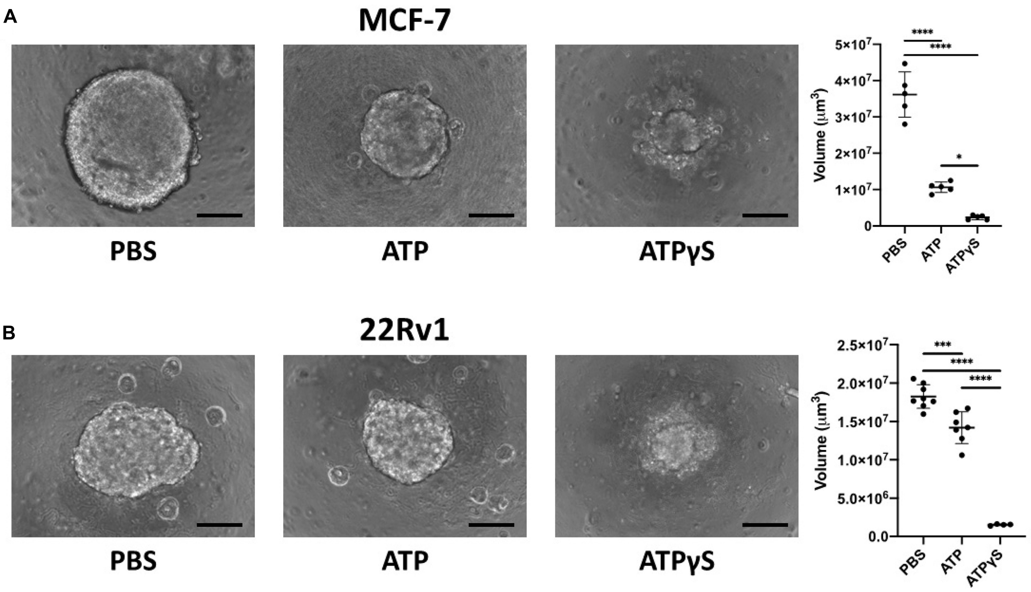


FIGURE 5
Extracellular ATP (eATP) inhibits MCF-7 and 22Rv1 breast cancer cell growth in 3D culture and non-hydrolyzable ATP analog ATPγS causes stronger reduction. We cultured 2000 MCF-7 **(A)** or 22Rv1 **(B)** cells in 3D culture conditions for 1 week with 1mM ATP, ATPγS or vehicle before acquiring images using a Keyence BZ-X710 microscope and determining sphere volume. ATP significantly inhibited the sphere size compared to vehicle, and its non-hydrolyzable analog significantly further reduced sphere size. One-way ANOVA with Tukey's post-test was used for pairwise comparisons. *$p < 0.05$; ***$p < 0.001$; ****$p < 0.0001$. Scale bar = 200 μm.

cells compared to ATP, which is subject to hydrolysis by CD39 and CD73 encoded by ENTPD1 and NT5E genes, respectively. Both conditions inhibited growth compared to PBS vehicle control (Figure 5A). Similar results were obtained in 22Rv1 cells (Figure 5B). These data show that eATP signaling inhibits the growth of breast and prostate cancer cells and that eATP hydrolysis is a mechanism that averts these effects. We also investigated how inhibition of A3R, encoded by the ADORA3 gene, affects growth in 3D culture using MRS-1220, a specific A3R inhibitor. We found dose-dependent growth inhibition in both MCF-7 and 22Rv1 cells in 3D culture with an EC50 of 39 nM in MCF-7 cells and 13 nM in 22Rv1 cells (Figure 6). Together, these data demonstrate the importance of eATP hydrolysis and the reliance on A3 signaling for ER^+ breast cancer and prostate cancer cells.

ER^+ breast cancer cell lines are uniquely sensitive to non-xanthine adenosine receptor antagonists in the drug repurposing Hub

The Drug Repurposing Hub measures differential sensitivity of numerous cell lines to pharmacologic agents (33). We analyzed non-xanthine A3 antagonists CGS-15943, SCH-58261, and MRS-1220 because of their ability to block adenosine receptors without phosphodiesterase inhibition (47). The results strongly supported our hypothesis. Breast cancer cells, especially ER^+ ones, are uniquely sensitive to these three drugs at 2.5 μM (Figure 7). Furthermore, 22Rv1 cells displayed similar sensitivity as ER^+ breast cancer cells, though there were too few prostate cancer cell lines to draw conclusions about prostatic cell lines as a whole. It should be noted, that using this technique, there is a limitation in the lack of connection of this data with the metastatic potential and targeting of the cancer. However, given the prevalence of bone metastasis in breast and prostate cancer, coupled with our other results, we generalize that treatment could lead to far-reaching impact. These data suggest that A3R inhibition may be a new therapeutic avenue for the treatment or prevention of ER^+ breast and prostate cancer bone metastases and further highlights the importance of eADO signaling in cancer.

Discussion

Metastasis is an inefficient process, and few disseminated cells successfully become overt metastases (48). Bone is a highly vascularized tissue (49) easily accessible by circulating cancer cells. An overwhelming majority of cancer deaths are caused by metastasis (50, 51) and bone is the most common metastatic site for ER^+ breast and prostate cancers (1, 2). Understanding

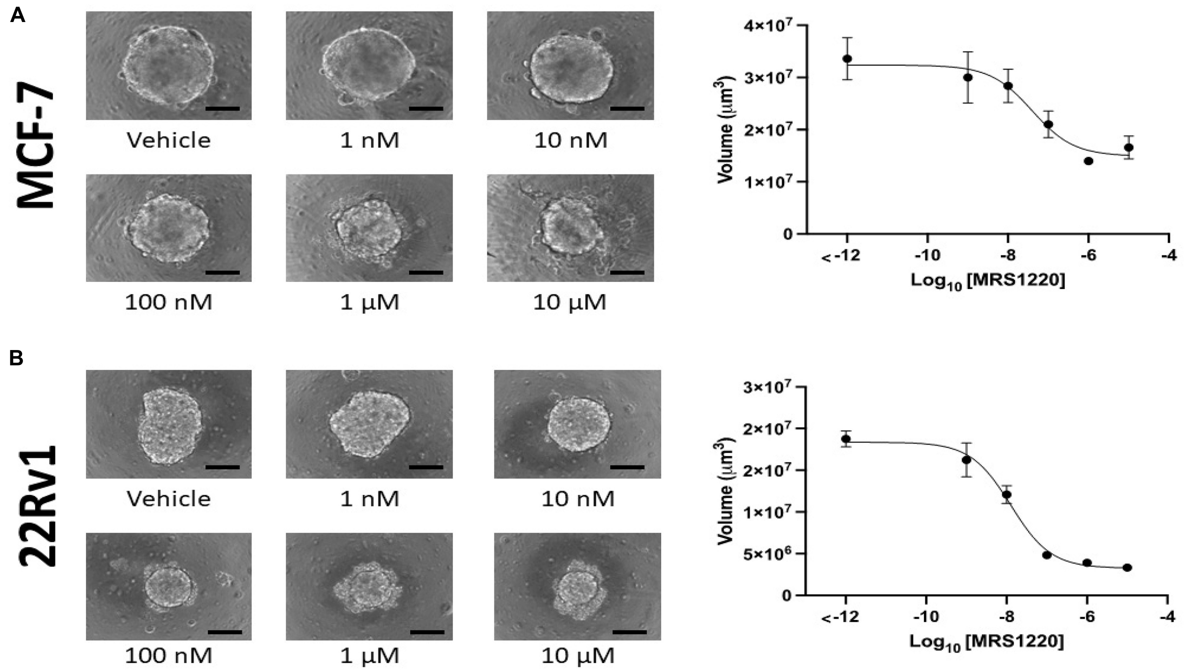


FIGURE 6
Adenosine A3 receptor antagonist MRS-1220 inhibits MCF-7 and 22Rv1 cells in a dose-dependent manner. We incubated 2000 MCF-7 **(A)** and 22Rv1 **(B)** cells in 3D culture conditions for 1 week before determining sphere volume. A dose-dependent inhibition was observed in both cell lines, with EC50 values of 39 nM (95% CI 11.27–110 nM) in MCF-7 cells and 13 nM (95% CI = 8.629–18.39 nM) in 22Rv1 cells. Scale bar = 200 μm.

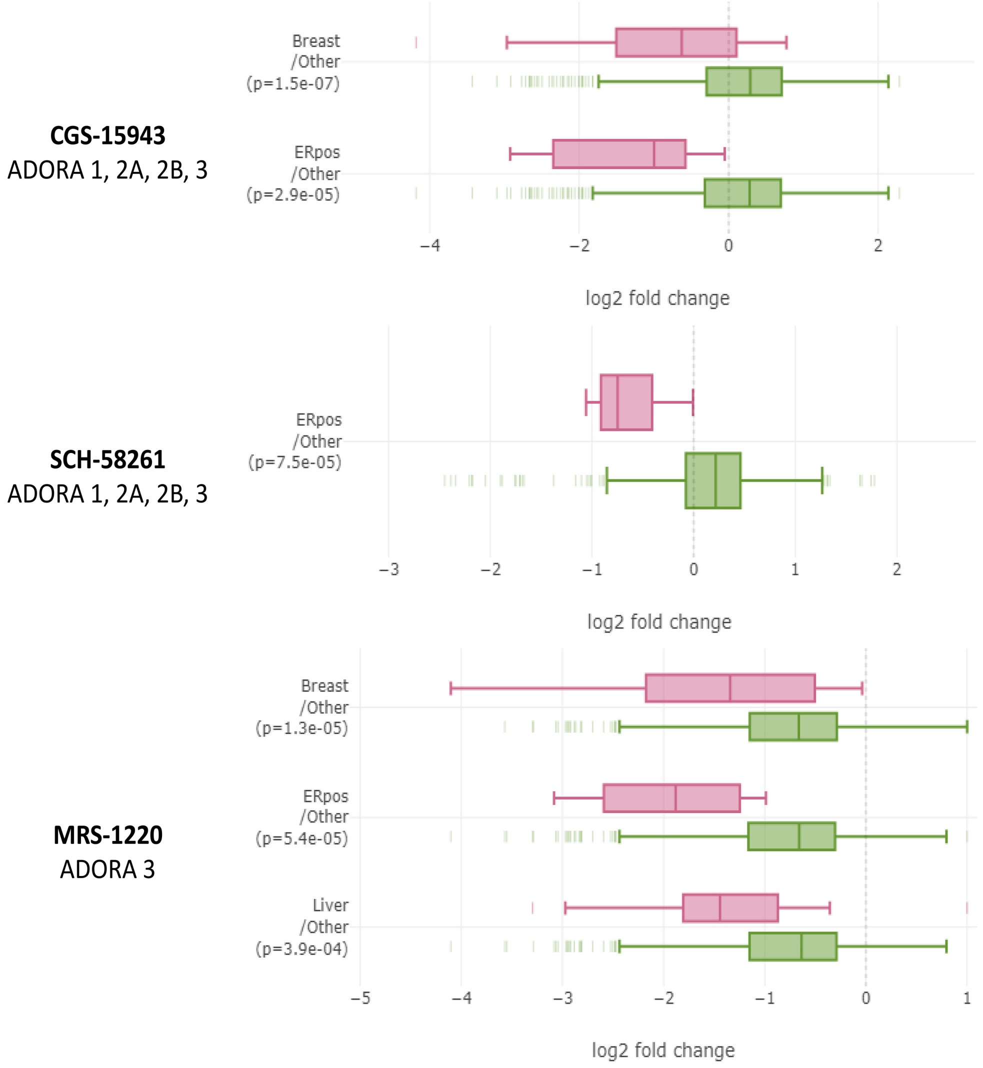


FIGURE 7
Breast cancer cell lines, especially ER$^+$ ones, are uniquely sensitive to non-xanthine adenosine receptor antagonists in large-scale PRISM screen performed by The Drug Repurposing Hub. In the Drug Repurposing Hub, numerous different cell lines are barcoded, pooled, and relative barcode frequency is collected after drug treatment. Non-xanthine A3 antagonists CGS-15943, SCH-68261, and MRS-1220 each decreased relative quantities of ER$^+$ breast cancer cell lines relative to other cell lines. Data were analyzed on DepMap portal using the limma R statistical package. Significant differences were considered by $p < 0.0005$.

the factors that prevent most breast and prostate cancer cells from colonizing this new environment and how some cells bypass these barriers is vital for preventing and treating bone metastases, and also determining which tumors may be low risk. Despite advances in bone metastasis treatment, clinical outcomes after bone metastases remain poor (5, 7, 52, 53). Prevention of breast cancer bone metastases by bisphosphonates is a rare example of a drug targeting a potential metastatic site, effectively reducing metastasis there (13).

Bisphosphonates have long been known to induce apoptosis of osteoclasts (54). We and others have reported that bisphosphonates also promoted osteocytes, the predominant bone cell, to release ATP to the extracellular environment through Cx43 hemichannels and that this decreases triple-negative breast cancer growth in bone (15–17, 55). We further found that eATP signaling inhibits and eADO promotes growth and migration in these cell lines. The present study provides new findings in several ways. We showed that expression of a three-gene expression signature comprising ENTPD1, NT5E, and ADORA3 in primary ER$^+$ breast and prostate cancers was correlated with bone metastases. The fact that these genes were much more highly expressed in bone metastases than in other locations or in primary tumors lends further support for their role in metastasizing bone, a tissue rich in Cx43 hemichannels that release ATP. The growth inhibitory effect of the non-hydrolyzable ATP analog ATPγS compared to eATP on 3D cultures of prostate (22Rv1) and ER$^+$ breast (MCF-7) directly showed the importance of these cells' ability to evade their environment from eATP. We also found that A3R inhibition by MRS-1220 inhibits growth in 3D culture of both cell lines and a wide range of ER$^+$ breast cancer cell lines. Altogether, our data may support a model shown in Figure 8. Osteocytes release ATP to the bone microenvironment that inhibits colonization of ER$^+$ breast and prostate cancers through the activation of one or more ATP receptors. However, in cells that have a greater ability to hydrolyze eATP to eADO through ENTPD1 and NT5E expression, there is less eATP-mediated inhibition. Instead, the generated eADO activates A3 receptor, enabling bone colonization. Future studies should be done, utilizing technology such as siRNA or CRISPR-KO/KD, to determine the direct role of ENTPD1, NTSE, and ADORA3 in cancer cell behavior.

There is a striking difference between breast and prostate cancer bone metastases, with tumors from breast usually displaying an osteolytic, bone destructive phenotype, while tumors from prostate usually adopting an osteoblastic phenotype with increased localized bone density (56). With our data consistent between two very different phenotypes, it is possible that skeletal metastases from other primary tumors share some of the same vulnerabilities and mechanisms.

Bone metastasis is a usually fatal complication that can occur with many cancer types. Unlike other locations, there are treatments that target bone rather than the cancer cells. So far, prophylactic bone metastasis trials have reported mixed results (13, 57, 58). However, these drugs may not be targeted at the right cohort of patients. Because of the long time span in which a metastasis can occur, many available genomic classifiers were designed to predict recurrence (59, 60). These often have limited predictive value for other outcomes. Our data suggest that there may be gene(s) in bone metastasis expression shared between cancers of multiple primary sites. Thus, the ENTPD1/NT5E/ADORA3 signaling axis has the potential to be used as a biomarker or therapeutic target to predict, prevent, or treat bone metastases from multiple sites. Future work should focus on the collection and analysis of this gene signature from primary and bone metastatic cancer sites as well asfrom a broader set of patients.

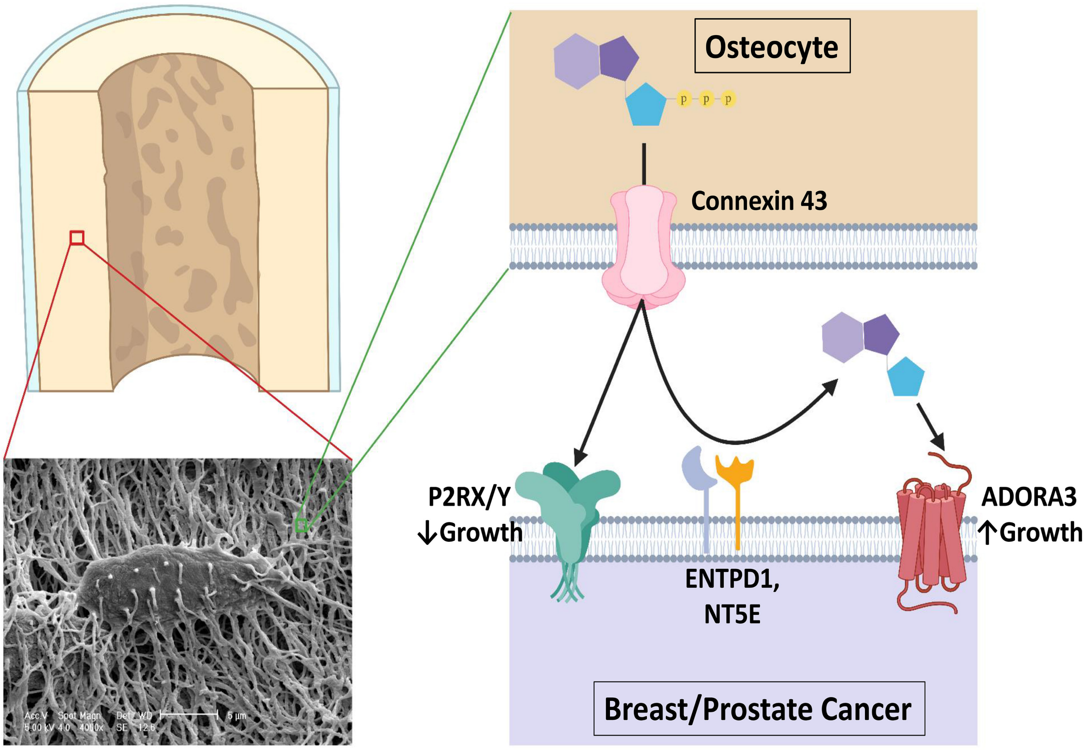


FIGURE 8
A proposed model of the role of purinergic signaling in breast and prostate cancer bone colonization. An estimated 42 billion human osteocytes reside in a lacuna-canalicular network with an estimated surface area of 215 m2 and an extracellular volume of 24 ml (70). Connexin 43 hemichannel activity is promoted by bisphosphonate treatment and in response to shear stress such as seen in exercise, through which ATP is released that usually inhibits breast and prostate cancer growth in bone through ATP receptor stimulation. However, CD39 and CD73 (encoded by ENTPD1 and NT5E) work in concert to hydrolyze the extracellular ATP in the bone microenvironment to ADO, where it is able to activate A3 receptors and promote growth. Figure was made using BioRender.

Inhibiting antibodies against CD39 (encoded by ENTPD1) and CD73 (encoded by NT5E) have recently been developed and are in clinical trials in an immunotherapeutic context (19). Adenosine receptor antagonism, especially of A2A, is also a promising immune stimulator (61, 62). Our data suggests that a separate mechanism inhibiting CD39 and CD73 may be particularly effective in treating or preventing bone metastasis if used in combination with an A3 inhibitor. These classes of drugs may have further enhancement in combination with bisphosphonate treatment.

Preventing bone metastasis may also reduce metastases to other locations. In the overwhelming majority of patients with metastatic ER^+ breast cancer, the initial presentation includes bone (35), and most patients who first present with skeletal metastases later develop metastases at other locations (63). Genetic evidence of bone metastases seeding other metastases has been found for both breast (64) and prostate (65–67) cancer. The bone microenvironment has been shown in experimental models to enhance the plasticity of ER^+ breast cancer cells (68) and strongly increase the ability of breast and prostate cancer cells to colonize in the lung and other organs from leg tumors (69). Thus, the importance of studying and preventing bone metastasis may be even higher than is currently appreciated.

Conclusion

A 3-gene signature composed of ENTPD1, NT5E, and ADORA3 is associated with a greater chance of bone metastasis in ER^+ breast and prostate cancers. These genes are more highly expressed in bone metastases than in other metastases or primary tumors. These genes encode enzymes that hydrolyze eATP to eADO, and an eADO receptor. In 3D culture, eATP decreased spheroid sizes of MCF-7 and 22Rv1 ER^+ breast and prostate cancer cell lines. ATPγS, which is resistant to hydrolysis, further decreased spheroid sizes. These cell lines are sensitive to MRS-1220, a specific A3R inhibitor. ER^+ breast cancer cell lines are sensitive to adenosine receptor inhibition.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Author contributions

DS, JJ, L-ZS, and VJ: conceptualization. DS, FA, JB, and KF: methodology and validation. DS and KF: software. DS: formal analysis, investigation, data curation, and writing—original draft preparation. JJ and L-ZS: resources. FA, L-ZS, KF, VJ, and JJ: writing—review and editing. JJ: project administration and funding acquisition. All authors read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Institutes of Health (NIH) Grants: CA196214 (to JJ and L-ZS), CA217297 and GM114142 (to VJ), and TL1TR002647 (to FA), The Cancer Bioinformatics Initiative: A UTSA/UTHSCSA Partnership: P20 CA165589 (to JB), Department of Defense (DoD): BC161273 (to JJ), and Welch Foundation grant: AQ-1507 (to JJ).

Acknowledgments

The authors thank Hongyun Cheng for technical assistance, Tim Huang at UTHSCSA for generously providing 22Rv1 cells, and Michael Brattain for generously providing MCF-7 cells.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.965429/full#supplementary-material

References

1. Kennecke H, Yerushalmi R, Woods R, Cheang MCU, Voduc D, Speers CH, et al. Metastatic behavior of breast cancer subtypes. *J Clin Oncol.* (2010) 28:3271–7. doi: 10.1200/JCO.2009.25.9820

2. Gandaglia G, Abdollah F, Schiffmann J, Trudeau V, Shariat SF, Kim SP, et al. Distribution of metastatic sites in patients with prostate cancer: a population-based analysis. *Prostate.* (2014) 74:210–6. doi: 10.1002/PROS.22742

3. Weinfurt KP, Li Y, Castel LD, Saad F, Timbie JW, Glendenning GA, et al. The significance of skeletal-related events for the health-related quality of life of patients with metastatic prostate cancer. *Ann Oncol.* (2005) 16:579–84. doi: 10.1093/ANNONC/MDI122

4. Jensen AT, Jacobsen JB, Nørgaard M, Yong M, Fryzek JP, Sørensen HT. Incidence of bone metastases and skeletal-related events in breast cancer patients: a population-based cohort study in Denmark. *BMC Cancer.* (2011) 11:29. doi: 10.1186/1471-2407-11-29/TABLES/3

5. Yong M, Jensen AØ, Jacobsen JB, Nørgaard M, Fryzek JP, Sørensen HT. Survival in breast cancer patients with bone metastases and skeletal-related events: a population-based cohort study in Denmark (1999-2007). *Breast Cancer Res Treat.* (2011) 129:495–503. doi: 10.1007/s10549-011-1475-5

6. Moreira DM, Howard LE, Sourbeer KN, Amarasekara HS, Chow LC, Cockrell DC, et al. Predicting time from metastasis to overall survival in castration-resistant prostate cancer: results from SEARCH. *Clin Genitourin Cancer.* (2017) 15:60–6.e2. doi: 10.1016/j.clgc.2016.08.018

7. Nørgaard M, Jensen AØ, Jacobsen JB, Cetin K, Fryzek JP, Sørensen HT. Skeletal related events, bone metastasis and survival of prostate cancer: a population based cohort study in denmark (1999 to 2007). *J Urol.* (2010) 184:162–7. doi: 10.1016/J.JURO.2010.03.034

8. Steeg PS. Targeting metastasis. *Nat Rev Cancer.* (2016) 16:201–18.

9. Hughes DE, Wright KR, Uy HL, Sasaki A, Yoneda T, Roodman DG, et al. Bisphosphonates promote apoptosis in murine osteoclasts in vitro and in vivo. *J Bone Miner Res.* (1995) 10:1478–87. doi: 10.1002/JBMR.5650101008

10. Plotkin LI, Bellido T. Bisphosphonate-Induced, hemichannel-mediated, anti-apoptosis through the Src/ERK pathway: a gap junction-independent action of connexin43. *Cell Commun Adhes.* (2009) 8:377–82. doi: 10.3109/15419060109080757

11. Saad F, Gleason DM, Murray R, Tchekmedyian S, Venner P, Lacombe L, et al. Long-Term efficacy of zoledronic acid for the prevention of skeletal complications in patients with metastatic hormone-refractory prostate cancer. *J Natl Cancer Inst.* (2004) 96:879–82. doi: 10.1093/JNCI/DJH141

12. Rosen LS, Gordon DH, Dugan W, Major P, Eisenberg PD, Provencher L, et al. Zoledronic acid is superior to pamidronate for the treatment of bone metastases in breast carcinoma patients with at least one osteolytic lesion. *Cancer.* (2004) 100:36–43. doi: 10.1002/CNCR.11892

13. Coleman R, Gray R, Powles T, Paterson A, Gnant M, Bergh J, et al. Adjuvant bisphosphonate treatment in early breast cancer: meta-analyses of individual patient data from randomised trials. *Lancet.* (2015) 386:1353–61. doi: 10.1016/S0140-6736(15)60908-4

14. Bonewald LF. The amazing osteocyte. *J Bone Miner Res.* (2011) 26:229–38. doi: 10.1002/jbmr.320

15. Cherian PP, Siller-Jackson AJ, Gu S, Wang X, Bonewald LF, Sprague E, et al. Mechanical strain opens connexin 43 hemichannels in osteocytes: a novel mechanism for the release of prostaglandin. *Mol Biol Cell.* (2005) 16:3100–6. doi: 10.1091/mbc.E04-10-0912

16. Zhou J, Riquelme M, Gao X, Ellies L, Sun L, Jiang J. Differential impact of adenosine nucleotides released by osteocytes on breast cancer growth and bone metastasis. *Oncogene.* (2015) 34:1831–42. doi: 10.1038/onc.2014.113

17. Zhou J, Riquelme M, Gu S, Kar R, Gao X, Sun L, et al. Osteocytic connexin hemichannels suppress breast cancer growth and bone metastasis. *Oncogene.* (2016) 35101:5597–607. doi: 10.1038/onc.2016.101

18. Boison D, Yegutkin GG. Adenosine metabolism: emerging concepts for cancer therapy. *Cancer Cell.* (2019). 36:582–96.

19. Perrot I, Michaud HA, Giraudon-Paoli M, Augier S, Docquier A, Gros L, et al. Blocking antibodies targeting the cd39/cd73 immunosuppressive pathway unleash immune responses in combination cancer therapies. *Cell Rep.* (2019) 27:2411–25.e9. doi: 10.1016/j.celrep.2019.04.091

20. Li X, Moesta AK, Xiao C, Nakamura K, Casey M, Zhang H, et al. Targeting CD39 in cancer reveals an extracellular ATP- and inflammasome-driven tumor immunity. *Cancer Discov.* (2019) 9:1754–73. doi: 10.1158/2159-8290.CD-19-0541

21. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of Image Analysis. *Nat Methods.* (2012) 9:671–5.

22. Amaral ML, Erikson GA, Shokhirev MN. BART: bioinformatics array research tool. *BMC Bioinformatics.* (2018) 19:296. doi: 10.1186/S12859-018-2308-X/FIGURES/3

23. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015) 43:e47. doi: 10.1093/nar/gkv007

24. Györffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat.* (2010) 123:725–31. doi: 10.1007/s10549-009-0674-9

25. Győrffy B. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Comput Struct Biotechnol J.* (2021) 19:4101–9. doi: 10.1016/J.CSBJ.2021.07.014

26. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. *Nature.* (2012) 490:61–70. doi: 10.1038/nature11412

27. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol.* (2020) 38:675–8. doi: 10.1038/s41587-020-0546-8

28. Lovero D, D'Oronzo S, Palmirotta R, Cafforio P, Brown J, Wood S, et al. Correlation between targeted RNAseq signature of breast cancer CTCs and onset of bone-only metastases. *Br J Cancer.* (2021) 2021:149–29. doi: 10.1038/s41416-021-01481-z

29. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, et al. The molecular taxonomy of primary prostate cancer. *Cell.* (2015) 163:1011–25. doi: 10.1016/J.CELL.2015.10.025

30. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* (2012) 2:401–4. doi: 10.1158/2159-8290.CD-12-0095

31. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* (2013) 6:pl1. doi: 10.1126/SCISIGNAL.2004088/SUPPL_FILE/2004088_TABLES2.XLS

32. Yu C, Mannan AM, Yvone GM, Ross KN, Zhang YL, Marton MA, et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat Biotechnol.* (2016) 34:419–23. doi: 10.1038/nbt.3460

33. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The drug repurposing hub: a next-generation drug library and information resource. *Nat Med.* (2017) 23:405–8. doi: 10.1038/nm.4306

34. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. *Cell.* (2017) 170:564–76.e16. doi: 10.1016/j.cell.2017.06.010

35. Solomayer EF, Diel IJ, Meyberg GC, Gollan C, Bastert G. Metastatic breast cancer: clinical course, prognosis and therapy related to the first site of metastasis. *Breast Cancer Res Treat.* (2000) 593:271–8. doi: 10.1023/A:1006308619659

36. Cao J, Wang T, Li Z, Liu G, Liu Y, Zhu C, et al. Prediction of metastatic prostate cancer by prostate-specific antigen in combination with T stage and gleason grade: nationwide, population-based register study. *PLoS One.* (2020) 15:e0228447. doi: 10.1371/JOURNAL.PONE.0228447

37. Kamel M, Khalil M, Alobuia W, Su J, Davis R. Incidence of metastasis and prostate-specific antigen levels at diagnosis in Gleason 3+4 versus 4+3 prostate cancer. *Urol Ann.* (2018) 10:203. doi: 10.4103/UA.UA_124_17

38. Egevad L, Granfors T, Karlberg L, Bergh A, Stattin P. Prognostic value of the Gleason score in prostate cancer. *BJU Int.* (2002) 89:538–42. doi: 10.1046/J.1464-410X.2002.02669.X

39. Gerhauser C, Favero F, Risch T, Simon R, Feuerbach L, Assenov Y, et al. Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell.* (2018) 34:996–1011. doi: 10.1016/J.CCELL.2018.10.016

40. Leapman MS, Cowan JE, Simko J, Roberge G, Stohr BA, Carroll PR, et al. Application of a prognostic gleason grade grouping system to assess distant prostate cancer outcomes. *Eur Urol.* (2017) 71:750–9. doi: 10.1016/J.EURURO.2016.11.032

41. Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One.* (2013) 8:e66855. doi: 10.1371/JOURNAL.PONE.0066855

42. Tosoian JJ, Birer SR, Jeffrey Karnes R, Zhang J, Davicioni E, Klein EE, et al. Performance of clinicopathologic models in men with high risk localized prostate

cancer: impact of a 22-gene genomic classifier. *Prostate Cancer Prostatic Dis.* (2020) 23:646–53. doi: 10.1038/s41391-020-0226-2

43. Soule HD, Vazquez J, Long A, Albert S, Brennan M. A human cell line from a pleural effusion derived from a breast carcinoma. *J Natl Cancer Inst.* (1973) 51:1409–16. doi: 10.1093/JNCI/51.5.1409

44. Sramkoski RM, Pretlow TG, Giaconia JM, Pretlow TP, Schwartz S, Sy MS, et al. A new human prostate carcinoma cell line, 22Rv1. *Vitr Cell Dev Biol Anim.* (1999) 35:403–9. doi: 10.1007/S11626-999-0115-4

45. Pretlow TG, Wolman SR, Micale MA, Pelley RJ, Kursh ED, Resnick MI, et al. Xenografts of primary human prostatic carcinoma. *J Natl Cancer Inst.* (1993) 85:394–8. doi: 10.1093/JNCI/85.5.394

46. Henry MD, Silva MD, Wen S, Siebert E, Solin E, Chandra S, et al. Spiculated periosteal response induced by intraosseous injection of 22Rv1 prostate cancer cells resembles subset of bone metastases in prostate cancer patients. *Prostate.* (2005) 65:347–54. doi: 10.1002/PROS.20300

47. Beavo JA, Rogers NL, Crofford OB, Hardman JG, Sutherland EW, Newman EV. Effects of xanthine derivatives on lipolysis and on adenosine 3',5'-monophosphate phosphodiesterase activity. *Mol Pharmacol.* (1970) 6:597–603.

48. Fidler IJ. Metastasis: quantitative analysis of distribution and fate of tumor emboli labeled With 125I-5-Iodo-2′ -deoxyuridine. *J Natl Cancer Inst.* (1970) 45:773–82. doi: 10.1093/JNCI/45.4.773

49. Tomlinson RE, Silva MJ. Skeletal blood flow in bone repair and maintenance. *Bone Res.* (2013) 1:311–22. doi: 10.4248/br201304002

50. Chaffer CL, Weinberg RA. A perspective on cancer cell metastasis. *Science.* (2011) 331:1559–64. doi: 10.1126/SCIENCE.1203543/SUPPL_FILE/1559.MP3

51. Dillekås H, Rogers MS, Straume O. Are 90% of deaths from cancer caused by metastases? *Cancer Med.* (2019) 8:5574–6. doi: 10.1002/CAM4.2474

52. Fizazi K, Carducci M, Smith M, Damião R, Brown J, Karsh L, et al. Denosumab versus zoledronic acid for treatment of bone metastases in men with castration-resistant prostate cancer: a randomised, double-blind study. *Lancet.* (2011) 377:813–22. doi: 10.1016/S0140-6736(10)62344-6

53. Stopeck AT, Lipton A, Body JJ, Steger GG, Tonkin K, De Boer RH, et al. Denosumab compared with zoledronic acid for the treatment of bone metastases in patients with advanced breast cancer: a randomized, double-blind study. *J Clin Oncol.* (2010) 28:5132–9. doi: 10.1200/JCO.2010.29.7101

54. Fleisch H, Russell RGG, Francis MD. Diphosphonates inhibit hydroxyapatite dissolution in vitro and bone resorption in tissue culture and in vivo. *Science.* (1969) 165:1262–4. doi: 10.1126/SCIENCE.165.3899.1262

55. Genetos DC, Kephart CJ, Zhang Y, Yellowley CE, Donahue HJ. Oscillating fluid flow activation of gap junction hemichannels induces atp release from MLO-Y4 osteocytes. *J Cell Physiol.* (2007) 212:207–14. doi: 10.1002/jcp.21021

56. O'Sullivan GJ. Imaging of bone metastasis: an update. *World J Radiol.* (2015) 7:202. doi: 10.4329/wjr.v7.i8.202

57. Coleman R, Finkelstein DM, Barrios C, Martin M, Iwata H, Hegg R, et al. Adjuvant denosumab in early breast cancer (D-CARE): an international, multicentre, randomised, controlled, phase 3 trial. *Lancet Oncol.* (2020) 21:60–72. doi: 10.1016/S1470-2045(19)30687-4

58. Smith MR, Saad F, Coleman R, Shore N, Fizazi K, Tombal B, et al. Denosumab and bone-metastasis-free survival in men with castration-resistant prostate cancer: results of a phase 3, randomised, placebo-controlled trial. *Lancet.* (2012) 379:39–46. doi: 10.1016/S0140-6736(11)61226-9

59. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med.* (2004) 351:2817–26. doi: 10.1056/NEJMOA041588/SUPPL_FILE/NEJMOA041588SA1.PDF

60. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med Genomics.* (2015) 8:54. doi: 10.1186/S12920-015-0129-6/FIGURES/10

61. Fong L, Hotson A, Powderly J, Sznol M, Heist RS, Choueiri TK, et al. Adenosine A2A receptor blockade as an immunotherapy for treatment-refractory renal cell cancer. *Cancer Discov.* (2019) 10:40–53. doi: 10.1158/2159-8290.CD-19-0980

62. Allard B, Allard D, Buisseret L, Stagg J. The adenosine pathway in immuno-oncology. *Nat Rev Clin Oncol.* (2020) 17:611–29. doi: 10.1038/s41571-020-0382-2

63. Coleman RE, Rubens P, Smith RD. Clinical course and prognostic factors following bone recurrence from breast cancer. *Br J Cancer.* (1998) 77:336–40.

64. Ullah I, Karthik GM, Alkodsi A, Kjällquist U, Stålhammar G, Lövrot J, et al. Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes. *J Clin Invest.* (2018) 128:1355–70. doi: 10.1172/JCI96149

65. Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature.* (2015) 520:353–7. doi: 10.1038/nature14347

66. Hong MKH, Macintyre G, Wedge DC, Van Loo P, Patel K, Lunke S, et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun.* (2015) 6:6605. doi: 10.1038/ncomms7605

67. Haider M, Zhang X, Coleman I, Ericson N, True LD, Lam HM, et al. Epithelial mesenchymal-like transition occurs in a subset of cells in castration resistant prostate cancer bone metastases. *Clin Exp Metastasis.* (2016) 33:239–48. doi: 10.1007/S10585-015-9773-7/TABLES/1

68. Bado IL, Zhang W, Hu J, Xu Z, Wang H, Sarkar P, et al. The bone microenvironment increases phenotypic plasticity of ER+ breast cancer cells. *Dev Cell.* (2021) 56:1100–17.e9. doi: 10.1016/J.DEVCEL.2021.03.008

69. Zhang W, Bado IL, Hu J, Wan YW, Wu L, Wang H, et al. The bone microenvironment invigorates metastatic seeds for further dissemination. *Cell.* (2021) 184:2471–86.e20. doi: 10.1016/J.CELL.2021.03.011

70. Buenzli PR, Sims NA. Quantifying the osteocyte network in the human skeleton. *Bone.* (2015) 75:144–50. doi: 10.1016/j.bone.2015.02.016

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 27 September 2022
DOI 10.3389/fmed.2022.976467

Self-supervised learning mechanism for identification of eyelid malignant melanoma in pathologic slides with limited annotation

Linyan Wang[1†], Zijing Jiang[2†], An Shao[1], Zhengyun Liu[3], Renshu Gu[2], Ruiquan Ge[2], Gangyong Jia[2], Yaqi Wang[4*] and Juan Ye[1*]

[1]Department of Ophthalmology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2]School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, [3]Department of Pathology, Lishui Municipal Central Hospital, Lishui, China, [4]College of Media Engineering, The Communication University of Zhejiang, Hangzhou, China

Purpose: The lack of finely annotated pathologic data has limited the application of deep learning systems (DLS) to the automated interpretation of pathologic slides. Therefore, this study develops a robust self-supervised learning (SSL) pathology diagnostic system to automatically detect malignant melanoma (MM) in the eyelid with limited annotation.

Design: Development of a self-supervised diagnosis pipeline based on a public dataset, then refined and tested on a private, real-world clinical dataset.

Subjects: A. Patchcamelyon (PCam)-a publicly accessible dataset for the classification task of patch-level histopathologic images. B. The Second Affiliated Hospital, Zhejiang University School of Medicine (ZJU-2) dataset – 524,307 patches (small sections cut from pathologic slide images) from 192 H&E-stained whole-slide-images (WSIs); only 72 WSIs were labeled by pathologists.

Methods: Patchcamelyon was used to select a convolutional neural network (CNN) as the backbone for our SSL-based model. This model was further developed in the ZJU-2 dataset for patch-level classification with both labeled and unlabeled images to test its diagnosis ability. Then the algorithm retrieved information based on patch-level prediction to generate WSI-level classification results using random forest. A heatmap was computed for visualizing the decision-making process.

Main outcome measure(s): The area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity were used to evaluate the performance of the algorithm in identifying MM.

OPEN ACCESS

EDITED BY
Jun Cheng,
Shenzhen University, China

REVIEWED BY
Qingyu Zhao,
Stanford University, United States
Sumit Kumar,
Amity University, India

*CORRESPONDENCE
Juan Ye
yejuan@zju.edu.cn
Yaqi Wang
wangyaqi@cuz.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Pathology,
a section of the journal
Frontiers in Medicine

RECEIVED 23 June 2022
ACCEPTED 26 August 2022
PUBLISHED 27 September 2022

CITATION
Wang L, Jiang Z, Shao A, Liu Z, Gu R, Ge R, Jia G, Wang Y and Ye J (2022) Self-supervised learning mechanism for identification of eyelid malignant melanoma in pathologic slides with limited annotation.
Front. Med. 9:976467.
doi: 10.3389/fmed.2022.976467

Results: ResNet50 was selected as the backbone of the SSL-based model using the PCam dataset. This algorithm then achieved an AUC of 0.981 with an accuracy, sensitivity, and specificity of 90.9, 85.2, and 96.3% for the patch-level classification of the ZJU-2 dataset. For WSI-level diagnosis, the AUC, accuracy, sensitivity, and specificity were 0.974, 93.8%, 75.0%, and 100%, separately. For every WSI, a heatmap was generated based on the malignancy probability.

Conclusion: Our diagnostic system, which is based on SSL and trained with a dataset of limited annotation, can automatically identify MM in pathologic slides and highlight MM areas in WSIs by a probabilistic heatmap. In addition, this labor-saving and cost-efficient model has the potential to be refined to help diagnose other ophthalmic and non-ophthalmic malignancies.

KEYWORDS

artificial intelligence - assisted bioinformatic analysis, self-supervised deep learning, pathology, tumor diagnosis, melanoma

Introduction

Malignant melanoma (MM) is an intractable cutaneous cancer originating from melanocytes with an extremely high mortality rate (65% of all skin cancer deaths) (1). Although eyelid melanoma accounts for only ∼1% of all cutaneous melanomas, it can camouflage melanocytic nevus (the most common benign eyelid tumor) both in the naked eye and under a microscope. Its primary diagnosis and management fall within the realm of ophthalmology. Despite the similar appearance, these two tumor types have markedly different biological behaviors, corresponding to distinct prognoses and treatments. Therefore, it is critically important to distinguish between the two diseases (2, 3). Like other types of tumor, the gold standard for MM diagnosis still relies on manual histopathological interpretation, which is subjective, laborious, tedious, and challenging for pathologists and ophthalmologists lacking experience encountering eyelid melanoma (3). Computer-aided diagnosis (CAD) in eyelid melanoma cases is urgently needed to make a comprehensive and objective pathological diagnosis (4).

The advancement of artificial intelligence (AI) technology has cast light both on natural images and medical areas. Compared with other fields, the automatic diagnosis based on histopathological images confronts more challenges due to the uniqueness of pathological data. Firstly, the digitization of traditional glass slides needs additional scanning equipment. Secondly, most pathological images are gigapixels, which are tremendously large: about 470 whole slide images (WSIs) scanned at 20× magnification (0.5 μm pixel^{-1}) contain roughly the same number of pixels as the entire ImageNet dataset (5). The diagnosis of pathology highly depends on its cellular characteristics, which means we need to annotate and analyze at the patch (a small tile cutting from WSI) level first. Such a procedure requires tremendous annotations by expert pathologists. Thirdly, based on the incidence of ocular tumors, the pathology slides are more valuable than fundus images or optical coherence tomography (OCT) images, which could be obtained in a routine follow-up. The lack of expertise to make high-quality annotations further restricted the number of usable pathology slides.

However, the availability of medical specialists to annotate digitized images and free text diagnostic reports does not scale with the need for large datasets required to train robust computer-aided diagnosis methods that can target the high variability of clinical cases and data produced. Most previous attempts in computational pathology are fully supervised learning studies. The automated system of pathological images requires a sufficient quantity of images with annotations (6–9). There are several drawbacks to this procedure. First, collecting unlabeled digitized slides only needs technicians to scan, but labeled images need extra experts with many years of medical education. If unlabeled medical images could be used in deep learning analysis, the usable datasets could be significantly expanded. Moreover, the laborious annotation has the potential to introduce manual label errors, as most current annotations were carried out at lower magnification.

Moreover, some boundaries of the tumor area are ambiguous with normal mixed cells and cancer cells, which even perplexes the annotation process. Last but not least, the unlabeled images by themselves may still include substantial clinical information. From the view above, generating a diagnostic system that can utilize labeled and unlabeled images may greatly benefit the diagnosis and treatment of the disease.

Self-supervised learning (SSL) is a new type of unsupervised learning algorithm to extract and analyze features of given data automatically. SSL has been applied to input data in various models, including RGB image (10), videos (11), medical image (12), mass spectrometry data (13), or multimodal data (14). The high efficiency of SSL makes it suitable for auxiliary medical uses. SSL requires only a limited quantity of labeled data and a relatively abundant quantity of unlabeled data for the machine to learn features. This perfectly meets the clinical conditions in which annotating pathological images is laborious, time-consuming, and probably inaccurate. We generate a diagnostic system based on Bootstrap Your Own Latent (BYOL), a new approach to SSL proven to achieve better performance compared to contrastive methods of other SSL algorithms (15). Generally, an original gigapixel-level pathological image is too complex for a deep learning system (DLS) to analyze. Therefore, we divided images into small patches. After pretreating these patch-level images, we input the patches of unlabeled images into the SSL network for extracting features as a pretraining task. Subsequently, combined with other labeled images, these learned features are repurposed to improve the classification of the network and thus increase data utilization. Using a random forest model, we extrapolated patch-level classification to whole-slide-image-level classification. Apart from the above, we also generate a heatmap of pathologic images to interpret the decision-making process.

This study aims to apply an SSL network to diagnose and classify MM and non-malignant areas from digital H&E stained pathological slides. To our knowledge, no other SSL networks have been used in detecting eyelid melanoma; we demonstrate the feasibility of using limited labeled data to establish a reliable eyelid MM detection model and describe a strategy for highlighting specific areas of concern.

Methods

This study was approved by the Second Affiliated Hospital, Zhejiang University School of Medicine (ZJU-2) Ethics Committee (No. Y2019-195) and the study adhered to the Declaration of Helsinki. In this study, we applied self-supervised learning to make eyelid melanoma identification. Our algorithm was first developed and tested in PatchCamelyon and then in the ZJU-2 dataset. Digitized pathologic images of slides were cut into small patches. The classification was based on these patch-level images and then extrapolated to WSIs. Besides, the algorithm also generated a heatmap to highlight the exact lesion area in WSI and improve the interpretability of the decision-making process of our model. The whole study workflow is summarized in **Figure 1**.

Datasets

A. PatchCamelyon (PCam): a publicly accessible dataset containing 327680 annotated color images (96 × 96 pixels) extracted from histopathologic scans of lymph node sections (16). The dataset uses agreed-upon metrics widely to compare different convolutional neural networks (CNNs) as the backbone. In this study, we used PCam as the benchmark for our model and compared the performance of our model to other CNNs or algorithms. The original data of PCam is shown in **Figure 2A**. The images were divided into training, validation, and testing.

B. ZJU-2 dataset: 192 whole-slide images (WSIs) from formalin-fixed paraffin-embedded (FFPE) pathological slices (**Table 1**). We retrospectively included 160 patients from the Second Affiliated Hospital, Zhejiang University School of Medicine, between January 2005 and December 2017, without other types of special selection. All slides were diagnosed by a minimum of two board-certified pathologists using H&E staining (if necessary, additional immunohistochemical staining was used) and traditional microscopy. There was no divergence in the diagnosis of all samples in this study. A separate technician working within the pathology department then scanned the selected slides and digitized these slides into WSIs at 400-fold magnification using a KF-PRO-005 (KFBio, Zhejiang, China). The WSIs were divided into four sets: pretraining, training, validation, and testing set. Besides the WSIs in the pretraining set, images from the other three sets were reviewed and labeled by an additional independent pathologist (>10 years of experience). By using window sliding, 192 whole slide images (WSIs) were cut into a total of 524,307 patches (256 × 256 pixels) for analysis (**Figure 2B**). The detailed image data partition is shown in **Table 1**. It's worth noting that only delineated tumor areas of MM slides were defined as malignant, and other patches were defined as non-malignant.

Self-supervised learning approach

We used Bootstrap Your Own Latent (BYOL) for learning features from unlabeled WSIs in this study (15). The study workflow was shown in **Figures 1A–C**. The architecture is shown in **Figure 1C**. In detail, it consists of two neural networks: online networks and target networks. It produces two augmented views (*v* and *v*') from a single image by applying two different distributions of image augmentations: one with a random horizontal flip and another with a random horizontal flip and gaussian blur (*t* and *t*'). Two identical CNNs with a different set of weights then output the representation (*y* and *y*') and projection (*z* and *z*') through a multilayer perceptron (MLP). On the online branch, we output the prediction *p*, making the architectural asymmetry. After normalizing *p* and *z*', we defined the mean squared error between normalized predictions and target projections, thereby generating the loss ($\mathcal{L}_{\theta,\xi}^{\mathrm{SSL-linear}}$). By reversely feeding *t* to the target network and *t*' to an online network, we computed the loss ($\mathcal{L}_{\theta,\xi'}^{\mathrm{SSL-linear}}$) and minimized the L2 loss $= \mathcal{L}_{\theta,\xi}^{\mathrm{SSL-linear}} + \mathcal{L}_{\theta,\xi'}^{\mathrm{SSL-linear}}$ with a stochastic optimization step, as depicted by the unidirectional gradient in

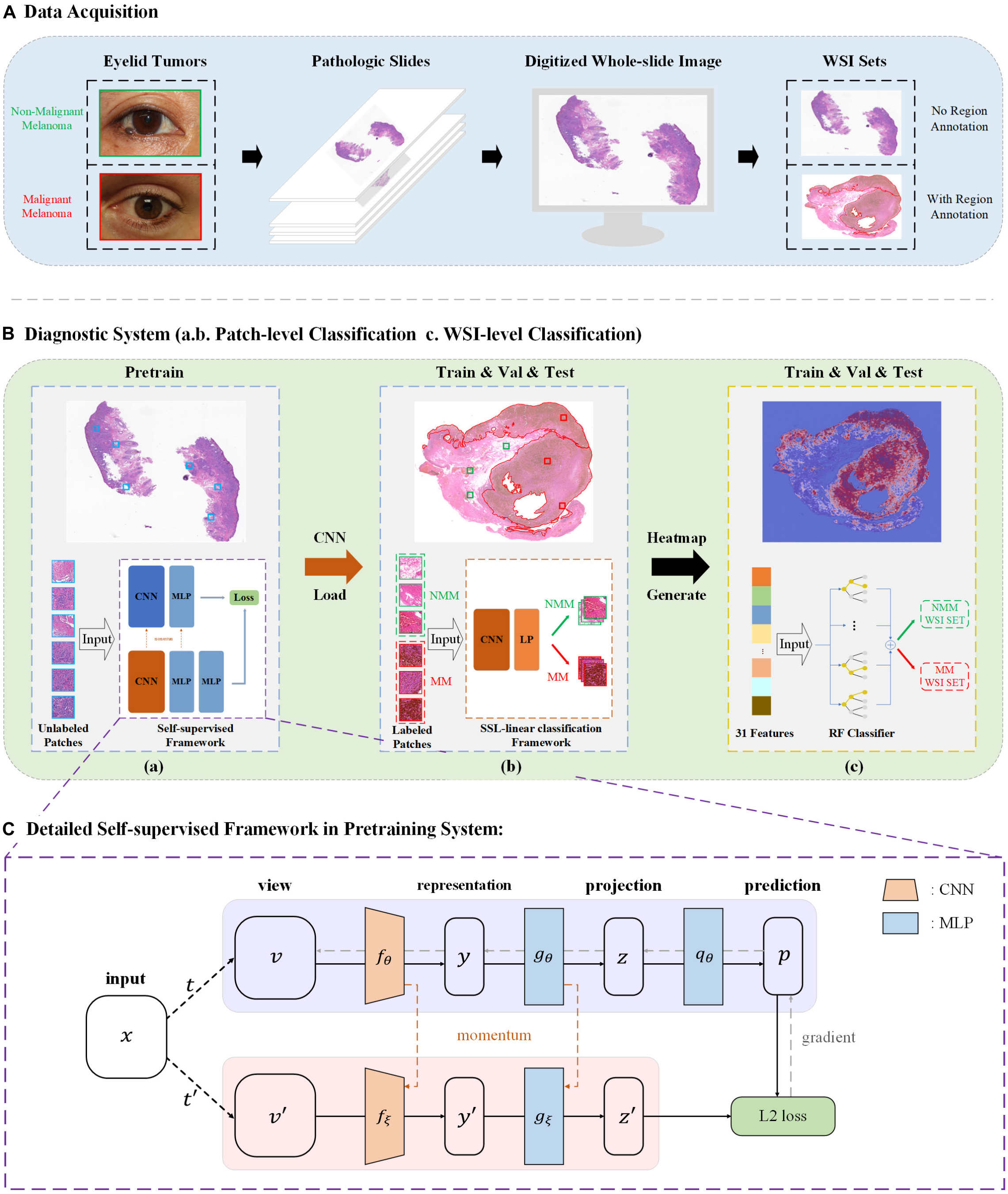


FIGURE 1
Study workflow. **(A)** Pathologic slides were acquired from eyelid tumors and transformed into digitized whole-slide images (WSIs). An experienced pathologist labeled ∼25% WSIs by delineating the tumor areas in WSIs. **(B)** Diagnostic system. **(a)** Pretraining is based on Bootstrap Your Own Latent (BYOL), a new approach to SSL. Patches from unlabeled WSIs were input into two identical convolutional neural networks (CNNs) with two different sets of weights for learning features and comparing the outputs with each other as pretraining. A load of learned image representation was then generated. **(b)** Training for patch-level classification. Patches from labeled WSIs (training and validation sets) were input into a CNN for training together with the load from the pretraining round, and training weights were acquired. The diagnostic ability of patch-level images was evaluated in the testing set. A value of the malignancy probability of every patch is then generated (not shown). **(c)** Extrapolation to image-level classification. Patches were embedded back into the corresponding WSIs, and by feeding back the malignancy probability of every patch, a probabilistic heatmap for WSIs was generated. Based on the predicted patch value, the threshold transformation was used to extract 31 features. The WSI-level classification based on random forest (RF) was then assigned. **(C)** BYOL architecture. In 2 CNNs ($f\theta$ and $f\xi$) with a different set of weights, θ are the trained weights, and ξ is an exponential moving average of θ. At the end of the training, parameter θ is acquired with the minimum of L2 loss, and *y* is used as the learned representation—Val, validation; MLP, multilayer perceptron; MM, malignant melanoma; NMM, non-malignant melanoma.

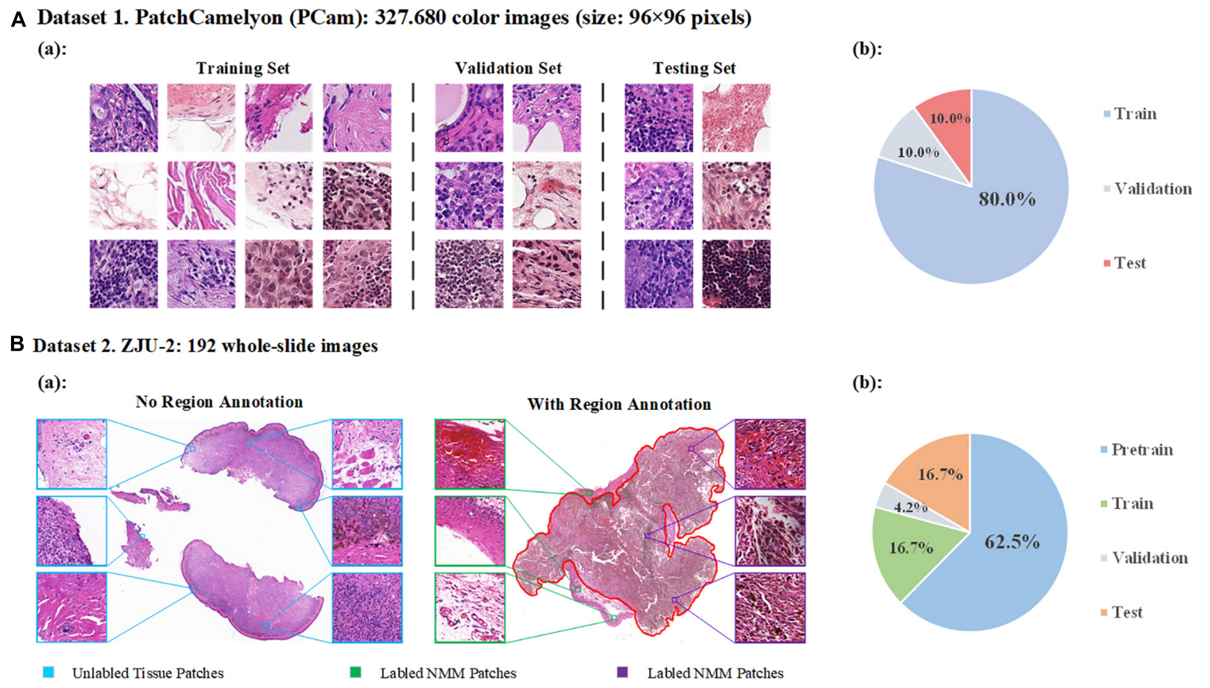


FIGURE 2

Data distribution. **(A)** Detailed data of PatchCamelyon (Pcam). **(a)** Examples of images in PatchCamelyon. **(b)** The data of original image grouping. **(B)** Dataset of ZJU-2. **(a)** Examples of pathological digitized WSIs with or without annotations and patches from WSIs. **(b)** Patches are divided into four sets: pretraining, training, validation, and testing sets.

Figure 1C. The Adam optimizer is used, and the learning rate is set as 3e-4. The learning rate of SSL-linear is 0.01, and the momentum is set at 0.9. Four NVIDIA TITAN Xp GPUs were used for model training.

Self-supervised learning-linear for patch-level classification

The self-supervised algorithm needs to use a CNN model as its base algorithm or backbone. When choosing the backbone candidates, we took the size of our datasets and the depth, stability, and memory cost of different CNNs into consideration when choosing the backbone candidates. So, we started with five commonly used CNNs (VGG16, ResNet18, ResNet50, DenseNet121, EfficientNetB7) for initial fully supervised learning tests in PCam (17, 18). After choosing the CNN with the best performance as the backbone to generate the SSL-linear, we moved on to the next experiment stage–comparing different self-supervised algorithms. The SSL-linear (No Pre) and SSL-linear (Frozen) methods were used as control groups to prove that both stages are necessary for the SSL-linear method. SSL-linear (No Pre) did not undergo a pretraining process, and SSL-linear (Frozen) froze the backbone CNN model's parameters during the training process, which is a traditional way of self-supervised learning. By comparing SSL-linear to the traditional classifiers, including ResNet50, SSL-linear was proved to be valid and feasible for pathological images. The algorithm was then applied to learn features and make classifications from patch-level images in four sets of the ZJU2 dataset.

Model performance was evaluated by accuracy (Acc), sensitivity (Sen), specificity (Spe), and the κ statistic (Cohen's kappa coefficient). For every patch, malignancy probability was calculated between 0 and 1 (1 refers to definitively malignant and is presented in red on the heatmap, while 0 refers to completely NM and is presented in blue) before feeding this estimate back into the WSI and generating the probabilistic heatmap for the full WSI.

Feature extraction and whole-slide-image-level classification using random forest

The original probabilistic heatmap was reprocessed, and 31 features were extracted, including the number of tumor areas; the proportion of tumor areas in the whole tissue; the largest area of the tumor; the longest axis of the largest tumor area; the prediction value across the tumor areas; the number of positive pixels; max, mean, variance, skewness, and kurtosis of pixel numbers in all tumor areas; perimeter, eccentricity, and solidity

TABLE 1 Summary of the ZJU-2 dataset.

Data set	Pretraining set		Training set		Validation set		Testing set		Total set	
	(Unlabeled)		(Labeled)		(Labeled)		(Labeled)			
Images (n)	WSI (patient)	Patch	WSI (patient)	Patch	WSI (patient)	Patch	WSI (patient)	Patch	WSI (patient)	Patch
Non-malignant	80 (76)		24 (24)	34187	4 (4)	1099	24 (24)	15597	132 (123)	
Malignant	40 (24)		8 (8)	35620	4 (4)	874	8 (8)	14762	60 (37)	
Total	120 (100)	422168	32 (32)	69807	8 (8)	1973	32 (32)	30359	192 (160)	524307

in tumor areas (6). These features were then used for WSI-level classification. The probabilistic input heatmap was a single-channel image the same size as the original WSI. Each pixel was refilled based on the prediction results (malignancy probability between 0 and 1). The 31 tumor features were encompassed with a threshold of 0.5. For all input objects, pixels greater or equal to the threshold value were assigned a pixel value of 255, while those below the threshold were set to 0. Following the extraction of these 31 features, WSI-level classification was applied. The random forest classifier shared the same training sets as SSL-linear, but SSL-linear analyzed patch-level images while the random forest classifier analyzed WSIs. The extracted 31 features with label information were sent into the random forest model for prediction.

Statistical analysis

In this study, we plotted receiver operating characteristic (ROC) curves to evaluate the performance of different classification algorithms. Classification metrics were calculated, including Acc, Sen, Spe, κ score (Cohen's kappa), balanced accuracy (B_Acc), and the area under the receiver operating characteristic curve (AUC) for each model. B_Acc is more sensitive to imbalanced data and can be used to address the inequality between malignant and NM data sets. All statistical analyses were conducted using the programming language Python (V.3.5.4).

Results

Classification ability in PatchCamelyon – the public dataset

In the PCam dataset, ResNet 50 outperformed the other four commonly used CNNs (VGG16, ResNet18, DenseNet121, and EfficientNetB7) in the supervised study task (**Table 2**, Group 1) and was chosen as the backbone to generate the SSL-linear algorithm. In the supervised study, among these 5 CNNs, ResNet50 had the highest AUC, 0.950, spe 90.1%, indicating the best performance. EfficientNetB7 had the highest Acc 88.4%; B_Acc 88.4%; κ score 0.767; Spe 92.0%. However, the training time of EfficientNetB7 (83.6 h) is approximately 5.5 times longer than ResNet50 (14.7 h). The volume of parameters of EfficientNetB7 (63.8 M) is 2.7 times larger than ResNet50 (23.5 M). The long training time and high demand for the memory capacity of graphical processing units (GPUs) make it impractical to use EfficientNetB7 in clinical settings. The comparison experiments of various networks also verified the rationality of the selection. Second, in a self-supervised study, we evaluated and compared the performance of SSL-linear and ResNet50 with different proportions of unlabeled pretraining

TABLE 2 Results of classification task in Patchcamelyon (PCam).

Pretrain: Train	Method	Acc (%)	B_Acc (%)	κ score	Sen (%)	Spe (%)	AUC
Ratio							
0:10 (Group 1)	VGG16	87.8	87.8	0.755	88	87.5	0.949
	ResNet18	85.9	85.9	0.718	82	90	0.929
	ResNet50	88.2	88.2	0.765	**86.3**	90.1	**0.950**
	DenseNet121	87.8	87.8	0.756	85.4	90.2	0.947
	EfficientNetB7	**88.4**	**88.4**	**0.767**	84.7	**92.0**	0.941
	Veeling et al. (17)	89.8					
	Mohamed et al. (18)	89.2					
5:5 (Group 2)	ResNet50 (No Pre)	84.2	84.2	0.685	**84.1**	84.4	0.923
	SSL-linear (No Pre)	84.8	84.8	0.697	78.4	91.2	0.925
	SSL-linear (Frozen)	75.7	75.7	0.514	81.7	69.7	0.828
	SSL-linear	**86.1**	**86.1**	**0.723**	82.3	**89.9**	**0.939**
7:3 (Group 3)	ResNet50 (No Pre)	83.2	83.2	0.664	75	91.3	0.921
	SSL-linear (No Pre)	83.8	83.8	0.677	75.4	**92.3**	0.931
	SSL-linear (Frozen)	74.6	74.6	0.493	**82.8**	66.5	0.82
	SSL-linear	**85.4**	**85.4**	**0.709**	82.6	88.3	**0.932**

The bold term represents the highest score within the same group.

and labeled training sets (Table 2, Group 2 and Group 3). Notably, the pretraining set was derived from the original training set in PCam. The single ResNet50 could not perform self-supervised learning from the pretraining set, so when we compared ResNet50 with SSL algorithms, ResNet50 only learned features from the training set, which was identical to patches in the training sets of other SSL algorithms. The results presented that in both the 5:5 and 7:3 proportions we used in this task, SSL linear achieved the best overall performance. The AUC, Acc, B_Acc, κ score were 0.939, 86.1%, 86.1%, 0.723 for the 5:5 proportion and 0.932, 85.4%, 85.4%, 0.709 for the 7:3 proportion, which were higher than other groups. It was worth noting that AUC, Acc, B_Acc, κ score, and spe of SSL-linear in the 7:3 proportion group were higher than ResNet50 in the 5:5 proportion group, indicating that SSL-linear utilized less labeled patches but achieved a better performance than ResNet50. Although the performance of SSL-linear didn't exceed that of the other four state-of-the-art supervised learning algorithms, SSL-linear utilized only half or even less labeled data to achieve accuracy with a gap smaller than 5%. The results proved that SSL-Linear was both valid and feasible for patch-level classification in pathological slide images, even with a limited amount of labeled data. Detailed information is reported in Table 2.

Patch-level classification of ZJU-2

The dataset distribution is summarized in Table 1. The whole set contained 422,168 patches from 120 unlabeled images and 102139 patches from 72 labeled images. The Acc, B_Acc, κ score, Spe, Sen, and AUC were calculated to evaluate and compare SSL-linear and five CNNs (Figure 3). After pretraining, SSL-linear achieved the best performance compared to five CNNS with identical training set, indicating the positive effect of the pretraining round. The Acc, B_Acc, κ score, Sen, and AUC were 90.9%, 90.7%, 0.817%, 85.2%, and 0.981 for SSL-linear, higher than other groups. Detailed information is reported in Figure 3.

Whole-slide-image-level classification of ZJU-2

In a real-world clinical setting, clinicians worry about the diagnosis of a certain slide instead of the small patches. Thus, we evaluated the WSI classification ability of our algorithm and compared it to five CNNs (Figure 4). The ROC curve was plotted, and AUC was calculated. The AUCs for SSL-linear, VGG16, ResNet18, and ResNet50 were 0.964, 0.935, 0.891, and 0.938, indicating that SSL-linear achieved the best performance in the WSI-level classification task. For 32 WSIs in the testing set of ZJU-2, SSL-linear failed to diagnose two malignant cases. Other metrics were calculated: Spe 100%; Sen 75%; Acc 93.8%; and κ score 0.818.

Visualization heatmap

To address the clinical scenario and increase the interpretability of the diagnosis results of our algorithm, we generated a probabilistic heatmap by integrating the corresponding patches. The melanoma area in the slides was highlighted red and indicated whether the surgical margin

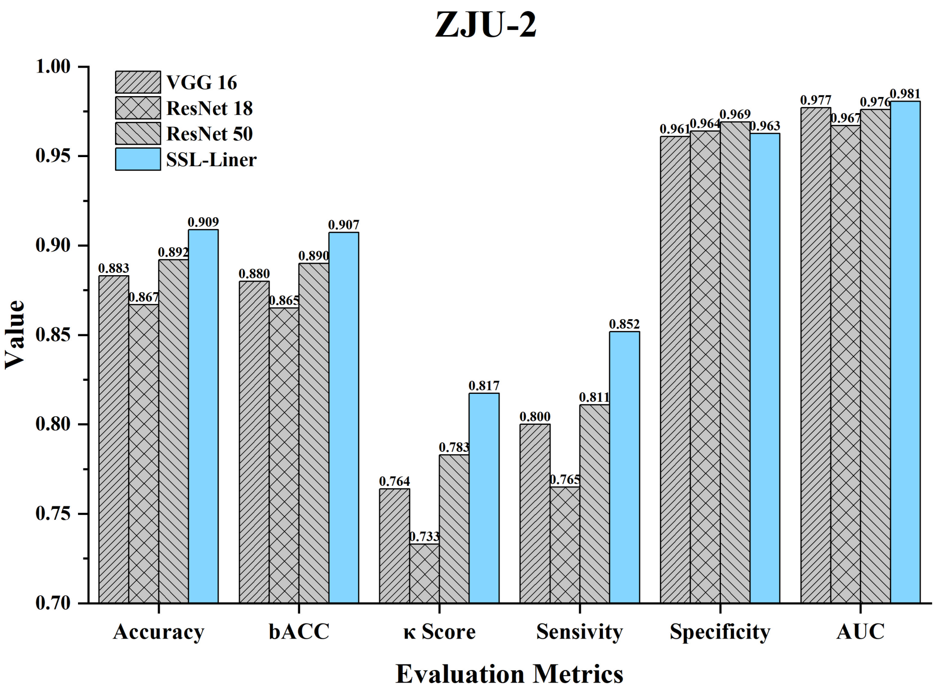


FIGURE 3
Comparison of different metrics for SSL-linear and 5 CNNs at the patch-level testing set of ZJU-2. κ, unweighted Cohen's kappa; Acc, accuracy; AUC, area under the receiver operating characteristic curve; B_Acc, balanced accuracy; CNN, convolutional neural network; ZJU-2, The Second Affiliated Hospital, Zhejiang University School of Medicine.

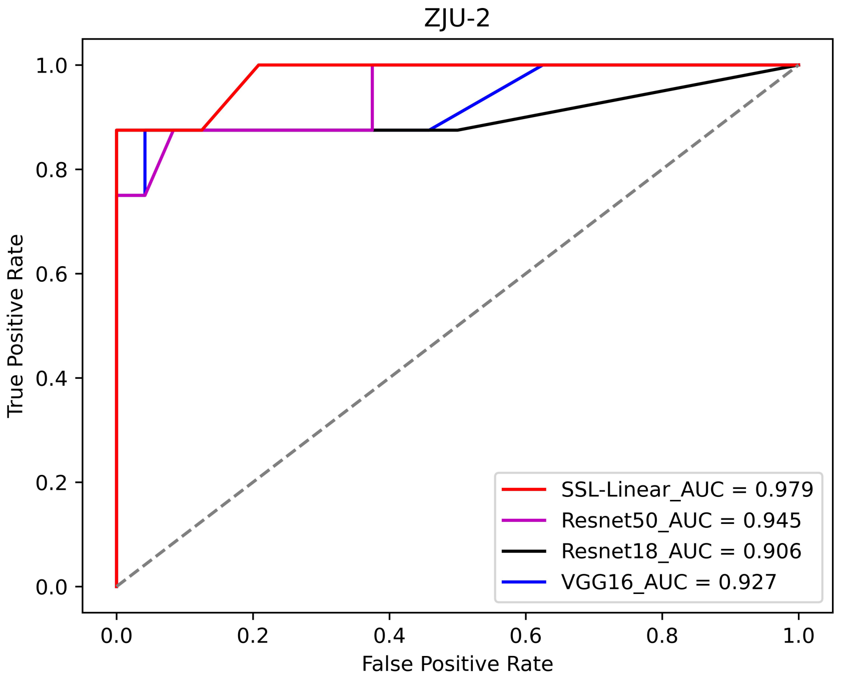


FIGURE 4
The receiver operating characteristic (ROC) curves of SSL-linear and 5 CNNs. Performance of SSL-linear, VGG16, ResNet18, and ResNet50 for melanoma detection for WSIs from ZJU-2. AUC, the area under the receiver operating characteristic curve; ZJU-2, The Second Affiliated Hospital, Zhejiang University School of Medicine.

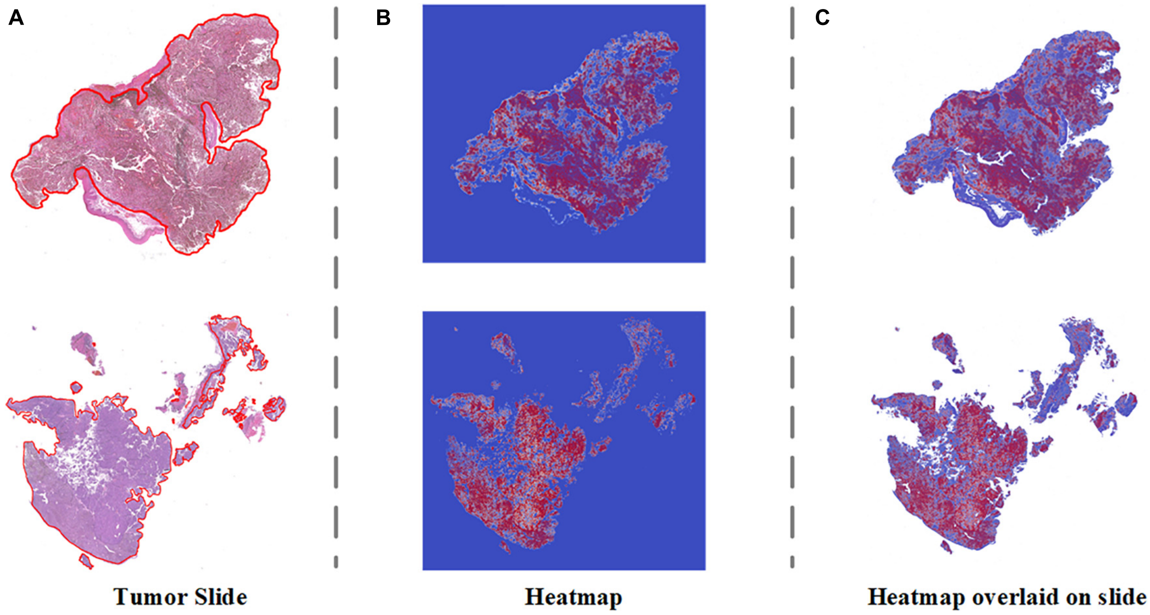


FIGURE 5
Visualization heatmap of pathological slides based on SSL. **(A)** The original pathological slide with tumor area delineated (H&E staining, ×40 scanned). **(B)** Probabilistic heatmap of the tumor slides generated by the algorithm. Red indicates higher malignancy. **(C)** Overlap of the tumor slide image and probabilistic heatmap.

was negative. Figure 5 demonstrates how our algorithm suggests melanoma areas by heightening the malignant zone. Figures 5A–C represent the original tumor slide image, the corresponding probabilistic heatmap and the overlap image, respectively. The overlapping image indicates that the prediction area of our algorithm corresponds to the delineation area.

Discussion

In this study, we trained a self-supervised learning model with a limited number of labeled images and developed a diagnostic system to detect eyelid MM in pathological slides. By comparing the classification ability of VGG16 ResNet18 and Resnet50 in PCam, ResNet50 was selected as the backbone for our pathologic diagnosis algorithm to generate SSL-linear. SSL-linear is based on BYOL and requires two identical CNNs (ResNet50 in this study) with a different set of weights in the pretraining round. In the patch-level classification task, SSL-linear displayed higher diagnostic accuracy even with fewer labeled images than the traditional ResNet50 classifier. We also introduced two state-of-the-art fully supervised algorithms (17, 18) to compare the performance of PCam. While the algorithms are closed source and utilize a training set 2 or 3 times larger than SSL-linear, the performance gap in Acc is relatively acceptable ($< 5\%$). It is valid and feasible for SSL-linear to make patch-level classifications in pathological slides. When applying to the ZJU-2 dataset, SSL-linear also demonstrated high diagnostic ability with the approximate 4:1 proportion of pretraining (unlabeled) and training (labeled) set in the patch-level and gigapixel WSI-level classification tasks. The computing systems that are used to solve problems in AI are opaque (19). This makes the diagnosis provided by the algorithm hard to convince both doctors and patients. To address this issue, we engineered our system to design a probabilistic heatmap highlighting malignant areas for pathologists. The emphasis on the area merits extra attention, and the indication of the negative margin is especially meaningful in highly lethal cancers like melanoma in our case.

Recently, there has been constant progress in self-supervised learning methods, such as contrastive learning, clustering, Simple Siamese networks, BYOL, etc., optimizing the performance of SSL algorithms and allowing transfer learning for different tasks (15, 20–22). In the medical field, SSL has been used to solve different problems with the type of input data (12, 13, 23, 24). Among these studies, pathological images have uniqueness for the following reasons. First, the pathology department encounters a huge quantity of slides, most of which won't be scanned to transfer into WSIs. Second, WSIs are at a gigapixel level with enormous information. Therefore, making annotations of pathologic slides is laborious, time-consuming, and requires a strong medical background. Although some slides were transferred into WSIs, most WSIs were not labeled; third, histopathological interpretation remains the gold standard for diagnosing some diseases. Currently, most research groups focus on improving the accuracy in the field of automatic pathological diagnosis, and different kinds of pathological images have been utilized. For instance, Ström et al. used labeled biopsy for algorithms to diagnose and grade prostate cancer (24); Kather et al. used deep learning to predict whether patients with gastrointestinal cancer respond well to immunotherapy (25).

Despite various motivations, most studies relied on sufficient ground truth labels of WSIs, which is difficult to attain in clinical scenarios. SSL does not require as many labeled WSIs as fully supervised learning and thus demonstrates the natural advantages of dealing with WSI's diagnosis. Some algorithms based on SSL have been applied to pathology. For example, Wataru et al. used their SSL-based algorithm to predict the pathological diagnosis of patients with interstitial lung disease. The algorithm achieved an AUC of 0.90 in the validation set and 0.86 in the testing set in diagnosing usual interstitial pneumonia with an approximate 1:2 proportion of pretraining and training set (4:1 proportion of pretraining and training set in the ZJU-2 dataset) (26, 27). However, due to the difference in task and labeling strategies, we cannot directly compare the performance of our algorithm to other studies.

Moreover, most studies in this field have focused on predicting common diseases. However, compared with common diseases, which are more unlikely for pathologists to misdiagnose, eyelid MM, the less common and dangerous cancer, presents a more urgent need for automated diagnosis or auxiliary diagnosis due to the lack of experience in encountering MM. Besides, our algorithm SSL-linear makes good use of unlabeled WSIs in reducing the burden of annotation and enhancing data utilization while achieving considerable performance in diagnosis.

To the best of our knowledge, it is the first study to apply self-supervised learning algorithms to ocular pathological research. With the approximate 4:1 proportion of pretraining and training set, SSL-linear achieved high accuracy at detecting MM area both in a patch (98.1%) and at WSI level (93.8%), out-competing the other five traditional CNNs. SSL-linear shows considerable diagnostic ability with limited labeled input data, not only easing the burden of annotating many gigapixel images but also providing relatively reliable diagnostic support for pathologists, especially those less experienced. Additionally, our diagnostic system takes only minutes to generate the output prediction results together with a clear probabilistic heatmap. For patients with MM, our diagnostic system can potentially reduce the probability of misdiagnosis and diagnostic omission, thus promoting the early treatment of MM. For clinicians, they could take advantage of telemedicine for rapid intraoperative consultation feedback. For pathologists, the highly malignant area indicated by the heatmap is also helpful in writing the pathological report and confirming the diagnosis.

Furthermore, it could raise the doctors' awareness of eyelid MM, a relatively less common cancer with a high mortality rate, and prioritize samples with higher malignant potential for senior pathologists. Despite the advantages of the automatic diagnostic system, human pathologists' work is still irreplaceable and has its own superiority. In a real clinical setting, the challenging cases will be reviewed by multiple pathologists with the help of immunohistochemistry, molecular information, or even genetic information in addition to H&E staining, while the algorithms only make classifications from the presentation of pathological slides. The primary purpose of developing a computer-aided diagnosis system is to assist human pathologists. The implementation of a self-supervised algorithm not only reduces the annotation burden and need for pathological expertise but also, which is more important, increases the data availability of future AI studies. The self-supervised design makes previously useless unlabeled data useful in the pretraining stage. It has the potential to be used in the broadening of disease types (e.g., basal cell carcinoma, squamous cell carcinoma, etc.) and task types (e.g., semantic segmentation of tumor areas in pathologic images based on SSL). From the technical aspect, the combination and comprehensive analysis of multimodal data (H&E staining, immunohistochemical staining, and non-image data like omics data) will be the future research focus.

Limitations

This study still had several limitations. First, the performance of the algorithm needed improvement as there is still a gap when compared to the state-of-the-art fully supervised learning algorithms. However, to the best of our knowledge, there has been no previous implied SSL algorithm to PCam as a benchmark. When SSL-linear was compared to closed-source fully supervised learning on a public data set, the performance gap was greatly affected by the disparity in training set size. In addition, to prove that SSL-linear achieves better performance than the traditional CNN, more groups with different proportions of pretraining and training sets could be organized both in the ZJU-2 data set and PCam. Second, this study did not include external validation, partly due to the lack of related case slides and difficulties in acquiring such pathological images from external. In the future, the diagnostic ability would be tested on data sets from independent sources (with different races, ages, etc.) to prove the generalization ability.

Additionally, the performance difference of the algorithm could be further investigated based on the above-mentioned different groups, not just malignant or non-malignant groups. Third, this study's total sample size of eyelid MM was relatively small compared with deep learning studies of other image types. This is limited by the inherent low incidence of eyelid MM. Despite this, all pathological slides are at gigapixel size with large information density and are different from each other; in other words, a total of 524,307 patches as input is relatively sufficient for SSL-linear to achieve considerable performance. Therefore, our sample size is acceptable for a pathological study.

Finally, our algorithm can only make a binary classification in this study. In the future, more disease types, including basal cell carcinoma or squamous cell carcinoma, will be introduced to validate the expendability.

In conclusion, SSL-linear was generated and demonstrated considerable performance with higher accuracy than traditional CNNs in distinguishing between benign and malignant eyelid lesions. With less labeled input data and an SSL framework, developing such a diagnostic system is relatively labor-saving and cost-efficient. The implementation of refined algorithms could be further applied to help diagnose various ophthalmic and non-ophthalmic malignancies.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

This study was approved by The Second Affiliated Hospital, Zhejiang University School of Medicine (ZJU-2) Ethics Committee (No. Y2019-195) and the study adhered to the Declaration of Helsinki. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

LW, ZJ, and JY were responsible for study design. LW, AS, and ZL were responsible for data collection and annotation. LW, ZJ, YW, ReG, and RuG analyzed the data. LW, ZJ, GJ, and JY drafted the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Key Research and Development Program of China (2019YFC0118401), National Natural Science Foundation Regional Innovation and Development Joint Fund (U20A20386), Provincial Key Research and Development Program of Zhejiang Province (2019C03020 and 2021C03199), Natural Science Foundation of Zhejiang Province of China (LQ19H150002), and Applied Research of Public Welfare Technology of Zhejiang Province (LGF22H120007).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Cummins DL, Cummins JM, Pantle H, Silverman MA, Leonard AL, Chanmugam A. Cutaneous malignant melanoma. *Mayo Clin Proc.* (2006) 81:500–7.

2. Mancera N, Smalley KSM, Margo CE. Melanoma of the eyelid and periocular skin: histopathologic classification and molecular pathology. *Surv Ophthalmol.* (2019) 64:272–88. doi: 10.1016/j.survophthal.2018.12.002

3. Malhotra R, Chen C, Huilgol SC, Hill DC, Selva D. Mapped serial excision for periocular lentigo maligna and lentigo maligna melanoma. *Ophthalmology.* (2003) 110:2011–8.

4. Catalyürek U, Beynon MD, Chang C, Kurc T, Sussman A, Saltz J. The virtual microscope. *IEEE Trans Inf Technol Biomed.* (2003) 7:230–48.

5. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med.* (2019) 25:1301–9. doi: 10.1038/s41591-019-0508-1

6. Wang L, Ding L, Liu Z, Sun L, Chen L, Jia R, et al. Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. *Br J Ophthalmol.* (2020) 104:318–23. doi: 10.1136/bjophthalmol-2018-313706

7. Li W, Yang Y, Zhang K, Long E, He L. Dense anatomical annotation of slit-lamp images improves the performance of deep learning for the diagnosis of ophthalmic disorders. *Nat Biomed Eng.* (2020) 4:767–77. doi: 10.1038/s41551-020-0577-y

8. Jiang YQ, Xiong JH, Li HY, Yang XH, Yu WT, Gao M, et al. Recognizing basal cell carcinoma on smartphone-captured digital histopathology images with a deep neural network. *Br J Dermatol.* (2020) 182:754–62. doi: 10.1111/bjd.18026

9. Bansal P, Vanjani A, Mehta A, Kavitha JC, Kumar S. Improving the classification accuracy of melanoma detection by performing feature selection using binary Harris hawks optimization algorithm. *Soft Comput.* (2022) 26:8163–81. doi: 10.1007/s00500-022-07234-1

10. Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe B, Matas J, Sebe N, Welling M editors. *European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9910 LNCS.* Berlin: Springer (2016). p. 69–84.

11. Sermanet P, Lynch C, Chebotar Y, Hsu J, Jang E, Schaal S, et al. Time-contrastive networks: self-supervised learning from video. In: *Proceedings of the*

IEEE International Conference on Robotics and Automation. Brisbane, QLD (2018). p. 1134–41. doi: 10.1109/ICRA.2018.8462891

12. Chatrian A, Colling RT, Browning L, Alham NK, Sirinukunwattana K, Malacrino S, et al. Artificial intelligence for advance requesting of immunohistochemistry in diagnostically uncertain prostate biopsies. *Modern Pathol.* (2021) 34:1780–94. doi: 10.1038/s41379-021-00826-6

13. Santilli AML, Jamzad A, Sedghi A, Kaufmann M, Logan K, Wallis J, et al. Domain adaptation and self-supervised learning for surgical margin detection. *Int J Comput Assist Radiol Surg.* (2021) 16:861–9. doi: 10.1007/s11548-021-02381-6

14. Chung JS, Zisserman A. Lip reading in profile. *Br Mach Vis Conf.* (2017) 2017:1–12. doi: 10.5244/c.31.155

15. Grill J-B, Strub F, Altché F, Tallec C, Richemond PH, Buchatskaya E, et al. Bootstrap your own latent: a new approach to self-supervised Learning. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY (2020).

16. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. (2017) 318:2199–210.

17. Veeling BS, Linmans J, Winkens J, Cohen T, Welling M. Rotation equivariant CNNs for digital pathology. *Artif Intell Lect Notes Bioinf.* (2018) 11071:210–8.

18. Mohamed M, Cesa G, Cohen TS, Welling MA. Data and compute efficient design for limited-resources deep learning. *arXiv* [Preprint]. (2020). arXiv:2004.09691.

19. Zednik C. Solving the black box problem: a normative framework for explainable artificial intelligence. *Philos Technol.* (2011) 34:265–88.

20. Chen X, He K. Exploring simple siamese representation learning[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Nashville, TN (2021). p. 15750–8.

21. Chen T, Kornblith S, Norouzi M, Hinton GA. Simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning, PartF16814.* Vienna: ICML (2020). p. 1575–85.

22. Caron M, Bojanowski P, Joulin A, Douze M. Deep clustering for unsupervised learning of visual features. *Lect Notes Comput Sci Artif Intell Lect Notes Bioinf.* (2018) 11218 LNCS:139–56.

23. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng.* (2021) 5:555–70. doi: 10.1038/s41551-020-00682-w

24. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol.* (2020) 21:222–32.

25. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med.* (2019) 25:1054–6.

26. Uegami W, Bychkov A, Ozasa M, Uehara K, Kataoka K, Johkoh T, et al. MIXTURE of human expertise and deep learning—developing an explainable model for predicting pathological diagnosis and survival in patients with interstitial lung disease. *Modern Pathol.* (2022) 35:1083–91. doi: 10.1038/s41379-022-01025-7

27. Srinidhi CL, Kim SW, Chen FD, Martel AL. Self-supervised driven consistency training for annotation efficient histopathology image analysis. *Med Image Anal.* (2022) 75:102256. doi: 10.1016/j.media.2021.102256

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fmed.2022.974485

Radiomics analysis of contrast-enhanced CT scans can distinguish between clear cell and non-clear cell renal cell carcinoma in different imaging protocols

Bettina Katalin Budai[1]*, Róbert Stollmayer[1], Aladár Dávid Rónaszéki[1], Borbála Körmendy[1], Zita Zsombor[1], Lõrinc Palotás[1], Bence Fejér[1], Attila Szendrõi[2], Eszter Székely[3], Pál Maurovich-Horvat[1] and Pál Novák Kaposi[1]

[1]Department of Radiology, Faculty of Medicine, Medical Imaging Centre, Semmelweis University, Budapest, Hungary, [2]Department of Urology, Faculty of Medicine, Semmelweis University, Budapest, Hungary, [3]Department of Pathology, Forensic and Insurance Medicine, Faculty of Medicine, Semmelweis University, Budapest, Hungary

OPEN ACCESS

EDITED BY
Jun Cheng,
Shenzhen University, China

REVIEWED BY
Mohammad Arafat Hussain,
Harvard Medical School, United States
Zhenyuan Ning,
Southern Medical University, China

*CORRESPONDENCE
Bettina Katalin Budai
budai.bettina@med.semmelweis-univ.hu

SPECIALTY SECTION
This article was submitted to Pathology, a section of the journal Frontiers in Medicine

RECEIVED 21 June 2022
ACCEPTED 28 September 2022
PUBLISHED 13 October 2022

CITATION
Budai BK, Stollmayer R, Rónaszéki AD, Körmendy B, Zsombor Z, Palotás L, Fejér B, Szendrõi A, Székely E, Maurovich-Horvat P and Kaposi PN (2022) Radiomics analysis of contrast-enhanced CT scans can distinguish between clear cell and non-clear cell renal cell carcinoma in different imaging protocols. *Front. Med.* 9:974485. doi: 10.3389/fmed.2022.974485

Introduction: This study aimed to construct a radiomics-based machine learning (ML) model for differentiation between non-clear cell and clear cell renal cell carcinomas (ccRCC) that is robust against institutional imaging protocols and scanners.

Materials and methods: Preoperative unenhanced (UN), corticomedullary (CM), and excretory (EX) phase CT scans from 209 patients diagnosed with RCCs were retrospectively collected. After the three-dimensional segmentation, 107 radiomics features (RFs) were extracted from the tumor volumes in each contrast phase. For the ML analysis, the cases were randomly split into training and test sets with a 3:1 ratio. Highly correlated RFs were filtered out based on Pearson's correlation coefficient ($r > 0.95$). Intraclass correlation coefficient analysis was used to select RFs with excellent reproducibility (ICC $\geq$ 0.90). The most predictive RFs were selected by the least absolute shrinkage and selection operator (LASSO). A support vector machine algorithm-based binary classifier (SVC) was constructed to predict tumor types and its performance was evaluated based-on receiver operating characteristic curve (ROC) analysis. The "Kidney Tumor Segmentation 2019" (KiTS19) publicly available dataset was used during external validation of the model. The performance of the SVC was also compared with an expert radiologist's.

Results: The training set consisted of 121 ccRCCs and 38 non-ccRCCs, while the independent internal test set contained 40 ccRCCs and 13 non-ccRCCs.

For external validation, 50 ccRCCs and 23 non-ccRCCs were identified from the KiTS19 dataset with the available UN, CM, and EX phase CTs. After filtering out the highly correlated and poorly reproducible features, the LASSO algorithm selected 10 CM phase RFs that were then used for model construction. During external validation, the SVC achieved an area under the ROC curve (AUC) value, accuracy, sensitivity, and specificity of 0.83, 0.78, 0.80, and 0.74, respectively. UN and/or EX phase RFs did not further increase the model's performance. Meanwhile, in the same comparison, the expert radiologist achieved similar performance with an AUC of 0.77, an accuracy of 0.79, a sensitivity of 0.84, and a specificity of 0.69.

Conclusion: Radiomics analysis of CM phase CT scans combined with ML can achieve comparable performance with an expert radiologist in differentiating ccRCCs from non-ccRCCs.

KEYWORDS

renal cell carcinoma, computed tomography, radiomics analysis, texture analysis, machine learning, artificial intelligence

Introduction

Kidney cancers are one of the most common malignancies in the world accounting for approximately 2.2% of annual cancer diagnoses (431 thousand/year) and 1.8% of cancer-related mortality (179 thousand/year) worldwide. It is almost twice as common in males than in females making it the 11th highest incidence of cancer in men and the 16th in women (1).

Due to the increasing accessibility of non-invasive diagnostic procedures nowadays up to 50% of renal neoplasms are incidentally discovered (2). Many of these small renal masses are benign, but because of their size, they are hard to characterize using imaging modalities increasing the importance of biopsy to select low-risk patients for active surveillance (3). At the time of diagnosis, approximately 15% of patients already have distant metastases (4). Accurate preoperative staging is crucial for making an appropriate treatment decision. For accurate staging – including the assessment of local invasiveness, lymph node involvement, and presence/absence of distant metastases – , contrast-enhanced thoraco-abdominopelvic CT examination is mandatory in patients with indeterminate renal mass (2, 5).

The histologic classification and grading of renal tumors are also important, as the prognostic and therapeutic implications vary among histologic subtypes. The current 2016 World Health Organization (WHO) classification differentiates between numerous types of kidney tumors including mesenchymal, metanephric, nephroblastic, neuroendocrine, and renal cell tumors among others (3).

Renal cell carcinoma (RCC) is the most common among the neoplastic diseases of the kidney, with approximately 90% of them being diagnosed as RCC (6). RCC is a collective term defining a heterogenous group of neoplasms including 14 subtypes (3) with drastically different histologic appearance, genetics, and prognosis, all originating from the renal tubular epithelium (6). The most common subtypes of RCC are clear cell renal cell carcinoma (ccRCC), papillary cell renal cell carcinoma (pRCC), and chromophobe cell renal cell carcinoma (chRCC), respectively, accounting for about approximately 75, 15, and 5% of all RCC cases (7).

Previous studies proved that the histologic subtype is an independent predictor of patient survival, and patients with ccRCC have a poorer prognosis compared to those with pRCC or chRCCs (8, 9), also patients with ccRCC are most likely to have distant metastasis at the time of radical nephrectomy (10). Due to the markedly higher biological aggressiveness of ccRCC compared to other subtypes, recent practice guidelines divide RCCs into two main groups as ccRCC and non-ccRCC (2, 11).

In the case of advanced RCC, treatment options have been rapidly expanded in the past decades. High-dose bolus interleukin-2 therapy has brought continued good results since approval for the treatment of metastatic RCC in 1992 (12) followed by the era of molecularly targeted therapies and more recently, the era of immunotherapeutic agents (13). Molecularly targeted therapies including Vascular Endothelial Growth Factor (VEGF) targeted tyrosine-kinase inhibitors such as bevacizumab, sunitinib, and pazopanib have been used with great success in patients with metastatic ccRCC, which is currently the recommended first-line standard-of-care treatment according to the European Society for Medical Oncology (ESMO) in patients with good risk (2). Then, novel immunotherapeutic agents revolutionized the treatment of advanced ccRCC (14). The ESMO guidelines recommend combined immune-checkpoint inhibitor antibody therapy (ipilimumab + nivolumab) as first-line treatment in patients with intermediate or poor-risk (2), and since the same results can be achieved using combined immune checkpoint inhibitors with lower toxicity, the usage of cytokine monotherapy diminished (14). Even though there is ample evidence available for the efficiency of sunitinib as a treatment for metastatic ccRCC, other less common renal carcinomas are less researched since they are most often excluded from the controlled phase III trials. Smaller prospective studies, however, suggest that VEGF inhibitors and mammalian target of rapamycin inhibitors are also beneficial in these cases (2). However, pRCCs show a worse response to VEGF-targeted antiangiogenic agents than ccRCCs (15).

Therefore, the non-invasive, imaging-based differentiation between tumor subtypes could facilitate the prediction of patient prognosis and guide clinicians in therapeutic decision-making and follow-up strategies (16). It has been proved that the different subtypes of RCCs have different contrast enhancement dynamics, ccRCCs have peak enhancement on the corticomedullary phase, meanwhile, pRCCs and chRCCs reach the peak during the nephrographic phase (17). Previous studies showed that, relative contrast enhancement of kidney tumors to the renal cortex (18) and CT imaging traits such as heterogeneous contrast enhancement, enhancement degree in corticomedullary phase, the presence of necrosis, and the presence of calcification show association with RCC subtypes (19). However, the morphology-based, conventional radiological evaluation of CT scans is subjective, has low specificity in differentiating RCC subtypes (20), and is highly dependent on the expertise of the radiologists (21).

In 2012, the term radiomics was introduced by Lambin et al. which refers to the automated analysis of medical images by the extraction of an extensive number of quantitative features that can objectively describe the given region of interest (ROI) (22). Radiomics as per definition is the mining and analysis of quantitative features from radiologic images, to improve clinical decision-making by identifying predictive imaging biomarkers and constructing different diagnostic and prognostic models. This novel technique has the potential to detect subtle differences in tissue texture that may not be detected by the human eye (22).

A typical radiomic study comprises the following main steps: medical image acquisition, image pre-processing, segmentation, feature extraction, feature selection, exploratory analysis, and model building and evaluation (23). Conventional radiomics analysis requires lesion segmentation in order to compute hand-crafted radiomics features. The segmentation can be performed either manually by using semi-automatic tools, or fully automatically with the help of convolutional neural networks. In radiomics studies of kidney tumors, the most widely used method is still the manual segmentation (24).

Radiomics analysis allows the extraction of a huge number of quantitative features from the selected volume of interest (VOI) that refer to the intensity histogram, the shape, or the texture of a certain lesion. The definitions and the mathematical formulas of radiomics features can differ between studies, therefore the Imaging Biomarker Standardization Initiative (IBSI) was established as an independent international collaboration aiming to standardize the extraction of quantitative imaging biomarkers to improve the reproducibility of radiomics studies. For a more detailed description of radiomics features, we refer the readers to the Reference Manual of the IBSI updated in 2020 (25). Radiomics analysis is most commonly applied to CT scans given its wide availability. CT texture analysis (CTTA) on contrast-enhanced CT scans also provides a quantitative description of the tissue contrast enhancement distribution after contrast-agent injection.

Radiomics is usually combined with machine learning algorithms for prediction model building. However, the usage of a large number of radiomics features often results in overfitting of the prediction model; therefore the number of features must be effectively reduced before model building (26). As an initial feature-selection step, it is recommended to filter out highly correlated, redundant features (23, 26). The most popular supervised feature selection methods are the model-based wrappers including the so-called recursive feature elimination algorithm that is used to select the optimal subset of predictive features that maximize the prediction performance; and the embedded algorithms such as the least absolute shrinkage and selection operator (LASSO) regression that allows selecting the most predictive features based on the feature importance score (23).

The most widely used conventional machine learning algorithms for prediction model building are logistic regression, LASSO, random forest (RFC), and support vector machine (SVC) classifiers (27).

Previously published studies have focused mainly on distinguishing between benign and malignant renal lesions (28–30) or on identifying aggressive tumor features of ccRCCs (31–37), and only a minority of studies have sought to distinguish between subtypes of RCC (20, 38–41). A few studies also showed that radiomics analysis combined with machine learning could facilitate the non-invasive diagnostics of kidney cancers including both classification of renal tumors, prediction of nuclear grade, identification of patients with poor prognosis, and prediction of treatment response (42, 43). However, most of the previously published studies had a single-center study design and used only internal validation for model evaluation and have not validated their results on external test cases (24, 43).

Yu et al. were among the first who used CT texture analysis for distinguishing between RCC subtypes (41). The authors performed radiomics analysis on 10 selected cross-sectional areas of the tumors in the nephrographic (NG) phase and extracted 43 features. Their SVC trained by all the 43 radiomics features achieved AUCs of 0.91, 0.92, and 0.85 in differentiating between ccRCCs vs. pRCCs, chRCCs and oncocytomas; pRCCs vs. ccRCCs, chRCCs and oncocytomas, and chRCCs vs. pRCCs, ccRCCs, and oncocytomas, respectively. Yu et al. demonstrated the ability of first-order statistics and texture features to predict RCC subtypes (41). By analyzing triphasic CT scans of 143 ccRCCs and 54 non-ccRCCs, Chen et al. illustrated that the radiomics features extracted from the corticomedullary (CM) phase have similar diagnostic ability compared to those extracted from the NG phase in differentiating between ccRCCs and non-ccRCCs (38). In their recent study, Wang et al. analyzed 147 ccRCCs and 43 non-ccRCCs and built a RFC, an SVC, and a logistic regression algorithm-based machine learning model from four selected radiomics features. The models achieved good to excellent results on the internal test dataset with AUC of 0.841–0.909 (20), and the authors also demonstrated that these radiomics-based machine learning models can overperform the diagnostic performance of an expert radiologist (AUC of 0.69). However, in these single-center studies, the machine learning prediction models were not validated on independent external test cases.

External validation of the machine learning models was completed in a two-center study by Li et al., who performed 3D texture analysis on both the unenhanced (UN), CM, and NG phase CT scans of 170 patients (40). In this study, either the Boruta or the minimum redundancy maximum relevance ensemble (mRMRe) algorithms were used to select the most relevant radiomics features. RFC models built in this study were tested on 85 independent external test cases from another hospital. The Boruta-based RFC achieved excellent performance with an AUC of 0.949 while the mRMRe-based RFC achieved an AUC of 0.851. The two sets of selected radiomics features differed significantly, suggesting that there is a huge difference in the performance of the feature selection algorithms, which significantly affects the performance of the machine learning classifier. These results also indicate that the CM features have higher diagnostic ability compared to NG phase features in the differentiation of ccRCCs from non-ccRCCs (40).

Kocak et al. were among the first, who validated their machine learning models' performance on publicly available datasets (39). In their retrospective study, the authors collected 48 ccRCCs, 13 pRCCs, and 7 chRCCs and performed CT texture analysis on UN and CM phase CT scans to differentiate between RCC subtypes. For external validation 26 cases (13 ccRCCs, 7 pRCCs, and 6 chRCCs) were selected from the TCGA public datasets including The Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) (44, 45), the TCGA-Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP) (44, 46), and the TCGA-Kidney Chromophobe (TCGA-KICH) (44, 47). The authors performed radiomics analysis on the largest cross-sectional areas of the tumors by extracting 275 radiomics features from both the UN and the CM phases. After feature selection, artificial neural network (ANN)-based and SVC-based prediction models were constructed for the differentiation between ccRCC and non-ccRCCs. The ANN algorithm-based model trained on CM phase features achieved an AUC of 0.822, while the SVC reached an AUC of 0.793 on the external test set (39).

Our study aimed to construct a 3D CTTA-based machine learning model for differentiating ccRCC from non-ccRCC that is generalizable and robust against different institutional imaging protocols. We aimed to demonstrate that our radiomics-based machine learning model can achieve comparable results with an expert radiologist. And the final aim of this study was to validate our prediction models on external test cases of a publicly available dataset to prove the models' reliability.

Materials and methods

Patient population

The institutional ethics committee of our university has approved the present study based on the World Medical Association guidelines and the Declaration of Helsinki, revised in 2000 in Edinburgh. As this is a retrospective study, the need for written patient consent was waived by the ethics committee. All patient data were analyzed anonymously.

Preoperative contrast-enhanced abdominal CT scans were retrospectively collected from patients who had undergone either radical or partial nephrectomy between 2008 January and May 2021 at our institution. Out of the patients who had undergone nephrectomy, 551 had available preoperative CT scans. The preoperative unenhanced UN, CM, and excretory (EX) phase CT scans in this study were obtained from the picture archiving and communication system (PACS) of our hospital. 346 cases were excluded due to the following exclusion criteria: diagnosed with benign kidney tumor ($n = 33$), diagnosed with other types of malignant kidney tumor ($n = 107$), nephrectomy due to other reason than tumor ($n = 75$), no available histopathologic report ($n = 30$), dual-phase (UN, CM, and EX) CT scan was not available ($n = 61$), underwent radiofrequency ablation ($n = 2$), damaged DICOM file ($n = 44$), incomplete coverage of the tumor ($n = 1$).

The final patient cohort included 209 patients diagnosed with either ccRCC, pRCC, or chRCC. The final histopathological diagnosis of RCC subtypes served as the reference standard. After nephrectomy, the whole tumor specimens were transferred to histological processing. The official pathology reports were retrospectively collected from the hospital information system. Three patients had two histologically proven tumors, therefore, the final dataset consisted of 161 ccRCCs, 34 pRCCs, and 17 chRCCs.

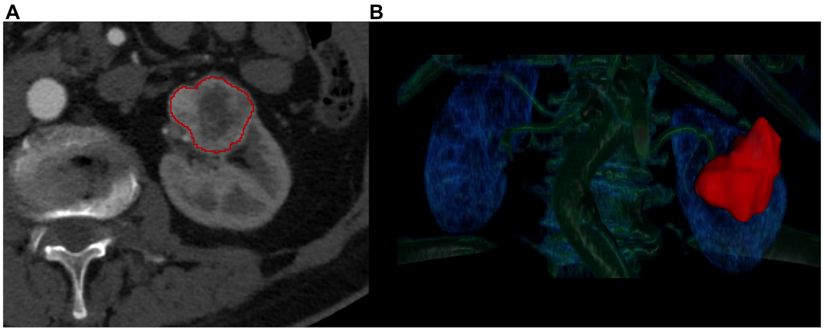


FIGURE 1
Manual segmentation of kidney tumors. The manual segmentation of the kidney tumors was completed on the corticomedullary phase axial CT scans **(A)**. The entire lesion volume was delineated slice-by-slice **(B)** in order to perform a three-dimensional radiomics analysis.

Imaging protocols

We examined the patients according to our routine diagnostic protocols with either a 16-slice Brilliance or a 64-slice Ingenuity Core 64 CT scanner (Philips Healthcare, Best, the Netherlands). The following acquisition parameters were used: tube voltage of 100–140 keV; automatic tube current modulation in the range of 105–977 mAs in CM, 93–918 mAs in UN, and 80–910 mAs in EX phase; collimation of 16 mm × 1.5 mm or 64 mm × 0.625 mm for the 16 and 64-slice scans, respectively. The 16-slice acquisitions were routinely reconstructed with filtered back projection (FBP) and 64-slice scans with the iDose4$^{\mathrm{TM}}$ hybrid iterative reconstruction kernel. The reconstructed slice thickness was 1.25–5 mm. A non-ionic, iodinated contrast agent (range of concentration: 350–370 mg/ml) was administered intravenously using a power injector with an injection rate of 1.5–3.5 ml/s, while the amount of the injected contrast media was adjusted to the body weight (0.5 g iodine/kg). After contrast agent administration, the CM phase was scanned at 30–45 s, and the EX phase at 300–480 s.

External test set

For the external validation of our machine learning prediction model, we included cases from the 2019 Kidney and Kidney Tumor Segmentation Challenge (KiTS19) public database (48, 49) that had available dual-phase (UN, CM, and EX phase) CT scans. We identified 75 cases with dual-phase CT scans, from those 69 cases were diagnosed with either ccRCC, pRCC, or chRCC. One case was excluded because the patient's position on the EX phase scan was prone instead of supine. The CT scans were performed by a variety of scanners including 19 different models from four vendors. The slice thickness varied between 1 and 7 mm, the tube voltage was between 100 and 140 keV, and the tube current varied between 95 and 747 mAs in the CM, 80–667 mAs in the UN, and 80–664 mAs in the EX phase scans. One patient had three tumors, and three had two tumors, therefore the final external test set consisted of 73 lesions. According to the available metadata, 50 of those were ccRCCs, 13 were pRCCs, and 10 were chRCCs. In the KiTS19 dataset, the binary segmentation masks were also available to all the tumors. The results shown here are in whole or part based upon data from the C4KC-KiTS dataset of The Cancer Imaging Archive (TCIA) (44, 48, 49).

Subjective classification

For the subjective, imaging feature-based analysis, an expert radiologist with over 10 years of experience in urologic imaging classified all the lesions of both internal and external test sets according to the RCC subtypes blinded to the patients' history, medical records, and to the results of tumor segmentation.

Image processing and radiomics analysis

Preoperative axial CT scans were anonymized and exported from the institutional PACS in Digital Imaging and Communications in Medicine (DICOM) format. The DICOM files were then converted to NIfTI file format for further image processing and analysis. The image processing and segmentation steps were completed by using the 3D Slicer software v.4.10.2 (50).

The entire volume of the tumors was segmented slice-by-slice on the CM phase scans. The segmentation of kidney tumors was performed by a trainee with 4 years of experience in tumor segmentation under the supervision of an expert radiologist with over 15 years of experience in abdominal and urologic imaging (**Figure 1**). The segmentation was performed by avoiding the edge of the tumor to avoid the inclusion of peripheral fat and partial volume effect. The UN and EX phase

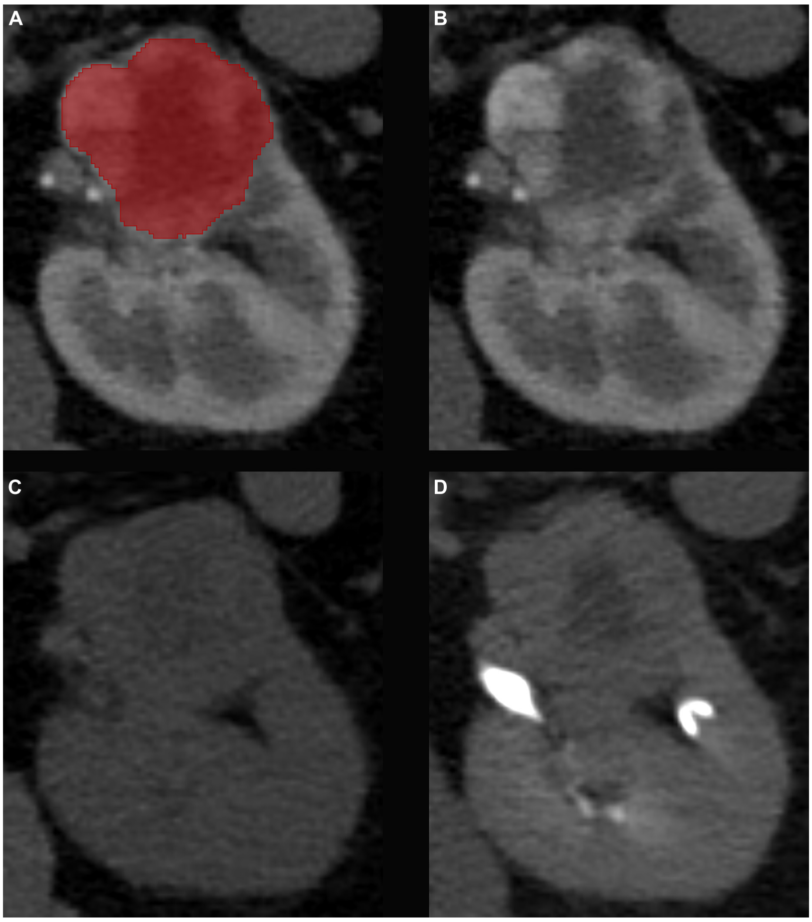


FIGURE 2
Results of the image co-registration. The manual segmentation of the kidney tumors was completed on the corticomedullary phase CT scans **(A)**. A non-rigid image co-registration was performed to fit the unenhanced **(B)** and excretory **(C)** phase CT scans to the corticomedullary phase as reference **(D)**.

CT scans were coregistered to the CM phase scans by using the Elastix extension of 3D Slicer (**Figure 2**).

To minimize the individual patient factors, the inter-scanner differences, and the difference between institutional imaging protocols, the voxel density values of the CM phase CT scans were normalized to the cortical density. In each case, the density of the renal cortex was measured by using 3–3 circular region of interests (ROI), then the mean cortical density was obtained by calculating the average value of the three measurements. In each case, the mean cortical density value was subtracted from the individual voxel intensity values.

For the radiomics analysis, the images were resampled to an isotropic voxel size of 1 mm × 1 mm × 1 mm to get rotation invariant radiomics features and to improve the robustness and reproducibility of the extracted features. The radiomics analysis was performed with the pyRadiomics package (51). A fixed bin width of 16 was used during the calculation of texture features. Altogether, 107 radiomics features were calculated from each phase scan, including 18 first-order histogram-based statistical features, 14 shape-based features, 24 gray-level co-occurrence matrix-based features (GLCM), 16 gray-level run-length matrix-based features (GLRLM), 16 gray-level size zone matrix-based features (GLSZM), 14 gray-level dependence matrix-based features (GLDM), and 5 neighboring gray-tone difference matrix-based features (NGTDM). Data is available in **Supplementary Table 1**.

Feature selection

Our feature selection method included three steps, all of which were completed by using solely the training set. First, highly correlated features were filtered out based on Pearson's correlation coefficients ($r > 0.95$). Then, reproducibility analysis was performed by using intraclass correlation coefficient (ICC) analysis. For the reproducibility analysis, the area of the segmented tumor masks was eroded by 1–1 voxel in each

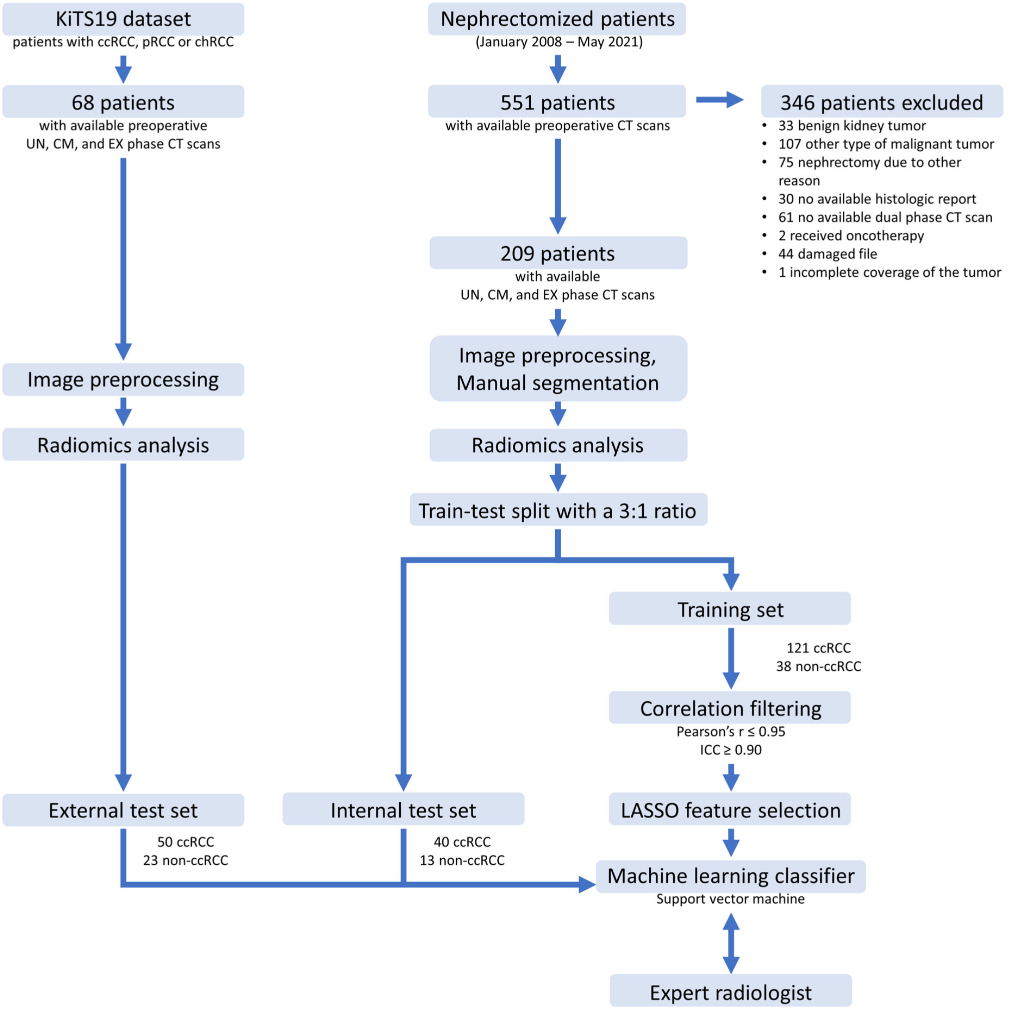

KiTS19 dataset
patients with ccRCC, pRCC or chRCC
68 patients
with available preoperative UN, CM, and EX phase CT scans
Image preprocessing
Radiomics analysis
External test set
50 ccRCC
23 non-ccRCC
Nephrectomized patients
(January 2008 – May 2021)
551 patients
with available preoperative CT scans
346 patients excluded
• 33 benign kidney tumor
• 107 other type of malignant tumor
• 75 nephrectomy due to other reason
• 30 no available histologic report
• 61 no available dual phase CT scan
• 2 received oncotherapy
• 44 damaged file
• 1 incomplete coverage of the tumor
209 patients
with available UN, CM, and EX phase CT scans
Image preprocessing, Manual segmentation
Radiomics analysis
Train-test split with a 3:1 ratio
Training set
121 ccRCC
38 non-ccRCC
Correlation filtering
Pearson's r ≤ 0.95
ICC ≥ 0.90
LASSO feature selection
Internal test set
40 ccRCC
13 non-ccRCC
Machine learning classifier
Support vector machine
Expert radiologist

FIGURE 3
Flowchart of the data analysis steps. ccRCC, clear cell renal cell carcinoma; pRCC, papillary cell renal cell carcinoma; chRCC, chromophobe renal cell carcinoma; KiTS, kidney tumor segmentation dataset; UN, unenhanced; CM, corticomedullary; EX, excretory; LASSO, least absolute shrinkage and selection operator; ICC, intraclass correlation coefficient.

direction as proposed previously (52, 53), and the radiomics feature extraction was repeated. The ICC was calculated for each radiomics feature based on a 2-way, single-rater, absolute agreement model. Only the features with excellent reproducibility defined as ICC value $\geq$0.90 were included in the wrapper-based feature selection step. The final step included either a least absolute shrinkage and selection operator (LASSO) algorithm, or a tuned ReliefF (TuRF) algorithm which selected the most relevant features based on their feature importance score. The optimal hyperparameter (λ) for LASSO feature selection was automatically determined on the training dataset by using the grid search method with 5-times repeated 5-fold stratified cross-validation. During hyperparameter tuning, negative mean squared error was used as a performance metric that the grid search tried to maximize.

Machine learning – Model building

For the machine learning-based analysis, the cases were randomly split into training and test sets with a 3:1 ratio. The radiomics features of the training dataset were standardized by centering around the mean with a unit standard deviation (SD). The test dataset was transformed using the hyperparameters from the training dataset. From the features selected by LASSO, SVC-based machine learning models were constructed to differentiate ccRCCs from non-ccRCCs. From the radiomics features selected by the TuRF algorithm, random forest classifier-based models were constructed. The hyperparameters of the classifiers were optimized with the grid search method based on the accuracy score during five-times repeated 5-fold stratified cross-validation on the training set. To overcome the class imbalance issue, balanced class-weights were used while fitting the models. The diagnostic performance of the models was evaluated on both the training set, the internal test set, and the external test set based on the receiver operating characteristic curve (ROC) analysis. During ROC analysis, the ccRCC data were set as the positive class, while the non-ccRCC as the negative class. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under curve (AUC) values were calculated. A two-tailed p-value < 0.05 indicated statistical significance. **Figure 3** shows the main steps of the data analysis.

Statistical analysis

The continuous variables in ccRCC and non-ccRCC patient groups were checked for homogeneity of variance with the F-test and normal distribution with the Shapiro-Wilk's test. Categorical variables were compared between the two groups with the chi-squared test and continuous variables with the Mann-Whitney U-test. The 95% confidence interval (CI) of the AUC values were calculated based on DeLong's method. The best threshold was determined based on the "closest top left" method; the point on the ROC curve closest to the top left corner of the plot was defined as min[(1-sensitivities)^2 + (1-specificities)^2]. The statistical comparisons between the ROC curves were performed according to the DeLong test. The threshold of $p < 0.05$ was applied to determine significance in all comparisons.

The statistical analysis was completed with "sklearn," "skrebate," "statmodels," and "scipy" packages written in Python (v.3.7.11.) computer language, and with "dplyr," "stats," "pROC," and "irr" packages written in R (v.3.6.3.) computer language.

Results

Patient population

The final study population contained 209 patients with 212 tumors (161 ccRCCs and 51 non-ccRCCs). There were no differences in patient age (p = 0.079) or sex (p = 0.9782) comparing ccRCCs with non-ccRCCs (**Table 1**). For the machine learning-based analysis, the cases were randomly split into training and test sets with a 3:1 ratio. The distribution of RCC subtypes in the training dataset was ccRCC in 121 cases and non-ccRCC in 38 cases (25 pRCC and 13 chRCC), meanwhile the internal independent test set contained 40 ccRCCs and 13 non-ccRCCs (9 pRCC and 4 chRCC).

From the KiTS19 public dataset 68 cases with 73 tumors were included in this study as an external test set. In the ccRCC group 33 (66.0%) patients were male and 17 (34.0%) were female, while in the non-ccRCC group, 10 (43.5%) were male and 13 (56.5%) were female (p = 0.069). The median age and interquartile range were 60.5 (23.2) years for ccRCCs and 53 (13.0) years for non-ccRCCs (p = 0.556).

Feature selection

During radiomics analysis, 107 radiomics features were extracted from both CM, EX, and UN phase scans. After filtering out the highly correlated and non-robust features, 39 CM, 38 EX, and 35 UN phase features remained. During hyperparameter tuning of the LASSO algorithm, the grid search defined 0.01 as the optimal λ value. In all three cases, an optimized LASSO algorithm (λ = 0.01) was used to select the most predictive radiomics features based on the feature importance score, which selected 10 CM phase, 5 EX phase, and 9 UN phase features. The selected radiomics features included both shape-based features, first-order statistics, and texture features in each case. The selected features are listed in **Table 2**.

Machine learning

The optimized SVC model (kernel: rbf, C: 500, gamma: 0.005) trained on the CM phase radiomics features achieved the highest prediction performance in differentiating ccRCCs from non-ccRCCs. During ROC analysis, its performance on the training set was AUC of 0.951 [95% CI: 0.913–0.989], accuracy of 0.925, sensitivity of 0.926, and specificity of 0.921 at threshold 0.655, it also achieved very good prediction rate on the internal independent test set with AUC of 0.873 [95% CI: 0.774–0.973], accuracy of 0.811, sensitivity of 0.90, and specificity of 0.539, and its diagnostic accuracy proved to be robust during validation on external test cases with AUC of 0.834, accuracy of 0.781, sensitivity of 0.800, and specificity of 0.739 (**Figure 4**). We also compared the diagnostic value of this model against the accuracy of an expert radiologist, which showed comparable results with no significant difference on either the internal (p = 0.866) or the external (p = 0.256) test sets (**Table 3**). On the internal test set, the expert radiologist achieved slightly better performance with an AUC of 0.886 (vs. 0.873), accuracy of 0.906 (vs. 0.811), sensitivity of 0.925 (vs. 0.90) and specificity of 0.846 (vs. 0.539), while on the external test set, the SVC slightly overperformed the expert radiologist, who achieved an AUC of 0.768 (vs. 0.834), accuracy of 0.795 (vs. 0.781), sensitivity of 0.84 (vs. 0.80), specificity of 0.696 (vs. 0.739) (**Figure 4**).

The optimized RFC model (criterion: entropy, n_estimators: 50) trained on the 10 CM phase radiomics features selected by the TuRF algorithm was able to distinguish between ccRCC vs. non-ccRCCs with an AUC of 1.000 on the training set, and also overperformed the SVC model on the internal test set (AUC of 0.874 vs. 0.811), however, showed poor results during external validation with an AUC of 0.663, which indicates overfitting and

TABLE 1 Distribution of demographics and tumor types in the patient cohorts.

	Study population			External test set		
	ccRCC	non-ccRCC	*P*-value	ccRCC	non-ccRCC	*P*-value
Number of cases (*n*)	161	51	–	50	23	–
Male, *n* (%)	107 (66.5%)	34 (66.7%)	0.978	33 (66.0%)	10 (43.5%)	0.069
Age, median (IQR) years	64.2 (15.6)	66.8 (16.6)	0.079	60.5 (23.2)	53 (13.0)	0.556

IQR, interquartile range; ccRCC, clear cell renal cell carcinoma.

TABLE 2 List of the selected radiomics features.

	Corticomedullary phase	Excretory phase	Unenhanced phase
Shape-based	Flatness; sphericity	Sphericity; SurfaceVolumeRatio	Sphericity; SurfaceVolumeRatio
First-order	10th percentile; energy; mean	Energy; median	Entropy; InterquartileRange;Median
Texture feature	GLCM_Correlation; GLRLM_GrayLevelNon-Uniformity; GLRLM_LongRunEmp; GLDM_DependenceNon-UniformityNorm; NGTDM_Coarseness	GLDM_DependenceEntropy	GLCM_InverseVariance; GLDM_DependenceEntropy; GLSZM_LargeAreaEmphasis; GLSZM_SizeZoneNon-UniformityNormalized

GLCM, gray-level co-occurrence matrix; GLRLM, gray-level run-length matrix; GLSZM, gray-level size zone matrix; GLDM, gray-level dependence matrix; NGTDM, neighboring gray-tone difference matrix.

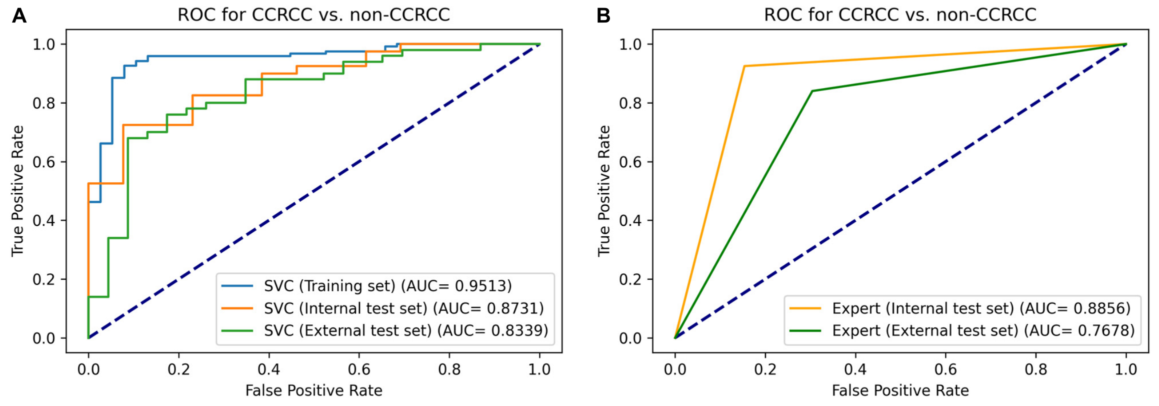


FIGURE 4
Receiver operating characteristic curves for distinguishing ccRCCs from non-ccRCCs. The performance of our support vector classifier **(A)** was similar to that of a radiologist specializing in urological imaging **(B)**. The radiomics-based machine learning model achieved an AUC of 0.951, 0.873, and 0.834 on the training set, internal test set, and external test set, respectively. Meanwhile, the expert radiologist reached an AUC of 0.886 on the internal test set, and an AUC of 0.768 on the external test set. ROC, receiver operating characteristic curve; SVC, support vector classifier; ccRCC, clear cell renal cell carcinoma.

demonstrates that the LASSO + SVC model can overperform the TuRF + RFC model in this task.

The optimized SVC (kernel: rbf, C: 75, gamma: 0.05) trained by the EX phase radiomics features showed worse performance on both the internal and external test sets with AUC of 0.719 and 0.64, respectively. As expected, the optimized SVC (kernel: linear, C:200) trained on the UN phase features showed even poorer performance with an AUC of 0.725 on the internal test set and AUC of 0.598 on the external test set. The UN and EX phase features were not able to increase the diagnostic performance of the SVC trained on CM phase features, the combined model (kernel: rbf, C: 500, gamma: 0.005) achieved an AUC of 0.862 and 0.711 on the internal and external test sets, respectively.

In **Supplementary Table 2**, we compare the results of our machine learning models with those reported in previously published studies.

Discussion

In this study, we constructed an externally validated radiomics-based machine learning prediction model for the

TABLE 3 Diagnostic performance of the machine learning models compared to that of an expert radiologist.

	AUC	Threshold	Accuracy	Sensitivity	Specificity	PPV	NPV
SVC - Training set	0.951 [0.913–0.989]	0.655*	0.925	0.926	0.921	0.974	0.796
SVC - Internal test set	0.873 [0.774–0.972]	0.655	0.811	0.900	0.539	0.857	0.636
SVC - External test set	0.834 [0.730–0.938]	0.655	0.781	0.800	0.739	0.870	0.630
RFC - Training set	1.000 [1.000–1.000]	0.500*	1.000	1.000	1.000	1.000	1.000
RFC - Internal test set	0.874 [0.755–0.993]	0.500	0.868	0.950	0.615	0.884	0.800
RFC - External test set	0.663 [0.529–0.796]	0.500	0.685	0.900	0.217	0.714	0.500
Expert – Internal test set	0.886 [0.776–0.996]	0.500	0.906	0.925	0.846	0.949	0.786
Expert – External test set	0.768 [0.659–0.877]	0.500	0.795	0.840	0.696	0.857	0.667

*The optimal threshold was determined based on the point closest to the top left corner of the graph.
AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value; RFC, random forest classifier; SVC, support vector classifier.

differentiation of ccRCC from non-ccRCC. Our SVC algorithm-based machine learning model trained by CM phase features achieved very good performance on the independent test cases from our institute with an AUC of 0.87 and its diagnostic ability also proved to be reproducible with an AUC of 0.83 during validation on external test cases from the KiTS19 dataset. In addition, we evaluated the accuracy of our SVC against that of an expert radiologist, which showed that the performance of the machine learning model is comparable (accuracy of 0.79 vs. 0.78 on the external dataset) which further supports the current literature and demonstrates the potential of CT texture analysis in this application.

The majority of the previously published studies focused on differentiating between benign and malignant kidney lesions (28–30) or identifying aggressive tumor features of ccRCCs (31–37), and only a handful of studies aimed to distinguish between the RCC subtypes (20, 38–41). It is important to highlight that previous studies used different softwares for radiomics feature extraction including both in-house developed algorithms (40, 41), and open-source tools such as the MaZda software (39) and the pyRadiomics package (38) which complicates the direct comparison of the previously published results. More importantly, most of the previous studies had a single-center study design and their models had not been validated on independent, external cases.

Yu et al. were among the first who used CT texture analysis for distinguishing between RCC subtypes (41). The authors performed radiomics analysis on 10 selected cross-sectional areas of the tumors in the NG phase and extracted 43 features. In each case, the average of the 10 values per feature was calculated. A 5-fold cross-validated, linear SVC was built to differentiate RCCs from oncocytomas for each radiomics feature separately. In distinguishing between ccRCCs vs. pRCCs, chRCCs and oncocytomas, first-order statistics “geometric mean” achieved the best predictive value with an AUC of 0.809. In the task of distinguishing between pRCCs vs. ccRCCs, chRCCs and oncocytomas, first-order statistics “median” reached the highest performance with an AUC of 0.811. While in the prediction of chRCCs vs. pRCCs, ccRCCs, and oncocytomas none of the features achieved good diagnostic performance: the highest AUC was 0.757. The SVC trained by all the 43 radiomics features achieved AUCs of 0.91, 0.92, and 0.85 in the three tasks, respectively, which may indicate that the prediction performance of the combination of the radiomics features is superior compared to the diagnostic value of individual features. Yu et al. demonstrated the ability of first-order statistics and texture features to predict RCC subtypes, however, in this single-center study all the scans were performed on the same CT scanner and the results were not validated on an independent test set (41). Yu et al. built SVC models from NG phase radiomics features, while in our study, we built an SVC prediction model from the combination of the most predictive CM features that proved to be reproducible when tested on independent external test cases (41). Our prediction model achieved comparable results compared to those reported by Yu et al. (AUC of 0.87 on the internal test set vs. AUC of 0.91 during cross-validation), which may indicate that the performance of CM features and NG features is comparable in predicting ccRCCs, although Yu et al. also included 10 oncocytomas in their dataset (41).

Chen et al. retrospectively collected triphasic CT scans from patients with RCCs (38). The final cohort in this study included 143 ccRCCs and 54 non-ccRCCs. To extract non-textural features, the authors calculated 13 different absolute and relative enhancement and attenuation ratios and values. After radiomics analysis, LASSO was used to select the most important features and to calculate texture-score with the linear combination of the selected features. Finally, three different prediction models were built, one logistic regression-based model from non-texture features, one model from texture features, and a third, combined logistic regression model. Among both the non-textural and the texture-feature-based models, the CM phase models achieved the highest performance with AUC = 0.823 and 0.887, while the performance of the combined model showed similar results in the CM and NG phases with AUCs of 0.891 and 0.900. The results of this study showed that adding non-texture features can improve the prediction performance of the texture feature-based model and that the CM phase and the NG phase radiomics features have similar diagnostic ability in differentiating between ccRCCs and non-ccRCCs. However, these models were not validated on an independent test set in this manuscript (38). The results of this study are comparable with the results of our SVC model trained on the CM phase radiomics features, especially with the results we reported on the training set (AUC = 0.951), however, we also validated the performance of our model on both independent internal (AUC = 0.873) and external test cases (AUC = 0.834).

In their recent study, Wang et al. analyzed 147 ccRCCs and 43 non-ccRCCs and built a RFC, an SVC, and a logistic regression algorithm-based machine learning model from four selected radiomics features (20). The authors reported very good results on the internal test dataset for each machine learning algorithm. Their RFC achieved the highest diagnostic performance with an AUC of 0.909 followed by the logistic regression classifier with an AUC of 0.906, while the SVC showed slightly worse results with an AUC of 0.841 (20). These results on the independent internal test set (AUC = 0.841–0.909) are very similar to the results of our SVC (AUC of 0.88) on the independent internal test set. However, all patients were scanned with the same CT scanner in this single-center study, and the models were not validated on external test cases. The diagnostic performance of an expert radiologist was also reported in this study, and the authors successfully demonstrated that radiomics-based machine learning models can overperform the accuracy of an expert radiologist. Although, the radiologist's performance reported in this manuscript was slightly inferior to that of our study (AUC of 0.69 vs. 0.76–0.88, sensitivity of 0.85 vs. 0.84–0.93, and specificity of 0.58 vs. 0.70–0.85).

In a two-center study by Li et al., external validation of the machine learning models was also completed (40). The authors performed 3D texture analysis on both the UN, CM, and NG phase CT scans of 170 patients. After the extraction of 3 × 52 texture features from the tumors, either the Boruta algorithm or the minimum redundancy maximum relevance ensemble (mRMRe) was used to select the most relevant features. Two RFCs were trained, one with the 8 CM phase features selected by the Boruta algorithm, and one by the combination of 7 nephrographic and one CM phase features selected by the mRMRe algorithm. The machine learning models were tested on 85 independent external test cases from another hospital. The Boruta-based model achieved an AUC of 0.949 and an accuracy of 92.9%, which significantly overperformed the mRMRe-based model which reached an AUC of 0.851 and an accuracy of 81.2%. Their results suggest that there is a huge difference between the performance of feature selection algorithms, as the two sets of selected features were markedly different. These results also indicate that the CM features have higher diagnostic ability compared to NG phase features in the differentiation of ccRCCs from non-ccRCCs (40). In our study, we extracted not just second-order texture features, but also first-order statistical parameters and shape-based features from the tumor volumes in the CM phase. The LASSO algorithm selected two shape-based, three first-order, and five texture features as the most important ones, which may indicate the importance of first-order statistics and shape-based features in addition to texture features. Although our results on the external test set are slightly worse than those reported by Li et al. (AUC of 0.834 vs. 0.949), it could be at least partly due to the fact that our independent test sets contained a significant number of atypical cases which is supported by that the accuracy of our SVC model proved to be comparable with that of an expert radiologist (accuracy of 0.78 vs. 0.79) (40).

Kocak et al. were among the first, who validated their machine learning models' performance on publicly available datasets (39). In their retrospective study, Kocak et al. collected 48 ccRCCs, 13 pRCCs, and 7 chRCCs and performed CT texture analysis on UN and CM phase CT scans to differentiate between RCC subtypes. For external validation, the authors selected 13 ccRCCs, 7 pRCCs, and 6 chRCCs from three publicly available datasets including The Cancer Genome Atlas-Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) (44, 45), the TCGA-Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP) (44, 46), and the TCGA-Kidney Chromophobe (TCGA-KICH) (44, 47). After manual segmentation, the authors performed texture analysis on the largest cross-sectional areas of the tumors by extracting 275 radiomics features from both the UN and the CM phases. After feature selection, the authors constructed artificial neural network (ANN)-based and SVC-based prediction models, that were evaluated based on ROC curve analysis and Matthews correlation coefficient (MCC) values. In the differentiation between ccRCC and non-ccRCCs, the ANN algorithm-based model combined with adaptive boosting trained on CM phase radiomics features, achieved an AUC = 0.870, accuracy of 86.7%, and MCC = 0.686 during internal validation, and AUC = 0.822, accuracy of 84.6%, and MCC = 0.728 on the external test set. Meanwhile, the SVC combined with adaptive boosting achieved an AUC = 0.852, accuracy of 89.7%, and MCC = 0.745 during internal validation, and AUC = 0.793, accuracy of 65.3%, and MCC = 0.426 on the external test set (39). Our results can be compared with those reported by Kocak et al., our SVC achieved slightly better performance both during internal validation (AUC 0.873 vs. 0.852 and external validation (AUC of 0.834 vs. 0.793), however, it is important to note that for external validation, we used 73 tumors from the KiTS19 dataset, while Kocak et al. validated their results on 26 selected cases from the TCGA datasets (39).

We confirmed the results of previous studies that the CM phase radiomics features are superior compared to the EX phase ones (29, 38). Interestingly, contrary to the results of Raman et al. (29), we were unable to prove that the addition of UN and/or EX phase radiomics features increase the predictive performance of the model, however, we did not analyze NG phase scans as these were not available in the KiTS19 dataset.

The limitation of our study is the relatively low number of patients, and that only the three most common RCC subtypes were studied, however, the other subtypes are rare. The distribution of the RCC subtypes was unbalanced, reflecting the unequal distribution in the global population. To handle unbalanced datasets, instead of using synthetic sampling methods, we used class-weight optimization during "training" and then we tested the model on independent cases from different institutions. Since the inclusion criteria in this study were not strict to avoid selection bias, the internal and external test datasets were also slightly unbalanced reflecting real-world conditions. We decided not to use synthetic sampling techniques to balance the groups of test sets, as we wanted our results on test sets to illustrate how the model would work in the daily clinical practice. The distribution of patients by sex in the training and test datasets were also slightly imbalanced, but it is well known that in the general population men are more likely to be affected by kidney cancer than women and that kidney cancer is about twice as common in men as in women. Accordingly, the number of male patients in our own study was slightly higher than the number of female patients in both our own dataset and the external test set. However, the distribution did not reach a significant level, i.e., the imbalance was similar between the ccRCC and non-ccRCC groups. Finally, nephrographic phase CT scans were not included in our study, as those were not available in the KiTS19 dataset.

In conclusion, we successfully built a support vector classifier-based machine learning model from CM phase radiomics features that was able to differentiate between ccRCCs and non-ccRCCs with good accuracy. The performance of our model was validated on both cases from our own institute during internal validation (AUC = 0.87), and cases from the KiTS19 dataset during external validation (AUC = 0.83), which proved our machine learning model's reliability and generalizability. We also compared the accuracy of the SVC with that of an expert radiologist (accuracy of 0.79 vs. 0.78 on the external dataset), which showed non-inferior results. Therefore, we conclude that radiomics analysis combined with machine learning could facilitate the non-invasive diagnosis of RCCs in clinical practice in an objective and automated way.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Semmelweis University Regional and Institutional Committee of Science and Research Ethics. The ethics committee waived the requirement of written informed consent for participation.

Author contributions

BB: conceptualization, investigation, data curation, formal analysis, methodology, and writing – original draft. RS, AR, BK, ZZ, and LP: data curation and writing – original draft. BF: methodology, data curation, and writing – review and editing. AS and ES: methodology and writing – review and editing. PM-H: resources and writing – review and editing. PK: conceptualization, investigation, supervision, and writing – review and editing. All authors read and approved the final version of the manuscript.

Funding

BB was supported by the ÚNKP-21-3-I New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.974485/full#supplementary-material

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J Clin.* (2021) 71:209–49. doi: 10.3322/caac.21660

2. Escudier B, Porta C, Schmidinger M, Rioux-Leclercq N, Bex A, Khoo V, et al. Renal cell carcinoma: esmo clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol.* (2019) 30:706–20. doi: 10.1093/annonc/mdz056

3. Moch H, Cubilla AL, Humphrey PA, Reuter VE, Ulbright TM. The 2016 WHO classification of tumours of the urinary system and male genital organs-part a: renal, penile, and testicular tumours. *Eur Urol.* (2016) 70:93–105. doi: 10.1016/j.eururo.2016.02.029

4. Wei H, Miao J, Cui J, Zheng W, Chen X, Zhang Q, et al. The prognosis and clinicopathological features of different distant metastases patterns in renal cell carcinoma: analysis based on the seer database. *Sci Rep.* (2021) 11:17822. doi: 10.1038/s41598-021-97365-6

5. Elkassem AA, Allen BC, Sharbidre KG, Rais-Bahrami S, Smith AD. Update on the role of imaging in clinical staging and restaging of renal cell carcinoma based on the ajcc 8th edition, from the ajr special series on cancer staging. *AJR Am J Roentgenol.* (2021) 217:541–55. doi: 10.2214/ajr.21.25493

6. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, et al. Renal cell carcinoma. *Nat Rev Dis Primers.* (2017) 3:17009. doi: 10.1038/nrdp.2017.9

7. Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep.* (2018) 23:313–26.e5. doi: 10.1016/j.celrep.2018.03.075

8. Capitanio U, Cloutier V, Zini L, Isbarn H, Jeldres C, Shariat SF, et al. A critical assessment of the prognostic value of clear cell, papillary and chromophobe histological subtypes in renal cell carcinoma: a population-based study. *BJU Int.* (2009) 103:1496–500. doi: 10.1111/j.1464-410X.2008.08259.x

9. Leibovich BC, Lohse CM, Crispen PL, Boorjian SA, Thompson RH, Blute ML, et al. Histological subtype is an independent predictor of outcome for patients with renal cell carcinoma. *J Urol.* (2010) 183:1309–15. doi: 10.1016/j.juro.2009.12.035

10. Cheville JC, Lohse CM, Zincke H, Weaver AL, Blute ML. Comparisons of outcome and prognostic features among histologic subtypes of renal cell carcinoma. *Am J Surg Pathol.* (2003) 27:612–24. doi: 10.1097/00000478-200305000-00005

11. Motzer RJ, Jonasch E, Agarwal N, Alva A, Baine M, Beckermann K, et al. Kidney cancer, version 3.2022, nccn clinical practice guidelines in oncology. *J Nat Compr Cancer Netw.* (2022) 20:71–90. doi: 10.6004/jnccn.2022.0001

12. Klapper JA, Downey SG, Smith FO, Yang JC, Hughes MS, Kammula US, et al. High-dose interleukin-2 for the treatment of metastatic renal cell carcinoma : a retrospective analysis of response and survival in patients treated in the surgery branch at the national cancer institute between 1986 and 2006. *Cancer.* (2008) 113:293–301. doi: 10.1002/cncr.23552

13. Hasanov E, Gao J, Tannir NM. The immunotherapy revolution in kidney cancer treatment: scientific rationale and first-generation results. *Cancer J.* (2020) 26:419–31. doi: 10.1097/ppo.0000000000000471

14. Lavacchi D, Pellegrini E, Palmieri VE, Doni L, Mela MM, Di Maida F, et al. Immune checkpoint inhibitors in the treatment of renal cancer: current state and future perspective. *Int J Mol Sci.* (2020) 21:4691. doi: 10.3390/ijms21134691

15. Choueiri TK, Plantade A, Elson P, Negrier S, Ravaud A, Oudard S, et al. Efficacy of sunitinib and sorafenib in metastatic papillary and chromophobe renal cell carcinoma. *J Clin Oncol.* (2008) 26:127–31. doi: 10.1200/jco.2007.13.3223

16. Shinagare AB, Krajewski KM, Braschi-Amirfarzan M, Ramaiya NH. Advanced renal cell carcinoma: role of the radiologist in the era of precision medicine. *Radiology.* (2017) 284:333–51. doi: 10.1148/radiol.2017160343

17. Young JR, Margolis D, Sauk S, Pantuck AJ, Sayre J, Raman SS. Clear cell renal cell carcinoma: discrimination from other renal cell carcinoma subtypes and oncocytoma at multiphasic multidetector Ct. *Radiology.* (2013) 267:444–53. doi: 10.1148/radiol.13112617

18. Bata P, Gyebnar J, Tarnoki DL, Tarnoki AD, Kekesi D, Szendroi A, et al. Clear cell renal cell carcinoma and papillary renal cell carcinoma: differentiation of distinct histological types with multiphase Ct. *Diagn Interv Radiol.* (2013) 19:387–92. doi: 10.5152/dir.2013.13068

19. Kim JK, Kim TK, Ahn HJ, Kim CS, Kim KR, Cho KS. Differentiation of subtypes of renal cell carcinoma on helical ct scans. *AJR Am J Roentgenol.* (2002) 178:1499–506. doi: 10.2214/ajr.178.6.1781499

20. Wang P, Pei X, Yin XP, Ren JL, Wang Y, Ma LY, et al. Radiomics models based on enhanced computed tomography to distinguish clear cell from non-clear cell renal cell carcinomas. *Sci Rep.* (2021) 11:13729. doi: 10.1038/s41598-021-93069-z

21. Sun XY, Feng QX, Xu X, Zhang J, Zhu FP, Yang YH, et al. Radiologic-radiomic machine learning models for differentiation of benign and malignant solid renal masses: comparison with expert-level radiologists. *AJR Am J Roentgenol.* (2020) 214:W44–54. doi: 10.2214/ajr.19.21617

22. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* (2012) 48:441–6. doi: 10.1016/j.ejca.2011.11.036

23. Shur JD, Doran SJ, Kumar S, Ap Dafydd D, Downey K, O'Connor JPB, et al. Radiomics in oncology: a practical guide. *Radiographics.* (2021) 41:1717–32. doi: 10.1148/rg.2021210037

24. Kocak B, Durmaz ES, Erdim C, Ates E, Kaya OK, Kilickesmez O. Radiomics of renal masses: systematic review of reproducibility and validation strategies. *AJR Am J Roentgenol.* (2020) 214:129–36. doi: 10.2214/ajr.19.21709

25. Zwanenburg A, Vallières M, Abdalah MA, Aerts H, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology.* (2020) 295:328–38. doi: 10.1148/radiol.2020191145

26. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging.* (2020) 11:91. doi: 10.1186/s13244-020-00887-2

27. Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui LA. Review of original articles published in the emerging field of radiomics. *Eur J Radiol.* (2020) 127:108991. doi: 10.1016/j.ejrad.2020.108991

28. Feng Z, Rong P, Cao P, Zhou Q, Zhu W, Yan Z, et al. Machine learning-based quantitative texture analysis of ct images of small renal masses: differentiation of angiomyolipoma without visible fat from renal cell carcinoma. *Eur Radiol.* (2018) 28:1625–33. doi: 10.1007/s00330-017-5118-z

29. Raman SP, Chen Y, Schroeder JL, Huang P, Fishman EK. Ct texture analysis of renal masses: pilot study using random forest classification for prediction of pathology. *Acad Radiol.* (2014) 21:1587–96. doi: 10.1016/j.acra.2014.07.023

30. Dana J, Lefebvre TL, Savadjiev P, Bodard S, Gauvin S, Bhatnagar SR, et al. Malignancy risk stratification of cystic renal lesions based on a contrast-enhanced ct-based machine learning model and a clinical decision algorithm. *Eur Radiol.* (2022) 32:4116–27. doi: 10.1007/s00330-021-08449-w

31. Schieda N, Thornhill RE, Al-Subhi M, McInnes MD, Shabana WM, van der Pol CB, et al. Diagnosis of sarcomatoid renal cell carcinoma with Ct: evaluation by qualitative imaging features and texture analysis. *AJR Am J Roentgenol.* (2015) 204:1013–23. doi: 10.2214/ajr.14.13279

32. Gurbani S, Morgan D, Jog V, Dreyfuss L, Shen M, Das A, et al. Evaluation of radiomics and machine learning in identification of aggressive tumor features in renal cell carcinoma (RCC). *Abdom Radiol.* (2021) 46:4278–88. doi: 10.1007/s00261-021-03083-y

33. Meng X, Shu J, Xia Y, Yang RA. Ct-based radiomics approach for the differential diagnosis of sarcomatoid and clear cell renal cell carcinoma. *Biomed Res Int.* (2020) 2020:7103647. doi: 10.1155/2020/7103647

34. Yi X, Xiao Q, Zeng F, Yin H, Li Z, Qian C, et al. Computed tomography radiomics for predicting pathological grade of renal cell carcinoma. *Front Oncol.* (2020) 10:570396. doi: 10.3389/fonc.2020.570396

35. Xv Y, Lv F, Guo H, Zhou X, Tan H, Xiao M, et al. Machine learning-based ct radiomics approach for predicting who/isup nuclear grade of clear cell renal cell carcinoma: an exploratory and comparative study. *Insights Imaging.* (2021) 12:170. doi: 10.1186/s13244-021-01107-1

36. Yang L, Gao L, Arefan D, Tan Y, Dan H, Zhang JA. Ct-based radiomics model for predicting renal capsule invasion in renal cell carcinoma. *BMC Med Imaging.* (2022) 22:15. doi: 10.1186/s12880-022-00741-5

37. Bektas CT, Kocak B, Yardimci AH, Turkcanoglu MH, Yucetas U, Koca SB, et al. Clear cell renal cell carcinoma: machine learning-based quantitative computed tomography texture analysis for prediction of fuhrman nuclear grade. *Eur Radiol.* (2019) 29:1153–63. doi: 10.1007/s00330-018-5698-2

38. Chen M, Yin F, Yu Y, Zhang H, Wen G. Ct-based multi-phase radiomic models for differentiating clear cell renal cell carcinoma. *Cancer Imaging.* (2021) 21:42. doi: 10.1186/s40644-021-00412-8

39. Kocak B, Yardimci AH, Bektas CT, Turkcanoglu MH, Erdim C, Yucetas U, et al. Textural differences between renal cell carcinoma subtypes: machine learning-based quantitative computed tomography texture analysis with independent external validation. *Eur J Radiol.* (2018) 107:149–57. doi: 10.1016/j.ejrad.2018.08.014

40. Li ZC, Zhai G, Zhang J, Wang Z, Liu G, Wu GY, et al. Differentiation of clear cell and non-clear cell renal cell carcinomas by all-relevant radiomics features from multiphase Ct: A Vhl mutation perspective. *Eur Radiol.* (2019) 29:3996–4007. doi: 10.1007/s00330-018-5872-6

41. Yu H, Scalera J, Khalid M, Touret AS, Bloch N, Li B, et al. Texture analysis as a radiomic marker for differentiating renal tumors. *Abdom Radiol.* (2017) 42:2470–8. doi: 10.1007/s00261-017-1144-1

42. Frank V, Shariati S, Budai BK, Fejér B, Tóth A, Orbán V, et al. Ct texture analysis of abdominal lesions – part ii: tumors of the kidney and pancreas. *Imaging.* (2021) 13:25–36. doi: 10.1556/1647.2021.00020

43. Suarez-Ibarrola R, Basulto-Martinez M, Heinze A, Gratzke C, Miernik A. Radiomics applications in renal tumor assessment: a comprehensive review of the literature. *Cancers.* (2020) 12:1387. doi: 10.3390/cancers12061387

44. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (Tcia): maintaining and operating a public information repository. *J Digit Imaging.* (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7

45. Akin, O, Elnajjar P, Heller M, Jarosz R, Erickson B, Kirk S, et al. *The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma Collection (TCGA-KIRC) (Version 3)*. The Cancer Imaging Archive. (2016). doi: 10.7937/K9/TCIA.2016.V6PBVTDR

46. Linehan M, Gautam R, Kirk S, Lee Y, Roche C, Bonaccio E, et al. *The Cancer Genome Atlas Cervical Kidney Renal Papillary Cell Carcinoma Collection (TCGA-KIRP) (Version 4)*. The Cancer Imaging Archive. (2016). doi: 10.7937/K9/TCIA.2016.ACWOGBEF

47. Linehan M, Gautam R, Sadow C, Levine SJ. *The Cancer Genome Atlas Kidney Chromophobe Collection (TCGA-KICH) (Version 3)*. The Cancer Imaging Archive. (2016). doi: 10.7937/K9/TCIA.2016.YU3RBCZN

48. Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, Li F, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: results of the kits19 challenge. *Med Image Anal.* (2021) 67:101821. doi: 10.1016/j.media.2020.101821

49. Heller N, Sathianathen N, Kalapara A, Walczak E, Moore K, Kaluzniak H, et al. *Data from C4KC-KiTS Dataset*. The Cancer Imaging Archive. (2019). doi: 10.7937/TCIA.2019.IX49E8NX

50. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging.* (2012) 30:1323–41. doi: 10.1016/j.mri.2012.05.001

51. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* (2017) 77:e104–7. doi: 10.1158/0008-5472.Can-17-0339

52. Cattell R, Chen S, Huang C. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. *Vis Comput Ind Biomed Art.* (2019) 2:19. doi: 10.1186/s42492-019-0025-6

53. Dercle L, Lu L, Schwartz LH, Qian M, Tejpar S, Eggleton P, et al. Radiomics response signature for identification of metastatic colorectal cancer sensitive to therapies targeting Egfr pathway. *J Nat Cancer Inst.* (2020) 112:902–12. doi: 10.1093/jnci/djaa017

TYPE Original Research
PUBLISHED 07 December 2022
DOI 10.3389/fmed.2022.1001383

A mucosal recovery software tool for endoscopic submucosal dissection in early gastric cancer

Yinuo Zhao[1†], Huogen Wang[2,3†], Yanyan Fan[4], Chaohui Jin[2], Qinwei Xu[5], Jiyong Jing[6], Tianqiao Zhang[7], Xuedong Zhang[4*] and Wanyuan Chen[8*]

[1]Department of Pathology, Rizhao People's Hospital, Rizhao, China, [2]Hithink RoyalFlush Information Network Co., Ltd., Hangzhou, China, [3]College of Computer Science and Technology, Zhejiang University, Hangzhou, China, [4]Department of Pathology, Liaocheng People's Hospital, Liaocheng, China, [5]Endoscopy Center, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China, [6]Department of Medical Education & Simulation Center, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou, Zhejiang, China, [7]Hangzhou No.14 High School, Hangzhou, Zhejiang, China, [8]Cancer Center, Department of Pathology, Zhejiang Provincial People's Hospital, Affiliated People's Hospital, Hangzhou Medical College, Hangzhou, Zhejiang, China

Background: Due to the limited diagnostic ability, the low detection rate of early gastric cancer (EGC) is a serious health threat. The establishment of the mapping between endoscopic images and pathological images can rapidly improve the diagnostic ability to detect EGC. To expedite the learning process of EGC diagnosis, a mucosal recovery map for the mapping between ESD mucosa specimen and pathological images should be performed in collaboration with endoscopists and pathologists, which is a time-consuming and laborious work.

Methods: 20 patients at the Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College from March 2020 to July 2020 were enrolled in this study. We proposed the improved U-Net to obtain WSI-level segmentation results, and the WSI-level results can be mapped to the macroscopic image of the specimen. For the convenient use, a software pipeline named as "Pathology Helper" for integration the workflow of the construction of mucosal recovery maps was developed.

Results: The MIoU and Dice of our model can achieve 0.955 ± 0.0936 and 0.961 ± 0.0874 for WSI-level segmentation, respectively. With the help of "Pathology Helper", we can construct the high-quality mucosal recovery maps to reduce the workload of endoscopists and pathologists.

Conclusion: "Pathology Helper" will accelerate the learning of endoscopists and pathologists, and rapidly improve their abilities to detect EGC. Our work can also improve the detection rate of early gastric cancer, so that more patients with gastric cancer will be treated in a timely manner.

KEYWORDS

mucosal recovery map, artificial intelligence, endoscopic submucosal dissection, early gastric cancer, Pathology Helper

Check for updates

OPEN ACCESS

EDITED BY
Jun Cheng,
Shenzhen University, China

REVIEWED BY
Tahira Nazir,
Riphah International University, Pakistan
Chen Li,
Northeastern University, China
Y. U. Weimiao,
Bioinformatics Institute (A*STAR), Singapore

*CORRESPONDENCE
Wanyuan Chen
chenwanyuan@hmc.edu.cn
Xuedong Zhang
xuedongzh@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Pathology, a section of the journal Frontiers in Medicine

RECEIVED 23 July 2022
ACCEPTED 23 November 2022
PUBLISHED 07 December 2022

CITATION
Zhao Y, Wang H, Fan Y, Jin C, Xu Q, Jing J, Zhang T, Zhang X and Chen W (2022) A mucosal recovery software tool for endoscopic submucosal dissection in early gastric cancer. *Front. Med.* 9:1001383. doi: 10.3389/fmed.2022.1001383

Core tip: In this article, we present a new approach to construct the high-quality mucosal recovery maps. We use approximately 20,000 patches to train a deep segmentation network to distinguish cancerous and intestinal metaplasia regions from normal ones. We also develop a mucosal recovery software tool to generates high-quality mucosal recovery maps. In clinical application, this technique can greatly reduce the workload of endoscopists and pathologists and rapidly improve their abilities to detect EGC.

Introduction

Gastric cancer (gastric carcinoma) is a malignant tumor originating from the gastric mucosal epithelium. In 2018, there were 1,033,701 new cases and 782,685 deaths due to gastric cancer, making it the 6th most commonly diagnosed and 3rd most fatal cancer worldwide (1). In 2015, 42% of the new cases of gastric cancer in the world occurred in China, representing a heavy disease burden of gastric cancer in the country (2). The prognosis of gastric cancer depends largely on the tumor stage. The 5-year survival rate for patients with early gastric cancer is 85–100% with endoscopic submucosal dissection (ESD) operation, while the 5-year survival rate for advanced gastric cancer is $<10\%$ (3). However, the early detection rate of gastric cancer is very low. Early detection, diagnosis, and treatment can effectively reduce the mortality of gastric cancer and improve the prognosis after timely treatment. In recent years, with the growth of public health awareness and the popularity of gastroscopy, there was an increase in the number of early gastric cancer detections, but not in the rate of EGC detection. The low rate of diagnosis of EGC may be due to the limited abilities in EGC diagnosis (4).

To expedite the learning process of EGC diagnosis, a mucosal recovery map for the mapping between ESD mucosa specimen and pathological images should be performed in collaboration with endoscopists and pathologists. The mucosal recovery map can show the size, boundary, depth of infiltration, and lymphatic vascular invasion of the lesion. However, it is a time-consuming and laborious work to prepare a mucosal recovery map. To finish a mucosal recovery map, the tumor area should be marked in each slide, and the tumor area should be mapped to the ESD mucosa specimen. If the lesions are large and irregular, it can take many hours to reconstruct a case (5, 6).

Fortunately, the rapid development of deep learning technology provides new ideas to construct mucosal recovery map. Recently, deep learning has been widely used in medical applications, such as computed tomography denoising (7), cell segmentation (8), COVID-19 diagnosis (9), histopathology image classification (10), and breast cancer diagnosis (11). Deep learning can automatically learn task-specific features directly from the data, which can dramatically shorten the time for data processing. In this study, a novel method is proposed for the construction of mucosal recovery maps based on deep learning which can reduce the work intensity of pathologists.

Materials and methods

This study was approved by the Ethics Committee of the Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College with the informed consent waived. The proposed method for the construction of mucosal recovery maps can be broken down into the following steps: (1) ESD postoperative specimens processing, (2) Pathological Image Segmentation, and (3) Sections mapping. The workflow was shown in **Figure 1**.

Endoscopic submucosal dissection postoperative specimens processing

A total of 20 patients at the Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College from March 2020 to July 2020 were enrolled in this study. All patients were diagnosed with EGC and treated with ESD resection. After ESD resection, all resected specimens were processed according to the guidelines of ESD (12). This procedure included stretching of the fresh specimen, fixation in formalin, sectioning of the fixed specimen, and macroscopic photography before and after sectioning. Firstly, the fresh specimen was stretched and pinned at outer borders upon a cork plate with standard pins, and a macroscopic image of the specimen was taken. Then the specimen was immediately fixed through immersion in 10% formalin for 24∼48 h and a second macroscopic image was taken. Finally, the fixed specimen was cut and sectioned into small sections at intervals of 2.0∼3.0 mm and a third macroscopic image was taken. After the pathological section made, all the sections are scanned into digital WSIs with a Motic scanner. The complete procedure is shown in **Figure 2**.

Pathological image segmentation

When a WSI is prepared properly, pathological image segmentation is the most critical step for the construction of mucosal recovery maps. In this study, a novel segmentation network is proposed for pathological image segmentation. The segmentation network can be broken down into the following steps: (1) Data annotation and preprocessing, and (2) Network construction and training.

Data annotation and preprocessing

The annotation work was carried out according to the Japanese classification of gastric carcinoma: 3rd English edition

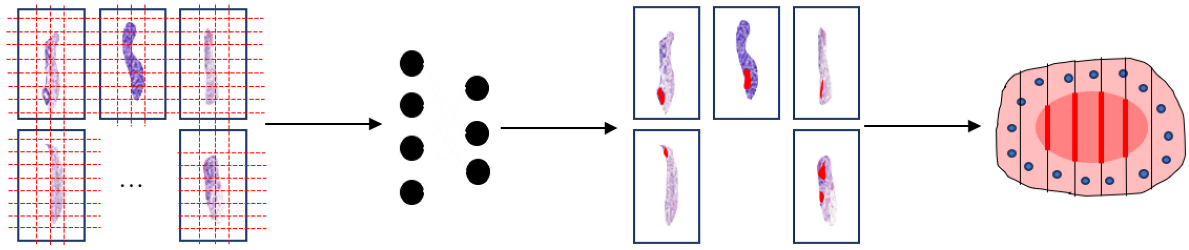


FIGURE 1
Work flow of the segmentation model. The whole slide images are split into patches **(far left)**. Then the patch-level annotation is obtained with the trained segmentation model **(near left)**. The patch-level annotation is mapped back to the WSI-level annotation based on their original location **(near right)**. Finally, these WSI-level annotations are shown on an image of the entire specimen.

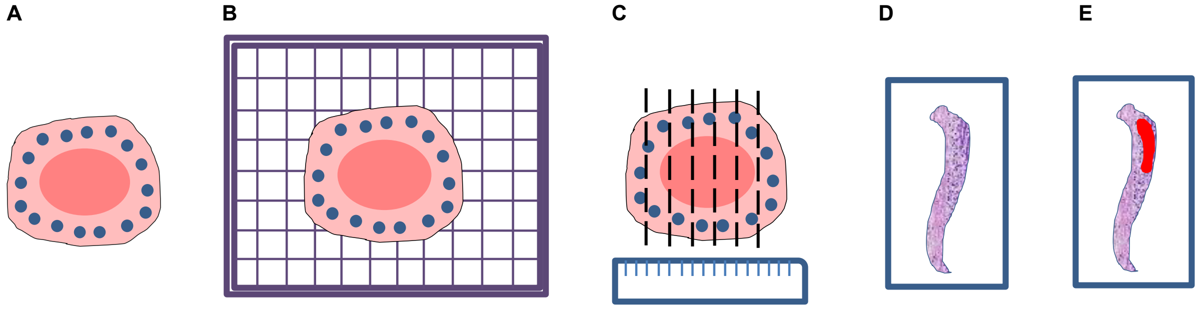


FIGURE 2
Processing flow of endoscopic submucosal dissection specimen: **(A)** Stretching and fixation; **(B)** macroscopic photography on a cork plate; **(C)** sectioning of the fixed specimen; **(D)** scanning; **(E)** annotation.

(13). All WSIs were manually labeled by a group of surgical pathologists by drawing around the cancerous regions (CR) and intestinal metaplasia regions (IR) with red and blue masks, respectively (Figure 3). These masks were modified, confirmed, and verified by another group of pathologists. In the corresponding mask generated, the cancerous regions, intestinal metaplasia regions, and normal mucosa regions (NR) were shown as red, blue, and green, respectively. Then, all annotated WSIs was divided into a training set and a testing set. The training set contained 112 WSIs from 11 patients, and the testing set contained 48 WSIs from 9 patients. Due to the limitations of GPU memory, all WSIs and corresponding masks were split into 512 × 512 pixel patches at 10x magnification (see as Figure 4), and all blank images were removed from the training set. There were 21,799 patches left in the training set and 9,784 patches left in the testing set. The overview of the dataset was shown in Table 1. Random oversampling was adopted for overcome the unbalance between the lesion area and normal area.

Network construction and training

Our segmentation network incorporates an SE block (14) into U-Net (15) as shown in Figure 5. U-Net is one of the famous Fully Convolutional Networks (FCNs) (16) used in biomedical image segmentation. The image-label pairs in the training set are fed into the segmentation network for training.

The ResNet-34 framework is employed as the backbone of U-Net. The architecture of our segmentation network is shown in Figure 6. The special residual blocks (Figure 6B) in ResNet are made up of several convolutional layers with the same number of output channels. Each convolutional layer is followed by a batch normalization layers and a rectified linear unit (ReLU). Then, a shortcut connection and element-wise addition is performed between input and output layers of the block, which make the network easier to optimize (17). Further, a Squeeze-and-Excitation (SE) block is incorporated into U-Net to boost the segmentation performance with increased generalization ability by exploiting adaptive channel-wise feature recalibration (14).

The loss function of our segmentation network is the combination of Jaccard distance loss (18) and cross-entropy loss (8). The loss function can be formulated as follows:

$$l = l_{\text{Jaccard distance}} + l_{\text{cross entropy}} \quad (1)$$

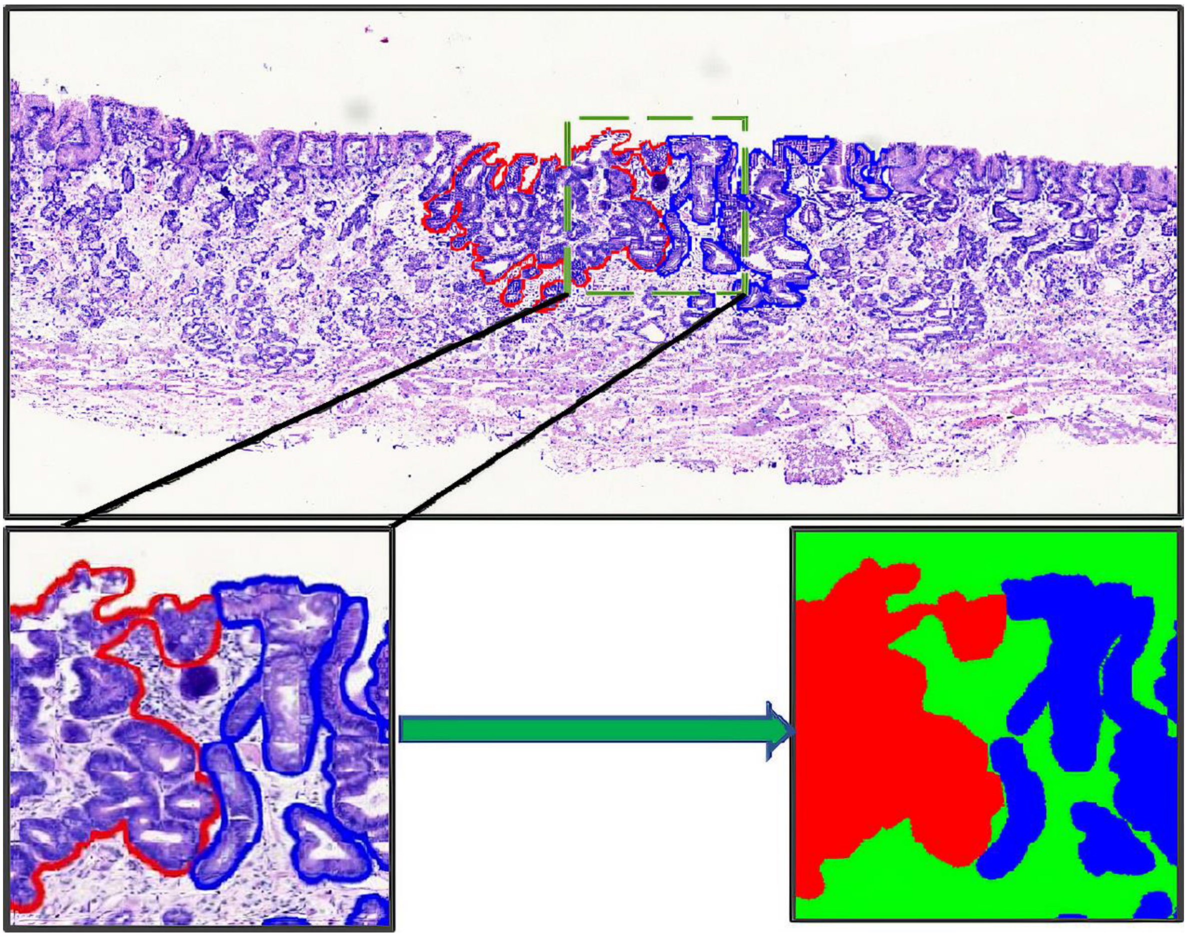


FIGURE 3
Whole slide image annotation. **(Above and lower left)** Endoscopic submucosal dissection specimen with cancerous regions outlined in red, and intestinal metaplasia regions in blue. **(Lower right)** Corresponding masks for using in deep learning.

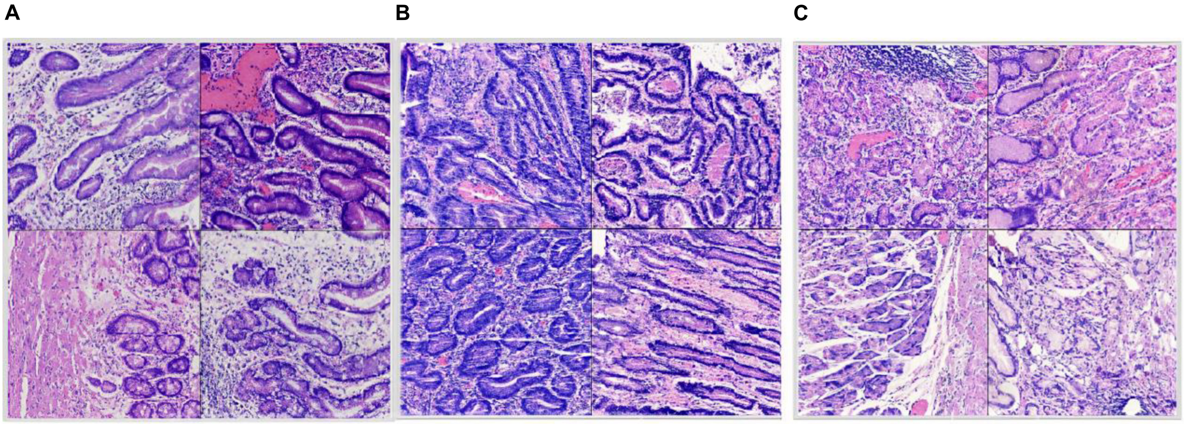


FIGURE 4
Split dataset: **(A)** Intestinal metaplasia region patches; **(B)** cancerous region patches; **(C)** normal region patches.

TABLE 1 The overview of the dataset for network training.

	Training set	Testing set
Cases	11	9
WSIs	112	48
Patches	21,799	9,784

$$l_{\text{Jaccard distance}} = 1 - \frac{\text{Intersection}}{\text{Union}} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{p_i \times y_i}{p_i + y_i - p_i \times y_i} \quad (2)$$

$$l_{\text{cross entropy}} = -\sum_{i}^{N} p_i \log\left(y_i\right) \quad (3)$$

In the formula (2) and (3), y_i and p_{i} are the i-th pixels of the labels and predictions, respectively, and N is the number of image pixels.

The network was trained using Adam optimizer with learning rate 3×10^{-4} for 50 epochs with a batch size of 4. The input size of our network was 512 × 512 pixels. To prevent overfitting, data augmentation was operated on all image-label pairs including rotation (rotation angle range 0∼359°), cropping (vertical and horizontal shift range in 0∼50 pixels), and vertical and horizontal flips. The network is implemented with Keras (TensorFlow backend) and trained on single GTX 1080Ti GPU.

Sections mapping

For each WSI in one case, the WSIs are split into 512 × 512 pixel patches. Then the patch-level annotation with the trained segmentation model, and map the patch-level annotation back to the WSI-level annotation based on their original location. Finally, the WSI-level annotation should be mapped back to the ESD specimens (see in **Figure 7**). Considering that pathological sections may be deformed and atrophied during processing, it was difficult to construct a perfect mucosal recovery map by simple stitching. In this study, GloFlow (19) was employed for slide stitching. GloFlow was a two-stage method for the fusion of pathological image using optical flow-based image registration with global alignment using a computationally tractable graph-pruning approach.

Results

Specimen preparation standard

To successfully complete the construction of mucosal recovery maps, the specimen needs to meet the following criteria: (1) The edge of the fixed ESD specimen should not be curly, (2) the surface of specimen should be dry and free of mucus, and (3) the photographs of specimen should be without reflections, and micro-structures should be clearly visible.

Data and result analysis

We have compared the performance of our model with U-Net on the testing set with 9,784 of patch images. The performance was quantified by using mean intersection over union (MIoU) and Dice Coefficient. MIoU is a standard metric for segmentation purposes, which computes the ratio between the intersection and the union of prediction and ground truth.

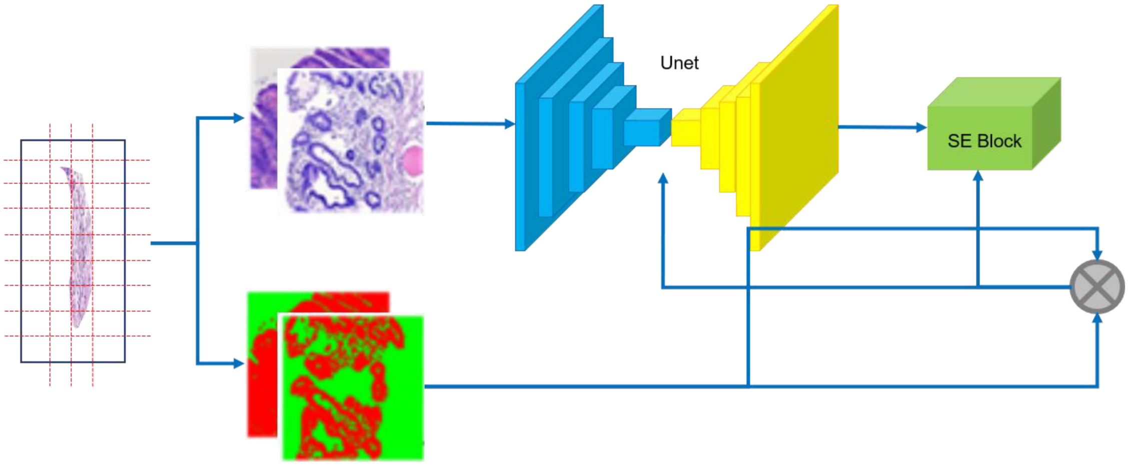


FIGURE 5
Training flow of CNN. The specimen **(left)** is photographed; the resulting whole slide images are annotated **(middle)**; image-label pairs are fed into the segmentation network **(right)**, which consists of the U-Net and the Squeeze and Excitation (SE) Block.

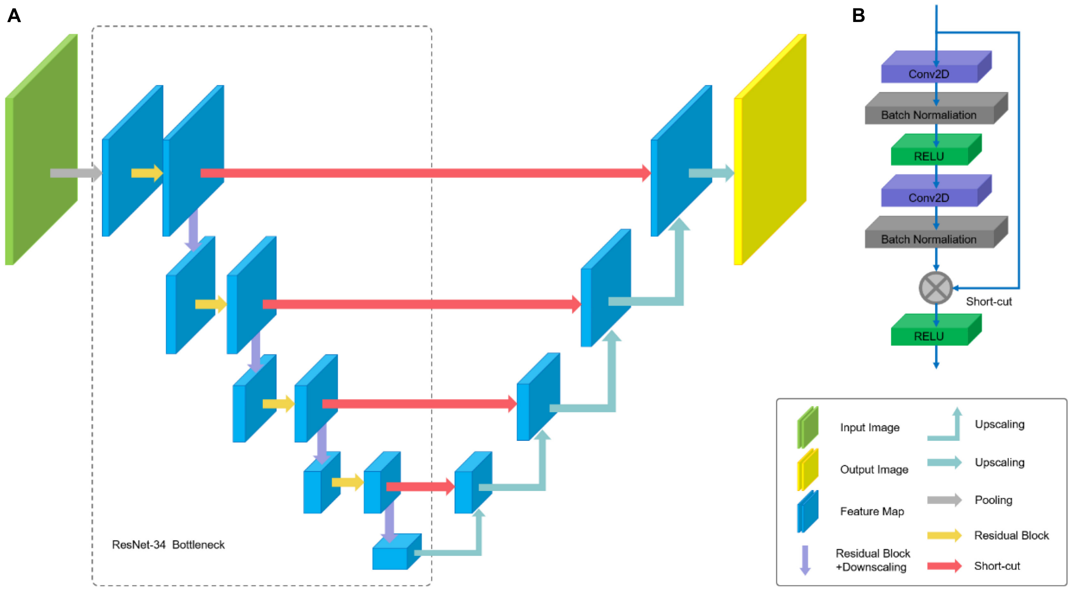


FIGURE 6
Architecture of U-Net: **(A)** U-Net-like architecture build with pre-trained ResNet-34; **(B)** residual Block.

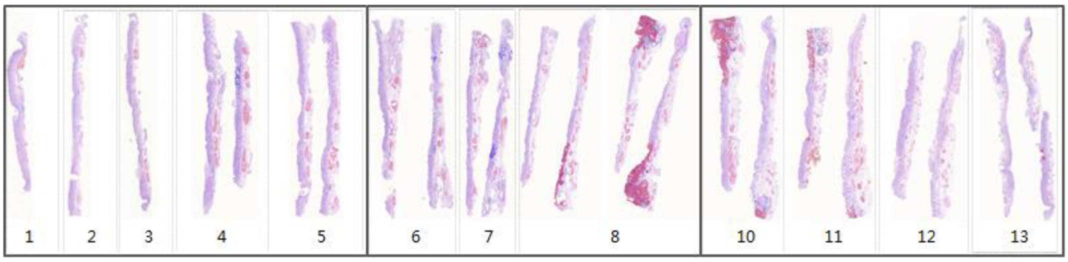


FIGURE 7
Whole slide image-level thumbnail results of 13 ESD specimens. Blue mark: Intestinal metaplasia. Red mark: Cancerous.

The Dice Coefficient is two times the Area of Overlap divided by the total number of pixels in both prediction and ground truth.

As shown in Table 2, the segmentation performance of our model and U-Net was listed. From the result in Table 2, our model can achieve better performance than U-Net. This is mostly due to the fact that a Squeeze-and-Excitation (SE) block can boost the segmentation performance with increased generalization ability by exploiting adaptive channel-wise feature recalibration. Some segmentation results and corresponding ground truth are shown in Figure 8.

TABLE 2 The segmentation performance of our model and U-Net.

Methods	MIoU	Dice
Our model	0.955 ± 0.0936	0.961 ± 0.0874
U-Net	0.921 ± 0.1761	0.932 ± 0.1585

The development of "Pathology Helper"

For the convenient use, a software pipeline named as "Pathology Helper" for integration the workflow of the construction of mucosal recovery maps was developed. The interface of "Pathology Helper" is shown as Figure 9.

Conclusion

In the clinical diagnosis and treatment of early gastric cancer, the detection rate (i.e., the number of early gastric cancers as a percentage of the total number of diagnosed gastric cancers) is an important index measuring the level of an endoscopic center. The detection rate varies from place to

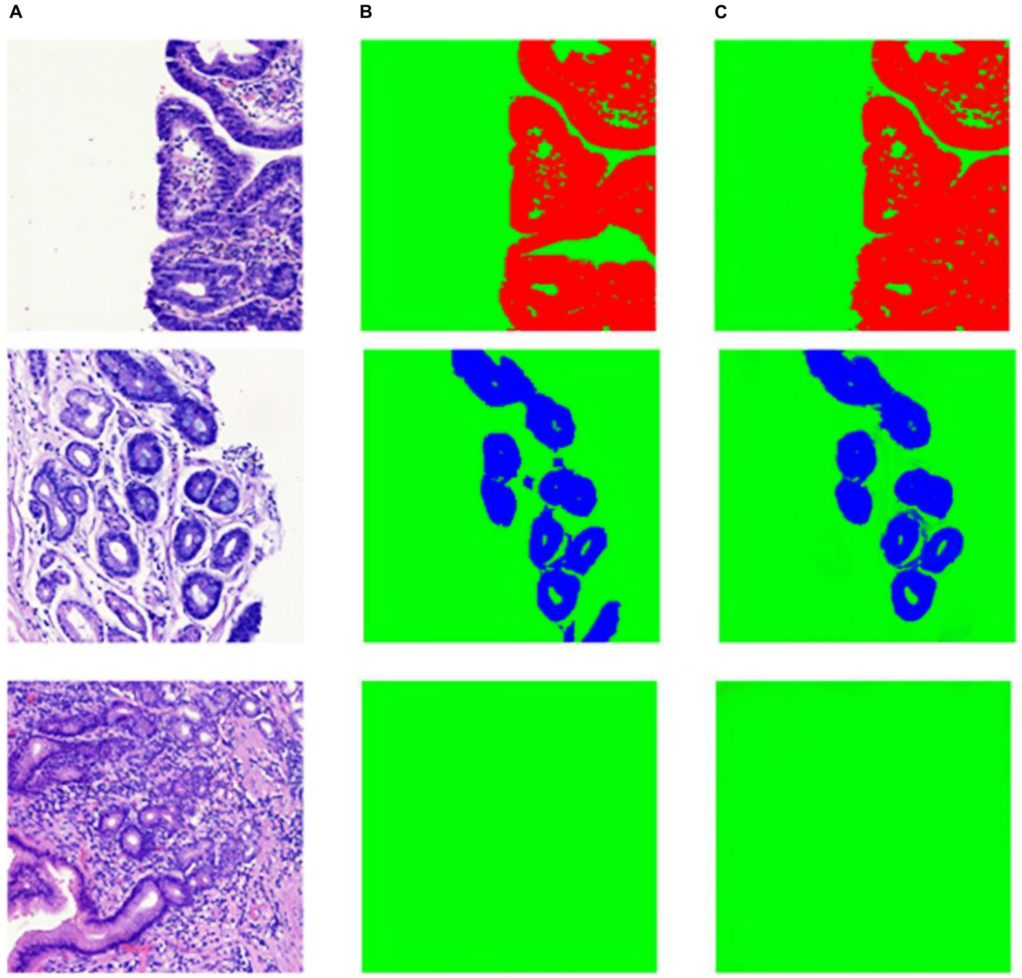


FIGURE 8
Example of segmentation result in patch-level. **(A)** Whole slide image patches; **(B)** annotation masks; **(C)** deep learning model prediction results.

place in China: It can reach 40% in developed cities along the southeast coast, but is less than 10% in remote areas. The overall detection rate in China is about 15%. Therefore, it is very important to improve the detection rate of early gastric cancer.

Mucosal recovery maps can help pathologists and endoscopists improve their understanding of endoscopy and pathomorphology. However, given a specimen of 6 cm × 5 cm × 0.2 cm and lesion area about 3 cm × 2 cm, it will take about 60 min for a skilled subspecialist in pathology to complete a finely made mucosal recovery map. If the histological classification of the cancer is complex, it may take even longer to complete the task. As a result, many endoscopists are unable to obtain high-quality mucosal recovery maps. In recent years, deep learning has been widely applied in the field of pathological diagnosis, thanks to the popularization of pathological section digitization. In 2017, Esteva et al. (20) used a convoluted neural network to analyze 129,450 pathological images of skin lesions and trained the model to distinguish skin squamous cell carcinoma from seborrheic keratosis, malignant melanoma, and benign nevus with the same accuracy as doctors. In 2019, Kather et al. (21) used a deep residual learning algorithm to identify microsatellite instability (MSI) directly from pathological slices. The accuracy of MSI recognition of colorectal cancer was 84%. There have also been artificial intelligence-assisted diagnostic studies on histopathology, including glioma grade (20), lymphoma classification (21), colorectal cancer polyp classification (22), and prostate cancer diagnosis (23). All these works matched or even went beyond the diagnostic level attained by human pathologists.

In this study, we design a novel segmentation network for pathological image segmentation. Starting with WSIs labeled by surgical pathologists in early gastrointestinal cancer, we trained a novel segmentation network for the automatical annotation of WSIs. Our segmentation network incorporates an SE block (14) into U-Net (15), one of the famous Fully Convolutional Networks (FCNs) (16) used in biomedical image segmentation.

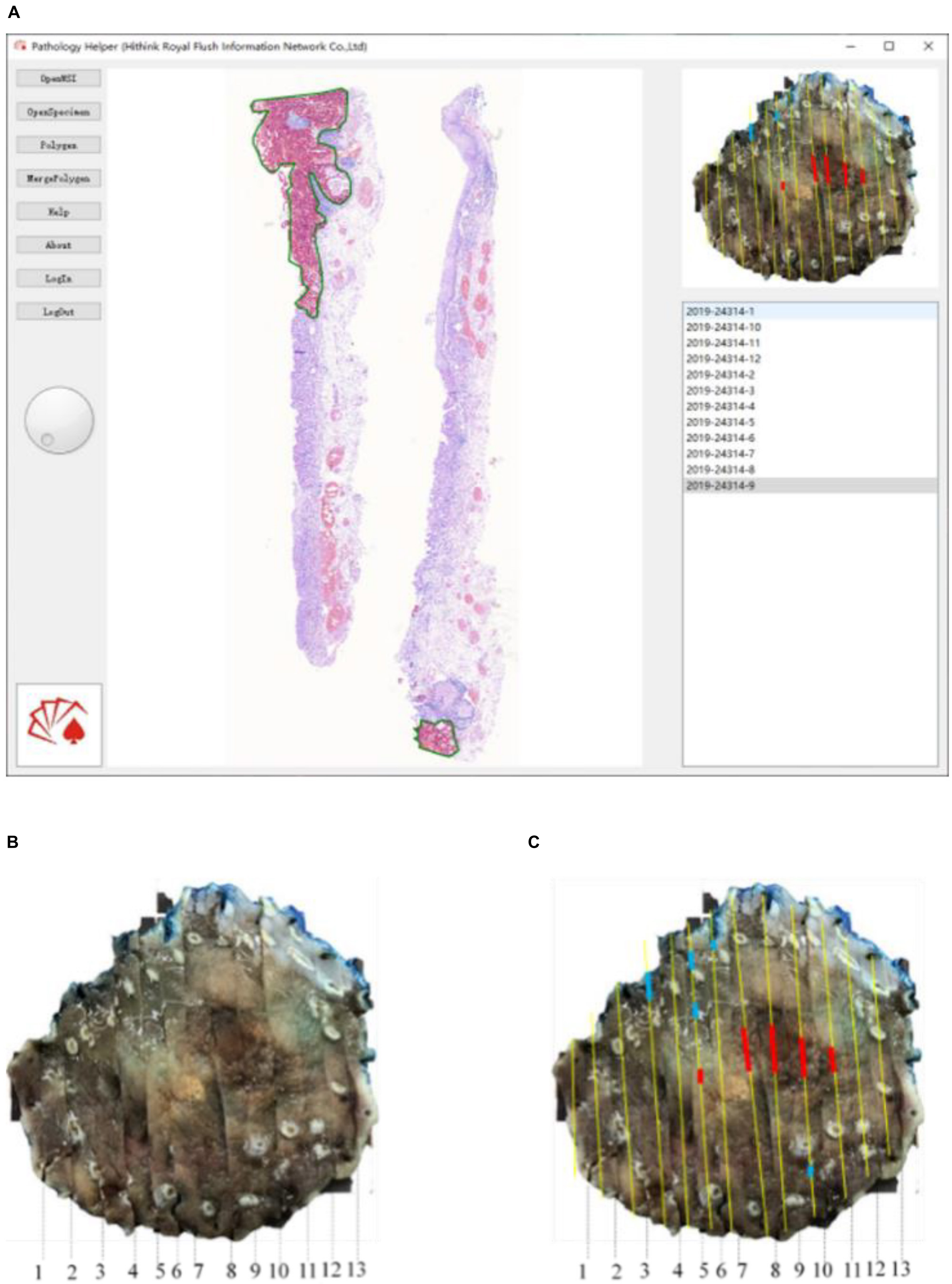


FIGURE 9

Whole slide image-level results mapping flow: **(A)** Mapping with Pathology Helper software; **(B)** specimen photo; **(C)** mapping result.

U-Net has had many successful applications, such as brain image segmentation (24), liver image segmentation (25), and cell counting, detection, and morphometry (26). However, it fails to take the differentiate between channel-wise features. In general, the SE block was proposed to be placed in InceptionNet (27) and ResNet (17) for boosting performance in classification and object detection *via* feature recalibration. Accordingly, we incorporate it into U-Net to boost the segmentation performance with increased generalization ability by exploiting adaptive channel-wise feature recalibration. The experiments show that our proposed network has better performance than U-Net alone. After pathological image segmentation, the WSI-level segmentation result is mapped back to the ESD specimen with the help of a mucosal recovery software tool "Pathology Helper".

"Pathology Helper" can help in the production of high-quality mucosal recovery maps. This will accelerate the learning of endoscopists and pathologists, and rapidly improve their abilities to detect EGC. Our work can also improve the detection rate of early gastric cancer, so that more patients with gastric cancer will be treated in a timely manner. However, this software tool still had several limitations. For example, the pathological image segmentation network was developed and trained on the dataset from a single large academic institution, which lacked multi-center or external data validation. Future research is required to determine if the same model trained can achieve high performance on larger or multi-institutional datasets.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College. The ethics committee waived the requirement of written informed consent for participation.

Author contributions

WC, XZ, and YZ: study conception and design. WC, XZ, YZ, YF, QX, and JJ: acquisition of data. HW, CJ, JJ, and TZ: analysis and interpretation of data and statistical analysis. YZ and HW: drafting of the manuscript. WC and XZ: critical revision of the manuscript for important intellectual content, and administrative, technical, or material support. WC: supervision. All authors were involved in writing the manuscript and had final approval of the submitted and published versions.

Funding

The present study was supported by Science and Technology Department Project of Zhejiang Province (LGF20F020018 and LGF21H160031), the Zhejiang Medical Technology Plan Project (2021KY498), Shanghai Committee of Science and Technology (20XD1402900, 21XD1423100, and 21JC1405200), and the Top-level Clinical Discipline Project of Shanghai Pudong (PWYgf2021-2).

Conflict of interest

HW and CJ were employed by Hithink RoyalFlush Information Network Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 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. doi: 10.3322/caac.21492

2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. *CA Cancer J Clin.* (2016) 66:115–32. doi: 10.3322/caac.21338

3. Orditura M, Galizia G, Sforza V, Gambardella V, Fabozzi A, Laterza MM, et al. Treatment of gastric cancer. *World J Gastroenterol.*(2014) 20:1635.

4. Zhang Q, Chen ZY, Chen CD, Liu T, Tang XW, Ren YT, et al. Training in early gastric cancer diagnosis improves the detection rate of early gastric cancer: an observational study in China. *Medicine.* (2015) 94:e384. doi: 10.1097/MD.0000000000000384

5. Reggiani Bonetti L, Manta R, Manno M, Conigliaro R, Missale G, Bassotti G, et al. Optimal processing of ESD specimens to avoid pathological artifacts. *Tech Coloproctol.* (2018) 22:857–66. doi: 10.1007/s10151-018-1887-x

6. Ebigbo A, Probst A, Messmann H, Märkl B, Nam-Apostolopoulos YC. Topographic mapping of a specimen after endoscopic submucosal dissection. *Endoscopy Int Open.* (2019) 7:E521–4. doi: 10.1055/a-0846-2043

7. Gholizadeh-Ansari M, Alirezaie J, Babyn P. Deep learning for low-dose CT denoising using perceptual loss and edge detection layer. *J Digit Imaging.* (2020) 33:504–15. doi: 10.1007/s10278-019-00274-4

8. Akram SU, Kannala J, Eklund L, Heikkilä J. Cell segmentation proposal network for microscopy image anlysis. In: *Proceedings of the Medical Image Computing and Computer-Assisted Intervention.* Athens: Springer (2016). p. 21–9.

9. Rahaman MM, Li C, Yao Y, Kulwa F, Rahman MA, Wang Q, et al. Identification of COVID-19 samples from chest X-Ray images using deep learning: a comparison of transfer learning approaches. *J X-ray Sci Technol.* (2020) 28:821–39. doi: 10.3233/XST-200715

10. Haoyuan C, Chen L, Xiaoyan L, Ge W, Weiming H, Yixin L, et al. GasHis-Transformer: a multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recognit.* (2022) 130:108827. doi: 10.1016/j.patcog.2022.108827

11. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. *arXiv.* (Preprint). (2016). doi: 10.48550/arXiv.1606.05718

12. Ono H, Yao K, Fujishiro M, Oda I, Uedo N, Nimura S, et al. Guidelines for endoscopic submucosal dissection and endoscopic mucosal1 resection for early gastric cancer. *Digest Endoscopy.* (2021) 33:4–20.

13. Japanese Gastric Cancer Association [JGCA]. Japanese classification of gastric carcinoma: 3rd english edition. *Gastric Cancer.* (2011) 14:101–12. doi: 10.1007/s10120-011-0041-5

14. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell.* (2020) 42:2011–23. doi: 10.1109/TPAMI.2019.2913372

15. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. *International conference on medical image computing and computer-assisted intervention.* Cham: Springer (2015). p. 234–41.

16. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell.* (2017) 39:640–51. doi: 10.1109/TPAMI.2016.2572683

17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.* Las Vegas, NV: IEEE (2016). p. 770–8.

18. Kosub S. A note on the triangle inequality for the jaccard distance. *Pattern Recognit Lett.* (2019) 120:36–8.

19. Krishna V, Joshi A, Bulterys PL, Yang E, Ng AY, Rajpurkar P. GloFlow: global image alignment for creation of whole slide images for pathology from video. *arXiv.* (Preprint). (2020). arXiv:2010.15269.

20. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* (2017) 542:115–8. doi: 10.1038/nature21056

21. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med.* (2019) 25:1054–6. doi: 10.1038/s41591-019-0462-y

22. Ertosun MG, Rubin DL. Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. *AMIA Annu Symp Proc.* (2015) 2015:1899–908.

23. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Informat.* (2016) 7:29. doi: 10.4103/2153-3539.186902

24. Kong X, Sun G, Wu Q, Liu J, Lin F. Hybrid pyramid u-net model for brain tumor segmentation. *International conference on intelligent information processing.* Cham: Springer (2018). p. 346–55.

25. Liu Z, Song YQ, Sheng VS, Wang L, Jiang R, Zhang X, et al. Liver CT sequence segmentation based with improved U-Net and graph cut. *Expert Syst Appl.* (2019) 126:54–63.

26. Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat Methods.* (2019) 16:67–70.

27. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition.* Las Vegas, NV: IEEE (2016). p. 2818–26.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 07 December 2022
DOI 10.3389/fmed.2022.1072109

Check for updates

OPEN ACCESS

EDITED BY
Jun Cheng,
Shenzhen University, China

REVIEWED BY
Gang Yin,
Sichuan Cancer Hospital, China
Haoda Lu,
Nanyang Technological
University, Singapore

*CORRESPONDENCE
Chen Li
lichen@bmie.neu.edu.cn

SPECIALTY SECTION
This article was submitted to
Pathology,
a section of the journal
Frontiers in Medicine

RECEIVED 17 October 2022
ACCEPTED 18 November 2022
PUBLISHED 07 December 2022

CITATION
Hu W, Chen H, Liu W, Li X, Sun H, Huang X, Grzegorzek M and Li C (2022) A comparative study of gastric histopathology sub-size image classification: From linear regression to visual transformer. *Front. Med.* 9:1072109. doi: 10.3389/fmed.2022.1072109

A comparative study of gastric histopathology sub-size image classification: From linear regression to visual transformer

Weiming Hu[1], Haoyuan Chen[1], Wanli Liu[1], Xiaoyan Li[2], Hongzan Sun[3], Xinyu Huang[4], Marcin Grzegorzek[4,5] and Chen Li[1]*

[1]Microscopic Image and Medical Image Analysis Group, College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China, [2]Department of Pathology, Liaoning Cancer Hospital and Institute, Cancer Hospital, China Medical University, Shenyang, China, [3]Department of Radiology, Shengjing Hospital, China Medical University, Shenyang, China, [4]Institute of Medical Informatics, University of Luebeck, Luebeck, Germany, [5]Department of Knowledge Engineering, University of Economics in Katowice, Katowice, Poland

Introduction: Gastric cancer is the fifth most common cancer in the world. At the same time, it is also the fourth most deadly cancer. Early detection of cancer exists as a guide for the treatment of gastric cancer. Nowadays, computer technology has advanced rapidly to assist physicians in the diagnosis of pathological pictures of gastric cancer. Ensemble learning is a way to improve the accuracy of algorithms, and finding multiple learning models with complementarity types is the basis of ensemble learning. Therefore, this paper compares the performance of multiple algorithms in anticipation of applying ensemble learning to a practical gastric cancer classification problem.

Methods: The complementarity of sub-size pathology image classifiers when machine performance is insufficient is explored in this experimental platform. We choose seven classical machine learning classifiers and four deep learning classifiers for classification experiments on the GasHisSDB database. Among them, classical machine learning algorithms extract five different image virtual features to match multiple classifier algorithms. For deep learning, we choose three convolutional neural network classifiers. In addition, we also choose a novel Transformer-based classifier.

Results: The experimental platform, in which a large number of classical machine learning and deep learning methods are performed, demonstrates that there are differences in the performance of different classifiers on GasHisSDB. Classical machine learning models exist for classifiers that classify Abnormal categories very well, while classifiers that excel in classifying Normal categories also exist. Deep learning models also exist with multiple models that can be complementarity.

Discussion: Suitable classifiers are selected for ensemble learning, when machine performance is insufficient. This experimental platform demonstrates that multiple classifiers are indeed complementarity and can improve the efficiency of ensemble learning. This can better assist doctors in diagnosis, improve the detection of gastric cancer, and increase the cure rate.

KEYWORDS

gastric histopathology, sub-size image, robustness comparison, algorithmic complementarity, image classification

1. Introduction

Gastric cancer is a serious threat to human health as a global killer disease. According to the most recent Global Cancer Statistics Report, gastric cancer has become the fifth most common cancer and the fourth leading cause of death (1). Histopathological examination of gastric cancer constitutes the gold standard for the detection of gastric cancer and is a prerequisite for its management (2).

Histopathological examinations begin by staining the sections with Hematoxylin and Eosin (H&E), which are used to visualize the nuclei and cytoplasm of tissue sections, highlighting the fine structure of cells and tissues for physician observation (3). The pathologist finds the diseased area by gross observation of the pathological slides with the naked eye. The pathologist then observes and diagnoses the diseased area of the pathological section using the low-power microscope of the microscope. Pathologists can use high-power microscopes for careful observation and judgment (4). For the entire pathological slice diagnosis process (5), the following problems can be found: slice information is easy to ignore (6). This shows that there is subjectivity throughout the process. The workload of pathologists is huge and the working hours are long, which is highly likely to lead to misdiagnosis (7). Therefore, there is an urgent need to address the issues more intensively.

However, computer-aided diagnosis technology has advanced rapidly in recent years, and the emergence of medical image classification technology in computer vision technology can achieve fast and efficient help for doctors to examine gastric cancer tissue sections (8). Image classification techniques have brought new breakthroughs to discriminate between benign and malignant cancer, distinguish between stages of tumor differentiation and differentiate tumor subtypes, as image classification techniques can provide valid information for pathologists to refer to during the diagnostic process (9). In addition, the development direction of image classification technology is mainly to enhance the accuracy of classification algorithms and improve the anti-interference ability, ensemble learning becomes an effective solution, and it becomes especially important to find multiple efficient classification algorithms with complementarity properties (10). Moreover, there is a lack of computer performance in practical work, and computer-aided medical image analysis often crops full-slice images into sub-size pictures (11). Therefore, we compare the image classification performance of a large number of algorithms on sub-size images in order to expect to find algorithms with complementarity properties for ensemble learning to improve medical image classification performance.

The database used in this study is GasHisSDB (12), containing 245,196 images, of which there are 97,076 abnormal images and 148,120 normal images. GasHisSDB is a database containing three sub-databases, including sub-database A (160 ×160 pixels), Sub-database B (120 ×120 pixels.), Sub-database C (80 ×80 pixels). GasHisSDB provides the ability to distinguish between classical machine learning classifier performance and deep learning classifier performance (13). Details are given in Section 2.1.

Classical machine learning methods still have excellent classification results in the field of image classification (14). Existing methods can extract different features of images and supply different performance of classifiers for image classification (15). Exploring different features using appropriate classifiers to obtain efficient classification results is the basis of using ensemble learning for medical images (16). Therefore, in this study, five different image features including two color features and three texture features are extracted for GasHisSDB. After extracting the features seven different classifiers are used for classification. Details are given in Sections 2.2 and 2.3.

In the field of medical image classification, deep learning algorithms are the most effective algorithms, and Convolutional Neural Network (CNN) is a widely used model for image classification, which can extract information from original medical images and classify normal and abnormal case images (17). Recently, Visual Transformer, which is originally applied to Natural Language Processing tasks, have become popular in computer vision, and Vision Transformer (ViT) have effective classification results when trained on large amounts of data and can significantly reduce the computer hardware and software resources required for training (18). CNN-based deep learning models, this study used VGG6, Inception-V3 and ResNet50. Visual transformer-based deep learning models in this study used VIT. The above four deep learning models use the same parameters with the same database: GasHisSDB. Details are given in Section 2.4.

This study makes the following contributions to the field of sub-size pathology image classification:

TABLE 1 Dataset scale of GasHisSDB.

Sub-database name	Cropping size	Abnormal	Normal
Sub-database A	160 × 160 pixels	13,124	20,160
Sub-database B	120 × 120 pixels	24,801	40,460
Sub-database C	80 × 80 pixels	59,151	87,500
Total		97,076	148,120

- Extensive testing is done and the complementarity of different classification methods is found.
- According to the complementarity, it can provide a basis for future ensemble learning research.

This paper is structured as follows: In Section 2, we detail the dataset used, classical classification methods, and deep learning methods. In Section 3, we show the comparative experimental setup, evaluation metrics and experimental results. In Section 4, we compare the experimental results and analyze them. In Section 5, we summarize the research and suggest future research directions.

2. Materials and methods

2.1. Dataset: GasHisSDB

The publicly available dataset GasHisSDB is used in this study to compare the performance of various learning models, expecting to discover the complementarity of various models in ensemble learning (12). The database contains three sub-datasets with a total of 245,196 images, and the size and number are shown in Table 1. The database is a sub-size gastric cancer pathology H&E staining image database, which contains two categories of images: normal and abnormal. The abnormal image contains more than 50% of the cancerous area, and the normal image is the image of the normal pathological slice tissue. Some examples of the GasHisSDB database are shown in Figure 1.

GasHisSDB contains images in png format acquired using electron microscopy. GasHisSDB contains two categories and the details of the two categories are shown below:

- Normal: each normal image does not contain cancerous regions. Each cell is almost free of anisotropy. In addition, the nuclei of the cells in the images have almost no mitosis and are arranged in a regular layer. Therefore, when observed under the light microscope, if no elimination of any cells and tissues is observed and the characteristics of a normal image are met, it can be judged as a normal image (19).
- Abnormal: Each abnormal image contains more than 50% of gastric cancer images. The general morphology of gastric cancer is mostly ulcerative. As the disease progresses, the cancer nest infiltrates from the mucosal layer to the muscular layer and plasma layer. The texture is hard and the cross-section is often grayish white. Under microscopic observation, the cancer cells can be arranged in nest-like, glandular vesicle-like, tubular or cord-like, and the boundary with the interstitium is usually clear. However, when cancer cells infiltrate the stroma, the borders between them are not clear. Based on these facts, abnormal pathological images can be judged when cells are observed to form unevenly sized, irregularly shaped, and irregularly arranged glandular or adenoid structures (19).

2.2. Methods of feature extraction

To extract a variety of virtual features of GasHisSDB is a prerequisite for classification using classical machine learning classifiers. In the comparison experiments, five methods are used to extract visual features from the database, including Color histogram, Luminance histogram, Histogram of Oriented Gradient (HOG), Local Binary Patterns (LBP), and Gray-level Co-occurrence Matrix (GLCM).

2.2.1. Color histogram

Among the different methods of feature extraction, the most common method to describe the color features of an image is the color histogram. The color histogram clearly represents the color spread in the image. The color histogram has the characteristic of being unaffected by image rotation and shift changes and by further normalization of image scale changes. It is especially applicable to describe images that are resistant to automatic segmentation and images that do not require consideration of the spatial location of subjects. However, the color histogram does not characterize the partial spread of colors in an image, the spatial location of each color, and specific objects. In this experiment, the luminance histogram is used as the luminance feature. The luminance feature is expressed as a histogram of the average of the three color components.

2.2.2. Texture features

The texture is a visual feature that reflects homogeneous phenomena in an image (20). That reflects the structure and arrangements of the surface structures on the surface of an object with slow or periodic changes (21). A texture feature is not a pixel-based feature. It requires statistical computation of regions containing multiple pixels, such as the grayscale distribution of pixels and their surrounding spatial neighbors, and local texture information. In addition, the global texture information is reflected as the repetition degree of local texture information.

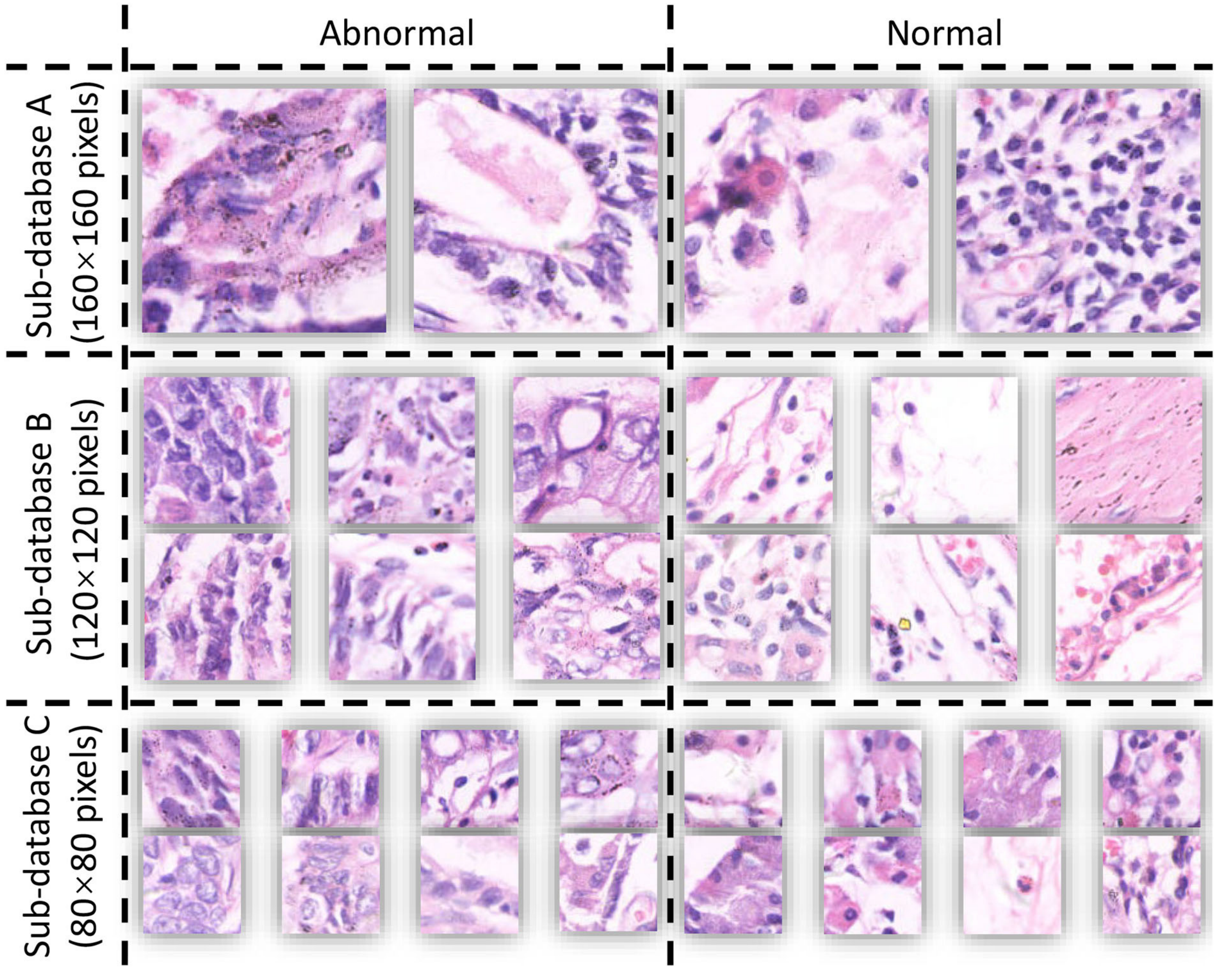

Abnormal
Normal
Sub-database A (160×160 pixels)
Sub-database B (120×120 pixels)
Sub-database C (80×80 pixels)

FIGURE 1
Example of GasHisSDB.

In this experiment, three texture features are extracted, which are HOG, LBP, and GLCM.

HOG is a feature descriptor commonly used in image processing for object detection. Features are constructed by computing a histogram of the gradient direction of local regions of an image. HOG has the property of operating on the local units of the image. So it has the advantage of maintaining excellent invariance in terms of geometric and optical distortion of the image. LBP has advantages such as gray invariance and rotation invariance, and the features are easy to compute. GLCM is defined by the joint probability density of pixels at two locations and is a second-order statistical feature about the variation of image brightness. It not only reflects the distribution of luminance. It also reflects the distribution of positions between pixels with the same or similar luminance. The main statistical values are: Contrast, Correlation, Energy, and Homogeneity.

2.3. Classical classification models

After the feature extraction step, complementarity comparison tests for image classification are performed using seven classical machine learning methods, including Linear Regression, *k*-Nearest Neighbor (*k*NN), naive Bayesian classifier, Random Forest (RF), linear Support Vector Machine (linear SVM), non-linear Support Vector Machine (non-linear SVM), and Artificial Neural Network (ANN).

Classical machine learning methods perform image classification by using virtual features. Linear Regression is a method to get a linear model as much as possible to accurately predict the real value output label. In Linear regression, the least square function is used to establish the relationship between one or more independent variables (22). An easy and commonly used supervised learning method is *k*NN. The main idea of *k*NN is to first find the nearest *k* samples based on the distance and then vote for the prediction result (23). The naive Bayesian classifier based on Bayesian decision theory in probability theory (24). RF is a parallel integrated learning method based on a decision tree learner. RF adds random attribute selection to the training process of decision trees (25). SVMs are divided into linear and non-linear. The difference between the two is mainly that the kernel functions of both are different (26). Linear SVM maps training examples to points in space to maximize the gap between the two categories. Then, the new examples are mapped to the same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVM can also use a kernel function to perform non-linear classification effectively, thereby implicitly mapping its input to a high-dimensional feature space. The ANN is a classification algorithm composed of a structure that simulates human brain neurons and is trained through a propagation algorithm (27).

2.4. Deep learning models

Complementarity comparison experiments use deep learning models for the classification of gastric cancer pathology images (28). First, the model is trained using training and validation sets generated from three sub-datasets of GasHisSDB. The test set is used in this experiment to evaluate the models' performance (29). Comparative analysis of multiple classification results is performed using the obtained evaluation metrics to determine if the classifiers would be complementarity in Ensemble learning (30). This experiment uses four deep learning models. Three of the models are based on CNNs, including VGG16, Inception-V3, and ResNet50. One more model corresponds to VT, which is ViT (31).

VGG is a convolutional neural network (CNN) improved by AlexNet, developed by Visual Geometry Group and Google DeepMind in 2014, and the most commonly used one in image classification is VGG16 (32). In 2014, Google's InceptionNet made its debut at the ILSVRC competition. Several versions of InceptionNet have been developed, with Inception-V3 being one of the more representative versions of this large family (33). He et al. proposed ResNet to address the difficulty of training deep networks due to gradient disappearance. The most commonly used in the field of image classification is ResNet50 (34). In recent years, Dosovitskiy et al. have proposed the ViT model using transformer. This model is not only very effective in the field of natural language processing, but also provides good results in the field of image classification. Effectively reduces the dependence of computer vision on CNN (35).

3. Experiment

3.1. Comparative experimental setup

The main process of complementarity experiments is divided into two parallel parts: The classification results of classical models and deep learning models are both analyzed and evaluated. The experimental flow is shown in Figure 2.

The various settings of the experimental platform are as follows:

1. Hardware configuration: The complementarity comparison experiment is conducted on a local computer with the Win10 operating system. The computer has 32 GB of running memory and is equipped with an 8 GB NVIDIA Quadro RTX 4000.
2. Data set partitioning: In this experiment, the training set, validation set and test set are divided in the ratio of 4:4:2.
3. Classical machine learning software configuration: The classical programming software use for machine learning is Matlab R2020a (9.8.0.132 350 2).
4. Deep learning software configuration: The Pytorch version 1.7.1 framework in Deep Learning Python 3.6 is very mature, and the code for this part of the experiment is done using them.
5. Classical machine learning parameter settings: The same parameters are used for all classification comparison experiments. In *k*NN, *k* is set to 9. The number of trees in RF is set to 10. The kernel function of the non-linear SVM is a Gaussian kernel. The ANN uses a 2-layer network with 10 nodes in the first layer and 3 nodes in the second layer. The number of epochs for ANN training is set to 500, the learning rate is set to 0.01, and the expected loss is set to 0.01.
6. Deep learning parameter settings: This part of the experiment focuses on classifying GasHisSDB using four deep learning methods to observe model complementarity. A learning rate of 0.00002 is used for each model, and the batch size is set to 32. One hundred epochs of experiments are performed to observe the classification results of this database on different models.

3.2. Evaluation metrics

The selection of evaluation indicators is important in complementarity comparison papers. In the experiments of this thesis, Accuracy (Acc) is the most significant metric, but also Precision (Pre), Recall (Rec), Specificity (Spe), and F1-score (F1) are selected. These selected metrics are very commonly used in comparison papers to analyze classifiers and thus better identify their complementarities to enhance and improve ensemble learning (36).

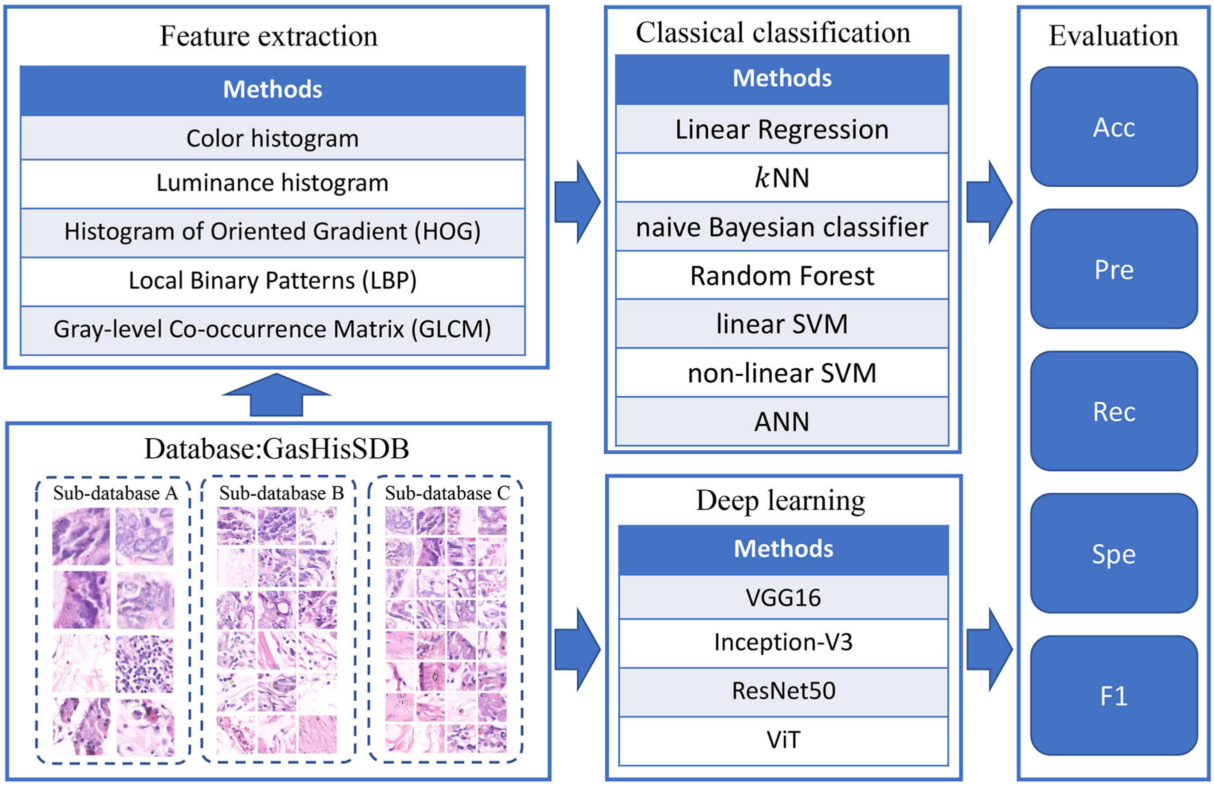


FIGURE 2
Workflow of the complementarity comparison experiment.

In the case of positive-negative binary classification, true positives (TP) correspond to the number of positive samples that are accurately predicted. The number of negative samples predicted to be positive is called false positive (FP). The number of positive samples predicted to be negative samples is called false negative (FN). True Negative (TN) is the number of negative samples predicted accurately (37).

The five evaluation indicators are described below and the formulas are shown in Table 2.

1. Acc: Accuracy is the ratio of the number of correct predictions to the total number of samples.
2. Pre: Precision is a measure of accuracy, indicating the proportion of examples classified as positive that are actually positive.
3. Recl: Recall is a measure of coverage, a measure of the number of positive examples classified as positive examples, indicating the proportion of all positive examples classified as pairs, which measures the ability of the classifier to identify positive examples.
4. Spe: Specificity indicates the proportion of all negative cases that were scored correctly, and measures the classifier's ability to identify negative cases.
5. F1: F1-Score combines Precision and Recall. Accuracy is the ratio of the number of correct predictions to the total number of samples.

TABLE 2 Evaluation metrics.

Assessment	Formula
Accuracy (Acc)	(TP + TN)/(TP + TN + FP + FN)
Precision (Pre)	TP/TP + FP
Recall (Rec)	TP/TP + FN
Specificity (Spe)	TN/TN + FP
F1-score (F1)	2 × (Pre ×Rec)/(Pre + Rec)

3.3. Experimental results

We set up an experimental platform to conduct various classification experiments on three sub-databases of the GasHisSDB. A large amount of experimental data is obtained for our experiments in order to investigate the complementarity of different methods (38).

The comparative results of classical machine learning methods are shown in Tables 3–5.

Table 6 show the comparison results of the deep learning methods.

4. Evaluation of results

4.1. Evaluation of classical machine learning methods

4.1.1. On 160 × 160 pixels sub-database

This section focuses on the classification results of classical machine learning methods for the 160 ×160 sub-database.

The color histogram has the highest number of items among all features. According to Table 3, the classical machine learning classifier on the color histogram, the best performer is RF with an accuracy of 85.99%. In addition, in color histogram, the classification accuracy of the three classifiers reached around 80%, which are LR, *k*NN, and ANN. All SVM classifiers perform poorly on color histogram features. However, color histogram on GasHisSDB, the naive Bayesian classifier, cannot get the classification effect because of the existence of a large number of low luminance statistics with zero values in the three color channels.

The luminance is the average of the colors. Its histogram does not yield better classification accuracy as a feature. Because of this, luminance histogram also has the above problem on the naive Bayesian classifier. The classification results of the naive Bayesian classifier for these two color features are therefore not presented in the Table 3. RF shows robustness in two features and obtains the highest accuracy rate of 79.13% using luminance histogram for classification. However, the LR, *k*NN, and ANN classifiers that perform better on color histogram significantly drop on luminance histogram.

The classification effect of HOG on all classifiers is not very effective and the accuracy is very close. The difference is not much distributed between 53 and 62%.

On the contrary, the distribution of LBP image classification accuracy is particularly scattered, with the highest Linear Regression classifier reaching 74.29%, followed by ANN reaching 71.84%. The lowest linear SVM classification effect is <50%.

The classification effect of the four statistic values of GLCM is 71.39% only for RF, and other classifiers are also above 60%. It is worth noting that the accuracy of non-linear SVM with other features except color histogram and GLCM has not changed at all, which is 60.58%. The accuracy of non-linear SVM classifier with color histogram is 56.09% and the accuracy of GLCM's non-linear SVM classifier is 67.76%.

4.1.2. On 120 × 120 pixels sub-database

Here, we focus on the comparison of the experimental results of the 120 × 120 pixels sub-database. The experimental results are shown in Table 3. In general, compared with 160×160 pixels sub-database classification results, 120 × 120 pixels sub-database classification results except for color histogram, the rest of the best classifiers remain unchanged.

The four better-performing classifiers on color histogram feature still perform better, and the accuracy rate fluctuates slightly, resulting in the *k*NN classifier reaching the best accuracy rate of 86.32%. The classification performance of the two SVM classifiers on the features of color histogram is still not ideal. Naive Bayesian classifier is still not suitable for color histogram and luminance histogram features. The linear SVM effect of luminance histogram classifier has been greatly improved in the classification of the 120 × 120 pixels sub-database. The accuracy of other classifiers on the features of luminance histogram has little change. The HOG feature still does not perform well in every classifier. The highest accuracy rate is only 62.35% of ANN. The classification results of LBP and GLCM features are similar to the classification effect on the 160 × 160 pixels sub-database. The best accuracy rate on LBP is a linear regression with a precision rate of 73.34%. The best accuracy rate on GLCM is that the RF reaches 71.15%. Similarly, the non-linear SVM of 120 × 120 pixels sub-database also has the problem of constant accuracy of multiple features.

4.1.3. On 80 × 80 pixels sub-database

The classification results of the 80 × 80 pixels sub-database are shown in Table 4. The overall best classifier on each feature remains the same as that of the best classifier for each feature corresponding to the 120 × 120 pixels sub-database except for HOG features that have a small gap between each classifier.

Compared with the classification results of the other two sub-databases, the classification effect of each classifier on color histogram and luminance histogram has no particularly large fluctuations. It confirms the consistency of the three databases of GasHisSDB.

The classification accuracy of color histogram is still polarized. The four excellent classifiers reach about 80%, and the other two are about 60%. The RF still showed robustness in the luminance histogram classification task. RF was the best classifier with an accuracy of 75.10%. The classification accuracy distribution of HOG features is denser than that of the other two sub-databases. The highest is only 59.87%. Due to the reduced sample size, each classifier has different degrees of accuracy reduction in addition to the naive Bayesian classifier for LBP features and GLCM features. The best classifier for LBP feature is still linear regression which reaches 70.92%. The highest accuracy rate of LBP feature has become 68.84% of *k*NN. In the classification results of the 80 × 80 pixels sub-database, the naive Bayesian classifier of color histogram and luminance

TABLE 3 Classification results of five image features using different classifiers in the 160 × 160 pixels sub-database of GasHisSDB [In (%)].

Freatures	Methods	Acc	Abnormal				Normal			
			Pre	Rec	Spe	F1	Pre	Rec	Spe	F1
Color histogram	LR	83.29	81.32	80.42	85.54	80.87	84.80	85.54	80.42	85.17
	*k*NN	85.52	82.95	84.35	86.43	83.64	87.58	86.43	84.35	87.01
	RF	**85.99**	81.65	87.83	84.55	84.63	89.88	84.55	87.83	87.13
	Linear SVM	41.12	33.92	35.96	45.16	34.91	47.40	45.16	35.96	46.25
	Non-linear SVM	56.09	Null	0.00	100.00	0.00	56.09	100.00	0.00	71.87
	ANN	78.89	77.78	72.68	83.75	75.14	79.67	83.75	72.68	81.66
Luminance histogram	LR	70.97	67.95	49.92	84.67	57.56	72.21	84.67	49.92	77.95
	*k*NN	77.10	70.30	72.60	80.03	71.43	81.78	80.03	72.60	80.90
	RF	**79.13**	72.17	76.60	80.78	74.32	84.14	80.78	76.60	82.42
	Linear SVM	42.34	40.50	98.67	5.68	57.43	86.74	5.68	98.67	10.66
	Non-linear SVM	60.58	Null	0.00	100.00	0.00	60.58	100.00	0.00	75.45
	ANN	71.23	64.74	59.34	78.97	61.92	74.90	78.97	59.34	76.88
HOG	LR	60.46	48.96	7.20	95.11	12.56	61.16	95.11	7.20	74.45
	*k*NN	61.42	51.31	41.65	74.28	45.98	66.17	74.28	41.65	69.99
	Naive Bayesian	54.43	45.11	71.84	43.11	55.42	70.17	43.11	71.84	53.40
	RF	60.85	50.33	53.01	65.95	51.63	68.32	65.95	53.01	67.11
	Linear SVM	53.28	44.82	80.14	35.79	57.49	73.47	35.79	80.14	48.13
	Non-linear SVM	60.58	Null	0.00	100.00	0.00	60.58	100.00	0.00	75.45
	ANN	**61.54**	54.30	15.40	91.57	23.99	62.45	91.57	15.40	74.26
LBP	LR	**74.29**	69.32	62.42	82.02	65.69	77.03	82.02	62.42	79.45
	*k*NN	70.21	66.11	50.11	83.28	57.01	71.95	83.28	50.11	77.20
	Naive Bayesian	57.71	47.78	78.28	44.32	59.34	75.82	44.32	78.28	55.94
	RF	70.27	62.16	62.84	75.10	62.50	75.64	75.10	62.84	75.37
	Linear SVM	48.17	36.83	44.02	50.87	40.10	58.27	50.87	44.02	54.32
	Non-linear SVM	60.58	Null	0.00	100.00	0.00	60.58	100.00	0.00	75.45
	ANN	71.84	67.38	55.41	82.54	60.81	73.99	82.54	55.41	78.03
GLCM	LR	67.73	59.71	55.75	75.52	57.67	72.40	75.52	55.75	73.93
	*k*NN	69.26	62.30	55.79	78.03	58.87	73.06	78.03	55.79	75.46
	Naive Bayesian	61.99	51.12	82.01	48.96	62.98	80.70	48.96	82.01	60.94
	RF	**71.39**	63.16	65.85	75.00	64.48	77.14	75.00	65.85	76.06
	Linear SVM	66.50	55.89	71.27	63.39	62.65	77.22	63.39	71.27	69.63
	Non-linear SVM	67.76	58.77	61.05	72.12	59.89	73.99	72.12	61.05	73.05
	ANN	68.69	60.64	58.65	75.22	59.63	73.65	75.22	58.65	74.43

The bold text in the table indicates the highest value of the classification result of different classifiers for the same feature.

histogram is not applicable, and, except for the GLCM feature, the problem that the accuracy of the non-linear SVM classifier does not change still exists.

4.2. Evaluation of deep learning methods

4.2.1. On 160 × 160 pixels sub-database

According to Table 5, on 160 × 160 pixels sub-database, all deep learning models have better classification results than classical machine learning methods. The VGG model with the longest training time and the largest model size has an accuracy above 95%. Inception-V3 and ResNet50 have better model size and training time than VGG16. However, Inception-V3 has lower accuracy than VGG16, and ResNet50 has the highest accuracy of 96.09%, which is the highest among all models. ViT is a Transformer-based classifier with an accuracy of 86.21%. However, it is still higher than the classification accuracy of all traditional machine learning methods on this sub-database. Significantly, ViT achieves such accuracy with only 1/4 of the

TABLE 4 Classification results of five image features using different classifiers in the 120 × 120 pixels sub-database of GasHisSDB [In (%)].

Freatures	Methods	Acc	Abnormal				Normal			
			Pre	Rec	Spe	F1	Pre	Rec	Spe	F1
Color histogram	LR	83.46	80.01	75.28	88.47	77.57	85.38	88.47	75.28	86.90
	*k*NN	**86.32**	82.28	81.55	89.24	81.92	88.75	89.24	81.55	88.99
	RF	86.08	80.36	83.87	87.43	82.08	89.84	87.43	83.87	88.62
	Linear SVM	46.28	39.48	77.62	27.06	52.34	66.36	27.06	77.62	38.45
	Non-linear SVM	62.00	Null	0.00	100.00	0.00	62.00	100.00	0.00	76.54
	ANN	81.20	78.14	70.14	87.98	73.93	82.78	87.98	70.14	85.30
Luminance histogram	LR	71.28	66.89	48.39	85.32	56.15	72.95	85.32	48.39	78.65
	*k*NN	76.43	68.19	71.17	79.65	69.65	81.84	79.65	71.17	80.73
	RF	**77.60**	69.36	73.57	80.08	71.40	83.17	80.08	73.57	81.60
	Linear SVM	58.54	47.45	84.62	42.55	60.80	81.86	42.55	84.62	55.99
	Non-linear SVM	62.00	Null	0.00	100.00	0.00	62.00	100.00	0.00	76.54
	ANN	71.18	62.97	58.65	78.86	60.73	75.68	78.86	58.65	77.23
HOG	LR	61.78	33.72	0.58	99.30	1.15	61.97	99.30	0.58	76.31
	*k*NN	62.02	50.04	39.56	75.79	44.18	67.17	75.79	39.56	71.22
	Naive Bayesian	54.83	44.29	73.06	43.66	55.15	72.56	43.66	73.06	54.52
	RF	60.55	48.15	49.62	67.25	48.87	68.53	67.25	49.62	67.88
	Linear SVM	50.91	39.90	57.60	46.81	47.14	64.30	46.81	57.60	54.18
	Non-linear SVM	62.00	Null	0.00	100.00	0.00	62.00	100.00	0.00	76.54
	ANN	**62.35**	54.37	5.77	97.03	10.43	62.69	97.03	5.77	76.17
LBP	LR	**73.34**	67.20	58.29	82.56	62.43	76.35	82.56	58.29	79.34
	*k*NN	70.27	64.05	49.64	82.92	55.93	72.87	82.92	49.64	77.57
	Naive Bayesian	57.39	46.41	78.43	44.49	58.31	77.09	44.49	78.43	56.42
	RF	70.13	60.88	59.90	76.41	60.39	75.66	76.41	59.90	76.03
	Linear SVM	46.21	29.70	30.40	55.89	30.05	56.71	55.89	30.40	56.30
	Non-linear SVM	62.00	Null	0.00	100.00	0.00	62.00	100.00	0.00	76.54
	ANN	71.19	64.72	53.19	82.23	58.39	74.13	82.23	53.19	77.97
GLCM	LR	67.54	58.21	51.65	77.27	54.74	72.28	77.27	51.65	74.69
	*k*NN	69.79	61.98	53.04	80.05	57.16	73.56	80.05	53.04	76.67
	Naive Bayesian	61.40	49.52	80.77	49.53	61.39	80.77	49.53	80.77	61.41
	RF	**71.15**	61.42	64.72	75.09	63.03	77.64	75.09	64.72	76.34
	Linear SVM	66.66	55.02	67.30	66.28	60.54	76.78	66.28	67.30	71.14
	Non-linear SVM	69.43	60.08	58.27	76.27	59.16	74.88	76.27	58.27	75.57
	ANN	68.10	58.45	55.56	75.79	56.97	73.56	75.79	55.56	74.66

The bold text in the table indicates the highest value of the classification result of different classifiers for the same feature.

training time and 1/3 of the model size compared to ResNet. Also, the accuracy curve is still trending upward and the loss function is still not fully converged.

4.2.2. On 120 × 120 pixels sub-database

According to the Table 5, the classification results are excellent on the sub-database of 120 × 120 pixels. Due to a large number of training samples, VGG16 is the classifier with the highest accuracy of 96.47% on this sub-database. However, the training time is doubled compared to that on the 160 × 160 sub-database. The accuracies of 95.83 and 95.94% are obtained for Inception-V3 and ResNet50, respectively. Due to the increase in the amount of training data, ViT also gained an accuracy improvement, rising to 89.44%.

4.2.3. On 80 × 80 pixels sub-database

According to Table 5, the classification results of the 80 × 80 subdatabase can be seen. It is the sub-database with the largest number of samples, and the accuracy of the four classifiers

TABLE 5 Classification results of five image features using different classifiers in the 80 × 80 pixels sub-database of GasHisSDB [In (%)].

Freatures	Methods	Acc	Abnormal				Normal			
			Pre	Rec	Spe	F1	Pre	Rec	Spe	F1
Color histogram	LR	82.22	78.17	77.59	85.35	77.88	84.93	85.35	77.59	85.14
	*k*NN	**85.24**	80.60	83.52	86.41	82.03	88.58	86.41	83.52	87.48
	RF	83.27	77.14	83.15	83.34	80.03	87.98	83.34	83.15	85.60
	Linear SVM	60.81	50.86	83.58	45.41	63.24	80.36	45.41	83.58	58.03
	Non-linear SVM	59.67	Null	0.00	100.00	0.00	59.67	100.00	0.00	74.74
	ANN	79.28	76.60	70.03	85.54	73.17	80.85	85.54	70.03	83.13
Luminance histogram	LR	70.16	66.79	51.77	82.60	58.33	71.70	82.60	51.77	76.76
	*k*NN	74.65	67.67	71.11	77.04	69.35	79.78	77.04	71.11	78.38
	RF	**75.10**	67.77	72.94	76.55	70.26	80.71	76.55	72.94	78.58
	Linear SVM	54.58	46.58	85.81	33.47	60.38	77.72	33.47	85.81	46.79
	Non-linear SVM	59.67	Null	0.00	100.00	0.00	59.67	100.00	0.00	74.74
	ANN	70.17	63.19	62.38	75.43	62.78	74.79	75.43	62.38	75.11
HOG	LR	**59.87**	53.42	3.96	97.67	7.37	60.07	97.67	3.96	74.39
	*k*NN	59.63	49.95	42.22	71.40	45.76	64.64	71.40	42.22	67.85
	Naive Bayesian	55.91	46.97	72.35	44.79	56.96	70.56	44.79	72.35	54.79
	RF	59.08	49.31	51.88	63.95	50.56	66.28	63.95	51.88	65.10
	Linear SVM	53.47	44.46	61.60	47.98	51.64	64.89	47.98	61.60	55.17
	Non-linear SVM	59.70	90.91	0.08	99.99	0.17	59.68	99.99	0.08	74.75
	ANN	59.67	50.08	2.49	98.32	4.75	59.87	98.32	2.49	74.42
LBP	LR	**70.92**	65.32	59.49	78.65	62.27	74.17	78.65	59.49	76.34
	*k*NN	68.48	63.20	52.32	79.41	57.24	71.13	79.41	52.32	75.04
	Naive Bayesian	59.09	49.55	77.69	46.52	60.51	75.52	46.52	77.69	57.57
	RF	68.16	60.13	62.49	71.98	61.29	73.95	71.98	62.49	72.95
	Linear SVM	43.10	27.68	25.48	55.01	26.53	52.20	55.01	25.48	53.56
	Non-linear SVM	59.67	Null	0.00	100.00	0.00	59.67	100.00	0.00	74.74
	ANN	68.57	62.75	54.32	78.21	58.23	71.69	78.21	54.32	74.81
GLCM	LR	65.56	57.32	57.24	71.19	57.28	74.65	71.21	64.23	72.89
	*k*NN	**68.84**	62.32	57.53	76.49	59.83	72.71	76.49	57.53	74.55
	naive Bayesian	62.12	51.96	80.87	49.45	63.27	79.27	49.45	80.87	60.91
	RF	68.39	60.13	64.23	71.21	62.11	74.65	71.21	64.23	72.89
	Linear SVM	66.82	57.14	71.04	63.97	63.33	76.57	63.97	71.04	69.71
	Non-linear SVM	68.31	61.03	59.26	74.42	60.13	72.99	74.42	59.26	73.70
	ANN	65.52	56.70	61.40	68.30	58.96	72.36	68.30	61.40	70.27

The bold text in the table indicates the highest value of the classification result of different classifiers for the same feature.

only changes slightly. VGG16 performs stably with an accuracy of 96.12%, which is the classification model with the highest accuracy. The lowest accuracy is still the ViT model with the least training time, at 90.23. It is worth noting that the training time of ViT is 13.26% of that of the highest accurate VGG16 on this sub-database.

4.3. Additional experiment

As stated in Section 4.2.1, ViT did not converge completely within 100 epochs. Experiments are added in this section to explore the performance of ViT, and the results are reflected in the last row of each sub-database in Table 5. The same parameter conditions were maintained for all additional experiments. In the additional experiments for the 160 × 160 sub-database, the control training time was similar to that of Inception-V3 and ResNet running 100 epochs. ViT runs 400 epochs and the accuracy reaches 92.23%. In the other two sub-databases with larger amount of data, again when controlling for the same training time as Inception-V3 and RseNet50. At this time, the accuracy of ViT models for the 120 × 120 pixel sub-database and the 80 × 80 pixel sub-database improves to 94.59 and 94.57%, respectively. The model size of ViT has a great advantage. Moreover, these image classification results reach the general level of medical image classification.

TABLE 6 Classification results of four deep learning classifiers on GasHisSDB [In (%)].

Sub-database size	Model	Quantity of epoch	Model size (MB)	Best eopch	Training time(s)	Acc	Category	Pre	Rec	Spe	F1
160 × 160 pixels	VGG16	100	268.16	100	13,873	95.90	Abnormal	93.8	96.0	95.9	94.9
							Normal	97.3	95.9	96.0	96.6
	Inception-V3	100	89.69	92	10,296	94.57	Abnormal	94.1	92.0	96.2	93.0
							Normal	94.9	96.2	92.0	95.5
	ResNet50	100	83.12	84	10,023	**96.09**	Abnormal	94.6	95.6	96.4	95.1
							Normal	97.1	96.4	95.6	96.7
	ViT	100	31.17	97	2,587	86.21	Abnormal	83.8	80.6	89.9	82.2
							Normal	87.7	89.9	80.6	88.8
		400	31.17	399	10,014	92.23	Abnormal	92.1	87.8	95.1	89.9
							Normal	92.3	95.1	87.8	93.7
120 × 120 pixels	VGG16	100	268.16	100	26,105	**96.47**	Abnormal	96.7	94.0	98.0	95.3
							Normal	96.4	98.0	94.0	97.2
	Inception-V3	100	89.69	98	19,719	95.83	Abnormal	94.6	94.4	96.7	94.5
							Normal	96.6	96.7	94.4	96.6
	ResNet50	100	83.12	94	19,087	95.94	Abnormal	96.2	93.0	97.8	94.6
							Normal	95.8	97.8	93.0	96.8
	ViT	100	31.17	100	4,077	89.44	Abnormal	87.0	84.9	92.2	85.9
							Normal	90.9	92.2	84.9	91.5
		500	31.17	496	20,410	94.59	Abnormal	93.5	93.4	95.3	93.2
							Normal	95.4	95.9	92.5	95.6
80 × 80 pixels	VGG16	100	268.16	90	62,152	**96.12**	Abnormal	94.2	96.3	96.0	95.2
							Normal	97.4	96.0	96.3	96.7
	Inception-V3	100	89.69	99	43,926	95.41	Abnormal	95.5	93.0	97.0	94.2
							Normal	95.3	97.0	93.0	96.1
	ResNet50	100	83.12	97	41,992	96.09	Abnormal	96.2	94.0	97.5	95.1
							Normal	96.0	97.5	94.0	96.7
	ViT	100	31.17	89	8,247	90.23	Abnormal	86.3	90.1	90.3	88.2
							Normal	93.1	90.3	90.1	91.7
		500	31.17	496	41,135	94.57	Abnormal	93.1	93.4	95.3	93.2
							Normal	95.6	95.3	93.4	95.4

The bold text in the table indicates the maximum value or the best index of the classification results of different categories.

4.4. t-SNE method analysis

To explore the possibility of ensemble learning between deep learning classifiers, we conducted a TSNE analysis of the top performing deep learning classifiers. the t-SNE method analysis was performed using the 160 × 160 pixels sub-database as an example and the results are shown in Figure 3.

This experimental platform use the t-SNE method to downscale the features extracted by the four deep learning methods into two-dimensional scatters displayed in the image. Representative images from the test set are selected in the figure, where the abnormal image suffers from misclassification in ViT, and its points after feature downscaling fall in the image normal population. This image performs well in the other three classifiers, and its feature-descended points fall in the image abnormal population. However, it can be observed that the selected normal image it performs well in Inception-V3, ResNet50, ViT, with the reduced points falling in the normal population, but performs poorly in VGG16.

5. Discussion

This chapter compares the classification results of different classifiers from the Linear Regression to Visual Transformer on the 160 × 160, 120 × 120, and 80 × 80 pixels sub-databases of the GasHisSDB. The classification performance of each method on GasHisSDB reflects complementarity.

Classical machine learning methods have a rigorous theoretical foundation. Their simplified ideas can show good classification results on some specific features and algorithms (39).

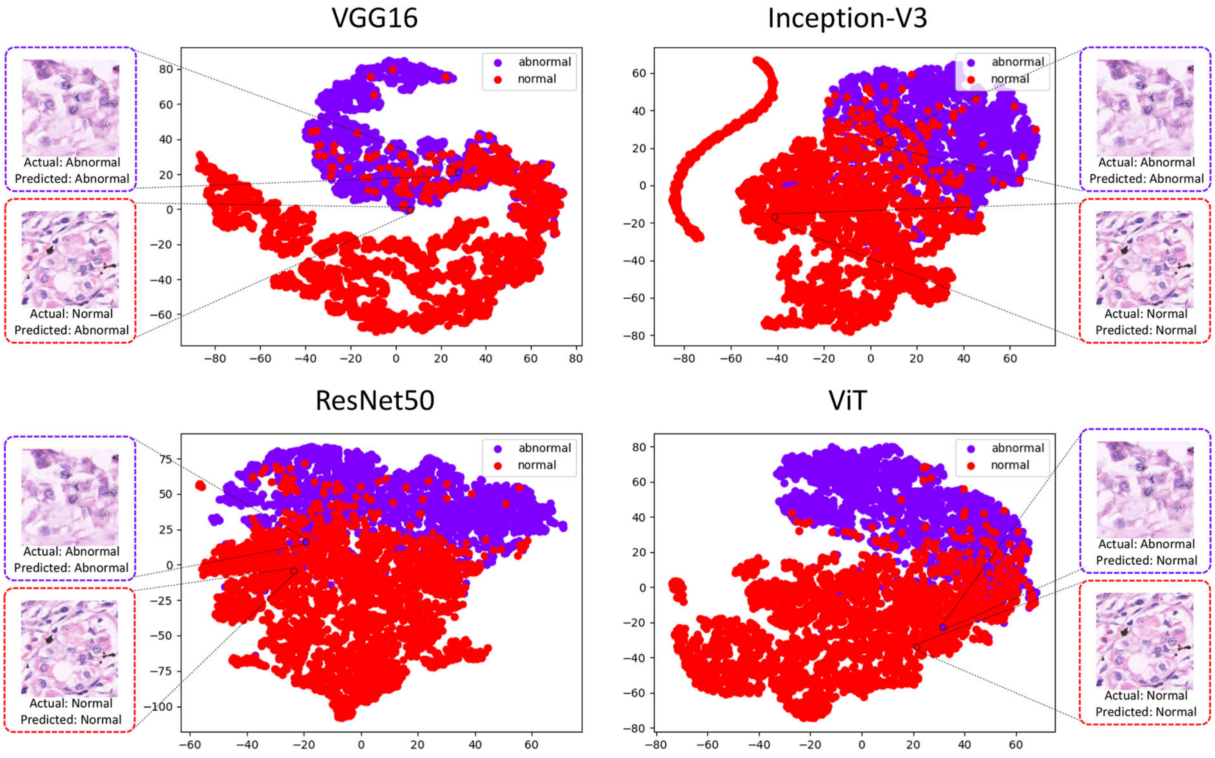


FIGURE 3
Plot of results from t-SNE analysis of four deep learning classification models.

This experimental platform shows that seven classifiers for GLCM classification on three sub-databases with little difference in accuracy, where the naive Bayesian classifier has significantly higher Rec than Spe for the abnormal category, and the linear SVM has slightly higher Rec than Spe. It shows that these two classifiers are better in classifying the abnormal category. However, the Spe of the other classification models are higher than the Rec, indicating that they are more effective in classifying the normal category. The same phenomenon occurs for every feature of every sub-database. There exist classifiers with high Rec values or high Spe values in the same condition. Such a result can be a powerful indication of the existence of this complementarity of these classifiers.

However, deep learning methods are still far ahead of classical machine learning methods in terms of image classification accuracy and experiment workload (40).

By analyzing the deep learning methods using the t-SNE method, there is a clear classification performance for their feature extraction. In Figure 3 it can also be seen that there is an aggregation of normal and abnormal images in the four classifiers. However, there is still inconsistency in the classification results and it can be understood that these methods can exist to some extent in a complementary manner (41).

The evaluation metrics for deep learning models are generally high, but complementarity in the field of machine learning also occurs in the field of deep learning (42). For example, the Spe of Inception-V3 and ResNet50 on sub-database C for abnormal category classification is high, but the high Rec of VGG16 can be well performed to the complementarity of the above two models.

The selection of suitable classifiers is the primary problem of ensemble learning, and after relevant experiments in the complementarity comparison experimental platform, it can be observed that these classifiers exhibit different performances (43). The complementarity possessed by these classifiers can adequately meet the needs of ensemble learning (44).

6. Conclusion and future work

In practice, machine performance often limits model training for large-size images, and finding multiple classification models with complementarity types is the basis for ensemble learning. For sub-sized images, this experiment tries a large number of classification models to find their complementarity and thus improve the efficiency of ensemble learning.

The experimental results show that complementarity in machine learning does exist for different classifiers of the same feature. Different classifiers for the same feature include classifiers that classify the abnormal category well and classifiers that classify the normal category well. This is a powerful indication of the complementarity among classifiers.

The evaluation metrics of the deep learning models are both very excellent. There are models that are less effective in classifying the abnormal category than the normal category. In this case, selecting the appropriate model that performs well for the abnormal category can contribute to ensemble learning. Complementarity can also be demonstrated in this situation.

There are still many excellent methods that have not been added to the experimental platform. Moreover, the recently popular ViT excels in the field of image processing, but ViT does not show significant experimental results on sub-size images. In the future, we will add more models to explore the complementarity nature of ensemble learning on sub-size images to improve the efficiency of ensemble learning.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

WH: method, experiment, and writing. HC and WL: experiment. XL and HS: medical knowledge. XH and MG: method. CL: method, writing, and funding. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 82220108007) and Beijing Xisike Clinical Oncology Research Foundation (No. Y-tongshu2021/1n-0379).

Acknowledgments

We would like to thank Miss. Zixian Li and Mr. Guoxian Li for their important discussion in this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2021) 71:209–49. doi: 10.3322/caac.21660

2. Wang FH, Shen L, Li J, Zhou ZW, Liang H, Zhang XT, et al. The Chinese society of clinical oncology (CSCO): clinical guidelines for the diagnosis and treatment of gastric cancer. *Cancer Commun*. (2019) 39:10. doi: 10.1186/s40880-019-0349-9

3. Cheng J, Han Z, Mehra R, Shao W, Cheng M, Feng Q, et al. Computational analysis of pathological images enables a better diagnosis of TFE3 Xp11. 2 translocation renal cell carcinoma. *Nat Commun*. (2020) 11:1778. doi: 10.1038/s41467-020-15671-5

4. Liang J, Yang X, Huang Y, Li H, He S, Hu X, et al. Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. *Med Image Anal*. (2022) 79:102461. doi: 10.1016/j.media.2022.102461

5. Tahiliani HT, Purohit AP, Desai SC, Jarwani PB. Retrospective analysis of histopathological spectrum of premalignant and malignant colorectal lesions. *Cancer Res Stat Treat*. (2021) 4:472–8. doi: 10.4103/crst.crst_87_21

6. Zhao P, Li C, Rahaman MM, Xu H, Yang H, Sun H, et al. A comparative study of deep learning classification methods on a small environmental microorganism image dataset (EMDS-6): from convolutional neural networks to visual transformers. *Front Microbiol*. (2022) 13:792166. doi: 10.3389/fmicb.2022.792166

7. Xue D, Zhou X, Li C, Yao Y, Rahaman MM, Zhang J, et al. An application of transfer learning and ensemble learning techniques for cervical histopathology image classification. *IEEE Access*. (2020) 8:104603–18. doi: 10.1109/ACCESS.2020.2999816

8. Nazarian S, Glover B, Ashrafian H, Darzi A, Teare J. Diagnostic accuracy of artificial intelligence and computer-aided diagnosis for the detection and characterization of colorectal polyps: systematic review and meta-analysis. *J Med Internet Res*. (2021) 23:e27370. doi: 10.2196/27370

9. Schmarje L, Santarossa M, Schroder SM, Koch R. A survey on semi-, self- and unsupervised learning for image classification. *IEEE Access*. (2021) 9:82146–68. doi: 10.1109/ACCESS.2021.3084358

10. Shinde PP, Shah S. A review of machine learning and deep learning applications. In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. Pune: IEEE (2018). p. 1–6. doi: 10.1109/ICCUBEA.2018.8697857

11. Li Y, Wu X, Li C, Li X, Chen H, Sun C, et al. A hierarchical conditional random field-based attention mechanism approach for gastric histopathology image classification. *Appl Intell*. (2022) 1–22. doi: 10.1007/s10489-021-02886-2

12. Hu W, Li C, Li X, Rahaman MM, Ma J, Zhang Y, et al. GasHisSDB: a new gastric histopathology image dataset for computer aided diagnosis of gastric cancer. *Comput Biol Med*. (2022) 142:105207. doi: 10.1016/j.compbiomed.2021.105207

13. Fu B, Zhang M, He J, Cao Y, Guo Y, Wang R. StoHisNet: a hybrid multi-classification model with CNN and transformer for gastric pathology images. *Comput Methods Programs Biomed*. (2022) 221:106924. doi: 10.1016/j.cmpb.2022.106924

14. Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recogn Lett*. (2021) 141:61–7. doi: 10.1016/j.patrec.2020.07.042

15. Ma P, Li C, Rahaman MM, Yao Y, Zhang J, Zou S, et al. A state-of-the-art survey of object detection techniques in microorganism image analysis: from classical methods to deep learning approaches. *Artif Intell Rev*. (2022) 1–72. doi: 10.1007/s10462-022-10209-1

16. Sun C, Li C, Zhang J, Rahaman MM, Ai S, Chen H, et al. Gastric histopathology image segmentation using a hierarchical conditional random field. *Biocybern Biomed Eng*. (2020) 40:1535–55. doi: 10.1016/j.bbe.2020.09.008

17. Zheng X, Wang R, Zhang X, Sun Y, Zhang H, Zhao Z, et al. A deep learning model and human-machine fusion for prediction of EBV-associated gastric cancer from histopathology. *Nat Commun*. (2022) 13:2970. doi: 10.1038/s41467-022-30459-5

18. Dai Y, Gao Y, Liu F. Transmed: transformers advance multi-modal medical image classification. *Diagnostics*. (2021) 11:1384. doi: 10.3390/diagnostics11081384

19. Japanese Gastric Cancer Association. Japanese classification of gastric carcinoma: 3rd English edition. *Gastric Cancer*. (2011) 14:101–12. doi: 10.1007/s10120-011-0041-5

20. Humeau-Heurtier A. Texture feature extraction methods: a survey. *IEEE Access*. (2019) 7:8975–9000. doi: 10.1109/ACCESS.2018.2890743

21. Kulwa F, Li C, Grzegorzek M, Rahaman MM, Shirahama K, Kosov S. Segmentation of weakly visible environmental microorganism images using pair-wise deep learning features. *Biomed Signal Process Control*. (2023) 79:104168. doi: 10.1016/j.bspc.2022.104168

22. Hope TM. Linear regression. In: *Machine Learning*. London: Elsevier (2020). p. 67–81. doi: 10.1016/B978-0-12-815739-8.00004-3

23. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In: *OTM Confederated International Conferences on the Move to Meaningful Internet Systems*." Catania: Springer (2003). p. 986–96. doi: 10.1007/978-3-540-39964-3_62

24. Yang FJ. An implementation of naive bayes classifier. In: *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*. Las Vegas, NV: IEEE (2018). p. 301–6. doi: 10.1109/CSCI46756.2018.00065

25. Shi T, Horvath S. Unsupervised learning with random forest predictors. *J Comput Graph Stat*. (2006) 15:118–38. doi: 10.1198/106186006X94072

26. Suthaharan S. Support vector machine. In: *Machine Learning Models and Algorithms for Big Data Classification*. Boston, MA: Springer (2016). p. 207–35. doi: 10.1007/978-1-4899-7641-3_9

27. Hopfield JJ. Artificial neural networks. *IEEE Circ Dev Mag*. (1988) 4:3–10. doi: 10.1109/101.8118

28. Zhang J, Li C, Kosov S, Grzegorzek M, Shirahama K, Jiang T, et al. LCU-Net: a novel low-cost U-Net for environmental microorganism image segmentation. *Patt Recogn*. (2021) 115:107885. doi: 10.1016/j.patcog.2021.107885

29. Zhang J, Ma P, Jiang T, Zhao X, Tan W, Zhang J, et al. SEM-RCNN: a squeeze-and-excitation-based mask region convolutional neural network for multi-class environmental microorganism detection. *Appl Sci*. (2022) 12:9902. doi: 10.3390/app12199902

30. Chen H, Li C, Wang G, Li X, Rahaman MM, Sun H, et al. GasHis-Transformer: a multi-scale visual transformer approach for gastric histopathological image detection. *Patt Recogn*. (2022) 130:108827. doi: 10.1016/j.patcog.2022.108827

31. Yang H, Zhao X, Jiang T, Zhang J, Zhao P, Chen A, et al. Comparative study for patch-level and pixel-level segmentation of deep learning methods on transparent images of environmental microorganisms: from convolutional neural networks to visual transformers. *Appl Sci*. (2022) 12:9321. doi: 10.3390/app12189321

32. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv Preprint*. (2014) arXiv:14091556.

33. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV (2016). p. 2818–26. doi: 10.1109/CVPR.2016.308

34. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV (2016). p. 770–8. doi: 10.1109/CVPR.2016.90

35. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint*. (2020) arXiv:201011929.

36. Liu W, Li C, Rahaman MM, Jiang T, Sun H, Wu X, et al. Is the aspect ratio of cells important in deep learning? A robust comparison of deep learning methods for multi-scale cytopathology cell image classification: from convolutional neural networks to visual transformers. *Comput Biol Med*. (2022) 141:105026. doi: 10.1016/j.compbiomed.2021.105026

37. Liu W, Li C, Xu N, Jiang T, Rahaman MM, Sun H, et al. CVM-Cervix: a hybrid cervical pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. *Patt Recogn*. (2022) 130:108829. doi: 10.1016/j.patcog.2022.108829

38. Zhou X, Li C, Rahaman MM, Yao Y, Ai S, Sun C, et al. A comprehensive review for breast histopathology image analysis using classical and deep neural networks. *IEEE Access*. (2020) 8:90931–56. doi: 10.1109/ACCESS.2020.2993788

39. Chen A, Li C, Zou S, Rahaman MM, Yao Y, Chen H, et al. SVIA dataset: a new dataset of microscopic videos and images for computer-aided sperm analysis. *Biocybern Biomed Eng*. (2022) 42:204–14. doi: 10.1016/j.bbe.2021.12.010

40. Shi Z, Zhu C, Zhang Y, Wang Y, Hou W, Li X, et al. Deep learning for automatic diagnosis of gastric dysplasia using whole-slide histopathology images in endoscopic specimens. *Gastric Cancer*. (2022) 25:751–60. doi: 10.1007/s10120-022-01294-w

41. Tsuneki M, Ichihara S, Kanavati F. Weakly supervised learning for poorly differentiated adenocarcinoma classification in gastric endoscopic submucosal dissection whole slide images. *medRxiv*. (2022). p. 1–15. doi: 10.1101/2022.05.28.22275729

42. Zhang J, Zhao X, Jiang T, Rahaman MM, Yao Y, Lin YH, et al. An application of pixel interval down-sampling (PID) for dense tiny microorganism counting on environmental microorganism images. *Appl Sci*. (2022) 12:7314. doi: 10.3390/app12147314

43. Li X, Li C, Rahaman MM, Sun H, Li X, Wu J, et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Art Intell Rev*. (2022) 55:4809–78. doi: 10.1007/s10462-021-10 121-0

44. Rahaman MM, Li C, Yao Y, Kulwa F, Rahman MA, Wang Q, et al. Identification of COVID-19 samples from chest X-Ray images using deep learning: a comparison of transfer learning approaches. *J X-ray Sci Technol*. (2020) 28:821–39. doi: 10.3233/XST-200715

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 06 January 2023
DOI 10.3389/fmed.2022.1029227

Predicting melanoma survival and metastasis with interpretable histopathological features and machine learning models

Justin Couetil[1†], Ziyu Liu[2†], Kun Huang[3], Jie Zhang[1*] and Ahmed K. Alomari[4*]

[1]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, United States, [2]Department of Statistics, Purdue University, West Lafayette, IN, United States, [3]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States, [4]Department of Pathology, Indiana University School of Medicine, Indianapolis, IN, United States

Check for updates

OPEN ACCESS

EDITED BY
Xin Qi,
Eisai, United States

REVIEWED BY
André Pedersen,
Norwegian University of Science and Technology, Norway
Cristian Scatena,
University of Pisa, Italy

*CORRESPONDENCE
Jie Zhang
✉ jizhan@iu.edu
Ahmed K. Alomari
✉ akalomar@iu.edu

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Pathology, a section of the journal Frontiers in Medicine

RECEIVED 26 August 2022
ACCEPTED 16 December 2022
PUBLISHED 06 January 2023

CITATION
Couetil J, Liu Z, Huang K, Zhang J and Alomari AK (2023) Predicting melanoma survival and metastasis with interpretable histopathological features and machine learning models.
Front. Med. 9:1029227.
doi: 10.3389/fmed.2022.1029227

Introduction: Melanoma is the fifth most common cancer in US, and the incidence is increasing 1.4% annually. The overall survival rate for early-stage disease is 99.4%. However, melanoma can recur years later (in the same region of the body or as distant metastasis), and results in a dramatically lower survival rate. Currently there is no reliable method to predict tumor recurrence and metastasis on early primary tumor histological images.

Methods: To identify rapid, accurate, and cost-effective predictors of metastasis and survival, in this work, we applied various interpretable machine learning approaches to analyze melanoma histopathological H&E images. The result is a set of image features that can help clinicians identify high-risk-of-metastasis patients for increased clinical follow-up and precision treatment. We use simple models (i.e., logarithmic classification and KNN) and "human-interpretable" measures of cell morphology and tissue architecture (e.g., cell size, staining intensity, and cell density) to predict the melanoma survival on public and local Stage I–III cohorts as well as the metastasis risk on a local cohort.

Results: We use penalized survival regression to limit features available to downstream classifiers and investigate the utility of convolutional neural networks in isolating tumor regions to focus morphology extraction on only the tumor region. This approach allows us to predict survival and metastasis with a maximum F1 score of 0.72 and 0.73, respectively, and to visualize several high-risk cell morphologies.

Discussion: This lays the foundation for future work, which will focus on using our interpretable pipeline to predict metastasis in Stage I & II melanoma.

KEYWORDS

computational pathology, histopathology, biomedical image processing, melanoma, neoplasm metastasis, survival prognosis, metastatic prognosis

1. Introduction

Melanoma is the fifth most common cancer, with about 110,000 new cases in the US alone in 2021, and its incidence is increasing approximately 1.4% each year. Most melanoma patients are considered cured when their superficial, thin, primary melanoma (Stage I) is surgically removed, resulting in a 99.4% 5-years survival rate (1). However, melanoma can recur as a locoregional disease or distant metastases in 6% of Stage I patients where cancer is limited to the superficial dermis, and in 20% of Stage II patients where cancer has invaded the deeper dermis and subcutis (2). Currently, clinicians still do not have accurate and cost-effective ways to predict tumor recurrence and metastasis from the primary tumor histopathological images of early-stage patients. In addition, current melanoma staging systems depend primarily on histopathologic features, and sometimes involve invasive sentinel lymph node biopsies. These procedures have not been shown to improve prognosis for early-stage patients (tumor invasion $<$ 1 mm) and therefore expose patients to unnecessary morbidity (3).

Currently, prognostication of localized melanoma (i.e., no distant metastases) relies on several histopathological criteria established by pathologists' examination of hematoxylin and eosin (H&E) stained tissue sections. These include Breslow depth and the presence of ulceration and microsatellitosis. Moreover, it also depends on the identification of tumor deposits in sentinel lymph nodes in cases where such procedure is performed. Tumor ulceration is the loss of full-thickness epithelium above the growing tumor and is an independent prognostic factor. Integrating these histopathologic findings with clinical information like the site of origin for tumors is important: acral (non-sun-exposed regions) and lentigo maligna melanomas both could have fusiform cells, but the prognoses are different, with thickness-matched acral melanoma being more aggressive (4). Moreover, prognosis varies by histologic subtype, where nodular and acral have generally worse outcomes than thickness-matched superficial spreading and desmoplastic subtypes (5, 6). However, personalized prognostication of early-stage melanoma ($<$ 0.75 mm) remains suboptimal. Ulceration, the hallmark of poor prognostic feature, is not a common finding for early-stage melanoma, and though most literature suggests that lymphocyte infiltration is an important marker for better prognoses, this relationship is uncertain for lesions under 0.75 mm depth of invasion (7).

Because histopathologic features remain suboptimal in predicting melanoma prognosis in early-stage patients, and early-stage patients make up about 80% of all newly diagnosed cases of melanoma (8), there is an pressing need for developing a machine learning based computational pathology pipeline to stratify patients. Rigorous measurement of cellular/nuclear morphological features of primary tumor pathological images may provide consistent performance across the heterogenous landscape of melanoma. Currently, published machine learning models using H&E images to study melanoma prognosis are mostly "black-box" models based on deep neural networks, specifically Convolutional Neural Networks (CNNs) (9). For instance, Forchhammer et al. applied CNNs trained on whole slide images to establish a model that stratified patients by their 10-years survival rates; however, improving risk classification beyond the existing staging guidelines has proven difficult for early-stage patients (10). CNN-based approaches have also been used to predict survival using locoregional/metastatic biopsies (11), which applies to less than 20% of all melanoma patients (12). Furthermore, these deep learning-based models identified abstract features that are neither visible nor directly associated with human-interpretable cell morphology and tissue structure, which is a major barrier for clinical adoption and generation of new hypotheses for research. It is imperative that pathologists and researchers understand the mechanisms behind the disease progression. In this paper, we present a pipeline with more interpretable machine learning methods that can be used alongside the very accurate, but less interpretable deep learning techniques.

Kulkarni and Robinson (13) published the only histopathology-based melanoma metastasis model to date. They achieved impressive accuracy (88–90%) for high/low risk stratification based on their deep learning models. Due to the lack of interpretability of the neural network, ablation studies were adopted to show that the ratio of lymphocyte area over tumor cell area was crucial for model accuracy. The individual contributions of the rest of the morphology feature set were not readily apparent. When it is difficult to understand what information the neural networks rely on to make their prediction, it is more difficult for pathologists, clinicians, and researchers to investigate further. This is a bottleneck for effective translational application of these neural networks. In addition, this work provided very accurate classification for patients with more advanced disease (skewed toward Stages II–III). However, Stage I patients comprise the majority of the general melanoma population, and metastasis is most likely to be missed in these individuals.

The work we present herein focuses on developing a machine learning pipeline to identify the reliable and interpretable H&E histopathology image features to predict 5-years survival and metastasis using the primary site biopsies from Stages I, II, and III melanoma patients. We have demonstrated that simple machine learning models (i.e., logistic regression, k-Nearest Neighbors, support vector machines, and random forest classifiers) using extracted interpretable features of cellular and nuclear morphology can generated accurate, sensitive, and specific prediction for 5-years survival and metastasis risks. We first applied deep learning methods (14) to identify tumor regions with CNN models, and extracted interpretable morphological features from only the tumor regions, and understand how this impacts downstream classifier performance. Because some of the morphological descriptors we used can be correlated to each other, we applied LASSO Cox regression to reduce the number of image features available to downstream classifiers to reduce the likelihood of overfitting and improve ease of interpretation by a pathologist.

One of the challenges for our study is that samples from large cancer databases, such as The Cancer Genome Atlas (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC), provide almost only survival information, without any metastasis information for Stages I and II patients. To tackle this challenge, we approached our work in two steps: (1) Training machine learning models for survival prediction with a merged cohort form the TCGA, CPTAC, and our own curated, high-quality local IU School of Medicine (IUSM) cohort; and (2) further refining it to predict melanoma metastasis on the IUSM cohort.

We demonstrated that our identified H&E image features can serve as accurate, rapid, and low-cost predictors of metastasis. Further, this approach can be seamlessly integrated into clinical workflows, given that digitized biopsies are an approved diagnostic tool (15) and interpretation of biopsies by a pathologist is standard of care in melanoma. To summarize, our work begins to bridge a significant clinical and research gap: the need for an adoptable and interpretable cell morphology machine learning pipeline to work alongside deep-learning approaches in the study of melanoma metastasis.

2. Data and materials and methods

2.1. Data description

To maximize sample size and test the model generalizability, we applied our pipeline to three melanoma cohorts: The Cancer Genome Atlas (TCGA) cohort, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) cohort (16), and the Indiana University School of Medicine (IUSM) cohort. For all three, we restrict our analyses to Stages I, II, and III melanoma patients. Stage IV patients, who already have distant metastases, were excluded. Further, slides that are misdiagnoses, microsatellite metastases, and those that cannot be confirmed as primary site biopsies due to lack of visible intact epidermis were removed. This rigorous quality control resulted in a sample size of 81 whole slide images (WSI) from 71 patients in the TCGA cohort, and 45 WSI from 19 patients in the CPTAC cohort. The TCGA and CPTAC cohorts only contain survival information, with no metastasis information. The IUSM cohort had 92 WSIs from 70 patients with both metastasis and survival information. This information is summarized in Table 1.

2.2. Feature extraction pipeline

Predicting metastasis and survival are two distinct but related tasks. Herein, we use the same feature extraction pipeline to predict 5-years metastasis for IUSM patients, and 5-years survival in the IUSM, TCGA, and CPTAC datasets (Figure 1). We modified the morphological feature set described in (17) by focusing on the morphological features and introducing two additional features (quantity and density) to describe lymphocytes and other small, hyperchromatic cells (e.g., pyknotic nuclei), each with 10 bins and five distribution statistics. We call this category *Small-Hyperchromatic cells*. In total, we have 135 morphological features extracted from WSIs to quantify the cell size and shape, as well as *Small-Hyperchromatic cell* density and counts, as well as statistics describing the distribution for each of these image features within each WSI (i.e., mean, standard deviation, skewness, kurtosis, and entropy).

2.2.1. H&E whole slide image pre-processing

Before image normalization and feature extraction, we first rescaled WSIs from different cohorts to the same resolution. The TCGA and the CPTAC images are scanned using different resolutions, so all image patches were resized to match the IUSM cohort resolution of 0.25 microns-per-pixel (mpp), which corresponds to a 400x magnification. For the TCGA and CPTAC cohorts, we tiled images into squares with dimension of 512px * 0.25/mpp, and then rescaled them into 512 × 512 pixels. To filter out black and white background patches from the WSI,

TABLE 1 Clinical follow up and slide quality among Stages I–III patients in three patient cohorts.

	Cohort; # patients (# slides)			
	Survival			Metastasis
	TCGA	CPTAC	IUSM	IUSM
Stages I, II, and III patients	129 (145)	63 (117)	70 (92)	70 (92)
Adequate quality slides	71 (81)	19 (45)	70 (92)	70 (92)
Follow up information (patients)				
Low risk–no event before 5 years	3	0	43	30
High risk–event before 5 years	7	8	7	26
Non-informative censoring before 5 years	61	11	20	14

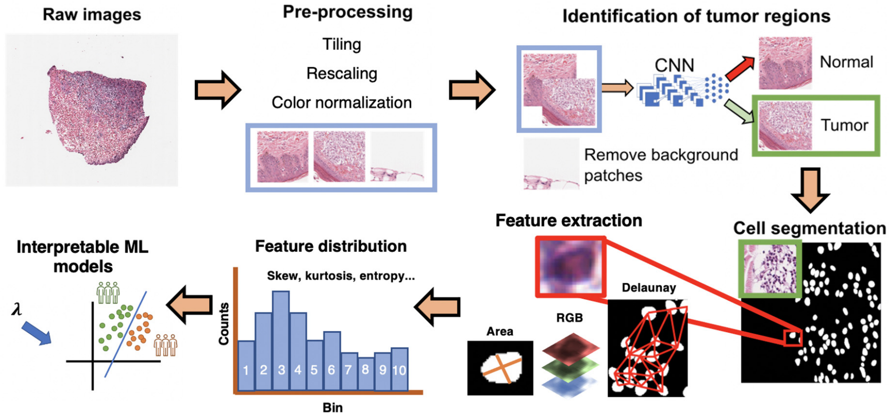


FIGURE 1
Interpretable cell morphology and machine learning pipeline. Whole slide images are tiled into 512 × 512 pixel patches and then rescaled for differences in resolution between datasets, background patches are removed by manually engineered color features, and tumor/normal regions are identified using a convolutional neural network that was trained with pathologist annotations. Tiles then normalized for hematoxylin and eosin (H&E) staining, and cells segmented with StarDist. Cell masks are analyzed with MATLAB to generate measures of cell morphology (e.g., area), tissue architecture (i.e., Delaunay distance), and identify small, hyperchromatic cells. 50 processed slides are randomly pooled, and k-means clustering ($k = 10$) is used to bin the distribution of each image feature and five distribution statistics are calculated (e.g., skew, kurtosis, mean). Finally, every cell of every slide is assigned the "bins" to generate a final dataset of N patients × 135 features.

we removed patches where the mean intensity of RGB channels together was greater than or equal to 230, less or equal to 40, or with a standard deviation less than 20. This removed both black and white background patches. IUSM images required further processing, using the following criteria to remove black and white patches: red channel mean intensity being 90% or less than the blue channel mean intensity, red channel mean intensity being less than 170, and green channel mean intensity being greater than 210. The results of these criteria were visually inspected for accuracy and consistency among the three datasets. We then applied the color normalization algorithm proposed by Macenko, Neithammer (18) to avoid batch-effects both within and across datasets. This algorithm is unsupervised and based on singular value decomposition of opacity density values. Cells in the WSI from all three cohorts are segmented using StarDist, which uses all three RGB color channels to segment cells (19). StarDist was better suited to this task than hierarchical multilevel thresholding based on our evaluation (**Supplementary Figure 1**). We further removed background patches by filtering out those that had fewer than certain number of cells segmented by StarDist. For the CPTAC and IUSM slides, the cutoff is 15 cells, and for the TCGA, the cutoff is 10. The results of this preprocessing were visually verified by our pathologist.

2.2.2. Identification of tumor regions with convolutional neural networks

Tumor biopsies contain variable amounts of normal tissue, therefore using the entire WSI to predict clinical outcomes may introduce additional bias. To study this, we focused our analysis on regions of the bulk tumor by using a CNN to triage tumor vs. normal image patches, and then applied our interpretable feature extraction pipeline to the CNN-identified tumor patches. The quality of specimen from the TCGA, CTPAC, and IUSM cohorts were very different. The IUSM cohort had the best quality, followed by TCGA and then CPTAC. We trained three different CNN's–one for each cohort–to understand how varying data quality impacted tumor vs. normal image patch classification on the entire (**Figure 2A**), we describe training process and model architecture.

To train and validate these CNNs, our pathologist used QuPath (20) to manually annotate normal tissue and tumor regions on WSIs from all three cohorts. Respectively, 20, 19, and 90 WSIs were annotated for the TCGA, CPTAC, and IUSM cohorts. We cropped image tiles from all three cohorts, resized if necessary to match resolutions (described previously), and assigned tumor/normal labels to image patches using our pathologist's QuPath annotations of tissue regions. Finally, we performed five-fold cross-validation using 1,000 tumor and 1,000 normal image tiles for training and validation (80/20% split). To ensure fair assessment of accuracy, we designed random sampling such that validation image tiles were not pulled from patients that appeared in the training set. It is important to note that some patients had multiple WSIs; therefore, naïve random sampling would mean that WSIs from a single patient could end up in both the testing and validation datasets. To prevent this, we ensured that the testing and validations splits were based on patients, not WSIs.

One of the most widely used CNN architectures for image classification and object detection is the "Inception" module, which employs convolutional kernels with different sizes, called scale filters (21). With our simple task and ample training data,

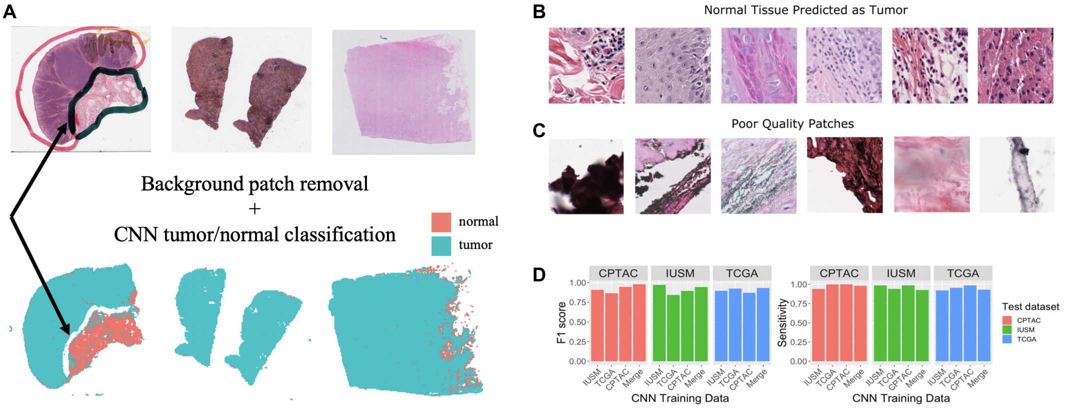


FIGURE 2
Training and assessment of convolutional neural network. Workflow and assessment of convolutional neural network. **(A)** Whole slide image patches are first filtered to remove background patches. This step removes artifacts, demonstrated by the ink from the whole slide images (WSI) that is not retained in the final set of patches (arrow). **(B)** Patches confused by the convolutional neural networks (CNN) can show mitotic figures and dense lymphocytic infiltrate. **(C)** Despite our filtering steps, there remain some artifact-ridden poor-quality patches. **(D)** Performance at epoch 200 for CNNs trained on the IU School of Medicine (IUSM), The Cancer Genome Atlas (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC), and merged datasets, comparing the Fl scores and sensitivity for tumor/normal classification achieved on validation sets from the IUSM, TCGA, and CPTAC cohorts.

we adopted the naïve version of "Inception" based model, herein called GoogLeNet, and modified the final layer of the network for binary tumor/normal tissue classification. We applied the ADAM (22) optimizer with learning rate 0.0002. Four different GoogLeNet models were trained using the CPTAC, IUSM, TCGA cohorts, and a final "Merged" cohort with 1,000 labeled patches from all three cohorts. Experimental results demonstrated that the GoogLeNet models achieve reasonable accuracy on all four validation datasets (> 0.8).

2.2.3. Cell-level feature extraction, aggregation, and investigation

We have previously developed a morphological feature extraction pipeline for H&E images, as described by Cheng, Zhang (17), and adopt it here to predict melanoma outcomes. For each WSI, we first used the *regionprops* function in MATLAB version R2022a to calculate area, major and minor axis length of the cells, major/minor axis ratio, staining intensities (RGB three color channels), and described the density of cells in the image by measuring the minimum, maximum, and mean distance to neighboring cells. Staining intensities were not use as image features, but were used to engineering a new cell category, *Small-Hyperchromatic cells*. This process is described later. The neighbor relationship was defined using the Delaunay triangulation method among cell centroids.

Once all WSIs were processed, we randomly sampled 50 images from the three datasets (IUSM, CPTAC, and TCGA), to perform k-mean clustering with 10 clusters for every image feature. This is analogous to generating 10 *"bins"* for a histogram. These 10 bins for each category of features represent a dataset-wide census of cell morphology and maintain representations of heterogeneity that simple statistics such as average cannot. Five statistics on the distribution of these histograms were calculated: mean, standard deviation, skewness, kurtosis, and entropy. This gave us a summarized data structure of 7 features × (10 bins + 5 statistics) = 105 cell features per patient. If a patient had multiple WSIs, we calculated the mean of all feature values across the WSIs, providing a single patient-level vector.

In addition to these 105 features, we engineered cutoffs to identify a specific category of small cells with hyperchromatic staining, which we refer to as *Small-Hyperchromatic cells*. This category tends to represent necrosis and dense inflammation by highlighting lymphocytes, and pyknotic nuclei. *Small-Hyperchromatic cells* were defined by an area less than 450 pixels and a ratio of the long and short cell axes less than 2 (favoring round cells rather than spindle cells). We calculated the quantity and density (proportion of these cells to all cells in each image patch) of these small-hyperchromatic cells for an additional 30 features: 10 bins and 5 statistics for the quantity and density of small hyperchromatic cells each. Adding this to the previously described 105 features provided a total of 135 features per WSI.

2.3. Univariate feature analysis

To investigate the ability of individual image features to stratify patients, we took the approach described by Lu, Xu (23), iterating through 100 cutoffs in the range of values for each image feature, which generates two strata for which to calculate Kaplan-Meier estimates and extract FDR-adjusted *p*-values. The resulting significant cutoffs are used to generate Kaplan-Meier curves for survival and metastasis. This analysis conducted in R v4.1.0 using packages survival v3.2-13 (24) and survminer v.0.4.9 (25).

2.4. Multivariate supervised classification for risk stratification

Despite merging data from three different patient cohorts, our sample size is still limited due to stringent quality control and a focus on Stages I through III. In the setting of a high feature dimension and small sample size, we use Lasso Cox regression from the glmnet v4.1-1 package in R (26) to reduce collinearity in the downstream supervised classification task (**Supplementary Figure 2**). The trained Lasso Cox model provides two feature sets based on the accuracy metric of concordance: "1se" feature set for a model whose variance is heavily regularized, where cross-validation accuracy is within one standard error of the maximum accuracy; and the "min" feature set corresponds to the model with the absolute highest cross-fold validation accuracy, which therefore usually provides more features than "1se."

To ensure model robustness and maximize the size of our training set, we used a modified five-fold cross-validation. We randomly shuffled and split all samples from all three cohorts into five equal-sized groups. For survival, we labeled "high risk" patients as those who suffered death/metastasis within 5 years, and "low risk" patients as those who had at least 5 years of uneventful follow-up. For metastasis, "high risk" patients were defined as those who suffered a metastasis at any time point, and "low risk" patients were metastasis-free for 5 years (for censored data) or beyond. Patients who were lost to follow-up (censored) before 5 years were not labeled as either high risk or low risk and removed for prognostic model training.

Traditionally in five-fold cross-validation, the feature weights (i.e., coefficients) generated by separate models are aggregated to provide a final model. This process is called "bagging." Here, instead, we used four of five groups to train a model which was then used to predict risk labels on the fifth hold-out group as validation. By repeating the process five times, all patients were used to train and test models, but there was no overlapping of patient data during each training and testing. Patients were resampled so that there was an equal proportion of high and low risk labels, with the total number of labels for each class being equal to the larger class prior to resampling. The performance metrics (F1 score sensitivity and specificity) were calculated by concatenating the results for the test set of each fold into a single matrix. We implemented this cross-validation process for random forests (RF), support-vector machines (SVM), k-Nearest Neighbors (KNN), and logarithmic classification. In certain instances, two models yielded similar accuracy, but they do not have the same level of interpretability; for example, logarithmic classification is more interpretable than SVM and KNN. The coefficients from logistic regression for each CNN-derived dataset and LASSO-derived feature set are visualized to investigate whether image features receive consistent coefficients in multivariate survival stratification task (**Supplementary Figures 4**, **5**). This is further discussed in the Results section.

2.5. Image feature visualization for interpretability

To visualize the features used for risk stratification, we generated "heatmaps" using the *ggplot2* (27) and *ggnewscale* (28). The heatmaps are placed side-by-side with the Hematoxylin and Eosin-stained WSI for inspection and direct interpretation by the pathologist.

2.6. Ethics statements

This study involves human subjects. The TCGA and CPTAC consortia provide their data to the public, and the data (follow up and histopathological images) is not linked to PHI. For the IUSM cohort of patients, secondary use of identifiable information and biospecimen is covered under our own broad institutional IRB.

3. Results

3.1. Assessment of GoogLeNet performance on tumor region identification

As described in the Methods section of this manuscript, four GoogLeNet CNNs were trained to recognize tumor tissue image patches. These four models were created using the CPTAC, TCGA, IUSM, as well as a balanced random selection of image patches from all three cohorts ("Merge"). All models perform well as indicated by the consistent high sensitivity and F1 scores across datasets (**Figure 2D**). For the few misclassified patches, visual inspection by our pathologist revealed that normal tissue predicted as tumor tended to contain mitotic figures, dense inflammation, and poor-quality image patches that remained despite our filtering process (**Figures 2B, C**).

3.2. Results of univariate analysis

3.2.1. Univariate Kaplan-Meier survival analysis

Using the univariate Kaplan-Meier log-rank test, we identified several features that can significantly stratify patients

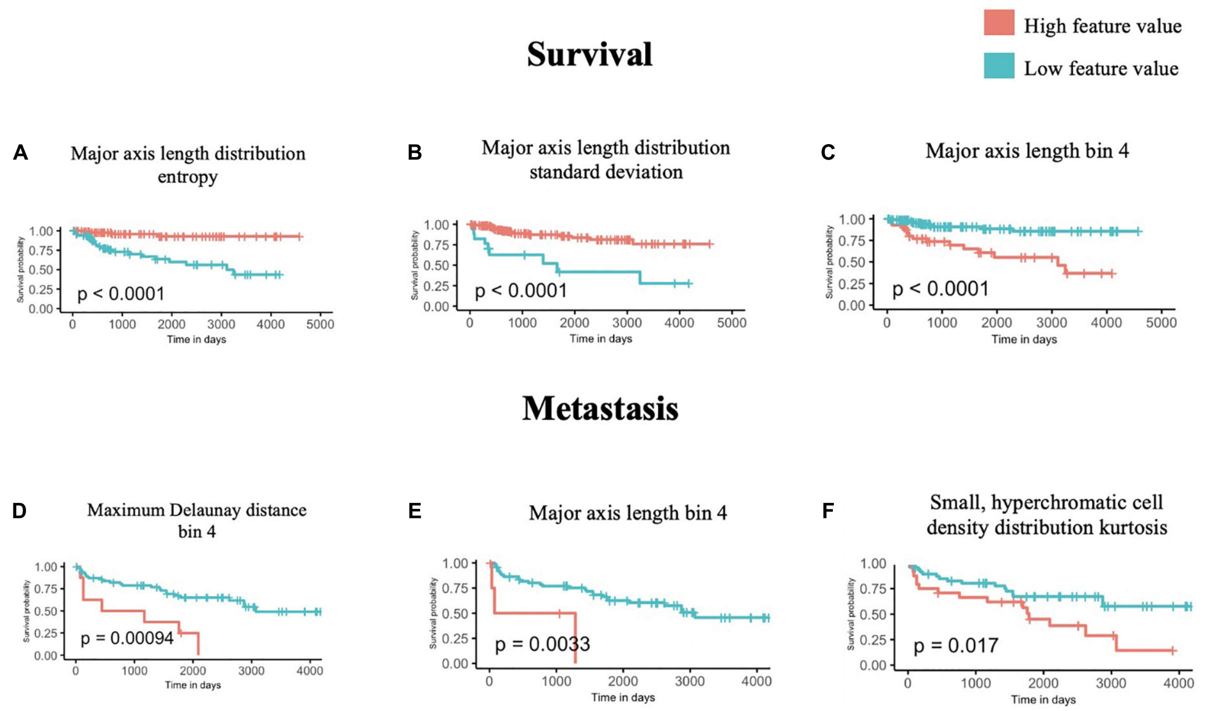


FIGURE 3
Univariate survival and metastasis analysis. After scanning through 100 cutoffs through the range in values for each feature, the cutoff providing the most significant *p*-values from log-rank tests of survival **(A–C)** and metastasis **(D–F)** times are used to generate Kaplan-Meier curves.

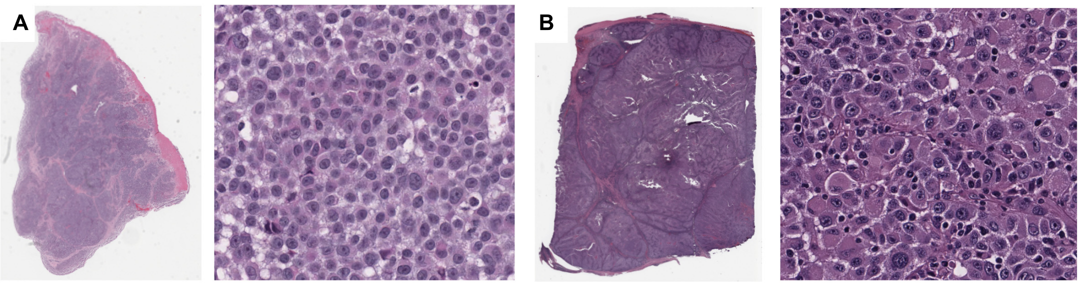


FIGURE 4
Feature visualization, *Maximum Delaunay distance bin 4*. At high power, **(A)** specimen TCGA-FR-A20 S, melanoma cells of dense/intermediate packing. **(B)** Specimen TCGA-ER-A19S, melanoma cells of very dense proliferation, going through stages of necrosis, and immune infiltration.

on their survival outcomes. The three most statistically significant ones are shown in Figures 3A–C, which are *Major axis length distribution entropy, Major axis length distribution standard deviation*, and *Major axis length bin 4*.

We further interpreted each of the identified features: As shown in Figure 3A, *Major axis length distribution entropy* significantly stratifies patient survival (log-rank $p < 0.0001$). Entropy is a measure of distribution uniformity, where high entropy represents a large variation in cell sizes, therefore, distributions with high entropy tend to have a high standard deviation. In Figure 3A, high entropy of the *Major axis length* distribution correlates with a better prognosis. This aligns with Figure 3B, which shows that a high standard deviation in the *Major axis length* also correlates with a good prognosis. Together, both features suggest that a high heterogeneity in cell sizes in a histopathological specimen (inflammatory, tumor, stromal, and otherwise) portend a better prognosis.

Major axis length bin 4 appears as a significant feature for both survival and metastasis (Figures 3C–E), with the same direction of effect, where a high proportion of this feature contends poor prognosis. The interpretation of this feature is summarized in Section "3.4 Morphological

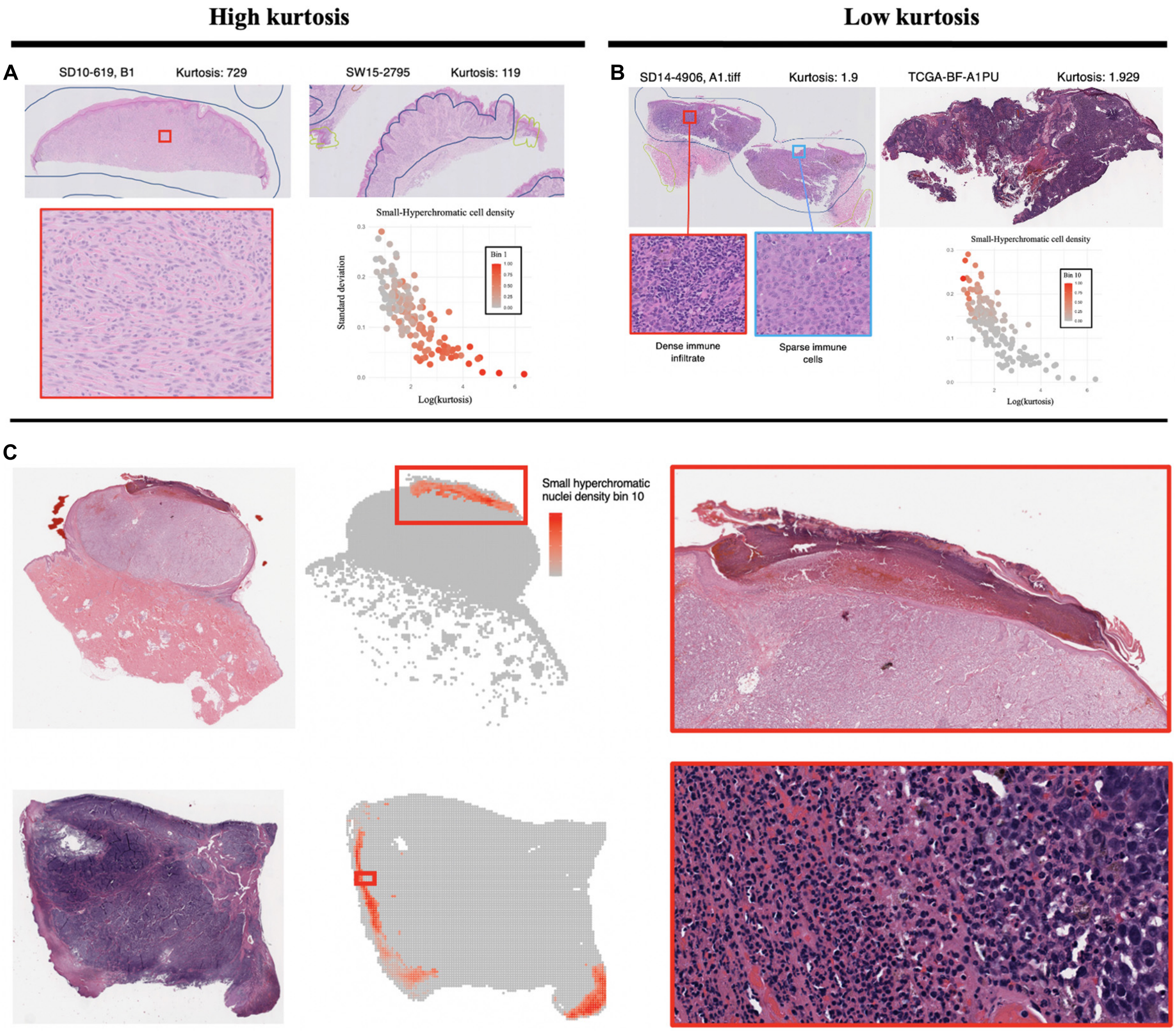


FIGURE 5
Visualization of small, hyperchromatic nuclei. **(A)** Demonstrates high kurtosis of Small-Hyperchromatic cells. These slides have uniformly low densities of immune infiltration. **(B)** Low kurtosis of Small-Hyperchromatic cells shows slides with a high variability of immune infiltration and necrosis across the entire specimen. Kurtosis of this feature is inversely correlated with standard deviation, and slides with high kurtosis have lower densities of Small-Hyperchromatic cells. In **(C)**, binl0 represents the highest density of Small-Hyperchromatic cells. Like the slides with low kurtosis in panel **(B)**, these high-density regions harbor necrosis of tissue with dense lymphocytic and neutrophilic infiltrate. Slides with low kurtosis have a higher standard deviation and higher density of Small-Hyperchromatic cells. The convolutional neural networks (CNN) used to filter out background patches in this visualization was trained on the clinical proteomic tumor analysis consortium (CPTAC) dataset.

features associated with 5-years survival/metastasis prediction."

3.2.2. Univariate Kaplan-Meier metastasis analysis

Using the same approach as above, we identified four features that are significantly associated with the prediction of 5-years metastasis (Figures 3D–F): *Maximum Delaunay distance bin 4*, *Major axis length bin 4*, and *Small-Hyperchromatic cell density distribution kurtosis*. *Maximum Delaunay distance bin 4* represents cells of intermediate packing density. High values of this feature were associated with a higher likelihood of metastasis in the univariate analysis. This feature is correlated with the *Minimum Delaunay distance bin 4* and *Mean Delaunay distance bin 4*, both of which are high risk for survival prediction: Spearman correlation coefficient (SCC) with *Maximum Delaunay distance bin 4* across all three datasets 0.55 and 0.899, respectively. We found that this density, defined by nuclei centroids, was seen in very different histomorphologies: In Figure 4, we show two regions with the similar density, but one is composed of small cells with intermediate packing Figure 4A, and the other is composed of distended rhabdoid cells (nuclei pushed to side of cell by cytoplasm) in a setting of very dense proliferation and immune infiltration Figure 4B.

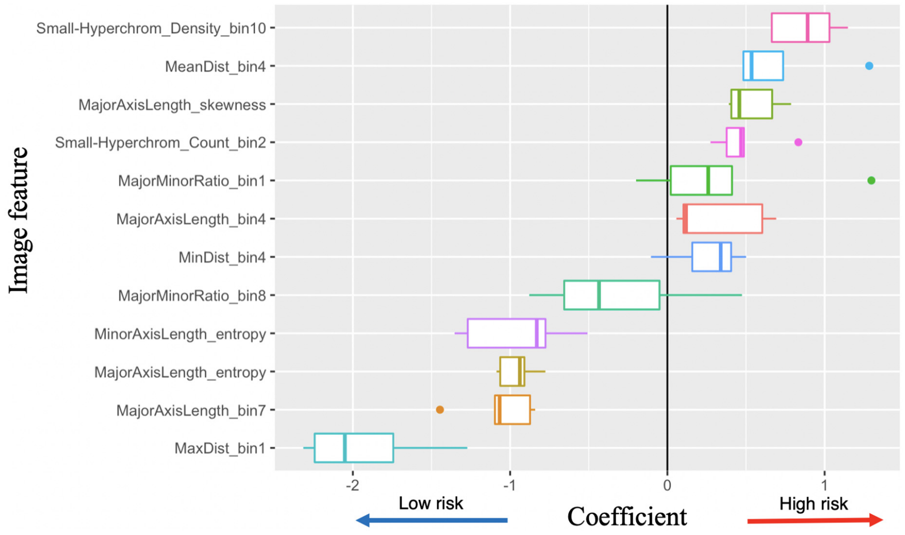


FIGURE 6
Image feature coefficients in 5-years survival logarithmic classifier based on Merge convolutional neural networks (CNN) and 1se LASSO feature set. Model weights for five logarithmic classification sub-models trained in the five-fold cross validation.

The *Small-Hyperchromatic cell distribution kurtosis* is a significant predictor in the Kaplan-Meier univariate analysis of metastasis (log-rank $P = 0.017$). Kurtosis measures the tailedness of a distribution. Further examination of slides with high kurtosis reveal specimen with a low density of inflammatory cell infiltration (mainly lymphocytes in this setting, Panel 5A). There is also a noted negative correlation between *Small-Hyperchromatic cell density kurtosis* and *standard deviation,* and we identify that slides with a high density of *Small-Hyperchromatic cells* tend to have distributions with low kurtosis and high standard deviation (**Figures 5A, B**). Aligning with kurtosis being a high-risk feature, *Small-Hyperchromatic cell distribution standard deviation* is a low-risk image feature in the multivariate survival models (**Figure 6**). Weak infiltration of tumors by lymphocytes is a well-established independent poor prognostic factor that pathologists assess (29), and for this reason, an image analysis pipeline for accurate quantification of tumor infiltrating lymphocytes has been studied (30).

To better understand what this *Small-Hyperchromatic cell* feature represents, we visualized the upper extreme of the density distribution (bin 10), demonstrating areas of necrosis and dense immune infiltration and ulceration on the peripheries of a nodular melanoma (**Figure 5C**). Ulceration occurs when tumors outgrow their blood supply and is an accepted marker for aggressive tumor biology and used for staging (31). As for the slides with low kurtosis (**Figure 5B**), they are associated with a high variability in the density of immune infiltration: In the same histologic specimen, there are regions with dense immune infiltration and necrosis, and other regions with sparse immune infiltration.

3.3. Multivariate risk stratification for 5-years survival and metastasis

As a baseline, stratifying patients based on their AJCC stage provided poor predictive values, with the F1 scores for survival and metastases being 0.44 and 0.51, respectively (**Table 2**). We experimented with several classification models using our image features, and several provided reasonable accuracy (**Supplementary Table 1**, **Supplementary Figure 3**). For 5-years survival prediction, the logistic classifier using the CNN trained on merged cohort for tumor region and Lasso-min feature set provided an F1 score of 0.72. For metastasis prediction, the KNN using the entire WSI and Lasso-1se feature set generated an F1 score of 0.73, while a comparable F1 score of 0.72 was achieved for the RF trained using the CNN trained on the IUSM cohort and Lasso-1se feature set (**Table 3**, **Supplementary Table 1**).

3.4. Morphological features associated with 5-years survival/metastasis prediction

3.4.1. Image features associated with 5-years survival

With the successful predictions on 5-years survival and metastasis, we further examined the image features. Logistic regression has the best interpretability, because the coefficients learned for each image feature can be visualized (**Figure 6**). There are several features with very high or low risks. The highest risk morphology for the Merge CNN-derived logistic regression survival model using the 1se feature set was the *Small-Hyperchromatic cell density bin 10,* while the lowest risk phenotypes were the *Maximum Delaunay distance bin 1, Major axis length bin 7, and Major axis length distribution entropy*.

Because both WSI and CNN-derived tumor-region-only survival models achieved high accuracy, we visualized and examined the coefficients assigned to all image features among all survival models, with most features show consistent direction of effect for the decreased or increased risk hazard (**Supplementary Figure 4**). We also visualized coefficient weights stratified by tumor-region-only (CNN) vs. WSI-derived models to check whether any features were weighted oppositely if background stroma was included. We did not find this to be the case, and tumor-region-only and WSI coefficients were consistent in direction of hazard coefficients (**Supplementary Figure 5**).

3.4.2. Image features associated with metastasis

The peak performance for metastasis is achieved by the KNN classifier using the entire the WSI and 1se LASSO feature set. This feature set contains: *Major axis length bin 4, Major axis length bin 7, Major minor ratio bin 1, Mean Delaunay distance bin 4, Max Delaunay distance bin 1, Small-Hyperchromatic cell count bin 2, Major axis length distribution skewness, and Major axis length distribution entropy.*

3.4.3. Visualization and interpretation of identified image features

Major axis length bin 4, mentioned previously, is a significant image feature for the univariate analysis of both survival and metastasis, with the same direction of effect: large values of this feature were associated with poor prognosis. This feature was maximized in specimen with small to intermediate sized melanocytes (**Figure 7**). Small cell melanoma has been associated with a poor prognosis previously in case series and case reports (32, 33), and our feature of *Major axis length bin 4* is consistent with this finding. Small cell melanoma, however is exceedingly rare, and though some IHC staining patterns of this variant have been described in patients with metastatic disease but of unknown primary lesions, it has not been systematically studied (34).

Additionally, *Major axis length bin 4* negatively correlated with *Major axis length standard deviation* (SCC-0.45). Not surprisingly, *Major axis length standard deviation* was also significant in the univariate analysis, where histologic specimen with low standard deviation was associated with poor prognosis, which also indicates the less variable melanocyte morphology on the H&E slides. *Major axis length distribution standard deviation* has a 0.892 SCC with *Major axis length distribution entropy*, which is a good prognostic feature used by the multivariate metastasis model. Therefore, the direction of effect of these features is consistent, and slides with many intermediate sized cells are associated with less variation in cell sizes across the entire specimen, which may indicate a poor prognosis for both survival and metastasis.

Major axis length bin 7 is correlated with a good prognosis in multivariate survival models (**Figure 6**). There is research to suggest that large nuclei are correlated with poor prognoses (35), which is slightly different from our features, because we segmented the entire cell rather than just the nucleus. As stated previously, small melanoma cells have also been associated with a poor prognosis. Enlargement of nuclei is typical in cancer histology. One hypothesis could be that *cell* enlargement to an extreme degree may represent a cell which replicates its DNA and cytoplasmic contents but cannot enter S phase and divide properly. Extremely large cells would therefore be a better prognostic factor. Macrophages, with their small nuclear to cytoplasmic ratio, could also contribute to this large-cell category. Our pipeline makes measuring small but systematic differences possible.

The *Small-Hyperchromatic cell distribution kurtosis* is also a significant predictor for univariate metastasis analysis and poor survival in multivariate survival analysis, which was already discussed in above univariate analysis section.

4. Discussion

The motivation behind this study is to develop an interpretable cell morphology pipeline and construct machine learning models for sensitive and specific 5-years survival (SN:86%, SP:78%) and metastasis (SN:72%, SP:71%) prognostics. We were able to generate several models that are highly sensitive and specific for both 5-years metastasis and survival risk prediction. Our work demonstrated that image features as the sole variables are powerful prognostic tools for prediction tasks, and the methodology is low cost, fast, and easy to implement.

We showed that the CNN-based approach used to isolate tumor regions improved predictive performance and reduced variability among classifiers in some instances. Moreover, no features extracted from only CNN-identified tumor regions had an opposite effect as the ones extracted from the whole

TABLE 2 Accuracy of American Joint Committee on Cancer (AJCC) to predict 5-years survival and metastasis, where stratification is by Stages I and II vs. Stage III.

Stages I and II vs. III	Sensitivity	Specificity	F1 score
Survival	0.455	0.735	0.444
Metastasis	0.414	0.778	0.511

TABLE 3 The best models for survival and metastasis prognosis, among convolutional neural networks (CNN)-trained tumor region-only and whole slide images (WSI), LASSO-derived feature sets, and classifiers.

Prognosis	Best model	Sensitivity	Specificity	F1 score
5-years survival	CNN on merged cohort, Lasso-min, logistic regression	0.86	0.78	0.72
5-years metastasis	CNN on IUSM cohort, Lasso-1se, random forest	0.78	0.57	0.72
	WSI, Lasso-1se, KNN	0.72	0.71	0.73

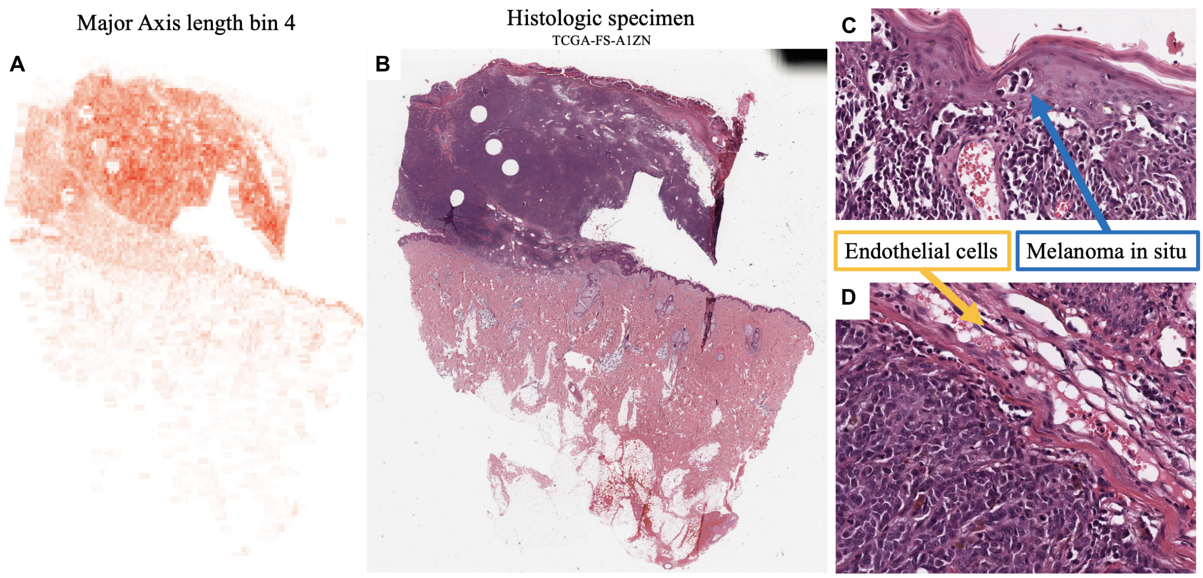


FIGURE 7
Feature visualization, Major Axis Length bin 4. **(A)** Major Axis bin 4 heatmap shown adjacent to the original **(B)** specimen. When inspected on high power, these cells are intermediate-to-small sized. **(C)** Melanoma in-situ with similarly sized spinosa cells. **(D)** Similarly sized endothelia. Tumor region identification by clinical proteomic tumor analysis consortium (CPTAC) convolutional neural networks (CNN).

slide images (Supplementary Figure 5). This demonstrated that the identified morphological descriptors are very robust to highly heterogenous cell quantities and morphologies in the histopathology slides. Given the enormous variety of melanoma histology and very small feature sets, we consider the sensitivity and specificity of this metastasis pipeline as promising for future development and adoption. Although the advantage of adopting the tumor selection step may not be obvious, we aim to test this same pipeline for our future cohort study: It will mostly contain patients with Stages I and II melanoma, and therefore, whose biopsies contain much more non-tumor tissue and very limited tumor region.

In this work, we discovered that the high density of *Small-Hyperchromatic cells* coincided with tumors that have more necrosis, ulceration, and pockets of dense inflammatory cell infiltration (Figure 5), and that cells with less immune infiltration overall, and the few that are present have a uniform distribution across the histological specimen. In specimen with greater degrees of immune infiltration, there is a large standard distribution of densities, characterized by pockets of dense inflammation and sparsely infiltrated areas. The density, kurtosis, and standard deviation of *Small-Hyperchromatic cell density* were all significant features for the prediction of metastasis and survival, and the direction of effect was consistent.

We found that slides with the densest regions of *Small-Hyperchromatic cells* coincide with large amounts of necrosis, especially ulceration, which is necrosis at the surface of the tumor (Figure 6C). Tumor ulceration is known as a poor prognostic factor for metastasis and survival, and what differentiates Stages IIa and IIb melanoma, and one of two criteria which differentiates Stages Ia/Ib. The interaction between the quantity and variability of immune infiltration and necrosis was not readily decipherable, and we hope to focus on this specifically in future research, by classifying cell types in the tumor and microenvironment, to quantify the colocalization of distinct inflammatory, stromal, and tumor cells directly.

Our model revealed that *Major Axis Length bin 4* was a significant feature used to predict both survival and metastasis. This feature corresponded to melanoma cells of intermediate to small size. Smaller melanoma cells have been reported to have a more aggressive clinical course (34). Larger cells (*Major Axis Length bin 7*) were associated with a better prognosis in our survival models. The relationship between cell size and prognosis in melanoma will be rigorously studied in a large cohort of patients in our future work.

Despite the successful development of machine learning models using interpretable features from primary biopsy histopathology for prognosis of melanoma, there are still limitations in our study: First, we had a limited cohort of Stage I/II patients, for whom a tool such as this would have the greatest impact. This is a problem common to melanoma metastasis research, more generally. Expanding our analysis to include clinicopathologic variables in a large cohort of Stage I/II is being carried out in our ongoing project. Additionally, though a good portion of the identified features for metastasis are interpretable or recapitulate those features known for survival prediction, for those that are still not discernable to human eyes, we believe that a large sample size will allow us to further validate and understand those features. For example, our pathologist was not able to identify a consistent pattern among cell morphologies among WSIs that maximized the *Minimum Delaunay distance skewness* feature, which may demonstrate that computer-quantified features are not always distinguishable to human eyes and may be superior to human in terms of refined feature extraction.

Third, the information captured by some features is correlated, and therefore may be redundant. For example, it is difficult to tell the difference between *Minimum Delaunay distance bin 10* and *Maximum Delaunay distance bin 1*. Rather than having three different distributions for maximum, minimum, and mean Delaunay distances among nuclei centroids, we can use a single distribution to describe cell density. Also, taken together, area and major/minor axis ratio together provide information about how large and ellipsoid a cell is, and the features *Major Axis* and *Minor Axis Length* may be redundant.

Finally, we did not explicitly model the interactions between specific cell types within the tissue. Existing research has quantified the architecture of the melanoma tumor and microenvironment by building "topological tumor graphs" that consist of a web of connected lymphocytes, fibroblasts, and cancer cells (36). Tumors with increased stroma and fibrous barriers separating lymphocytes from tumor were associated with a *worse* prognosis. Our work employs statistics (i.e., kurtosis, entropy, standard deviation) to describe the cell heterogeneity within a single histological specimen. We do not however, explicitly measure cell-cell interactions and spatial arrangements. Identifying cell types and establishing a metrics for their interactions is part of our ongoing work. In our future work, we plan to incorporate similar metrics into models to improve prognostic accuracy with a larger cohort.

This research has important implications for the future. Our research team has applied this interpretable cell morphology machine learning pipeline to several cancer types with success (17, 37). We have made improvements on the framework to improve model stability by reducing collinear variables and investigating the role of CNNs in focusing the analysis to tumor regions. The next step for our research is to assemble a large retrospective cohort of approximately Stages I and II patients with at least 5 years of clinical follow up. Being able to accurately predict 5-years metastasis risk in a large cohort of early-stage melanoma patients would transform melanoma clinical care. Currently, there is a shortage of dermatologists, and melanoma is a common, potentially aggressive cancer. This prognostic tool could help diagnose future melanoma metastasis at an earlier stage, which could potentially improve a patient's chance of survival, as response to treatment in advanced melanoma is inversely correlated with tumor burden (38). Triaging early-stage patients would also provide researchers with a means to identify a patient population for studying the biology of metastasis and tumor dormancy.

Our pipeline could also be applied to the study of immunotherapy response. The current clinical gold standard is PD-L1 staining of tumor tissue, but it is poorly predictive of patients who will respond to immunotherapy, nor those who will have adverse events due to the immune checkpoint inhibition (39). AI approaches to predict these by analyzing histopathology and radiological images have been published, but most employ DL learning approaches and interpretability/explainability is still a key issue (40). As in our discussion of melanoma prognosis, we believe that deep-learning and more interpretable approaches are both needed for effective clinical translation.

5. Conclusion

In this study, we were able to develop two models, which use a set of interpretable morphological features, to predict melanoma 5-years survival and metastasis with maximum F1 scores of 0.72 and 0.73 respectively. The maximum sensitivity of our metastasis model is 0.72, and although this level of sensitivity is not superior to the published deep learning-based methods, our models are transparent on the features identified and are much more interpretable than deep-learning approaches. We demonstrated the interpretability of image features and models by recapitulating several known prognostic features. We believe that the accuracy of our metastasis model will improve with a larger cohort of patients. Overall, our methods proved quite interpretable and accurate, laying the foundation for a robust, clinically relevant, accurate, low-cost, and rapid metastasis prediction tool for early-stage melanoma that can complement deep-learning techniques.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Indiana University School of Medicine. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

JC and ZL contributed equally to the work in this manuscript and analyzed the results and prepared the figures. JC performed the prognostic model training and wrote the manuscript. ZL trained the convolutional neural networks. AA and JC interpreted the histologic images. AA provided the IUSM cohort slides and follow-up information. JZ, AA, and KH formulated the questions and supervised the project. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the IUSCCC Pilot Grant, Indiana University Precision Health Initiative, and UL1TR002529 (CTSI TL1, NIH, S. Moe, S. Wiehe).

Acknowledgments

We thank the IUSM Department of Dermatopathology technicians and administrative staff for their support of this project, through the staining, scanning, and collection of the histologic specimens.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.1029227/full#supplementary-material

References

1. NCI. *Cancer Stat Facts: Melanoma of the Skin*. Bethesda, MD: NCI (2021).

2. von Schuckmann L, Hughes M, Ghiasvand R, Malt M, van der Pols J, Beesley V, et al. Risk of melanoma recurrence after diagnosis of a high-risk primary tumor. *JAMA Dermatol.* (2019) 155:688–93. doi: 10.1001/jamadermatol.2019.0440

3. Falk Delgado A, Zommorodi S, Falk Delgado A. Sentinel lymph node biopsy and complete lymph node dissection for melanoma. *Curr Oncol Rep.* (2019) 21:54. doi: 10.1007/s11912-019-0798-y

4. Kuchelmeister C, Schaumburg-Lever G, Garbe C. Acral cutaneous melanoma in caucasians: clinical features, histopathology and prognosis in 112 patients. *Br J Dermatol.* (2000) 143:275–80. doi: 10.1046/j.1365-2133.2000.03651.x

5. Pollack L, Li J, Berkowitz Z, Weir H, Wu X, Ajani U, et al. Melanoma survival in the united states, 1992 to 2005. *J Am Acad Dermatol.* (2011) 65:S78–86. doi: 10.1016/j.jaad.2011.05.030

6. Di Carlo V, Stiller C, Eisemann N, Bordoni A, Matz M, Curado M, et al. Does the morphology of cutaneous melanoma help to explain the international differences in survival? Results from 1 578482 adults diagnosed during 2000-2014 in 59 countries (CONCORD-3). *Br J Dermatol.* (2022) 187:364–80. doi: 10.1111/bjd.21274

7. Thurin M, Marincola F. *Molecular Diagnostics for Melanoma: Methods and Protocols*. Totowa, NJ: Humana Press (2013). doi: 10.1007/978-1-62703-727-3

8. Lim H, Collins S, Resneck J, Bolognia J, Hodge J, Rohrer T, et al. The burden of skin disease in the united states. *J Am Acad Dermatol.* (2017) 76: 958–72.e2. doi: 10.1016/j.jaad.2016.12.043

9. Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. *J Intern Med.* (2020) 288:62–81. doi: 10.1111/joim.13030

10. Forchhammer S, Abu-Ghazaleh A, Metzler G, Garbe C, Eigentler T. Development of an image analysis-based prognosis score using google's teachable machine in Melanoma. *Cancers (Basel).* (2022) 14:2243. doi: 10.3390/cancers14092243

11. Peng Y, Chu Y, Chen Z, Zhou W, Wan S, Xiao Y, et al. Combining texture features of whole slide images improves prognostic prediction of recurrence-free survival for cutaneous melanoma patients. *World J Surg Oncol.* (2020) 18:130. doi: 10.1186/s12957-020-01909-5

12. Herbert A, Koo M, Barclay M, Greenberg D, Abel G, Levell N, et al. Stage-specific incidence trends of melanoma in an English region, 1996–2015: longitudinal analyses of population-based data. *Melanoma Res.* (2020) 30:279–85. doi: 10.1097/CMR.0000000000000489

13. Kulkarni P, Robinson E, Sarin Pradhan J, Gartrell-Corrado R, Rohr B, Trager M, et al. Deep learning based on standard H&E images of primary melanoma

tumors identifies patients at risk for visceral recurrence and death. *Clin Cancer Res.* (2020) 26:1126–34. doi: 10.1158/1078-0432.CCR-19-1495

14. Coudray N, Ocampo P, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med.* (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5

15. FDA. *FDA Allows Marketing of First Whole slide Imaging System for Digital Pathology*. Silver Spring, MD: FDA (2017).

16. Edwards N, Oberti M, Thangudu R, Cai S, McGarvey P, Jacob S, et al. The CPTAC data portal: a resource for cancer proteomics research. *J Proteome Res.* (2015) 14:2707–13. doi: 10.1021/pr501254j

17. Cheng J, Zhang J, Han Y, Wang X, Ye X, Meng Y, et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer Res.* (2017) 77:e91–100. doi: 10.1158/0008-5472.CAN-17-0313

18. Macenko M, Neithammer M, Marron J, Borland D, Woosley J, Xiaojun G, et al. A method for normalizing histology slides for quantitative analysis. *Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Boston, MA: IEEE (2009). doi: 10.1109/ISBI.2009.5193250

19. Schmidt U, Weigert M, Broaddus C, Myers EW. Cell detection with star-convex polygons. *Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Granada: Springer (2018). doi: 10.1007/978-3-030-00934-2_30

20. Bankhead P, Loughrey M, Fernández J, Dombrowski Y, McArt D, Dunne P, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep.* (2017) 7:16878. doi: 10.1038/s41598-017-17204-5

21. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA (2015). doi: 10.1109/CVPR.2015.7298594

22. Kingma D, Ba J. Adam: a method for stochastic optimization. *arXiv.* (2014). [Preprint].

23. Lu Z, Xu S, Shao W, Wu Y, Zhang J, Han Z, et al. Deep-learning-based characterization of tumor-infiltrating lymphocytes in breast cancers from histopathology images and multiomics data. *JCO Clin Cancer Inform.* (2020) 4:480–90. doi: 10.1200/CCI.19.00126

24. Therneau T, Grambsch PM. *Modeling Survival Data: extending the Cox Model*. New York, NY: Springer (2000).

25. Kassambara A, Kosinski M, Biecek P. *survminer: Drawing Survival Curves Using 'ggplot2'. R Package Version 0.4.9.* (2021). Available online at: https://CRAN.R-project.org/package=survminer

26. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for cox's proportional hazards model via coordinate descent. *J Stat Softw.* (2011) 39:1–13. doi: 10.18637/jss.v039.i05

27. Wickham H. *ggplot2: Elegent Graphics for Data Analysis*. New York, NY: Springer-Verlag (2016). doi: 10.1007/978-3-319-24277-4_9

28. Campitelli E. *ggnewscale: Multiple Fill and Colour Scales in 'ggplot2'. R Package Version 0.4.8.* (2022). Available online at: https://CRAN.R-project.org/package=ggnewscale

29. Clemente C, Mihm M, Bufalino R, Zurrida S, Collini P, Cascinelli N. Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. *Cancer.* (1996) 77:1303–10. doi: 10.1002/(SICI)1097-0142(19960401)77:7<1303::AID-CNCR12>3.0.CO;2-5

30. Acs B, Ahmed F, Gupta S, Wong P, Gartrell R, Pradhan J, et al. An open source automated tumor infiltrating lymphocyte algorithm for prognosis in melanoma. *Nat Commun.* (2019) 10:5440. doi: 10.1038/s41467-019-13043-2

31. Keung E, Gershenwald J. The eighth edition American joint committee on cancer (AJCC) melanoma staging system: implications for melanoma treatment and care. *Expert Rev Anticancer Ther.* (2018) 18:775–84. doi: 10.1080/14737140.2018.1489246

32. Ronen S, Czaja R, Ronen N, Pantazis C, Iczkowski K. Small cell variant of metastatic melanoma: a mimicker of lymphoblastic leukemia/lymphoma. *Dermatopathology.* (2019) 6:231–6. doi: 10.1159/000503703

33. Barnhill R, Flotte T, Fleischli M, Perez-Atayde A. Cutaneous melanoma and atypical spitz tumors in childhood. *Cancer.* (1995) 76:1833–45. doi: 10.1002/1097-0142(19951115)76:10<1833::AID-CNCR2820761024>3.0.CO;2-L

34. Satturwar S, Pantanowitz L, Patel R, Cantley R. Cytologic features of small cell melanoma. *Diagn Cytopathol.* (2022) 50:E63–70. doi: 10.1002/dc.24889

35. Barzilai A, Goldberg I, Yulash M, Pavlotsky F, Zuckerman A, Trau H, et al. Silver-stained nucleolar organizer regions (AgNORs) as a prognostic value in malignant melanoma. *Am J Dermatopathol.* (1998) 20:473–7. doi: 10.1097/00000372-199810000-00008

36. Failmezger H, Muralidhar S, Rullan A, de Andrea C, Sahai E, Yuan Y. Topological tumor graphs: a graph-based spatial model to infer stromal recruitment for immunosuppression in melanoma histology. *Cancer Res.* (2020) 80:1199–209. doi: 10.1158/0008-5472.CAN-19-2268

37. Cheng J, Han Z, Mehra R, Shao W, Cheng M, Feng Q, et al. Computational analysis of pathological images enables a better diagnosis of TFE3 Xp11.2 translocation renal cell carcinoma. *Nat Commun.* (2020) 11:1778. doi: 10.1038/s41467-020-15671-5

38. Nishino M, Giobbie-Hurder A, Ramaiya N, Hodi F. Response assessment in metastatic melanoma treated with ipilimumab and bevacizumab: CT tumor size and density as markers for response and outcome. *J Immunother Cancer.* (2014) 2:40. doi: 10.1186/s40425-014-0040-2

39. Ribas A, Hu-Lieskovan S. What does PD-L1 positive or negative mean? *J Exp Med.* (2016) 213:2835–40. doi: 10.1084/jem.20161462

40. Laleh N, Ligero M, Perez-Lopez R, Kather J. Facts and hopes on the use of artificial intelligence for predictive immunotherapy biomarkers in cancer. *Clin Cancer Res.* (2022). [Epub ahead of print]. doi: 10.1158/1078-0432.CCR-22-0390

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 24 January 2023
DOI 10.3389/fmed.2023.1114673

OPEN ACCESS

EDITED BY
Jun Cheng,
Shenzhen University, China

REVIEWED BY
Wenxu Zhang,
City University of Hong Kong, Hong Kong SAR, China
Hongzhong Tang,
Xiangtan University, China

*CORRESPONDENCE
Xiaoyan Li
✉ lixiaoyan@cancerhosp-ln-cmu.com
Chen Li
✉ lichen@bmie.neu.edu.cn

SPECIALTY SECTION
This article was submitted to Pathology,
a section of the journal Frontiers in Medicine

RECEIVED 02 December 2022
ACCEPTED 06 January 2023
PUBLISHED 24 January 2023

CITATION
Shi L, Li X, Hu W, Chen H, Chen J, Fan Z, Gao M, Jing Y, Lu G, Ma D, Ma Z, Meng Q, Tang D, Sun H, Grzegorzek M, Qi S, Teng Y and Li C (2023) EBHI-Seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. *Front. Med.* 10:1114673. doi: 10.3389/fmed.2023.1114673

EBHI-Seg: A novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks

Liyu Shi[1], Xiaoyan Li[2]*, Weiming Hu[1], Haoyuan Chen[1], Jing Chen[1], Zizhen Fan[1], Minghe Gao[1], Yujie Jing[1], Guotao Lu[1], Deguo Ma[1], Zhiyu Ma[1], Qingtao Meng[1], Dechao Tang[1], Hongzan Sun[3], Marcin Grzegorzek[4,5], Shouliang Qi[1], Yueyang Teng[1] and Chen Li[1]*

[1]Microscopic Image and Medical Image Analysis Group, College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China, [2]Department of Pathology, Cancer Hospital of China Medical University, Liaoning Cancer Hospital and Institute, Shengyang, China, [3]Shengjing Hospital, China Medical University, Shenyang, China, [4]Institute of Medical Informatics, University of Lübeck, Lübeck, Germany, [5]Department of Knowledge Engineering, University of Economics in Katowice, Katowice, Poland

Background and purpose: Colorectal cancer is a common fatal malignancy, the fourth most common cancer in men, and the third most common cancer in women worldwide. Timely detection of cancer in its early stages is essential for treating the disease. Currently, there is a lack of datasets for histopathological image segmentation of colorectal cancer, which often hampers the assessment accuracy when computer technology is used to aid in diagnosis.

Methods: This present study provided a new publicly available *Enteroscope Biopsy Histopathological Hematoxylin and Eosin Image Dataset for Image Segmentation Tasks* (EBHI-Seg). To demonstrate the validity and extensiveness of EBHI-Seg, the experimental results for EBHI-Seg are evaluated using classical machine learning methods and deep learning methods.

Results: The experimental results showed that deep learning methods had a better image segmentation performance when utilizing EBHI-Seg. The maximum accuracy of the Dice evaluation metric for the classical machine learning method is 0.948, while the Dice evaluation metric for the deep learning method is 0.965.

Conclusion: This publicly available dataset contained 4,456 images of six types of tumor differentiation stages and the corresponding ground truth images. The dataset can provide researchers with new segmentation algorithms for medical diagnosis of colorectal cancer, which can be used in the clinical setting to help doctors and patients. EBHI-Seg is publicly available at: https://figshare.com/articles/dataset/EBHI-SEG/21540159/1.

KEYWORDS

colorectal histopathology, enteroscope biopsy, image dataset, image segmentation, EBHI-Seg

1. Introduction

Colon cancer is a common deadly malignant tumor, the fourth most common cancer in men, and the third most common cancer in women worldwide. Colon cancer is responsible for 10% of all cancer cases (1). According to prior research, colon and rectal tumors share many of the same or similar characteristics. Hence, they are often classified collectively (2). The present study categorized rectal and colon cancers into one colorectal cancer category (3). Histopathological examination of the intestinal tract is both the gold standard for the diagnosis of colorectal cancer and a prerequisite for disease treatment (4).

The advantage of using the intestinal biopsy method to remove a part of the intestinal tissue for histopathological analysis, which is used to determine the true status of the patient, is that it considerably reduces damage to the body and rapid wound healing (5). The histopathology sample is then sectioned and processed with Hematoxylin and Eosin (H&E). Treatment with H&E is a common approach when staining tissue sections to show the inclusions between the nucleus and cytoplasm and highlight the fine structures between tissues (6, 7). When a pathologist performs an examination of the colon, they first examine the histopathological sections for eligibility and find the location of the lesion. The pathology sections are then examined and diagnosed using a low magnification microscope. If finer structures need to be observed, the microscope is adjusted to use high magnification for further analysis. However, the following problems usually exist in the diagnostic process: the diagnostic results become more subjective and varied due to different doctors reasons; doctors can easily overlook some information in the presence of a large amount of test data; it is difficult to analyze large amounts of previously collected data (8). Therefore, it is a necessary to address these issues effectively.

With the development and popularization of computer-aided diagnosis (CAD), the pathological sections of each case can be accurately and efficiently examined with the help of computers (9). Now, CAD is widely used in many biomedical image analysis tasks, such as microorganism image analysis (10–18), COVID-19 image analysis (19), histopatholgical image analysis (20–27), cytopathological image analysis (28–31) and sperm video analysis (32, 33). Therefore, the application of computer vision technology for colorectal cancer CAD provides a new direction in this research field (34).

One of the fundamental tasks of CAD is the aspect of image segmentation, the results of which can be used as key evidence in the pathologists' diagnostic processes. Along with the rapid development of medical image segmentation methodology, there is a wide demand for its application to identify benign and malignant tumors, tumor differentiation stages, and other related fields (35). Therefore, a multi-class image segmentation method is needed to obtain high segmentation accuracy and good robustness (36).

The present study presents a novel *Enteroscope Biopsy Histopathological H&E Image Dataset for Image Segmentation Tasks* (EBHI-Seg), which contains 4456 electron microscopic images of histopathological colorectal cancer sections that encompass six tumor differentiation stages: normal, polyp, low-grade intraepithelial neoplasia, high-grade intraepithelial neoplasia, serrated adenoma, and adenocarcinoma. The segmentation coefficients and evaluation metrics are obtained by segmenting the images of this dataset using different classical machine learning methods and novel deep learning methods.

2. Related work

The present study analyzed and compared the existing colorectal cancer biopsy dataset and provided an in-depth exploration of the currently known research findings. The limitations of the presently available colorectal cancer dataset were also pointed out.

The following conclusions were obtained in the course of the study. For existing datasets, the data types can be grouped into two major categories: Multi and Dual Categorization datasets. Multi Categorization datasets contain tissue types at all stages from Normal to Neoplastic. In Trivizakis et al. (37), a dataset called "Collection of textures in colorectal cancer histology" is described. It includes 5,000 patches of size 74 × 74 μm and contains seven categories. However, because there were only 10 images, it is too small for a data sample and lacked generalization capability. In Chen et al. (23), a dataset called "NCT-CRC-HE-100K" is proposed. This is a set of 100,000 non-overlapping image patches of histological human colorectal cancer (CRC) and normal tissue samples stained with (H&E) that was presented by the National Center for Tumor Diseases (NCT). These image patches are from nine different tissues with an image size of 224 × 224 pixels. The nine tissue categories are adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium. This dataset is publicly available and commonly used. However, because the image sizes are all 224 × 224 pixels, the dataset underperformed in some global details that need to be observed in individual categories. Two datasets are utilized in Oliveira et al. (38): one containing colonic H&E-stained biopsy sections (CRC dataset) and the other consisting of prostate cancer H&E-stained biopsy sections (PCa dataset). The CRC dataset contains 1,133 colorectal biopsy and polypectomy slides grouped into three categories and labeled as non-neoplastic, low-grade and high-grade lesions. In Kausar et al. (39), a dataset named "MICCAI 2016 gland segmentation challenge dataset (GlaS)" is used. This dataset contained 165 microscopic images of H&E-stained colon glandular tissue samples, including 85 training and 80 test datasets. Each dataset is grouped into two parts: benign and malignant tumors. The image size is 775 × 522 pixels. Since this dataset has only two types of data and the number of data is too little, so that it performs poorly on some multi-type training.

Dual Categorization datasets usually contain only two types of tissue types: Normal and Neoplastic. In Wei et al. (40), a dataset named "FFPE" is proposed. This dataset obtained its images by extracting 328 Formalin-fixed Paraffin-embedded (FFPE) whole-slide images of colorectal polyps classified into two categories of : hyperplastic polyps (HPs) and sessile serrated adenomas (SSAs). This dataset contained 3,125 images with an image size of 224 × 224 pixels and is small in type and number. In Bilal et al. (41), two datasets named "UHCW" and "TCGA" are proposed. The first dataset is a colorectal cancer biopsy sequence developed at the University Hospital of Coventry and Warwickshire (UHCW) for internal validation of the rectal biopsy trial. The second dataset is the Cancer Genome Atlas (TCGA) for external validation of the trial. This dataset is commonly used as a publicly available cancer dataset and stores genomic data for more than 20 types of cancers. The two dataset types are grouped into two categories: Normal and Neoplastic. The first dataset contains 4,292 slices, and the second dataset contained 731 slices with an image size of 224 × 224 pixels.

All of the information for the existing datasets is summarized in Table 1. The issues associated with the dataset mentioned above included fewer data types, small amount of data, inaccurate dataset ground truth, etc. The current study required an open-source multi-type colonoscopy biopsy image dataset.

TABLE 1 A dataset for the pathological classification of colorectal cancer.

	Dataset name	Categorization	Amount	Size	Year
Multi categorization	Collection of textures in colorectal cancer histology	Lymphoid follicles, mucosal glands, debris, adipose, tumor epithelium simple stroma, complex stroma, background patches with no tissue	5,000	74 × 74 μm (0.495 micrometer per pixel)	2016
	HE-NCT-CRC-100K	MUS, NORM, STR, TUM ADI, BACK, DEB, LYM, MUC	100,000	224 ×224 pixels	2016
	MICCAI'16 gland seg-mentation challenge dataset	Benign tumors, malignant tumors	85	775 ×522 pixels	2017
	CRC dataset	Non-neoplastic, low-grade, high-grade lesions	1,133	512 ×512 pixels	2021
Dual categorization	FFPE	HPs, SSAs	3,152	224 ×224 pixels	2021
	The Cancer Genome Atlas dataset	Normal, Neoplastic	731	224 ×224 pixels	2021
	University Hospitals Coventry and Warwick-shire dataset	Normal, Neoplastic	4,292	224 ×224 pixels	2021

3. Basic information for EBHI-Seg

3.1. Dataset overview

The dataset in the present study contained 4,456 histopathology images, including 2,228 histopathology section images and 2,228 ground truth images. These include normal (76 images and 76 ground truth images), polyp (474 images and 474 ground truth images), low-grade intraepithelial neoplasia (639 images and 639 ground truth images), high-grade intraepithelial neoplasia (186 images and 186 ground truth images), serrated adenoma (58 images and 58 ground truth images), and adenocarcinoma (795 images and 795 ground truth images). The basic information for the dataset is described in detail below. EBHI-Seg is publicly available at: https://figshare.com/articles/dataset/EBHI-SEG/21540159/1.

In the present paper, H&E-treated histopathological sections of colon tissues are used as data for evaluating image segmentation. The dataset is obtained from two histopathologists at the Cancer Hospital of China Medical University [proved by "Research Project Ethics Certification" (No. 202229)]. It is prepared by 12 biomedical researchers according to the following rules: Firstly, if there is only one differentiation stage in the image and the rest of the image is intact, then the differentiation stage became the image label; Secondly, if there is more than one differentiation stage in the image, then the most obvious differentiation is selected as the image label; In general, the most severe and prominent differentiation in the image was used as the image label.

Intestinal biopsy was used as the sampling method in this dataset. The magnification of the data slices is 400×, with an eyepiece magnification of 10× and an objective magnification of 40×. A Nissan Olympus microscope and NewUsbCamera acquisition software are used. The image input size is 224 × 224 pixels, and the format is *.png. The data are grouped into five types described in detail in Section 2.2.

3.2. Data type description

3.2.1. Normal

Colorectal tissue sections of the standard category are made-up of consistently ordered tubular structures and that does not appear infected when viewed under a light microscope (42). Section images with the corresponding ground truth images are shown in Figure 1A.

3.2.2. Polyp

Colorectal polyps are similar in shape to the structures in the normal category, but have a completely different histological structure. A polyp is a redundant mass that grows on the surface of the body's cells. Modern medicine usually refers to polyps as unwanted growths on the mucosal surface of the body (43). The pathological section of the polyp category also has an intact luminal structure with essentially no nuclear division of the cells. Only the atomic mass is slightly higher than that in the normal category. The polyp category and corresponding ground truth images are shown in Figure 1B.

3.2.3. Intraepithelial neoplasia

Intraepithelial neoplasia (IN) is the most critical precancerous lesion. Compared to the normal category, its histological images show increased branching of adenoid structures, dense arrangement, and different luminal sizes and shapes. In terms of cellular morphology, the nuclei are enlarged and vary in size, while nuclear division increases (44). The standard Padova classification currently classifies intraepithelial neoplasia into low-grade and high-grade INs. High-grade IN demonstrate more pronounced structural changes in the lumen and nuclear enlargement compared to low-grade IN. The images and ground truth diagrams of high-grade and low-grade INs are shown in Figures 1C, D.

3.2.4. Adenocarcinoma

Adenocarcinoma is a malignant digestive tract tumor with a very irregular distribution of luminal structures. It is difficult to identify its border structures during observation, and the nuclei are significantly enlarged at this stage (45). An adenocarcinoma with its corresponding ground truth diagram is shown in Figure 1E.

3.2.5. Serrated adenoma

Serrated adenomas are uncommon lesions, accounting for 1% of all colonic polyps (46). The endoscopic surface appearance of serrated

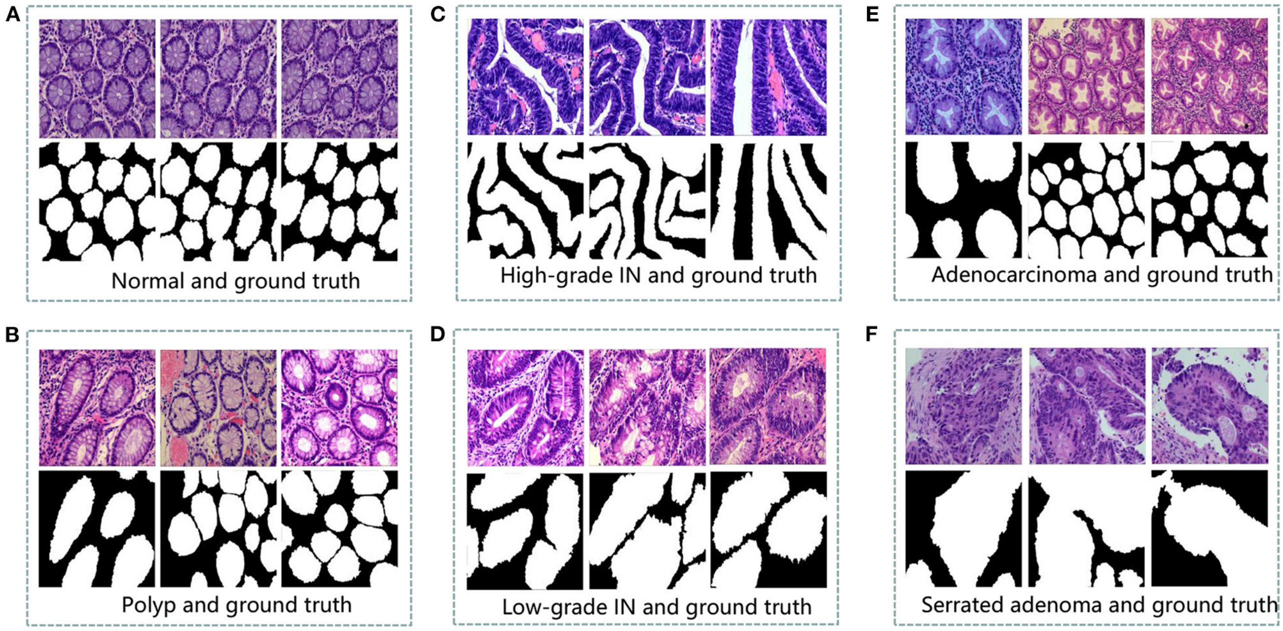

A
Normal and ground truth
B
Polyp and ground truth
C
High-grade IN and ground truth
D
Low-grade IN and ground truth
E
Adenocarcinoma and ground truth
F
Serrated adenoma and ground truth

FIGURE 1
An example of histopathological images database. **(A)** Normal and ground truth, **(B)** Polyp and ground truth, **(C)** High-grade Intraepithelial Neoplasia and ground truth, **(D)** Low-grade Intraepithelial Neoplasia and ground truth, **(E)** Adenocarcinoma and ground truth, and **(F)** Serrated adenoma and ground truth.

adenomas is not well characterized but is thought to be similar to that of colonic adenomas with tubular or cerebral crypt openings (47). The image of a serrated adenoma with a corresponding ground truth diagram is shown in Figure 1F.

4. Evaluation of EBHI-Seg

4.1. Image segmentation evaluation metric

Six evaluation metrics are commonly used for image segmentation tasks. The Dice ratio metric is a standard metric used in medical images that is often utilized to evaluate the performance of image segmentation algorithms. It is a validation method based on spatial overlap statistics that measures the similarities between the algorithm segmentation output and ground truth (48). The Dice ratio is defined in Equation (1).

$$\mathrm{DiceRatio} = \frac{2\,|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}| + |\mathrm{Y}|}. \tag{1}$$

In Equation (1), for a segmentation task, X and Y denote the ground truth and segmentation mask prediction, respectively. The range of the calculated results is [0,1], and the larger the result the better.

The Jaccard index is a classical set similarity measure with many practical applications in image segmentation. The Jaccard index measures the similarity of a finite set of samples: the ratio between the intersection and concatenation of the segmentation results and ground truth (49). The Jaccard index is defined in Equation (2).

$$\mathrm{JaccardIndex} = \frac{|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X} \cup \mathrm{Y}|}. \tag{2}$$

The range of the calculated results is [0,1], and the larger the result the better.

TABLE 2 Confusion matrix.

Ground truth	Predict mask	
	Positive	Negative
Positive	TP	TN
Negative	FP	FN

Recall and precision are the recall and precision rates, respectively. The range of the calculated results is [0,1]. A higher output indicates a better segmentation result. Recall and precision are defined in Equations (3), (4),

$$\mathrm{Precison} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{3}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4}$$

where TP, FP, TN, and FN are defined in Table 2.

The conformity coefficient (Confm Index) is a consistency coefficient, which is calculated by putting the binary classification result of each pixel from [$-\infty$,1] into continuous interval [$-\infty$,1] to calculate the ratio of the number of incorrectly segmented pixels to the number of correctly segmented pixels to measure the consistency between the segmentation result and ground truth. The conformity coefficient is defined in Equations (5), (6),

$$\mathrm{ConfmIndex} = (1 - \frac{\theta_{\mathrm{AE}}}{\theta_{\mathrm{TP}}}), \theta_{\mathrm{TP}} > 0, \tag{5}$$

$$\mathrm{ConfmIndex} = \mathrm{Failure}, \theta_{\mathrm{TP}} = 0, \tag{6}$$

Where θ_{AE}= θ_{FP}+θ_{FN} represents all errors of the fuzzy segmentation results. θ_{TP} is the number of correctly classified pixels. Mathematically, ConfmIndex can be negative infinity if θ_{TP}=0. Such a segmentation result is definitely inadequate and treated as failure without the need of any further analysis.

4.2. Classical machine learning methods

Image segmentation is one of the most commonly used methods for classifying image pixels in decision-oriented applications (50). It groups an image into regions high in pixel similarity within each area and has a significant contrast between different regions (51). Machine learning methods for segmentation distinguish the image classes using image features. (1) *k*-means algorithm is a classical division-based clustering algorithm, where image segmentation means segmenting the image into many disjointed regions. The essence is the clustering process of pixels, and the *k*-means method is one of the simplest clustering methods (52). Image segmentation of the present study dataset is performed using the classical machine learning method described above. (2) Markov random field (MRF) is a powerful stochastic tool that models the joint probability distribution of an image based on its local spatial action (53). It can extract the texture features of the image and model the image segmentation problem. (3) OTSU algorithm is a global adaptive binarized threshold segmentation algorithm that uses the maximum inter-class variance between the image background and the target image as the selection criterion (54). The image is grouped into foreground and background parts based on its grayscale characteristics independent of the brightness and contrast. (4) Watershed algorithm is a region-based segmentation method, that takes the similarity between neighboring pixels as a reference and connects those pixels with similar spatial locations and grayscale values into a closed contour to achieve the segmentation effect (55). (5) Sobel algorithm has two operators,

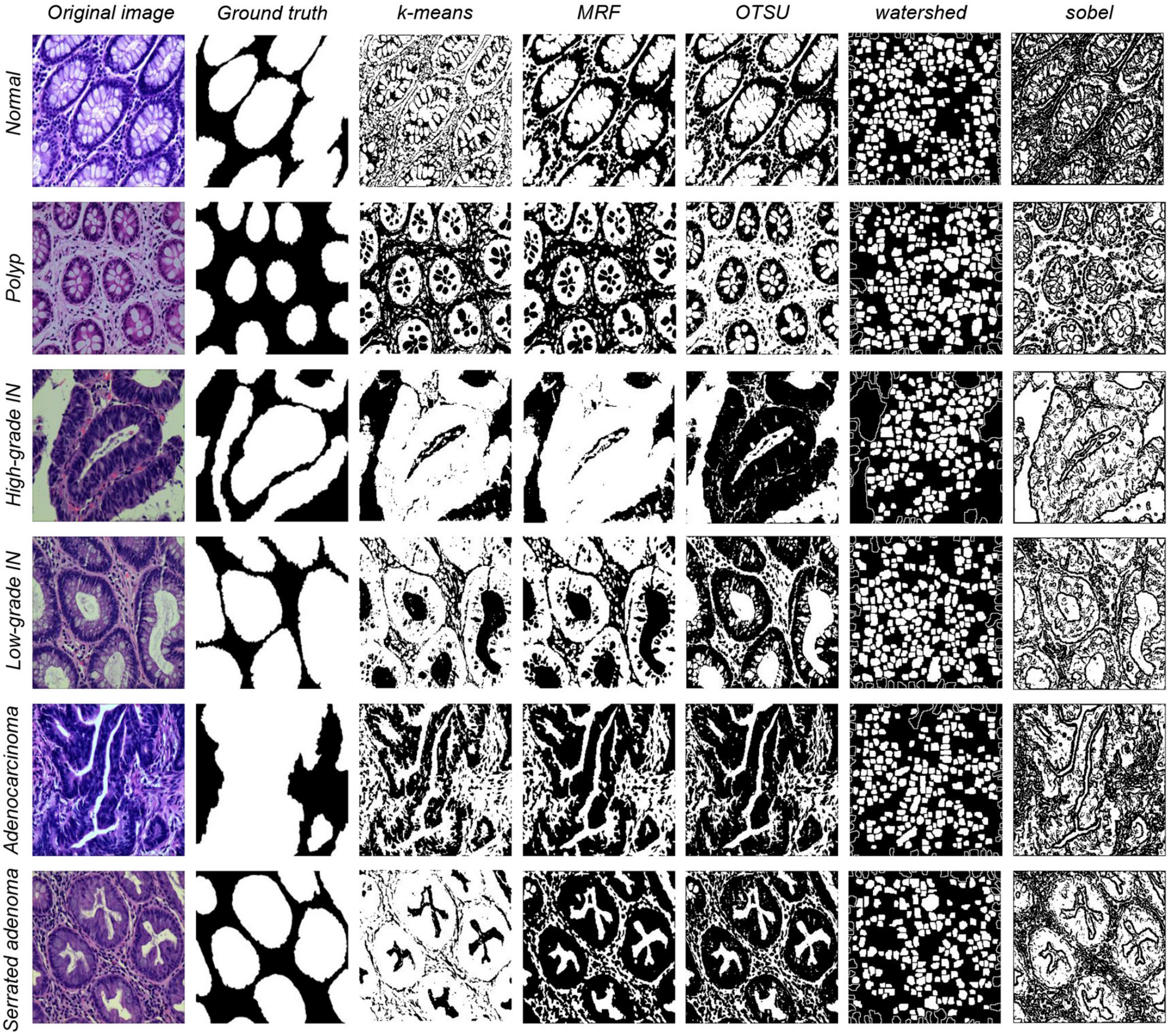


FIGURE 2
Five types of data segmentation results obtained by different classical machine learning methods.

TABLE 3 Evaluation metrics for five different segmentation methods based on classical machine learning.

		Dice ratio	JaccardIndex	Conformity coefficient	Precision	Recall
Normal	*k*-means	0.648	0.488	−0.184	0.646	0.663
	MRF	0.636	0.473	−0.230	0.637	0.658
	OTSU	0.410	0.265	−2.871	0.515	0.351
	Watershed	0.461	0.300	−1.375	0.668	0.356
	Sobel	0.652	0.487	−0.102	0.763	0.579
Polyp	*k*-means	0.592	0.430	−0.528	0.546	0.663
	MRF	0.511	0.362	−2.133	0.540	0.502
	OTSU	0.400	0.259	−3.108	0.413	0.399
	Watershed	0.433	0.277	−1.675	0.551	0.362
	Sobel	0.583	0.416	−0.499	0.626	0.562
High-grade IN	*k*-means	0.626	0.478	−0.467	0.650	0.620
	MRF	0.550	0.441	−30.85	0.614	0.526
	OTSU	0.249	0.150	−12.06	0.373	0.191
	Watershed	0.472	0.309	−1.258	0.738	0.350
	Sobel	0.634	0.469	−0.200	0.728	0.577
Low-grade IN	*k*-means	0.650	0.492	−0.172	0.651	0.663
	MRF	0.554	0.404	−1.808	0.643	0.504
	OTSU	0.886	0.811	0.6998	0.832	0.979
	Watershed	0.464	0.303	−1.345	0.676	0.357
	Sobel	0.656	0.492	−0.079	0.771	0.582
Adenocarcinoma	*k*-means	0.633	0.481	−0.414	0.655	0.645
	MRF	0.554	0.404	−1.808	0.643	0.504
	OTSU	0.336	0.215	−5.211	0.454	0.282
	Watershed	0.458	0.298	−1.437	0.700	0.349
	Sobel	0.553	0.388	−0.733	0.692	0.484
Serrated adenoma	*k*-means	0.636	0.473	−0.230	0.637	0.658
	MRF	0.571	0.419	−0.898	0.656	0.547
	OTSU	0.393	0.248	−2.444	0.565	0.315
	Watershed	0.449	0.290	−1.494	0.656	0.345
	Sobel	0.698	0.541	0.7484	0.662	0.572

where one detects horizontal edges and the other detects vertical flat edges. An image is the final result of its operation. Sobel edge detection operator is a set of directional operators that can be used to perform edge detection from different directions (56). The segmentation results are shown in Figure 2.

The performance of EBHI-Seg for different machine learning methods is observed by comparing the images segmented using classical machine learning methods with the corresponding ground truth. The segmentation evaluation metrics results are shown in Table 3. The Dice ratio algorithm is a similarity measure, usually used to compare the similarity of two samples. The value of one for this metric is c onsidered to indicate the best effect, while the value of the worst impact is zero. The Table 3 shows that *k*-means has a good Dice ratio algorithm value of up to 0.650 in each category. The MRF and Sobel segmentation results also achieved a good Dice ratio algorithm value of around 0.6. In terms of image precision and recall segmentation coefficients, *k*-means is maintained at approximately 0.650 in each category. In the classical machine learning methods, *k*-means has the best segmentation results, followed by MRF and Sobel. OTSU has a general effect, while the watershed algorithm has various coefficients that are much lower than those in the above methods. Moreover, there are apparent differences in the segmentation results when using the above methods.

In summary, EBHI-Seg has significantly different results when using different classical machine learning segmentation methods. Different classical machine learning methods have an obvious differentiation according to the image segmentation evaluation metrics. Therefore, EBHI-Seg can effectively evaluate the segmentation performance of different segmentation methods.

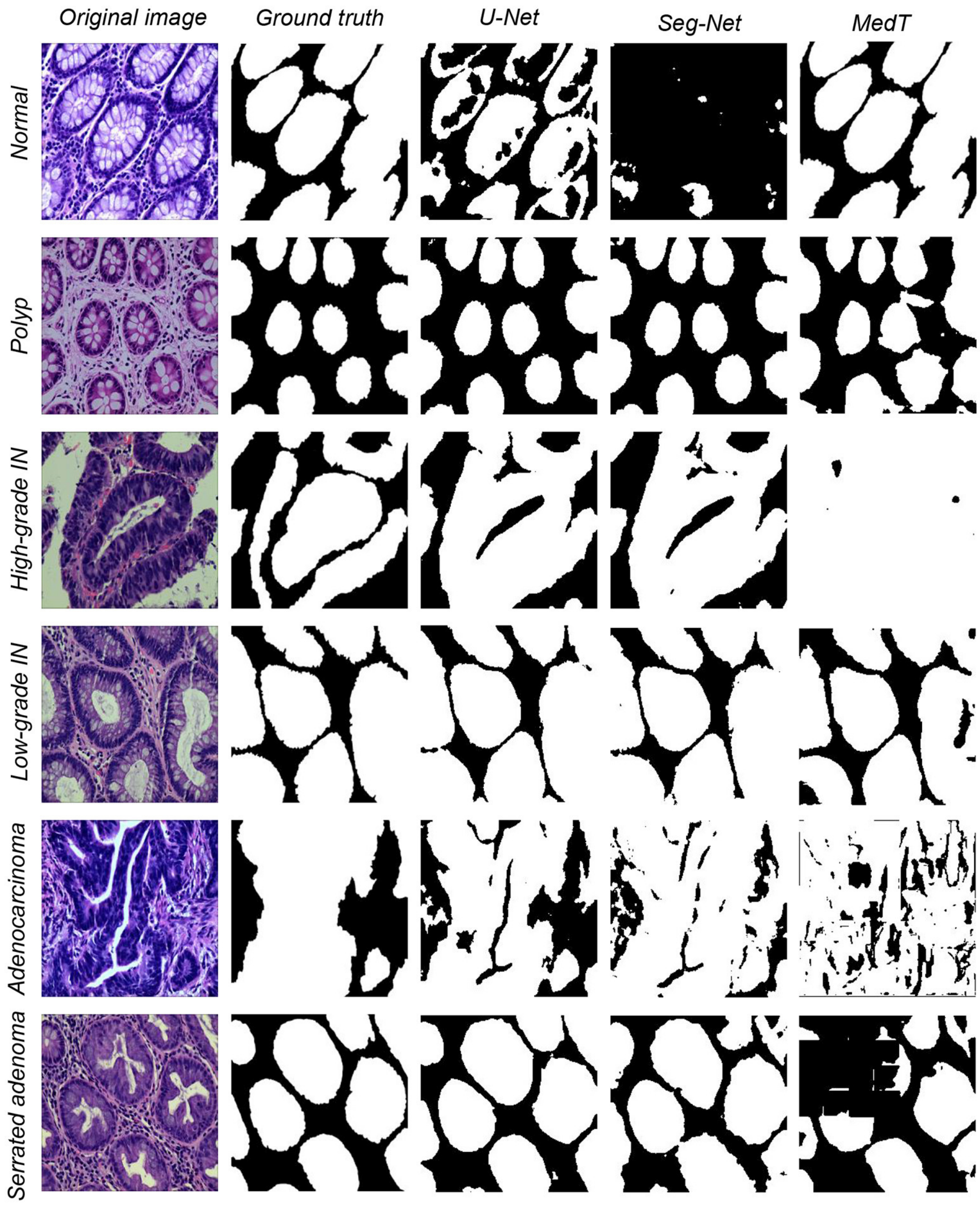


FIGURE 3
Three types of data segmentation results obtained by different deep learning methods.

4.3. Deep learning methods

Besides the classical macine learning metheds tested above, some popular deep learning methods are also tested. (1) Seg-Net is an open source project for image segmentation (57). The network is identical to the convolutional layer of VGG-16, with the removal of the fully-connected hierarchy and the addition of max-pooling indices resulting in improved boundary delineation. Seg-Net performs better in large datasets. (2) U-Net network structure was first proposed in 2015 (58) for medical imaging. U-Net is lightweight, and its simultaneous detection of local and global information is helpful for both information extraction and diagnostic results from clinical medical images. (3) MedT is a network published in 2021, which is a transformer structure that applies an attention mechanism based on medical image segmentation (59). The segmentation results are shown in Figure 3.

The segmentation effect is test on the present dataset using three deep learning models. In the experiments, each model is trained using the ratio of the training set, validation set, and test set of 4 : 4 : 2. All of the information for the existing datasets is summarized in Table 4. The model learning rate is set to $3e - 6$, epochs are set to 100, and batch-size is set to 1. The optimizer is Adam, the loss function is crossentropyloss and the activation function is ReLU. The dataset segmentation results of using three different models are shown in Figure 3. The experimental segmentation evaluation metrics are shown in Table 5. Overall, deep learning performs much better than classical machine learning methods. Among them, the evaluation indexes of the training results using the U-Net and Seg-Net models can reach 0.90 on average. The evaluation results of the MedT model are slightly worse at a level, between 0.70 and 0.80. The training time is longer for MedT and similar for U-Net and Seg-Net.

Based on the above results, EBHI-Seg achieved a clear differentiation using deep learning image segmentation methods. Image segmentation metrics for different deep learning methods are significantly different so that EBHI-Seg can evaluate their segmentation performance.

4.4. Experimental environment

This section presents the hardware configuration data required for this experiment as well as the software version.

Processor: Intel Core i7-8700 @ 3.20GHz Six Core

Graphics (GPU): NVIDIA GeForce RTX 2080

Graphics (CPU): Intel UHD Graphics 630

Hard Drive: SM961 NVMe SAMSUNG 512GB (Solid State Drive)

Motherboard: Dell 0NNNCT (C246 chipset)

Mainframe: Dell Precision 3630 Tower Desktop Mainframe

Software Versions: CUDA 11.2, torch 1.7.0, torchvision 0.8.0, python 3.8.

TABLE 4 Deep learning of the number of different types of training images.

	Train	Test	Predict
Normal	30	30	16
Polyp	190	190	94
Low-grade IN	256	256	127
High-grade IN	74	74	38
Serrated adenoma	23	23	12
Adenocarcinoma	318	318	159

5. Discussion

5.1. Discussion of image segmentation results using classical machine learning methods

Six types of tumor differentiation stage data in EBHI-Seg were analyzed using classical machine learning methods to obtain the results in Table 3. Base on the Dice ratio metrics, *k*-means, MRF and Sobel show no significant differences among the three methods around 0.55. In contrast, Watershed metrics are ∼0.45 on average, which is lower than the above three metrics. OTSU index is around ∼0.40 because the foreground-background is blurred in some experimental samples and OTSU had a difficulty extracting a suitable segmentation threshold, which resulted in undifferentiated test results. Precision and Recall evaluation indexes for k-means, MRF, and Sobel are also around 0.60, which is higher than those for OTSU and Watershed methods by about 0.20. In these three methods, *k*-means and MRF are higher than Sobel in the visual performance of the images. Although Sobel is the same as these two methods in terms of metrics, it is difficult to distinguish foreground and background images in real

TABLE 5 Evaluation metrics for three different segmentation methods based on deep learning.

		Dice ratio	JaccardIndex	Conformity coefficient	Precision	Recall
Normal	U-Net	0.411	0.263	−2.199	0.586	0.328
	Seg-Net	0.777	0.684	−0.607	0.895	0.758
	MedT	0.676	0.562	−0.615	0.874	0.610
Polyp	U-Net	0.965	0.308	−1.514	0.496	0.470
	Seg-Net	0.937	0.886	0.858	0.916	0.965
	MedT	0.771	0.643	0.336	0.687	0.920
High-grade IN	U-Net	0.895	0.816	0.747	0.847	0.961
	Seg-Net	0.894	0.812	0.757	0.881	0.913
	MedT	0.824	0.707	0.556	0.740	0.958
Low-grade IN	U-Net	0.911	0.849	0.773	0.879	0.953
	Seg-Net	0.924	0.864	0.826	0.883	0.977
	MedT	0.889	0.808	0.730	0.876	0.916
Adenocarcinoma	U-Net	0.887	0.808	0.718	0.850	0.950
	Seg-Net	0.865	0.775	0.646	0.792	0.977
	MedT	0.735	0.595	0.197	0.662	0.864
Serrated adenoma	U-Net	0.938	0.886	0.865	0.899	0.983
	Seg-Net	0.907	0.832	0.794	0.859	0.963
	MedT	0.670	0.509	−0.043	0.896	0.544

images.The segmentation results for MRF are obvious but the running time for MRF is too long in comparison with other classical learning methods. Since classical machine learning methods have a rigorous theoretical foundation and simple ideas, they have been shown to perform well when used for specific problems. However, the performance of different methods varied in the present study.

5.2. Discussion of image segmentation results using deep learning methods

In general, deep learning models are considerably superior to classical machine learning methods, and even the lowest MedT performance is still higher than the highest accuracy of classical machine learning methods. In EBHI-Seg, the Dice ratio evaluation index of MedT reaches ∼0.75. However, the MedT model size was larger and as a result the training time was too long. U-Net and Seg-Net have higher evaluation indexes than MedT, both of about 0.88. Among them, Seg-Net has the least training time and the lowest training model size. Because the normal category has fewer sample images than other categories, the evaluation metrics of the three deep learning methods in this category are significantly lower than those in other categories. The evaluation metrics of the three segmentation methods are significantly higher in the other categories, with Seg-Net averaging above 0.90 and MedT exceeding 0.80.

6. Conclusion and future work

The present stduy introduced a publicly available colorectal pathology image dataset containing 4456 magnified 400× pathology images of six types of tumor differentiation stages. EBHI-Seg has high segmentation accuracy as well as good robustness. In the classical machine learning approach, segmentation experiments were performed using different methods and evaluation metrics analysis was carried out utilizing segmentation results. The highest and lowest Dice ratios are 0.65 and 0.30, respectively. The highest Precision and Recall values are 0.70 and 0.90, respectively, while the lowest values are 0.50 and 0.35, respectively. All three models performed well when using the deep learning method, with the highest Dice ratio reaching above 0.95 and both Precision and Recall values reaching above 0.90. The segmentation experiments using EBHI-Seg show that this dataset effectively perform the segmentation task in each of the segmentation methods. Furthermore, there are significant differences among the segmentation evaluation metrics. Therefore, EBHI-Seg is practical and effective in performing image segmentation tasks.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: EBHI-Seg is publicly available at: https://figshare.com/articles/dataset/EBHISEG/21540159/1.

Author contributions

LS: data preparation, experiment, result analysis, and paper writing. XL: data collection and medical knowledge. WH: data collection, data preparation, and paper writing. HC: data preparation and paper writing. JC, ZF, MGa, YJ, GL, DM, ZM, QM, and DT: data preparation. HS: medical knowledge. MGr and YT: result analysis. SQ: method. CL: data collection, method, experiment, result analysis, paper writing, and proofreading. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 82220108007) and the Beijing Xisike Clinical Oncology Research Foundation (No. Y-tongshu2021/1n-0379).

Acknowledgments

We thank Miss. Zixian Li and Mr. Guoxian Li for their important discussion in this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. (2021) 71:209–49. doi: 10.3322/caac.21660

2. Lee YC, Lee YL, Chuang JP, Lee JC. Differences in survival between colon and rectal cancer from SEER data. *PLoS ONE*. (2013) 8:e78709. doi: 10.1371/journal.pone.0078709

3. Pamudurthy V, Lodhia N, Konda VJ. Advances in endoscopy for colorectal polyp detection and classification. In: *Baylor University Medical Center Proceedings. Vol. 33*. Taylor & Francis (2020). p. 28–35. doi: 10.1080/08998280.2019.1686327

4. Thijs J, Van Zwet A, Thijs W, Oey H, Karrenbeld A, Stellaard F, et al. Diagnostic tests for *Helicobacter pylori*: a prospective evaluation of their accuracy,

without selecting a single test as the gold standard. *Am J Gastroenterol*. (1996) 91:10. doi: 10.1016/0016-5085(95)23623-6

5. Labianca R, Nordlinger B, Beretta G, Mosconi S, Mandalà M, Cervantes A, et al. Early colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. (2013) 24:vi64–vi72. doi: 10.1093/annonc/mdt354

6. Fischer AH, Jacobson KA, Rose J, Zeller R. Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harbor Protocols*. (2008) 2008:pdb-prot4986. doi: 10.1101/pdb.prot4986

7. Chan JK. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. *Int J Surg Pathol*. (2014) 22:12–32. doi: 10.1177/1066896913517939

8. Gupta V, Vasudev M, Doegar A, Sambyal N. Breast cancer detection from histopathology images using modified residual neural networks. *Biocybernetics Biomed Eng*. (2021) 41:1272–87. doi: 10.1016/j.bbe.2021.08.011

9. Mathew T, Kini JR, Rajan J. Computational methods for automated mitosis detection in histopathology images: a review. *Biocybern Biomed Eng*. (2021) 41:64–82. doi: 10.1016/j.bbe.2020.11.005

10. Li C, Wang K, Xu N. A survey for the applications of content-based microscopic image analysis in microorganism classification domains. *Artif Intell Rev*. (2019) 51:577–646. doi: 10.1007/s10462-017-9572-4

11. Zhang J, Li C, Yin Y, Zhang J, Grzegorzek M. Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer. *Artif Intell Rev*. (2022) 2022:1–58. doi: 10.1007/s10462-022-10192-7

12. Zhang J, Li C, Kosov S, Grzegorzek M, Shirahama K, Jiang T, et al. LCU-Net: a novel low-cost U-Net for environmental microorganism image segmentation. *Pattern Recogn*. (2021) 115:107885. doi: 10.1016/j.patcog.2021.107885

13. Zhao P, Li C, Rahaman MM, Xu H, Yang H, Sun H, et al. A comparative study of deep learning classification methods on a small environmental microorganism image dataset (EMDS-6): from convolutional neural networks to visual transformers. *Front Microbiol*. (2022) 13:792166. doi: 10.3389/fmicb.2022.792166

14. Kulwa F, Li C, Zhang J, Shirahama K, Kosov S, Zhao X, et al. A new pairwise deep learning feature for environmental microorganism image analysis. *Environ Sci Pollut Res*. (2022) 2022:1–18. doi: 10.1007/s11356-022-18849-0

15. Ma P, Li C, Rahaman MM, Yao Y, Zhang J, Zou S, et al. A state-of-the-art survey of object detection techniques in microorganism image analysis: from classical methods to deep learning approaches. *Artif Intell Rev*. (2022) 2022:1–72. doi: 10.1007/s10462-022-10209-1

16. Kulwa F, Li C, Grzegorzek M, Rahaman MM, Shirahama K, Kosov S. Segmentation of weakly visible environmental microorganism images using pair-wise deep learning features. *Biomed Signal Process Control*. (2023) 79:104168. doi: 10.1016/j.bspc.2022.104168

17. Zhang J, Li C, Rahaman MM, Yao Y, Ma P, Zhang J, et al. A comprehensive survey with quantitative comparison of image analysis methods for microorganism Biovolume measurements. *Arch Comput Methods Eng*. (2022) 30, 639–73. doi: 10.1007/s11831-022-09811-x

18. Zhang J, Li C, Rahaman MM, Yao Y, Ma P, Zhang J, et al. A comprehensive review of image analysis methods for microorganism counting: from classical image processing to deep learning approaches. *Artif Intell Rev*. (2021) 2021:1–70. doi: 10.1007/s10462-021-10082-4

19. Rahaman MM, Li C, Yao Y, Kulwa F, Rahman MA, Wang Q, et al. Identification of COVID-19 samples from chest X-ray images using deep learning: a comparison of transfer learning approaches. *J X-ray Sci Technol*. (2020) 28:821–39. doi: 10.3233/XST-200715

20. Chen H, Li C, Wang G, Li X, Rahaman MM, Sun H, et al. GasHis-Transformer: a multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recogn*. (2022) 130:108827. doi: 10.1016/j.patcog.2022.108827

21. Li Y, Wu X, Li C, Li X, Chen H, Sun C, et al. A hierarchical conditional random field-based attention mechanism approach for gastric histopathology image classification. *Appl Intell*. (2022) 2022:1–22. doi: 10.1007/s10489-021-02886-2

22. Hu W, Li C, Li X, Rahaman MM, Ma J, Zhang Y, et al. GasHisSDB: a new gastric histopathology image dataset for computer aided diagnosis of gastric cancer. *Comput Biol Med*. (2022) 2022:105207. doi: 10.1016/j.compbiomed.2021.105207

23. Chen H, Li C, Li X, Rahaman MM, Hu W, Li Y, et al. IL-MCAM: an interactive learning and multi-channel attention mechanism-based weakly supervised colorectal histopathology image classification approach. *Comput Biol Med*. (2022) 143:105265. doi: 10.1016/j.compbiomed.2022.105265

24. Hu W, Chen H, Liu W, Li X, Sun H, Huang X, et al. A comparative study of gastric histopathology sub-size image classification: from linear regression to visual transformer. *Front Med*. (2022) 9:1072109. doi: 10.3389/fmed.2022.1072109

25. Li Y, Li C, Li X, Wang K, Rahaman MM, Sun C, et al. A comprehensive review of Markov random field and conditional random field approaches in pathology image analysis. *Arch Comput Methods Eng*. (2022) 29:609–39. doi: 10.1007/s11831-021-09591-w

26. Sun C, Li C, Zhang J, Rahaman MM, Ai S, Chen H, et al. Gastric histopathology image segmentation using a hierarchical conditional random field. *Biocybern Biomed Eng*. (2020) 40:1535–55. doi: 10.1016/j.bbe.2020.09.008

27. Li X, Li C, Rahaman MM, Sun H, Li X, Wu J, et al. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artif Intell Rev*. (2022)2022:1–70. doi: 10.1007/s10462-021-10121-0

28. Rahaman MM, Li C, Wu X, Yao Y, Hu Z, Jiang T, et al. A survey for cervical cytopathology image analysis using deep learning. *IEEE Access*. (2020) 8:61687–710. doi: 10.1109/ACCESS.2020.2983186

29. Mamunur Rahaman M, Li C, Yao Y, Kulwa F, Wu X, Li X, et al. DeepCervix: a deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. *Comput Biol Med*. (2021) 136:104649. doi: 10.1016/j.compbiomed.2021.104649

30. Liu W, Li C, Rahaman MM, Jiang T, Sun H, Wu X, et al. Is the aspect ratio of cells important in deep learning? A robust comparison of deep learning methods for multi-scale cytopathology cell image classification: from convolutional neural networks to visual transformers. *Comput Biol Med*. (2021) 2021:105026. doi: 10.1016/j.compbiomed.2021.105026

31. Liu W, Li C, Xu N, Jiang T, Rahaman MM, Sun H, et al. CVM-Cervix: a hybrid cervical pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. *Pattern Recogn*. (2022) 2022:108829. doi: 10.1016/j.patcog.2022.108829

32. Chen A, Li C, Zou S, Rahaman MM, Yao Y, Chen H, et al. SVIA dataset: a new dataset of microscopic videos and images for computer-aided sperm analysis. *Biocybern Biomed Eng*. (2022) 2022:10. doi: 10.1016/j.bbe.2021.12.010

33. Zou S, Li C, Sun H, Xu P, Zhang J, Ma P, et al. TOD-CNN: an effective convolutional neural network for tiny object detection in sperm videos. *Comput Biol Med*. (2022) 146:105543. doi: 10.1016/j.compbiomed.2022.105543

34. Pacal I, Karaboga D, Basturk A, Akay B, Nalbantoglu U. A comprehensive review of deep learning in colon cancer. *Comput Biol Med*. (2020) 126:104003. doi: 10.1016/j.compbiomed.2020.104003

35. Miranda E, Aryuni M, Irwansyah E. A survey of medical image classification techniques. In: *2016 International Conference on Information Management and Technology (ICIMTech)*. Bandung: IEEE (2016). p. 56–61.

36. Kotadiya H, Patel D. Review of medical image classification techniques. In: *Third International Congress on Information and Communication Technology*. Singapore: Springer (2019). p. 361–9. doi: 10.1007/978-981-13-1165-9_33

37. Trivizakis E, Ioannidis GS, Souglakos I, Karantanas AH, Tzardi M, Marias K. A neural pathomics framework for classifying colorectal cancer histopathology images based on wavelet multi-scale texture analysis. *Sci Rep*. (2021) 11:1–10. doi: 10.1038/s41598-021-94781-6

38. Oliveira SP, Neto PC, Fraga J, Montezuma D, Monteiro A, Monteiro J, et al. CAD systems for colorectal cancer from WSI are still not ready for clinical acceptance. *Sci Rep*. (2021) 11:1–15. doi: 10.1038/s41598-021-93746-z

39. Kausar T, Kausar A, Ashraf MA, Siddique MF, Wang M, Sajid M, et al. SA-GAN: stain acclimation generative adversarial network for histopathology image analysis. *Appl Sci*. (2021) 12:288. doi: 10.3390/app12010288

40. Wei J, Suriawinata A, Ren B, Liu X, Lisovsky M, Vaickus L, et al. Learn like a pathologist: curriculum learning by annotator agreement for histopathology image classification. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, HI: IEEE (2021). p. 2473–83.

41. Bilal M, Tsang YW, Ali M, Graham S, Hero E, Wahab N, et al. AI based pre-screening of large bowel cancer via weakly supervised learning of colorectal biopsy histology images. *medRxiv*. (2022) doi: 10.1101/2022.02.28.22271565

42. De Leon MP, Di Gregorio C. Pathology of colorectal cancer. *Digest Liver Dis*. (2001) 33:372–88. doi: 10.1016/S1590-8658(01)80095-5

43. Cooper HS, Deppisch LM, Kahn EI, Lev R, Manley PN, Pascal RR, et al. Pathology of the malignant colorectal polyp. *Hum Pathol*. (1998) 29:15–26. doi: 10.1016/S0046-8177(98)90385-9

44. Ren W, Yu J, Zhang ZM, Song YK, Li YH, Wang L. Missed diagnosis of early gastric cancer or high-grade intraepithelial neoplasia. *World J Gastroenterol*. (2013) 19:2092. doi: 10.3748/wjg.v19.i13.2092

45. Jass JR, Sobin LH. *Histological Typing of Intestinal Tumours*. Berlin; Heidelberg: Springer Science & Business Media (2012). doi: 10.1007/978-3-642-83693-0_2

46. Spring KJ, Zhao ZZ, Karamatic R, Walsh MD, Whitehall VL, Pike T, et al. High prevalence of sessile serrated adenomas with BRAF mutations: a prospective study of patients undergoing colonoscopy. *Gastroenterology*. (2006) 131:1400–7. doi: 10.1053/j.gastro.2006.08.038

47. Li SC, Burgart L. Histopathology of serrated adenoma, its variants, and differentiation from conventional adenomatous and hyperplastic polyps. *Arch Pathol Lab Med*. (2007) 131:440–5. doi: 10.5858/2007-131-440-HOSAIV

48. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, et al. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Acad Radiol*. (2004) 11:178–89. doi: 10.1016/S1076-6332(03)00671-8

49. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*. (1901) 37:547–79.

50. Naz S, Majeed H, Irshad H. Image segmentation using fuzzy clustering: a survey. In: *2010 6th International Conference on Emerging Technologies (ICET)*. Islamabad: IEEE (2010). p. 181–6.

51. Zaitoun NM, Aqel MJ. Survey on image segmentation techniques. *Procedia Comput Sci*. (2015) 65:797–806. doi: 10.1016/j.procs.2015.09.027

52. Dhanachandra N, Manglem K, Chanu YJ. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Comput Sci*. (2015) 54:764–71. doi: 10.1016/j.procs.2015.06.090

53. Deng H, Clausi DA. Unsupervised image segmentation using a simple MRF model with a new implementation scheme. *Pattern Recogn*. (2004) 37:2323–35. doi: 10.1016/S0031-3203(04)00195-5

54. Huang C, Li X, Wen Y. AN OTSU image segmentation based on fruitfly optimization algorithm. *Alexandria Eng J*. (2021) 60:183–8. doi: 10.1016/j.aej.2020.06.054

55. Khiyal MSH, Khan A, Bibi A. Modified watershed algorithm for segmentation of 2D images. *Issues Informing Sci Inf Technol*. (2009) 6:1077. doi: 10.28945/1077

56. Zhang H, Zhu Q, Guan Xf. Probe into image segmentation based on Sobel operator and maximum entropy algorithm. In: *2012 International Conference on Computer Science and Service System*. Nanjing: IEEE(2012). p. 238–41.

57. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. (2017) 39:2481–95. doi: 10.1109/TPAMI.2016.2644615

58. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *International CONFERENCE on Medical Image Computing and Computer-Assisted Intervention*. Berlin; Heidelberg: Springer (2015). p. 234–41. doi: 10.1007/978-3-662-54345-0_3

59. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: gated axial-attention for medical image segmentation. *In: International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer (2021). p. 36–46. doi: 10.1007/978-3-030-87193-2_4

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fmed.2022.1070072

OPEN ACCESS

EDITED BY
Jun Cheng,
Shenzhen University, China

REVIEWED BY
Jun Jiang,
Mayo Clinic, United States
Yuhan Zhang,
The Chinese University of Hong Kong,
Hong Kong SAR, China

*CORRESPONDENCE
Limei Guo
✉ guolimei@bjmu.edu.cn
Xueying Shi
✉ shixueying@bjmu.edu.cn
Lin Luo
✉ luol@pku.edu.cn

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Pathology, a section of the journal Frontiers in Medicine

RECEIVED 14 October 2022
ACCEPTED 22 December 2022
PUBLISHED 26 January 2023

CITATION
Zhang H, He Y, Wu X, Huang P, Qin W, Wang F, Ye J, Huang X, Liao Y, Chen H, Guo L, Shi X and Luo L (2023) PathNarratives: Data annotation for pathological human-AI collaborative diagnosis.
Front. Med. 9:1070072.
doi: 10.3389/fmed.2022.1070072

PathNarratives: Data annotation for pathological human-AI collaborative diagnosis

Heyu Zhang[1†], Yan He[2†], Xiaomin Wu[1], Peixiang Huang[1], Wenkang Qin[1], Fan Wang[1], Juxiang Ye[3], Xirui Huang[2], Yanfang Liao[2], Hang Chen[1], Limei Guo[3*], Xueying Shi[3*] and Lin Luo[1*]

[1]College of Engineering, Peking University, Beijing, China, [2]Department of Pathology, Longgang Central Hospital of Shenzhen, Shenzhen, China, [3]Department of Pathology, School of Basic Medical Science, Peking University Health Science Center, Peking University Third Hospital, Beijing, China

Pathology is the gold standard of clinical diagnosis. Artificial intelligence (AI) in pathology becomes a new trend, but it is still not widely used due to the lack of necessary explanations for pathologists to understand the rationale. Clinic-compliant explanations besides the diagnostic decision of pathological images are essential for AI model training to provide diagnostic suggestions assisting pathologistsŠ practice. In this study, we propose a new annotation form, PathNarratives, that includes a hierarchical decision-to-reason data structure, a narrative annotation process, and a multimodal interactive annotation tool. Following PathNarratives, we recruited 8 pathologist annotators to build a colorectal pathological dataset, CR-PathNarratives, containing 174 whole-slide images (WSIs). We further experiment on the dataset with classification and captioning tasks to explore the clinical scenarios of human-AI-collaborative pathological diagnosis. The classification tasks show that fine-grain prediction enhances the overall classification accuracy from 79.56 to 85.26%. In Human-AI collaboration experience, the trust and confidence scores from 8 pathologists raised from 3.88 to 4.63 with providing more details. Results show that the classification and captioning tasks achieve better results with reason labels, provide explainable clues for doctors to understand and make the final decision and thus can support a better experience of human-AI collaboration in pathological diagnosis. In the future, we plan to optimize the tools for the annotation process, and expand the datasets with more WSIs and covering more pathological domains.

KEYWORDS

pathology, human-AI collaboration, data annotation, multimodal data, colorectal cancer

1. Introduction

Pathological diagnosis is the gold standard for most diseases, especially oncology, and is the cornerstone of clinical treatment (1). It studies the etiology, pathogenesis, and morphological changes of tissues and drives decisions about discovering, treating, and preventing diseases. With the development of deep learning and artificial intelligence (AI) technologies (2, 3), computational pathology has made significant strides in helping pathologists with auxiliary diagnostics and increasing their productivity in smart medicine applications such as classifying tumor subtypes (4–7), detecting cancerous regions (8), and segmenting lesion areas (9–11), especially for small and easily neglected lesion areas (12).

Artificial intelligence for pathology has stimulated a growing demand for high-quality pathological image datasets. Deep-learning-based computational pathology requires model training with numerous gigapixel whole-slide images (WSIs) scanned from H&E-stained specimens and annotated with diagnostic labels (13, 14). Due to its professionalism, pathological annotation usually relies on professional pathologists and is time-consuming and costly (15). The form and granularity of annotations imply the types of potential applications a dataset can support. For example, some large-scale datasets with WSI-level weak labels are used for weakly supervised classification tasks (16–18), while some datasets with region-level annotations can support more tasks of lesion segmentation with multiclassification types or even verbal explanations (19–21). Nevertheless, existing public datasets are not directly applicable for clinical use because most focus on the ground truth labels about what the diseases and lesions are, rather than why and how they are discovered and decided. As a result, the trained AI models can hardly provide enough diagnostic explanations for pathologists to understand the rationale.

There still exist challenges in collecting why and how annotations because pathologists' diagnostic thinking logics are not well recorded and structured. Furthermore, the descriptions of a lesion's decisive morphological characteristics are not consolidated due to the diverse captioning habits of pathologists. Most importantly, interactive annotation approaches must provide a flexible and systematic experience while avoiding additional workload for pathologist annotators (22).

In this study, we propose PathNarrative, a new annotation form that can collect both diagnostic labels and rich logical reasoning data for pathological AI to better collaborate with human pathologists. PathNarratives introduces an annotation protocol for pathologists to record both the decision-layer lesions and the reason-layer decisive features of diagnostic logic. It defines a hierarchical multimodal data structure to manage the decision-to-reason labels and their relations, a narrative annotation process, and an interactive tool to support annotators working in a flexible and multimodal way with clinical tags, voice, and pencil to not only mark the lesions but also point out the relative decisive features. Meanwhile, the underlying field-of-view (FOV) moving and pausing behaviors can be recorded simultaneously to together form the hierarchical annotation. Following the PathNarratives protocol, we recruited eight pathologist annotators and built a colorectal pathological dataset containing 174 WSIs with hierarchical decision-to-reason annotations. We further conduct experiments on the dataset with classification and captioning tasks to explore the clinical scenarios of human-AI collaboration in pathological diagnosis.

The major contributions of this study are as follows:

(1) A new annotation protocol, PathNarratives, that can obtain and manage clinical-compliant fine-grain multimodality labels, diagnostic thinking logic, and decision explanations. The hierarchical data structure involves decision-layer and reason-layer labels compliant with standard pathology clinical guides. A hierarchical terminology for the colorectal tumor is also proposed. Multimodality information labels are supported for flexible annotation.

(2) A comprehensive colorectal dataset of gigapixel WSIs with fine-grained annotations following PathNarratives was constructed. Each WSI involves the decision-to-reason hierarchical labels and the multimodality information.

(3) Exploration of the application scenarios of the PathNarratives colorectal dataset in diagnosis and experiments results show that finer labels improved performance in the classification and capitalization tasks. The explainable results supported doctors' efforts to better understand and experience human-AI collaboration in pathological diagnosis.

The rest of the study is arranged as follows. Section "2 Related study" the related study on datasets, narrative annotation, and relative AI applications. Section "3 Data annotation protocol" introduces the pathological data annotation protocol. Section "4 Dataset" presents the annotated colorectal dataset. Section 5 "Classification and captioning tasks on narratives-annotated dataset" shows the application scenarios and experiments on the dataset. Section "6 Conclusion" concludes and discusses future study.

2. Related study

2.1. Pathological datasets

Some pathology datasets are typically weakly labeled with simple metastatic disease circled at the WSI level and only applied to a single decision scenario (23–27). For example, CAMELYON16 (23) and CAMELYON17 (24) datasets have been widely used in research for automated detection and classification of breast cancer to enable automated evaluation of patient staging while reducing the subjectivity of the diagnosis. Similarly, the authors compiled TCIA (25) containing clinical information from epithelial ovarian cancer (EOC) and peritoneal serous papillary carcinoma (PSPC) to explore and develop methods for predicting the therapeutic effect of bevacizumab in patients with EOC and PSPC. The breast cancer dataset BreCaHAD (26) divides WSIs into six tissue classifications including mitosis, apoptosis, tumor nucleus, non-tumor nucleus, tubule, and non-tubule, to support multiclassification tasks. Another breast cancer dataset, BreaKHis (27), is designed for baseline classification of tumor benign-malignant and discrimination of subtype characteristic tissues. These dataset annotations only stay at the decision level of the metastatic region; the granularity is not detailed and persuasive enough.

Several pathological datasets aim to provide better clinical captioning to reflect pathology reports in computational pathology (19–21), including two categories. One was taken from existing digital resources, such as pathology textbooks and clinical and research journal article databases, which are typically represented by PathVQA (19) and ARCH (20). Such datasets are massive in volume but low in acquisition cost, poor in quality, and inconsistent in standards. These two datasets are often used for pre-training representational learning. During compilation, PathVQA also emphasizes templated and open-ended generation of visual question answers, compared to ARCH's extracted image and image-related text pairs. Another type is obtained by picking patches from WSI, such as PatchGastricADC22

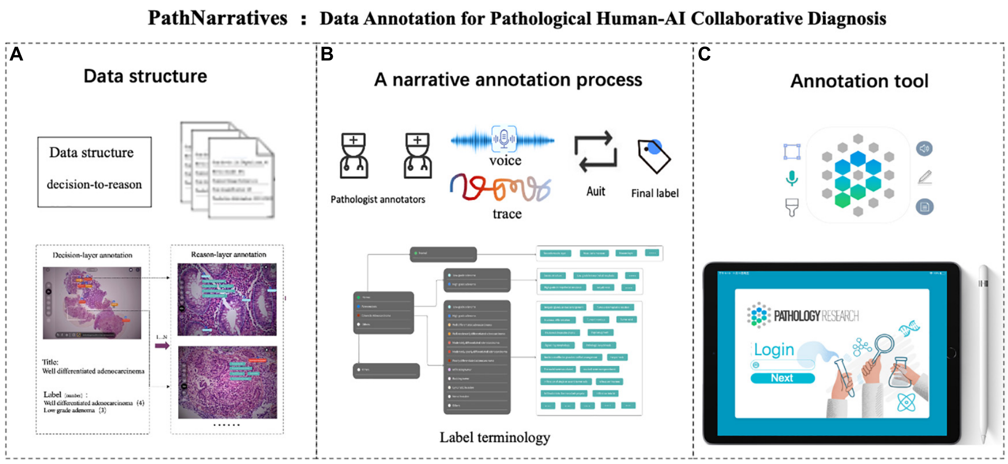


FIGURE 1
PathNarratives protocol including a hierarchical fine-grain data structure and a multimodal annotation process with an interactive annotation tool. **(A)** Decision-to-reason data structure. **(B)** Narrative annotation with label terminology. **(C)** Interactive annotation tool: ParVis.

(28) and BCIDR (21). Among them, PatchGastricADC22 is derived from the actual clinical case diagnosis reports from the same hospital. Each instance has two magnifications, so the quality and resolution are consistent. Each WSI contains unorderly collected patches. Patches that belong to the same WSI have the same caption. Since there are only independent patches, there is no way to understand the mutual reasons for different patches in the doctor's diagnosis. BCIDR allows more pathologists to participate in the annotation. The patches are extracted from eight typical regions and added captions, which makes their captions more focused on the detailed information at the cellular level. Thus, all of these datasets do not focus on region-level reasonable diagnostics. PathLAKE (22) proposes an annotation best practice that includes hierarchical case-level, region-level, and cell-level labels on breast cancer annotation but does not take the doctors' diagnostic logic or the experience of multimodal inputs into consideration.

2.2. Narrative annotation model

Narrative annotation focuses on the description of the relationship between entities, and entity relationships are collected during the annotation phase. Attributes, relationships, and entities in the same image are often closely related (29–32). Localized Narratives (30) connect vision and language by artificially using mouse scribing to join action connections between entities and make the captioning in content more hierarchical. It asks annotators to describe an image with their voice while simultaneously hovering their mouse over the region they are describing. Using this mouse trajectory and voice inputs, the narrative dataset performs better in the caption task. Similarly, TReCS (31) exploits using detailed and reasonable language descriptions paired with mouse traces to generate images. More realistic images could be generated using descriptions and traces compared to those without traces. The interactions and relationships between objects contribute to a visual understanding of the main components of object-centric events (33). MITR (32) shows a framework to jointly model images, text, and human attention traces, which connects what to say with where to look by modeling human attention traces. The process of narrative annotation also contains helpful information in essence. By exploring the visual attention of doctors browsing and the process of scanning trajectories, Chakraborty et al. (34) found there are strongly correlated between the feature regions of algorithm tasks and lesions in the image to a certain extent, which reflects their diagnostic logic. The annotators draw the object's bounding box with the mouse and add class labels through voice. Significant speed gains are achieved while maintaining high-quality annotations (35). In addition to manually adding entity relations during the annotation process, the models for video action recognition can also be considered partially auto-generating narrative relations of the entity bounding boxes (36–38).

2.3. Applications of AI in pathology

Medical classification and segmentation have also actively been explored (39–45). Gurcan et al. (39) reviewed pathological image analysis methods for computer-assisted diagnosis, including pretreatment, nucleus and gland segmentation, feature extraction, and classification. Veta et al. (40) discussed histological image analysis methods for breast cancer and conducted additional discussions on mitosis detection and proliferation assessment.

TABLE 1 Basic information of participating pathologists.

Pathologist	Years-of-working	Subspecialty
P1	More than 15 years	Histopathology
P2	3–5 years	Histopathology
P3	3–5 years	Histopathology
P4	5–10 years	Digestive
P5	10–15 years	Digestive
P6	10–15 years	Histopathology
P7	3–5 years	Histopathology
P8	More than 15 years	Histopathology
P9	3–5 years	Digestive

Luo et al. (42) combined the characteristics of tumor cells and their surrounding organizational form environment to predict patient survival outcome information experimentally. HAG (43) was proposed to fuse multiresolution information and speed up prediction without reducing accuracy. Abu Haeyeh et al. (44) normalized the staining of RCC and used a weakly supervised multi-instance learning method. The results show that they can classify benign-malignant and determine tumor subtypes to support medical treatment management. Zhou et al. (45) chose TCGA, combing features at different magnifications, to achieve the classification and localization of colorectal tumors.

Pathological captioning tasks are being studied recently to automatically generate diagnostic texts based on patient medical images, assist inexperienced doctors, and reduce clinical errors (46). The typical representative is still PathVQA (19). PathVQA first reviews related research in medical radiology, such as VQA-Med (47) and VQA-RAD (48), and then explores the experiments of vision questions and answers tasks in pathology. The PathVQA automatically generates what, why, and other question-answer pairs to conduct the learning model by extracting pathological images and corresponding text information. In contrast to PathVQA, PatchGastricADC22 extracts patches from endoscopic biopsy specimens of gastric adenocarcinoma and trains an attention-based pipeline model to predict image features. The physician diagnostic logics of WSIs or lesion regions have not been extensively explored in the caption task at present.

3. Data annotation protocol

3.1. Overview

We first analyzed the clinical routines of pathological diagnosis to formulate the annotation data structure and the protocol of PathNarratives, as shown in **Figure 1**. To be specific, we consulted the WHO pathological clinical guideline (49), analyzed the pathology report templates from the pathology departments of two top-tier hospitals, and observed two pathologists for their diagnosis browsing and thinking practices with permission (P4 and P9 in **Table 1**). The goal was to explore how pathological decisions are made, explained, and concluded into reports, and what granularity of interpretable annotations can be collected in a natural process.

We then defined the PathNarratives protocol, which includes a hierarchical decision-to-reason data structure, a multimodal annotating process, and an interactive annotation tool. It allows annotators to work in a flexible and multimodal way to mark and circle lesion areas, look for typical characteristics and outline them, and describe the basis of judgment, by using clinical tags, voice, pencil lining, and FOV moving. Following this, the collected data can cover the types of diagnostic disease and lesion, the decisive morphological features, and the corresponding pathologists' logical narrations and viewing behaviors.

3.2. Data structure

Decision-to-reason annotation

Concluding a pathological diagnosis report involves two layers of information. The decision-layer information is about the slide-wise diagnostics (one report may involve several slides of the patient) and descriptions of lesion regions that appear explicitly in the pathology report. In contrast, the reason-layer information demonstrates the underlying typical features and reasons that pathologists use to judge the lesion and diagnose it. Although the reason-layer information is essential to explain the rationale, it is usually implicit in pathologists' knowledge systems and does not show in the report. Only when pathologists discuss with other doctors will they refer to both the decision-layer and reason-layer information of the diagnosis, using multimodal ways such as texts, voice, screenshots, and mouse/pencil moving.

Besides the two layers of information, we discovered that doctors' behaviors such as browsing, view zooming-in/out, view shifting, view pausing, and mouse/pencil hovering represent their attention focus and thinking logic during the pathological diagnosis process. Such behavior data also provide informative inputs for AI learning and, therefore, are also considered in our data structure.

The decision-to-reason data structure to manage the hierarchical multimodal annotation is shown in **Figure 2**. The decision-layer represents the labels around WSIs and lesion regions, where each WSI can involve multiple lesion regions (one-to-many mapping, shown as 1. . .N in **Figure 2**). The reason-layer is related to the corresponding multiple features labeled with descriptions to explain the rationale behind judging each lesion decision (one-to-many mapping, shown as 1. . .N). Multimodal annotations are supported as clinical tags, free texts, voice, and pencil/mouse moving traces of the doctor's annotating behaviors, which are timestamp synchronized and associated with both the layers of data (many-to-many mapping, shown as N. . .N). Multiple annotations together form one comprehensive pathology report (many-to-many mapping, shown as N. . .N).

Unified terminology

We also considered the need for unified terminology of the two layers of labels in the data structure design, where the colorectal tumor is chosen in this study. During the pathological shadowing, we found that if we allowed two pathologists to input free-text reasoning labels, their expressions could vary severely even when they agreed on the tumor types and reasons for the same lesion of a colorectal WSI. For example, pathologist 4 (P4 in **Table 1**) described the features as a "gland fused with a sieve," while Pathologist 9 (P9 in **Table 1**) described the same one as a "sieve hole." Further interviews with the two doctors proved that they meant the same thing, though their textual expressions looked quite different. The variability of labels affects not only the performance of the AI model but also the normalization of data, and therefore, unified terminology is necessary.

We analyzed pathological books, published specifications, and pathology report templates from hospitals and consulted senior pathologists (P1 and P9 in **Table 1** with more than 15 years of diagnostic experience) to build the decision-to-reason unified terminology, as shown in **Table 2** (refer **Supplementary material** for the full version). We first referred to the 2019 WHO Blue Book (World Health Organization) (50), which defines the classification of digestive system tumors and borrowed the colorectal classification terms to form the overall classifications as "normal, adenocarcinoma and adenoma." Besides the WHO Blue Book, comprehensive pathology report templates from two top-tier hospitals in China are also considered to further define the finer classification of the decision-layer label, e.g., "Adenocarcinoma"

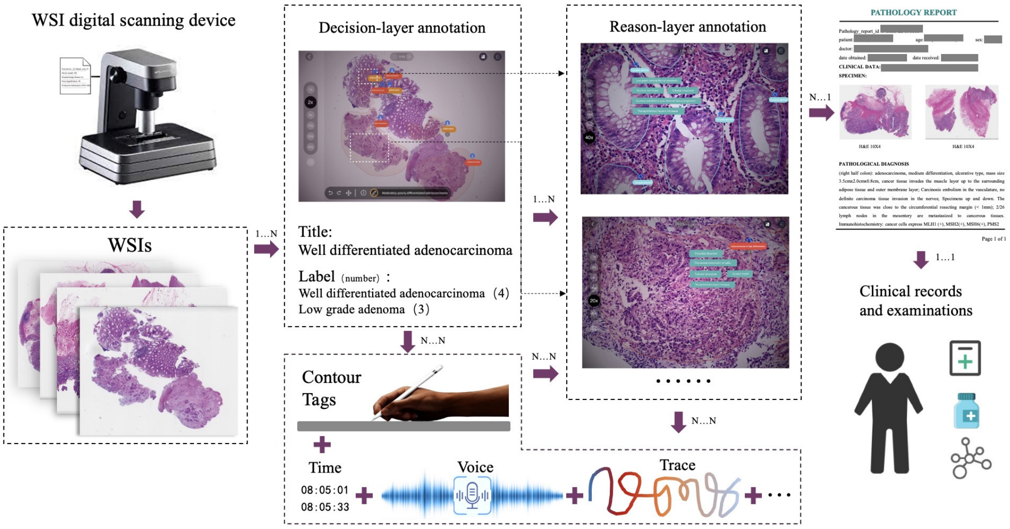

WSI digital scanning device
WSIs
1...N
Decision-layer annotation
Title:
Well differentiated adenocarcinoma
Label (number) :
Well differentiated adenocarcinoma (4)
Low grade adenoma (3)
N...N
Contour
Tags
Time
08:05:01
08:05:33
Voice
Trace
1...N
N...N
Reason-layer annotation
......
N...N
N...1
PATHOLOGY REPORT
Pathology_report_id [redacted]
patient: [redacted] age: [redacted] sex: [redacted]
doctor: [redacted]
date obtained: [redacted] date received: [redacted]
CLINICAL DATA: [redacted]
SPECIMEN:
H&E 10X4
H&E 10X4
PATHOLOGICAL DIAGNOSIS
Page 1 of 1
1...1
Clinical records
and examinations

FIGURE 2
The decision-to-reason multimodality data structure of PathNarratives. The decision-layer represents the labels around WSIs and lesion regions, where each WSI corresponds to multiple lesion regions. The reason-layer represents the corresponding multiple features marked and described to judge the lesion regions.

in the classification is categorized into subtypes such as "Poorly differentiated adenocarcinoma" and "Moderately differentiated adenocarcinoma." In addition, some terms that frequently occur in pathology reports describing features of lesions, such as "Tumor invasion," "Tumor budding," "vascular invasion," and "nerve invasion," are also set as decision-layer labels to better accommodate pathologists' habits and clinical needs.

Reason-layer label terminology was designed under the decision-layer labels. As the WHO book and pathology reports do not involve detailed reasoning information, we invited the senior pathologists to summarize the main features into the reason-layer annotation description from textbooks (51) with consideration of the decision labels and pathology reports. As shown in Table 2, "Poorly differentiated adenocarcinoma" in the decision-layer is further associated with detailed reason-layer labels describing diagnostic features such as "Irregular arrangements of glands" and "Mucinous differentiation." Specifically, the decision-layer labels under the "Normal" category are used to describe normal colorectal elements such as "Fatty tissue," "smooth muscle," and "Lymphatic vessel." The terminology terms are ordered from histomorphology to cell morphology for pathologists' convenience in browsing and selecting from it.

3.3. Annotation process and tool

The PathNarratives annotation process includes a coarse-grain phase and a fine-grain phase that follow the decision-to-reason labeling structure. The design of the two phases is to accommodate the different clinical application needs such that in the coarse-grain annotation phase, an annotator browses a WSI and circles large lesion areas to tag with the classification labels and then makes a preliminary slide-wise diagnosis description, as shown in Figure 3A. This annotation phase can be completed quickly by doctors and an overview diagnosis can be provided. Then, in the fine-grain annotation phase, an annotator needs to circle the finer subtype decisions of lesions with typical features as completely as possible and explain the decisive reasons. They can use a decision-layer subtype label pencil to circle the typical lesion features, and then either attach

TABLE 2 Label terminology partial (in total, there are 3 classification labels, 12 subtypes, and 77 reason-layer labels).

Classification label	Decision-layer subtype label	Reason-layer label
Adenocarcinoma	Poorly differentiated adenocarcinoma	Irregular arrangement of glands Mucinous differentiation Vacuolated nuclei ...
	Moderately differentiated adenocarcinoma	
	...	
	Tumor invasion	Infiltration of single or several tumor cells Invasion into the muscularis mucosae ...
	Tumor budding ...	Tumor budding (grade 1) ...
Adenoma	Low-grade adenoma	Low-grade intraepithelial neoplasia Glands lack mature differentiation ...
	High-grade adenoma	...
Normal	Normal	Fatty tissue Smooth muscle Lymphatic vessel ...

Specifically, adenocarcinoma is mapped to 9 decision-layer subtypes and 34 reason-layer labels; adenoma is mapped to 2 decision-layer subtypes and 25 reason-layer labels; normal is mapped to 1 subtype and 18 reason-layer labels.

reason tags or record voice explanations to explain the diagnostics. The decision and reason labels can be directly picked from the predefined label terminology, as shown in Figures 3B, C. The fine-grain phase is more sophisticated and requires more time and labor. Images and annotations can be replayed, compared, and audited afterward, as shown in Figures 3D–F, respectively.

The above annotation process is carried out using our self-developed software ParVis for the convenience of pathologist annotators, auditors, and project managers cooperating on an annotation project. The software comprises a mobile client for doctors' daily annotation/audit and a web server for annotation project management. Administrators create projects, upload pathology images, set roles and access rights, and manage terminologies through the web server. Pathologists use the mobile client to join projects, submit annotations, review them, and audit the results.

According to the annotation process, ParVis has four major functions: label, playback, review, and audit. On the label module interface in Figures 3A–C, a pathology annotator can start labeling a WSI for coarse annotation of the slide-level diagnosis description and use different colors of classification pencils to mark lesion area contours as in Figure 3A. For further fine-grain annotation, ParVis provides different colors of subtype pencils for the annotator to circle the contours of typical lesion features as in Figure 3B, and the icons of "mic" or "tag" can be clicked to describe the features with voice or text to generate the reason-layer labels in Figure 3C. In addition to colors, the pencil tool supports flexible shapes for marking lesion areas, such as "curve," "rectangle," or "brush." ParVis also provides a "ruler" to measure the area size according to the needs of pathological reports. The fundamental functions such as magnification rate, eagle view, screenshot, location, and metadata view are also provided as basic functions.

ParVis forms the structural multimodal annotation data for further analysis, playback, review, and audit. It also periodically records the timestamps of browsing and moving behavior events during labeling (with doctors' prior permission) for further synchronization. The behavioral tracking includes events such as "FOV center change," "voice recording," "magnification," "pencil switching," "undo," and "delete" over time during the doctor labeling process. These data can support application modules of playback (to replay the annotation process), comparison (for medical students to review and learn from multiple experts or teachers to examine multiple Students' work simultaneously), and audit (for auditors to review and refine the annotations), as shown in Figures 3D–F. Most importantly, the synchronized events such as magnification and focus center shifting implicitly recorded can be used to analyze physician behaviors. For example, visualizing the FOV center trajectory shows the length of stay is positively correlated with the difficulty of the lesion area, which is consistent with the conclusion in Wang and Schmid (37). Behavioral data indicate the logical thinking of doctors and their attention to assist the interpretability of AI.

The audit is an essential step for the annotation process to ensure data quality and consistency, which needs to be conducted by senior pathologists. The ParVis audit module is designed following the general practice of the pathology department. A senior pathologist clicks the Audit button and selects the items marked by primary pathologists and checks for missing or wrong annotations. If there is a problem, they need to revise, add, or delete the labels to finalize the submission. We use Kappa, Dice, and BLEU to evaluate the consistency of different levels of annotations in section "4.1 Data source and overall statistics."

During the annotation practice, we kept optimizing the process according to observed issues. One important issue is the cost of fine-grain annotation to label all the reasoning tags, which is tedious and expensive for pathologists even though it provides more details and explanations. Since many adjacent glands or lesions share similar characteristics, we added a "Bundle pencil" tool to support annotators to circle adjacent lesion regions of similar reasoning tags, so that a pathologist can simply apply a one-off description to all the lesions and features within the bundling circle. This setting saves annotation time to a considerable degree in practice.

4. Dataset

4.1. Data source and overall statistics

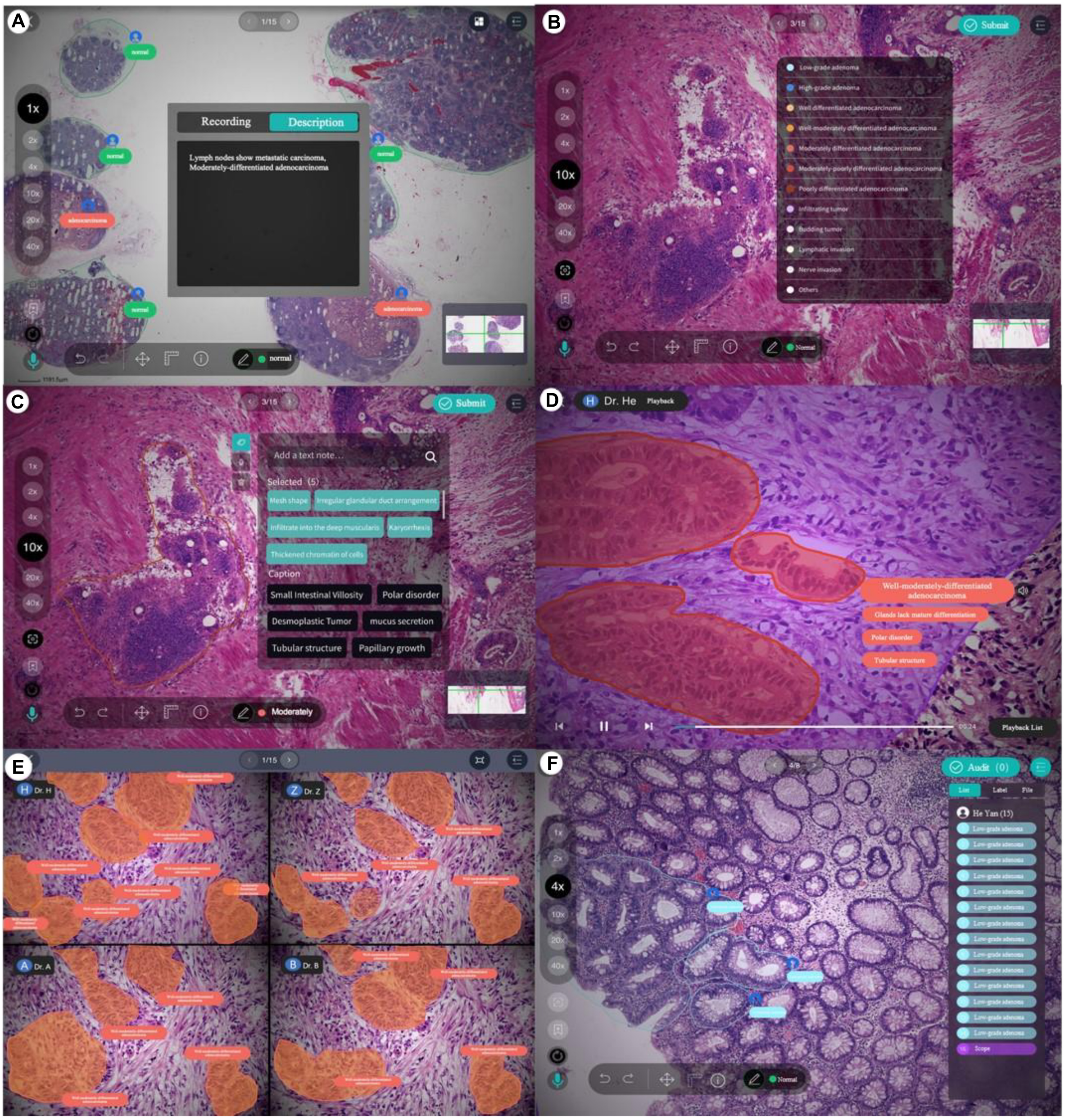
Based on the PathNarratives protocol, we recruited eight pathologist annotators (P1–P8 in Table 1) to build a colorectal tumor dataset, CR-PathNarratives, which includes 174 annotated colorectal WSIs with a length of 8,000–90,000 pixels and width 6,000–60,000 pixels, all with the decision-to-reason and multimodal data structure.

We selected colorectal cancer because it is characterized by high incidence and mortality. Colorectal cancer has become the second leading cause of cancer death worldwide, with 930,000 deaths in 2020. In 2020, the new incidence rate of colorectal cancer in China was 12.2% and the fatality rate of colorectal cancer was 9.5% (52). In addition, colorectal tissue sections present explicit morphological variance and cover wide categories of tumor types with well-established pathological diagnostic guidelines and standards for database design and practice.

The WSIs were obtained from one first author's cooperative hospital with approval. The chief pathologist selected 891 H&E-stained slides from 300 patients and randomly sampled 300 pieces to scan into WSIs at 20X objective magnification. At present, the collection of annotated data containing 174 WSIs has been completed.

We conducted the basic statistics of CR-PathNarratives on the distributions of classification types, decision-layer subtype labels, reason-layer labels, labeled areas, and diagnostic captions composed with reasoning labels. The dataset covers all three class types: adenocarcinoma, adenoma, and normal. The detailed categories and numbers are shown in Table 3.

Each WSI contains a simple overall caption, several decision-layer labels, and tens to hundreds of reason-layer labels. In total, in 174 WSIs, 108 contain adenocarcinoma areas ranging from well differentiated to poorly differentiated, 38 contain adenoma areas, 17 contain both adenoma and adenocarcinoma, and 45 are normal slides with only normal areas labeled. There are in total 11 types of decision-layer labels and 75 reason-layer labels, including free-text tags. For the whole dataset, there are 23,532 regions manually circled, and some are grouped as 539 bundles in total (a bundle consists of multiple or single regions sharing the same features and captions, which can effectively reduce the labeling efforts, as mentioned in section "3.3 Annotation process and tool"). In total, there are 878 different kinds of captions associated with all the labeled regions, and each caption comprises 4.4 label terms on average (max = 19 and min = 1), as shown in Figure 4F.



Lymph nodes show metastatic carcinoma, Moderately-differentiated adenocarcinoma

FIGURE 3
Decision-to-reason annotation functions of ParVis. **(A)** Label module: Lesion area circled on WSI by an annotator in the coarse-grain annotation phase, with preliminary diagnosis descriptions. **(B)** Label module: Decision-layer subtype labels as different colors of pencils in the fine-grain annotation phase. **(C)** Label module: Reason-layer features circled and labeled by clicking on the reason-layer terminology tags or recording voice explanations, in the fine-grain annotation phase. **(D)** Playback module to replay the annotation events on the WSI image which is structural and can be searched and analyzed. **(E)** Comparison module to view and compare different doctors' annotations. **(F)** Audit module for senior doctors to review and correct previous annotations.

Whole-slide image-wise statistics show that on average, a WSI contains 3.1 labeled bundles (max = 41 and min = 1) that reflect 135.2 regions. For further AI algorithm computation, each WSI scanned at 20 × magnification was cut into patches of 256*256 pixels. In statistics, the averaged labeled regions contain 76 patches (the diversity ranged from max = 2,477 to min = 1). On average, one WSI is associated with 8.93 different kinds of captions (max = 40 and min = 1) and involves 12.03 reason label terms (max = 42 and min = 1).

We also investigated the texts and captions frequently used in annotation statistics. The most commonly used label terms are "Stratified or pseudostratified arrangement of nuclei" (7.21%), "Rod-shaped nuclei" (7.02%), "Increased layers of epithelial cells" (5.78%), and "Chromatin condensation of cells" (5.50%). For reason-layer labels, the most commonly used captions are "Mitosis visible, mucous differentiated, vacuolated nuclei," "Markedly reduced cytoplasm, stratified or pseudostratified arrangement of nuclei, increased layers of epithelial cells, rod-shaped nuclei, oval nucleus," and "Cribriform."

We also evaluated the consistency of doctors' annotations for the quality of the datasets. For 10% of the annotated samples (18 WSIs), we asked a senior doctor P4 to review and label the same WSIs annotated by a senior doctor P5 and a junior doctor P2. Three levels

TABLE 3 Subtype distribution and data scale table.

Classification	Decision subtype	Number of WSIs in the subtype	Total
Adenocarcinoma	Well differentiated	20	108*
	Poorly differentiated	23	
	Moderately differentiated	26	
	Well-moderately differentiated	16	
	Moderately-poorly differentiated	23	
Adenoma	High-grade adenoma	25	38*
	Low-grade adenoma	13	
Normal		45	45

*Indicates that 17 lesion slides contain both adenocarcinoma and adenoma.

of annotation consistency are analyzed as shown in Figure 4 (WSI number sorted by their consistency value for illustration): consistency of WSI classification in (c), (d), consistency of lesion regions for coarse-grain classification labels in (c) vs. fine-grain subtype labels in (d), and consistency of reason descriptions of lesion features in (e), measured with the Kappa, Dice, and BLEU values, respectively. For the consistency of WSI classification, the types decided by both doctors are all the same for the 18 WSIs, which achieves an overall Kappa = 1. For the consistency of lesion regions, the patch-level classification labels and decision subtype labels achieve an average Kappa of 0.91 (max = 1, min = 0.66) and 0.85 (max = 1, min = 0.42), respectively, while the pixel-level consistency of the same-label lesion area achieves Dice values of 0.96 (max = 1, min = 0.85) and 0.92 (max = 1, min = 0.61) for classification and subtype labels, respectively. Both the patch-level Kappa value and the Dice value are with an average beyond 0.85, and the variance among different WSI is considered due to the difficulty levels of different cases. For consistency of reason descriptions represented by lesion caption, the BLEU1 value is mostly beyond 0.4 with an average of 0.78, as shown in Figure 4E.

Annotation auditing is widely used in clinical practice. When inconsistency occurs, the primary annotator needs to double check, and if there is still a dissenting opinion, the senior and primary annotators need to communicate with each other to achieve a consensus.

4.2. Decision-to-reason annotation

The two layers of decision-to-reason data are shown as examples in Figure 5. A doctor would rather look at the typical reason-layer features first to quickly conclude the diagnosis and lesion areas, and then spend much more time explaining with subtype details, typical features, and reasons. For example, the doctor looked at the lesions on a WSI that present visual features such as "Cribriform," "nucleus stratified or pseudostratified arrangement," and "polar disorder" and then quickly marked the whole WSI as "moderately differentiated adenocarcinoma" and circled two adenocarcinoma regions and one adenoma region. Then they refined to circle more reasoning feature regions and select the detailed reason-layer labels for fine-grain annotation.

Artificial intelligence training requires the annotations to be as complete as possible. Coarse-grain labeling is simpler and costs less time because doctors roughly scan the lesions and add labels to the low-resolution WSI, which takes only tens of seconds. In contrast, though it contributes necessarily detailed reasoning information, fine-grain labeling inevitably takes a longer time in marking all the circles and label terms. Experiments show the time of coarse-grain labeling per WSI is on average 1.7' as shown in Figure 4G, ranging from 0.29' to 2.97', while the time spent for fine-grain labeling is on average 46.17', ranging from 14.69' to 98.83' as shown in Figure 4H, which is 20+ times of that for coarse-grain one.

Fortunately, by applying the proposed "Bundle pencil" to group similar small lesion regions for the one-off application of the same labels as shown in Figure 5C, the fine-grain annotation time can be significantly reduced down to 1/6–1/2 of the original one. We also found it uses more time for the doctor to label adenomas than to label adenocarcinomas because the lesion areas of adenocarcinomas are often tangled and cannot be labeled separately. It also took much time to zoom back and forth to inspect a large lesion area and label all the typical details at different views. Based on this finding, we proposed the following methods to further reduce the burden of doctors. (1) Use the "Bundle pencil" to circle lesion areas with similar features and (2) Future exploration of AI technologies to provide automatic hints for circling and labeling.

Taking the WSI shown in Figure 5 as an example, the WSI was marked with 12 adenocarcinoma areas, 9 adenoma areas, and an overall cost of 1'12" for coarse-grain labeling, and the adenoma was described with the text "Low-grade intraepithelial neoplasia." During fine-grain annotation, the doctor marked 83 well-differentiated adenocarcinomas, 45 low-grade adenomas, and added 8 bundle tags, which overall cost 7'42". In another example case, annotating a WSI takes a doctor 12" to circle 3 lesion regions with classification labels, while annotating the fine-grain 488 typical features with diagnostic reasons take up to 31'24" for no-bundle-circle annotation vs. about half of it for bundle-circle annotation. In contrast, by simply applying the "Bundle pencil" to group similar small lesion regions and one-off label them, the annotation time is significantly reduced to 14'52", which is less than half of the previous time.

4.3. Multimodal data

Besides decision-to-reason data, CR-PathNarratives also covers multimodal annotation data. Each WSI in the PathNarratives dataset has visual information on the image feature regions and language information of the physician's annotations described in section "4.2 Decision-to-reason annotation." On the contrary, the PathNarratives

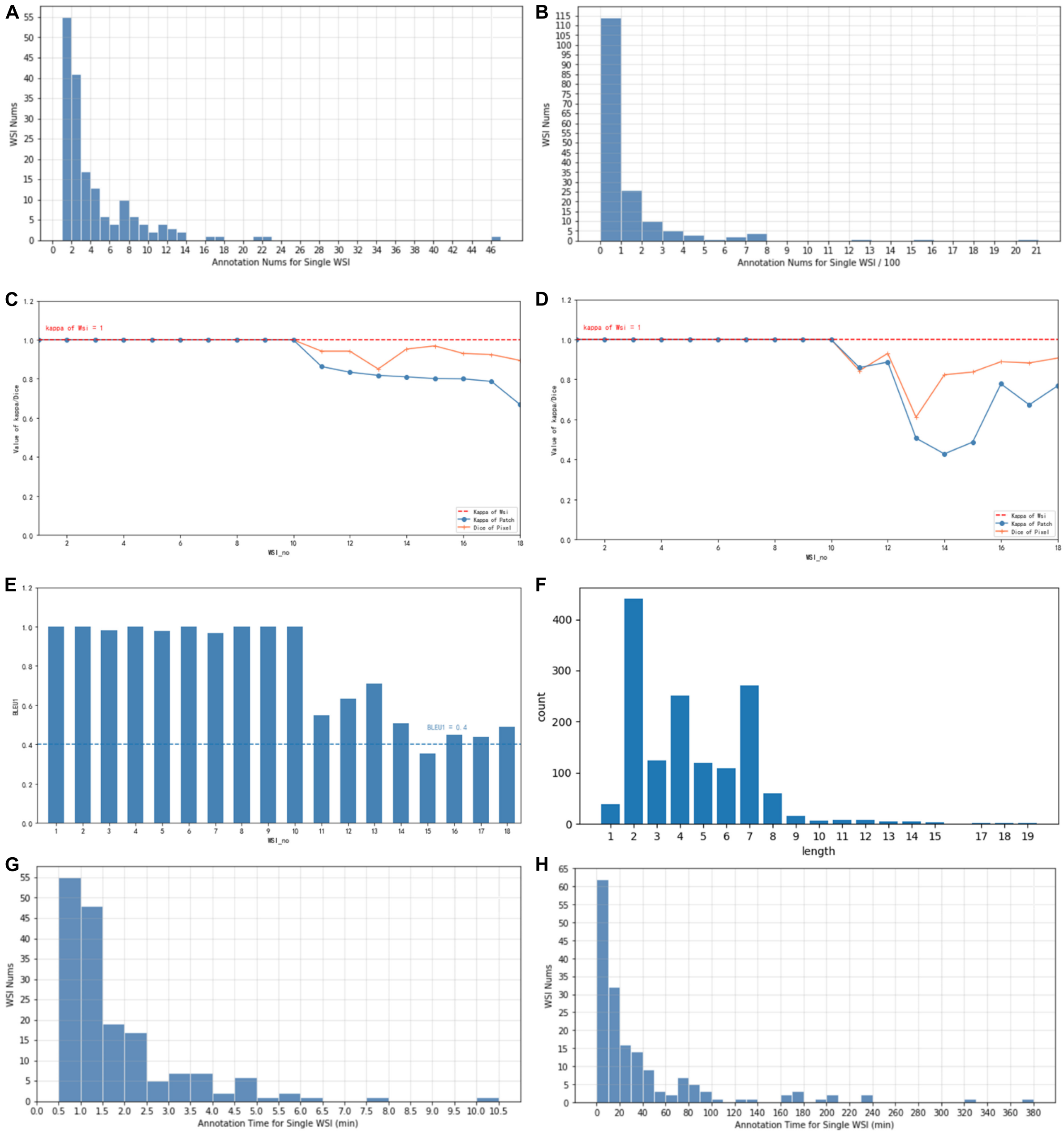
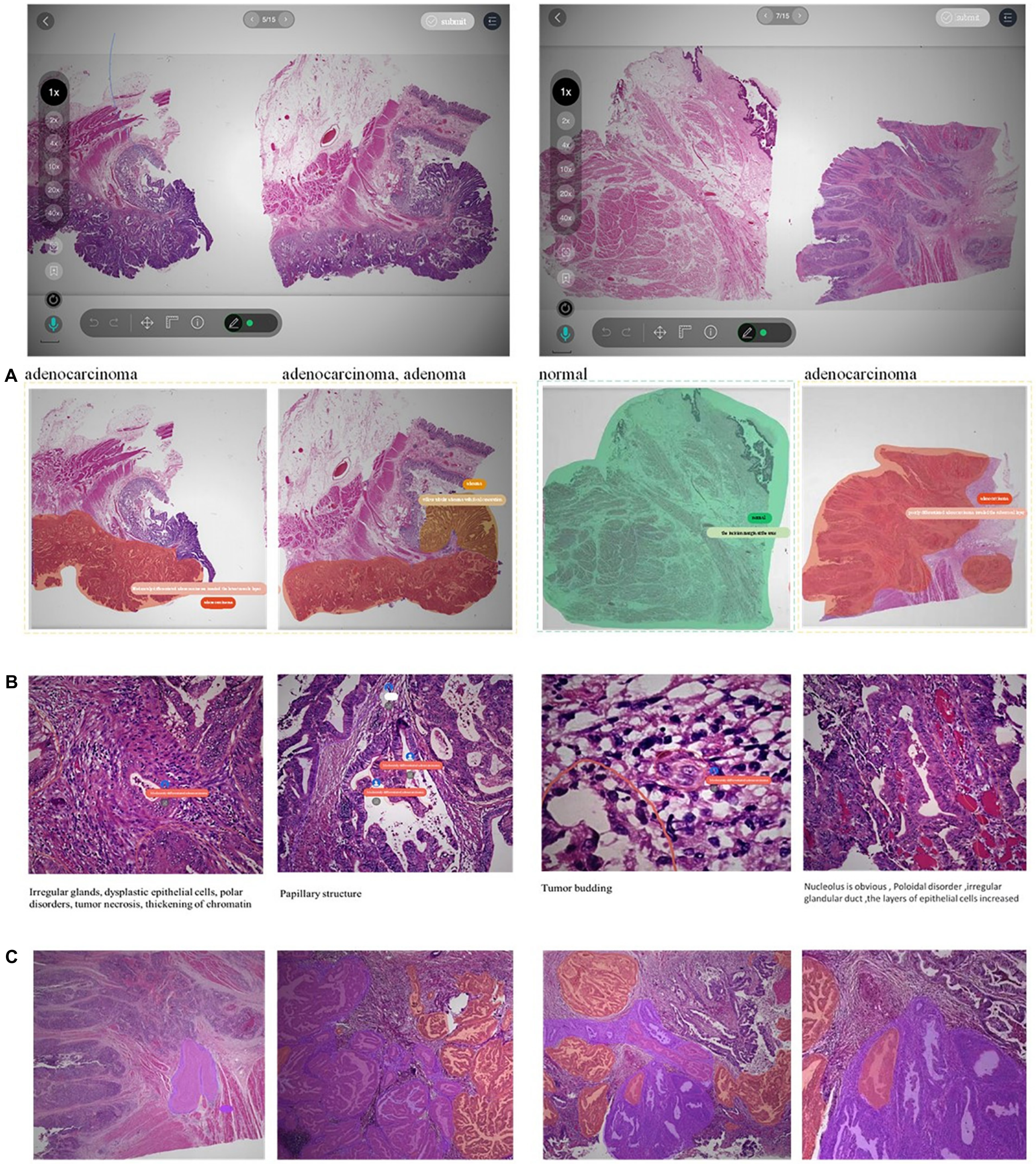


FIGURE 4
(A) The typical number of annotations contained in a single WSI of coarse-grain phase. **(B)** The typical number of annotations contained in a single WSI of fine-grain phase. **(C)** Consistency of different annotators with Kappa and Dice values of coarse-grain annotation data. **(D)** Consistency of different annotators with Kappa and Dice values of fine-grain annotation data. **(E)** Consistency of different annotators with BLEU1 of description data. **(F)** Distribution of caption length and number. **(G)** The time spent on a single WSI annotation of the coarse-grain phase. **(H)** The time spent on a single WSI annotation of the fine-grain phase.

dataset also contains voice information and behavioral trajectory information, according to doctors' preferences. From the example shown in Figure 6, we found that voice information mainly consists of the following two types of purposes: explaining diagnosis by thinking or labeling *via* voice. We observed that after his annotation, the doctor turns on the voice record button and tries to elaborate on his observation for teaching purposes, e.g., "Open the whole WSI and find that the right side is somewhat abnormal. Click to enlarge and observe to confirm the adenocarcinoma. On the left side, there are irregular glandular and tubular arrangements and invasion of the muscle layer." Junior physicians can replay and listen to learn the voice-input recordings about WSI colorectal diagnostic methods, which shares similarity to the AI learning process. The voice-transcribed text labels contain richer information among the marked areas and complement the textual label terms. However, our experiment does not involve the special natural language processing needs for pathological text recognition, which is an in-depth research area. Instead, we only recruited human medical students to perform that transcribing tasks.

The behavior-tracking data of doctors are stored in a structured time-series record of labeled behaviors such as time stamps, visual field centers, magnifications, labeling tools, toggle label colors,



Irregular glands, dysplastic epithelial cells, polar disorders, tumor necrosis, thickening of chromatin

Papillary structure

Tumor budding

Nucleolus is obvious , Poloidal disorder ,irregular glandular duct ,the layers of epithelial cells increased

FIGURE 5
Examples of decision-to-reason annotation data. **(A)** The annotation of coarse-grain phase. **(B)** The annotation of the fine-grain phase. **(C)** The annotation of fine-grain phase with bundle label.

markers, coordinates, deletions, and modifications during their labeling process. When a doctor labeled a WSI, we continuously recorded his FOV window changing, visual scan path, and resolution zoom in and out information. We visualized the doctor's attention distribution of diagnosis by aggregating the pixels of the doctor's viewport boxes, combining them with the center points of the viewport boxes, checking the time, zooming into incorporating scan path, and plotting a behavioral trajectory heatmap as shown in Figure 4, 5. The attention heatmaps echo the areas that the doctors observed the most with higher heat scores. In comparison, tracks of junior physicians demonstrate more back-and-forth browsing and reluctance than those of the senior pathologists who are experienced to make diagnoses rapidly.

5. Classification and captioning tasks on the narratives-annotated dataset

To investigate the potential clinical applications that the CR-PathNarratives dataset can support, we selected a classification task

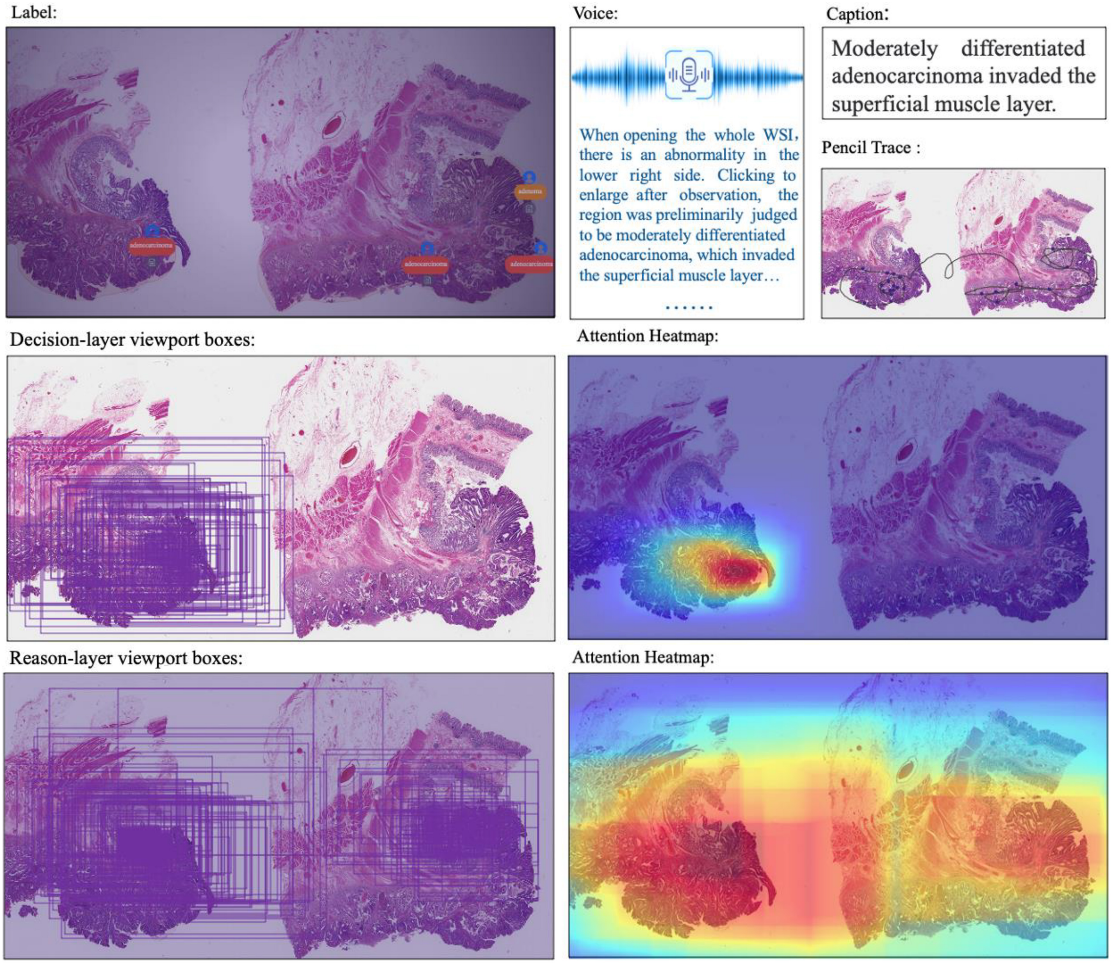

Label:
adenocarcinoma
adenoma
adenocarcinoma
adenocarcinoma
Voice:
When opening the whole WSI, there is an abnormality in the lower right side. Clicking to enlarge after observation, the region was preliminarily judged to be moderately differentiated adenocarcinoma, which invaded the superficial muscle layer…
……
Caption:
Moderately differentiated adenocarcinoma invaded the superficial muscle layer.
Pencil Trace :
Decision-layer viewport boxes:
Attention Heatmap:
Reason-layer viewport boxes:
Attention Heatmap:

FIGURE 6
Multimodal information in the dataset: WSI, annotated information, voice, caption, ratio of decision-labeled and center point trace of the viewport, ratio of detailed labeled, and center point trace of viewport and corresponding generated heat map.

and a captioning task and trained the baseline AI models. We also conducted an evaluation of Human-AI collaboration experience to explore the doctor subjects' trust and acceptance when being provided with comprehensive decision-to-reason suggestions by AI models. The experimental baseline AI model is shown in Figure 7.

Task 1: Classification of coarse-grain and fine-grain labeling data

Task definition

Given a WSI with coarse-grain classification labels vs. fine-grain subtype labels defined in Table 2, the goal is to compare their performances of classification (normal, adenocarcinoma, and adenoma) to explore the impact on different levels of labeling details. For ideal clinical use, false negatives should be avoided, which means a WSI containing adenocarcinoma should not be misjudged as an adenoma or benign case.

Methods

Each WSI is assigned a universal ID. We used the OpenSlide tool (53) to extract patches of 256*256 pixels from WSIs at 20 × magnification. Macenko stain normalization (54) is used for pre-processing to ensure uniform WSI quality. The OTSU algorithm (55) is used to separate foreground and background, ensuring that all valid patches come from the foreground tissues.

The training and test sets are first divided into the WSI grade to avoid patches from the same patient being included in both sets. The total cropped tissue patches for training were counted, where a patch is regarded as a labeling type if its central pixel falls into the region labeled with that type. For each WSI, the patches with one labeling type were randomly sampled according to the overall ratio of the type in the dataset. Normal patches are guaranteed to come from normal WSIs rather than normal areas of tumor slides. The test set is composed of four WSIs with two adenoma and two adenocarcinoma ones, cropped as patches with stride 256 in X and Y directions without overlap area. The numbers of sampled patches are shown in Table 4.

ResNet-50 (49) is used for patch feature extraction and classification in our experiments. The same setting (batch size = 128, classes_num = 3) is used to perform the classification of the tumor, carcinoma, and normal cases. We used Adam to optimize the model with an initial learning rate of zero and â taken from the set of (0.9, 0.999). After five warm-up epochs, the learning rate reached 0.001. Then, CosineAnnealingLR was chosen as the learning rate decay

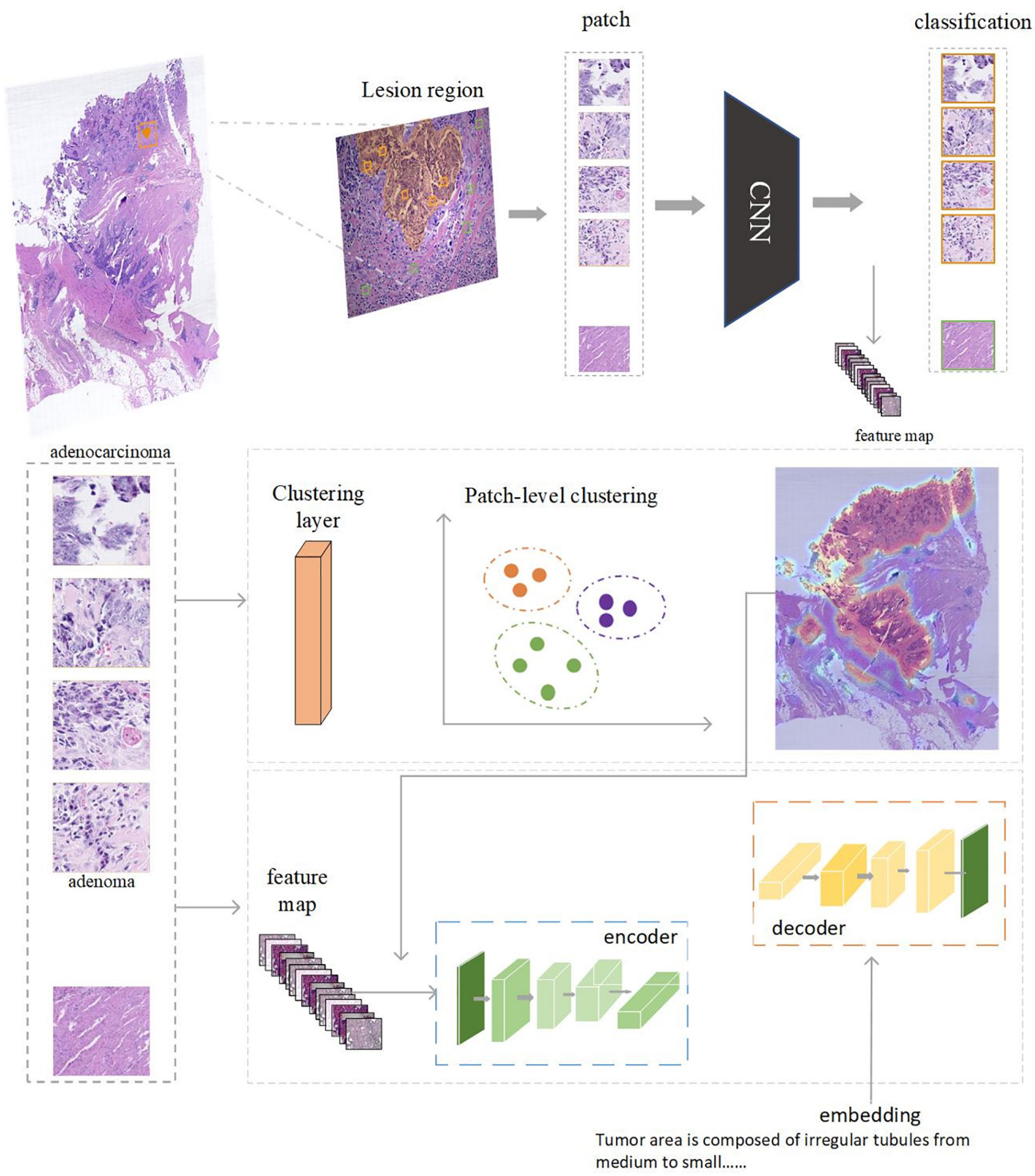


FIGURE 7
The AI model consists of two parts: feature extraction and classification using CNN. Captioning with transformers. Clustering can be performed based on the regions annotated by doctors. After each patch is classified and clustered, the captioning of clustered areas can be performed.

TABLE 4 The number of sampled patches for the training set and test set for the classification task.

	Normal	Adenocarcinoma	Adenoma
Training set with coarse-grain classification labels	133,312	133,321	133,286
Training set with fine-grain subtype labels	133,312	133,322	133,252
Test set	15,244	4,197	10,603

strategy, and after 25 epochs, it decayed to zero. Experiments were run with PyTorch on a machine with a V100 graphics card.

Evaluation

We evaluated the performance with precision, recall, and accuracy indicators. Precision is to measure how many of the positive predictions are positive. Recall tells how many positive cases in the test set are predicted correctly. Accuracy reflects the overall ratio of correct predictions (adenoma, adenocarcinoma, and normal).

Results

Table 5 shows that fine-grain prediction enhances the overall classification accuracy from 79.56 to 85.26%, with a +5.7% improvement compared with the coarse-grain one. In specific, for normal class, the recall measure of fine-grain prediction outperforms

TABLE 5 Confusion matrix of prediction results for models trained with coarse-grain classification labels vs. fine-grain decision-layer subtype labels.

		Prediction			
	Ground truth	Normal	Adenoma	Adenocarcinoma	#Recall
Coarse-grain class data	Normal	5317	483	2039	67.83%
	Adenoma	28	1687	16	97.46%
	Adenocarcinoma	624	26	5517	89.46%
	#Precision	89.07%	76.82%	72.86%	**79.56%**
Fine-grain subtype data	Normal	6202	297	1340	79.12%
	Adenoma	69	1622	40	93.70%
	Adenocarcinoma	565	8	5594	90.71%
	#Precision	90.37%	84.17%	80.21%	**85.26%**

Recall and precision numbers are calculated, and the two boxed numbers represent the overall accuracies of the two models, respectively.

TABLE 6 Partial caption prediction result.

BLEU4	Predicted caption	Original caption
0	The cytoplasm was markedly reduced, karyorrhexis, thickened chromatin, screen mesh. Moderately differentiated adenocarcinoma.	Nuclei remain polar, nucleus stratified or pseudostratified arrangement, tubular structure, increased epithelial cell hierarchy, low grade intraepithelial neoplasia. Low grade adenomas.
0.3	Irregular glandular duct arrangement, cribriform structure, the nucleus of tumor cells are round, nucleoli were more prominent, necrosis. Moderately differentiated adenocarcinoma.	Some tumor cells with round nucleus, nucleoli were more prominent, some cribriform arrangement, some papillary arrangement, necrosis, some tumor cells rod-shaped, stratified arrangement. Moderately differentiated adenocarcinoma.
0.45	Irregular glandular duct arrangement. Moderately differentiated adenocarcinoma.	Infiltration into the submucosa. Moderately differentiated adenocarcinoma.
0.5	Nuclei rod-shaped, nucleus stratified or pseudostratified arrangement, tubular structure. Low grade adenomas.	Nuclei rod-shaped, nucleus stratified, tubular. Low grade adenomas.
0.99	Nuclei rod-shaped, nucleus stratified or pseudostratified arrangement, tubular structure. Low grade adenomas.	Nuclei rod-shaped, nucleus stratified or pseudostratified arrangement, tubular structure. Low grade adenomas.

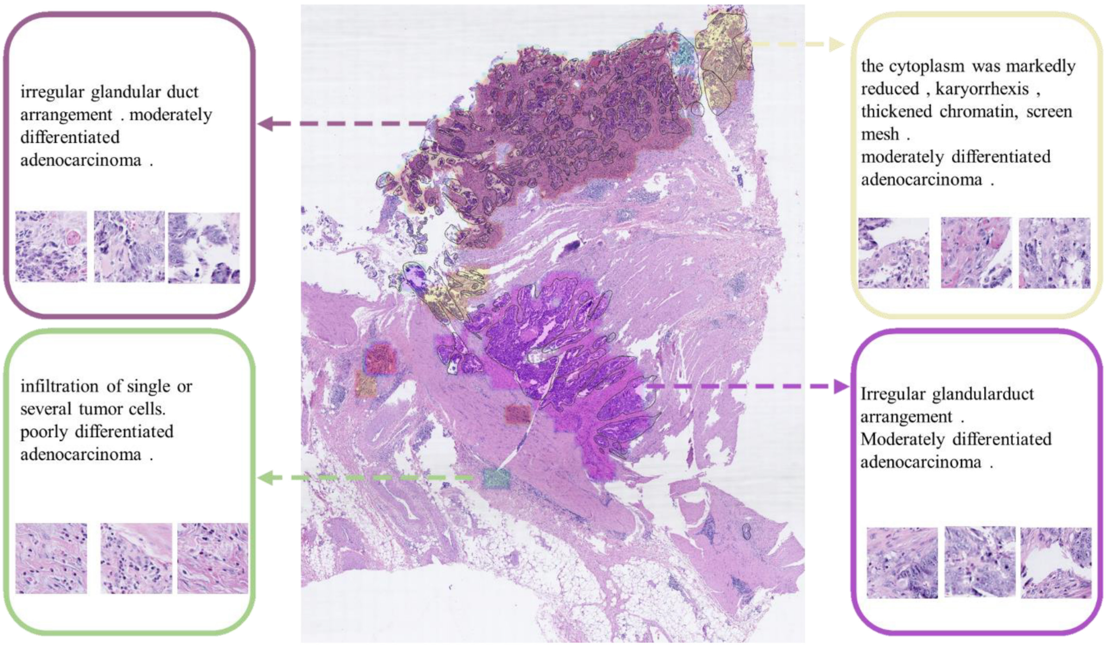

irregular glandular duct arrangement . moderately differentiated adenocarcinoma .
the cytoplasm was markedly reduced , karyorrhexis , thickened chromatin, screen mesh .
moderately differentiated adenocarcinoma .
infiltration of single or several tumor cells.
poorly differentiated adenocarcinoma .
Irregular glandularduct arrangement .
Moderately differentiated adenocarcinoma .

FIGURE 8
Captions generated for each clustered area. These captions are used to describe the lesion detail features of the clustered area.

that of the coarse-grain prediction up to +11.29%, from 67.83 to 79.12%. For adenocarcinoma, coarse-grain prediction results in a small false negative, reaching the recall of 89.46%, while fine-grain one further improves it up to 90.71%. The fine-grain recall measure of adenoma is also good at 93.70%, though is −3.76% inferior to the coarse-grain one, and one possible reason is that some tumor stroma characteristics are difficult to identify. In conclusion, experimental results show that fine-grain annotations can achieve an overall good performance of classification and indicate more details of the present lesions.

Task 2: Caption generation for explaining diagnosis rationale

Task definition

Besides classification, we further verify the effectiveness of reason-layer data in explaining details for the classification rationale in order to support clinical scenarios of pathologists-AI collaboration. We designed a captioning experiment to compare the descriptions annotated by the doctor with the region captions generated by the AI model. We also conducted a subjective evaluation for doctors to review the captions generated.

Methods

The captioning model consists of a Resnet-18 (49) backbone network and a transformer (56). Between the two modules, we inserted a clustering filter module to aggregate patches belonging to the same lesion area into ac luster. The model accepts random patches as input, extracts features *via* the backbone network, and predicts the classification type (normal, adenoma, and adenocarcinoma) of the patch. The clustering filter will then aggregate adjacent abnormal patches into clusters representing the lesion areas. Each cluster contains several patches, which are regarded as a bag of unordered patches. All the patch features in this bag are fed into the transformer to generate the corresponding caption.

All the labeled lesion areas were divided into several patches with corresponding captions for training purposes. For tokenization purposes, patches in each caption bag are sampled to a fixed number. Specifically in the experiment, we set the number of patches per caption as up to 64. During the testing phase, the DBSCAN (57) clustering filter was used after the backbone was completed. Each cluster generated by the clustering filter was into the transformer to generate the caption. We used a Tesla V100 graphics card for training with batch size = 4; AdamW was used as the optimizer with a learning rate of 1e-5. In the test stage, we sampled up to 256 patches per cluster for caption prediction.

Evaluation

The bilingual evaluation understudy (BLEU) (58) score was adopted for quantitative region-level algorithm evaluation. BLEU value is used to measure the similarity between a set of machine-generated translation sentences and a set of human-translated sentences. A higher score reflects a better agreement between the caption produced by the model and the ground-truth description by the annotator.

$$bleu_n = \frac{\sum_{c\in candidates}\sum_{n-gram\in c} Count_{clip}(n-gram)}{\sum_{c'\in candidates}\sum_{n-gram'\in c'} Count_{clip}(n-gram')} \quad (1)$$

Results

We used four grades of BLEU values B1, B2, B3, and B4 to quantify the captioning results. Experiments showed that the model achieved B1 =0.56, B2 = 0.49, B3 = 0.44, and B4 = 0.36, for which the predicted captions demonstrated good similarity to the ground truth descriptions (BLEU around or higher than 0.4). Some examples are shown in **Table 6** for better illustration.

Task 3: Human-AI collaboration experience

We also engaged physicians in qualitative evaluation of the captions at the cluster level. For a certain WSI for testing in Task 1, after completing the ResNet-based classification, we used DBSCAN to cluster the patches and visualize the clustering result as shown in **Figure 8**. All lesion regions are clustered into 13 large typical areas, represented by different colors in **Figure 8**. Eight pathologists (P1-P8 in **Table 1**) were recruited to rate the trust in the algorithm for classification and generating caption results with the subjective Likert Scale (59). For AI-assisted diagnosis, the baseline average score was 3.88 for the trustworthiness and confidence of AI classification results, while with the visualization results of the AI classification algorithm trained by the CR-PathNarratives dataset, the trust and confidence scores in AI-assisted diagnosis provided with more details raised from 3.88 to 4.63. By providing more auxiliary diagnostic information step by step (reason-layer text description, reason-layer text description, and behavior trajectory thermal map), pathologists' trust in AI auxiliary diagnosis increased from 4.25 to 4.38. It shows that CR-PathNarratives with decision-to-reason detail benefit the interpretability of AI by doctors.

In conclusion, our dataset can be applied to the basics of classification and captioning scenarios. Experiments show that adding more comprehensive reason information not only achieves better classification gains, identifies detailed features such as cancer stroma, and reduces the false positive rate, but also enhances the trustworthiness and confidence of doctors to understand and collaborate with pathological AI models.

6. Conclusion

Pathological diagnosis is the gold standard for tumor diagnosis. The continuous development and progress of AI have brought new possibilities for pathology diagnosis. However, there is a relative lack of datasets in the field of computational pathology. We proposed a data annotation protocol PathNarratives with a hierarchical decision-to-reason data structure and a multimodal annotating process and tool. This data annotation schema focuses on the labeling process of the physician with audit capability, records the behavioral information of the physician, and supports analyzing and discovering the diagnostic ideas and logic of physicians. Based on the protocol we have built the colon-rectal dataset, CR-PathNarratives, which contains 174 H&E-stained WSIs. Each WSI was annotated with decision-to-reason labels and multimodal information on vision, language, voice, and behavioral trajectories. Voice explanations and behavioral trajectories make the data more descriptive. Furthermore, we use the decision-to-reason labels of this dataset to perform classification (adenoma, adenocarcinoma, and normal) experiments, as well as region-level and cluster-level captioning experiments for lesion description. Experiments show that our dataset can be applied to multiscenario algorithmic experiments. Refined annotations facilitate machine learning of more detailed information and reduce the false positive rate of classification. Visualization of comprehensive reasoning details enhances the trustworthiness and confidence of doctors to collaborate with pathological AI models, aiming for better human-AI collaboration.

In the future, we plan to optimize the tools for the annotation process, such as adding automated suggestion hints to speed up the annotation. The WSIs in the datasets are expected to be expanded on 300–800 slides, and then we consider using the proposed annotation model to prepare datasets in other pathological domains. Advanced algorithmic models can be further investigated, e.g., better utilizing behavior tracking as training inputs to optimize the classification results.

Data availability statement

The original contributions presented in this study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding authors.

Author contributions

HZ, XW, LL, YH, LG, and XS conceived, designed, and coordinated the writing of the whole manuscript. HZ, XW, PH, and WQ contributed to data collection and experiments. HZ, FW, and HC were responsible for software. PH, WQ, FW, HC, JY, XH, and YL revised literature and wrote the different parts of the manuscript. All authors contributed to critically revised and approved the final version of this manuscript.

Funding

This study was supported in part by the Foundation of Shenzhen Science and Technology Innovation Committee (JCYJ20180507181527806) and Beijing Municipal Science and Technology Commission (Z131100004013036).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.1070072/full#supplementary-material

References

1. Rorke L. Pathologic diagnosis as the gold standard. *Cancer.* (1997) 79:665–7. doi: 10.1002/(SICI)1097-0142(19970215)79:4<665::AID-CNCR1>3.0.CO;2-D

2. Tsuneki M. Deep learning models in medical image analysis. *J Oral Biosci.* (2022) 64:312–20. doi: 10.1016/j.job.2022.03.003

3. Litjens G, Sánchez C, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep.* (2016) 6:26286. doi: 10.1038/srep26286

4. Javed S, Mahmood A, Fraz M, Koohbanani N, Benes K, Tsang Y, et al. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Med Image Anal.* (2020) 63:101696. doi: 10.1016/j.media.2020.101696

5. Hou L, Samaras D, Kurc T, Gao Y, Davis J, Saltz J. Patch-based convolutional neural network for whole slide tissue image classification. *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV: IEEE (2016). p. 2424–33. doi: 10.1109/CVPR.2016.266

6. Korbar B, Olofson A, Miraflor A, Nicka C, Suriawinata M, Torresani L, et al. Deep learning for classification of colorectal polyps on whole-slide images. *J Pathol Inform.* (2017) 8:30. doi: 10.4103/jpi.jpi_34_17

7. Coudray N, Ocampo P, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med.* (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5

8. Kather J, Weis C, Bianconi F, Melchers S, Schad L, Gaiser T, et al. Multi-class texture analysis in colorectal cancer histology. *Sci Rep.* (2016) 6:27988. doi: 10.1038/srep27988

9. Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A, Sethi A. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Trans. Med Imaging.* (2017) 36:1550–60. doi: 10.1109/TMI.2017.2677499

10. Louis D, Feldman M, Carter A, Dighe A, Pfeifer J, Bry L, et al. Computational pathology: a path ahead. *Arch Pathol Lab Med.* (2016) 140:41–50. doi: 10.5858/arpa.2015-0093-SA

11. Bulten W, Bándi P, Hoven J, Loo R, Lotz J, Weiss N, et al. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci Rep.* (2019) 9:864. doi: 10.1038/s41598-018-37257-4

12. Shi J, Gao Z, Zhang H, Puttapirat P, Wang C, Zhang X, et al. Effects of annotation granularity in deep learning models for histopathological images. *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*. San Diego, CA: IEEE (2019). p. 2702–8. doi: 10.1109/BIBM47256.2019.8983158

13. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Med Image Anal.* (2016) 33:170–5. doi: 10.1016/j.media.2016.06.037

14. Litjens G, Bandi P, Ehteshami Bejnordi B, Geessink O, Balkenhol M, Bult P, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *Gigascience.* (2018) 7:giy065. doi: 10.1093/gigascience/giy065

15. Banville H, Albuquerque I, Hyvärinen A, Moffat G, Engemann D, Gramfort A. Self-supervised representation learning from electroencephalography signals. *2019 IEEE 29th international workshop on machine learning for signal processing (MLSP)*. Pittsburgh: IEEE (2019). p. 1–6. doi: 10.1109/MLSP.2019.8918693

16. Campanella G, Hanna M, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam K, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med.* (2019) 25:1301–9. doi: 10.1038/s41591-019-0508-1

17. Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X. Transmil: transformer based correlated multiple instance learning for whole slide image classification. *Adv Neural Inform Process Syst.* (2021) 34:2136–47.

18. Rony J, Belharbi S, Dolz J, Ayed I, McCaffrey L, Granger E. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *Arxiv.* [Preprint]. (2019).

19. He X, Zhang Y, Mou L, Xing E, Xie P. Pathvqa: 30000+ questions for medical visual question answering. *Arxiv.* [Preprint]. (2020). doi: 10.36227/techrxiv.13127537.v1

20. Gamper J, Rajpoot N. Multiple instance captioning: learning representations from histopathology textbooks and articles. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Nashville, TN: IEEE (2021). p. 16549–59. doi: 10.1109/CVPR46437.2021.01628

21. Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat Mach Intell.* (2019) 1:236–45. doi: 10.1038/s42256-019-0052-1

22. Wahab N, Miligy I, Dodd K, Sahota H, Toss M, Lu W, et al. Semantic annotation for computational pathology: multidisciplinary experience and best practice recommendations. *J Pathol.* (2022) 8:116–28. doi: 10.1002/cjp2.256

23. Bejnordi B, Veta M, Van Diest P, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA.* (2017) 318:2199–210. doi: 10.1001/jama.2017.14580

24. Bandi P, Geessink O, Manson Q, Van Dijk M, Balkenhol M, Hermsen M, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Trans Med Imaging.* (2018) 38:550–60. doi: 10.1109/TMI.2018.2867350

25. Wang C, Chang C, Khalil M, Lin Y, Liou Y, Hsu P, et al. Histopathological whole slide image dataset for classification of treatment effectiveness to ovarian cancer. *Sci Data.* (2022) 9:25. doi: 10.1038/s41597-022-01127-6

26. Aksac A, Demetrick D, Ozyer T, Alhajj R. BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis. *BMC Res Notes.* (2019) 12:82. doi: 10.1186/s13104-019-4121-7

27. Spanhol F, Oliveira L, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. *IEEE Trans Biomed Eng.* (2015) 63:1455–62. doi: 10.1109/TBME.2015.2496264

28. Tsuneki M, Kanavati F. Inference of captions from histopathological patches. *Arxiv.* [Preprint]. (2022).

29. Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G, Cucchiara R. From show to tell: a survey on image captioning. *Arxiv.* [Preprint]. (2021).

30. Pont-Tuset J, Uijlings J, Changpinyo S, Soricut R, Ferrari V. Connecting vision and language with localized narratives. In: Vedaldi A, Bischof H, Brox T, Frahm J editors. *Computer vision – ECCV 2020. ECCV 2020. lecture notes in computer science*. Cham: Springer (2020). p. 647–64. doi: 10.1007/978-3-030-58558-7_38

31. Koh J, Baldridge J, Lee H, Yang Y. Text-to-image generation grounded by fine-grained user attention. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. Waikoloa, HI: IEEE (2021). p. 237–46. doi: 10.1109/WACV48630.2021.00028

32. Meng Z, Yu L, Zhang N, Berg T, Damavandi B, Singh V, et al. Connecting what to say with where to look by modeling human attention traces. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Nashville, TN: IEEE (2021). p. 12679–88. doi: 10.1109/CVPR46437.2021.01249

33. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis.* (2017) 123:32–73. doi: 10.1007/s11263-016-0981-7

34. Chakraborty S, Ma K, Gupta R, Knudsen B, Zelinsky G, Saltz J, et al. Visual attention analysis of pathologists examining whole slide images of Prostate cancer. *2022 IEEE 19th International symposium on biomedical imaging (ISBI)*. Kolkata: IEEE (2022). p. 1–5. doi: 10.1109/ISBI52829.2022.9761489

35. Gygli M, Ferrari V. Efficient object annotation via speaking and pointing. *Int J Comput Vis.* (2020) 128:1061–75. doi: 10.1007/s11263-019-01255-4

36. Jhuang H, Gall J, Zuffi S, Schmid C, Black M. Towards understanding action recognition. *Proceedings of the IEEE international conference on computer vision*. Sydney, NSW: IEEE (2013). p. 3192–9. doi: 10.1109/ICCV.2013.396

37. Wang H, Schmid C. Action recognition with improved trajectories. *Proceedings of the IEEE international conference on computer vision*. Sydney, NSW: IEEE (2013). p. 3551–8. doi: 10.1109/ICCV.2013.441

38. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu, HI: IEEE (2017). p. 6299–308. doi: 10.1109/CVPR.2017.502

39. Gurcan M, Boucheron L, Can A, Madabhushi A, Rajpoot N, Yener B. Histopathological image analysis: a review. *IEEE Rev Biomed Eng.* (2009) 2:147–71. doi: 10.1109/RBME.2009.2034865

40. Veta M, Pluim J, Van Diest P, Viergever M. Breast cancer histopathology image analysis: a review. *IEEE Trans Biomed Eng.* (2014) 61:1400–11. doi: 10.1109/TBME.2014.2303852

41. Saha M, Chakraborty C. Her2Net: a deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. *IEEE Trans Image Process.* (2018) 27:2189–200. doi: 10.1109/TIP.2018.2795742

42. Luo X, Zang X, Yang L, Huang J, Liang F, Rodriguez-Canales J, et al. Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J Thoracic Oncol.* (2017) 12:501–9. doi: 10.1016/j.jtho.2016.10.017

43. Yan J, Chen H, Wang K, Ji Y, Zhu Y, Li J, et al. Hierarchical attention guided framework for multi-resolution collaborative whole slide image segmentation. *International conference on medical image computing and computer-assisted intervention*. Berlin: Springer (2021). p. 153–63. doi: 10.1007/978-3-030-87237-3_15

44. Abu Haeyeh Y, Ghazal M, El-Baz A, Talaat I. Development and evaluation of a novel deep-learning-based framework for the classification of renal histopathology images. *Bioengineering.* (2022) 9:423. doi: 10.3390/bioengineering9090423

45. Zhou C, Jin Y, Chen Y, Huang S, Huang R, Wang Y, et al. Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning. *Comput Med Imaging Graph.* (2021) 88:101861. doi: 10.1016/j.compmedimag.2021.101861

46. Pavlopoulos J, Kougia V, Androutsopoulos I, Papamichail D. Diagnostic captioning: a survey. *Knowl Inform Syst.* (2022) 64:1691–722. doi: 10.1007/s10115-022-01684-7

47. Abacha A, Hasan S, Datla V, Liu J, Demner-Fushman D, Müller H. VQA-Med: overview of the medical visual question answering task at ImageCLEF 2019. *Working Notes of CLEF 2019 - conference and labs of the evaluation forum*. Lugano: CEUR-WS.org (2019).

48. Lau J, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. *Sci Data.* (2018) 5:180251. doi: 10.1038/sdata.2018.251

49. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. . *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV: IEEE (2016). p. 770–8. doi: 10.1109/CVPR.2016.90

50. Nagtegaal I, Odze R, Klimstra D, Paradis V, Rugge M, Schirmacher P, et al. The 2019 WHO classification of tumours of the digestive system. *Histopathology.* (2020) 76:182. doi: 10.1111/his.13975

51. Ponz de Leon M. Colorectal cancer at the beginning of the new millennium. In: World Health Organization editor. *Colorectal Cancer.* (Berlin: Springer) (2002). p. 285–9. doi: 10.1007/978-3-642-56008-8_14

52. Dyba T, Randi G, Bray F, Martos C, Giusti F, Nicholson N, et al. The European cancer burden in 2020: incidence and mortality estimates for 40 countries and 25 major cancers. *Eur J Cancer.* (2021) 157:308–47. doi: 10.1016/j.ejca.2021.07.039

53. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: a vendor-neutral software foundation for digital pathology. *J Pathol Inform.* (2013) 4:27. doi: 10.4103/2153-3539.119005

54. Macenko M, Niethammer M, Marron J, Borland D, Woosley J, Guan X, et al. A method for normalizing histology slides for quantitative analysis. *2009 IEEE international symposium on biomedical imaging: from nano to macro*. Boston, MA: IEEE (2009). p. 1107–10. doi: 10.1109/ISBI.2009.5193250

55. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybernet.* (1979) 9:62–6. doi: 10.1109/TSMC.1979.4310076

56. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. *Adv Neural Inform Process Syst.* (2017) 30:5998–6008.

57. Khan K, Rehman S, Aziz K, Fong S, Sarasvady S. DBSCAN: past, present and future. *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. Bangalore: IEEE (2014). p. 232–8. doi: 10.1109/ICADIWT.2014.6814687

58. Papineni K, Roukos S, Ward T, Zhu W. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the association for computational linguistics*. Stroudsburg: Association for Computational Linguistics (2002). p. 311–8. doi: 10.3115/1073083.1073135

59. Likert R, Roslow S, Murphy G. A simple and reliable method of scoring the Thurstone attitude scales. *J Soc Psychol.* (1934) 5:228–38. doi: 10.1080/00224545.1934.9919450

Frontiers in Medicine

Translating medical research and innovation into improved patient care

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

Discover the latest Research Topics

See more →

frontiers

Frontiers in
Medicine

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact