An accurate prediction of the origin for bone metastatic cancer using deep learning on digital pathological images

Lianghui Zhu,[a,i] Huijuan Shi,[b,i] Huiting Wei,[b] Chengjiang Wang,[a] Shanshan Shi,[a] Fenfen Zhang,[b] Renao Yan,[a] Yiqing Liu,[a] Tingting He,[a] Liyuan Wang,[b] Junru Cheng,[a] Hufei Duan,[a] Hong Du,[c] Fengjiao Meng,[d] Wenli Zhao,[e] Xia Gu,[f] Linlang Guo,[g] Yingpeng Ni,[h] Yonghong He,[a,]*** Tian Guan,[a,]** and Anjia Han[b,]*

[a]Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen, Guangdong, China
[b]Department of Pathology, the First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China
[c]Department of Pathology, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China
[d]Department of Pathology, Zhongshan People's Hospital, Zhongshan, China
[e]Department of Pathology, The First People's Hospital of Huizhou, Huizhou, China
[f]Department of Pathology, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China
[g]Department of Pathology, Zhujiang Hospital, Southern Medical University, Guangzhou, China
[h]Department of Pathology, Jieyang People's Hospital (Jieyang Affiliated Hospital, Sun Yat-Sen University), Jieyang, China

eBioMedicine
2023;87: 104426

Published Online XXX
https://doi.org/10.1016/j.ebiom.2022.104426

Summary

Background Determining the origin of bone metastatic cancer (OBMC) is of great significance to clinical therapeutics. It is challenging for pathologists to determine the OBMC with limited clinical information and bone biopsy.

Methods We designed a regional multiple-instance learning algorithm to predict the OBMC based on hematoxylin-eosin (H&E) staining slides alone. We collected 1041 cases from eight different hospitals and labeled 26,431 regions of interest to train the model. The performance of the model was assessed by ten-fold cross validation and external validation. Under the guidance of top3 predictions, we conducted an IHC test on 175 cases of unknown origins to compare the consistency of the results predicted by the model and indicated by the IHC markers. We also applied the model to identify whether there was tumor or not in a region, as well as distinguishing squamous cell carcinoma, adenocarcinoma, and neuroendocrine tumor.

Findings In the within-cohort, our model achieved a top1-accuracy of 91.35% and a top3-accuracy of 97.75%. In the external cohort, our model displayed a good generalizability with a top3-accuracy of 97.44%. The top1 consistency between the results of the model and the immunohistochemistry markers was 83.90% and the top3 consistency was 94.33%. The model obtained an accuracy of 98.98% to identify whether there was tumor or not and an accuracy of 93.85% to differentiate three types of cancers.

Interpretation Our model demonstrated good performance to predict the OBMC from routine histology and had great potential for assisting pathologists with determining the OBMC accurately.

Funding National Science Foundation of China (61875102 and 61975089), Natural Science Foundation of Guangdong province (2021A15-15012379 and 2022A1515 012550), Science and Technology Research Program of Shenzhen City (JCYJ20200109110606054 and WDZC20200821141349001), and Tsinghua University Spring Breeze Fund (2020Z99CFZ023).

Keywords: Deep learning; Digital pathology; Bone metastatic cancer; Origin; Regional multiple-instance learning

*Corresponding author.
**Corresponding author.
***Corresponding author.
E-mail addresses: hananjia@mail.sysu.edu.cn (A. Han), guantian@sz.tsinghua.edu.cn (T. Guan), heyh@sz.tsinghua.edu.cn (Y. He).

[i]These authors contributed equally.

Research in context

Evidence before this study

Many advanced cancers metastasize to bone and determining the origin of bone metastatic cancer is of great significance to perform precise therapy for patients. Traditionally, pathologists need to combine the whole-body imaging, medical history, and immunohistochemistry staining results to determine the origin of bone metastatic carcinoma. However, nearly 20% cases are still hard to diagnose even after comprehensive tests. Recently, deep learning has been widely applied in digital pathology images. It demonstrated excellent performance in tumor recognition, classification, grading, metastasis recognition, and prognosis analysis. Only one publication of Lu used a weekly supervised learning method to identify the occult primary site of tumors by hematoxylin-eosin staining slides. However, to the best of our knowledge, no researches about determining the origin of bone metastatic cancer from routine histology with the approach of deep learning have been reported up to now. Unlike slides of primary lung cancer, breast cancer or other common cancers where large cancer foci cluster together, many bone metastatic cancer slides contain several scattered cancer foci, some of which are too small to be detected.

Added value of this study

We assembled a multi-center dataset comprising 1259 cases, labeled 27,998 regions of interest, and designed a regional multiple-instance learning algorithm to determine the origin of bone metastatic cancer from digital histology. Both in the within-cohort and the external cohort, our model demonstrated high top1 accuracy and top3 accuracy. To further demonstrate its practicability, we tested 175 cases of unknown origin with specific immunohistochemistry markers which were selected based on the top3 potential origins predicted by the model and the morphology of tissues. It also exhibited high consistency between the results predicted by the model and indicated by the immunohistochemistry markers. Additionally, we found that this model could be trained to identify whether there was tumor or not in a region, as well as distinguishing squamous cell carcinoma, adenocarcinoma, and neuroendocrine tumor.

Implications of all the available evidence

For the cases where sufficient tissues for necessary immunohistochemistry tests are unavailable, our model trained on thousands of samples have great potential for assisting pathologists with determining the OBMC accurately. Furthermore, the accurate top3 predictions of the model are likely to reduce the number of immunohistochemistry markers selected in the routine diagnosis process.

Introduction

Bone is a common site of cancer metastasis.[1] Bone metastasis occurs in 70% patients with advanced breast cancer, 85% with prostate cancer and 40% with lung cancer.[2] Patients with bone metastases are often clinically manifested as bone pain, pathological fracture, spinal cord compression, and hypercalcemia.[3,4] Determining the origin of bone metastatic cancer (OBMC) is essential to perform precise therapy for patients.[5,6]

The OBMC is difficult for pathologists to identify based on the limited bone biopsies and correlated clinical information. In practice, pathologists combine the whole-body imaging, medical history, and immunohistochemistry (IHC) staining results to draw a conclusion of the OBMC.[7–9] Due to the incompleteness of clinical information, pathologists have to select multiple IHC markers to determine the OBMC (see Fig. 1a and b). However, limited bone fine needle biopsies make it impossible for pathologists to perform adequate IHC tests. Even after comprehensive processes of clinical and pathological examinations, about 20% cases are still hard to diagnose.[10] How to make an exact and fast determination of the OBMC poses a great challenge in medical science.

With the rapid development of deep learning theory and technology, related methods have been widely used in the field of medicine.[11–13] The powerful learning ability, excellent transfer performance and strong robustness make deep learning very suitable for medical image analysis. Particularly in recent years, deep learning has been widely used in the analysis of digital pathological images, including tumor recognition,[14,15] classification,[16] grading,[17] metastasis recognition,[18] and prognosis analysis.[19] These explorations illustrate great potentials and advantages of deep learning in the computational pathology.

As for the problem of tumor metastasis, Wang adopted GoogLeNet,[20] ResNet101,[21] and VGG-net[22] architectures to detect whether breast cancer has metastasized to lymph nodes and the performance was comparable with the pathologist interpreting the slides.[23] Chuang trained a model on the dataset of whole slide images (WSIs) in four-fold to judge whether lymph nodes have micro metastases of rectal cancers.[24] Pham designed a two-step combined convolution neural network to detect nodal metastasis of lung cancer with a low false positive rate.[25] Jaakko trained a residual network to identify metastatic cutaneous squamous cell carcinoma, as well as assessing its metastatic risk and prognosis. Recently,[26] with the help of slide-level multiple-instance learning algorithm, Lu made it possible to identify the occult primary site of tumors.[27] Trained by digital images of Hematoxylin and eosin (H&E) stained sections alone, their model reached an accuracy of 83%.

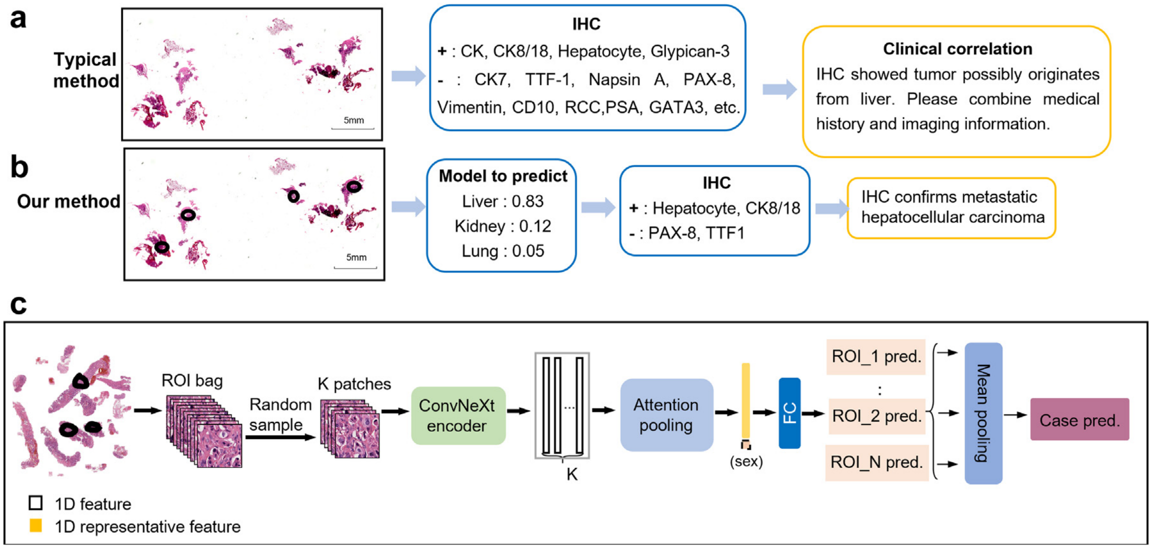
a
Typical method
5mm
IHC
+ : CK, CK8/18, Hepatocyte, Glypican-3
- : CK7, TTF-1, Napsin A, PAX-8, Vimentin, CD10, RCC,PSA, GATA3, etc.
Clinical correlation
IHC showed tumor possibly originates from liver. Please combine medical history and imaging information.
b
Our method
5mm
Model to predict
Liver : 0.83
Kidney : 0.12
Lung : 0.05
IHC
+ : Hepatocyte, CK8/18
- : PAX-8, TTF1
IHC confirms metastatic hepatocellular carcinoma
c
ROI bag
Random sample
K patches
ConvNeXt encoder
K
Attention pooling
(sex)
FC
ROI_1 pred.
ROI_2 pred.
ROI_N pred.
Mean pooling
Case pred.
□ 1D feature
🟨 1D representative feature

***Fig. 1:* The work-flow of diagnosing the OBMC.** (a) The typical workflow for pathologists to make differential diagnosis of the OBMC. Pathology doctors diagnosed the OBMC based on the pathology of H&E slides, the results of dozens of IHC tests and clinical correlation. (b) The new workflow of applying the method of RMIL to determine the OBMC. Based on the labeled ROI in a WSI, our model provided three most likely OBMCs. Only three IHC stains need to be used to confirm the final origin. (c) The workflow of RMIL.

However, to the best of our knowledge, no researches on determining the OBMC with the approach of deep learning have been reported till now. Owing to the characteristic of bone metastatic cancers (i.e., many bone metastatic carcinoma foci scatter in the whole image, some of which are too small to be detected; the morphology of tumors in bone tissues is not as clear as that in other tissues as a result of decalcification; and bone tissues are frequently squeezed and deformed when taken from patients), to predict the primary site of tumors simply with slide labels poses a great challenge for the model.

In this study, we proposed a regional multiple-instance learning algorithm (RMIL) to overcome this challenge. We collected 1041 cases of bone metastatic carcinoma with eight common primary sites (breast, prostate, thyroid, lung, liver, kidney, stomach and intestine) and labeled 26,431 regions of interest (ROI) on H&E images to train the model. Most of previous researches using weekly supervised method regard a whole slide as a bag.[27–30] Here, a labeled region (LR) in this work was defined as a bag, which greatly enlarged the number of bags in case of insufficient slides. Instead of feeding all patches of a bag into the model pretrained on the dataset of natural images as mentioned by Lu et al., we randomly selected a certain number of patches in a bag. In the former algorithm, the number of patches in a bag was so large that all parameters of the feature extractor had to be frozen, which only extracted basic features. However, in our study, we aimed to train the feature extracting network by bone metastatic cancer images so as to make more accurate predictions.

We also built an external dataset to confirm the robustness of our model. Furthermore, to demonstrate that our method could truly help pathologists determine the OBMC, we evaluated RMIL on another 175 cases of bone metastatic carcinoma with unknown origins. The main contributions of this study are as follows:

(1) We established a large data set of bone metastatic carcinoma, including 1259 cases, 2449 whole slide images (WSIs) and 27,998 ROI labeled by experts.
(2) Aiming at predicting the OBMC, we designed RMIL, which could be trained in an end-to-end manner and exhibited excellent performance on the test dataset. On the external dataset, this method also achieved high accuracy, indicating its good adaptability across different staining protocols and imaging scanners.
(3) The results predicted by RMIL on 175 cases of bone metastatic carcinoma with unknown origins demonstrated that RMIL could be used as an auxiliary tool for differential diagnosis of the OBMC and narrow down the selection of IHC markers for determining the OBMC.

Methods

Data set

We collected both bone biopsies and surgical specimens from eight healthcare centers in China and the diagnosis time of these cases was from 1998 to 2022. Detailed description how the study cohort was recruited including inclusion and exclusion diagram

(Supplementary Fig. S1) could be available in supplementary data. All immunohistochemical antibodies in the research were also listed in supplementary data.

The overall dataset used in this project was composed of 1473 cases of patients (Supplementary Table S1). The dataset used to determine the OBMC was comprised of 1259 cases (27,998 labeled ROI) from eight hospitals (Table 1). We divided these data into three subsets, dataset A, dataset B and dataset C. All H&E slides were sectioned formalin-fixed and paraffin-embedded tissues.

Dataset A containing 1041 cases (26,431 ROI) was randomly split into a training set (80% ROI), a validation set (10% ROI) and a test set (10% ROI) for model development and evaluation. The primary sites of bone metastatic cancer in Dataset A came from eight common organs including lung, breast, prostate, kidney, liver, thyroid, stomach and intestine.

Dataset B composed of 43 cases (383 ROI) was used to assess the adaptability of the model across different hospitals as well as image acquisition devices. There were only six primary sites of bone metastatic carcinoma (lung, breast, kidney, liver, stomach and intestine) in dataset B due to the limited number of cases.

Dataset C comprised of 175 cases (1184 ROI) was used to evaluate the consistency between the results predicted by RMIL and the origins suggested by IHC tests. Eight primary sites which were the same as described in dataset A could be found in Dataset C. The cases of unknown OBMC on dataset C referred to the patients without valuable imaging information for OBMC during their visits to the doctors, the cases without IHC tests because the patients were discharged from the hospital, or the cases with IHC tests, yet the results of which were unable to indicate the OBMC due to the limited variety of IHC antibodies.

Dataset D containing 1203 cases (Supplementary Table S2) was used to train the model of predicting whether there was metastatic cancer in bone tissue. The category of non-bone metastases involved 52 cases and every 20 patches randomly selected in a WSI formed a bag, which was different from the bag produced in the category of bone metastases. All the cases in dataset A were totally the same as those of bone metastases in dataset D.

Dataset E used to train the model of classifying adenocarcinoma, squamous and neuroendocrine carcinoma was composed of 1073 cases (Supplementary Table S3), 911 adenocarcinoma cases of which came from dataset A.

The incidence rates of breast cancer and prostate cancer are greatly related with sex which may contribute to the model making accurate predictions; therefore, we collected sex information self-reported by patients from operation records and pathological reports.

Dataset	Primary site	Lung			Liver			Kidney			Breast			Stomach			Prostate			Thyroid			Intestine			Total		
		C	W	R	C	W	R	C	W	R	C	W	R	C	W	R	C	W	R	C	W	R	C	W	R	C	W	R
DA	FAHSYU	288	644	11,742	119	268	2249	71	189	1765	142	328	2874	13	33	385	57	134	1823	97	212	1694	45	104	1186	832	1912	23,718
	EDFAHSYU	4	7	99	1	1	4	–	–	–	3	7	54	1	2	35	1	3	25	–	–	–	3	7	113	13	27	330
	FAHGZMU	14	14	125	1	1	9	1	1	13	1	1	5	1	1	11	2	2	29	–	–	–	–	–	–	20	20	192
	GZFPH	28	41	397	7	15	166	3	5	35	6	6	82	5	5	34	11	16	242	3	3	26	–	–	–	63	91	982
	HZFH	6	6	61	1	1	5	3	3	32	5	5	33	–	–	–	2	2	14	3	3	25	–	–	–	20	20	170
	ZJHSMU	6	6	82	4	4	31	–	–	–	3	3	43	–	–	–	4	4	43	9	9	100	1	1	15	27	27	314
	ZSPH	26	26	282	7	7	65	8	8	93	17	17	206	–	–	–	5	5	57	3	3	22	–	–	–	66	66	725
	Total	372	744	12,788	140	297	2529	86	206	1938	177	367	3297	20	41	465	82	166	2233	115	230	1867	49	112	1314	**1041**	**2163**	**26,431**
DB	JYPH	20	39	186	10	22	102	3	6	22	5	12	39	1	1	5	–	–	–	–	–	–	4	5	29	43	85	383
DC	FAHSYU	(Lung)			(Liver)			(Kidney)			(Breast)			(Stomach)			(Prostate)			(Thyroid)			(Intestine)			Total		
	IHC	83	91	548	18	21	226	8	10	49	10	16	82	2	2	24	7	9	45	4	5	32	10	10	71	142	165	977
	DOC	28	31	173	–	–	–	–	–	–	–	–	–	2	2	10	2	2	21	–	–	–	1	1	3	33	36	207
	Total	111	122	721	18	22	226	8	10	49	10	16	82	4	4	34	9	11	66	4	5	32	11	11	74	**175**	**201**	**1184**
Total		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	**1259**	**2449**	**27,998**

Note: DA: Dataset A, DB: Dataset B, DC: Dataset C. C: Case, W: WSI, R: ROI. FAHSYU: the First Affiliated Hospital, Sun Yat-sen University, EDFAHSYU: The East Division of the First Affiliated Hospital, Sun Yat-sen University, FAHGZMU: The First Affiliated Hospital of Guangzhou Medical University, GZFPH: Guangzhou First People's Hospital, HZFH: Huizhou First Hospital, ZJHSMU: ZhuJiang Hospital of Southern Medical University, ZSPH: Zhongshan City People's Hospital, JYPH: Jieyang City People's Hospital. Due to uncertainties of the ground truth labels of dataset C, we used brackets to enclose the potential OBMC. Dataset A was used to develop and evaluate the model. Dataset B was used to assess the adaptability of the model across different hospitals as well as image acquisition devices. Dataset C was used to evaluate the consistency between the results predicted by RMIL and the origins suggested by IHC tests.

Table 1: **Statistics of the dataset used to develop the model of predicting the OBMC. According to different purposes, this dataset was divided into three subsets—dataset A, dataset B, and dataset C.**

Digitization and annotation

Slides in dataset A, dataset C, dataset D and dataset E were scanned with an SQS-1000 scanner (Sqray company, Shenzhen, China) at × 20 magnification and dataset B was scanned with an SQS-2000 at × 20 magnification. All private information of a patient was eliminated when downloading the case. ALL slides were de-identified before scanning and digitization. We used all available data from eight health centers.

To verify the origin of bone metastases, two experts with 20 years' pathology experience were invited to read slides. They combined HE slides, immunohistochemistry results and patients' clinical information to make diagnoses. If these two pathologists held different opinions, a third expert was involved to read the slide as well as adding more immunohistochemical markers to confirm the origin if necessary.

Three pathologists with more than five years of clinical experience drew one to twenty ROI in every whole slide image. All ROI in a slide shared the same label. The only requirement of annotation is that more than 50% of a LR be tumor areas.

For dataset C, three pathologists combined the H&E morphology and the top3 origins predicted by our model and then conducted IHC tests. Another three expert pathologists would be invited to make a unified judgment on uncertain cases.

Image preprocessing

We cropped ROI into 256 × 256 patches (without overlap) at × 20 magnification. Before training, we adopted techniques of data augmentation, including random flipping and color jitter.

The architecture of the model

Many bone metastatic carcinoma foci scatter in the whole image, some of which are too small to be detected. The morphology of tumors in bone tissues is not as clear as that in other tissues as a result of decalcification and bone tissues are frequently squeezed and deformed when taken from patients. With slide-level labels alone, the model trained on a dataset that is not large enough may fail to learn the characteristics of carcinomas. To overcome this limitation, pathologists were involved to annotate representative regions. This method can not only improve the accuracy of classification but also the efficiency of annotation. When labeling data, pathologists only need to roughly draw ROI ranging from one to twenty rather than categorize every patch in a region or carry out further image screening. In this study, a fixed number of patches were randomly picked in a LR, in which every selected patch was regarded as an instance.

A patch with a shape of 256 × 256 was encoded into a 1024-dimentinal feature vector by a network, which we called feature extractor. In this work, we studied the effects of three types of feature extractors (ResNet,[21] Swin Transformer,[30] and ConvNeXt[31]) on the results. All features in a bag were merged into a representative vector by the operation of attention-based pooling.[32] The classification layer embedded the representative vector concatenated with sex and finally output the region-level predictions. We managed to train the model in an end-to-end manner, that is, all components of the model including feature extractors and attention-pooling networks could be trained simultaneously.

The workflow of RMIL is as shown in Fig. 1c. We use $\boldsymbol{h_k}$ to represent the feature vector of the k-th patch (k = 1, 2, ...N). The attention score a_k that indicates the importance of the kth patch to the region-level prediction is computed as

$$a_k = \frac{\exp(\boldsymbol{W_a}(\tanh(\boldsymbol{V_a h_k}) \odot \sigma(\boldsymbol{U_a h_k})))}{\sum_{j=1}^{N} \exp\left(\boldsymbol{W_a}\left(\tanh\left(\boldsymbol{V_a h_j}\right) \odot \sigma\left(\boldsymbol{U_a h_j}\right)\right)\right)} \tag{1}$$

$\boldsymbol{V_a}$ and $\boldsymbol{U_a} \in \mathbb{R}^{L \times D}$ are the same dimensions where L = 768 and D = 1024. Here, D represents the dimension of the feature of each patch. There is another independent weight parameter $\boldsymbol{W_a} \in \mathbb{R}^{1 \times 768}$ in the attention network. The representative vector of a LR can be expressed as $\boldsymbol{v_{roi}} \in \mathbb{R}^{1 \times D}$.

$$\boldsymbol{v_{roi}} = \sum_{k=1}^{N} a_k \boldsymbol{h_k} \tag{2}$$

Then v_{roi} is concatenated with the patient's sex which is encoded by binary values to produce a new vector $\boldsymbol{v'_{roi}} \in \mathbb{R}^{1 \times (D+1)}$. After feeding $\boldsymbol{v'_{roi}}$ into a classification layer, the network outputs the final regional probability $\boldsymbol{P_{roi}}$. In clinical practice, pathologists prefer case-level diagnosis rather than region-level prediction. We assumed that the contribution of each LR to the case level classification was equal. Therefore, we averaged the probability of each LR and then regarded the origin with highest probability as the case level prediction. The case-level probabilities are $\boldsymbol{P_{case}}$.

$$\boldsymbol{P_{case}} = \frac{1}{M}\left(\sum_{1}^{M} \boldsymbol{P_{roi}^{m}}\right) \tag{3}$$

Here, M represents the total number of ROI in a case and $\boldsymbol{P_{roi}^{m}}$ is the probability of the m-th LR.

In this work, we studied the effects of the number of patches selected in a LR, the type of backbone and the sex addition on the performance of the model. We also compared the results predicted by RMIL on our dataset with those of TOAD proposed by Lu.

Training details

To overcome the problem of distribution imbalance of training set, we calculated the loss with the method of weighted cross entropy.[33] The weight of loss is negatively correlated with the number of ROI of each category during training and we adjusted the weight of lung from 0.4 to 1 due to its high frequency in practice. Finally, the weights were set to 2, 2.5, 3, 1, 2.5, 3.5, 8 and 4 corresponding to the primary site of breast, prostate, thyroid, lung, liver, kidney, stomach and intestine.

Our batch size was 16 and the training epochs were 50. We trained the model via the Adam optimizer with a learning rate of 1e-4 and a decayed rate of 1e-4.

Visualization

To visualize the attention score which interpreted the importance of each region or patch to the final prediction, we drew a heatmap of the WSI, as well as the region. To generate a more fine-grained heatmap, we cropped the WSI into 256 × 256 patches with 75% overlap and a LR into the same size with an overlap of 90%.

Statistics analysis

We estimated the performance of models with evaluation metrics of one-versus-rest recall and precision. For the prediction of every primary site, we applied the receiver operating characteristic curve (ROC curve) and the area under ROC curves (AUROC) to evaluate the model. The comprehensive evaluation metrics in this paper includes the average accuracy and the micro averaged AUROC. We also used the top3 accuracy which calculates the percentage of records for which the targets are in the top3 predictions. A high top3 accuracy indicates that the top3 predictions of the model can be useful to narrow down the OBMC and reduce the number of subsequent tests. Due to uncertainties of the ground truth labels of dataset C, we assessed the concordance between the predictions of the model and the results of IHC tests with top-k (k = 1,3) agreement, as well as Cohen kappa score which measures the prefect agreement and the agreement by chance between two raters.[34] In the experiment of detecting weather there was carcinoma in the bone biopsy, sensitivity, specificity and F1-score were adopted to evaluate the performance of the model. To reduce the impact of random data set division, we carried out the ten-fold cross validation. Two-sided P tests without adjustments were applied to all statistical analyses and the significance threshold was 0.05.

Software and package

The software liberties and packages we used included python 3.7.10, opencv-python 4.2.0, pytorch 1.8.1, matplotlib 3.2.2, numpy 1.19.2, pandas 1.3.3, scipy 1.6.2, tensorflow 2.6.0, torchvision 0.9.1, and scikit-learn 0.24.2. The software used to annotate whole slide images was ImageViewerG 1.1.7.

Role of funders

Funders were not involved in data collection, analysis, interpretation, trial design, patient recruitment, or any aspect pertinent to the study. They had no role in the writing of the manuscript or the decision to submit it for publication.

Ethics

This retrospective study was approved by IEC for Clinical Research and Animal Trials of the First Affiliated Hospital of Sun Yat-sen University (No. [2022]429). All patients had signed informed consent before they accepted operations biopsies and pathological examinations.

Results

The number of selected patches

In the study, we designed a series of values for experiments on dataset A to determine the optimum number of patches selected. The results were shown in Fig. 2. In the first stage, with the increase of patch number, the average accuracy (ACC) and micro area under the receiver operator characteristic curve (micro-AUROC) gradually increased. In the second stage, when the patch number reached five, the ACC and the micro-AUROC achieved the maximum values of 92.48% ± 0.22% and 99.47% ± 0.08% respectively. In the third stage, ACC and micro-AUROC were fluctuating as the number of patches rose, but remained under the maximum values. This phenomenon implied it was some part of the whole patches from a LR that ultimately determined the category.

To figure out the key patches that exert major impacts on the final classification, we input all the patches of a LR into the model and calculated the attention score of each patch for visualization (Fig. 3). The results revealed that the attention scores of cancer cells in the test set were significantly higher than those of non-cancer cells.

Backbone

The abilities of encoders to extract features from pathological images are different. ResNet[21] is a series of classic network in image processing. In this work, we studied the influence of seven encoders on the performance of the model. All encoders were pretrained on ImageNet before end-to-end training on dataset A.

The recall, precision, ACC, and ROC curve of each encoder were shown in Table 2. It was evident that ConvNeXt outperformed ResNet on all metrics. For the recall metrics of eight primary sites, ConvNeXt-tiny scored highest on four of them (prostate: 89.53% ± 2.02%, lung: 95.93% ± 0.80%, kidney: 90.17% ± 2.54%, stomach: 87.94% ± 2.27%). ConvNeXt-

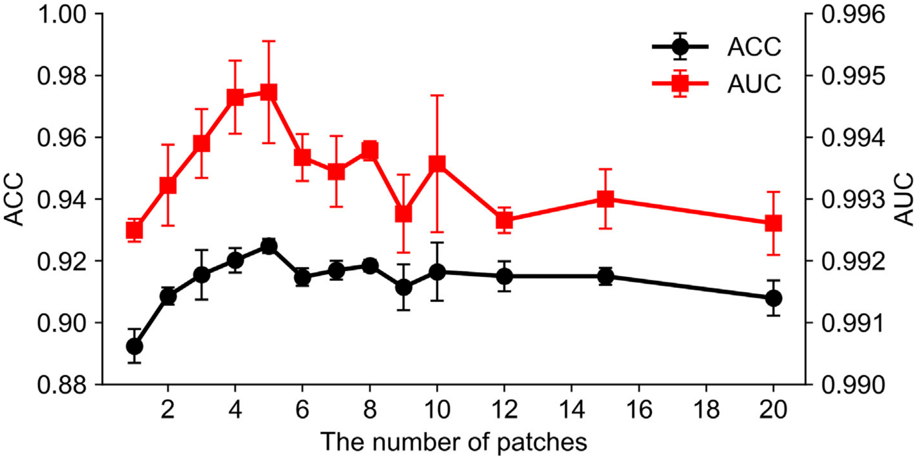

***Fig. 2:* The influence of the number of patches in a bag.** In order to realize end-to-end training, we randomly selected a fixed number of patches from every bag for model training. To obtain the optimum number, a series of numbers of patches were investigated with associated 95% confidence intervals. n = 2720 ROI.

tiny also performed best on four of the eight precision metrics (thyroid: 94.51% ± 1.89%, kidney: 91.50% ± 0.31%, stomach: 91.33% ± 1.83%, intestine: 89.71% ± 2.43%). Moreover, ConvNeXt-tiny achieved the highest ACC of 92.52% ± 0.24% and micro-AUROC of 99.51% ± 0.06% among the networks. The comprehensive research results suggested that ConvNeXt-tiny was the optimum encoder of our model.

The influence of sex on the performance of the model

The influence of sex on the performance of the model was shown in Fig. 4. Without inputting the information of sex, the AUROC of bone metastatic breast cancer was 97.71% ± 0.55%. While it raised to 98.80% ± 0.21% with P value of 0.023 (t test) after adding sex as input. Sex also had a positive influence on the prediction of bone metastatic prostate carcinoma. With or without sex, the AUROCs were 99.27% ± 0.27% and 98.37% ± 0.31% respectively (P = 0.014 with t test). Furthermore, the overall performance of the model with sex as input was superior to that of the model without sex as input. With or without sex, the ACCs were 92.51% ± 0.29% and 90.60% ± 1.08% respectively (P = 0.029 with t test).

The overall performance of the model

The optimal number of selected patches was set to five and ConvNeXt-tiny was the most appropriate backbone of our model. Besides, we integrated the sex information into the region-level feature before the final classification layer. With ten-fold cross validation experiments on dataset A, our model achieved an ACC of 93.31% ± 0.48% and a micro-AUROC of 99.57% ± 0.03% at ROI level. More details of each individual primary site were shown in Fig. 5a–c. The top-k predictions were useful to narrow down potential OBMC in clinical and reduce the number of ancillary tests necessary for identifying the primary sites. Consequently, our approach provided pathology doctors with the top3 possible OBMCs and the Top3-ACC of our model was 99.05% ± 0.13%. At the case level, the model achieved an ACC of 91.35% ± 2.94% and a top3-ACC of 97.75% ± 1.64%.

Generalization on the external cohort

To assess the generalizability of our method across various healthcare centers with different staining protocols and image scanners, we built dataset B from an external medical center (Table 1). At the ROI level, our model produced an ACC of 73.96% ± 1.23% and a top3-ACC of 95.21% ± 0.43%. At the case level, the ACC was 81.38% ± 1.30% and the top3-ACC was 97.44% ± 0.55%. Compared with the significant drop from 91.35% ± 2.94% to 81.38% ± 1.30% (P < 0.01 with t test) in ACC, the top3-ACC did not decrease notably (test: 97.75% ± 1.64%, external test: 97.44% ± 0.55%, P = 0.76 with t test). Our model displayed good adaptability to varied staining protocols and image scanners in terms top3-ACC. More detailed performance of RMIL on the external test was displayed in Supplementary Fig. S2.

We then calculated the metrics of TOAD.[27] In the within-cohort, the ACC of patient level was 72.25% ± 5.06% and the top3 accuracy was 92.60% ± 2.89%. In the external cohort, the ACC of patient level was 71.83% ± 4.41% and the top3 ACC was 93.40% ± 1.22%. The comparison of the results of RMIL versus TOAD implied the regions labeled by pathologists and the features trained by bone metastatic cancers could improve the accuracy of predictions. More details could be available in Supplementary Figs. S3 and S4.

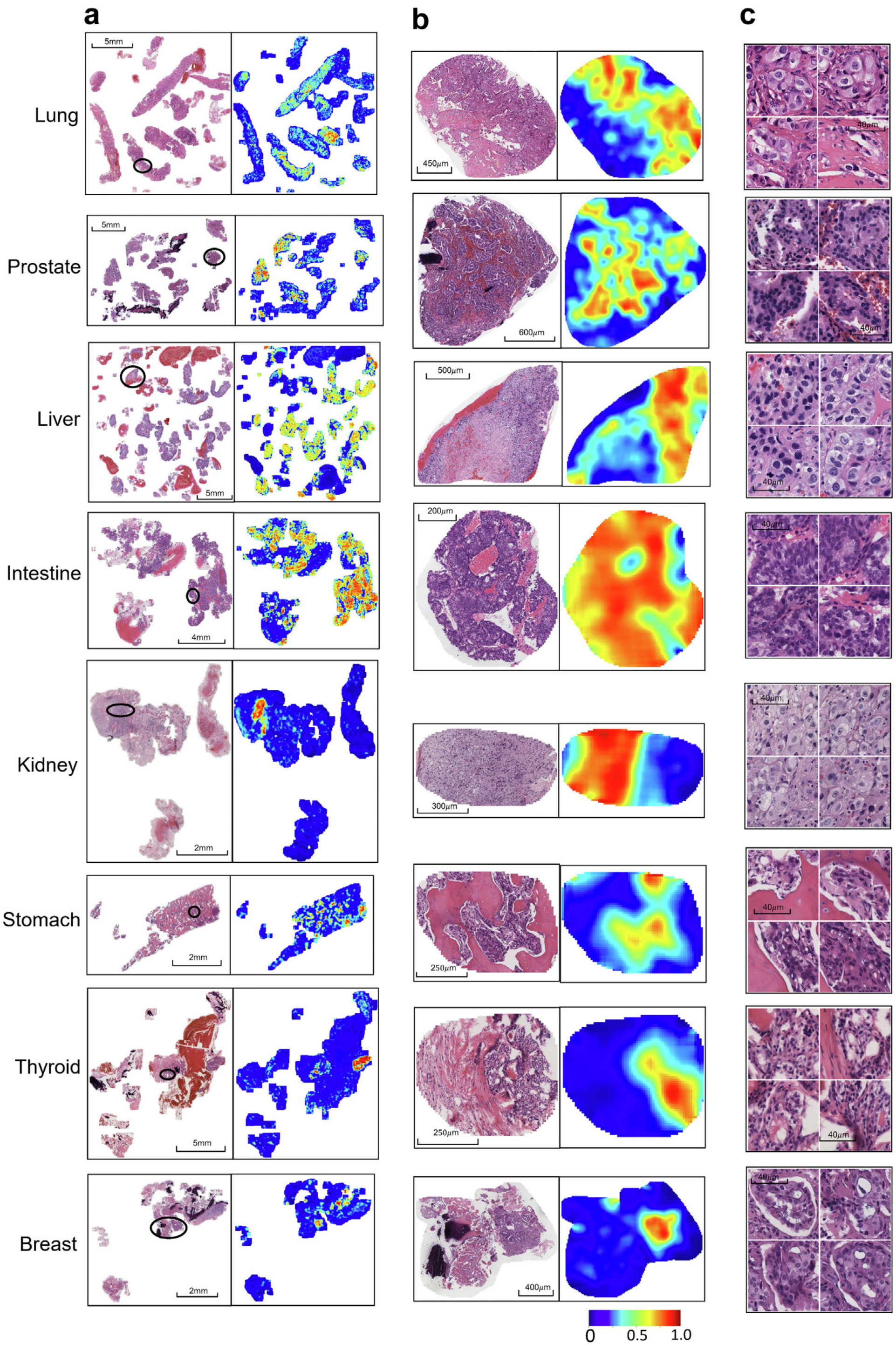

Fig. 3: **Attention heatmaps.** (a) Attention heatmaps of the WSIs. (b) Attention heatmaps of the ROI. These ROI can be found in the black marks of their corresponding left WSIs. (c) Metastatic carcinomas from the regions with high attention scores. (a)–(c) the OBMCs from top to bottom are lung, prostate, liver, intestine, kidney, stomach, thyroid and breast.

Recall (%)

	Res-34	Res-50	Res-101	Conv-B	Conv-S	Conv-T	Swin-T
Breast	83.43 ± 3.08	79.79 ± 4.84	82.18 ± 2.17	88.89 ± 2.96	84.96 ± 5.57	88.60 ± 2.14	**91.38 ± 2.99**
Prostate	83.78 ± 3.86	82.60 ± 3.28	78.47 ± 2.40	85.99 ± 3.10	84.66 ± 2.02	**89.53 ± 2.02**	87.17 ± 1.78
Thyroid	92.48 ± 1.14	92.16 ± 2.75	90.69 ± 0.45	**93.79 ± 1.38**	93.14 ± 3.92	93.79 ± 2.32	92.32 ± 0.26
Lung	90.46 ± 1.28	88.82 ± 2.91	89.26 ± 2.00	95.14 ± 0.57	95.55 ± 0.99	**95.93 ± 0.80**	92.91 ± 2.01
liver	82.63 ± 3.27	85.19 ± 2.41	83.14 ± 1.28	83.01 ± 2.75	**89.27 ± 0.94**	89.14 ± 2.66	86.97 ± 0.35
Kidney	88.83 ± 1.07	84.50 ± 5.60	83.50 ± 5.21	89.5 ± 3.33	89.00 ± 2.81	**90.17 ± 2.54**	87.33 ± 3.53
Stomach	81.56 ± 6.01	82.27 ± 1.13	75.18 ± 3.00	80.85 ± 5.90	85.11 ± 8.57	**87.94 ± 2.27**	85.82 ± 4.09
Intestine	78.27 ± 8.67	76.05 ± 2.09	69.63 ± 4.49	86.17 ± 3.44	86.67 ± 2.47	84.44 ± 2.98	**88.40 ± 2.40**

Precision (%)

	Res-34	**Res-50**	**Res-101**	**Conv-B**	**Conv-S**	**Conv-T**	**Swin-T**
Breast	78.76 ± 1.33	81.62 ± 1.96	79.94 ± 1.45	87.92 ± 1.10	**89.41 ± 3.70**	87.23 ± 3.30	82.64 ± 6.61
Prostate	82.57 ± 8.83	80.82 ± 2.58	88.27 ± 3.98	89.18 ± 3.12	**93.99 ± 1.62**	88.48 ± 4.10	92.49 ± 0.30
Thyroid	88.72 ± 1.67	83.04 ± 3.87	82.83 ± 3.31	90.50 ± 5.03	92.93 ± 4.00	**94.51 ± 1.89**	92.42 ± 3.32
Lung	94.67 ± 1.23	93.83 ± 2.67	92.11 ± 1.07	94.09 ± 1.65	93.83 ± 1.18	95.32 ± 1.14	**95.36 ± 1.17**
liver	86.58 ± 4.08	80.87 ± 4.22	76.15 ± 5.60	**91.22 ± 2.40**	85.29 ± 3.62	87.95 ± 1.21	89.42 ± 2.54
Kidney	81.42 ± 1.75	80.72 ± 6.63	79.38 ± 10.14	89.33 ± 1.20	90.95 ± 2.82	**91.50 ± 0.31**	91.10 ± 4.20
Stomach	65.33 ± 8.83	68.43 ± 3.72	61.92 ± 14.19	83.38 ± 14.77	90.49 ± 3.50	**91.33 ± 1.83**	81.08 ± 8.21
Intestine	77.61 ± 8.21	70.46 ± 4.32	81.16 ± 9.98	84.15 ± 2.09	83.02 ± 9.10	**89.71 ± 2.43**	78.21 ± 3.09

	Res-34	**Res-50**	**Res-101**	**Conv-B**	**Conv-S**	**Conv-T**	**Swin-T**
ACC (%)	87.53 ± 1.31	85.99 ± 0.43	85.34 ± 0.71	91.21 ± 0.71	91.41 ± 0.96	**92.52 ± 0.24**	90.87 ± 0.81
AUROC (%)	98.73 ± 0.13	98.64 ± 0.07	98.42 ± 0.15	99.39 ± 0.06	99.38 ± 0.03	**99.51 ± 0.06**	99.35 ± 0.10

Note: Res-34: ResNet-34. Res-50: ResNet-50. Res-101: ResNet-101. Conv-B: ConvNeXt-base, Conv-S: ConvNeXt-small. Conv-T: ConvNeXt-tiny. Swin-T: Swin Transformer-tiny. We reported the metrics of recall, precision, ACC and AUROC of each primary site with 95% confidence intervals on the test set (n = 2720 ROI). The number in bold in each row represents the high test score among seven feature extracting models.

***Table 2:* The performance of the model across different backbones.**

Evaluation on unknown OBMC cases

To verify our method of helping pathologists determine potential OBMCs and narrow down necessary tests, we evaluated RMIL on another 175 cases of unknown OBMCs. Combining H&E morphology and the top3 origins predicted by our model, pathologists made new slides from the same blocks to perform extra IHC tests. Results of 142 cases in Fig. 5g illustrated that the predictions of the model highly agreed with the origins suggested by IHC markers (K = 73.14% ± 2.56%), and

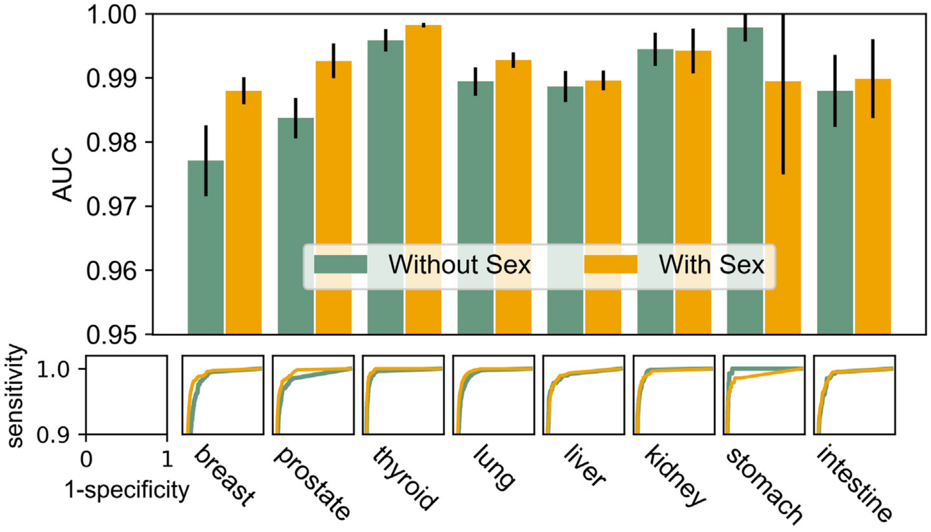

***Fig. 4:* The influence of sex on the performance of the model.** The top figure demonstrates the AUROCs (95% confidence intervals) of every primary site with or without the sex as an input of the model. The lower row shows the ROC curves of each primary site with or without sex as an input to the model. n = 2720 ROI.

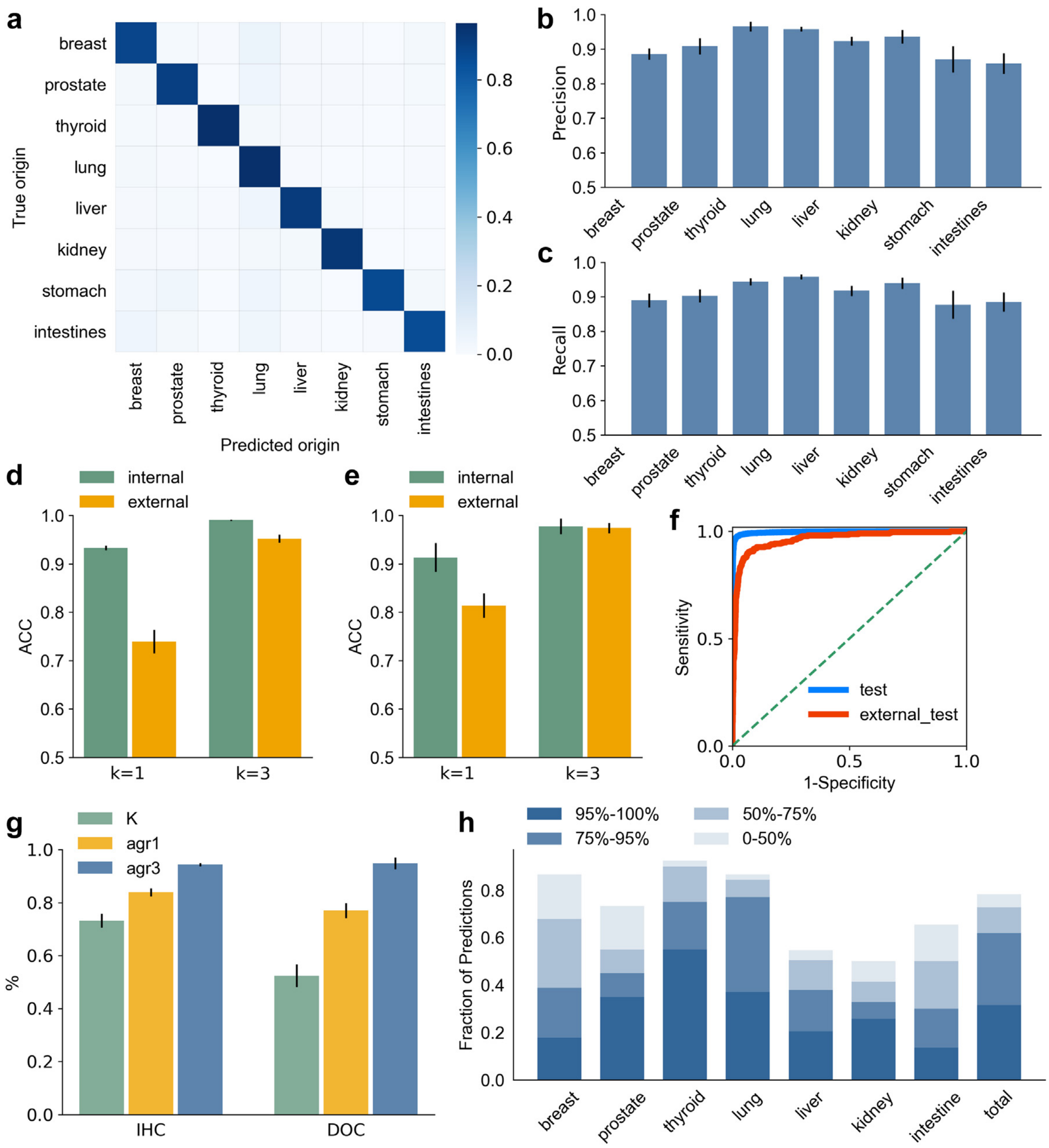

Fig. 5: **The performance of the model.** (a) The ROI-level confusion matrix of predictions of the model. Given to the imbalance in class distribution, the value of each row in the confusion matrix was divided by the total number of this category. (b) The ROI-level precision of each primary site. (c), The ROI-level recall of each primary site. (a)–(c) n = 2720 ROI. (d) The ROI-level top-k accuracies for the predictions of the OBMCs on the test set (n = 2720 ROI) and the external set (n = 383 ROI). (e) The case-level top-k accuracies for the predictions of the OBMCs on the test set (n = 136 cases) and the external set (n = 43 cases). (f) Micro averaged one-versus-rest ROC curves for the classification of the OBMC, evaluated on the test set (n = 2720 ROI) and the external test set (n = 383 ROI). The micro averaged AUROC was 99.57% (95% CI: 99.54%, 99.60%) on the test and 95.19% (95% CI: 94.81%, 95.57%) on the external test. (g) We calculated the consistency between the predictions of the model and the results indicated by IHC tests on 145 unknown cases (IHC). For the left 33 cases which could not be indicated by IHC, we assessed the agreement between the predictions of the model and the judgements made by pathologists according to the morphologies of H&E slides (DOC). The metrics of agreement include Cohen Kappa score (K), top1 agreement (agr1), and top3 agreement (agr3). (h) The fractions of samples (y axis) that were correctly classified at or above a certain confidence threshold. Due to the limited cases, seven primary sites were included in this analysis (n = 142 cases). (b)–(e), (g) Error bars indicate 95% confidence intervals.

the agreement achieved 83.90% ± 1.57% and 94.33% ± 0.62% respectively when taking into account top1 and top3 predictions. For the left 33 cases whose origins could not be indicated by the IHC markers, pathologists made determinations based on the morphology and their experience. The consistency between the predictions of the model and the pathologists' judgments was 76.97% ± 2.83% for top1 94.85% ± 2.18% for top3, and 52.37% ± 4.25% for K value.

We further analyzed the confidence of top1 predictions on dataset C. As shown in Fig. 5h, 32.37% of the cases had a confidence level greater than 95%, 31.65% had a confidence level between 75% and 90%, 17.27% had a confidence level between 50% and 75%, and 5.76% had a confidence level lower than 50%. The accuracy of top1 predictions with a confidence level lower than 50% was only 22.06% ± 1.50%, while the accuracies of the results with three other confidence levels were all 100% (supplementary Fig. S5). The confidence of top3 predictions on dataset C was displayed in Fig. 6a. 31.58% of the cases had a confidence level greater than 95%, 32.17% had a confidence level between 75% and 90%, 15.8% had a confidence level between 50% and 75%, and 16.62% had a confidence level lower than 50% which significantly increased the agreement by over 10 points (from 83.90% of top1 to 94.33% of top3).

Top3 predictions

We then calculated the frequency of each primary site occurred in top3 predictions. The results in Fig. 6b showed that among eight sites, lung occurred most frequently (24.39% ± 2.32% on the test set and 28.44% ± 0.78% on the unknown-test set), breast occurred the second (18.33% ± 2.96% on the test set and 20.38% ± 5.09% on the unknown-test set) followed by liver (13.64% ± 3.94% on the test set and 13.29% ± 2.11% on the unknown-test set), which were accordant with the proportions of cases with

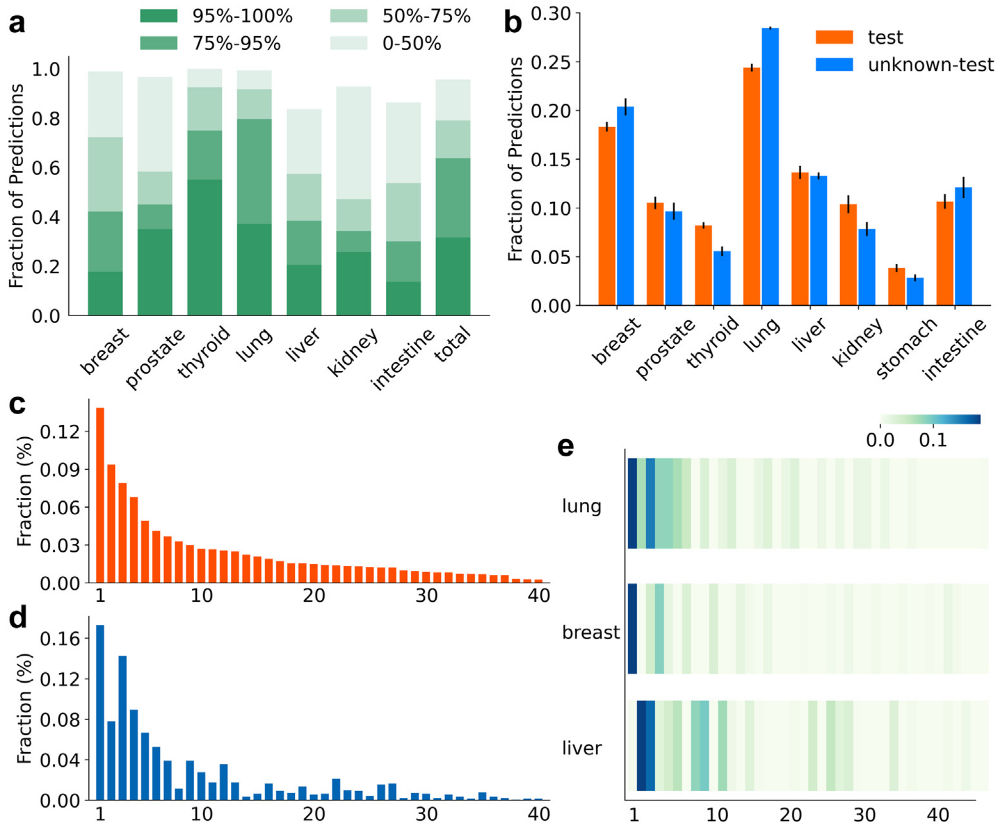

***Fig. 6:* Top3 Predictions.** (a) The fractions of samples (y axis) that were correctly predicted in top3 results at or above a certain confidence threshold. Due to the limited cases (n = 142 cases), seven primary sites were included in this analysis. (b) The frequency of each primary site occurred in tops3 predictions. The test set contained 136 cases and the unknown-test set contained 142 cases. Error bars indicate 95% confidence intervals. (c) The frequency of the combinations of three origins occurred in top3 predictions on the test set (n = 136 cases). (d) The frequency of the combinations of three origins occurred in top3 predictions on the unknown set (n = 142 cases). (e) The frequency of combinations of three origins occurred in top3 predictions when the ground truths were lung, breast, and liver respectively (n = 278 cases). (c)–(e) The x-axis means the combination of three primary sites which was available in supplementary Table S4.

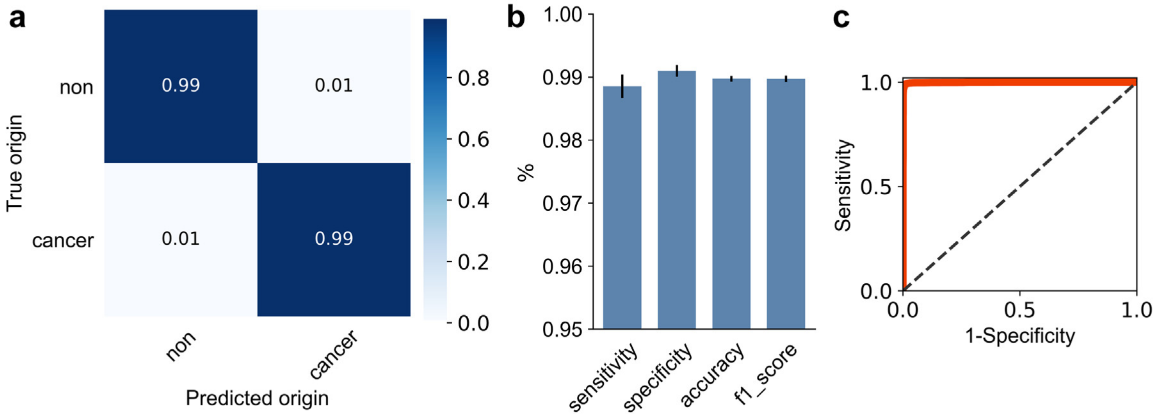

Fig. 7: **The performance of the model to predict whether there are metastatic carcinomas in bone tissue.** (a) The ROI-level confusion matrix of predictions of the model. The value of each row in the confusion matrix was divided by the total number of this category. (b) The ROI-level sensitivity, specificity, accuracy and f1-score of the model. The positive category refers to the ROI with metastatic carcinomas. Error bars indicate 95% confidence intervals. (c) The ROC curve with an AUROC of 99.90% (95% CI: 99.89%, 99.91%). (a)–(c), n = 2734 ROI.

corresponding origins. To figure out which three combinations were constantly predicted by our model, we measured the frequency of 56 combinations from eight primary sites. The combination of breast, lung and intestine was the most common in top3 predictions on the test set (13.87%), as well as on the unknown-test set (17.30%). When predicting the cases with a ground truth of lung, our model also provided breast, lung and intestine (19.68%) as the most common combination. More details could be available in Fig. 6, Supplementary Tabale S4, and Supplementary Fig. S6.

Predicting whether there is metastatic cancer in bone tissue

Pathologists will evaluate whether there is a metastatic tumor in bone tissue according to the histomorphology before identifying the OBMC. With the same framework as RMIL, we obtained a sensitivity of 98.85% ± 0.19%,a specificity of 99.10% ± 0.09%, an ACC of 98.98% ± 0.09%, and an f1-score of 98.97% ± 0.05%. The performance proved that it could act as an RMIL-aided tool to locate tumor areas, reducing the workload of labeling ROI. Individual category assessments were available in Fig. 7.

Identification of cancer type

RMIL was able to distinguish adenocarcinoma, squamous cell carcinoma, and neuroendocrine tumor. It produced an accuracy of 93.85% ± 0.73% and a micro-AUROC of 98.83% ± 0.24% (Fig. 8), manifesting its broad prospects in the differentiation of cancer type.

Discussion

Identifying the OBMC is critical for the selection of optimal method of treatments for patients, such as

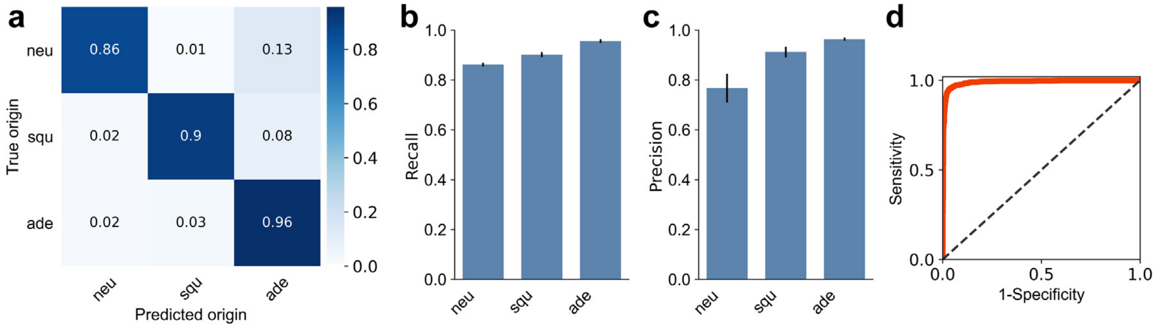

Fig. 8: **The performance of the model to classify adenocarcinoma, squamous, and neuroendocrine carcinoma.** (a) The ROI-level confusion matrix of predictions of the model. The value of each row in the confusion matrix was divided by the total number of this category. (b) The ROI-level precision of each category. (c) The ROI-level recall of each category. (a)–(c) "ade" represents adenocarcinoma, "squ" represents squamous cell carcinoma, and "neu" represents neuroendocrine tumor. (b)–(c) Error bars indicate 95% confidence intervals. (d) The ROC curve with an AUROC of 98.83% (95% CI: 98.59%, 99.07%). (a)–(d) n = 1090 ROI.

target therapeutics, chemotherapy, radiotherapy and immunotherapy.[35–38] We designed a deep learning-based algorithm to determine the OBMC by routine H&E histology along with sex information. Practically, when detailed clinical data of patients with bone metastases are not available, dozens of IHC tests are required to make diagnoses. For cases with tiny bone biopsies, sufficient bone tissues for necessary IHC tests are unavailable. Our model trained on thousands of samples can make reasonable predictions of the OBMC and greatly simplify the laborious and time-consuming work-ups to find primary sites of tumors. We labeled 26,431 ROI on H&E slides as input and implemented end-to-end model training. The top1 prediction of the OBMC of high accuracy is of great reference value in primary hospitals where advanced imaging equipment and complete IHC reagents are not available. The top3 predictions can also narrow down the origins of some complex cases which are challenging for pathologists to identify directly from morphology.

Instead of feeding all patches of a bag into the model pretrained on the dataset of natural images as mentioned by Lu et al., we randomly selected a certain number of patches in a bag. In the previous work of embedding MIL algorithms,[27,39–42] the number of patches in a bag was so large that all parameters of the feature extractor had to be frozen, which only extracted basic features. However, in our study, we aimed to train the feature extracting network by bone metastatic cancer images so as to make more accurate predictions. Large number of patches in a bag (e.g., greater than 256) will make the speed of training very slow.[41,43] The changing number of images in a batch also slow down the speed of training. A fixed number of patches in a bag can solve this problem. In addition, we are able to set batch size to 16 or even greater, which makes the training converge more smoothly.

In our study, the excellent performance of ConvNeXt might be related to the enlargement of the receptive filed, the expansion of channels, and the increasement of the depth. When comparing the characteristics of metastatic cancer from different origins, it is necessary to consider both histological features of tumor (e.g., nest structure, glandular structure, sieve texture, strip texture and necrosis) and microscopic characteristics of cell (e.g., mitosis, nuclear size and shape). With a larger receptive field, we are able to obtain comprehensive tissue level features. Also, wider channels ensure ConvNeXt to grasp more abundant features and deeper blocks help it project pathological image features to higher dimensions to solve complex classification problems.

Breast and prostate cancers have obvious sex tendencies. In 182 cases of breast cancer bone metastases, 181 patients were female; and in 82 cases of prostate cancer bone metastases, all of them were male. With sex as input, almost the performance metrics of all origins were more or less improved except stomach (Fig. 4). The number of stomach cancer cases might have an impact on this phenomenon. Sex is benefit for the origin determination of poorly differentiated cases. The morphologies of poorly-differentiated lung cancer and breast cancer were hard to be distinguished, which turned out be easier for the model with the sex of male as input.

Considering the great challenge to predict the OBMC from morphology with the top1 result, top3 results also exhibited valuable and practical significance. For example, the origin of a case on dataset C was predicted as intestine with a confidence level of 54%, lung of 36% and liver of 4.5% (Supplementary Fig. S7). Pathologists made corresponding IHC tests depending on the top3 origins predicted by the model and morphology. This case was diagnosed as metastatic enteric-type adenocarcinoma of the lung with supportive IHC staining reaction: positive for CK20, CDX2, CK7 and TTF-1. Enteric-type adenocarcinoma has striking histomorphology and immunophenotypic similarities with colorectal adenocarcinomas, but has its own specific features on the level of molecular pathologic mechanisms, which rarely occur in the areas of upper respiratory tract and lung.[44,45] After carefully examining the training set, we also found an enteric-type adenocarcinoma of lung case. It implied that top3 predictions could provide more comprehensive origin indications and remind doctors of information they might have neglected.

The limitation of our method is that pathologists are required to label ROI before RMIL makes predictions. However, in the studies of Lu and other researchers,[27,42,43] ROI are not required. Although our model displayed good generalizability on the external dataset, we need to collect more data from other widely distributed medical centers to examine the model. Another limitation of our study is that in our datasets, we did not include the cases beyond eight origins because the number of such cases were really small. In the dataset of unknown origins, these cases might occur, however our model would still predict the origin as one of the eight types. Of course, with more and more data collected, “other cancer types” will be included in our dataset or even be separately subdivided into uterine cancer, ovarian cancer, etc.

The developed model will be applied broadly in the secondary hospitals, tertiary hospitals and rural hospitals with the promotion of digital scanners. This method is easily embedded in mature labeling software matched with scanners. After digitizing slides, clinicians and pathologists can use the analysis function to predict the top3 possible OBMCs. Combining the predictions of our model and medical information, doctors are able to make accurate diagnoses with less resources and time.

In conclusion, our research serves as an accurate assistive tool for pathologists to predict OBMC with only routine H&E slides and the sex of the patient.

Contributors

A.J.H., L.H.Z., Y.H.H., and H.J.S. conceived the study. L.H.Z., A.J.H., H.J.S., and Y.H.H. designed the experiments. L.H.Z., C.J.W., and Y.Q.L. completed all codes of data processing, model training and testing. L.H.Z., H.J.S., and C.J.W. performed the experiments analysis. L.H.Z., T.G., S.S.S., Y.Q.L., T.T.H., L.Y.W, L.L.G., and Y.P.N. curated dataset A. J.R.C curated dataset B. H.F.D., T.G., S.S.S., T.T.H., and X. G. curated dataset C, dataset D, and dataset E. H.D., F.J.M. and W.L.Z. annotated regions of interest. H.J.S., A.J.H., H.D. and H.T.W. made judgements of the left 33 cases which could not be indicated by IHC markers. F.F.Z., H.T.W., and L.Y.W. counterstained old slides, as well as prepared new H&E slides. L.H.Z. and H.J.S. prepared the manuscript. Y.H.H., A.J.H., T.G., and R.A.Y. revised the writing. A.J.H., Y.H.H. and T.G. supervised the research. A.J.H., L.H.Z., H.J.S., C.J.W. and S.S.S. have accessed and verified data. All authors have read and agreed to publish the paper.

Data sharing statement

The pathological images used in this paper is not publicly available. However, they can be obtained by contacting the corresponding author (hananjia@mail.sysu.edu.cn) for scientific research purposes. Our method was developed in the framework of PyTorch. The source codes are available at https://drive.google.com/file/d/1Z7a8VltHqlTt83d3KVBpvxsyXQhK2QK5/view?usp=sharing.

Declaration of interests

There are no competing interests in this work.

Acknowledgements

We thank Qiang Huang for providing slide scanners; Zhuobin Liang, Qiming He, and Jiawen Li for supplying interfaces to process raw data; Maomao Zeng for the contrastive suggestions on the design of the experiments; This work was supported in part by National Science Foundation of China (61875102 and 61975089), Natural Science Foundation of Guangdong Province (2021A15-15012379 and 2022A1515012550), Science and Technology Research Program of Shenzhen City (JCYJ20200109110606054 and WDZC20200821141349001), and Tsinghua University Spring Breeze Fund (2020Z99CFZ023).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2022.104426.

References

1 von Moos R, Costa L, Gonzalez-Suarez E, Terpos E, Niepel D, Body JJ. Management of bone health in solid tumours: from bisphosphonates to a monoclonal antibody. *Cancer Treat Rev*. 2019;76:57–67.

2 Coleman RE, Croucher PI, Padhani AR, et al. Bone metastases. *Nat Rev Dis Prim*. 2020;6:83.

3 Oster G, Lamerato L, Glass AG, et al. Natural history of skeletal-related events in patients with breast, lung, or prostate cancer and metastases to bone: a 15-year study in two large US health systems. *Support Care Cancer*. 2013;21:3279–3286.

4 Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. *Clin Cancer Res*. 2006;12:6243s–6249s.

5 D'Oronzo S, Coleman R, Brown J, Silvestris F. Metastatic bone disease: pathogenesis and therapeutic options Up-date on bone metastasis management. *J Bone Oncol*. 2019;15:1–12.

6 Coleman R, Hadji P, Body JJ, et al. Bone health in cancer: ESMO clinical practice guidelines. *Ann Oncol*. 2020;31:1650–1663.

7 Gutzeit A, Antoch G, Kuhl H, et al. Unknown primary tumors: detection with dual-modality PET/CT - initial experience. *Radiology*. 2005;234:227–234.

8 Penson A, Camacho N, Zheng YY, et al. Development of genome-derived tumor type prediction to inform clinical cancer care. *Jama Oncol*. 2020;6:84–91.

9 Rassy E, Pavlidis N. Progress in refining the clinical management of cancer of unknown primary in the molecular era. *Nat Rev Clin Oncol*. 2020;17:541–554.

10 Destombe C, Botton E, Le Gal G, et al. Investigations for bone metastasis from an unknown primary. *Joint Bone Spine*. 2007;74:85–89.

11 Chen H, Zhang Y, Zhang WH, et al. Low-dose CT via convolutional neural network. *Biomed Opt Express*. 2017;8:679–694.

12 Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Z Med Phys*. 2019;29:102–127.

13 Shen YQ, Shamout FE, Oliver JR, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat Commun*. 2021;12:1–13.

14 Chuang WY, Chang SH, Yu WH, et al. Successful identification of nasopharyngeal carcinoma in nasopharyngeal biopsies using deep learning. *Cancers*. 2020;12:1–11.

15 Saldanha OL, Quirke P, West NP, et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat Med*. 2022;28:1232–1239.

16 Xie XF, Fu CC, Lv L, et al. Deep convolutional neural network-based classification of cancer cells on cytological pleural effusion images. *Modern Pathol*. 2022;35:609–614.

17 Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*. 2020;21:233–241.

18 Huang SC, Chen CC, Lan J, et al. Deep neural network trained on gigapixel images improves lymph node metastasis detection in clinical settings. *Nat Commun*. 2022;13:3347.

19 Yao JW, Zhu XL, Jonnagaddala J, Hawkins N, Huang JZ. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med Image Anal*. 2020;65:1–14.

20 Szegedy C, Liu W, Jia YQ, et al. Going deeper with convolutions. *Proc Cvpr Ieee*. 2015;1–9.

21 He K, Zhang X, Ren S, Sun J. *Deep residual learning for image recognition. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016:1–9.

22 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *Com Sci*. 2014:1–14.

23 Bejnordi BE, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA, J Am Med Assoc*. 2017;318:2199–2210.

24 Chuang WY, Chen CC, Yu WH, et al. Identification of nodal micrometastasis in colorectal cancer using deep learning on annotation-free whole-slide images. *Modern Pathol*. 2021;34:1901–1911.

25 Pham HHN, Futakuchi M, Bychkov A, Furukawa T, Kuroda K, Fukuoka J. Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach. *Am J Pathol*. 2019;189:2428–2439.

26 Knuutila JS, Riihil P, Karlsson A, et al. 266 Identification of metastatic primary cutaneous squamous cell carcinoma using artificial intelligence analysis of whole slide images. *J Invest Dermatol*. 2021;141:S194.

27 Lu MY, Chen TY, Williamson DFK, et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature*. 2021;594:106–110.

28 Xu Y, Mo T, Feng QW, Zhong PL, Lai MD, Chang EIC. Deep learning of feature representation with multiple instance learning for medical image analysis. *ICASSP*. 2014:1626–1630.

29 Ilse M, Tomczak JM, Welling M. *Attention-based deep multiple instance learning. International conference on machine learning. PMLR*. 2018:2127–2136.

30 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. *ICCV*. 2021:9992–10002.

31 Liu Z, Mao HZ, Wu CY, Feichtenhofer C, Darrel T, Xie S. *A ConvNet for the 2020s. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022:11966–11976.

32 Xu G, Song ZG, Sun Z, et al. CAMEL: a weakly supervised learning framework for histopathology image segmentation. *ICCV*. 2019:10681–10690.

33 Jadon S. A survey of loss functions for semantic segmentation. In: *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (Cibcb)*. 2020:115–121.

34 McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22:276–282.

35 Massard C, Loriot Y, Fizazi K. Carcinomas of an unknown primary origin—diagnosis and treatment. *Nat Rev Clin Oncol*. 2011;8:701–710.

36 Sturge J, Caley MP, Waxman J. Bone metastasis in prostate cancer: emerging therapeutic strategies. *Nat Rev Clin Oncol*. 2011;8(6): 357–368.

37 Clemons M, Gelmon KA, Pritchard KI, Paterson AHG. Bone-targeted agents and skeletal-related events in breast cancer patients with bone metastases: the state of the art. *Curr Oncol*. 2012;19(5):259–268.

38 Tsuya A, Fukuoka M. Bone metastases in lung cancer. *Clin Calcium*. 2008;18(4):455.

39 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;5:555–570.

40 Chikontwe P, Kim M, Nam SJ, Go H, Park SH. Multiple instance learning with center embeddings for histopathology classification. *Med Imag Comput Comput Assist Intervention (MICCAI)*. 2020;12265:519–528.

41 Chen Z, Zhang J, Che SL, Huang JZ, Han X, Yuan YX. Diagnose like A pathologist: weakly-supervised pathologist-tree network for slide-level immunohistochemical scoring. *Aaai Conf Artif Inte*. 2021;35:47–54.

42 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 Inf Process Syst*. 2021;34:2136–2147.

43 Chen RJ, Chen C, Li Y, et al. *Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022: 16144–16155.

44 Leivo I. Intestinal-type adenocarcinoma: classification, immunophenotype, molecular features and differential diagnosis. *Head Neck Pathol*. 2017;11:295–300.

45 Magalhaes MAO, Irish JC, Weinreb I, Perez-Ordonez B. Adenosquamous carcinoma of hypopharynx with intestinal-phenotype. *Head Neck Pathol*. 2015;9:114–118.