Generalizing Few-Shot Classification of Whole-Genome Doubling Across Cancer Types

Sherry Chao
Harvard University
Cambridge, MA, USA
schao@g.harvard.edu

David Belanger
Google Brain
Cambridge, MA, USA
dbelanger@google.com

Abstract

The study and treatment of cancer is traditionally specialized to the cancer's primary site of origin. However, certain phenotypes are shared across cancer types and have important implications for clinical care. To date, automating the identification of these characteristics from routine clinical data - irrespective of the type of cancer - is impaired by tissue-specific variability and limited labeled data. Whole-genome doubling is one such phenotype; whole-genome doubling events occur in nearly every type of cancer and have significant prognostic implications. Using digitized histopathology slide images of primary tumor biopsies, we train a deep neural network model end-to-end to accurately generalize few-shot classification of whole-genome doubling across 17 cancer types. By taking a meta-learning approach, cancer types are treated as separate but jointly-learned tasks. This approach outperforms a traditional neural network classifier and quickly generalizes to both held-out cancer types and batch effects. These results demonstrate the unrealized potential for meta-learning to not only account for between-cancer type variability but also remedy technical variability, enabling real-time identification of cancer phenotypes that are too often costly and inefficient to obtain.

1. Introduction

Genomic characteristics of a patient's cancer, such as gene mutations and aneuploidy, are increasingly used to improve the course of care [55, 66, 67, 18, 63]. In spite of their clinical benefit, these characteristics not only are difficult to measure from routinely-collected patient data but also necessitate measurements across cancer types, made difficult due to cancer's inherently heterogeneous nature and the limited size of patient cohorts. There is an unmet need to build tools that automate fast identification of cancer phenotypes from routinely-collected patient data, irrespective of cancer type, particularly when the phenotype (i) has prognostic or therapeutic implications and (ii) is expensive or slow to measure under traditional means.

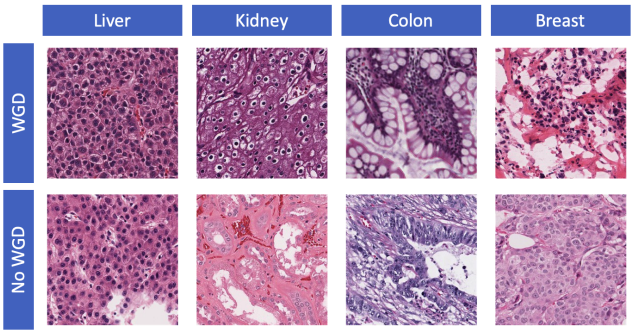

Figure 1. Representative histopathology image tiles from four cancer types, depicting tumor biopsy samples with (**top**) and without (**bottom**) whole-genome doubling.

Whole-genome doubling (WGD) is one such phenotype. WGD is a genome-wide aberration characterized by the presence of at least twice the normal number of chromosomes and is associated with advanced metastasis and overall poor prognosis [1]. Patients with WGD events are more prone to aneuploidy, which lends itself to more aggressive treatment regimens for multiple cancer types [13, 33]. Moreover, WGD itself confers unique vulnerabilities that can be therapeutically targetable [44, 58, 10, 48]. The prevalence and prognostics of WGD merits knowledge of WGD status in determining the course of care; however, measuring WGD is inefficient. Karyotyping costs $11k/diagnosis and DNA sequencing costs $10k/genome, both of which take several weeks to complete [35, 51]. The medical oncology community would significantly benefit from automating WGD identification via more time- and cost-efficient means.

We propose inferring WGD from digitized histopathology images of tumor biopsies, a routinely-collected source of patient data (Figure 1). Across cancer types, the tissue morphology is a manifestation of the genomic characteris-

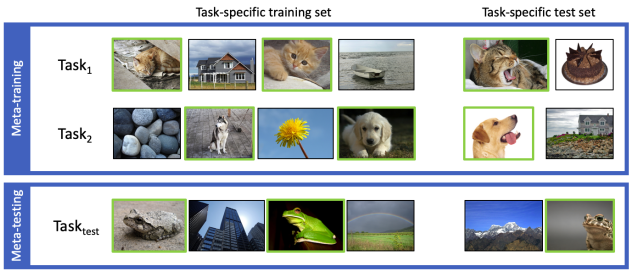

Figure 2. Overview of meta-learning. The model is trained on multiple tasks ("meta-training"), and at deployment time, the model is presented with a small set of labeled examples and quickly bootstraps a task-specific classifier ("meta-testing"). In contrast, under standard supervised learning, the model remains static at deployment time. Positive-labeled examples are highlighted in green.

tics of the tumor. However, histopathology images from different cancer types exhibit tissue-specific characteristics (e.g., colon, lung, skin) even if they share the same WGD status. Traditionally, good performance on cancer-related classification tasks has been achieved via training separate models for each cancer type [28]. This approach has several shortcomings: (i) it necessitates acquiring many training examples from all cancer types, as each model learns from a single cancer type, and (ii) the models are not interchangeable, i.e., a model trained to classify WGD for lung cancer is unable to classify WGD for breast cancer. Successfully integrating machine learning into the clinic necessitates a model that can sufficiently handle inter-cancer diversity [32].

Recent work by Fu et al. [17] to classify WGD from histopathology images across cancer types shows good performance on only seven out of 27 cancer types. We propose using meta-learning to automate the classification of WGD across cancer types (Figure 2). In the meta-learning regime, models are learning to learn from few examples. Let us consider a toy example of a standard meta-learning framework. We are given three small datasets: Dataset A contains images of cats/non-cats, Dataset B contains images of dogs/non-dogs, and Dataset C contains images of horses/non-horses. Each dataset has been curated to train a classifier on its respective label (i.e., cat, dog, horse). Instead of training on each task individually, however, we instead train on how to *learn* to learn the tasks from only a few training examples. At meta-training, the model is presented with eight labeled images from each dataset before being asked to correctly classify eight new unlabeled images from each dataset. Therefore, at meta-test time, when the model is given a fourth dataset, Dataset D, which contains images of frogs/non-frogs, it will ideally have learned to learn a new task. Namely, after the model is presented with eight labeled images from Dataset D, it will be able to accurately classify subsequent images from Dataset D as either a frog or a non-frog.

In meta-learning for classification, several approaches have been proposed, including matching networks and prototypical networks [59, 53]. In this work, we adapt the model-agnostic meta-learning (MAML) framework to the problem of WGD classification from histopathology images across multiple cancer types [16]. We take a multi-task view by treating WGD classification for each cancer type as a separate, learnable task. Under identical models, the MAML training approach is able to outperform standard supervised learning and generalize well to unseen cancer types in the held-out meta-test set.

We subsequently extend this approach to accounting for batch effects, or distributional shifts across histopathology images due to technical variation in data collection. Batch effects are pervasive in biomedical datasets. Whereas under standard practices (e.g., fitting a simple model such as linear regression), we mitigate batch effects by incorporating a batch effect-specific term, the complexity of deep neural network classifiers invalidates this solution because the interactions between variables entangles the batch effect with the effect of interest. Meta-learning addresses this problem by treating each batch as its own task-specific dataset such that the model is focused on learning to learn the task (e.g., WGD classification) instead of learning specifics about the batch (e.g., image resolution or brightness).

Ultimately, we extend the application of meta-learning beyond classifying different labels (one **label** per task) to two novel use cases in the field of medical imaging:

1. Classification of the same label across different cancer types (one **cancer type** per task).

2. Classification of the same label across different batches (one **batch** per task).

Thus, in a real-world scenario where a clinician would like to quickly classify the WGD status for tumor biopsies of a particular cancer type or batch, he/she need only label a small handful of histopathology images in order for the meta-learner to automate labeling of subsequent samples.

2. Related Work

Much effort has been devoted to automating cancer diagnosis by training neural networks to discern tumor from normal tissue in histopathology images [45, 25, 5, 54, 57, 30, 29, 70, 19, 7]. Cancer diagnosis efforts have further delved into cancer subtyping of individual cancer types [23, 37, 42, 62, 9, 11]. Such work has progressed in tandem with detailed tissue segmentation approaches [21, 60]. More recently, increasing efforts are being made in multi-class classification, namely aggregating cancer types in an attempt to accurately diagnose the correct cancer type from all possible cancer types [24, 40, 46].

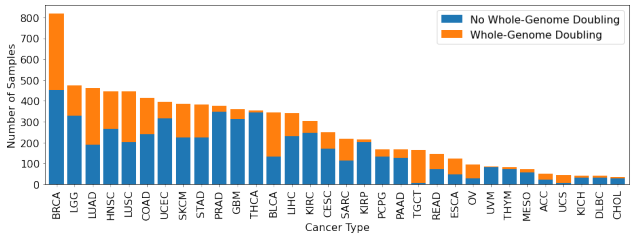

Figure 3. Summary of the distribution of WGD status by cancer type for samples from The Cancer Genome Atlas. Refer to Appendix A in the Supplementary Materials for acronym descriptions.

Applying machine learning to histopathology images in order to infer characteristics of a patient's cancer is a growing research area [15]. One avenue of applications has focused on predicting survival and prognosis [65, 52, 49, 26, 68, 2, 50, 19]. Another avenue has focused on using morphological features to infer molecular features about a patient's cancer [17, 39, 3, 38, 4, 28, 34]. Moreover, studies demonstrate we are now able to predict phenotypes such as microsatellite instability and tumor mutational burden from histopathology images of certain cancer types [14, 27, 6, 28, 64]. While these advancements have important therapeutic implications, their applicability is cancer type-specific. For instance, Fu et al. [17] generate embeddings to predict WGD, but the results do not generalize across cancer types.

The application of meta-learning to medicine is relatively nascent. In medical imaging, Hu et al. [22] train a meta-learning algorithm, Reptile, on mini-ImageNet to classify diabetic retinopathy from eye screenings. In drug discovery, Olier et al. [41] develop Meta-QSAR to predict chemical compound activity against a target protein. In genomics, Runge et al. [47] employ a meta-learner to design RNA sequences that fold into target structures. These approaches have not been applied to cancer, which is ripe for meta-learning given cancer's inherent heterogeneity and the widespread availability of histopathology images. While efforts have been made to cluster cancer types by morphological similarity to improve predictive power, this approach requires manual curation of datasets [43]. In this work, we attempt to fill the gap by marrying a meta-learning training regime to a generalizable cancer classification task.

3. Cohort

3.1. Cohort Selection

The data collected for this analysis comprise RGB-channel images of hematoxylin and eosin (H&E) stained histopathology slides and corresponding WGD status labels for 3,596 samples across 17 cancer types from The Cancer Genome Atlas (TCGA). TCGA is a public database of clinical and genomic data for over 20,000 patient samples spanning 33 cancer types (Figure 3). The analysis focuses on primary tumor, for which we have the diagnostic H&E-stained whole slide image. Ground truth WGD status labels were annotated via analysis of DNA sequencing data [56]. To ensure both image and label availability, we took the intersection of TCGA images provided by the National Cancer Institute (NCI) Genomic Data Commons (GDC) Data Portal and WGD labels provided by Taylor et al. [56].

Cancer types were subsequently chosen based on the number of available images, selecting for cancer types whose number of images was within one standard deviation of the median number of images (BLCA, COAD, ESCA, HNSC, KIRC, LIHC, READ, STAD, UCEC). In order to be able to include common cancer types with more than 450 images (BRCA, LUAD, LUSC) while remaining within our storage constraints, we randomly subsampled 25% (BRCA) or 50% (LUAD, LUSC) of the images to yield 200-250 images per common cancer type. Furthermore, rare cancer types with less than 100 images (ACC, CHOL, KICH, OV, UCS) were included to study performance in real-world scenarios with limited labeled data. Refer to Appendix A in the Supplementary Materials for acronym descriptions.

3.2. Cohort Overview

In total, 42% of slide images, or 1,522 images, were WGD-positive, ranging between 19% and 82% of images by cancer type. The breakdown by clinical features is as follows (percentages do not sum to 100% due to rounding):

1. **Tumor Stage**: Stage I (30.0%), Stage II (26.7%), Stage III (26.4%), Stage IV (14.0%), NA (3.0%)
2. **Gender**: Male (51.0%), Female (49.0%), NA (0.01%)
3. **Age**: $\leq$39 (2.8%), 40-49 (8.0%), 50-59 (19.3%), 60-69 (27.8%), 70-79 (22.2%), $\geq$80 (7.7%), NA (12.3%)
4. **Race**: White (66.4%), Black (10.4%), Asian (10.0%), NA (13.3%)

Refer to Appendix B in the Supplementary Materials for a detailed description of the TCGA dataset by cancer type with relevant clinical features, including tumor stage, patient gender, age, and race.

3.3. Feature Extraction

Of the 3,596 svs images selected for this study, 3,507 were successfully converted to jpeg images; the 89 images that failed during conversation were due to missing metadata in the svs file. To address sample imbalance (i.e., multiple images per patient), we randomly sampled one image to retain from each patient with multiple images, yielding 3,467 images in total.

Since slides are digitized at multiple magnifications, it was important to determine which magnification would be

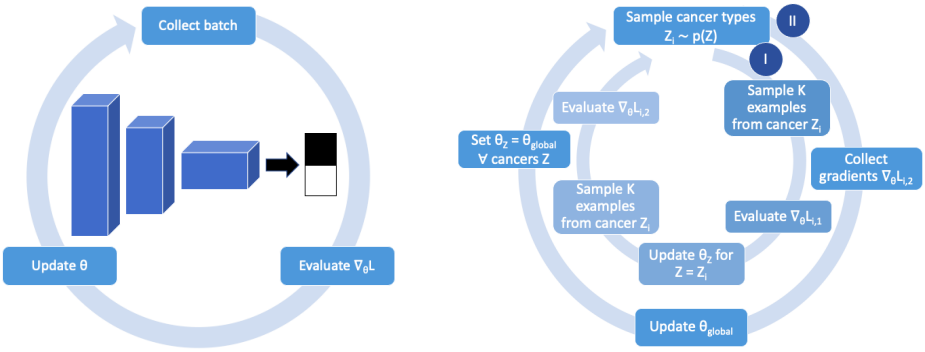

Figure 4. Comparison of training regimes. While traditional approaches to model optimization optimize solely the global parameters (**left**), meta-learning also optimizes parameters that are local to the particular task (e.g., cancer type or batch) (**right**).

most useful for WGD classification. Preliminary training of a ResNet18-based model on WGD classification of colorectal cancer slides at various magnifications (5x, 10x, 20x) showed the best performance at 10x magnification, which yielded the highest slide-level accuracy and AUC. Thus, we selected 10x as our default magnification and extracted images at 10x magnification for training and evaluation.

3.4. Feature Tiles

Each histopathology slide was segmented into adjacent, non-overlapping tiles of dimension $3 \times 256 \times 256$ ($C \times H \times W$), wherein only tiles with less than 50% whitespace were retained. On average, each histopathology image was segmented into 3,155 tiles for a total of 10.9 million tiles. Across the samples, the standard deviation was 1,807 tiles with first, second, and third quartiles of 1,766 tiles, 3,096 tiles, and 4,325 tiles, respectively.

4. Methods

4.1. Model

Let data set $\mathcal{D}$ consist of Z cancer types, each of which is comprised of N_z slide images $s_{z,i}$ and binary labels $y_{z,i}$, $i \in \{1, \ldots, N_z\}$. As described in Section 3.4, each slide image $s_{z,i}$ is segmented into $T_{z,i}$ non-overlapping tiles $x_{z,i,t}$, $t \in \{1, \ldots, T_{z,i}\}$, with C color channels, H pixel height, and W pixel width. The data can be summarized as follows:

$$\mathcal{D} = \{D_1, \ldots, D_Z\}$$

$$D_z = \{(s_{z,i}, y_{z,i})\}_{i=1}^{N_z}$$

$$s_{z,i} = \{x_{z,i,1}, \ldots, x_{z,i,T_{z,i}}\}$$

$$x_{z,i,t} \in \mathbb{R}^{C \times H \times W}$$

$$y_{z,i} = \begin{cases} 1, & \text{if whole-genome doubling} \\ 0, & \text{otherwise} \end{cases}$$

Formally, we train the network to maximize the following likelihood function:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, p\left(Y \mid S, \theta\right) = \underset{\theta}{\operatorname{argmax}} \prod_{z}^{Z} \prod_{i}^{N_z} p(y_{z,i} \mid s_{z,i}, \theta)$$

The predictive distribution $p(y \mid s)$ is parameterized by a deep neural network comprised of a ResNet18 and two fully-connected layers [20]. The ResNet18 architecture includes residual connections, which help to facilitate training of models with many layers. We initialize training with a ResNet18 model pre-trained on the ImageNet dataset [12]. The pre-trained ResNet18 undergoes additional pre-training on images sampled from the meta-train set to tailor its learning to histopathology images. The final fully-connected layers are trained from scratch on the histopathology images while the lower layers are fine-tuned to these images. The fully-connected layers have a hidden size of 512 with dropout and tanh activation function. Because each slide is comprised of a set of tiles, we employ LogSumExp pooling across tiles as a smooth approximation to max pooling, enabling learning across multiple tiles. LogSumExp pooling of the tile-level prediction scores yields the final slide-level prediction score.

4.2. Training

Model training is performed by optimizing the model parameters in order to maximize the log likelihood, yielding the maximum likelihood estimation. Each model is trained on the meta-train set (BLCA, BRCA, COAD, HNSC, LUAD, LUSC, READ, STAD) (2,056 images), tuned on the meta-validation set (ESCA, LIHC) (458 images), and evaluated on the meta-test set (ACC, CHOL, KICH, KIRC, OV, UCS, UCEC) (953 images). Each cancer type is subsequently split into train and test sets, e.g., for ACC, eight samples are randomly selected for the meta-test train set; the remaining samples are selected for the meta-test test set. Prior to training, tile images are normalized by channel to the mean and standard deviation of the meta-train set.

Algorithm 1 Meta-Training

Require: $p(\mathcal{Z})$: distribution over cancer types
Require: α, η: step size hyperparameters
  Initialize θ randomly
  repeat
    Sample batch of cancer types $\mathcal{Z}_i \sim p(\mathcal{Z})$
    for all $\mathcal{Z}_i$ **do**
      Sample K examples from $\mathcal{D}_{\mathcal{Z}_i}$
      Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{Z}_i}(f_\theta)$ with respect to the K examples
      Compute adapted parameters using gradient descent: $\theta_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{Z}_i}(f_\theta)$
      Sample K additional examples from $\mathcal{D}_{\mathcal{Z}_i}$ for the global update
    end for
    Update $\theta \leftarrow \theta - \eta \nabla_\theta \sum_{\mathcal{Z}_i \sim p(\mathcal{Z})} (\mathcal{L}_{\mathcal{Z}_i}(f_{\theta'}))$
  until forever

In our baseline scenario ("CNN"), the model is pre-trained on the meta-train set before undergoing meta-validation and meta-test. In our meta-learning scenario ("MAML"), the model is pre-trained and meta-trained on the meta-train set before undergoing meta-validation and meta-test. Thus, the CNN and MAML models are identical, differing solely in training regime (Figure 4). The models are subject to the same evaluation scheme, and performance is averaged over 40 random initializations. To compare the performance of CNN and MAML, we employ the Wilcoxon signed-rank test [61]. We employ this test because it is able to exploit the paired structure of the experiments, where each machine learning system has been run on the same set of train-test splits.

4.2.1 Pre-Training

During pre-training, the model is trained on WGD classification using all cancer types in the meta-train set. Slides and tiles in the train set are shuffled for every epoch and augmented via random vertical/horizontal flips and color jitter. To prevent overfitting, we apply 50% random dropout at each fully-connected layer. We train for up to 200 epochs with a minibatch size of 24 slides (50 randomly sampled tiles per slide) and a learning rate of 0.0001 with an Adam optimizer [31]. We reduce the learning rate by a factor of 0.1 upon validation loss plateau with a patience of five epochs. To encourage regularization, model parameters are saved when the binary cross entropy loss on the validation set improves upon that of the previous epoch.

4.2.2 Meta-Training

Meta-training proceeds according to Algorithm 1 using the pre-trained embeddings described in Section 4.2.1. At every step, the parameters of the local (cancer type-specific) models are set to the parameters of the global model. We sample a batch of meta-train cancer types $\mathcal{Z}_i$, and from each of these cancer types, we sample a batch of K examples. Using these examples, we perform one gradient update of the local parameters. Next, we sample a second batch of K examples from each cancer type $\mathcal{Z}_i$. Using these examples, we perform one forward pass with their respective local models and store the gradient of the loss with respect to the parameters. Once one batch of cancers is complete, the global parameters are updated using the stored gradients.

In all experiments, we meta-train for up to 50 epochs with a learning rate of 0.0001 for both the local parameter update and the global parameter update. For each update, we sample a batch of 16 slides (50 randomly sampled tiles per slide) from five out of the eight cancer types in the train set, ensuring uniform sampling of the cancer types. Eight slides are used for the local update, and the remaining eight slides are used for the global update. For the local parameter update, we use an Adam optimizer, while for the global parameter update, we use stochastic gradient descent [31].

4.3. Meta-Validation and Meta-Test

Following pre-training (CNN) and meta-training (MAML), we assess the model's few-shot classification performance using a train set size of eight slides (50 randomly sampled tiles per slide) per cancer type (i.e., eight-shot learning). The remaining slides are allocated to the test set. The model takes a fixed number of gradient steps on the meta-validation/-test train set before being evaluated on the meta-validation/-test test set. We employ a learning rate of 0.0001 with an Adam optimizer [31].

4.4. Hyperparameter Tuning

The amount of dropout d and number of gradient steps g is tuned based on the average meta-validation test set binary cross entropy loss from taking g gradient steps and applying d dropout on the meta-validation train set. Once the optimal hyperparameters are determined, we evaluate the model by taking g gradient steps and applying d dropout on the meta-test train set and measuring WGD classification AUC on the meta-test test set for each cancer type in the meta-test set. In all experiments, the baseline CNN classifier performed optimally with five gradient steps and 0% dropout, and the MAML classifier performed optimally with 20 gradient steps and 25% dropout.

4.5. Experiments

4.5.1 Cancer Types

To assess the utility of meta-learning for generalizing few-shot WGD classification across cancer types, we applied the meta-training approach by treating WGD classification

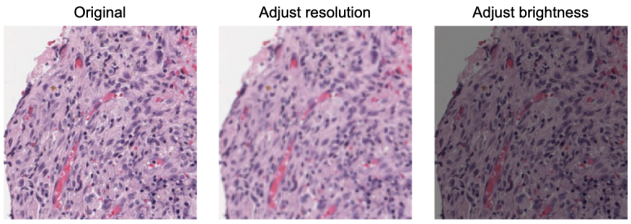

Figure 5. Given the original unperturbed images (**left**), we assess the ability of the meta-learning training regime to generalize few-shot classification across cancer types when faced with two batch effects: lower resolution (**center**) and lower brightness (**right**).

for each cancer type in the meta-train set as separate tasks. During meta-training, the model is encouraged to learn to learn WGD classification for different cancer types that exhibit different tissue morphology, with the intent of extracting WGD-specific signal. The meta-test test set classification performance of the MAML classifier trained under the meta-learning regime is subsequently compared to that of the baseline CNN classifier trained via standard fine-tuning of a pre-trained deep neural network (see Section 4.1).

4.5.2 Batch Effects

To assess the utility of meta-learning for generalizing few-shot WGD classification across batches, we applied two transformations to the images in the meta-test set, reflective of real-world technical variations in image capture:

1. **Resolution.** Reduction of the effective pixel width and height by 50% to mimic a systematic distributional shift to a lower-resolution input distribution.

2. **Brightness.** Reduction of the pixel intensity by 50% to mimic a systematic distributional shift to a dimmer input distribution.

Analogous to the experiment described in Section 4.5.1, we applied the meta-training approach by treating WGD classification for each cancer type in the meta-train set as separate tasks. During meta-test, however, we assess the classification performance of the MAML classifier trained under the meta-learning regime and the CNN classifier on the *batch-adjusted* images from the meta-test set.

5. Results

5.1. Cancer Types

Following model training to minimize binary cross-entropy loss, we evaluate model performance based on the meta-test test set AUCs, which compares the prediction scores outputted by the model to the ground truth WGD labels. Table 1 depicts the classification performance of the seven cancer types in the meta-test set. On average, the baseline CNN classifier achieves an AUC of 0.6888, ranging from AUC of 0.6611 to 0.7000. In contrast, the MAML classifier achieves an AUC of 0.6944, ranging from AUC of 0.6843 to 0.7022, achieving better performance on average than the CNN classifier on five of the meta-test cancer types. Notably, the MAML approach outperforms the baseline CNN approach on four out of the five rare cancer types. Taken together, the MAML approach outperforms the baseline CNN approach on the meta-test set (Wilcoxon signed-rank one-sided p-value=0.0411).

	CNN	MAML
ACC	0.6873 ± 0.0540	**0.6988 ± 0.0581**
CHOL	**0.6890 ± 0.0532**	0.6845 ± 0.0643
KICH	0.6928 ± 0.0312	**0.7022 ± 0.0303**
KIRC	0.6611 ± 0.0609	**0.6843 ± 0.1018**
OV	0.6950 ± 0.0393	**0.7020 ± 0.0435**
UCEC	**0.7000 ± 0.0602**	0.6859 ± 0.0816
UCS	0.6846 ± 0.0387	**0.6908 ± 0.0667**
META-TEST	0.6888 ± 0.0506	**0.6944 ± 0.0773**

Table 1. Results comparing the WGD classification AUC under a baseline standard (CNN) or meta-learning (MAML) training regime. Results are shown for the held-out meta-test set, by cancer type and combined for the entire meta-test set, from 40 random initializations (average ± 1 standard deviation).

	CNN	MAML
ACC	0.6316 ± 0.1011	**0.6930 ± 0.0689**
CHOL	**0.6733 ± 0.0570**	0.6730 ± 0.0548
KICH	**0.7081 ± 0.0261**	0.6941 ± 0.0285
KIRC	0.6269 ± 0.0805	**0.6401 ± 0.0855**
OV	0.7097 ± 0.0728	**0.7136 ± 0.0751**
UCEC	0.6794 ± 0.0496	**0.6877 ± 0.0569**
UCS	**0.6697 ± 0.0330**	0.6649 ± 0.0886
META-TEST	0.6713 ± 0.0716	**0.6809 ± 0.0717**

Table 2. Results comparing the WGD classification AUC under a baseline standard (CNN) or meta-learning (MAML) training regime. Results are shown for the resolution-adjusted held-out meta-test set, by cancer type and combined for the entire resolution-adjusted meta-test set, from 40 random initializations (average ± 1 standard deviation).

5.2. Batch Effects

5.2.1 Image Resolution

Table 2 depicts the classification performance of the resolution-adjusted seven cancer types in the meta-test set. On average, the baseline CNN classifier achieves an AUC of 0.6713, ranging from AUC of 0.6269 to 0.7097. In contrast, the MAML classifier achieves an AUC of 0.6809, ranging from AUC of 0.6401 to 0.7136, achieving better performance on average than the CNN classifier on

	CNN	MAML
ACC	0.6742 ± 0.0801	$\mathbf{0.7062 \pm 0.0655}$
CHOL	0.6819 ± 0.0475	$\mathbf{0.7147 \pm 0.0388}$
KICH	$\mathbf{0.7101 \pm 0.0216}$	0.7079 ± 0.0335
KIRC	0.6670 ± 0.0795	$\mathbf{0.7060 \pm 0.0825}$
OV	0.6989 ± 0.0774	$\mathbf{0.7000 \pm 0.0952}$
UCEC	0.6899 ± 0.0312	$\mathbf{0.6961 \pm 0.0379}$
UCS	$\mathbf{0.6900 \pm 0.0495}$	0.6837 ± 0.0874
META-TEST	0.6884 ± 0.0620	$\mathbf{0.6973 \pm 0.0704}$

Table 3. Results comparing the WGD classification AUC under a baseline standard (CNN) or meta-learning (MAML) training regime. Results are shown for the brightness-adjusted held-out meta-test set, by cancer type and combined for the entire brightness-adjusted meta-test set, from 40 random initializations (average $\pm$ 1 standard deviation).

four of the meta-test cancer types. Taken together, the MAML approach outperforms the baseline CNN approach on the meta-test set (Wilcoxon signed-rank one-sided p-value=0.0312). Due to the coarse-grained nature of the resolution-adjusted images and associated loss of pixel information, the MAML classifier learns better on the original unperturbed meta-test set images than the resolution-adjusted meta-test set images (Wilcoxon signed-rank one-sided p-value=0.0254). This result is consistent with our feature extraction analysis to identify the optimal magnification, which showed superior performance on 10x magnification images compared to 5x magnification images.

5.2.2 Image Brightness

Table 3 depicts the classification performance of the brightness-adjusted seven cancer types in the meta-test set. On average, the baseline CNN classifier achieves an AUC of 0.6884, ranging from AUC of 0.6670 to 0.7101. In contrast, the MAML classifier achieves an AUC of 0.6973, ranging from AUC of 0.6837 to 0.7147, achieving better performance on average than the CNN classifier on five of the meta-test cancer types. Notably, the MAML approach outperforms the baseline CNN approach on three out of the five rare cancer types. Taken together, the MAML approach outperforms the baseline CNN approach on the meta-test set (Wilcoxon signed-rank one-sided p-value=0.0370). Moreover, the MAML classifier is able to learn equally well on the original unperturbed meta-test set images and the brightness-adjusted meta-test set images, with no significant difference in performance (Wilcoxon signed-rank two-sided p-value=0.9967).

6. Discussion

In this work, we demonstrate that machine learning enables signal extraction from medical imaging data mired in tissue site-specific idiosyncrasies. Unlabeled data is often abundant in healthcare settings because label acquisition is expensive. The meta-learning training regime enables fast learning with only a handful of examples. In the case of WGD classification, the MAML classifier outperforms the baseline CNN classifier on the meta-test set when it is trained on only eight training examples per cancer type.

In addition, we introduce two batch effects into our dataset in order to further study the utility of the meta-learning approach. It is generally difficult to correct for batch effects with deep neural network classifiers because the interactions between variables entangles the batch effect with the effect of interest. However, the meta-learning approach is able to learn despite systematically-imposed differences between the meta-train and meta-test sets. For every image in the meta-test set, we perturb the image by (i) reducing the brightness by 50%, or (ii) reducing the resolution by 50%. In both cases, the MAML classifier outperforms the baseline CNN classifier on WGD classification. Furthermore, the MAML classifier's performance is comparable between that of the original unperturbed images and that of the lower-brightness images.

Ultimately, accounting for variations *between* cancer types is made possible by fast learning on only a handful of labeled images, which was successfully extended to accounting for technical variations *within* cancer types. From a clinical perspective, the ability to accurately and cost-effectively stratify patients enables a more fine-grained study of and tailored approach to treatment. From a machine learning perspective, fast adaptation to new tasks is key to mitigating heterogeneity in high-dimensional data that is nonspecific to the signal of interest.

As natural extensions of this work, we will expand this analysis to include all 33 cancer types from TCGA and multiple labels beyond WGD status. For instance, the pan-cancer detection of WGD from histopathology images suggests the possibility of detecting other genomic aberrations, such as DNA mismatch repair deficiency (MMRd). Genomic aberrations are increasingly used to predict response to cancer therapies, such as immune checkpoint blockade (ICB). While MMRd is linked to ICB response, it is not understood at a morphological level [69]. Given MMRd status shows negative correlation with WGD status, we hypothesize that complex tissue-level features associated with MMRd may be orthogonal to those associated with WGD [1]. In an effort to advance morphology-guided treatment decisions, we envision applying meta-learning to generalize MMRd status and WGD status classification across cancer types from histopathology images.

Finally, we hope to devise methods that can learn from multiple slices and magnifications, as higher magnifications may capture intra-cellular patterns, while lower magnifications may capture inter-cellular patterns [36]. By facilitat-

ing a more complete picture of the tumor, we envision these technologies can be seamlessly integrated into the clinic for real-time histology assessment and decision support [8].

Acknowledgments

The authors thank Marc Schwartz, Jonathan Chen, Max Jan, and Victor Quach for helpful discussions, and the anonymous reviewers for their valuable feedback. This work was supported by the Simons Foundation and the Google Cloud Research Credits Program.

Software and Data

The code to reproduce all results is publicly available: https://github.com/chsher/CAML. TCGA histopathology images can be accessed via the NCI GDC Data Portal: https://portal.gdc.cancer.gov. WGD labels can be obtained from the Supplementary Materials of Taylor et al. [56].

References

[1] Craig M. Bielski, Ahmet Zehir, Alexander V. Penson, Mark T. A. Donoghue, Walid Chatila, Joshua Armenia, Matthew T. Chang, Alison M. Schram, Philip Jonsson, Chaitanya Bandlamudi, Pedram Razavi, Gopa Iyer, Mark E. Robson, Zsofia K. Stadler, Nikolaus Schultz, Jose Baselga, David B. Solit, David M. Hyman, Michael F. Berger, and Barry S. Taylor. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature Genetics*, 50(8): 1189–1195, 2018. doi: 10.1038/s41588-018-0165-1. 1, 7

[2] Wouter Bulten, Maschenka Balkenhol, Jean-Joël Awoumou Belinga, Américo Brilhante, Aslı Çakır, Lars Egevad, Martin Eklund, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, et al. Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists. *Modern Pathology*, pages 1–12, 2020. 3

[3] Erik A Burlingame, Mary McDonnell, Geoffrey F Schau, Guillaume Thibault, Christian Lanciault, Terry Morgan, Brett E Johnson, Christopher Corless, Joe W Gray, and Young Hwan Chang. Shift: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. *Scientific reports*, 10(1):1–14, 2020. 3

[4] Dmitrii Bychkov, Nina Linder, Aleksei Tiulpin, Hakan Kücükel, Mikael Lundin, Stig Nordling, Harri Sihto, Jorma Isola, Tiina Lehtimäki, Pirkko-Liisa Kellokumpu-Lehtinen, et al. Deep learning identifies morphological features in breast cancer predictive of cancer erbb2 status and trastuzumab treatment efficacy. *Scientific reports*, 11(1):1–10, 2021. 3

[5] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 2

[6] Rui Cao, Fan Yang, Si-Cong Ma, Li Liu, Yu Zhao, Yan Li, De-Hua Wu, Tongxin Wang, Wei-Jia Lu, Wei-Jing Cai, et al. Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer. *Theranostics*, 10(24):11080, 2020. 3

[7] Chi-Long Chen, Chi-Chung Chen, Wei-Hsiang Yu, Szu-Hua Chen, Yu-Chan Chang, Tai-I Hsu, Michael Hsiao, Chao-Yuan Yeh, and Cheng-Yu Chen. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nature communications*, 12(1):1–13, 2021. 2

[8] Po-Hsuan Cameron Chen, Krishna Gadepalli, Robert MacDonald, Yun Liu, Shiro Kadowaki, Kunal Nagpal, Timo Kohlberger, Jeffrey Dean, Greg S Corrado, Jason D Hipp, et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature medicine*, 25(9):1453–1457, 2019. 8

[9] Jun Cheng, Zhi Han, Rohit Mehra, Wei Shao, Michael Cheng, Qianjin Feng, Dong Ni, Kun Huang, Liang Cheng, and Jie Zhang. Computational analysis of pathological images enables a better diagnosis of tfe3 xp11. 2 translocation renal cell carcinoma. *Nature communications*, 11(1):1–9, 2020. 2

[10] Yael Cohen-Sharir, James M McFarland, Mai Abdusamad, Carolyn Marquis, Sara V Bernhard, Mariya Kazachkova, Helen Tang, Marica R Ippolito, Kathrin Laue, Johanna Zerbib, et al. Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. *Nature*, 590(7846):486–491, 2021. 1

[11] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018. doi: 10.1038/s41591-018-0177-5. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[13] Sally M Dewhurst, Nicholas McGranahan, Rebecca A Burrell, Andrew J Rowan, Eva Grönroos, David Endesfelder, Tejal Joshi, Dmitri Mouradov, Peter Gibbs, Robyn L Ward, et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer discovery*, 4(2):175–185, 2014. 1

[14] Amelie Echle, Heike Irmgard Grabsch, Philip Quirke, Piet A van den Brandt, Nicholas P West, Gordon GA Hutchins, Lara R Heij, Xiuxiang Tan, Susan D Richman, Jeremias

Krause, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*, 159(4):1406–1416, 2020. 3

[15] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, pages 1–11, 2020. 3

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017. URL http://arxiv.org/abs/1703.03400. 2

[17] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1 (8):800–810, 2020. 2, 3

[18] Stefanie Groeneveld-Krentz, Michael P Schroeder, Michael Reiter, Malwine J Pogodzinski, Helia J Pimentel-Gutiérrez, Renia Vagkopoulou, Jana Hof, Christiane Chen-Santel, Karin Nebral, Jutta Bradtke, et al. Aneuploidy in children with relapsed b-cell precursor acute lymphoblastic leukaemia: clinical importance of detecting a hypodiploid origin of relapse. *British journal of haematology*, 185(2): 266–283, 2019. 1

[19] Fangjian Han, Li Yu, and Yi Jiang. Computer-aided diagnosis system of lung carcinoma using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 690–691, 2020. 2, 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385. 4

[21] David Joon Ho, Dig VK Yarlagadda, Timothy M DAlfonso, Matthew G Hanna, Anne Grabenstetter, Peter Ntiamoah, Edi Brogi, Lee K Tan, and Thomas J Fuchs. Deep multi-magnification networks for multi-class breast cancer image segmentation. *Computerized Medical Imaging and Graphics*, 88:101866, 2021. 2

[22] Shi Hu, Jakub Tomczak, and Max Welling. Meta-learning for medical image classification. In *1st Conference on Medical Imaging with Deep Learning*, 2018. 3

[23] Osamu Iizuka, Fahdi Kanavati, Kei Kato, Michael Rambeau, Koji Arihiro, and Masayuki Tsuneki. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific reports*, 10(1):1–11, 2020. 2

[24] Shivam Kalra, Hamid R Tizhoosh, Sultaan Shah, Charles Choi, Savvas Damaskinos, Amir Safarpoor, Sobhan Shafiei, Morteza Babaie, Phedias Diamandis, Clinton JV Campbell, et al. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *NPJ digital medicine*, 3(1):1–15, 2020. 2

[25] Fahdi Kanavati, Gouji Toyokawa, Seiya Momosaki, Michael Rambeau, Yuka Kozuma, Fumihiro Shoji, Koji Yamazaki, Sadanori Takeo, Osamu Iizuka, and Masayuki Tsuneki. Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific reports*, 10(1):1–11, 2020. 2

[26] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019. 3

[27] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019. 3

[28] Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8):789–799, 2020. 2, 3

[29] Young-Gon Kim, Sungchul Kim, Cristina Eunbee Cho, In Hye Song, Hee Jin Lee, Soomin Ahn, So Yeon Park, Gyungyub Gong, and Namkug Kim. Effectiveness of transfer learning for enhancing tumor classification with a convolutional neural network on frozen sections. *Scientific Reports*, 10(1):1–9, 2020. 2

[30] Susanne Kimeswenger, Philipp Tschandl, Petar Noack, Markus Hofmarcher, Elisabeth Rumetshofer, Harald Kindermann, Rene Silye, Sepp Hochreiter, Martin Kaltenbrunner, Emmanuella Guenova, et al. Artificial neural networks and pathologists recognize basal cell carcinomas based on different histological patterns. *Modern Pathology*, pages 1–9, 2020. 2

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[32] Andreas Kleppe, Ole-Johan Skrede, Sepp De Raedt, Knut Liestøl, David J Kerr, and Håvard E Danielsen. Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer*, pages 1–13, 2021. 2

[33] Anastasia Y Kuznetsova, Katarzyna Seget, Giuliana K Moeller, Mirjam S de Pagter, Jeroen ADM de Roos, Milena Dürrbaum, Christian Kuffer, Stefan Müller, Guido JR Zaman, Wigard P Kloosterman, et al. Chromosomal instability,

tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. *Cell cycle*, 14 (17):2810–2820, 2015. 1

[34] Alona Levy-Jurgenson, Xavier Tekpli, Vessela N Kristensen, and Zohar Yakhini. Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Scientific reports*, 10(1): 1–11, 2020. 3

[35] Yonghong Li, Lori A Anderson, Edward I Ginns, and James J Devlin. Cost effectiveness of karyotyping, chromosomal microarray analysis, and targeted next-generation sequencing of patients with unexplained global developmental delay or intellectual disability. *Molecular diagnosis & therapy*, 22(1): 129–138, 2018. 1

[36] Jonathan TC Liu, Adam K Glaser, Kaustav Bera, Lawrence D True, Nicholas P Reder, Kevin W Eliceiri, and Anant Madabhushi. Harnessing non-destructive 3d pathology. *Nature biomedical engineering*, pages 1–16, 2021. 7

[37] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, pages 1–16, 2021. 2

[38] Wenqi Lu, Simon Graham, Mohsin Bilal, Nasir Rajpoot, and Fayyaz Minhas. Capturing cellular topology in multi-gigapixel pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 260–261, 2020. 3

[39] Nikhil Naik, Ali Madani, Andre Esteva, Nitish Shirish Keskar, Michael F Press, Daniel Ruderman, David B Agus, and Richard Socher. Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. *Nature communications*, 11(1):1–8, 2020. 3

[40] Javad Noorbakhsh, Saman Farahmand, Sandeep Namburi, Dennis Caruana, David Rimm, Mohammad Soltanieh-ha, Kourosh Zarringhalam, Jeffrey H Chuang, et al. Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images. *Nature communications*, 11(1):1–14, 2020. 2

[41] Iván Olier, Noureddin Sadawi, G. Richard J. Bickerton, Joaquin Vanschoren, Crina Grosan, Larisa N. Soldatova, and Ross D. King. Meta-qsar: a large-scale application of meta-learning to drug design and discovery. *CoRR*, abs/1709.03854, 2017. URL http://arxiv.org/abs/1709.03854. 3

[42] Liron Pantanowitz, Gabriela M Quiroga-Garza, Lilach Bien, Ronen Heled, Daphna Laifenfeld, Chaim Linhart, Judith Sandbank, Anat Albrecht Shach, Varda Shalev, Manuela Vecsler, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *The Lancet Digital Health*, 2(8):e407–e416, 2020. 2

[43] Jeonghyuk Park, Yul Ri Chung, Seo Taek Kong, Yeong Won Kim, Hyunho Park, Kyungdoc Kim, Dong-Il Kim, and Kyu-Hwan Jung. Aggregation of cohorts for histopathological diagnosis with deep morphological analysis. *Scientific reports*, 11(1):1–11, 2021. 3

[44] Ryan J Quinton, Amanda DiDomizio, Marc A Vittoria, Kristỳna Kotỳnková, Carlos J Ticas, Sheena Patel, Yusuke Koga, Jasmine Vakhshoorzadeh, Nicole Hermance, Taruho S Kuroda, et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature*, pages 1–6, 2021. 1

[45] Patricia Raciti, Jillian Sue, Rodrigo Ceballos, Ran Godrich, Jeremy D Kunz, Supriya Kapur, Victor Reuter, Leo Grady, Christopher Kanan, David S Klimstra, et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Modern Pathology*, 33(10):2058–2066, 2020. 2

[46] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, page 102032, 2021. 2

[47] Frederic Runge, Danny Stoll, Stefan Falkner, and Frank Hutter. Learning to design rna. In *Proceedings of the 36th International Conference on Learning Representations*, 2019. 3

[48] Ram W Sabnis. Novel kif18a inhibitors for treating cancer, 2020. 1

[49] Charlie Saillard, Benoit Schmauch, Oumeima Laifa, Matahi Moarii, Sylvain Toldo, Mikhail Zaslavskiy, Elodie Pronier, Alexis Laurent, Giuliana Amaddeo, Hélène Regnault, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology*, 72 (6):2000–2013, 2020. 3

[50] Akira Saito, Hidenori Toyoda, Masaharu Kobayashi, Yoshinori Koiwa, Hiroki Fujii, Koji Fujita, Atsuyuki Maeda, Yuji Kaneoka, Shoichi Hazama, Hiroaki Nagano, et al. Prediction of early recurrence of hepatocellular carcinoma after resection using digital pathology images assessed by machine learning. *Modern Pathology*, 34(2):417–425, 2021. 3

[51] Katharina Schwarze, James Buchanan, Jilles M Fermont, Helene Dreau, Mark W Tilley, John M Taylor, Pavlos Antoniou, Samantha JL Knight, Carme Camps, Melissa M Pentony, et al. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the united kingdom. *Genetics in Medicine*, 22 (1):85–94, 2020. 1

[52] Yunlang She, Zhuochen Jin, Junqi Wu, Jiajun Deng, Lei Zhang, Hang Su, Gening Jiang, Haipeng Liu, Dong Xie, Nan Cao, et al. Development and validation of a deep learning model for non–small cell lung cancer survival. *JAMA network open*, 3(6):e205842–e205842, 2020. 3

[53] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017. URL http://arxiv.org/abs/1703.05175. 2

[54] Zhigang Song, Shuangmei Zou, Weixun Zhou, Yong Huang, Liwei Shao, Jing Yuan, Xiangnan Gou, Wei Jin, Zhanbo Wang, Xin Chen, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nature communications*, 11(1):1–9, 2020. 2

[55] Lova Sun, Melina E Marmarelis, and Corey J Langer. Systemic therapy for mutation-driven nsclc. In *Seminars in Radiation Oncology*, volume 31, pages 140–148. Elsevier, 2021. 1

[56] Alison M Taylor, Juliann Shih, Gavin Ha, Galen F Gao, Xiaoyang Zhang, Ashton C Berger, Steven E Schumacher, Chen Wang, Hai Hu, Jianfang Liu, et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer cell*, 33(4):676–689, 2018. 3, 8

[57] Yuri Tolkach, Tilmann Dohmgörgen, Marieta Toma, and Glen Kristiansen. High-accuracy prostate cancer pathology using deep learning. *Nature Machine Intelligence*, 2(7):411–418, 2020. 2

[58] Anand Vasudevan, Klaske M Schukken, Erin L Sausville, Vishruth Girish, Oluwadamilare A Adebambo, and Jason M Sheltzer. Aneuploidy as a promoter and suppressor of malignant growth. *Nature Reviews Cancer*, 21(2):89–103, 2021. 1

[59] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *CoRR*, abs/1606.04080, 2016. URL http://arxiv.org/abs/1606.04080. 2

[60] Xiyue Wang, Yuqi Fang, Sen Yang, Delong Zhu, Minghui Wang, Jing Zhang, Kai-yu Tong, and Xiao Han. A hybrid network for automatic hepatocellular carcinoma segmentation in h&e-stained whole slide images. *Medical Image Analysis*, 68:101914, 2021. 2

[61] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992. 5

[62] Ann-Christin Woerl, Markus Eckstein, Josephine Geiger, Daniel C Wagner, Tamas Daher, Philipp Stenzel, Aurélie Fernandez, Arndt Hartmann, Michael Wand, Wilfried Roth, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *European urology*, 78(2):256–264, 2020. 2

[63] John Xie, Adeem Nachabe, Lindsey J Hathaway, Bachir Farah, Bachir Berbari, Yuwen Li, Theresa C Brown, Janet L Schmid, Francisco Socola, Nakhle S Saba, et al. The prognostic implications of tetraploidy/near-tetraploidy in acute myeloid leukemia: a case series and systematic review of the literature. *Leukemia & Lymphoma*, 62(1):203–210, 2021. 1

[64] Hongming Xu, Sunho Park, Jean René Clemenceau, Nathan Radakovich, Sung Hak Lee, and Tae Hyun Hwang. Deep learning approach to predict tumor mutation burden (tmb) and delineate its spatial heterogeneity from whole slide images. *bioRxiv*, page 554527, 2020. 3

[65] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65: 101789, 2020. 3

[66] Mark Yarchoan, Alexander Hopkins, and Elizabeth M Jaffee. Tumor mutational burden and response rate to pd-1 inhibition. *The New England journal of medicine*, 377(25):2500, 2017. 1

[67] Chih-Hsiang Yu, Tze-Kang Lin, Shiann-Tarng Jou, Chien-Yu Lin, Kai-Hsin Lin, Meng-Yao Lu, Shu-Huey Chen, Chao-Neng Cheng, Kang-Hsi Wu, Shih-Chung Wang, et al. Mlpa and dna index improve the molecular diagnosis of childhood b-cell acute lymphoblastic leukemia. *Scientific reports*, 10 (1):1–11, 2020. 1

[68] Xingzhi Yue, Neofytos Dimitriou, and Ognjen Arandjelovic. Colorectal cancer outcome prediction from h&e whole slide images using machine learning and automatically inferred phenotype profiles. *NA*, 2019. URL https://arxiv.org/abs/1902.03582. 3

[69] Pengfei Zhao, Li Li, Xiaoyue Jiang, and Qin Li. Mismatch repair deficiency/microsatellite instability-high as a predictor for anti-pd-1/pd-l1 immunotherapy efficacy. *Journal of hematology & oncology*, 12(1):1–14, 2019. 7

[70] Changjiang Zhou, Yi Jin, Yuzong Chen, Shan Huang, Rengpeng Huang, Yuhong Wang, Youcai Zhao, Yao Chen, Lingchuan Guo, and Jun Liao. Histopathology classification and localization of colorectal cancer using global labels by weakly supervised deep learning. *Computerized Medical Imaging and Graphics*, 88:101861, 2021. 2