Technology in Cancer Research & Treatment
Volume 20: 1-14
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/15330338211027901
journals.sagepub.com/home/tct

Deep Learning Models for Gastric Signet Ring Cell Carcinoma Classification in Whole Slide Images

Fahdi Kanavati[1], Shin Ichihara[2], Michael Rambeau[3], Osamu Iizuka[3], Koji Arihiro[4], and Masayuki Tsuneki[1,3]

Abstract

Signet ring cell carcinoma (SRCC) of the stomach is a rare type of cancer with a slowly rising incidence. It tends to be more difficult to detect by pathologists, mainly due to its cellular morphology and diffuse invasion manner, and it has poor prognosis when detected at an advanced stage. Computational pathology tools that can assist pathologists in detecting SRCC would be of a massive benefit. In this paper, we trained deep learning models using transfer learning, fully-supervised learning, and weakly-supervised learning to predict SRCC in Whole Slide Images (WSIs) using a training set of 1,765 WSIs. We evaluated the models on two different test sets ($n = 999$, $n = 455$). The best model achieved a ROC-AUC of at least 0.99 on all two test sets, setting a top baseline performance for SRCC WSI classification.

Keywords

signet ring cell carcinoma, gastric cancer, deep learning, weakly-supervised learning, multiple-instance learning, transfer learning

Received: March 02, 2021; Revised: May 19, 2021; Accepted: May 30, 2021.

Introduction

According to the Global Cancer Statistics 2018,[1] stomach cancer was responsible for over 1 million new cases in 2018 with an estimated 783,000 deaths, making it the fifth most frequently diagnosed cancer and the third leading cause of cancer death in the world. Importantly, incidence rates are markedly elevated in Eastern Asia (e.g., Japan and Republic of Korea), whereas the rates in Northern America and Northern Europe are generally low and are equivalent to those seen across the African regions.[1] However, a series of studies has shown that the incidence of signet ring cell carcinoma (SRCC) of stomach (a subtype of poorly cohesive carcinoma) has been slowly increasing, especially in the United States.[2-4] The great majority of SRCC occurs in the stomach, with the rest arising in other organs (e.g., breast, gallbladder, pancreas, urinary bladder, and colon).[5]

SRCC is an invasive gastric adenocarcinoma and can be accompanied by diffuse growth of adenocarcinoma cells associated with a wide range of desmoplastic reactions, in particular when infiltrating into the submucosa or beyond.[6] This type of growth is defined as diffuse cancer according to the Lauren classification.[7] In the early stage of the disease, intramucosal SRCC appears as layered cancer cells in the superficial portions of the mucosa without desmoplasia.[8-10] The typical signet-ring cells contain intracytoplasmic mucin that compresses the nucleus to the periphery of the cell wall, and glandular formations are rarely observed. Due to these morphological appearances, some of the SRCC cells often appear to mimic crushed oxyntic glands, crushed mucous neck cells, the goblet cells of the intestinal metaplasia, and gastric xanthoma (histiocytic aggregation).[11] This makes SRCC more likely to be missed on routine histopathological diagnoses. False negatives have a detrimental

[1] Medmain Research, Medmain Inc., Fukuoka, Japan
[2] Department of Surgical Pathology, Sapporo Kosei General Hospital, Sapporo, Hokkaido, Japan
[3] Medmain Inc., Fukuoka, Japan
[4] Department of Anatomical Pathology, Hiroshima University Hospital, Hiroshima, Japan

Corresponding Author:
Masayuki Tsuneki, Medmain Research, Medmain Inc., Fukuoka, 810-0042, Japan.
Email: tsuneki@medmain.com

impact on the quality and accuracy of the pathological diagnosis, and it should be addressed urgently.

Computational pathology has been gaining momentum over the past decade, in particular due to the large increase in resources that allow the digitization and processing of Haematoxylin and Eosin (H&E) stained glass slides of surgical and biopsy specimens into Whole Slide Images (WSIs). Machine learning, in particular deep learning, has found many applications in computational pathology, such as cancer detection and classification, cell detection and segmentation, and gene mutation expression for a variety of organs and pathologies.[12-31]

Preparing a large fully-annotated training dataset for WSI cancer classification is a tedious, time-consuming task. This is because WSIs are extremely large, with heights and widths in the tens of thousands of pixels, as a result of being scanned at magnifications of ×20 or ×40 in order to reveal cellular-level details. The large image size makes it difficult to train and apply a CNN directly to WSIs due to GPU memory constraints. To bypass the computational constraints, the typically adopted approach is to divide the WSI into a set of fixed-sized tiles.[13,14,20,32] Training of the CNN is done by using the resulting labeled tiles as input. Classification of a WSI is done by applying the CNN in a sliding window fashion, classifying the individual tiles, then aggregating all their classification outputs into a final WSI classification. The aggregation could be as simple as taking the maximum probability output of the tiles or using an RNN model.[13,18] Obtaining a dataset of labeled tiles can either be done by asking pathologists to draw contours on WSIs or to classify pre-extracted, fixed-sized tiles. The latter requires pre-fixing the tile size and having pathologists classify millions of tiles. This is a tedious task. The former is preferable as the tile size can be modified later, and viewing the WSI provides context to pathologists and allows them to draw contours on large cancer infiltration areas; however, it can still be tedious especially with complex cancer infiltration patterns requiring annotations of individual cells. Once annotated, a single WSI can produce thousands of labeled tiles for training. A large dataset of labeled tiles is a requirement for fully-supervised learning.

On the other hand, weakly-supervised learning is an alternative approach and requires only weakly-labeled data.[33] Given that diagnoses of WSIs are readily available from reports, additional annotations by pathologists are not required. Weakly-supervised learning methods, such as multiple instance learning (MIL),[34] can operate directly on the WSIs by using the diagnoses as slide-level labels. This is a highly attractive solution. One particular advantage of MIL is that it can reduce the labeling requirement. MIL was initially proposed in the context of drug discovery,[34,35] and subsequently found many applications in computer vision,[36] including histopathology classification and segmentation.[13,20,23,30,37-41] The caveat in histopathology applications, however, is that the method tends to require a large training dataset of WSIs in order to work well. This has been demonstrated recently by Campanella *et al*[13] using a dataset of 44,732 WSIs to classify prostate cancer, basal cell carcinoma, and breast cancer metastases, with a reported Receiver Operator Curve (ROC) area under the curve (AUC) of about 0.98 on 3 test sets of about 1,500 WSIs each. They observed that at least 10,000 WSIs were necessary for training to obtain a good performance. Both weakly- and fully-supervised learning could be used on a dataset that has a combination of detailed cellular-level annotations and slide-level labels.

Only recently has SRCC detection been investigated.[42,43] Li *et al*[42] set up the MICCAI DigestPath2019 challenge where [1] task was SRCC instance detection. A training dataset was made publicly available consisting of a total of 455 images (of which 77 had SRCC). The images were crops of size 2000 × 2000 pixels extracted from WSIs. A total of 12,381 instances of SRCC were manually annotated; however, the dataset still contains unannotated instances of SRCC. Li *et al*[42] proposed a semi-supervised framework for SRCC detection where the goal was to train a deep learning network to detect individual SRCC instances using the combination of annotated and unannotated SRCC instances. The model was then evaluated on a test set consisting of 227 images (of which 12 had SRCC). The 1st runner up at the challenge proposed using a specialized loss[43] to separate the contribution of annotated and unannotated training samples resulting in an improvement in SRCC instance detection on the test set. Although there might be some interest from a research perspective in detecting all instances of SRCC in a specimen in order to calculate measurements, such as the karyoplasmic ratio or the degree of atypia, and study their correlations with outcomes. However, from a clinical perspective, all that matters is detecting whether a specimen has SRCC.

In this paper, our aim is the clinical application of detecting SRCC in WSIs. It is not quite known for this particular application which training method is the most appropriate. Annotating individual SRCC cells is a tedious task, and a method that uses minimal annotations would be more desirable if it does not involve a compromise in performance. To this end, we trained several deep learning models using a combination of transfer learning, fully-supervised learning, and weakly-supervised learning. We used a training dataset consisting of a total of 1,765 WSIs of which 100 WSIs had an SRCC diagnosis. A group of pathologists non-exhaustively annotated individual cells suspected of SRCC in all of the 100 WSIs. We performed an investigation of different training methods in order to best understand which aspects contribute to obtaining a good SRCC WSI classification given the available data.

Methods and Materials

Our proposed method for SRCC WSI classification consists of using a CNN trained on tiles extracted from WSIs and using a combination of transfer learning, fully-supervised learning, and weakly-supervised learning to train the models. Figure 1 provides an overview of the training methods.

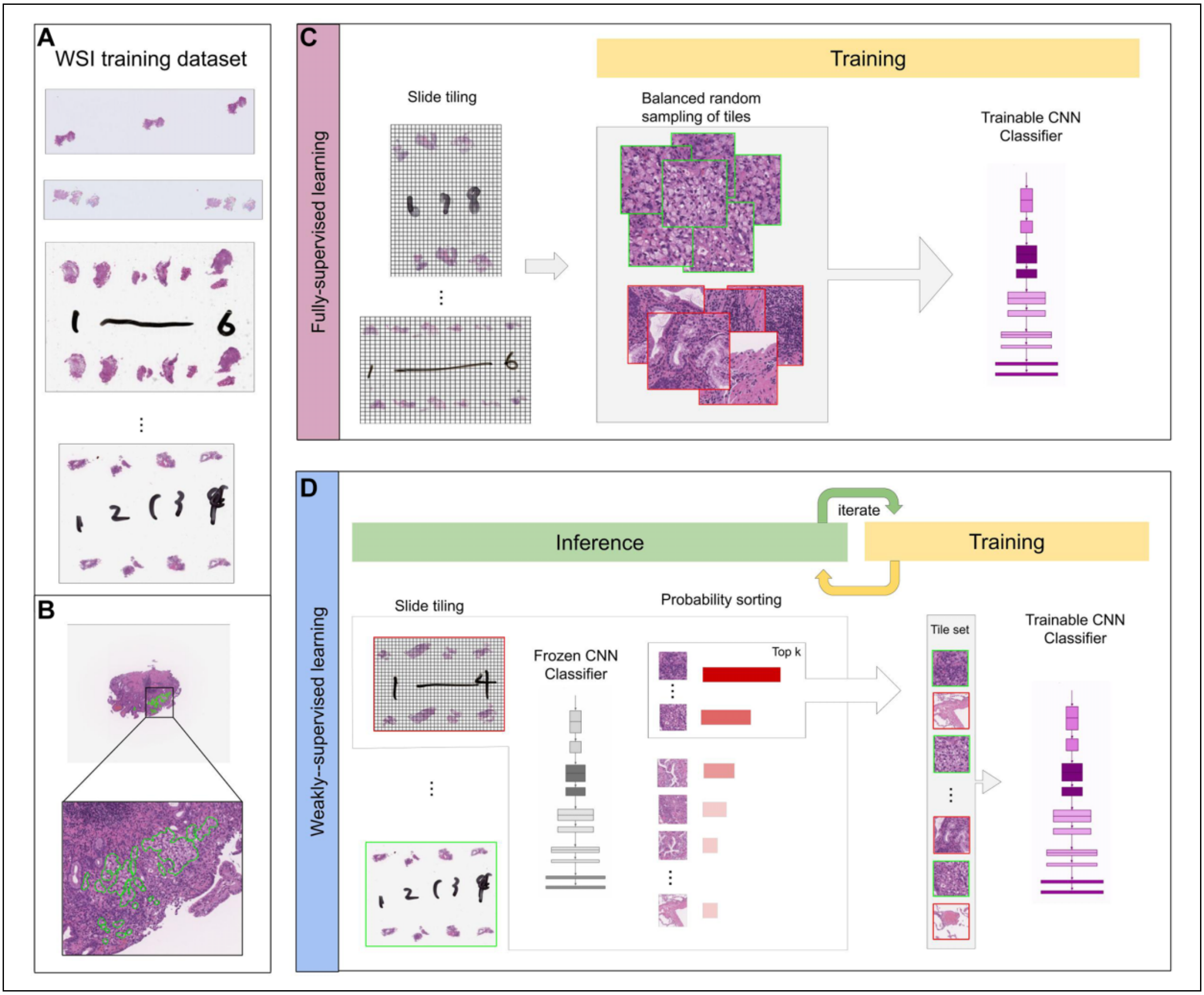

Figure 1. Overview of the training methods. A, Shows examples of biopsy WSIs in the training dataset. B, Shows an example of the SRCC annotations overlaid digitally on WSIs. The annotations were used to guide the extraction of tiles. C, Shows an overview of the fully-supervised method where balanced batches of tiles are extracted from the WSI to train the CNN classifier. D, Shows an overview of the weakly-supervised method. The method alternates between two steps: inference and training. During inference a frozen CNN classifier is run in a sliding window fashion on each WSI and the top k tiles with the highest probabilities are placed into the training tile set. Once the training tile set reaches a certain size *T*, the training step is triggered.

Problem Formulation

In histopathology, a pathologist diagnoses a WSI as having cancer if it is seen in any sub-region of the WSI; otherwise, it is diagnosed as not having cancer. This means that if a WSI with cancer were subdivided into a dense grid of smaller fixed-size tiles, then at least one of those tiles must have cancer, even though initially we do not know which tiles have cancer. If the WSI does not have cancer, then none of those tiles have cancer. This type of problem can be formulated generally with MIL. The MIL formulation adopts the concept of labeled bags that contain a collection of instances. A WSI i is considered as a bag H_i and any tile j sampled from it is considered as a instance $x_{ij} \in H_i$. In the binary setting, a bag H_i can either have a positive label ($y_i = 1$) or a negative label ($y_i = 0$). Similarly, an instance j from bag i can either have a positive label ($y_{ij} = 1$) or a negative label ($y_{ij} = 0$). The label of a bag i is positive if at least 1 instance in the bag has a positive label. If instance labels y_{ij} are known, then the bag label y_i can be obtained as $y_i = \max_j(y_{ij})$. However, in practice, as training is carried out on the instance level, the bag label y_i is used to derive the labels y'_{ij} of the instances x_{ij} for which there are no labels. The goal is then to train a model $f(x)$ that can classify all the instances. The MIL formulation allows training from data that has either purely labeled bags or a mix of labeled instances and bags, with one end of the extreme where only bags are labeled, and the other end of the extreme where all instances are labeled. In this paper we are interested in classifying SRCC, so a positive label corresponds to a WSI having SRCC, and a negative label to the absence of SRCC.

Training Methods

Fully-supervised (FS) learning. When we have labels for all the instances, there is no need to use the bag labels to derive the instance labels, and the MIL formulation becomes the classical fully-supervised (FS) learning method. The training dataset is $\{(x_{ij}, y_{ij})\}_{i=1,\ldots,N, j=1,\ldots,N(i)}$, where N is the number of WSIs and $N(i)$ is the number of tiles from the i^{th} WSI. All positive WSIs would need to be annotated, potentially at cellular-level, such that labels are available for all the tiles.

Weakly-supervised (WS) learning. When we have labels only for bags or a mix of bags and instances, we can train the model using MIL. The bag label is used to infer the label of the instances. The training alternates between 2 steps: inference and training. Using the model trained so far, the inference step is used to extract a list of candidate tiles for training. During an epoch (one sweep through the entire dataset), we perform a balanced sampling (see Sec.) of tiles by randomly selecting in turn either a positive ($y_i = 1$) or a negative ($y_i = 0$) WSI. We then run inference on the WSI in a sliding window fashion and select the top k tiles with the highest probabilities. This is done by sorting the probabilities in descending order and selecting the top k instances. From a positive WSI, the top k correspond to tiles that the model is most confident that they contain cancer and their probabilities should be closer to 1. From a negative WSI, the top k correspond to tiles that the model assigned the highest probabilities to, and they should be closer to 0. At each iteration, the top k tiles are added to the set of training tiles. Once the size of the set reaches a certain threshold T, the set is shuffled and then fed into the model as batches for training. The model can alternate between inference and training many times within an epoch (if T is less than the number of iterations in an epoch) or only once at the end of the epoch. At each iteration step, y_i alternates between 0 and 1, and for selected instances that do not have a label, they are assigned the label $y'_{ij} = y_i \forall x'_{ij} \in \hat{H}'_i(k)$, where

$$\hat{H}'_i(k) = \underset{H'_i \subset H_i, |H'_i| = k}{\operatorname{argmax}} \sum_{x_{ij} \in H'_i} f(x_{ij})$$

is the subset of top k tiles.

Weakly-supervised with fully-supervised pre-training. We can train the model by first training it with the FS method, and then refining the model further by training it for additional epochs using the WS method.

Class imbalance. The training set was highly imbalanced, where WSIs with the negative class far outnumbered WSIs with the positive class (SRCC). To improve predictive performance on the positive class, we created a balanced sampler by over-sampling tiles from the positive class. This was done by having the tile sampler alternate from picking a fixed number of tiles from either a positive or a negative WSI. For FS, k tiles are picked randomly, whereas with WS, the top k tiles are picked based on their probabilities. The over-sampling ensures that all the negative WSIs are used for training during each epoch.

Deep Learning Model

We used the EfficientNet Convolutional Neural Network (CNN) architecture,[44] which has achieved state-of-the-art accuracy on computer vision datasets while having a smaller number of parameters and a floating point operations per second (FLOPS) values that is an order of magnitude smaller compared to other existing architectures. The architecture uses compound scaling along width, depth, and image resolution of a baseline network, with mobile inverted bottleneck convolution (MBConv) as convolutional units. Different scales of EfficientNet have been trained on the ImageNet dataset.[45] We used the EfficientNet-B1 model architecture which has 7.8 M parameters.

For transfer learning (TL), we initialized the weights of all the convolutional layers with the pre-trained weights on ImageNet. The final classification layer was a fully-connected layer with single output and a sigmoid activation function, and its weights were randomly initialized using the Glorot uniform initializer.[46] During the first epoch, all the weights were frozen except for the weights of the final classification layer; this is so as to prevent random initial weight of the classification layer from destroying the pre-trained weights. After the first epoch, all the weights were unfrozen to become trainable.

Tile Extraction

Tiles were extracted on the fly from the WSIs by direct indexing of locations without loading the entire WSI into memory. For a WSI, the locations were pre-computed as follows: first, we performed tissue detection by thresholding the image using Otsu's method; this step allowed eliminating a large portion of the white background and reducing unnecessary sampling of tile instances from the background. If annotation are available, then they could be used to further reduce the valid tissue sampling regions. Then, given a stride that allows subdividing the WSI into a grid, we extracted grid cell locations only from the valid tissue regions. These grid cells location were then used to extract tiles at the desired tile size and magnifications. For all the models, we used a fixed tile size of 224×224 pixels, and a stride during training of 112×112 pixels.

As tiles were extracted from the WSIs, we randomly applied data augmentation on the fly in the form of tile flips, 90° rotations, translations, and color shifts in order increase robustness and add regularization effect to the training.

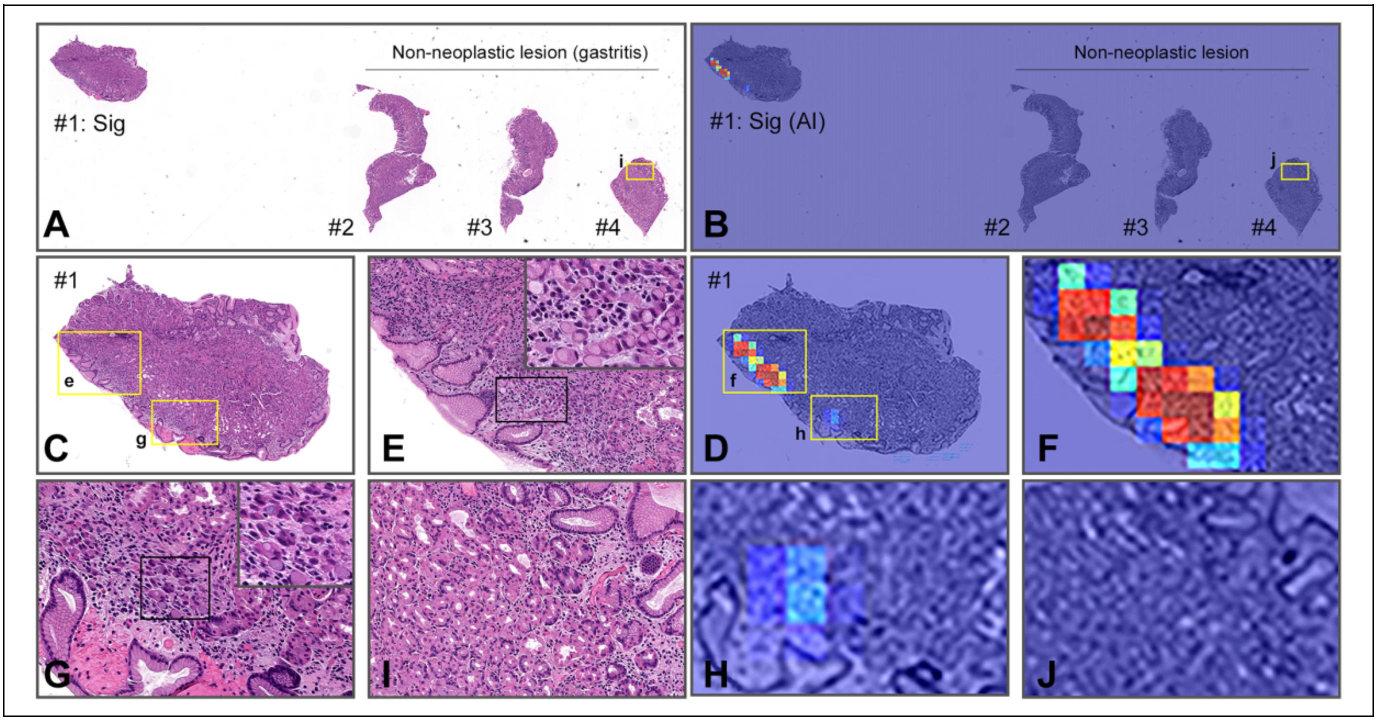

Figure 2. Representative true positive case. There are four endoscopic biopsy fragments in this WSI (A). According to the pathological diagnostic report, #1 is signet ring cell carcinoma and #2-#4 are gastritis (non-neoplastic lesion) (A). When viewed under low magnification, highlighting is visible only in #1 on heatmap image (B). When the highlighted area in #1 is magnified (C), strong and low-signal areas are seen (D); a large number of signet ring cell carcinoma cells (E) are observed in the strong-signal area (F) and a small number of signet ring cell carcinoma cells (G) are seen in the low-signal area (H). Enlargement of the tissue in #4 confirms that it does not contain any signet ring cell carcinoma cells (I and J).

WSI Classification

The models were trained as classifiers on the tile level; however, to obtain a WSI classification, the model was applied in a sliding window fashion using a stride of 112 × 112 pixels, then the WSI was assigned the label of the maximum probability of its tiles.

Heatmap Visualization

We generated two types of heatmaps from the model using two methods: classification probability and Gradient-weighted Class Activation Mapping (Grad-CAM).[47] The former consists in the tiling of the classification probability outputs by mapping each input tile's 1 × 1 classification probability to a stride × stride pixels output tile. This can result in a blocky heatmap visualization, especially with large strides. The smaller the stride, the more fine-grained the output; however, this comes at an increase in prediction time. The latter, Grad-CAM, is a method that uses the gradients of the target output flowing into the final convolutional layer to produce a coarse localization map. With the EfficientNet-B1 model, this produces a 7 × 7 output for a 224 × 224 input tile. Using a stride that is smaller than the input tile size, the outputs can be further smoothed by averaging the overlapping tiles; however, this too results in an increase in prediction time. Figures 2 and 3 show examples of probability heatmaps with a stride of 112 × 112 pixels, whereas Figures 5 and 6 show Grad-CAM visualizations with a stride of 32 × 32 pixels.

Implementation Details

The deep learning models were implemented and trained using TensorFlow.[48] We used OpenSlide[49] to read WSIs on the fly without pre-extracting all the tiles. AUCs were calculated in python using the scikit-learn package[50] and plotted using matplotlib.[51] The 95% CIs of the AUCs were estimated using the bootstrap method[52] with 1,000 iterations.

Datasets

Hospital A and B. For the present retrospective study, 2,824 cases of gastric epithelial lesions HE (hematoxylin & eosin) stained specimens, each from a distinct patient, were collected from the surgical pathology files of Hiroshima University Hospital (Hospital A) and Tokyo IUHW Mita Hospital (Hospital B) after being reviewed by surgical pathologists. The experimental protocols were approved by the Institutional Review Board (IRB) of the Hiroshima University (No. E-1316) and International University of Health and Welfare (No. 19-Im-007). All research activities complied with all relevant ethical regulations and were performed in accordance with relevant guidelines and regulations of each hospital. Informed consent to use

Table 1. ROC AUCs and Log Losses With Their Associated Confidence Intervals (CIs) for the 2 Test Sets: Hospital A and B and DigestPath2019.

		ROC AUC	log loss
Hospital A & B (n = 999)	FS ×5	0.9891 (0.9791-0.9964)	1.5268 (1.367-1.7253)
	FS ×10	0.9966 (0.9932-0.9989)	1.4356 (1.3189-1.5564)
	FS ×20	0.9931 (0.9879-0.9971)	3.0527 (2.8842-3.2541)
	WS ×10	0.9992 (0.9981-0.9999)	0.0608 (0.0492-0.0737)
	FS + WS ×10	0.9986 (0.9971-0.9996)	0.2737 (0.2357-0.322)
	FS w/o TL ×10	0.9807 (0.9703-0.9894)	1.999 (1.8697-2.1183)
	WS-noanno ×10	0.9778 (0.9611-0.9908)	0.381 (0.3231-0.4342)
DigestPath2019 (n = 455)	FS ×5	0.9724 (0.956-0.9836)	0.3667 (0.2477-0.5072)
	FS ×10	0.9868 (0.9739-0.9963)	0.1584 (0.1037-0.2235)
	FS ×20	0.9618 (0.9396-0.9769)	0.4678 (0.3869-0.5885)
	WS ×10	0.9728 (0.9486-0.9907)	0.293 (0.2063-0.3798)
	FS + WS ×10	0.9912 (0.9841-0.9974)	0.0911 (0.0636-0.119)
	FS w/o TL ×10	0.9529 (0.9211-0.9741)	0.2173 (0.1698-0.2808)
	WS-noanno ×10	0.9619 (0.9389-0.9877)	0.4207 (0.2886-0.5324)

histopathological samples and pathological diagnostic reports for research purposes had previously been obtained from all patients prior to the surgical procedures at both hospitals and an opportunity for refusal to participate in research was guaranteed by an opt-out manner.

The combined dataset obtained from both hospitals consisted of 2,824 WSIs of which were divided into sets of 1,765, 60, and 999 for training, validation, and test, respectively. The training set consisted of 100 SRCC, 571 other adenocarcinoma, and 1,094 non-neoplastic lesion, the validation set consisted of 20 SRCC, 20 other adenocarcinoma, and 20 non-neoplastic lesions, and the test set consisted of 78 SRCC, 82 other adenocarcinoma and 839 non-neoplastic lesion. Given that the goal is to train a binary classifier, the cases were grouped into SRCC vs non-SRCC (other adenocarcinoma and non-neoplastic lesions). All cases were solely composed of endoscopic biopsy specimen WSIs. The 100 SRCC WSIs were manually annotated by a group of 11 surgical pathologists who perform routine histopathological diagnoses by drawing around the areas that corresponded to SRCC. The pathologists carried out detailed cellular-level annotations on cells that fit the description of SRCC cells as defined by the World Health Organization (WHO) classification of tumors (i.e., the following three tumor cell morphologies were adopted: (1) a cell with an intracytoplasmic cyst filled with acid mucin, giving the classical signet-ring appearance; (2) a tumor cell with eosinophilic cytoplasmic granules containing neutral mucin with a slightly eccentric nucleus; and (3) a tumor cell in which the cytoplasm is distended, with secretory granules of acid mucin appearing like a goblet cell).[53,54] The other adenocarcinomas subset included the following subtypes: tubular (tub), poorly differentiated (por) and papillary (pap) types which did not include SRCC cells in WSIs.[54] The non-neoplastic subset included the following categories: ulcer, gastritis, regenerative mucosa, fundic gland polyp and almost normal gastric mucosa. Each annotated WSI was observed by at least two pathologists, with the final checking and verification performed by a senior pathologist. All the WSIs were scanned at a magnification of ×20.

DigestPath2019. The DigestPath2019 data (note 1) was obtained from the signet ring task of the DigestPath2019 grand challenge competition, part of the MICCAI 2019 Grand Pathology Challenge Li *et al*.[42] We used the provided training dataset as a test set given that the classification labels were available. The dataset consisted of 455 images from 99 patients, of which 77 images from 20 patients contained SRCC. The provided images were 2000 × 2000 pixels crops extracted at a magnification of ×40 from WSIs. The original intended task of the challenge was to detect all instances of SRCCs; however, we only perform SRCC classification on the images. The size of the images was then adjusted based on the expected magnification of a given model.

Experiments and Results

Set-Up

We trained using three different training methodologies: fully-supervised (FS), weakly-supervised (WS), and fully-supervised pre-training followed by weakly-supervised (FS-WS). This resulted in seven different models: FS ×5, FS ×10, FS ×20, FS w/o TL ×10, WS ×10, WS-noanno ×10, and FS+WS ×10.

For the FS method, we training the models using WSIs at three different magnifications ×5, ×10, and ×20. For the magnification at ×10, we trained using the FS method with and without transfer learning (w/o TL). During training the balanced tile sampler ensured that at least $k = 40$ tiles were randomly selected from each WSIs during a given epoch.

For the WS methods, we trained the models at a magnification of at ×10. In addition, we trained two versions of the model where in one version we only sampled from the annotated regions from the positive WSI, and in the other version we sampled tiles without using any of the annotations (WS-noanno). We used a top k value of 1, and $T = 128$, meaning

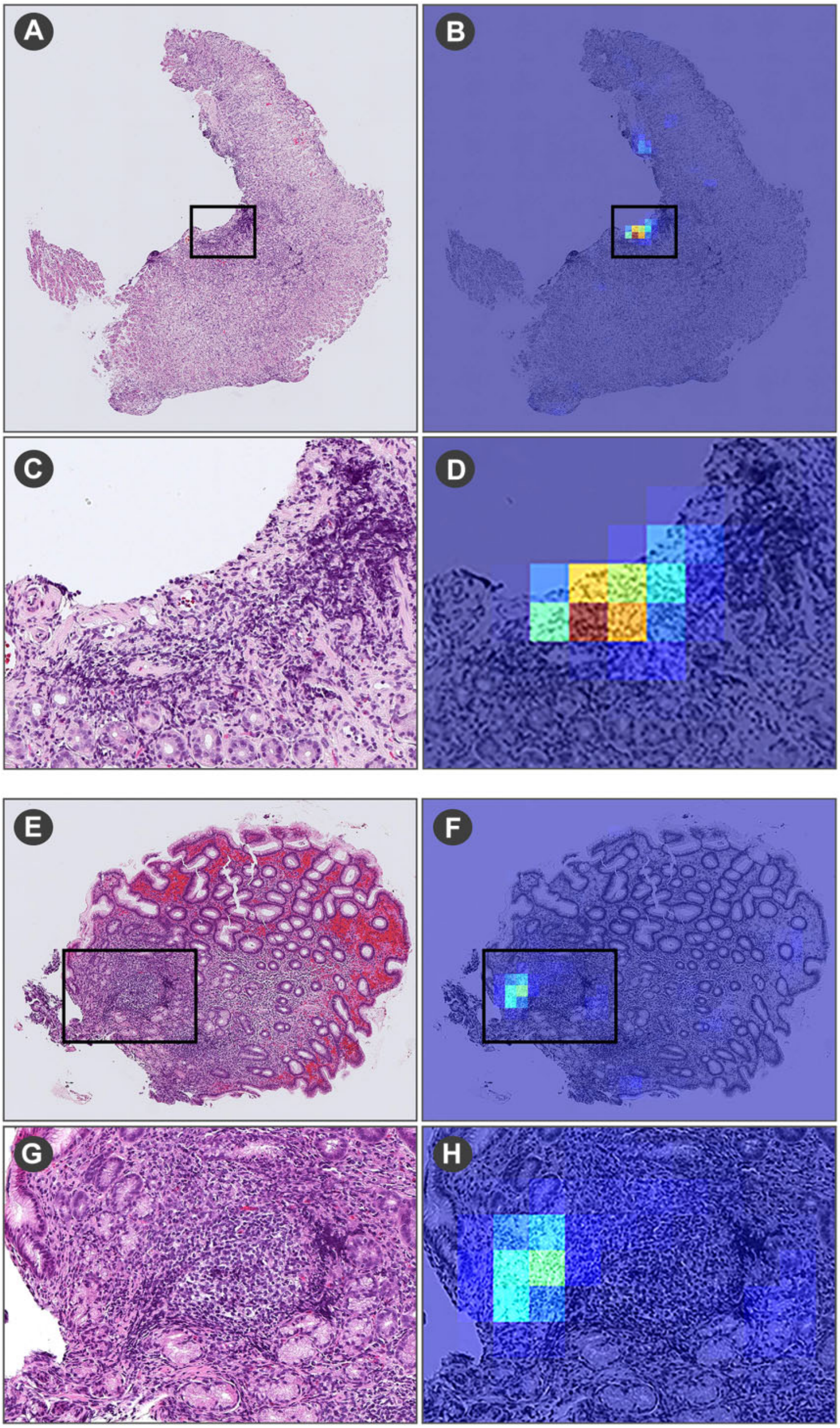

Figure 3. Representative false positive cases. A, is a case chronic gastritis (non-neoplastic lesion). A-D, Pathologically, the false positives might be due to the lymphocytes being mixed around the smooth muscle cells and blood vessels of the muscularis mucosae and the nuclear density of the lymphocytes being similar to SRCC. E, is a case of chronic gastritis (non-neoplastic lesion). E-H, The false positive area includes pyloric glands disrupted by inflammation. Pathologically, the false positive area is suggested as a pyloric gland by comparison with other adjacent pyloric gland(s). However, on practical diagnosis, if such a finding is observed, additional investigation should be performed to confirm that it is a pyloric gland.

that training was run multiple times during an epoch. To look into the effect of the top *k* parameter, we also trained multiple versions of the WS-noanno model for a range of top $k \in \{1, 5, 10, 15, 20\}$, while keeping all the other hyperparameters fixed.

We evaluated the models on two test sets: Hospital A & B (n = 999, 78 SRCC, 82 other adenocarcinoma and 839 non-neoplastic lesions) and DigestPath2019 (n = 455, 77 SRCC).

Model Hyperparameters

All models were trained with the same hyperparameters. We used the Adam optimization algorithm[55] with $beta_1 = 0.9$ and $beta_2 = 0.999$, a learning rate of 0.001 with a decay of 0.95 every 2 epochs, and the binary cross entropy loss/log loss. We used a batch size of 32. The performance of a given model was tracked on a validation set. We used an early stopping approach, to avoid overfitting, with a patience of 10 epochs, meaning that training would stop if no improvement was observed for 10 epochs past the lowest validation loss. The model with the lowest validation loss was chosen as the final model.

Model Evaluation

We performed predictions on the WSIs of the test set by using a sliding window with an input tile size of 224 × 224 pixels and a stride of 112 × 112 pixels. The WSI classification probability was obtained by max-pooling the probabilities of all of its tiles. We computed the ROC curves and their corresponding AUCs as well as the log losses from all the models. Figure 7 and Table 1 summarize the results on the test sets. Figure 2 shows an example true positive classifications on four endoscopic biopsy fragments, whereas Figure 3 shows examples of false positive classifications and Figure 4 shows a representative example of false negative classification.

The models displayed good generalization performance on the DigestPath2019 independent test set, which consisted of WSI crops obtained from a different source than the one used for training our models. We used the training set provided by DigestPath2019 as it was publicly available. We could not perform a direct comparison with the reported results of Li *et al*[42] as the test set is not publicly available.

The WS training method achieved a statistically significantly lower log loss compared to the FS method. Figure 6 shows Grad-CAM visualization of the seven models on four positive images from the DigestPath2019. The models do no seem to pick up on the same areas. The FS+WS ×10 models picked up more SRCC cells than the WS ×10 model.

For the WS method, guiding the sampling of the positive tiles from the annotated regions improved the predictive performance as compared to without using any of the annotations (WS ×10 vs WS-noanno ×10).

Transfer learning was helpful in increasing predictive performance, given that the model trained without transfer learning (FS ×10 w/o TL) mostly achieved the lowest performance on all two test sets.

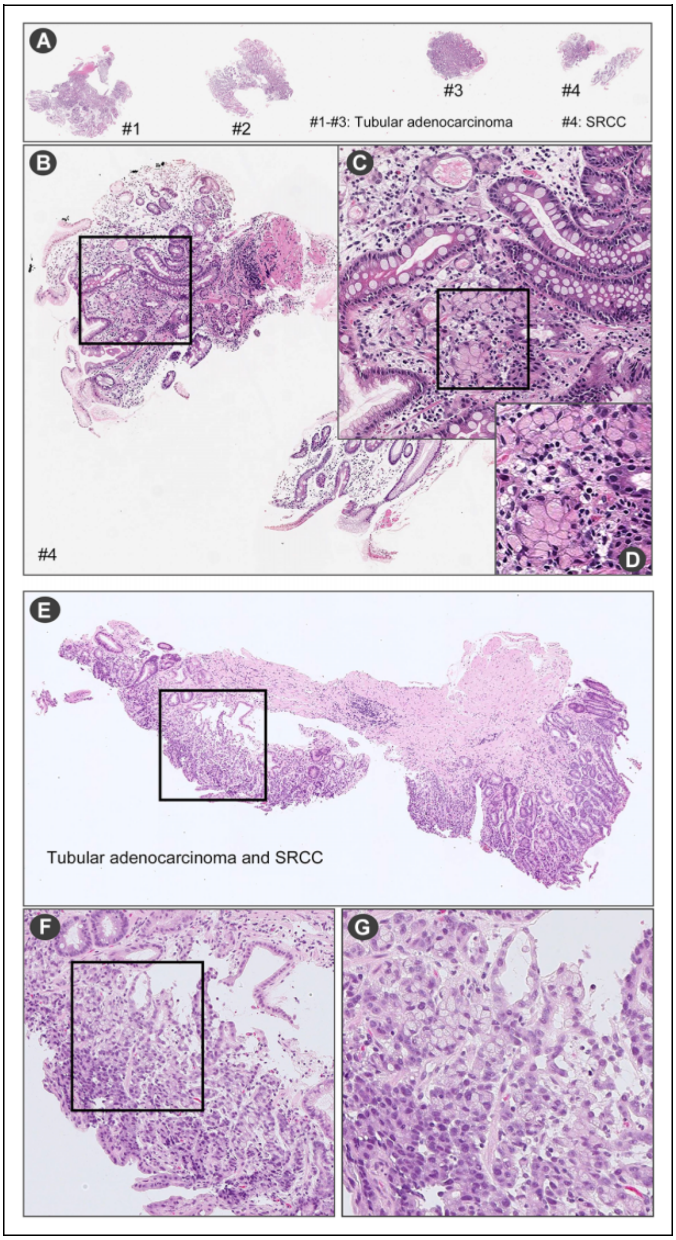
A
#1
#2
#3
#4
#1-#3: Tubular adenocarcinoma
#4: SRCC
B
#4
C
D
E
Tubular adenocarcinoma and SRCC
F
G

Figure 4. Representative false negative cases. In (A) there are four endoscopic biopsy fragments (#1-#4). According to the pathological diagnostic report, (A) #4 has SRCC. In the fragment of (B) #4, a few SRCC cells were observed (C) at high magnification (D). (E) is endoscopic biopsy fragment. According to the pathological diagnostic report, this fragment has tubular adenocarcinoma and SRCC. When viewed under high magnification (F and G), SRCC cells were observed.

The model trained at ×20 has a higher false positive rate compared to the model trained at ×10. The model trained at ×5 similarly had a higher false positive rate compared to the model trained at ×10.

An examination of some of the false positive cases showed that they were mostly due to cells exhibiting similar appearance to SRCC. In the chronic gastritis case in Figure 3, the nuclear density of the lymphocytes mimics the appearance of SRCC,

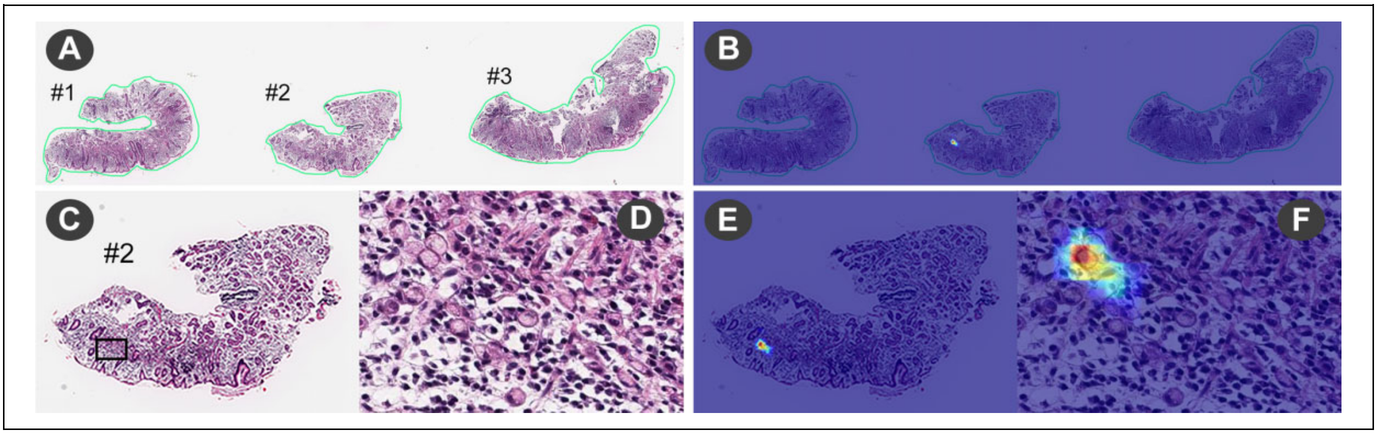

Figure 5. Representative Grad-CAM heatmap image for true-positive detection of SRCC cells. (A) Shows non-neoplastic annotations (green lines) of gastric endoscopic biopsy specimens (#1-#3) by pathologists. Tissue fragments #1 and #3 are gastritis and #2 has SRCC cells (A, C, D). Pathologists missed SRCC cells on fragment #2 (A). SRCC cells were visualized only in fragment #2 by Grad-CAM heatmap image (B). At high magnification, in fragment #2, Grad-CAM hotspots (E, F) were overlapped with infiltrating area of SRCC cells (C, D).

which most likely led to the false positive. Figure 5 shows a Grad-CAM visualization of a case used as part of the validation set where a tissue fragment was incorrectly annotated as gastritis (non-neoplastic lesion). It was initially thought to be a false-positive case; however, another inspection by expert pathologists revealed that it is a true-positive detection of SRCC. It was missed by the pathologists performing the annotations potentially due to the presence of only a small number of SRCC cells within a background of chronic inflammatory cells infiltration that have some morphological similarities to SRCC cells, making them difficult to spot. Nonetheless, the models were able to make a correct detection.

Influence of the Top k *Parameter*

Figure 8 shows the ROC curves for the two test sets. There was a noticeable trend where an increasing k value led to a decrease in the AUC and a noticeable increase in the false positive rate.

Running Time

The models overall took between 2-4 days to train on a machine with a single Nvidia Titan V GPU. The prediction time per WSI is dependent of the number of tiles that contain tissue, and it can range from 1k to 10k tiles. Prediction was at an average rate of 150 tiles per second on a single GPU.

Discussion

In this paper we have presented a deep learning application for SRCC WSI classification. The models, based on the EfficientNet-B1 architecture, achieved high ROC AUC performance on two test sets, one of which originated from a different medical institution. We analyzed the performance of different training methodologies and WSI magnifications. Results showed that a WS training method with WSIs at a magnification of ×10 achieved the highest predictive performance.

The use of WS training method achieved better performance than using the FS method alone. This is most likely due to the WS method training on tiles that have the highest probability from both the positive and negative WSIs, while the FS method trains on randomly sampled tiles. At each training iteration, the WS trains on the most confident tiles for the positive class and the most likely to be a false positive for the negative class. This prioritizes the training on reducing the false positive rate, especially given that the WSI aggregation method is max-pooling. As a single false positive tile would result in a false positive classification for the WSI.

An interesting observation from Figure 6 was that the FS+WS ×10 models picked up more SRCC cells than the WS ×10 model; this is most likely due to it having encountered more instances of SRCC tiles during the FS pre-training phase, with the WS training later serving to reduce the false positive rate.

Guiding the sampling of tiles for the WS method improved the predictive performance as observed from comparing WS ×10 vs WS-noanno ×10. This was to be expected given that there was only a small number of positive WSIs. Achieving a high predictive performance without any annotations that restrict the regions from which to sample requires a significantly larger dataset, as from the entire WSI of potentially thousands of tiles only 1 tile is selected for training. Campanella *et al*[13] observed that at least 10,000 are required to achieve a good performance.

When only WSI labels are available, using only the tile with the maximum probability ($k = 1$) led to the best performance based on the results in Figure 8. An increase in k led to an increase in the false positive rate. The only valid assumption that can be made is that at least 1 tile has to be positive when the label of the WSI is positive; however, by using $k > 1$, we are assuming that there are at least k tiles with a positive label, which is a strong assumption and more likely to be incorrect especially when there are WSIs in the training set that only have a few SRCC cells such that they are contained in a number

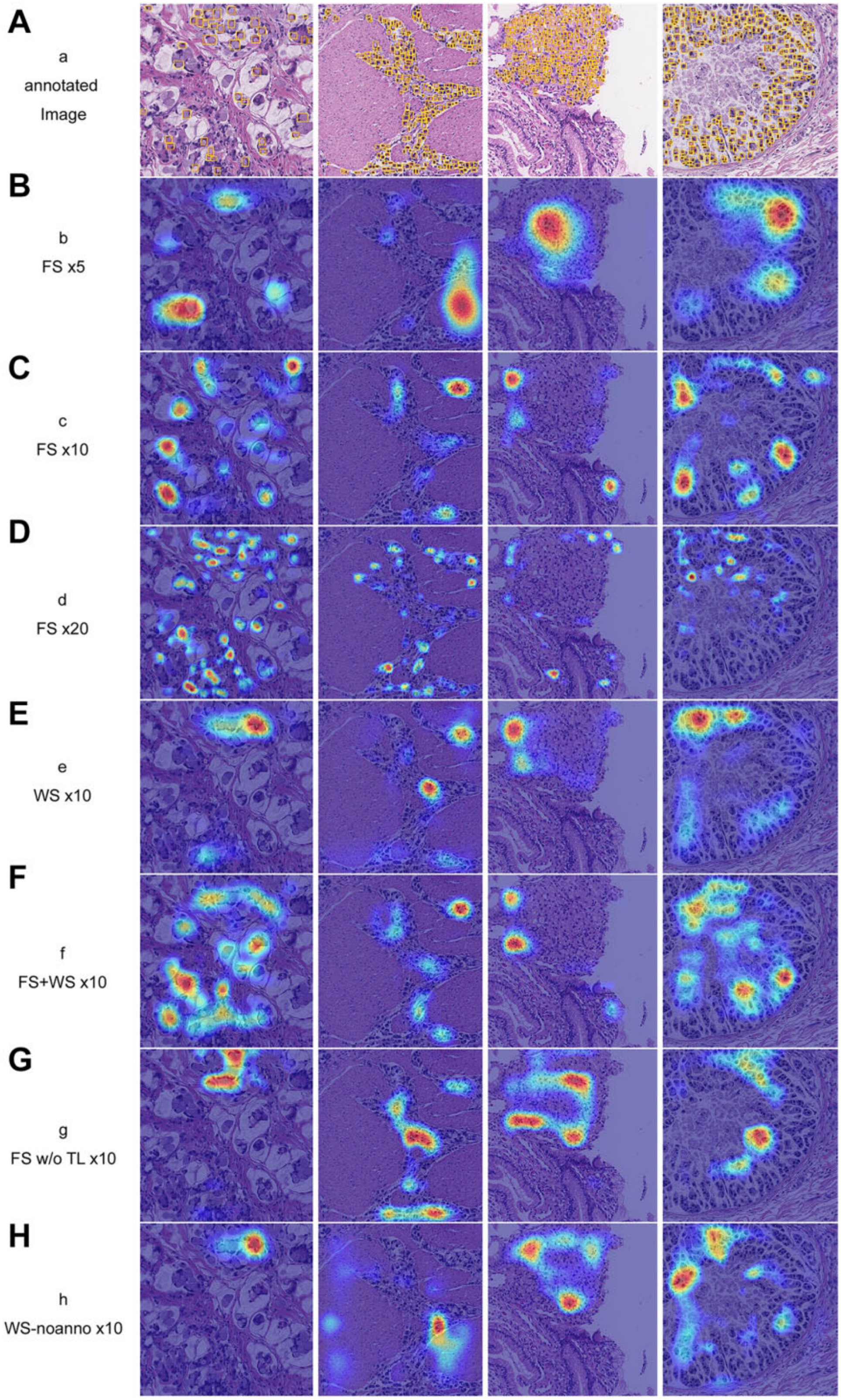

Figure 6. Grad-CAM visualization on positive images from the DigestPath2019 dataset. Row (A) shows four annotated images with yellow bounding boxes on SRCC instances. Rows (B-H) show the Grad-CAM outputs from the 7 different models.

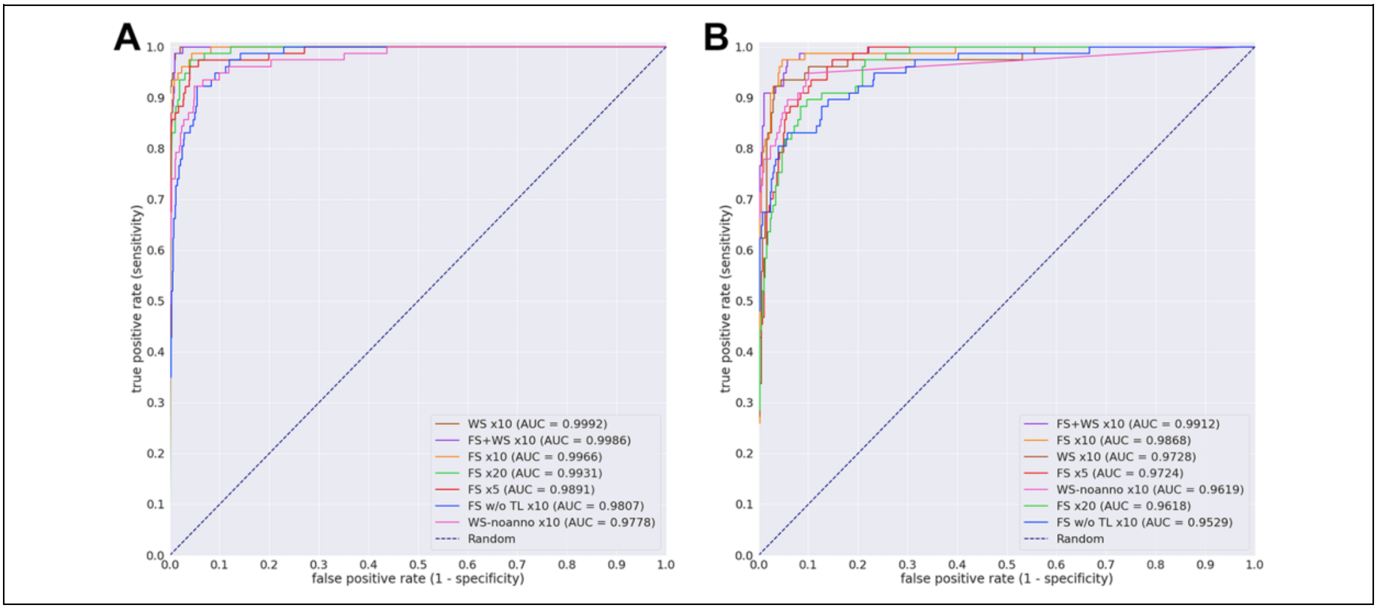

Figure 7. ROC curves from the 7 different models on the 2 test sets: (A) Hospital A and B and (B) DigestPath2019.

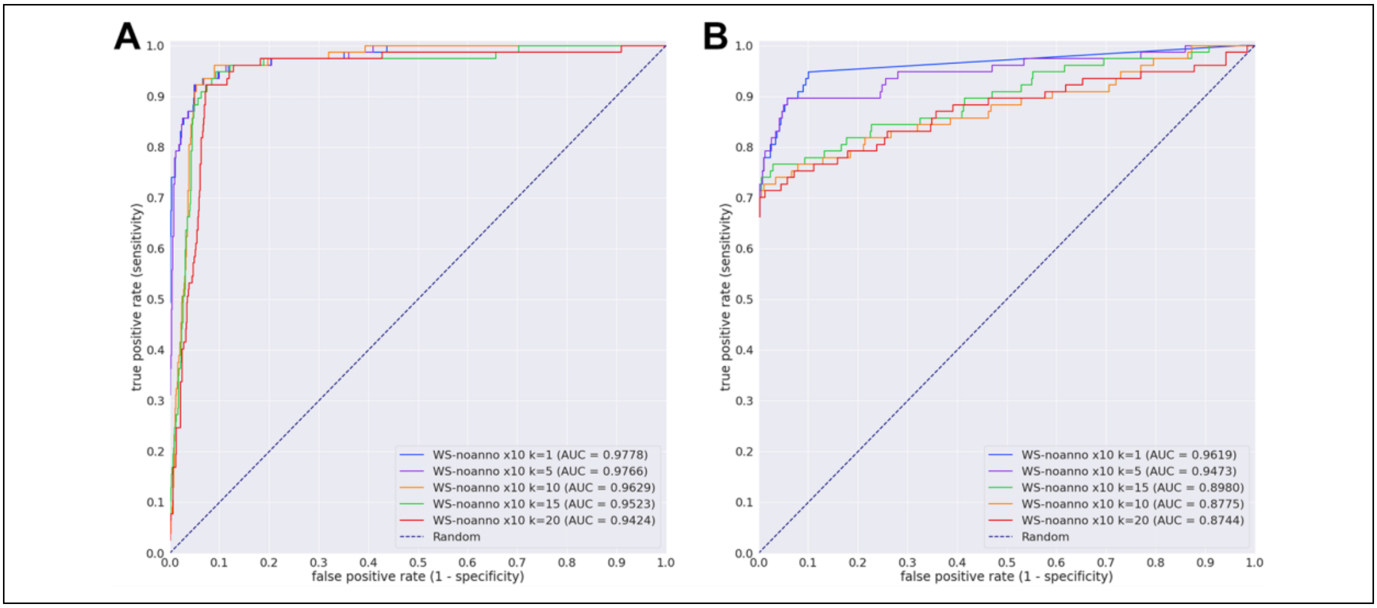

Figure 8. ROC curves from varying the top k across the range $\{1,5,10,15,20\}$ for the WS method using only slide-level labels (WS-noanno).

$< k$ of tiles. In such cases, then, it would occur that the number of tiles being sampled as positive is larger the actual number of positive tiles present. Our training set contained a few WSIs that only had a small area of infiltrating SRCC cells. This most likely led to the observed increase in the false positive rate.

The training dataset only contained a small number of positive WSIs ($n = 100$), and the use of transfer learning has helped in increasing predictive performance, given that the model trained without transfer learning (FS ×10 w/o TL) mostly achieved the lowest performance on all two test sets.

Training at ×10 seems to yield better performance than training at ×20. The model trained at ×10 had a lower false positive rate; this is most likely due to the ×10 model having more context information from the neighboring tissues. In order to confirm an SRCC diagnosis, pathologists typically view a WSI at a low magnification (e.g., ×4 or ×5) and then at a higher magnification to check the cellular morphology. It is more difficult for pathologists to distinguish between SRCC cells and mimicker cells (lymphocytes and histiocytes) if they are viewed in isolation without viewing the neighboring tissues. The lack of context information from the neighboring tissues could be the reason why the ×20 model had a slightly lower predictive performance than the ×10 model. However, going at magnification of ×5 also results in a increase in the false positive rate, and this is most likely due to the loss of cellular-level detail, making it harder to properly detect SRCC. Nonetheless, the model was still capable of predicting SRCC. This result was particularly interesting to pathologists given that they would view the WSI with a magnification of at least ×10 before confirming an SRCC diagnosis.

As a certain element of randomness is involved when training the models, some of the variations in the predictive performance between the models could be attributed to it. However, the majority of the training methods achieved an acceptable high performance, signifying that it is possible to train an SRCC WSI classifier. One potential limitation is that we do not know the extent of how well the models generalize to WSIs from different source, given that most of the WSI test sets came from the same source as the training set. Nonetheless, the good performance on the DigestPath2019 dataset, even though it only consisted of WSI crops, is highly promising. As the model do not achieve AUC of 1.0, then based on the intended application of the models, the threshold can be adjusted to obtain a desired sensitivity and specificity, so there could be a potential risk of over- or under-diagnosis based on the chosen threshold. In addition, we do not know how well the models perform on challenging cases, such as intramucosal SRCC in-situ[56] and mimicker non-neoplastic cells like xanthoma[57] cells, as neither the training or test sets contained any of these.

Conclusion

In this study, we evaluated several different training methods for the task of SRCC WSI classification, and each method has a different requirement on amount of manual annotations. Annotating WSIs can be extremely tedious because of the massive size of the WSI. We have shown that a weakly-supervised method using minimal amounts of annotations can be used to train a WSI SRCC classification model with similar performance as a fully-supervised method, meaning that detailed manual annotations are not required to obtain a model that could be used in a clinical setting. Patients with SRCC tend to have poorer prognosis than patients with other types of gastric carcinoma.[58,59] However, recent studies have shown that the incidence of SRCC has been constantly increasing.[2,4,60] Pathologists sometimes find SRCC more difficult to diagnose compared to other types of gastric carcinoma.[10] An AI model that can assist pathologists in detecting SRCC would be extremely beneficial as it can help them reduce diagnosis errors as well as potentially detect SRCC at an earlier stage and, as a result, significantly improve patient prognosis.[61]

Authors' Note

F.K. and M.T. designed the studies; F.K., M.R., O.I. and M.T. performed experiments and analyzed the data; S.I. performed pathological diagnoses and helped with pathological discussion; K.A. provided pathological cases; F.K., S.I. and M.T. wrote the manuscript; M.T. supervised the project. All authors reviewed the manuscript. The experimental protocols were approved by the Institutional Review Board (IRB) of the Hiroshima University (No. E-1316) and International University of Health and Welfare (No. 19-Im-007). All research activities complied with all relevant ethical regulations and were performed in accordance with relevant guidelines and regulations of each hospital. Informed consent to use histopathological samples and pathological diagnostic reports for research purposes had previously been obtained from all patients prior to the surgical procedures at both hospitals and an opportunity for refusal to participate in research was guaranteed by an opt-out manner.

Acknowledgments

We are grateful for the support provided by Professors Takayuki Shiomi & Ichiro Mori at Department of Pathology, Faculty of Medicine, International University of Health and Welfare; Dr. Ryosuke Matsuoka at Diagnostic Pathology Center, International University of Health and Welfare, Mita Hospital; and Dr. Naoko Aoki (pathologist). We thank the pathologists who have been engaged in the annotation work for this study.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: F.K., M.R., O.I. and M.T. are employees of Medmain Inc.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Masayuki Tsuneki https://orcid.org/0000-0003-3409-5485

Note

1. https://digestpath2019.grand-challenge.org/

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(6):394-424.
2. Henson DE, Dittus C, Younes M, Nguyen H, Albores-Saavedra J. Differential trends in the intestinal and diffuse types of gastric carcinoma in the United States, 1973-2000: increase in the signet ring cell type. *Arch Pathol Lab Med*. 2004;128(7):765-770.
3. Moran MS, Schnitt SJ, Giuliano AE, et al. Society of Surgical Oncology—American Society for Radiation Oncology consensus guideline on margins for breast-conserving surgery with whole-breast irradiation in stages I and II invasive breast cancer. *Ann Surg Oncol*. 2014;21(3):704-716.
4. Taghavi S, Jayarajan SN, Davey A, Willis AI. Prognostic significance of signet ring gastric cancer. *J Clin Oncol*. 2012;30(28):3493.
5. Yokota T, Kunii Y, Teshima S, et al. Signet ring cell carcinoma of the stomach: a clinicopathological comparison with the other histological types. *Tohoku J Exp Med*. 1998;186(2):121-130.
6. Nagtegaal ID, Odze RD, Klimstra D, et al. The 2019 who classification of tumours of the digestive system. *Histopathology*. 2020;76(2):182-188.

7. Laurén P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-typecarcinoma. *Acta Pathol Microbiol Scand*. 1965;64(1):31-49. doi:10.1111/apm.1965.64.1.31
8. Bamba M, Sugihara H, Kushima R, et al. Time-dependent expression of intestinal phenotype in signet ring cell carcinomas of the human stomach. *Virchows Archiv*. 2001;438(1):49-56.
9. Tatematsu M, Furihata C, Katsuyama T, et al. Gastric and intestinal phenotypic expressions of human signet ring cell carcinomas revealed by their biochemistry, mucin histochemistry, and ultrastructure. *Cancer Res*. 1986;46(9):4866-4872.
10. Yamashina M. A variant of early gastric carcinoma. Histologic and histochemical studies of early signet ring cell carcinomas discovered beneath preserved surface epithelium. *Cancer*. 1986;58(6):1333-1339.
11. Arnold C, Lam-Himlin D, Montgomery EA. *Atlas of Gastrointestinal Pathology: A Pattern Based Approach to Neoplastic Biopsies*. Lippincott Williams & Wilkins; 2018:176-177.
12. 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*. 2017;318(22):2199-2210.
13. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301-1309.
14. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559.
15. Gertych A, Swiderska-Chadaj Z, Ma Z, et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Sci Rep*. 2019;9(1):1483.
16. Graham S, Vu QD, Raza SEA, et al. Hover-net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal*. 2019;58:101563.
17. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; June 17-19, 1997, San Juan, PR, USA: IEEE; 2016:2424-2433.
18. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Sci Rep*. 2020;10(1):1-11.
19. Kalra S, Tizhoosh H, Choi C, et al. Yottixel—an image search engine for large archives of histopathology whole slide images. *Med Image Anal*. 2020;65:101757.
20. Kanavati F, Toyokawa G, Momosaki S, et al. Weakly-supervised learning for lung carcinoma classification using deep learning. *Sci Rep*. 2020;10(1):1-11.
21. Korbar B, Olofson AM, Miraflor AP, et al. Deep learning for classification of colorectal polyps on whole-slide images. *J Pathol Inform*. 2017;8:30.
22. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*. 2016;32(12):i52-i59.
23. Li C, Wang X, Liu W, Latecki LJ, Wang B, Huang J. Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Med Image Anal*. 2019a;53:165-178.
24. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep*. 2016;6:26286.
25. Luo X, Zang X, Yang L, et al. Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J Thorac Oncol*. 2017;12(3):501-509.
26. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Med Image Anal*. 2016;33:170-175.
27. Saltz J, Gupta R, Hou L, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep*. 2018;23(1):181-193.
28. Shi X, Su H, Xing F, Liang Y, Qu G, Yang L. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. *Med Image Anal*. 2020;60:101624.
29. Syrykh C, Abreu A, Amara N, et al. Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning. *NPJ Digit Med*. 2020;3(1):1-8.
30. Wang S, Zhu Y, Yu L, et al. RMDL: recalibrated multi-instance deep learning for whole slide gastric image classification. *Med Image Anal*. 2019;58:101549.
31. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci Rep*. 2019;9(1):3358.
32. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*. 2017.
33. Zhou ZH. A brief introduction to weakly supervised learning. *Natl Sci Rev*. 2018;5(1):44-53.
34. Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. *Artif Intell*. 1997;89(1-2):31-71.
35. Maron O, Lozano-Pérez T. A framework for multiple-instance learning. In: *Advances in Neural Information Processing Systems*; 31 July, 1998; Cambridge, MA. Massachusetts Institute of Technology Press; 1998:570-576.
36. Babenko B, Yang MH, Belongie S. Robust object tracking with online multiple instance learning. *IEEE Trans Pattern Anal Mach Intell*. 2010;33(8):1619-1632.
37. Cheplygina V, de Bruijne M, Pluim JP. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal*. 2019;54:280-296.
38. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*. 2018.
39. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. *Comput Struct Biotechnol J*. 2018;16:34-42.
40. Sudharshan P, Petitjean C, Spanhol F, Oliveira LE, Heutte L, Honeine P. Multiple instance learning for histopathological breast cancer image classification. *Expert Syst Appl*. 2019;117:103-111.

41. Xu Y, Zhu JY, Eric I, Chang C, Lai M, Tu Z. Weakly supervised histopathology cancer image segmentation and classification. *Med Image Anal*. 2014;18(3):591-604.
42. Li J, Yang S, Huang X, et al. Signet ring cell detection with a semi-supervised learning framework. In: *International Conference on Information Processing in Medical Imaging*; 2-7 June, 2019, Hong Kong, China: Springer; 2019 b:842-854.
43. Lin T, Guo Y, Yang C, Yang J, Xu Y. Decoupled gradient harmonized detector for partial annotation: application to signet ring cell detection. *arXiv preprint arXiv:2004.04455*. 2020.
44. Tan M, Le QV. Efficientnet: rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning*; PMLR; 2019. http://arxiv.org/abs/1905.11946. Cite arxiv:1905.11946Comment: Published in ICML 2019.
45. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*; June 20-25, 2009, Miami, FL, USA. IEEE; 2009:248-255.
46. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*; March 31, 2010 Sardinia, Italy: Proceedings of Machine Learning Research; 2010:249-256.
47. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*; June 20-23, 1995, Cambridge, MA, USA. IEEE; 2017:618-626.
48. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. Published 2015. Accessed November 18, 2020. Software available from tensorflow.org. https://www.tensorflow.org/
49. 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.
50. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
51. Hunter JD. Matplotlib: a 2d graphics environment. *Comput Sci Eng*. 2007;9(3):90-95. doi:10.1109/MCSE.2007.55
52. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. CRC Press; 1994.
53. Arai T. Where does signet-ring cell carcinoma come from and where does it go? *Gastric Cancer*. 2019;22(4):651-652. doi:10.1007/s10120-019-00960-w
54. Bosman FT, Carneiro F, Hruban RH, Theise ND. *WHO Classification of Tumours of the Digestive System*. 4th ed. World Health Organization; 2010.
55. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980.* 2014.
56. Tsugeno Y, Nakano K, Nakajima T, et al. Histopathologic analysis of signet-ring cell carcinoma in situ in patients with hereditary diffuse gastric cancer. *Am J Surg Pathol*. 2020;44(9):1204-1212.
57. Drude JR, Balart LA, Herrington JP, Beckman EN, Burns TW. Gastric xanthoma: histologic similarity to signet ring cell carcinoma. *J Clin Gastroenterol*. 1982;4(3):217-221.
58. Liu X, Cai H, Sheng W, et al. Clinicopathological characteristics and survival outcomes of primary signet ring cell carcinoma in the stomach: retrospective analysis of single center database. *PLoS One*. 2015;10(12):e0144420.
59. Pernot S, Voron T, Perkins G, Lagorce-Pages C, Berger A, Taieb J. Signet-ring cell carcinoma of the stomach: impact on prognosis and specific therapeutic challenge. *World J Gastroenterol*. 2015;21(40):11428.
60. Bamboat ZM, Tang LH, Vinuela E, et al. Stage-stratified prognosis of signet ring cell histology in patients undergoing curative resection for gastric adenocarcinoma. *Ann Surg Oncol*. 2014;21(5):1678-1685. doi:10.1245/s10434-013-3466-8
61. Machlowska J, Pucułekek M, Sitarz M, Terlecki P, Maciejewski R, Sitarz R. State of the art for gastric signet ring cell carcinoma: from classification, prognosis, and genomic characteristics to specified treatments. *Cancer Manage Res*. 2019;11:2151.