Weakly supervised learning for multi-organ adenocarcinoma classification in whole slide images

Masayuki Tsuneki*, Fahdi Kanavati
Medmain Research, Medmain Inc., Akasaka, Chuo-ku, Fukuoka, Japan

* tsuneki@medmain.com

Abstract

The primary screening by automated computational pathology algorithms of the presence or absence of adenocarcinoma in biopsy specimens (e.g., endoscopic biopsy, transbronchial lung biopsy, and needle biopsy) of possible primary organs (e.g., stomach, colon, lung, and breast) and radical lymph node dissection specimen is very useful and should be a powerful tool to assist surgical pathologists in routine histopathological diagnostic workflow. In this paper, we trained multi-organ deep learning models to classify adenocarcinoma in biopsy and radical lymph node dissection specimens whole slide images (WSIs). We evaluated the models on five independent test sets (stomach, colon, lung, breast, lymph nodes) to demonstrate the feasibility in multi-organ and lymph nodes specimens from different medical institutions, achieving receiver operating characteristic areas under the curves (ROC-AUCs) in the range of 0.91 -0.98.

Introduction

Adenocarcinoma is a type of carcinoma that has the propensity to differentiate into glandular, ductal, and acinar cells in several organs (e.g., stomach, colon, lung, and breast). According to the Global Cancer Statistics 2020 [1], number of new deaths and % of all sites for stomach, colon, lung, and breast cancers were as follows: 768,793 cases (7.7%) in stomach, 576,858 cases (5.8%) in colon, 1,796,144 cases (18.0%) in lung, and 684,996 cases (6.9%) in breast. Adenocarcinoma is the most common type of cancer affecting these four organs, so that adenocarcinoma classification in the primary organs especially on biopsy specimens is one of the most important histopathological inspection in clinical workflow to determine the strategies of cancer treatment. Moreover, lymph nodes are the most common site of metastatic adenocarcinoma, and can be constituted the first clinical manifestation of the cancer. The important clinical practice of the surgical pathologist is to identify the presence or absence of a malignant process in the lymph node. If cancer cells are identified within the efferent lymph vessels and extra-nodal tissues, it is necessary to note in the pathological report because of the possible prognostic significance. Histopathological evaluation of lymph node metastasis is very important for staging of tumors, documentation of tumor recurrence, and prediction of the most
corresponding author on reasonable request. The datasets that support the findings of this study are available from International University of Health and Welfare, Mita Hospital (Tokyo, Japan) and Kamachi Group Hospitals (Fukuoka, Japan), but restrictions apply to the availability of these data, which were used under a data use agreement which was made according to the Ethical Guidelines for Medical and Health Research Involving Human Subjects as set by the Japanese Ministry of Health, Labour and Welfare, and so are not publicly available. The data contains potentially sensitive information. However, the data are available from the authors upon reasonable request for private viewing and with permission from the corresponding medical institutions within the terms of the data use agreement and if compliant with the ethical and legal requirements as stipulated by the Japanese Ministry of Health, Labour and Welfare. Contact person: Professor Dr. Takayuki Shiomori, Department of Pathology, Faculty of Medicine, International University of Health and Welfare (Tokyo, Japan) phone: +81-476-20-7701 E-mail: t_shiomori@iuhw.ac.jp (2) Ethical board of Kamachi Group Hospitals (Wajiro, Shinkuki, Shinkomonji, and Shinmizumaki Hospital) Contact person: Dr. Shigeo Nakano, Head of Department of Surgical Pathology at Kamachi Group Hospitals (Fukuoka, Japan) Phone: +81-92-608-0001 E-mail: sdnakano@harajuku-reha.com

Funding: This study is based on results obtained from a project, JPNP14012, subsidized by the New Energy and Industrial Technology Development Organization (NEDO). The founder provided support in the form of salaries for authors M.T. and F.K., but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the ‘author contributions’ section.

Competing interests: M.T. and F.K. are employees of Medmain Inc. (Fukuoka, Japan). This does not alter our adherence to PLOS ONE policies on sharing data and materials.

probable primary site for a metastatic cancer of uncertain primary site. However, in the routine practical diagnosis, frequently there are numerous number of lymph nodes to be inspected in a single glass slide and there are number of radical lymph node dissection specimen glass slides in the same patient, which should be a workload burden for surgical pathologists.

The incorporation of deep learning models in routine histopathological diagnostic workflow is on the horizon and is a promising technology, allowing the potential of reducing the burden of time-consuming diagnosis and increasing the detection rate of anomalies including cancers. Deep learning has been widely applied in tissue classification and adenocarcinoma detection on whole-slide images (WSIs), cellular detection and segmentation, and the stratification of patient outcomes [2–15]. Previous works have looked into applying deep learning models for adenocarcinoma classification separately for different organ, such as stomach [15–17], colon [15, 18], lung [16, 19], and breast [20, 21] histopathological specimen WSIs. Although these existing models exhibited very high ROC-AUCs for each organ, they cannot classify adenocarcinoma across organs accurately.

In this study, we trained deep learning models using weakly-supervised learning to predict adenocarcinoma in WSIs of stomach, colon, lung, and breast biopsy specimens for primary tumors as well as radical lymph node dissection specimens for metastatic carcinoma using training datasets for stomach, colon, lung, and breast biopsy specimen WSIs without annotations. We evaluated the models on each primary organ biopsy specimen (stomach, colon, lung, and breast) as well as radical lymph node dissection specimens to evaluate presence or absence of metastatic adenocarcinoma, achieving and ROC-AUC from 0.91 to 0.98. Our results suggest that deep learning algorithms might be useful for histopathological diagnostic aids for adenocarcinoma classification in primary organs and lymph node metastatic cancer screening.

Materials and methods
Clinical cases and pathological records
In the present retrospective study, a total of 8,896 H&E (hematoxylin & eosin) stained histopathological specimen slides of human adenocarcinoma and non-adenocarcinoma (adenoma and non-neoplastic) were collected from the surgical pathology files of five hospitals: International University of Health and Welfare (IUHW), Mita Hospital (Tokyo, Japan) and Kamachi Group Hospitals (total four hospitals: Wajiro, Shinkuki, Shinkomonji, and Shinmizumaki Hospital) (Fukuoka, Japan) after histopathological review by surgical pathologists. Adenoma cases were included as adenoma is a common differential diagnosis and exhibits some similarities to adenocarcinoma. The histopathological specimens were selected randomly to reflect a real clinical settings as much as possible. Prior to the experimental procedures, each WSI diagnosis was observed by at least two pathologists with the final checking and verification performed by senior pathologists. All WSIs were scanned at a magnification of x20 using the same Leica Aperio AT2 Digital Whole Slide Scanner (Leica Biosystems, Tokyo, Japan) and were saved as SVS file format with JPEG2000 compression.

Dataset
Hospitals which provided histopathological specimen slides were anonymised by randomly assigning a letter (e.g., Hospital-A, B, C, D, and E). Table 1 breaks down the distribution of training sets from four domestic hospitals (Hospital-A, B, C, and D). Table 2 shows the distribution of 1K (1,000 WSIs), 2K (2,000 WSIs), and 4K (4,000 WSIs) training sets. Validation sets were selected randomly from the training sets and the numbers of validation sets were given in parentheses (Table 2). The distribution of test sets from five domestic hospitals (Hospital-A, B, C, D, and E) was summarized in Table 3. In both training and test sets, stomach, colon, lung,
and breast WSIs solely consisted of biopsy (stomach and colon: endoscopic biopsy, lung: transbronchial lung biopsy (TBLB), breast: needle biopsy) specimens and lymph node WSIs consisted of radical dissection specimens (Tables 1–3). The distribution of lymph nodes using test sets were summarized in Table 4. All training sets WSIs were not manually annotated and the training algorithm only used the WSI labels which were extracted from the histopathological diagnostic reports after reviewing surgical pathologists; meaning that the only information available for the training was whether the WSI contained adenocarcinoma or non-adenocarcinoma but no information available about the location of the cancerous lesions.

Deep learning models

In this study, we used the EfficientNetB1 [22] as the architecture of our models. We observed no further improvements from using larger models. We used the partial fine-tuning approach.
to train them. This method starts with an existing pre-trained models on ImageNet and fine-tunes only the affine parameters of the batch normalization layers and the final classification layer while leaving the remaining weights frozen. Fig 1 shows an overview of the training method.

As we only had WSI labels, we used a weakly supervised method to train the models. The training method is similar to the one described in [24].

WSIs typically have large areas of white background that is not required for training the model and can easily be eliminated with preprocessing via thresholding using Otsu’s method [25]. This creates a mask of the tissue regions from which it would then be possible to sample tiles in real-time using the OpenSlide library [26] by providing coordinates from the tissue regions.

For a given WSI, we obtained a single prediction on the slide-level using the following approach: we divided the WSIs into a grid with a fixed stride, and we applied the model in a sliding window fashion over the grid, resulting in a predictions for the entire tissue regions.
We then took the maximum probability from all the tiles as used that as a slide-level probability of the WSI having ADC.

During training, we initially performed a balanced random sampling of tiles from the tissue regions for first two epochs; this meant that we alternated between a positive WSI and a negative WSI and selecting an equal number of tiles from each. After the second epoch, we switched into hard mining of tiles, whereby we alternated between a positive WSI and a negative WSI; however, this time performing a sliding window inference on the entire tissue regions and then selecting the top $k$ tiles with the highest probabilities for being positive. If the WSI is negative, this effectively selects the tiles most likely to be false positives. The selected tiles were placed in a training subset, and once that subset contained $N$ tiles, a training was run whereby the model weights get updated. We used $k = 8$, $N = 256$, and a batch size of 32.

In addition, during training, we performed data augmentation of the images by performing random shifts in brightness, contrast, hue and saturation, and rotation angles as well as horizontal and vertical flipping.

We optimised the model weights by minimising the binary cross-entropy loss using the Adam optimization algorithm [27] with the following parameters: $beta_1 = 0.9$, $beta_2 = 0.999$ and a learning rate of 0.001. We applied a learning rate decay of 0.95 every 2 epochs. We used early stopping by tracking the performance of the model on a validation set; this allows stopping the training when no improvement was observed for more than 10 epochs. The model with the lowest validation loss was chosen as the final model.

**Software and statistical analysis**

The deep learning models were implemented and trained using TensorFlow [28]. AUCs were calculated in python using the scikit-learn package [29] and plotted using matplotlib [30]. The 95% CIs of the AUCs were estimated using the bootstrap method [31] with 1000 iterations.
The true positive rate (TPR) was computed as
\[
TPR = \frac{TP}{TP + FN}
\]  
(1)
and the false positive rate (FPR) was computed as
\[
FPR = \frac{FP}{FP + TN}
\]  
(2)
Where TP, FP, and TN represent true positive, false positive, and true negative, respectively. The ROC curve was computed by varying the probability threshold from 0.0 to 1.0 and computing both the TPR and FPR at the given threshold.

Compliance with ethical standards
The experimental protocol was approved by the ethical board of International University of Health and Welfare (No. 19-Im-007) and Kamachi Group Hospitals (No. 173). All research activities complied with all relevant ethical regulations and were performed in accordance with relevant guidelines and regulations in the all hospitals mentioned above.

Availability of data and material
The datasets generated during and/or analysed during the current study are not publicly available due to specific institutional requirements governing privacy protection but are available from the corresponding author on reasonable request. The datasets that support the findings of this study are available from International University of Health and Welfare, Mita Hospital (Tokyo, Japan) and Kamachi Group Hospitals (Fukuoka, Japan), but restrictions apply to the availability of these data, which were used under a data use agreement which was made according to the Ethical Guidelines for Medical and Health Research Involving Human Subjects as set by the Japanese Ministry of Health, Labour and Welfare, and so are not publicly available. The data contains potentially sensitive information. However, the data are available from the authors upon reasonable request for private viewing and with permission from the corresponding medical institutions within the terms of the data use agreement and if compliant with the ethical and legal requirements as stipulated by the Japanese Ministry of Health, Labour and Welfare.

Results
Insufficient AUC performance of WSI adenocarcinoma evaluation using existing stomach adenocarcinoma classification model
Prior to the training of multi-organ adenocarcinoma model, we have demonstrated the existing stomach adenocarcinoma classification model [15] AUC performance on test sets (Table 3). Table 5 and Fig 2A show that stomach and colon endoscopic biopsy WSIs exhibited high ROC-AUC and low log loss values but not in lung TBLB, breast needle biopsy, and radical lymph node dissection WSIs. Thus, we have trained the models using different WSI number of training sets (Table 2).

High AUC performance of WSI evaluation of adenocarcinoma histopathology images
We trained models using weakly-supervised (WS) learning which could be used with weak labels (WSI labels) [24]. We trained using the EfficientNetB1 convolutional neural network
(CNN) architecture at magnification x10. The models were applied in a sliding window fashion with input tiles of 224x224 pixels and a stride of 256 (Fig 1). To train the deep learning models, we used a total of 1,000 (1K), 2,000 (2K), and 4,000 (4K) training set WSIs (Table 2). This resulted in three different models: (1) WS-1K: 224, x10 EfficientNetB1, (2) WS-2K: 224, x10 EfficientNetB1, and (3) WS-4K: 224, x10 EfficientNetB1. We evaluated the models on test sets from domestic hospitals (Table 3). For each test set (stomach endoscopic biopsy, colon endoscopic biopsy, lung TBLB, breast needle biopsy, and radical lymph node dissection), we computed the ROC-AUC, log loss, accuracy, sensitivity, and specificity (using a probability threshold of 0.5) and summarized the results in Tables 6 and 7 and Fig 2B–2D. The models trained using 2K and 4K training sets have a higher ROC-AUCs compared to the model trained using 1K and existing stomach adenocarcinoma model (Table 6, Fig 2). However, there was no obvious difference between the model trained using 2K and 4K training sets (Table 6, Fig 2C and 2D). In test sets from domestic hospitals, the model (WS-4K: 224, x10 EfficientNetB1) achieved very high ROC-AUCs (0.912–0.978) with low values of log loss (0.203–0.437) (Table 6). In all test sets, the model (WS-4K: 224, x10 EfficientNetB1) achieved very high accuracy (0.853–0.929), sensitivity (0.796–0.911), and specificity (0.825–0.931) (Table 7). As shown in Fig 2, Tables 5–7, the model (WS-4K: 224, x10 EfficientNetB1) is fully applicable for multi-organ adenocarcinoma classification in wide variety of organs (stomach, colon, lung, breast, and lymph node) WSIs. Figs 3–7 show representative cases of true positive, true-negative, false positive, and false negative, respectively from using the model (WS-4K: 224, x10 EfficientNetB1).

### True positive adenocarcinoma prediction of stomach, colon, lung, and breast biopsy WSIs

Our model (WS-4K: 224, x10 EfficientNetB1) satisfactorily predicted adenocarcinoma in stomach endoscopic biopsy (Fig 3A–3C), colon endoscopic biopsy (Fig 3D–3G), lung TBLB (Fig 3H and 3I), and breast needle biopsy (Fig 3J and 3K) specimens. Importantly, the heatmap images showed true negative predictions of internal non-neoplastic tissue fragments (#2 in Fig 3A and 3B; #3 in Fig 3D and 3E; 3H–3K) which were confirmed by surgical pathologists.

### True positive adenocarcinoma prediction of radical lymph node dissection (lymphadenectomy) WSIs

A lymphadenectomy (radical lymph node dissection) is a surgical procedure to evaluate evidence of metastatic cancer. In routine histopathological diagnosis, the histopathological inspection of lymph nodes is one of the very important but time-consuming task to avoid the risk of medical oversight. Therefore, in clinical settings, the multi-organ adenocarcinoma model is more useful when performing histopathological diagnosis of lymphadenectomy.
specimen WSIIs. Our model (WS-4K: 224, x10 EfficientNetB1) perfectly predicted metastatic lung adenocarcinoma (Fig 4A–4D) and breast invasive ductal carcinoma (Fig 4E–4J). The heatmap images showed true negative predictions (Fig 4B) of internal non-neoplastic lymph nodes (Fig 4A). Importantly, adenocarcinoma localization areas in both metastatic lung

Fig 2. ROC curves with AUCs from four different models (A–D) on the test sets: (A) Existing stomach adenocarcinoma classification model and weakly supervised (WS) learning models based on 1K (B), 2K (C), and 4K (D) training sets with tile size 224 px and magnification at x10.

https://doi.org/10.1371/journal.pone.0275378.g002
adenocarcinoma (Fig 4C) and breast invasive ductal carcinoma (Fig 4G) are positively predicted by heatmap images (Fig 4D and 4H).

**Table 6. ROC-AUC and log loss results for adenocarcinoma classification on test sets using trained models.**

| test sets                        | WS-1K: 224, x10 EfficientNetB1 | WS-2K: 224, x10 EfficientNetB1 | WS-4K: 224, x10 EfficientNetB1 |
|----------------------------------|---------------------------------|---------------------------------|---------------------------------|
|                                  | ROC-AUC                         | log loss                        | ROC-AUC                         | log loss                        |
| Stomach endoscopic biopsy        | 0.886 [0.864–0.912]             | 0.415 [0.356–0.473]             | 0.913 [0.894–0.932]             | 0.355 [0.315–0.404]             |
| Colon endoscopic biopsy          | 0.973 [0.964–0.981]             | 0.209 [0.175–0.242]             | 0.977 [0.969–0.986]             | 0.197 [0.167–0.226]             |
| Lung TBLB                        | 0.879 [0.859–0.900]             | 0.501 [0.443–0.555]             | 0.931 [0.915–0.946]             | 0.342 [0.300–0.386]             |
| Breast needle biopsy             | 0.919 [0.900–0.935]             | 0.358 [0.315–0.404]             | 0.919 [0.901–0.936]             | 0.371 [0.325–0.423]             |
| Lymph node radical dissection    | 0.929 [0.903–0.951]             | 0.427 [0.380–0.486]             | 0.953 [0.939–0.978]             | 0.228 [0.188–0.257]             |

**Table 7. Scores of accuracy, sensitivity, and specificity on test sets using the best model (WS-4K: 224, x10 EfficientNetB1).**

| test sets                        | WS-4K: 224, x10 EfficientNetB1 |
|----------------------------------|---------------------------------|
|                                  | Accuracy | Sensitivity | Specificity |
| Stomach endoscopic biopsy        | 0.859 [0.837–0.877] | 0.813 [0.766–0.850] | 0.882 [0.856–0.905] |
| Colon endoscopic biopsy          | 0.929 [0.912–0.944] | 0.907 [0.878–0.935] | 0.943 [0.924–0.960] |
| Lung TBLB                        | 0.853 [0.831–0.875] | 0.885 [0.861–0.915] | 0.825 [0.792–0.855] |
| Breast needle biopsy             | 0.853 [0.831–0.876] | 0.796 [0.755–0.837] | 0.892 [0.868–0.917] |
| Lymph node radical dissection    | 0.928 [0.912–0.944] | 0.911 [0.866–0.955] | 0.931 [0.913–0.948] |

Our model (WS-4K: 224, x10 EfficientNetB1) showed true negative predictions of metastatic adenocarcinoma in lymph nodes without evidence of cancer metastasis (Fig 5). In Fig 5A, there were numbers of lymph nodes with broad ranging of size (small to large) and shape (round to irregular) which were not predicted as metastatic lymph nodes (Fig 5B). Moreover, in Fig 5C, the lymph node was enlarged due to lymphadenitis (Fig 5E) but without evidence of metastatic adenocarcinoma which were not predicted as metastatic lymph nodes (Fig 5D).

**True negative adenocarcinoma prediction of radical lymph node dissection (lymphadenectomy) WSIs**

Histopathologically, Fig 6A shows no evidence of metastatic adenocarcinoma. Our model (WS-4K: 224, x10 EfficientNetB1) exhibited false positive predictions of adenocarcinoma (Fig 6).
Fig 3. Representative true positive adenocarcinoma classification of stomach, colon, lung, and breast biopsy test cases using the model (WS-4K: 224, x10 EfficientNetB1). In the adenocarcinoma whole slide images (WSIs) of stomach endoscopic biopsy (A), colon endoscopic biopsy (D), lung transbronchial lung biopsy (TBLB) (H), and breast core needle biopsy (J) specimens, the heatmap images show true positive prediction of adenocarcinoma cells (B, E, F, G, I, K) which correspond respectively to H&E histopathology (A, C, D, F, G, H, J). The heatmap images show true negative predictions of non-neoplastic tissue fragments (#2 in (B) and #3 in (E)) and true positive predictions of adenocarcinoma tissue fragments (#1 in (B) and #1-#2 in (E)) which correspond respectively to H&E histopathology of adenocarcinoma area (C, F, G). The heatmap uses the jet color map where blue indicates low probability and red indicates high probability.

https://doi.org/10.1371/journal.pone.0275378.g003

6B, 6D and 6F). These tissue areas (Fig 6C and 6E) showed dense hematoxylic artifacts induced by crushing during specimen handling procedures which could be the primary cause of false positive due to its morphological similarity to irregular shaped and dense nuclei in adenocarcinoma cells.
False negative adenocarcinoma prediction of radical lymph node dissection (lymphadenectomy) WSI

In Fig 7A, histopathologically, only two metastatic colon adenocarcinoma foci were observed in the left-most lymph node (Fig 7C). After double checking two independent pathologists, there were no more metastatic adenocarcinoma cells in Fig 7A. However, the heatmap image did not predict any adenocarcinoma cells (Fig 7B).
Fig 5. Representative true negative metastatic adenocarcinoma classification of radical lymph node dissection (lymphadenectomy) test sets using the model (WS-4K: 224, x10 EfficientNetB1). Histopathologically, in (A), there were diverse size (small to large) and shape (round to irregular) of lymph nodes without evidence of metastatic adenocarcinoma. The heatmap image (B) shows true negative prediction of metastatic adenocarcinoma. Histopathologically, in (C), there were lymph nodes with lymphadenitis (E) but without evidence of metastatic adenocarcinoma (C, E). The heatmap image (D) shows true negative prediction of metastatic adenocarcinoma. The heatmap uses the jet color map where blue indicates low probability and red indicates high probability.

https://doi.org/10.1371/journal.pone.0275378.g005

Fig 6. Representative example of metastatic adenocarcinoma false positive prediction outputs on a case from the radical lymph node dissection (lymphadenectomy) test set using the model (WS-4K: 224, x10 EfficientNetB1). Histopathologically, (A) has no sign of metastatic adenocarcinoma. The heatmap image (B) exhibits false positive predictions of adenocarcinoma (D, F) where the tissue consists of dense hematoxylic artifacts induced by crushing during specimen handling procedures (C, E), which most likely is the primary cause of the false positive prediction due to its morphological similarity to adenocarcinoma cells with irregular shaped and dense nuclei. The heatmap uses the jet color map where blue indicates low probability and red indicates high probability.

https://doi.org/10.1371/journal.pone.0275378.g006
Discussion

In the present study, we trained multi-organ deep learning models for the classification of adenocarcinoma in WSIs using weakly-supervised learning. The models were trained on WSIs obtained from four medical institutions and were then applied on multi-organ test sets obtained from five medical institutions to demonstrate the generalisation of the model on unseen data. The deep learning model (WS-4K: 224, x10 EfficientNetB1) achieved ROC-AUCs in the range of 0.91-0.98.

So far, we have been investigating adenocarcinoma classification on histopathological WSIs in diverse organs (e.g., stomach [15–17], colon [15, 18], lung [19, 24], and breast [20, 21]). These models are specific to each organ, and versatile adenocarcinoma histopathological classification model(s) which can be applied in multi-organ have not been developed to date. The global adenocarcinoma classification model in multi-organ may play key roles in first-screening processes especially radical lymph node dissection specimens which consist of a large number of lymph nodes in a single WSI in routine pathological diagnosis in the clinical laboratories.

Prior to the training, we have demonstrated the versatility of the existing models. For example, the existing stomach adenocarcinoma classification model [15] exhibited scores of high ROC-AUC and low log loss for the stomach and colon endoscopic biopsy test sets, but not for the lung, breast, and lymph node test sets (Table 5). Therefore, we have trained the deep learning models from scratch by the weakly-supervised learning approach in this study.

We have collected histopathological H&E stained specimens from as many medical institutions as possible to ensure diversities of histopathological variability and specimen quality in training sets (Table 1). In the training sets, we did not include radical lymph node dissection specimens because we would like to train the model based on the primary organs and predict metastatic adenocarcinoma in lymph nodes. In all training sets (1K, 2K, and 4K), WSIs from each organ (stomach, colon, lung, and breast) were equally distributed (Table 2).

In this study, we showed that it was possible to exploit the use of a moderate size training sets of 2,000 (2K) and 4,000 (4K) WSIs to train deep learning models using a weakly-supervised learning, and we have obtained high ROC-AUC performance on primary organ (stomach, colon, lung, and breast) and radical lymph node dissection test sets, which is highly promising in terms of the generalisation performance of our models to classify adenocarcinoma in multi-organs. Using the weakly-supervised learning method allowed us to train on
our datasets and obtain high performance without manually performed annotations. This means that it is possible to train a very high performance model for any type of cancer classification in multi-organ without having to have detailed cellular level or rough annotations or requiring an extremely large number of WSI. We have demonstrated the usefulness of weakly-supervised learning approach for lung carcinoma classification [24]. Importantly, there were no significant difference in ROC-AUC and log loss results between 2K and 4K training sets, meaning that small number (total 2,000 WSIs) of training datasets were enough for adenocarcinoma classification in multi-organ.

Our model satisfactorily predicted adenocarcinoma areas not only in primary organs (stomach, colon, lung, and breast) (Fig 3) but also in radical lymph node dissection specimens (Fig 4). In routine histopathological diagnosis, inspecting cancer metastasis in lymph nodes is laborious because usually there are a lot of lymph nodes with wide variety of sizes and shapes in glass slides. Our model can localise the prediction of adenocarcinoma invasion and visualise them as heatmap images (Fig 4) which would be a great tool for primary screening or double-check purpose in clinical workflow in laboratories. Importantly, our model can evaluate adenocarcinoma-free (non-metastatic) lymph nodes (Fig 5) which reflected high specificity (0.931) (Table 7). This is an important finding to apply our model in clinical workflow. This study is not without limitations. One limitation is the use of a single scanner type for the majority of collected cases. Another limitation is the presence of false positive/negatives. The false positives seem to be primarily caused by the dense hematoxylic artifacts induced by crushing during specimen handling procedures which have morphological similarities to adenocarcinoma cell clusters with irregular shaped and dense nuclei (Fig 6). Another major limitation is that the models were not validated in independent cohorts from different institutions.

**Acknowledgments**

We are grateful for the support provided by Dr. Shin Ichihara at Department of Surgical Pathology, Sapporo Kosei General Hospital (Sapporo, Japan); Dr. Makoto Abe at Department of Pathology, Tochigi Cancer Center (Tochigi, Japan); Dr. Shigeo Nakano at Kamachi Group Hospitals (Fukuoka, Japan); Professor Takayuki Shiomi at Department of Pathology, Faculty of Medicine, International University of Health and Welfare (Tokyo, Japan); Dr. Ryosuke Matsumo at Diagnostic Pathology Center, International University of Health and Welfare, Mita Hospital (Tokyo, Japan). We thank pathologists who have been engaged in reviewing cases and clinicopathological discussion for this study.

**Author Contributions**

**Conceptualization:** Masayuki Tsuneki.

**Data curation:** Masayuki Tsuneki, Fahdi Kanavati.

**Formal analysis:** Masayuki Tsuneki, Fahdi Kanavati.

**Investigation:** Masayuki Tsuneki, Fahdi Kanavati.

**Methodology:** Masayuki Tsuneki, Fahdi Kanavati.

**Project administration:** Masayuki Tsuneki.

**Resources:** Masayuki Tsuneki.

**Software:** Masayuki Tsuneki, Fahdi Kanavati.

**Supervision:** Masayuki Tsuneki.
Validation: Masayuki Tsuneki, Fahdi Kanavati.
Visualization: Masayuki Tsuneki, Fahdi Kanavati.
Writing – original draft: Masayuki Tsuneki, Fahdi Kanavati.
Writing – review & editing: Masayuki Tsuneki.

References
1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021; 71(3):209–249.
2. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature communications. 2016; 7:12474. https://doi.org/10.1038/ncomms12474 PMID: 27527408
3. 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; 2016. p. 2424–2433.
4. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis. 2016; 33:170–175. https://doi.org/10.1016/j.media.2016.06.037 PMID: 27423409
5. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports. 2016; 6:26286. https://doi.org/10.1038/srep26286 PMID: 27212078
6. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics. 2016; 32(12):i52–i59. https://doi.org/10.1093/bioinformatics/btw252 PMID: 27307644
7. Korbar B, Olson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep learning for classification of colorectal polyps on whole-slide images. Journal of pathology informatics. 2017; 8. https://doi.org/10.4103/jpi.jpi_34_17 PMID: 28626201
8. Luo X, Zang X, Yang L, Huang J, Liang F, Rodriguez-Canales J, et al. Comprehensive computational pathological image analysis predicts lung cancer prognosis. Journal of Thoracic Oncology. 2017; 12 (3):501–509. https://doi.org/10.1016/j.jtho.2016.10.017 PMID: 27826035
9. Cougdray N, Ocampo PS, Sakellaropoulos T, Narula S, Snuderl M, Fenyo D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature medicine. 2018; 24(10):1559–1567. https://doi.org/10.1038/s41591-018-0177-5 PMID: 30224757
10. 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. Scientific reports. 2019; 9(1):1–8. https://doi.org/10.1038/s41598-019-40041-7 PMID: 30833650
11. Gertych A, Swiderska-Chadaj Z, Ma Z, Ing N, Markiewicz T, Cierwinka S, et al. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. Scientific reports. 2019; 9(1):1483. https://doi.org/10.1038/s41598-019-37689-0 PMID: 30728398
12. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama. 2017; 318(22):2199–2210. https://doi.org/10.1001/jama.2017.14585
13. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell reports. 2018; 23 (1):181–193. https://doi.org/10.1016/j.celrep.2018.03.086 PMID: 29617659
14. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VVK, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine. 2019; 25(8):1301–1308. https://doi.org/10.1038/s41591-018-0508-1 PMID: 31308507
15. Iizuka O, Kanavati F, Kato K, Rambeau M, Arhiro K, Tsuneki M. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Scientific reports. 2020; 10(1):1–11. https://doi.org/10.1038/s41598-020-58467-9 PMID: 32001792
16. Kanavati F, Tsuneki M, A deep learning model for gastric diffuse-type adenocarcinoma classification in whole slide images. arXiv preprint arXiv:210412478. 2021;
17. Kanavati F, Ichihara S, Rambeau M, Iizuka O, Arhiro K, Tsuneki M. Deep learning models for gastric signet ring cell carcinoma classification in whole slide images. Technology in Cancer Research & Treatment. 2021; 20:1533033211027901. https://doi.org/10.1177/1533033211027901 PMID: 34191660
|   | Reference                                                                                           |
|---|----------------------------------------------------------------------------------------------------|
| 18. | Tsuneki M, Kanavati F. Deep learning models for poorly differentiated colorectal adenocarcinoma classification in whole slide images using transfer learning. Diagnostics. 2021; 11(11):2074. https://doi.org/10.3390/diagnostics11112074 PMID: 34829419 |
| 19. | Kanavati F, Toyokawa G, Momosaki S, Takeoka H, Okamoto M, Yamazaki K, et al. A deep learning model for the classification of indeterminate lung carcinoma in biopsy whole slide images. Scientific Reports. 2021; 11(1):1–14. https://doi.org/10.1038/s41598-021-87644-7 PMID: 33854137 |
| 20. | Kanavati F, Tsuneki M. Breast invasive ductal carcinoma classification on whole slide images with weakly-supervised and transfer learning. bioRxiv. 2021.; https://doi.org/10.3390/cancers13215368 PMID: 34771530 |
| 21. | Kanavati F, Ichihara S, Tsuneki M. A deep learning model for breast ductal carcinoma in situ classification in whole slide images. Virchows Archiv. 2022; p. 1–14. PMID: 35076741 |
| 22. | Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR; 2019. p. 6105–6114. |
| 23. | Kanavati F, Tsuneki M. Partial transfusion: on the expressive influence of trainable batch norm parameters for transfer learning. arXiv preprint arXiv:210205543. 2021.; |
| 24. | Kanavati F, Toyokawa G, Momosaki S, Rambeau M, Kozuma Y, Shoji F, et al. Weakly-supervised learning for lung carcinoma classification using deep learning. Scientific reports. 2020; 10(1):1–11. https://doi.org/10.1038/s41598-020-66333-x PMID: 32518413 |
| 25. | Otsu N. A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics. 1979; 9(1):62–66. https://doi.org/10.1109/TSMC.1979.4310076 |
| 26. | Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayan M. OpenSlide: A vendor-neutral software foundation for digital pathology. Journal of pathology informatics. 2013; 4. https://doi.org/10.4103/2153-3539.119005 PMID: 24244884 |
| 27. | Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.; |
| 28. | Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015. Available from: https://www.tensorflow.org/; |
| 29. | Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830. |
| 30. | Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007; 9(3): 90–95. https://doi.org/10.1109/MCSE.2007.55 |
| 31. | Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994. |