Title

Tailoring pretext tasks to improve self-supervised learning in histopathologic subtype classification of lung adenocarcinomas.

Authors

Ding, Ruiwen
Yadav, Anil
Rodriguez, Erika
et al.

Publication Date

2023-11-01

DOI

10.1016/j.compbiomed.2023.107484

Peer reviewed

HHS Public Access

Author manuscript

Comput Biol Med. Author manuscript; available in PMC 2024 June 04.

Published in final edited form as:

Comput Biol Med. 2023 November ; 166: 107484. doi:10.1016/j.compbiomed.2023.107484.

Tailoring Pretext Tasks to Improve Self-Supervised Learning in Histopathologic Subtype Classification of Lung Adenocarcinomas

Ruiwen Ding[1,*], **Anil Yadav**[1], **Erika Rodriguez**[2], **Ana Cristina Araujo Lemos da Silva**[3], **William Hsu**[1]

[1]Medical & Imaging Informatics, Department of Radiological Sciences, David Geffen School of Medicine at University of California, Los Angeles (UCLA), Los Angeles, CA, USA

[2]Department of Pathology & Laboratory Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

[3]Federal University of Uberlândia, MG, Brazil

Abstract

Lung adenocarcinoma (LUAD) is a morphologically heterogeneous disease with five predominant histologic subtypes. Fully supervised convolutional neural networks can improve the accuracy and reduce the subjectivity of LUAD histologic subtyping using hematoxylin and eosin (H&E)-stained whole slide images (WSIs). However, developing supervised models with good prediction accuracy usually requires extensive manual data annotation, which is time-consuming and labor-intensive. This work proposes three self-supervised learning (SSL) pretext tasks to reduce labeling effort. These tasks not only leverage the multi-resolution nature of the H&E WSIs but also explicitly consider the relevance to the downstream task of classifying the LUAD histologic subtypes. Two tasks involve predicting the spatial relationship between tiles cropped from lower and higher magnification WSIs. We hypothesize that these tasks induce the model to learn to distinguish different tissue structures presented in the images, thus benefiting the downstream classification. The third task involves predicting the eosin stain from the hematoxylin stain, inducing the model to learn cytoplasmic features relevant to LUAD subtypes. The effectiveness of the three proposed SSL tasks and their ensemble was demonstrated by comparison with other state-of-the-art pretraining and SSL methods using three publicly available datasets. Our work can be extended to any other cancer type where tissue architectural information is important. The model could be used to expedite and complement the process of routine pathology diagnosis tasks. The code is available at https://github.com/rina-ding/ssl_luad_classification.

Keywords

self-supervised learning; pretext task; histopathology; lung adenocarcinoma; histologic subtype classification

[*]Corresponding author. 924 Westwood Blvd Suite 420, Los Angeles, CA, 90024, USA.

Declaration of Competing Interest

None

1. Introduction

Lung cancer is the leading cause of cancer death in the United States (U.S.), and around 85% of all lung malignancies are non-small cell lung cancers (NSCLC) [1]. Lung adenocarcinoma (LUAD) is the most common subtype of NSCLC and is morphologically heterogeneous. The most recent World Health Organization cancer guidelines classify invasive nonmucinous LUAD into five subtypes, including lepidic, acinar, papillary, micropapillary, and solid. Moreover, LUAD frequently exhibits a heterogeneous mixture of multiple histologic subtypes in the same tumor [2]. The heterogeneous histologic subtypes are associated with different prognostic impacts on patient outcome, with lepidic having a good prognosis, acinar and papillary having an intermediate prognosis, and micropapillary and solid having a poor prognosis [3]. In resected tumors, each histologic subtype is usually manually quantified by pathologists, and then the most predominant histologic subtype is derived based on the percentage of each subtype [2]. However, this grading system might not be sufficient for capturing the aggressiveness of the disease. For example, a small amount of micropapillary pattern can be associated with poor prognosis even though the predominant subtype may not be micropapillary [4]. In addition, the subjective nature of predominant histologic subtype grading is associated with poor to intermediate inter-reader agreement. A study reported pulmonary pathologists having good kappa scores of 0.70 to 0.84 for ‘typical’ cases representing the five histological subtypes and poor kappa scores of 0.24 to 0.52 for more ‘difficult’ ones [5].

Fully supervised convolutional neural networks (CNNs) can improve the accuracy and reduce the subjectivity of LUAD histologic subtyping using Hematoxylin and Eosin (H&E)-stained whole slide images (WSIs) [6], [7]. However, training these models requires labeled data, which is time-consuming and labor-intensive. One approach is to pretrain the model using a large unrelated dataset such as ImageNet [8] and then transfer the learned knowledge to a new model re-trained using the specific and smaller dataset. Although many tasks can benefit from transfer learning using ImageNet compared to training from scratch [9], medical tasks such as histopathology classification may not benefit as much since the natural images from ImageNet may not be as relevant to the medical images. Recently, self-supervised learning (SSL) approaches [10] have been used to first train the model on a pretraining task (pretext task) from the medical images that are not labeled by experts and then transfer the learned knowledge to the downstream task where expert labels are provided. Koohbanani et al. showed that pretext tasks that leverage the multi-resolution nature of pathology images were more useful for the downstream task of pathology image classification than generic pretext tasks such as predicting image rotation or flipping [11]. Another category of SSL is contrastive learning, where the model is trained to maximize the similarity between comparable samples [12].

In this work, we propose three SSL tasks designed to focus the model on leveraging the multi-resolution nature of the H&E WSIs and to consider explicitly the relevance to the downstream task of classifying the LUAD histologic subtypes. WSIs can be acquired at different magnification levels, capturing local cellular features at higher magnification and global spatial morphology at lower magnification. As shown in Figure 1a, each LUAD

histologic subtype has a different gland architecture, tissue morphology, and cytoplasm features, which should be reflected in the learned representations from the pretext tasks. To this end, we devised an SSL task that predicts whether an image tile cropped at a higher magnification level from the WSI is contained in another image cropped at a lower magnification level and the order of concatenation of the low-high magnification tiles ($P_{Contained}$, Figure 1b). We devised a second SSL task where the model predicts the grid position of the higher magnification tile from the lower magnification tile (P_{Grid}) (Figure 1c). The hypothesis behind both tasks is that they induce the model to learn to match different tissue structures presented in the WSI, mimicking the process of pathologists distinguishing subtypes by identifying unique tissue structures, and thus benefit the downstream classification where those structures are also present. We also proposed a third SSL task where an E-stained image is predicted from an input hematoxylin-stained image (P_{Stain}, Figure 1d). The rationale is that since each LUAD subtype is characterized by different patterns such as cytoplasmic features and they are stained pink by the E stain, P_{Stain} can induce the model to learn these features and benefit the downstream classification task. Once the SSL model was trained for each SSL task, the learned weights were transferred to its downstream model for further finetuning on expert-labeled data. In the end, to leverage the advantage of all three proposed SSL tasks, the final prediction of the LUAD subtype was derived using weighted average prediction from each downstream model pretrained with the corresponding proposed SSL task. Our main contributions are as follows:

1. Existing studies do not explicitly consider the relevance of an SSL task to the downstream task in various domains. This work is among the first to demonstrate the benefit of devising SSL tasks that are closely tailored to the downstream tasks.
2. Specifically, we devised three novel SSL tasks relevant to the downstream task of LUAD histologic subtype classification. These tasks force the model to understand tissue structures and identify features across magnifications.
3. Using three different publicly available datasets National Lung Screening Trial (NLST) [13], The Cancer Genome Atlas (TCGA) [14], and Clinical Proteomic Tumor Analysis Consortium (CPTAC) [15], we showed the effectiveness of the three SSL tasks and the ensemble of the three by comparing them with other pretraining methods including ImageNet-pretrained weights, state-of-the-art pathology-specific SSL tasks, and state-of-the-art contrastive learning methods.

2. Methods

2.1 Overview

The three SSL tasks: $P_{Contained}$, P_{Grid}, and P_{Stain} are summarized in Figure 1b, c, d. For each SSL task, a CNN-based model was used to learn the task on self-labeled (non-expert labeled) data. Specifically, self-labeled means the labels are derived from existing information (e.g., image metadata) that did not require expert-derived labels. The learned weights were then transferred to the downstream LUAD histologic subtype classification task models for finetuning, where expert annotation was used. Once all three downstream models were

trained, a final ensemble prediction was derived using the weighted average predictions from the three trained individual models (Figure 1e). Each SSL task and the downstream task are elaborated on in the following sections. Pseudocode of $P_{Contained}$, P_{Grid}, and P_{Stain} are in Figure 2.

2.2 Proposed SSL Task 1: $P_{Contained}$

In the first proposed SSL task $P_{Contained}$, a learning model takes a pair of channel-wise concatenated lower and higher magnification image tiles cropped from a WSI as input. The order of the concatenation is random. A ResNet18 [16] model is trained to predict whether the higher magnification tile is contained in the lower magnification tile and the order of the two images in the concatenation. Specifically, the model predicts a total of four classes: “contained and higher magnification tile comes first in the concatenation”, “not contained and higher magnification tile comes first in the concatenation”, “contained and lower magnification tile comes first in the concatenation”, and “not contained and lower magnification tile comes first in the concatenation” (Figure 1b). A lower magnification image is first generated from a higher magnification image using down-sampling factor d. Non-overlapping higher magnification tiles are then cropped from the lower magnification tile. Each lower magnification tile generates a total of d^2 non-overlapping tiles at the higher magnification level. All tiles, regardless of magnification level, have the same dimensions. The mathematical relationship between a lower and a higher magnification tile given d is:

$$(x_H, y_H, w_H, h_H) = (x_L \times d, y_L \times d, w_L \times d, h_L \times d)$$

where x_L, y_L are the lower magnification tile L’s upper left corner coordinates, w_L and h_L are L’s width and height, and x_H, y_H are the higher magnification tile H’s upper left corner coordinates, w_H and h_H are H’s width and height. Figure 1b shows an example of the higher magnification tile being contained in the lower magnification tile (“contained”) and the higher magnification tile being ordered before the lower magnification tile. For the “not contained” image pair, the higher magnification tile was randomly sampled from another lower magnification tile from the same patient.

2.3 Proposed SSL Task 2: P_{Grid}

The second proposed SSL task P_{Grid} is a variant of $P_{Contained}$. A ResNet-18 model takes in a pair of channel-wise-concatenated lower and higher magnification image tiles and predicts the grid class of the higher magnification tile relative to the lower magnification tile. The mathematical relationship between a lower and a higher magnification is the same as in $P_{Contained}$. The higher magnification tile always comes first, and the lower magnification tile always comes second in concatenation. For a WSI with down-sampling factor d, since there are d^2 non-overlapping higher magnification tiles from the lower magnification tile of the same size, each of the d^2 higher magnification tiles resides in one of the d^2 spatially ordered grid. This pretext task asks the model to classify which of the grid the higher magnification tile resides in its lower magnification one (Figure 1c).

2.4 Proposed SSL Task 3: P_{Stain}

In H&E images, cell nuclei are dyed blue or purple by hematoxylin (H), while other contents, such as cytoplasm, are dyed pink by eosin (E). The third proposed pretext task P_{Stain}, depicted in Figure 1d, takes in an H-stained image tile and predicts the E stain of that same image tile using a U-Net model with ResNet18 as backbone [17]. Since each LUAD histologic subtype is characterized by patterns such as cytoplasmic features [3], P_{Stain} might induce the model to learn these features and benefit the downstream classification task. H and E stains were separated using sparse non-negative matrix factorization (SNMF) [18]. Each single dye staining is restored by solving a matrix decomposition problem, where the staining is decomposed into the dye spectra and the contribution of the dye to the image pixel intensity.

2.5 Preventing the Model from Learning Shortcuts during SSL

When designing any SSL task, we must ensure that the model learns desired information without trivial shortcuts [19]. For the proposed SSL tasks $P_{Contained}$ and P_{Grid} that both involve learning the spatial relationships between lower and higher magnification tiles, it is possible that low-level cues such as boundary patterns of each low and high magnification tile can serve as a model learning shortcut. For $P_{Contained}$, on the low and high magnification tiles, respectively, random data transformation was applied, consisting of random horizontal and vertical flips, random rotation, and random brightness jittering, during batch generation in training reduces the risk of shortcut learning. For P_{Grid}, random rotation was used.

2.6 Downstream Task: Subtype Classification

The downstream classification tasks pretrained using $P_{Contained}$ and P_{Grid}, denoted $D_{Contained}$ and D_{Grid} respectively, take in a channel-wise concatenated image formed by two identical image tiles of the same size as in $P_{Contained}$ and P_{Grid}, since the input dimensions of the pretrained and downstream models need to be consistent (both with dimensions $224 \times 224 \times 6$). D_{Stain} takes in a single image tile as input. All three downstream models use ResNet18 [16] architecture. Once all three downstream models are trained, they are ensembled by multiplying each of the three trained individual models' ($D_{Contained}$, D_{Grid}, and D_{Stain}) prediction probabilities by a fixed weight found by grid search on the validation sets with average F1 score as the metric (Figure 1e). Formally, the prediction probability $p_{Ensemble}$ of the ensemble model can be expressed as:

$$p_{Ensemble} = p_{Contained} \times w_{Contained} + p_{Grid} \times w_{Grid} + p_{Stain} \times w_{Stain}$$

where $p_{Contained}$, p_{Grid}, and p_{Stain} are the prediction probabilities of proposed models $D_{Contained}$, D_{Grid}, and D_{Stain} and $w_{Contained}$, w_{Grid}, and w_{Stain} are their corresponding weights found by grid search, and $w_{Contained} + w_{Grid} + w_{Stain} = 1$.

3. Experiments

3.1 Data

Three publicly available datasets from the NLST (146 patients, 407 WSIs), TCGA (325 patients, 355 WSIs), and CPTAC (139 patients, 667 WSIs) were used. For all three datasets, the inclusion criteria include the patients being stage I or II LUAD and having at least one H&E WSI. Cases with substantial slide artifacts, such as pen marks on the WSIs, were excluded. Table 1 provides more details about each dataset, and Figure 3 illustrates the participant selection process.

3.1.1 Annotation—Two pathologists (E.R. and A.S.) were involved in the tile-level annotation of six tissue classes (five LUAD subtypes plus non-tumor). Specifically, each WSI was presented to the pathologists as a uniform grid of tiles cropped at either 10x or 20x magnification depending on the WSI's objective magnification (Supplementary Table 1), and they randomly chose tiles from the grid to annotate. During annotation, the pathologists had access to the original multi-resolution WSI where they could zoom in or out of any region. Due to tissue heterogeneity and the uniform grid tiling of the WSI, multiple subtypes may be present in the same tile. In this scenario, the most predominant class was used as the label. If one pathologist (A) was unsure of the label, the tile was marked as needing consensus from the other pathologist (B). That tile was then reviewed and labeled independently by pathologist B. A total of 30 tiles were marked as needing consensus across three datasets, and they were not included in modeling. An inter-reader agreement experiment was conducted to calculate Cohen's kappa score [20], [21] between the pathologists' labels on those 30 tiles. Table 3 summarizes the number of annotated tiles that were used in modeling.

3.1.2 Data Used for SSL Experiments—For model training and evaluation of the three proposed SSL tasks, both NLST and TCGA were used. In both $P_{Contained}$ and P_{Grid}, non-overlapping lower magnification tiles were cropped at 2.5x with size 512 × 512, from which 16 non-overlapping higher magnification tiles were cropped at 10x with size 512 × 512 for patients with 40x objective magnification WSIs; non-overlapping lower magnification tiles were cropped at 5x with size 1024 × 1024, from which 16 non-overlapping higher magnification tiles were cropped at 20x with size 1024 × 1024 for patients with 20x objective magnification WSIs. In P_{Stain}, each tile (H&E-stained) was cropped at either 10x magnification with size 512 × 512 or 20x magnification with size 1024 × 1024 depending on the WSI's objective magnification, from which the corresponding H-stained tiles and E-stained tiles were derived. In $P_{Contained}$ and P_{Grid}, a pair of tiles is defined as one sample; in P_{Stain}, a single tile is one sample. In all three SSL tasks, 310,201 (80%) samples were used for training, and 77,520 (20%) were used for validation. All tiles were resized to be 224 × 224.

3.1.3 Data Used for Downstream Subtype Classification Experiments—For model training and evaluation of the downstream task of LUAD histologic subtype classification, NLST, TCGA, and CPTAC data were used. NLST and TCGA tiles were annotated at either 10x magnification with size 512 × 512 or 20x magnification with size

1024 × 1024, and all CPTAC tiles were annotated at 20x magnification with size 1024 × 1024. The total pixel area a tile covers is the same across three datasets (Supplementary Table 1). All tiles were resized to be 224 × 224. Many more non-tumor tiles were annotated compared to other classes in the NLST and TCGA datasets. To ensure class balance during model learning, non-tumor tiles were under sampled in training and validation sets such that the final number of non-tumor tiles was equal to the number of lepidic tiles. The annotated tiles across three datasets were mixed for model training (60%, 988 tiles), validation (20%, 336 tiles), and testing (20%, 373 tiles) (Table 2).

3.2 Comparison Approaches

Our proposed SSL tasks were compared with other state-of-the-art pretraining methods P in terms of their benefit for the downstream classification. $P_{FromScratch}$ is when no pretraining was used. $P_{ImageNet}$ represents when ImageNet pre-trained weights were used as the pretraining [8]. $P_{MagLevel}$ is an SSL task that predicts the magnification level of a tile (4 classes, 40x, 10x, 2.5x, and 1.25x) [11]. P_{JigMag} is another SSL task where the model takes in a sequence of image tiles cropped at different magnification levels with various orders and predicts the arrangement of those tiles. The tiles have a contextual relationship with each other. Starting from the current higher magnification level, the next lower magnification tile was cropped and downsampled such that its center is also the center of the current higher magnification tile [11]. There are 24 different arrangements of 4 tiles with different magnification levels. Therefore P_{JigMag} is a 24-class prediction task. The original paper did not provide details on how the tiles were inputted into the model, but we did channel-wise concatenation of the 4 tiles. Both $P_{MagLevel}$ and P_{JigMag} used cross-entropy loss. P_{BYOL}, $P_{SimSiam}$ are contrastive learning methods BYOL [22] and SimSiam [23] that aim to learn representations which are invariant to transformations on images such as random cropping, rotations, and flipping. The models learn losses that minimize the negative cosine similarity between learned feature representations for different transformations. The difference between the two models is that BYOL uses a momentum encoder network to prevent the model from collapsing. SimSiam showed that the use of momentum encoder is unnecessary and the stop-gradient operation on one branch of the Siamese network is critical. The sample size for training and validating $P_{MagLevel}$, P_{JigMag}, P_{BYOL}, and $P_{SimSiam}$ was the same as in proposed SSL tasks.

3.3 Model Training and Evaluation

3.3.1 SSL Models—In all three proposed SSL tasks $P_{Contained}$, P_{Grid}, and P_{Stain} and two of the baseline SSL tasks $P_{MagLevel}$ and P_{JigMag}, the models were trained using an Adam optimizer with batch size 32, a learning rate and weight decay of 0.0001. The models of $P_{Contained}$, P_{Grid}, $P_{MagLevel}$, and P_{JigMag} used cross entropy loss and the model of P_{Stain} used Absolute Error Loss (L1) loss computed between the predicted and original images.

For P_{BYOL} and $P_{SimSiam}$, the batch size was 128, and the learning rate was 0.005. Stochastic gradient descent (SGD) optimizer with a momentum of 0.9 was used.

All models had a dropout layer (p = 0.2) before the linear prediction layer. All models were trained on 80% of samples/tiles, validated on the remaining 20% of samples/tiles (Table 2),

and had early stopping monitored by validation loss with a patience of 10 epochs and a maximum of 200 epochs.

3.3.2 Downstream Subtype Classification Models—1,697 annotated tiles were used to train and evaluate models for the downstream tasks D. Data transformation, including random horizontal and vertical flips and random rotations, was applied during the generation of each training batch to increase the diversity of training data. Stratified five-fold cross-validation was used. Within each fold, tiles were split into 60% training, 20% validation, and 20% testing. Performance was measured using the F1 score on test sets with paired t-test as a statistical significance test to compare model performance (where $p<0.05$ was considered significant). The F1 score was computed as $2\times\frac{Precision\times Recall}{Precision+Recall}=\frac{TP}{TP+\frac{1}{2}\left(FP+FN\right)}$ where TP is true positive, FP is false positive, and FN is false negative. Benjamini and Hochberg method was used to correct for multiple comparisons [24]. The models had a dropout layer ($p = 0.2$) before the linear prediction layer and had early stopping monitored by validation loss with a patience of 10 epochs. The hyperparameters were the same as in $P_{Contained}$ except that the batch size was 16.

4. Results

4.1 Inter-reader Agreement on Tile-level Annotation

Cohen's kappa score was 0.72 (0.53 – 0.92) on the 30 tiles marked as needing consensus from the other pathologist. This score indicates good agreement between the two readers on the tiles that they were relatively unsure of. There was disagreement on a total of 6 out of 30 tiles. The most common disagreements were between acinar and micropapillary tiles and nontumor and lepidic tiles.

4.1 Avoiding Shortcuts in SSL Tasks

Table 4 shows that the downstream classification results of $D_{Contained}$ and D_{Grid} are better when applying shortcut-avoiding techniques, forcing the model to learn useful cues from images.

4.2 Downstream Classification Results Under Different Annotation Budgets

4.2.1 100% annotation budget.—Table 5 summarizes the average F1 score and standard deviation for the downstream task of classifying LUAD subtypes using different pretraining methods when given the entire training and validation set. $P_{Contained}$ achieved the highest average F1 score for lepidic, papillary, and micropapillary. P_{Stain} achieved the highest average F1 score for acinar, solid, and non-tumor. While P_{Grid} did not have the highest F1 score on any class, it achieved similar performance as $P_{Contained}$ on lepidic, had a higher F1 score for the micropapillary class as compared to P_{Stain}, and had a higher F1 score for non-tumor as compared to $P_{Contained}$. The average F1 score was not significantly different between $D_{Contained}$, D_{Grid}, and D_{Stain}. The model of $D_{Ensembled}$ achieved better performance on all classes except for papillary, as compared to the individual model of $D_{Contained}$, D_{Grid}, and D_{Stain}. Even though the model of $D_{Contained}$ had higher average F1 score on papillary, its standard deviation was larger than the one of $D_{Ensemble.}$

$P_{Contained}$, P_{Grid}, and P_{Stain} substantially improved the downstream task $D_{Contained}$, D_{Grid}, and D_{Stain} respectively, compared to $D_{FromScratch}$. The models of $D_{ImageNet}$, $D_{MagLevel}$, D_{JigMag}, D_{BYOL}, and $D_{SimSiam}$ had improved results upon $D_{FromScratch}$ but did not outperform $D_{Contained}$, D_{Grid}, D_{Stain}, or $D_{Ensemble}$, demonstrating the informative value of our pretext learning tasks. Further, as shown in Figure 4a, the model from $D_{Ensemble}$ had statistically significantly higher average F1 score as compared to all six baseline models in all classes except for lepidic. The model from $D_{Ensemble}$ showed slightly improved performance on lepidic when compared with the models of $D_{ImageNet}$, but the difference was not statistically significant.

4.2.1 50% annotation budget.—

Table 6 summarizes the downstream results when the annotation budget is only half of the training and validation cases. $P_{Contained}$ achieved the highest average F1 score for lepidic, acinar, and papillary; P_{Stain} achieved the highest average F1 score for micropapillary, solid, and non-tumor. The model of $D_{Ensembled}$ achieved better performance on lepidic, acinar, papillary, and micropapillary, as compared to the individual model of $D_{Contained}$, D_{Grid}, and D_{Stain}.

The model from $D_{Ensemble}$ had a significantly better average F1 score when compared with all six other baseline models for acinar, papillary, and solid.

The percent improvement of average F1 scores across all six classes from $D_{FromScratch}$ to $D_{Ensemble}$ was 0.309, which was much higher than the improvement when 100% annotation budget was used (0.228).

4.3 Error Analysis

According to Figure 5a, the most likely classes to be misclassified are non-tumor and lepidic, acinar and papillary, and micropapillary and papillary when the annotation budget is 100%. When the annotation budget is 50%, as illustrated in Figure 5b, the most likely classes to be misclassified are between acinar and papillary, acinar and micropapillary, and micropapillary and papillary. These trends are consistent with prior literature stating that papillary and micropapillary subtypes and papillary and acinar subtypes are particularly challenging to distinguish [25].

4.4 Visualizing Downstream Classification Results

As shown in Figure 1e, each WSI can be fed into the ensemble model during inference to generate predictions for all tiles, overlaying the results on the original WSI. Figure 6 shows some example predictions for each dataset. The pathologists independently derived the predominant subtype label. While some WSIs show a more heterogeneous prediction map, in general, the model predictions are consistent with the pathologists-derived predominant subtype label.

In addition, correctly classified example tiles from the test sets of three proposed models $D_{Contained}$, D_{Grid}, and D_{Stain} were interpreted and visualized using Gradient-weighted Class Activation Mapping (Grad-CAM) [26] as shown in Figure 7. In general, for subtypes (acinar, papillary, and micropapillary) that have prominent architectural features, such as acini and papillae, the models focused on these features. However, the model attention is

less consistent in lepidic and non-tumor since no prominent structures exist. As for the solid tile, D_{Grid} had negative attention on the stromal region, which indicates that the model could recognize the region as not being a solid subtype.

5. Discussion

Early-stage LUAD is the most common subtype of NSCLC, which exhibits heterogeneous biological behaviors and aggressiveness within the same tumor. Each patient has one of five predominant histologic subtypes (lepidic, acinar, papillary, micropapillary, and solid). Each has a different prognostic impact on recurrence and predictive value for response to adjuvant therapy [3], [27], [28]. Fully supervised CNNs can improve the accuracy and reduce the subjectivity of LUAD histologic subtype classification [6], [7]. However, they rely on a large amount of expert annotation. In this work, we used three SSL tasks to train the models on a non-expert annotated large dataset and learn representations from SSL tasks that are closely relevant to the downstream task of LUAD histologic subtype classification.

All three proposed SSL tasks outperformed all baseline pretraining methods in terms of their benefit to the downstream classification of the LUAD histologic subtype. In most classes, the proposed ensemble model achieved better results than each model individually, indicating the value of combining downstream models pre-trained using different SSL tasks. Each baseline pretraining method including $P_{ImageNet}$, $P_{MagLevel}$, P_{JigMag}, P_{BYOL}, and $P_{SimSiam}$ improved the downstream task results as compared to when not having any pretraining, but they did not outperform our proposed SSL tasks. For $P_{ImageNet}$, even though the dataset was a large, labeled dataset ImageNet, the inherent difference between natural images from ImageNet and the pathology images might reduce the benefit of this pretraining method. For $P_{MagLevel}$ and P_{JigMag}, they were shown to benefit the downstream pathology prediction tasks more as compared to pathology-agnostic SSL tasks. However, they were not designed with a specific downstream task in mind, and thus, the learned representations might not be as closely relevant to the downstream task. Results show that D_{JigMag} generally performed better than $D_{MagLevel}$ under 100% annotation budget, consistent with the findings in Koohbanani et al [11]. Further, in general, results show that the contrastive learning methods P_{BYOL} and $P_{SimSiam}$ do not benefit the downstream results as much as $P_{MagLevel}$ and P_{JigMag} do. These two contrastive learning methods aim to learn feature representations by maximizing the feature similarity between a positive image pair formed by two transformed versions of the original image. The purpose is to make the model learn representations invariant to various transformations. However, false positive pairs might exist in heterogeneous pathology images such as LUAD WSIs during the positive view generation step [29]. For example, if the image transformation involves random cropping, then there is a chance that the two randomly cropped regions are semantically different (different tissue categories), but the model is still forced to learn similar representations from them. That might be misleading for the model learning process.

A unique aspect of our proposed SSL tasks is that they were devised based on understanding features and relationships that are helpful in LUAD histologic subtype classification. As shown in Figure 1a, each subtype has different tissue morphology, gland architecture,

and cytoplasmic features. Our SSL tasks were devised to learn those tissue features from unlabeled data. Both proposed SSL tasks $P_{Contained}$ and P_{Grid} are related to learning the spatial relationships between a lower and higher magnification tile cropped from a multi-resolution WSI. The hypothesis is that these two tasks induce the model to learn to distinguish different tissue structures presented in the WSIs, thus benefiting the downstream classification where those structures are also present. The third proposed SSL task P_{Stain} aims to learn the cytoplasm structures by predicting an E-stain from an H-stain. This task can benefit the downstream classification since each LUAD histologic subtype is characterized by different cytoplasmic features [3].

Further, the percent improvement of average F1 scores from $D_{FromScratch}$ to $D_{Ensemble}$ indicates that our SSL tasks can benefit the downstream performance more when less downstream data is annotated. This trend is also commonly observed in other SSL works [11], [30]. Therefore, one can use SSL based on their downstream annotation budget and the amount of available unlabeled data. In general, there is no specific rule on what a reasonable annotation budget is. Typically, the desired amount of annotated data depends on the prediction task's complexity and the data's representativeness. If the task is simple and the data is homogeneous and representative of the task, then SSL might not be needed. However, if the task is relatively complex, data is heterogeneous, and unlabeled data is abundant, then SSL can be a good option. Our results also indicate that tailoring the SSL task to the downstream task is important. The SSL task should be designed to allow the model to learn downstream-relevant information. For example, if color is an important feature for the downstream task, one can design an SSL task related to learning different color schemes in the images. The pipeline of our work can be extended to any other prediction tasks where tissue architectural information is important, such as prostate cancer Gleason pattern grading [31].

In general, F1 scores on each class show that for the proposed ensemble model, lepidic, solid, and non-tumor are the easier classes, whereas acinar, papillary, and micropapillary are the harder classes. The confusion between acinar and papillary, between papillary and micropapillary, was also reflected in Figure 5 and the disagreement between the pathologists during tile labeling. The reason might be that acinar, papillary, and micropapillary patterns all have similar architectural features, such as acini and papillae [3], [32], which adds more confusion and challenge to the model learning. These trends align with the findings from Gertych et al. and Wei et al. that the models had better performance in solid, lepidic, and non-tumor classes than the others [6], [7]. Although we used fewer expert annotations, the F1 scores for each LUAD histologic subtype are on par with or better than the ones in Gertych et al. [7] and Wei et al. [6], indicating that our SSL tasks can potentially not only reduce labeling efforts but also improve model performance.

The clinical impact of our SSL-pretrained LUAD histologic subtype classification model is that it can expedite and complement the process of routine pathology diagnosis tasks. As illustrated in Figure 1e and Figure 6, WSIs can be given to the model, which generates tile-level predictions overlaid on the original WSIs. The prediction heatmaps can then be presented to the pathologists as the starting point of the diagnosis task. The pathologists will further verify or modify the model predictions as needed.

A limitation of this work is that it currently lacks generalizability on external validation sets. For example, when using CPTAC data as an external validation set in the downstream task, the results were much worse (average F1 scores around 0.5, see Supplementary Table 2) than when mixing CPTAC data with NLST and TCGA data for training and evaluation (average F1 scores around 0.9), which is what we did in this work. As presented by Howard et al. [33], there are many site-specific characteristics, including sample preprocessing, slide staining, scanner parameters, and population differences. We tried different stain normalization techniques, such as Macenko [34] and Reinhard [35] normalization and color jittering as data transformation during training, but the external validation results were not improved. These results indicate that other site-specific factors may contribute to the suboptimal results. For example, NLST and TCGA WSIs were acquired from formalin-fixed paraffin-embedded (FFPE) blocks, whereas CPTAC WSIs were obtained from frozen specimens. In addition, the scanner parameters differ between NLST, TCGA, and CPTAC. Unfortunately, we have yet to find a method effective at mitigating these site-specific differences. Another limitation is that the model prediction of the LUAD histologic subtype is at the tile level, which cannot accurately characterize very heterogeneous regions.

6. Conclusion

In summary, we demonstrated the benefit of tailoring SSL tasks to the downstream task by proposing three SSL pretext tasks that induce the model to learn important tissue features and morphology closely relevant to the downstream task of LUAD histologic subtype classification. Extensive experiments were conducted to show the advantage of our proposed SSL tasks over the other state-of-the-art pretraining methods. Although the disease domain of this work is in LUAD, one can easily use our proposed SSL tasks in any other prediction tasks of any domains where tissue architectural information is important. Going beyond LUAD histologic subtype classification, as part of our future work, we will further characterize and quantify the tumor microenvironment features specific to each LUAD subtype and leverage both such features and deep features extracted from the trained subtype classifier to predict patient outcomes such as overall survival and recurrence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the National Cancer Institute (NCI) for granting access to public datasets such as NLST, TCGA, and CPTAC. This work was supported by NIH/National Cancer Institute R01 CA210360, NIH/National Cancer Institute R01 CA226079, NIH/National Cancer Institute U2C CA271898, and the V Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1]. Siegel R, Ward E, Brawley O, and Jemal A, “Cancer statistics, 2011,” CA Cancer J Clin, vol. 61, no. 4, pp. 212–236, Jul. 2011, doi: 10.3322/caac.20121. [PubMed: 21685461]

[2]. Nicholson AG et al. , “The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015,” Journal of Thoracic Oncology, vol. 17, no. 3, pp. 362–387, 2022, doi: 10.1016/j.jtho.2021.11.003. [PubMed: 34808341]

[3]. Kuhn E, Morbini P, Cancellieri A, Damiani S, Cavazza A, and Comin CE, “Adenocarcinoma classification: Patterns and prognosis,” Pathologica, vol. 110, no. 1, pp. 5–11, 2018. [PubMed: 30259909]

[4]. Miyoshi T et al. , “Early-Stage Lung Adenocarcinomas With a Micropapillary Pattern, a Distinct Pathologic Marker for a Significantly Poor Prognosis,” Am J Surg Pathol, vol. 27, no. 1, pp. 101–109, 2003. [PubMed: 12502932]

[5]. Thunnissen E et al. , “Reproducibility of histopathological subtypes and invasion in pulmonary adenocarcinoma. An international interobserver study,” Modern Pathology, vol. 25, no. 12, pp. 1574–1583, Dec. 2012, doi: 10.1038/MODPATHOL.2012.106. [PubMed: 22814311]

[6]. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, and Hassanpour S, “Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks,” Sci Rep, vol. 9, no. 1, pp. 1–8, 2019, doi: 10.1038/s41598-019-40041-7. [PubMed: 30626917]

[7]. Gertych A et al. , “Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides,” Sci Rep, vol. 9, no. 1, pp. 1–12, 2019, doi: 10.1038/s41598-018-37638-9. [PubMed: 30626917]

[8]. Deng J, Dong W, Socher R, Li L-J, Li Kai, and Fei-Fei Li, “ImageNet: A large-scale hierarchical image database,” pp. 248–255, Mar. 2010, doi: 10.1109/CVPR.2009.5206848.

[9]. Huh M, Agrawal P, and Efros AA, “What makes ImageNet good for transfer learning?,” Aug. 2016, doi: 10.48550/arxiv.1608.08614.

[10]. Krishnan R, Rajpurkar P, and Topol EJ, “Self-supervised learning in medicine and healthcare,” Nature Biomedical Engineering 2022 6:12, vol. 6, no. 12, pp. 1346–1352, Aug. 2022, doi: 10.1038/s41551-022-00914-1.

[11]. Koohbanani NA, Unnikrishnan B, Khurram SA, Krishnaswamy P, and Rajpoot N, “Self-Path: Self-Supervision for Classification of Pathology Images with Limited Annotations,” IEEE Trans Med Imaging, vol. 40, no. 10, pp. 2845–2856, 2021, doi: 10.1109/TMI.2021.3056023. [PubMed: 33523807]

[12]. Chen T, Kornblith S, Norouzi M, and Hinton G, “A Simple Framework for Contrastive Learning of Visual Representations,” 37th International Conference on Machine Learning, ICML 2020, vol. PartF168147–3, pp. 1575–1585, Feb. 2020, Accessed: Apr. 05, 2023. [Online]. Available: https://arxiv.org/abs/2002.05709v3

[13]. Gatsonis CA et al. , “The national lung screening trial: Overview and study design,” Radiology, vol. 258, no. 1, pp. 243–253, Jan. 2011, doi: 10.1148/RADIOL.10091808/-/DC1. [PubMed: 21045183]

[14]. Weinstein JN et al. , “The Cancer Genome Atlas Pan-Cancer analysis project,” Nature Genetics 2013 45:10, vol. 45, no. 10, pp. 1113–1120, Sep. 2013, doi: 10.1038/ng.2764. [PubMed: 24071849]

[15]. Gillette MA et al. , “Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma,” Cell, vol. 182, no. 1, pp. 200–225.e35, Jul. 2020, doi: 10.1016/J.CELL.2020.06.013. [PubMed: 32649874]

[16]. He K, Zhang X, Ren S, and Sun J, “Deep Residual Learning for Image Recognition,” Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.

[17]. Weng W and Zhu X, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” IEEE Access, vol. 9, pp. 16591–16603, May 2015, doi: 10.1109/ACCESS.2021.3053408.

[18]. Vahadane A et al. , “Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images,” IEEE Trans Med Imaging, vol. 35, no. 8, pp. 1962–1971, Aug. 2016, doi: 10.1109/TMI.2016.2529665. [PubMed: 27164577]

[19]. Doersch C, Gupta A, and Efros AA, “Unsupervised visual representation learning by context prediction,” Proceedings of the IEEE International Conference on Computer Vision, vol. 2015 Inter, pp. 1422–1430, 2015, doi: 10.1109/ICCV.2015.167.

[20]. McHugh ML, “Interrater reliability: the kappa statistic,” Biochem Med (Zagreb), vol. 22, no. 3, pp. 276–282, Oct. 2012. [PubMed: 23092060]

[21]. Cohen J, “A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES 1”.

[22]. Grill JB et al. , “Bootstrap your own latent a new approach to self-supervised learning,” Adv Neural Inf Process Syst, vol. 2020-Decem, 2020.

[23]. Chen X and He K, “Exploring Simple Siamese Representation Learning,” in Computer Vision and Pattern Recognition , 2021. Accessed: Nov. 07, 2022. [Online]. Available: https://github.com/facebookresearch/simsiam

[24]. Benjamini Y and Hochberg Y, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/J.2517-6161.1995.TB02031.X.

[25]. Warth A et al. , “Interobserver variability in the application of the novel IASLC/ATS/ERS classification for pulmonary adenocarcinomas,” European Respiratory Journal, vol. 40, no. 5, pp. 1221–1227, Nov. 2012, doi: 10.1183/09031936.00219211. [PubMed: 22408209]

[26]. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, and Batra D, “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization.” pp. 618–626, 2017. Accessed: Aug. 02, 2023. [Online]. Available: http://gradcam.cloudcv.org

[27]. Zhang Y et al. , “The prognostic and predictive value of solid subtype in invasive lung adenocarcinoma,” Scientific Reports 2014 4:1, vol. 4, no. 1, pp. 1–6, Nov. 2014, doi: 10.1038/srep07163.

[28]. Tsao MS et al. , “Subtype classification of lung adenocarcinoma predicts benefit from adjuvant chemotherapy in patients undergoing complete resection,” Journal of Clinical Oncology, vol. 33, no. 30, pp. 3439–3446, 2015, doi: 10.1200/JCO.2014.58.8335. [PubMed: 25918286]

[29]. Stacke K, Unger J, Lundström C, and Eilertsen G, “Learning Representations with Contrastive Self-Supervised Learning for Histopathology Applications,” Dec. 2021, Accessed: Apr. 05, 2023. [Online]. Available: https://arxiv.org/abs/2112.05760v2

[30]. Bengar JZ, van de Weijer J, Twardowski B, and Raducanu B, “Reducing Label Effort: Self-Supervised meets Active Learning,” pp. 1631–1639, 2021, [Online]. Available: http://arxiv.org/abs/2108.11458

[31]. Epstein JI, “An Update of the Gleason Grading System,” J Urol, vol. 183, no. 2, pp. 433–440, Feb. 2010, doi: 10.1016/J.JURO.2009.10.046. [PubMed: 20006878]

[32]. Solis LM et al. , “Histologic patterns and molecular characteristics of lung adenocarcinoma associated with clinical outcome,” Cancer, vol. 118, no. 11, pp. 2889–2899, Jun. 2012, doi: 10.1002/CNCR.26584. [PubMed: 22020674]

[33]. Howard FM et al. , “The impact of site-specific digital histology signatures on deep learning model accuracy and bias,” Nature Communications 2021 12:1, vol. 12, no. 1, pp. 1–13, Jul. 2021, doi: 10.1038/s41467-021-24698-1.

[34]. Macenko M et al. , “A method for normalizing histology slides for quantitative analysis,” Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009, pp. 1107–1110, 2009, doi: 10.1109/ISBI.2009.5193250.

[35]. Reinhard E, Ashikhmin M, Gooch B, and Shirley P, “Color transfer between images,” IEEE Comput Graph Appl, vol. 21, no. 5, pp. 34–41, Sep. 2001, doi: 10.1109/38.946629.

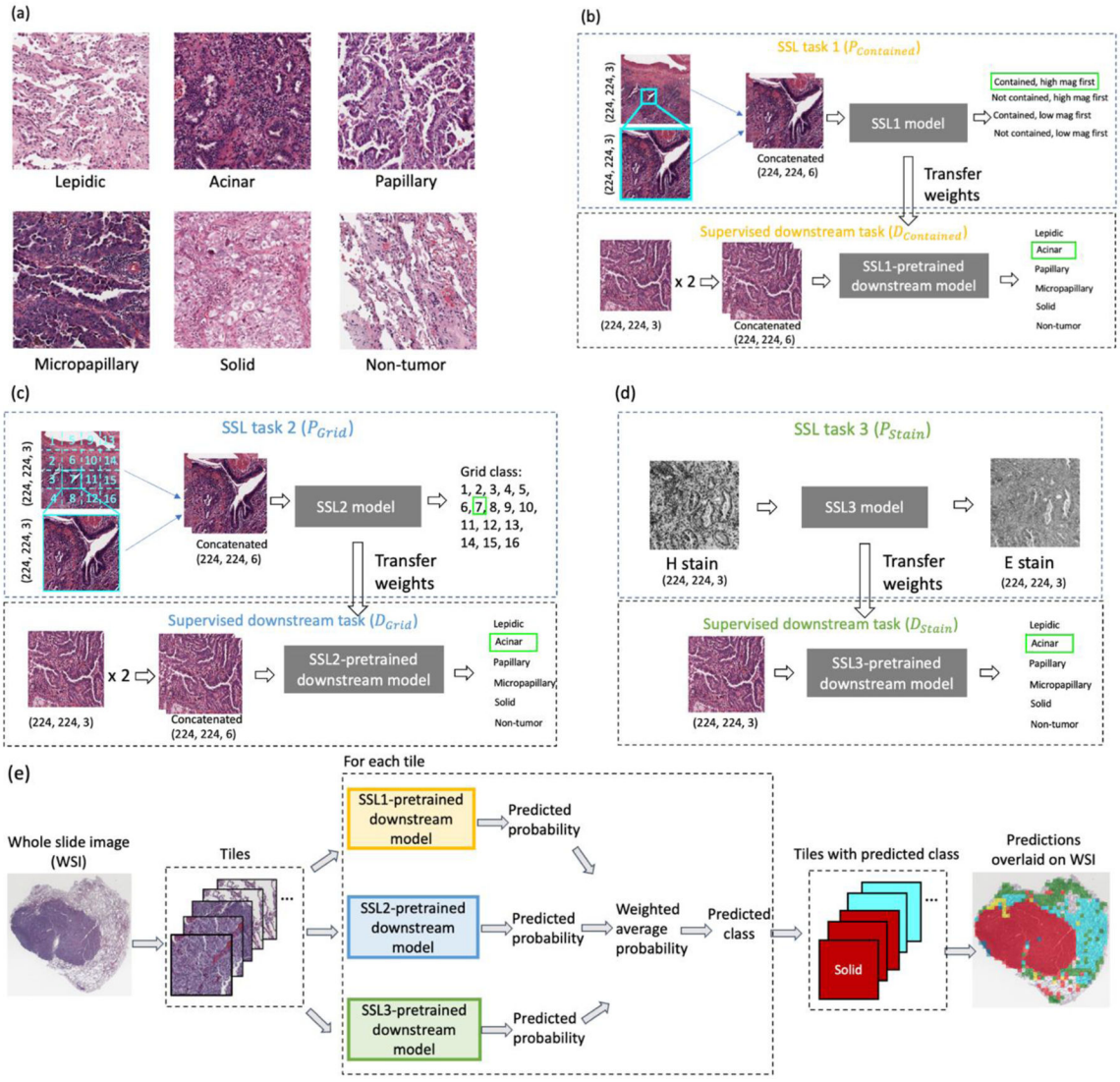


Figure 1.

Overview of the SSL-pretrained LUAD histologic subtype classification pipeline. (a) Example tiles for each class. (b) Workflow of proposed SSL task $P_{Contained}$. (c) Workflow of proposed SSL task P_{Grid}. (d) Workflow of proposed SSL task P_{Stain}. (e) Workflow of inference on new data.

Algorithm 1 $P_{Contained}$

Input

- N pairs of low (L_i) and high (H_i) magnification image tiles (L_i, H_i) randomly concatenated channel-wise as either $[L_i, H_i]$ or $[H_i, L_i]$; H_i may or may not be contained in L_i
- True class labels Y_i. The four classes are:
  - H_i is contained in L_i and concatenation order is $[H_i, L_i]$
  - H_i is not contained in L_i and concatenation order is $[H_i, L_i]$
  - H_i is contained in L_i and concatenation order is $[L_i, H_i]$
  - H_i is not contained in L_i and concatenation order is $[L_i, H_i]$
- The maximum number of training steps T
- A ResNet18 classifier consisting of an encoder E parameterized by θ and a linear predictor P parameterized by ϕ where P is sequentially ordered after E

Output Updated θ, to be transferred to the downstream classifier

for $t = 1, ..., T$ **do**

1. Sample a mini-batch of (L_i, H_i) and corresponding Y_i
2. Apply random horizontal flip, random vertical flip, random rotation, and random brightness jittering on L_i and H_i, respectively
3. Update both θ and ϕ by taking a step on mini-batch cross-entropy loss between the predicted (Y_i') and true (Y_i) class labels, using Adam optimizer

end for

Algorithm 2 P_{Grid}

Input

- N pairs of low (L_i) and high (H_i) magnification image tiles (L_i, H_i) concatenated channel-wise $[H_i, L_i]$ where H_i is contained in L_i; $i \in (0, d^2]$
- d, the downsampling factor from H_i to L_i in the whole slide image
- True class labels Y_i: A total of d^2 classes, with each class representing one of the d^2 spatial grids H_i resides in relative to L_i
- The maximum number of training steps T
- A ResNet18 classifier consisting of an encoder E parameterized by θ and a linear predictor P parameterized by ϕ where P is sequentially ordered after E

Output Updated θ, to be transferred to the downstream classifier

for $t = 1, ..., T$ **do**

1. Sample a mini-batch of (L_i, H_i) and corresponding Y_i
2. Apply random rotation on L_i and H_i respectively
3. Update both θ and ϕ by taking a step on mini-batch cross-entropy loss between the predicted (Y_i') and true (Y_i) class labels, using Adam optimizer

end for

Algorithm 3 P_{Stain}

Input

- N hematoxylin-stained image tiles H_i and their corresponding eosin-stained image tiles E_i
- The maximum number of training steps T
- A ResNet18-UNet consisting of an encoder E parameterized by θ and a decoder D parameterized by ϕ where D is sequentially ordered after E

Output Updated θ, to be transferred to the downstream classifier

for $t = 1, ..., T$ **do**

1. Sample a mini-batch of H_i and E_i
2. Update both θ and ϕ by taking a step on mini-batch L1 loss between the predicted (E_i') and (E_i), using Adam optimizer

end for

Figure 2.

Pseudocode for each proposed SSL task $P_{Contained}$, P_{Grid}, and P_{Stain}.

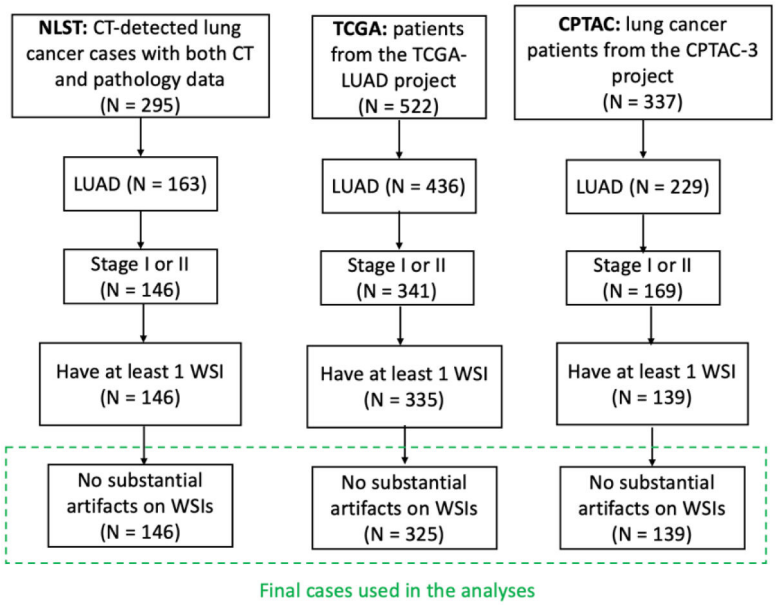
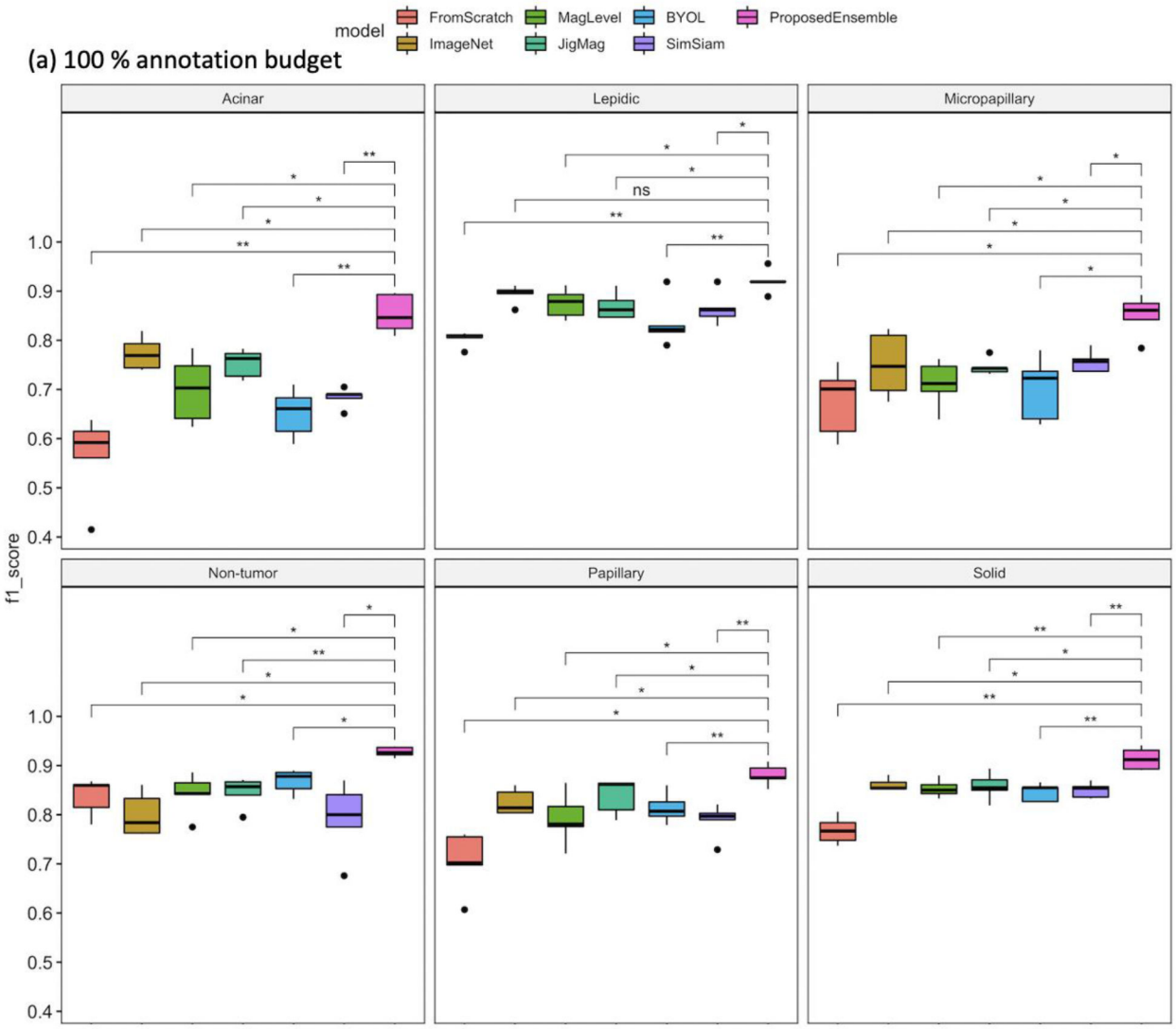


Figure 3.
Patient selection diagrams for NLST, TCGA, and CPTAC cohorts.



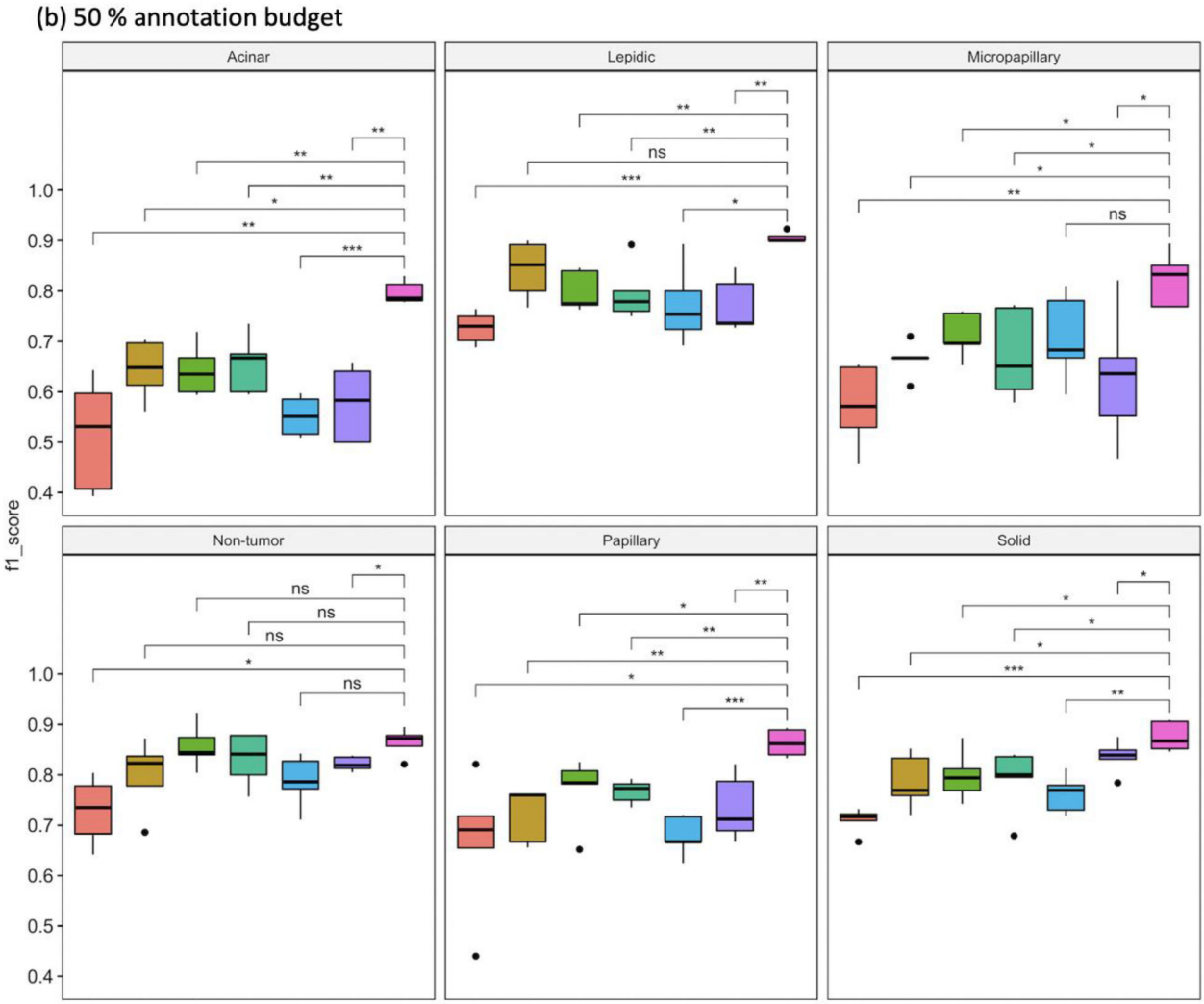


Figure 4.

Box plots from 5-fold cross-validation test sets for each downstream model initialized with different pretraining methods. Statistical significance after correcting for multiple comparisons was represented by * ($p \leq 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), **** ($p < 0.0001$), or ns ($p > 0.05$).

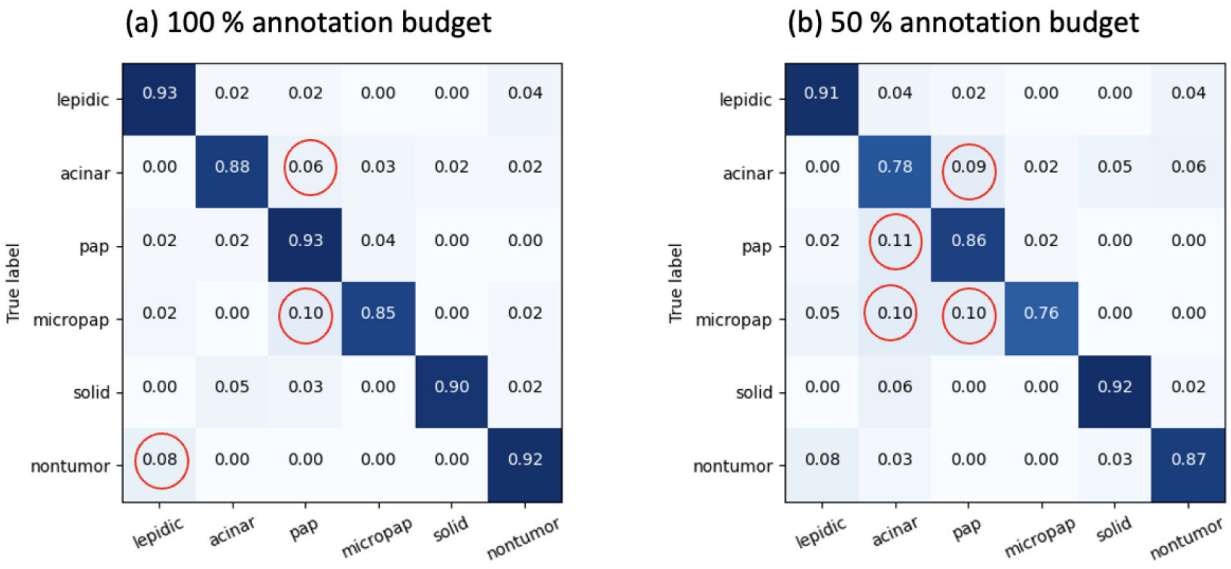


Figure 5.

Confusion matrix from one of the test set folds in the ensemble models under 100% and 50% annotation budget.

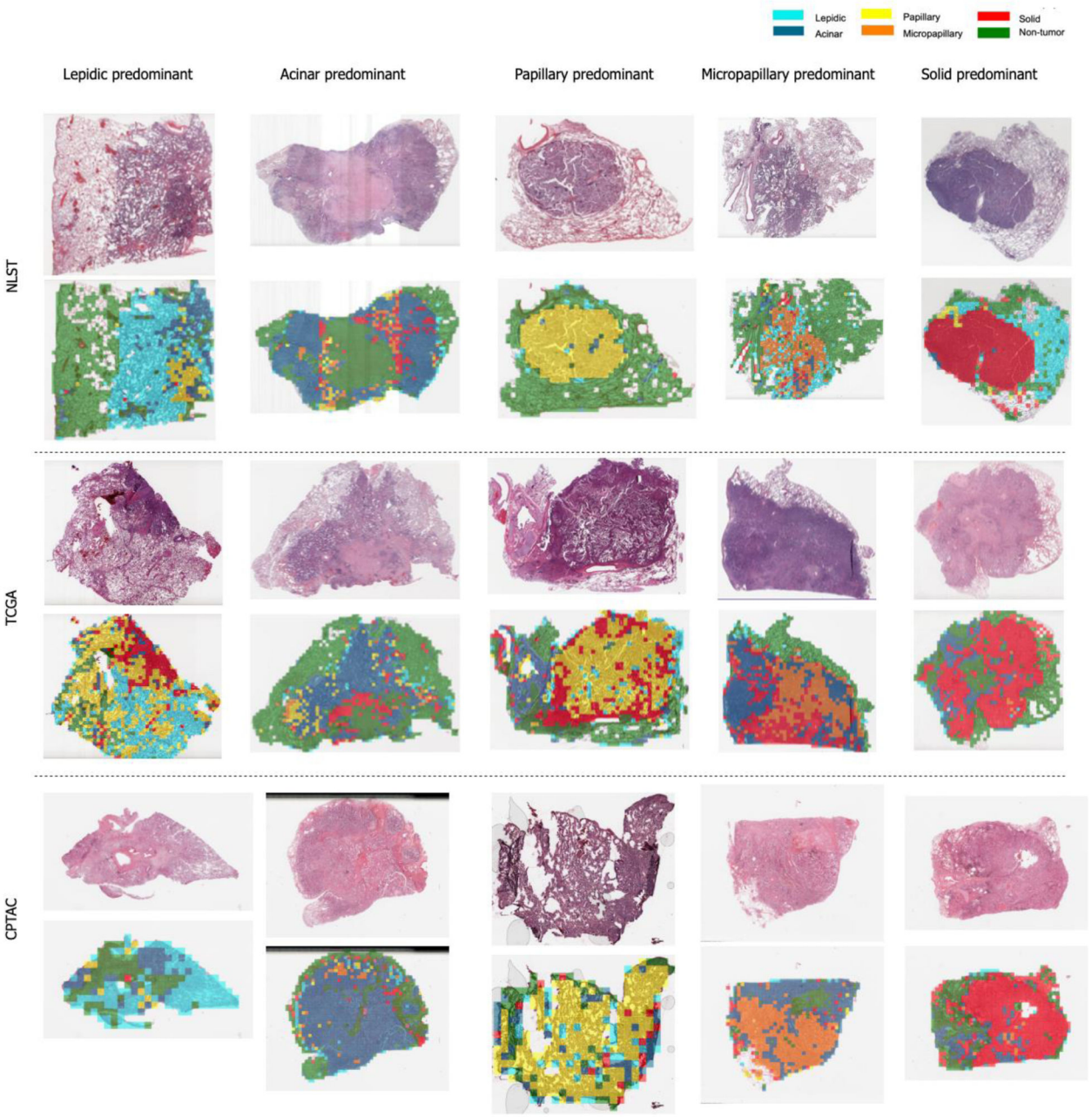


Figure 6.

Tile-level prediction of LUAD histologic subtypes overlaid on the original WSI by the ensemble model under 100% annotation budget. The predominant histologic subtype of each WSI was derived by the pathologists blinded from the model prediction.

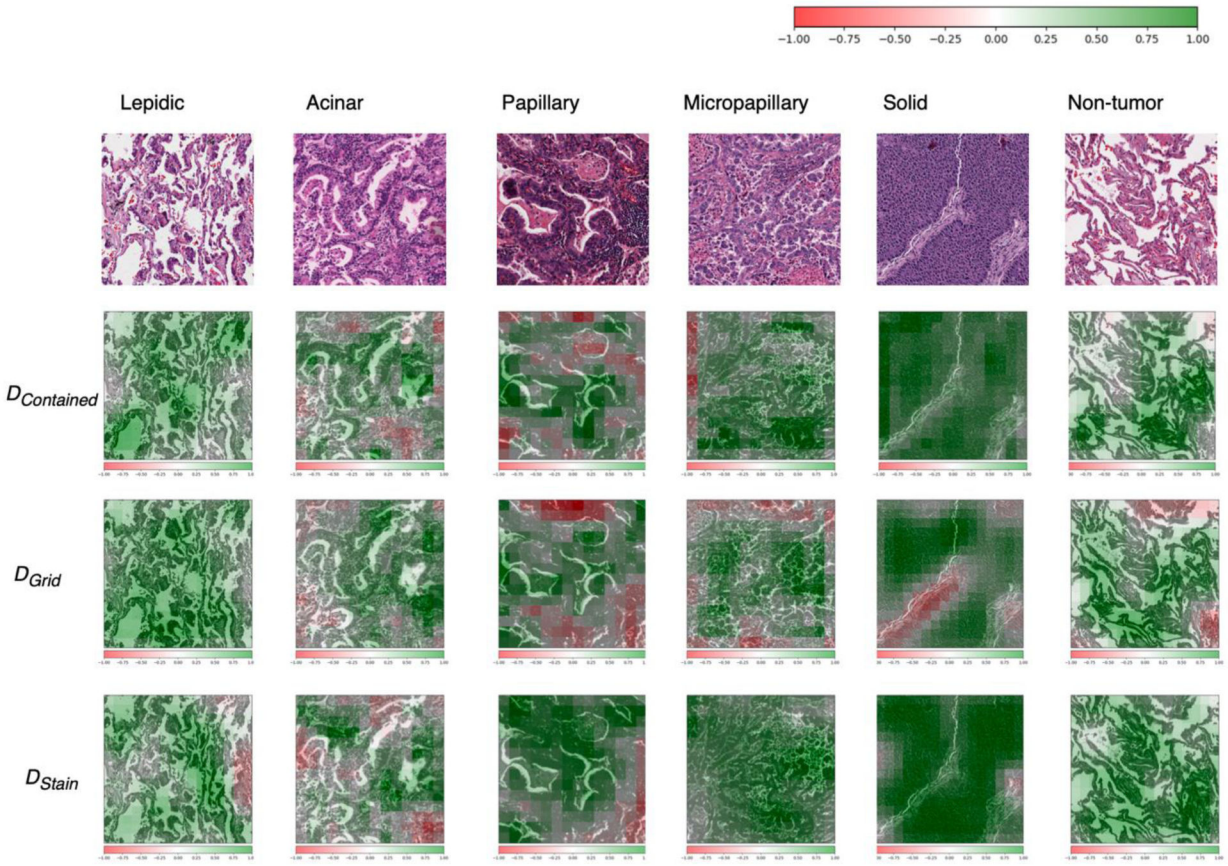


Figure 7.

GradCAM visualization of correctly classified example test set tiles predicted by $D_{Contained}$, D_{Grid}, and D_{Stain}. Negative values (red) indicate that the pixels negatively contribute to the prediction of that class, while positive values (green) indicate that the pixels positively contribute to the prediction.

Table 1.

Patient characteristics of the three datasets used in this study.

	# Patients	# WSIs	Mean age	# Stage I	# Females
NLST	146	407	63.7	121 (82.9%)	72 (49.3%)
TCGA	325	355	65.3	227 (69.8%)	175 (53.9%)
CPTAC	139	667	64.5	91 (65.5%)	50 (36.0%)

Table 2.

Number of samples/tiles used in SSL and downstream tasks respectively, stratified into train, validation, and/or test sets.

	SSL (# of samples)		**Downstream (# of tiles)**		
	Training (80%)	***Validation (20%)***	***Training (60%)***	***Validation (20%)***	***Testing (20%)***
NLST	131,693	32,916	392	134	164
TCGA	178,508	44,604	116	41	46
CPTAC	0	0	480	161	163
Total	310,201	77,520	988	336	373

Table 3.

Total number of tiles used for each class in downstream classification.

	Lepidic	Acinar	Papillary	Micropapillary	Solid	Non-tumor	Total
NLST	123	122	111	49	133	152	690
TCGA	30	34	35	32	37	35	203
CPTAC	120	160	136	120	136	132	804

Table 4.

Comparison between without versus with shortcut-avoiding techniques in downstream tasks $D_{Contained}$ and D_{Grid}. Results are average F1 score on test sets (n = 373) for downstream classification task D.

		Lepidic (n = 55)	**Acinar (n = 64)**	**Papillary (n = 58)**	**Micropapillary (n = 41)**	**Solid (n = 63)**	**Non-tumor (n = 92)**
$D_{Contained}$ **(proposed)**	Binary prediction of "contained" vs "not contained"	0.877 ± 0.0322	0.717 ± 0.0243	0.838 ± 0.0345	0.752 ± 0.0113	0.860 ± 0.0127	0.829 ± 0.0580
	4-class prediction (binary + tile order)	0.887 ± 0.0344	0.785 ± 0.0275	0.862 ± 0.0365	0.807 ± 0.0382	0.878 ± 0.0210	0.834 ± 0.0231
	4-class prediction + data transform	**0.904 ± 0.0207**	**0.829 ± 0.0349**	**0.897 ± 0.0352**	**0.844 ± 0.0341**	**0.900 ± 0.0362**	**0.849 ± 0.0345**
D_{Grid} **(proposed)**	16-class prediction	0.815 ± 0.0220	0.718± 0.0458	0.780 ± 0.0344	0.720 ± 0.0376	0.844 ± 0.0070	0.813 ± 0.0322
	16-class prediction + data transform	**0.900 ± 0.0194**	**0.797 ± 0.0209**	**0.861 ± 0.0264**	**0.812 ± 0.0471**	**0.874 ± 0.0233**	**0.860 ± 0.0336**

Table 5.

Average F1 score on test sets (n = 373) for downstream classification task D, using 100 % expert annotation budget during training and validation. The best individual model was underlined, and the overall best model (individual or ensemble) was bolded.

$D_{Pretrain}$	**Lepidic (n = 55)**	**Acinar (n = 64)**	**Papillary (n = 58)**	**Micropapillary (n = 41)**	**Solid (n = 63)**	**Non-tumor (n = 92)**
$D_{FromScratch}$	0.803 ± 0.0137	0.564 ± 0.0786	0.704 ± 0.0551	0.676 ± 0.0638	0.768 ± 0.0248	0.837 ± 0.0340
$D_{ImageNet}$	0.894 ± 0.0167	0.773 ± 0.0298	0.825 ± 0.0231	0.751 ± 0.0589	0.861 ± 0.0113	0.800 ± 0.0401
$D_{MagLevel}$	0.875 ± 0.0262	0.700 ± 0.0610	0.792 ± 0.0475	0.711 ± 0.0431	0.854 ± 0.0160	0.842 ± 0.0374
D_{JigMag}	0.870 ± 0.0239	0.753 ± 0.0258	0.838 ± 0.0320	0.746 ± 0.0150	0.858 ± 0.0246	0.846 ± 0.0278
D_{BYOL}	0.835 ± 0.0437	0.652 ± 0.0442	0.814 ± 0.0277	0.700 ± 0.0607	0.846 ± 0.0162	0.868 ± 0.0221
$D_{SimSiam}$	0.865 ± 0.0299	0.683 ± 0.0178	0.788 ± 0.0313	0.756 ± 0.0197	0.850 ± 0.0140	0.792 ± 0.0664
$D_{Contained}$**(proposed)**	0.904 ± 0.0207	0.829 ± 0.0349	**0.897 ± 0.0352**	0.844 ± 0.0341	0.900 ± 0.0362	0.849 ± 0.0345
D_{Grid}**(proposed)**	0.900 ± 0.0194	0.797 ± 0.0209	0.861 ± 0.0264	0.812 ± 0.0471	0.874 ± 0.0233	0.860 ± 0.0336
D_{Stain}**(proposed)**	0.893 ± 0.0327	0.833 ± 0.0202	0.863 ± 0.0426	0.801 ± 0.0242	0.907 ± 0.0182	0.925 ± 0.0174
$D_{Ensemble}$**(proposed)**	**0.918 ± 0.0239**	**0.851 ± 0.0317**	0.881 ± 0.0192	**0.849 ± 0.0330**	**0.915 ± 0.0192**	**0.928 ± 0.00906**

Table 6.

Average F1 score on test sets (n = 373) for downstream classification task D, using 50 % expert annotation budget during training and validation. The best individual model was underlined, and the best model of any kind (individual or ensemble) was bolded.

$D_{Pretrain}$	**Lepidic (n = 55)**	**Acinar (n = 64)**	**Papillary (n = 58)**	**Micropapillary (n = 41)**	**Solid (n = 63)**	**Non-tumor (n = 92)**
$D_{FromScratch}$	0.727 ± 0.0281	0.514 ± 0.100	0.665 ± 0.125	0.572 ± 0.0739	0.710 ± 0.0226	0.729 ± 0.0593
$D_{ImageNet}$	0.842 ± 0.0519	0.644 ± 0.0529	0.721 ± 0.0492	0.664 ± 0.0313	0.787 ± 0.0491	0.799 ± 0.0641
$D_{MagLevel}$	0.800 ± 0.0357	0.643 ± 0.0461	0.770 ± 0.0611	0.712 ± 0.0405	0.798 ± 0.0443	0.857 ± 0.0396
D_{JigMag}	0.796 ± 0.0509	0.654 ± 0.0519	0.766 ± 0.0208	0.675 ± 0.0807	0.790 ± 0.0583	0.831 ± 0.0470
D_{BYOL}	0.773 ± 0.0698	0.551 ± 0.0353	0.679 ± 0.0356	0.707 ± 0.0787	0.762 ± 0.0343	0.788 ± 0.0463
$D_{SimSiam}$	0.772 ± 0.0493	0.577 ± 0.0672	0.735 ± 0.0592	0.628 ± 0.119	0.835 ± 0.0298	0.822 ± 0.0129
$D_{Contained}$**(proposed)**	0.883 ± 0.0178	0.767 ± 0.0412	0.849 ± 0.0221	0.726 ± 0.0505	0.843 ± 0.0504	0.849 ± 0.0849
D_{Grid}**(proposed)**	0.873 ± 0.0313	0.723 ± 0.0245	0.787 ± 0.0531	0.757 ± 0.0701	0.855 ± 0.0428	0.802 ± 0.0669
D_{Stain}**(proposed)**	0.864 ± 0.0362	0.766 ± 0.0355	0.815 ± 0.0558	0.768 ± 0.0716	**0.890 ± 0.0140**	**0.900 ± 0.0515**
$D_{Ensemble}$**(proposed)**	**0.903 ± 0.0150**	**0.786 ± 0.0144**	**0.856 ± 0.0244**	**0.822 ± 0.0483**	0.876 ± 0.0264	0.864 ± 0.0251