DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Yoni Schirris\textsuperscript{a,b}, Efstratios Gavves\textsuperscript{b,d}, Iris Nederlof\textsuperscript{a}, Hugo Mark Horlings\textsuperscript{a}, Jonas Teuwen\textsuperscript{a,b,c,*}

\textsuperscript{a}Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands
\textsuperscript{b}University of Amsterdam, Science Park 402, 1098 XH Amsterdam, the Netherlands
\textsuperscript{c}Radboud University Medical Center, Department of Medical Imaging, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands
\textsuperscript{d}Ellogon AI B.V., The Netherlands

Abstract

We propose a Deep learning-based weak label learning method for analysing whole slide images (WSIs) of Hematoxylin and Eosin (H&E) stained tumor cells not requiring pixel-level or tile-level annotations using Self-supervised pre-training and heterogeneity-aware deep Multiple Instance Learning (DeepSMILE). We apply DeepSMILE to the task of Homologous recombination deficiency (HRD) and microsatellite instability (MSI) prediction. We utilize contrastive self-supervised learning to pre-train a feature extractor on histopathology tiles of cancer tissue. Additionally, we use variability-aware deep instance learning to learn the tile feature aggregation function while modeling tumor heterogeneity. Compared to state-of-the-art genomic label classification methods, DeepSMILE improves classification performance for HRD from 70.43 ± 4.10% to 83.79 ± 1.25% AUC and MSI from 78.56 ± 6.24% to 90.32 ± 3.58% AUC in a multi-center breast and colorectal cancer dataset, respectively. These improvements suggest we can improve genomic label classification performance without collecting larger datasets. In the future, this may reduce the need for expensive genome sequencing techniques, provide personalized therapy recommendations based on widely available WSIs of cancer tissue, and improve patient care with quicker treatment decisions - also in medical centers without access to genome sequencing resources.

Keywords:
Self-supervised learning, Multiple Instance Learning, Computational pathology, Histogenomics

1. Introduction

Early recognition of abnormalities in the DNA Damage Response (DDR) machinery in tumors can greatly support personalized medicine by identifying patients that may benefit from therapies that exploit DDR-related genomic alterations (van der Velden et al. (2019); Pearl et al. (2015); Pilié et al. (2019)). For example, Homologous Recombination Deficiency (HRD) and MicroSatellite Instability (MSI) can function as a biomarker indicating therapy sensitization in, respectively, breast cancer (Lord and Ashworth (2016)) and colorectal cancer (Kather et al. (2018); Mauri et al. (2020); Arena et al. (2020)) patients.

Currently, the most common techniques for determining HRD or MSI are next-generation whole genome and exome sequencing methods (Davies et al. (2017); Zhu et al. (2018)), targeted DNA sequencing methods using Polymerase Chain Reaction (Boland et al. (1998)), and immunohistochemistry methods (Kawakami et al. (1998)).
These techniques, however, are expensive, time-consuming, and not all globally accessible (Snowsill et al., 2017). Additionally, the molecular features that indicate functional deficiencies in the homologous recombination pathway, thus indicating targeted therapy sensitivity, are inconclusive and debated (Davies et al., 2017). Therefore, these techniques are not routinely applied in the clinic.

Generally, the golden standard for solid tumor diagnostics is the use of Hematoxylin & Eosin-stained (H&E) Whole-Slide Images (WSIs). In contrast to genome sequencing methods, H&E WSIs are easily accessible, inexpensive, and reflect the cellular and tissue morphology that result from genomic alterations. However, the morphology of the broad range of DDR defects has not yet been described. Therefore, H&E WSIs are, at present, not used to detect DDR defects in routine clinical diagnostics. This is a missed opportunity as employing H&E WSIs to detect DDR defects could assist personalized medicine by guiding early patient stratification for additional diagnostic tests or guiding therapy decisions. The digitization of WSIs has opened up doors for computational analysis to perform this task.

Recent work has shown great promise for deep learning methods for the computational analysis of digitized H&E WSIs for genomic status classification (Echle et al., 2020; Kather et al., 2020; Fu et al., 2020; Coudray et al., 2018). Although end-to-end supervised methods for a Convolutional Neural Network (CNN) on gigapixel WSIs exist, they require specialized implementations to circumvent the large memory requirements required for loss backpropagation with gigapixel images, and this leads to a low training and inference speed (Pinckaers et al., 2020; Chen et al., 2021). In contrast, two-stage methods stop the gradient at the tile feature extraction or tile aggregation level. Generally, these methods do not pass the entire WSI through a neural network but instead split the WSI into many small tiles which are the input to the network. Either the task is framed as tile-supervision, i.e., supervised learning from each tile to the WSI-level label after which the proportion of positively predicted tiles is said to be the WSI-level prediction (Kather et al., 2019, 2020; Echle et al., 2020; Fu et al., 2020), or it is framed as WSI-supervision in which the tiles are compressed using a pre-trained feature extractor to perform supervised classification of the WSI directly using all latent feature vectors of its constituting tiles as input. This can be framed as either full WSI-supervision (Tellez et al., 2020; Aswolinskiy et al., 2021) or weak WSI-supervision (Lu et al., 2021).

However, these methods have their limitations. Since the cellular and tissue morphology related to DDR defects are unknown, there are no spatial annotations to guide training. This is why tile-supervision uses the WSI-level label as a tile-level label, which results in a noisy supervisory signal since the signal of the DDR defect is, likely, not present in every tile. Noisy supervision, in turn, leads to large data requirements (Echle et al., 2020). Current WSI compression techniques report that the top performing self-supervised learning (SSL) method is a bidirectional generative adversarial network, which are notoriously difficult to train (Tellez et al., 2019) and are outperformed by supervised pre-training on large annotated datasets (Tellez et al., 2020). The field of SSL has grown quickly lately, however, and claims to close the gap between supervised and unsupervised learning in the natural image domain (Grill et al., 2020; Chen et al., 2020). In practice, though, most deep learning methods in histopathology do not use a domain-specific feature extractor altogether by employing an ImageNet pre-trained feature extractor (Coudray et al., 2018; Kather et al., 2019, 2020; Fu et al., 2020; Lu et al., 2021). Since natural scenes and medical images have strongly different data distributions, using an ImageNet pre-trained network might not be optimal (Ke et al., 2021; Raghu et al., 2019), and thus employing the latest SSL methods on unlabeled domain-specific data is a promising avenue to increase performance, generalizability, and robustness (Hendrycks et al., 2019).

In this paper, we investigate the effectiveness of SimCLR (Chen et al., 2020), which makes latent representations of heavily augmented versions of the same tile similar while making the latent representations of heavily augmented versions of different tiles distinct. SimCLR is used to pre-train a feature extractor without any spatial annotations or WSI-level labels. This feature extractor is evaluated and compared to the commonly used ImageNet pre-trained feature extractor on the downstream task of HRD and MSI classification. We research its relative effectiveness when finetuning the feature extractor with noisy WSI label tile-supervision, and in the case when we freeze the feature extractor to perform Attention-Based
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Deep Multiple Instance Learning (DeepMIL, [Ilse et al. (2018)]) on the tiles of a WSI. Finally, we propose a feature variability-aware variant of DeepMIL, which models tumor heterogeneity and further increases performance.

1.1. Contributions

The main contributions of this work can be summarized as follows: 1) We outperform existing tile-supervised WSI-label learning methods for HRD and MSI classification for a breast cancer and colorectal cancer dataset, respectively, using DeepMIL on top of tile latent features extracted by a feature extractor that is pre-trained on histopathology tiles using SimCLR. 2) We introduce VarMIL, which extends DeepMIL with a feature variability module to model tumor heterogeneity. VarMIL outperforms DeepMIL for HRD and MSI classification in a breast and colorectal cancer dataset, respectively. 3) We combine the above modules and propose a Deep learning-based weak label learning method for histopathology not requiring pixel-level or tile-level annotations using Self-supervised pre-training and heterogeneity-aware deep Multiple Instance Learning (DeepSMILE).

The paper is structured as follows: Section 2 provides an overview of previous work and positions our work within it. Section 3 describes the data used, followed by section 4 describing our model details. Section 5 describes all experiments. We conclude with the discussion and conclusion in section 6.

2. Related work

We distinguish between two-stage (not end-to-end) and one-stage (end-to-end) methods. DeepSMILE falls in the former category, and we position our work with respect to other two-stage methods in Table 1.

Two-stage methods generally cut the WSI into many small cropped patches, the tiles, which are processed to classify the WSI. In two-stage methods, one can distinguish between tile-supervised WSI-label methods, fully supervised, and weakly supervised methods.

2.1. Two-stage: tile-supervised WSI-label methods

Two-stage tile-supervised WSI-label methods assign the WSI-level label to each tile of a region of interest of a WSI and treat the tile classification as a supervised learning task. Most commonly, the patient-level label is computed as the fraction of positively classified tiles. This method has been applied to gene mutation classification in lung cancer ([Coudray et al. (2018)]), MSI status classification in colorectal and stomach cancer ([Kather et al. (2019); Echle et al. (2020)]), and a large variety of genomic, transcriptomic, and survival labels in pan-cancer tissue ([Kather et al. (2020); Fu et al. (2020); Echle et al. (2020)].

2.2. Two-stage: fully supervised methods

Two-stage fully supervised methods reduce the dimensionality of the WSI by replacing each tile in-place with its latent feature representation as encoded by a pre-trained feature extractor. The resulting compressed WSIs are subsequently used as input for fully supervised methods. Several methods can be employed to pre-train the feature extractor. Self-supervised methods ([Tellez et al., 2019]) and multi-task supervised pre-training on other tasks and data ([Tellez et al., 2020] [Aswolinskiy et al., 2021]) have been described for this task in the literature.

2.3. Two-stage: weakly supervised methods

In contrast to fully supervised methods, two-stage weakly supervised methods use a MIL approach. In a MIL approach, it is assumed that some of the unlabeled tiles (or instances) in the labeled WSI (or bag) contain the signal for the WSI-level label. These methods propose a linear or non-linear combination of the latent features or predicted scores of a selection of the tiles in a WSI to represent the WSI latent features. This WSI latent feature is then used to classify the WSI, so that a WSI-level loss can be computed and backpropagated through the classification network.

[Durand et al. (2016)] compute a score for each instance and use the top and bottom K scores to compute the bag-level prediction. [Courtiol et al. (2019)] extend the work by [Durand et al. (2016)] to tissue segmentation and WSI-level label classification using an MLP on the top and bottom instance scores. As in our work, [Ilse et al. (2018)] learn the bag-level classification by computing an attention-weighted mean of all tile feature vectors. This latent representation of the WSI is fed into an MLP to compute the bag label. [Lu et al. (2019)] used DeepMIL and showed
that pre-training a feature extractor with contrastive predictive coding (Oord et al., 2019) on histopathology tiles improves downstream malignancy classification performance when compared to an ImageNet pre-trained feature extractor, yet this has not been evaluated on full WSIs. Finally, Campanella et al. (2019) produce tile-level feature vectors and class scores using a network trained with a max-pooling MIL approach. The top 20 scored feature vectors are passed to an RNN to produce the WSI-level label to detect metastasis. Similarly, Valieris et al. (2020) have applied this method to predict HRD and mismatch repair deficiency, the DDR deficiency that is closely related to MSI.

2.4. End-to-end methods: contextless

Xie et al. (2020) cluster the latent features of tiles of a WSI into representative parts and concatenate each part's representative tile. This concatenation is mapped to the WSI-level label by a fully connected layer. For prostate cancer classification, they obtain similar results to the MIL-RNN method by Campanella et al. (2019). Although training is end-to-end, this method only concatenates tile-level features and thus does not model tile interactions to include higher-level spatial context.

2.5. End-to-end methods: context-aware

Pinckaers et al. (2020b) reduce the memory requirements of a CNN by 97% with a streaming CNN, which allows applying the CNN directly to megapixel histopathology images. Pinckaers et al. (2020a) further extended the method to accommodate gigapixel WSIs, applied to prostate cancer detection, reaching similar results to the MIL-RNN method by Campanella et al. (2019). Chen et al. (2021) instead leverage the unified memory mechanism and other GPU optimization techniques to overcome the memory constraints. Although these methods reach state-of-the-art results, streaming CNNs are approximately 16× slower than MIL methods during training and inference on WSIs at low resolution, which scales non-linearly with increasing resolution.

Our proposed method, DeepSMILE, is a two-stage weakly supervised method applied to HRD and MSI classification. Compared to the existing two-stage weakly supervised methods, it introduces the use of contrastive self-supervised learning methods to pre-train a histopathology-specific feature extractor. Furthermore, we introduce an extended version of DeepMIL that models tumor heterogeneity. We describe our model in detail in section 4.

3. Materials and methods

3.1. Data

We use digitized H&E tissue slides of breast (BC) and colorectal (CRC) cancer tissue to classify a tumor's genomic labels. These WSIs are large gigapixel images that can exceed 100,000×100,000 px, whereas the genomic labels are binary representations of complex genomic features derived from DNA sequencing results. This section describes the data collection and pre-processing steps for each dataset, and how the genomic labels are obtained.

3.1.1. Colorectal cancer tissue dataset: TCGA-CRC

We use the pre-processed colorectal tumor tiles extracted from Formalin-Fixed, Paraffin-Embedded (FFPE) WSIs from the Cancer Genome Atlas (TCGA) with accompanying binarized MSI labels from Kather (2019). We use the train-test split as given, and perform a 5-fold patient-level train-validation split. This dataset consists of 155,176 (train: 93,408, test: 61,768) tiles from 360 patients (train: 260, test: 100). The train and test set respectively consist of 39 and 26 MSI, and 221 and 74 MicroSatellite Stable (MSS) patients. The number of tiles per WSI ranges from 6 to 4,064 with an average of 375 tiles per WSI. We refer to this dataset as TCGA-CRC.

3.1.2. Breast cancer tissue dataset: TCGA-BC

We use all FFPE WSIs for BC tissue from TCGA (referred to as TCGA-BC), and retrieve genomic DDR-related labels, including HRD Score, from Knijnenburg et al. (2018). The HRD Score (Timms et al., 2014; Marquard et al., 2015) is a discrete score computed as the sum of the number of subchromosomal regions with allelic imbalance extending to the telomere (Birkbak et al., 2012), the number of chromosomal breaks between adjacent regions of at least 10 megabases (Popova et al., 2012), and the number of regions with a loss of heterozygosity event of intermediate size (Abkevich et al., 2012). Since the relationship of the HRD Score to actual homologous recombination functionality and cellular morphology is not known, we test two binarization
strategies to define the classification task of distinguishing HRD-high (assumed to be HR deficient and thus sensitive to targeted therapy) and HRD-low patients (assumed to be HR proficient and thus not sensitive to targeted therapy).

First, we split the score at the median \( m_{HRD} \), similar to previous work [Kather et al., 2020]. Secondly, we aim to provide a better supervisory signal by splitting the set in tertiles \( t_{HRD} \) and assigning a top-tertile or bottom-tertile label, discarding patients with an HRD score close to the median. That is,

\[
m_{HRD} = \begin{cases} 1, & \text{if } HRD_m > q(0.5; D^{TCGA-BC}) = 21 \\ 0, & \text{otherwise} \end{cases} \tag{1}
\]

where \( HRD_m \) is the HRD score of patient \( m \) and \( q(\alpha; D^d) \) indicates the \( \alpha \)-quantile of the HRD scores of the combined train and validation set of dataset \( d \). For \( t_{HRD} \):

\[
t_{HRD} = \begin{cases} 1, & \text{if } HRD_m \geq q(0.66; D^{TCGA-BC}) = 13 \\ 0, & \text{if } HRD_m \leq q(0.33; D^{TCGA-BC}) = 30 \end{cases} \tag{2}
\]

discard patient otherwise

The train-test splits for TCGA-BC are generated for those 940 patients provided by Knijnenburg et al. [2018] for which genomic labels and FFPE WSIs are available. The splits are patient-level stratified for the following subset of genomic labels: mHRD, core base mutation status, core strand mutation status, any mutation status, any MMR mutation status. We create a held-out test split of 25% (238 patients with 262 WSIs), leaving 75% (702 patients with 742 WSIs) for training and intermediate validation. Of the 75% for training and validation, we create a 5-fold train-validation split, so that 15% of all patients are used for validation and 60% for training.

Since not each patient has an HRD Score computed, the test set finally has 121 positive mHRD WSIs, 134 negative mHRD WSIs, 78 positive tHRD WSIs, and 105 negative tHRD WSIs. The train-validation set has 345 positive mHRD WSIs, 380 negative mHRD WSIs, 253 positive tHRD WSIs, and 244 negative tHRD WSIs.

The WSIs are tiled at a spacing of 1.143 microns per pixel (mpp) with a resolution of 224 × 224, resulting in an edge length of 256 µm. We selected the zoom level best representing this spacing, before downsampling. Background tiles were subsequently excluded. A tile is considered a background tile when more than half of its pixels have a pixel value over 240 and a Sobel gradient above 15 after conversion to grayscale as in Fu et al. [2020]. The number of tiles from TCGA-BC WSIs ranges from 131 to 11,029, with an average of 3,625 tiles per WSI.

When training the classifiers on the TCGA-BC dataset, we subsample 500 tiles randomly from each WSI that has more than 500 tiles to speed up the classification training.
process, similar to [Kather et al. (2020)]. After choosing the subsample, this is fixed for the rest of the training.

3.2. Evaluation

The model performance is evaluated using the area under the receiver operating characteristic curve (AUC), using the mean patient-level prediction scores for patients with multiple WSIs, compared to the genomic binary labels. For each experiment, we report the mean and standard deviation of the AUC scores on the test split for 5 different folds and visualize the curves. The complete experimental pipeline is summarized in Algorithm 1. We use common training, intermediate validation, and final validation steps as presented in Algorithm 1 line 13-28. We select the model with the highest patient-level AUC on the validation set for final validation on the test set.

4. Model

In this section we describe our model, which is a two-stage weakly supervised method, using self-supervised pre-training and feature variability-aware deep multiple instance learning. We will refer to this model as DeepSMILE (from Self-supervised heterogeneity-aware Multiple Instance LEarning). Our model is visualized in Figure 1 and its comparison to existing methods is summarized in Table 1.

We split this section into two parts. In the first part in section 4.1.1 we present how we train an in-domain pathology-specific feature extractor with SimCLR to extract the latent feature vector of each tile. In the subsequent section 4.2.3 we present VarMIL, which comprises the second stage of the model. This is a MIL approach extending DeepMIL, detailed in section 4.2.2, that models the variance of each latent feature of all instances in a bag.

In the experiments (section 5), we will compare DeepSMILE to a tile-supervised WSI-label learning baseline with an ImageNet pre-trained feature extractor as presented in section 4.2.1.

4.1. Feature extraction

4.1.1. Self-supervised pre-training

In the feature extraction stage, we use a CNN to map each tile of a WSI to a lower-dimensional representation.

Algorithm 1: Summary of experimental pipeline. Square brackets contain section number providing details for that step.

```plaintext
1 Download WSIs with matching genomic labels [sec 3.1]
2 Extract tissue-containing tiles from WSI [sec 3.1.2]
3 Split tiles on patient level into train and test set, create k-fold train-validation split from train set [sec 3.1.2]
4 Perform self-supervised pre-training on tiles from train set, save model weights, extract features from training and test tiles and save to disk [sec 4.1.1]
5 if pipeline == "baseline" then
6     model = ImageNet pre-trained CNN with tile-level MLP classifier [sec 4.2.1]
7     data_loader = load tiles as samples, with patient-level label per tile
8 else if pipeline == "DeepSMILE" then
9     model = VarMIL WSI-level classifier [sec 4.2.3]
10    data_loader = load latent features of all tiles of a WSI as samples, with patient-level label per sample
11 end
12 optimizer = ADAM [sec 4]
13 for fold in k_folds do
14    for epoch in epochs do
15    for (x, y, step) in data_loader do
16        pred = model.forward(x)
17        loss = WeightedCE(y, pred) [sec 4]
18        loss.backward()
19        optimizer.step()
20        if step % evaluate_every == 0 then
21            val_pred = model.forward(X_val)
22            AUC = compute_auc(val_pred, Y_val) [sec 3.2]
23            save(AUC, model)
24        end
25    end
26    Pick model with top AUC on val set [sec 3.2]
27    Use best model to evaluate performance on test set
28 end
29 Report k-fold $\mu \pm \sigma$ of patient-level AUC [sec 5]
```
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Schirris et al.

Figure 1: A visual summary of DeepSMILE. Top-left: We extract all non-background tiles from a WSI, and from each tile extract the latent feature vector using a domain-specific feature extractor pre-trained with SimCLR. Bottom-left: During pre-training, SimCLR transforms each image twice with random augmentations, extracts latent features using the backbone of interest, maps each vector to a projection head, and uses these to compute the loss. After training, the projection MLP is discarded and the backbone of interest is used to compress tiles. Top-right: VarMIL computes the attention-weighted mean and variance of all tile latent feature vectors. The concatenated mean and variance are passed to a linear classification layer to compute the WSI-level genomic label classification, and the loss is backpropagated through the right side of the system. Bottom-right: VarMIL computes the attention weight from each latent feature vector, and computes an attention-weighted mean and variance of all tile latent feature vectors of the WSI.

Formally, this neural network $z_{\theta}$ parametrized by $\theta$ maps the $i^{th}$ tile of patient $m$, $X^m_i$, into an $H$-dimensional vector $Z^m_i$:

$$Z^m_i := z_{\theta}(X^m_i)$$  \hspace{1cm} (3)

To find the parameters $\theta$ we use SimCLR, which is a contrastive self-supervised learning method. We thus obtain an in-domain histopathology-specific feature extractor without requiring pixel-level annotations or WSI-level labels. This is particularly relevant when the cellular and tissue morphologies related to the genomic label are unknown, as is the case with HRD and MSI.

The feature extractor $z_{\theta}$ is initialized using He initialization [He et al., 2015]. The loss function, data augmentations and their parameters are identical to those in Chen et al. (2020). We use the normalized temperature-scaled cross-entropy (NT-Xent) loss on projected latent feature vectors of images that are augmented using a random crop, random horizontal flip, color jitter, and a random grayscale transformation. A visual representation of SimCLR is provided in the lower-left frame of Figure 1.

The parameters $\theta$ are optimized using the Adam optimizer with a learning rate of $3 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, with no learning rate schedule or weight decay.

Similar to previous work (Kather et al., 2019, 2020; Echle et al., 2020) we use different networks for TCGA-CRCk and TCGA-BC, where we use Resnet18 (He et al., 2016) and ShufflenetV2 (Ma et al., 2018) respectively.
This design choice was made to be able to directly compare the results. This difference in networks and dataset size leads to different choices in batch size and number of epochs. On TCGA-BC we train $z_0$ for 45 epochs with a batch size of 950, whereas we use 100 epochs and a batch size of 128 for TCGA-CRCk. The training takes approximately three days using four Nvidia Titan RTX cards (24GB VRAM).

4.2. Classification
To predict the genomic label, which is a whole-slide level label, we introduce VarMIL, an extension of DeepMIL which additionally takes the intra-WSI inter-tile variance of extracted tile features into account. We compare with two baseline models: DeepMIL and a tile-supervision baseline with a tile-level majority vote.

For each classification method, we used the weighted cross-entropy loss where the weight is selected to balance the classes.

As our method extends DeepMIL, we first introduce the baseline models in sections 4.2.1 and 4.2.2.

4.2.1. Baseline: tile supervised WSI-label learning
We compare against a tile-supervised method as used in, e.g., [Kather et al.] (2019, 2020) and [Fu et al.] (2020). In these models, the patient-level label is assigned to each tile from the tumor bed, and the classification task is defined to predict the WSI-level label directly from each individual tile. The final WSI-level prediction is defined as a variant of the majority vote, computed as the proportion of positively labeled tiles, that is:

$$p^m = \frac{1}{N} \sum_{i=1}^{I_m} \text{argmax} p^m_i,$$

where $I_m$ is the number of tiles for patient $m$ and $p^m_i$ is the two-dimensional class probability vector of tile $i$ for patient $m$ with the last index belonging to the positive class.

On TCGA-BC we follow the approach of [Kather et al.] (2020) and train a ShufflenetV2 for 4 epochs with a learning rate of $5 \times 10^{-3}$ and a batch size of 512. Whereas, on TCGA-CRCk, we follow the approach of [Kather et al.] (2019) and train a ResNet18 for 50 epochs with a batch size of 256, learning rate of $10^{-6}$ (selected following a hyperparameter search). For both datasets, we use the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.99$, without a learning rate schedule, and evaluate the performance every 100 steps.

The convolutional layers are initialized using either an ImageNet pre-trained network or a self-supervised pre-trained network on histopathology tiles as presented in section 4.1.1. The final fully connected layers are re-initialized using He initialization. In contrast to [Kather et al.] (2020) we fine-tune all layers instead of only the last layers.

4.2.2. Weak label classification: DeepMIL
DeepMIL is a permutation-invariant MIL model that represents the bag-level latent features as the attention-weighted average of instance-level feature vectors. This bag-level representation is subsequently classified by a fully connected layer. DeepMIL is applied to the encoded tile feature vectors from the first stage as described in section 4.1.

The DeepMIL algorithm predicts for each patient $m$ an attention weight vector $d^m := (d^m_1, d^m_2, \ldots, d^m_I)$ where $I_m$ are the number of available tiles for this patient. This is computed as the softmax of the output of a two-layer multilayer perceptron (MLP) with weights $(W_i, b_i)_{i=1,2}$ on top of all tile-level feature vectors $Z^m$:

$$d^m := \text{softmax} (W_2 \tanh (W_1 Z^m + b_1 1_{1 \times I_m}) + b_2),$$

where the dimensionalities of the weights $(W_i, b_i)_{i=1,2}$ depend on the output dimension $H$ of the feature extractor of the first stage and the MLP dimension $v$. In particular $W_1 \in \mathbb{R}^{I \times H}, b_1 \in \mathbb{R}^I$ and $W_2 \in \mathbb{R}^{1 \times v}$ and $b_2 \in \mathbb{R}$. We define $1_{1 \times I_m}$ as $[1 \cdots 1] \in \mathbb{R}^{1 \times I_m}$, meaning that $b_1$ is added tile-wise to each $v$-dimensional column of $W_1 Z^m$.

The output of the MLP is 1-dimensional for each tile, and therefore $d^m \in \mathbb{R}^{I_m}$.

Subsequently the WSI-level representation $\bar{Z}^m \in \mathbb{R}^H$ is computed as the matrix-vector product between the attention vector $d^m$ and the tile feature vector matrix $Z^m \in \mathbb{R}^{H \times I_m}$ from (3):

$$\bar{Z}^m := Z^m d^m = \sum_{i=1}^{I_m} d^m_i Z^m_i,$$

Finally, the two class output probabilities $p^m = (p^m_+, p^m_-)$ are computed by a linear layer with trainable
weights \((W_\ell, b_\ell)\) on top of the WSI-level latent representation \(Z^m\):

\[
p^m = \text{sigmoid} \left( W_\ell Z^m + b_\ell \right), \tag{7}
\]

where \(W_\ell \in \mathbb{R}^{2 \times H}\) and \(b \in \mathbb{R}^2\). As in [Ilse et al. (2018)] we select \(\nu = 128\) and use the Adam optimizer with \(\beta_1 = 0.9\) and \(\beta_2 = 0.99\). The parameters are initialized using the He initialization. The batch size, learning rate, weight decay, and maximum tiles per WSI rate were found using hyperparameter search (see supplementary material). For the learning rate and weight decay, this results in values of \(5 \times 10^{-4}\) and \(10^{-4}\), respectively. On TCGA-BC we train for 16 epochs with a batch size of 16 on a subsample of 500 tiles per WSI and evaluate the performance every 5 iterations. For TCGA-CRCk we train for 16 epochs on all tiles with a batch size of 16, evaluated every 10 iterations.

To allow minibatch training, we pad \(Z^m\) with zero vectors up to a size of 550 when subsampling 500 tiles for TCGA-BC and up to a size of 11,000 for TCGA-CRCk.

4.2.3. Our method: Weak label classification with VarMIL

A limitation of DeepMIL is that the attention-weighted mean is unable to capture tile interactions and global, high-level features. This results in an aggregated feature vector that only represents local, tile-level features. However, global features, such as tumor border shape and intratumor heterogeneity, might be indicative of HRD or MSI. Therefore, we extend DeepMIL with an attention-weighted variance module, termed VarMIL, which computes the variability in features across tiles within a single WSI as a measure of tissue heterogeneity (see supplementary material for a more detailed motivation). In addition to \(\tilde{Z}^m\) from the original DeepMIL framework, we propose to model the feature variability of the tiles by adding a learned attention-weighted variance. The weighted variance for any patient \(m\) is defined as

\[
Z^m_{\sigma} = \frac{I_m}{I_m - 1} \sum_{i=1}^{I_m} \hat{a}^m_i (Z^m_i - \hat{Z}^m)^2 \tag{8}
\]

We then concatenate \(Z^m_\sigma\) and \(\tilde{Z}^m\) into a single vector, that is

\[
\hat{Z}^m := \begin{bmatrix} \tilde{Z}^m \ Z^m_\sigma \end{bmatrix}, \tag{9}
\]

which is the novel WSI-level latent representation that is passed to a linear layer with trainable weights \((W_\varphi, b_\varphi)\):

\[
p^m = \text{sigmoid} \left( W_\varphi \hat{Z}^m + b_\varphi \right), \tag{10}
\]

where \(W_\varphi \in \mathbb{R}^{2 \times \hat{H}}\) and \(b_\varphi \in \mathbb{R}^2\). We use the same hyperparameters and initialization as used for DeepMIL (section 4.2.2).

5. Experiments

5.1. Experiment 1: Self-supervised learning and VarMIL improve HRD Score classification in TCGA-BC

We perform an ablation study to show the added value of each of our proposed components on the downstream task of binarized HRD score classification on the TCGA-BC dataset (for dataset details, see section 3.1.2). More specifically, we present the AUC scores (computed with scikit-learn v0.22.1) and ROC curves for tile-supervised WSI-label learning, DeepMIL, and VarMIL on top of an ImageNet or in-domain SimCLR pre-trained feature extractor.

Table 2 and Figure 2 show that neither of the proposed components separately improve classification performance. We see, though, that a combination of a SimCLR pre-trained feature extractor with DeepMIL or VarMIL significantly improves performance. Compared to the ImageNet tile-supervision baseline, SimCLR-DeepMIL increases performance by 7.95% and 11.68% AUC for mHRD and tHRD, respectively, both with a lower standard deviation across training folds. Compared to SimCLR-DeepMIL, SimCLR-VarMIL further increases the performance by 0.57% and 0.7% AUC for mHRD and tHRD, respectively, with a further decreased variance across folds.

We present the ROC curves for all models for mHRD and tHRD classification performance in Figure 3 and Figure 4. These show that the reported improvements are consistent over the ROC curve.

5.2. Experiment 2: Self-supervised learning and VarMIL improve MSI classification on TCGA-CRCk

We evaluate the baseline and all models with a SimCLR pre-trained extractor initialization on the TCGA-CRCk dataset (for dataset details, see section 3.1.1) to classify
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Table 2: Comparison of DeepSMILE (SimCLR-VarMIL) to the baseline (ImageNet pre-trained tile-supervised WSI-label learning). Top row shows the results of our baseline model, second row shows same method but with a SimCLR pre-trained network. Third (fifth) row shows weak label learning with DeepMIL (VarMIL) on features extracted with an ImageNet pre-trained feature extractor, while the fourth (sixth) row uses a SimCLR pre-trained feature extractor. The top row of the rightmost column reproduces the same results of Kather et al. (2019) (reported as 77 (95% CI, 62–87)) with the same method on the same dataset. The following rows show the added effect of each our our proposed modules, illustrating that VarMIL on SimCLR extracted features outperforms the baseline.

| Extractor initialization | Classification Method | TCGA-BC mHRD | TCGA-BC tHRD | TCGA-CRCk MSI |
|--------------------------|-----------------------|-------------|-------------|--------------|
| ImageNet                 | Logistic Reg          | 64.23 ± 2.75| 70.43 ± 4.10| 78.56 ± 6.24 |
| SimCLR                   | Logistic Reg          | 63.67 ± 1.40| 69.60 ± 7.70| 77.97 ± 4.27 |
| ImageNet                 | DeepMIL               | 63.01 ± 1.13| 71.61 ± 3.24|              |
| SimCLR                   | DeepMIL               | 72.18 ± 2.51| 82.09 ± 1.64| 88.45 ± 3.74 |
| ImageNet                 | VarMIL                | 62.80 ± 1.16| 72.32 ± 1.50|              |
| SimCLR                   | VarMIL                | 72.75 ± 0.93| 83.79 ± 1.25| 90.32 ± 3.58 |

Figure 2: Results of Table 2 visualized with violin plots. Each color represents a model variation from one row from Table 2. Each violin plot visualizes the 5 ROCAUC scores obtained through evaluation on the test set of the color-defined model trained on 5 different train splits of the dataset defined on the x axis. Left: mHRD classification in TCGA-BC with all model variations. Middle: tHRD classification in TCGA-BC with all model variations. Right: MSI classification in TCGA-CRCk with the ImageNet-baseline and all SimCLR model variations.

MSI versus MSS tumors. Since this is a published preprocessed dataset, we can directly compare the results with those from Kather et al. (2019). From Table 2 we see that our ImageNet-initialized tile-supervision baseline achieves an AUC of 78.56 ± 6.24%, similar to the published results in Kather et al. (2019). Similar to Experiment 1 (section 5.1), we see that a SimCLR pre-trained feature extractor does not increase performance for the tile-supervision method. When we use the standard DeepMIL classification framework on top of SimCLR pre-trained extracted features, the mean AUC increases by 9.89% from 78.56 ± 6.24% to 88.45 ± 3.74% compared to the ImageNet tile-supervision baseline. VarMIL further increases the performance to 90.32 ± 3.58%. The ROC curves in Figure 5 show that the improvement is consistent over the entire curve.
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Schirris et al.

6. Discussion and conclusions

We proposed DeepSMILE, which uses a histopathology-specific self-supervised pre-trained feature extractor with VarMIL, a classification network that correctly deals with the weak label and learns an aggregation function over the tiles while modeling intratumor heterogeneity.

First, SimCLR was shown to be effective in the histopathology domain, even when using the transformations and hyperparameters that were optimized for the ImageNet dataset. Only when the tile latent features were extracted with a domain-specific SimCLR pre-trained feature extractor we noticed an improved performance with DeepMIL and VarMIL. This performance improvement was not seen when using DeepMIL or VarMIL on top of tile latent feature vectors extracted with an ImageNet pre-trained feature extractor. The tile-supervision baseline, however, did not benefit from a histopathology-specific feature extractor. The SimCLR-initialized tile-supervision model achieved similar results to the ImageNet-initialized tile-supervision model, indicating that the noisy supervision and tile prediction aggregation limits the maximum achievable model perfor-
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Schirris et al.

More specifically, when using a SimCLR pre-trained feature extractor, the attention-weighted mean of tile feature vectors as WSI-level latent representation is expressive for HRD and MSI classification. For all classification tasks on all datasets, DeepMIL increased AUC by up to 11.66% compared to tile-supervised WSI-label learning.

Finally, an attention-weighted variance of tile feature vectors adds valuable information for HRD and MSI classification. VarMIL increased performance by 0.5% to 2% AUC with a lower variance across training folds. For the MSI classification task, our achieved performance of 90.32% is significantly higher than the performance achieved with ImageNet tile-supervision on the same dataset by either Kather et al. (2019), Kather et al. (2020), or Echle et al. (2020). Additionally, the method presented by Valieris et al. (2020) shows no improvement on MSI classification for the TCGA stomach cancer dataset compared to Kather et al. (2019), while our method improves performance on TCGA-CRCk by 11.76%. Although a perfect comparison can not be made, the results hint at our method performing at least on par, or better, than the one introduced by Valieris et al. (2020). DeepSMILE achieved similar performance on HRD classification on TCGA-BC as Kather et al. (2020), while we did not use tumor bed annotations, stain normalization, or quality assessment of the WSIs.

6.1. Limitations

We acknowledge several limitations to the experimental setup and our proposed method.

Firstly, although the improvements of our proposed modules show similar dynamics in TCGA-CRCk and TCGA-BC, we can not reproduce the HRD classification results in TCGA-BC from Kather et al. (2020) with the ImageNet-initialized tile-supervision baseline. Since we follow a very similar training scheme and obtain the same results with the baseline for MSI classification on the pre-processed dataset published by Kather et al. (2019), the observed difference in TCGA-BC is likely due to different preprocessing steps. In our work, we view the tiles at a resolution of 1.143 mpp instead of 0.5 mpp. Additionally, we do not use tumor bed annotations. However, according to Kather et al. (2020), using only tiles from the tumor bed increases performance by a mere 1% AUC for HRD classification, which does not explain the gap we see. Therefore, we expect that viewing the tiles at a higher resolution would provide more information and could increase performance further. Future work could explicitly investigate the effect of different tile resolutions on downstream task classification performance to find the optimal preprocessing settings for this task.

Similarly, we have not explicitly investigated the added effect of self-supervised learning and VarMIL relative to varying pre-processing steps like tumor annotation or H&E normalization. We expect self-supervised pre-training to reduce the positive effect of stain normalization, and VarMIL to reduce the positive effect of tumor bed annotations. Since we see similar improvements with DeepSMILE in TCGA-BC (no stain normalization, no tumor bed annotation) and TCGA-CRCk (stain normalized, tumor bed annotated), however, our results indicate that the improvements are found independent of stain normalization and tumor bed annotations.

Thirdly, the performance of our proposed method can likely be improved by using wider and deeper feature extractors, pre-training for longer on a larger set of pan-cancer WSIs, and with domain-optimized hyperparameters and data augmentations for self-supervised pre-training in the histopathology domain.

Additionally, although we use an in-domain feature extractor by using self-supervised pre-training, the feature extractor is not further fine-tuned for the task at hand when using DeepMIL or VarMIL. Since we subsample 500 random tiles from the WSI, this would allow for end-to-end training and fine-tuning of the feature extractor, which may further improve performance.

Finally, although VarMIL models inter-tile feature heterogeneity, the model does not take into account feature interactions or spatial information of the tiles. Future research could investigate and compare the effectiveness and computational requirements of Neural Image Compression (Tellez et al. (2019)) and end-to-end context-aware learning (Pinckaers et al. (2020a); Chen et al. (2021)) for genomic label classification to our proposed method. Additionally, it could be interesting to investigate the performance of vision transformers (Dosovitskiy et al., 2020) on top of extracted latent feature vectors of all tiles of a WSI to model tile interactions with low computational complexity.
6.2. Conclusions

We have shown that self-supervised learning and weak label learning methods in computational histopathology can lift the performance of HRD and MSI classification directly from H&E WSIs without the necessity to collect larger datasets. In the future, these methods may reduce the need for expensive genome sequencing techniques, provide personalized therapy recommendations based on widely available H&E WSIs, and improve patient care with quicker treatment decisions - also in medical centers without access to genome sequencing resources.

Acknowledgements

The collaboration project is co-funded by the PPP Allowance made available by Health Holland[^1] Top Sector Life Sciences & Health, to stimulate public-private partnerships.

Conflicts of interest

The authors declare the following competing interests. E.G. is a shareholder of Ellogon.ai. H.M. Horlings received a consultation fee paid to the institute from Roche Diagnostics, outside the scope of the work described here. The other authors report no conflicts of interest.

References

Abkevich, V., Timms, K., Hennessy, B., Potter, J., Carey, M., Meyer, L., Smith-McCune, K., Broaddus, R., Lu, K., Chen, J., et al., 2012. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. British journal of cancer 107, 1776–1782.

Arena, S., Corti, G., Durinikova, E., Montone, M., Reilly, N.M., Russo, M., Lorenzato, A., Arcella, P., Lazzari, L., Rospo, G., et al., 2020. A subset of colorectal cancers with cross-sensitivity to olaparib and oxaliplatin. Clinical Cancer Research 26, 1372–1384.

Aswolinskiy, W., Tellez, D., Raya, G., van der Woude, L., Looijen-Salamon, M., van der Laak, J., Grunberg, K., Ciompi, F., 2021. Neural image compression for non-small cell lung cancer subtype classification in h&e stained whole-slide images, in: Tomaszewski, J.E., Ward, A.D. (Eds.), Medical Imaging 2021: Digital Pathology, SPIE. p. 1. doi:10.1117/12.2581943

Birkbak, N.J., Wang, Z.C., Kim, J.Y., Eklund, A.C., Li, Q., Tian, R., Bowman-Colin, C., Li, Y., Greenecolozzi, A., Iglehart, J.D., et al., 2012. Telomeric allelic imbalance indicates defective dna repair and sensitivity to dna-damaging agents. Cancer discovery 2, 366–375.

Boland, C.R., Thibodeau, S.N., Hamilton, S.R., Sidransky, D., Eshleman, J.R., Burt, R.W., Meltzer, S.J., Rodriguez-Bigas, M.A., Fodde, R., Ranzani, G.N., et al., 1998. A national cancer institute workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer research 58, 5248–5257.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine 25, 1301–1309. doi:10.1038/s41591-019-0508-1

Chen, C.L., Chen, C.C., Yu, W.H., Chen, S.H., Chang, Y.C., Hsu, T.I., Hsiao, M., Yeh, C.Y., Chen, C.Y., 2021. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. Nature Communications 12, 1193. doi:10.1038/s41467-021-21467-y

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR. p. 1597–1607.

Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyő, D., Moreira, A.L., Razavian, N., Tsirigos, A., 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine 24, 1559–1567. doi:10.1038/s41591-018-0177-5

[^1]: https://www.health-holland.com
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Schirris et al.

Courtiol, P., Maussion, C., Moarri, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al., 2019. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature Medicine 25, 1519–1525. doi:10.1038/s41591-019-0583-3

Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A.M.e.a., 2017. Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. Nature Medicine 23, 517–525. doi:10.1038/nm.4292

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Durand, T., Thome, N., Cord, M., 2016. Weldon: Weakly supervised learning of deep convolutional neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 4743–4752. doi:10.1109/CVPR.2016.513

Echle, A., Grabsch, H.I., Quirke, P., van den Brandt, P.A., West, N.P., Hutchins, G.G.A., Heij, L.R., Tan, X., Richman, S.D., Krause, J., et al., 2020. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. Gastroenterology 159, 1406–1416.e11. doi:10.1053/j.gastro.2020.06.021

Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., Gerstung, M., 2020. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer , 1–11doi:10.1038/s43018-020-0085-8

Grill, J.B., Strub, F., Alché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al., 2020. Bootstrap your own latent: A new approach to self-supervised learning. arXiv:2006.07733 [cs, stat] ArXiv: 2006.07733.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE. p. 1026–1034. doi:10.1109/ICCV.2015.123

Hendrycks, D., Mazeika, M., Kadavath, S., Song, D., 2019. Using self-supervised learning can improve model robustness and uncertainty. arXiv preprint arXiv:1906.12340.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning, in: International Conference on Machine Learning, PMLR. p. 2127–2136.

Kather, J.N., 2019. Histological images for msi vs. mss classification in gastrointestinal cancer, tfe samples. URL: https://doi.org/10.5281/zenodo.2530835 doi:10.5281/zenodo.2530835

Kather, J.N., Halama, N., Jaeger, D., 2018. Genomics and emerging biomarkers for immunotherapy of colorectal cancer, in: Seminars in cancer biology, Elsevier. pp. 189–197.

Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A.J., Bankhead, P., et al., 2020. Pan-cancer image-based detection of clinically actionable genetic alterations. Nature Cancer , 1–11doi:10.1038/s43018-020-0087-6

Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al., 2019. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine 25, 1054–1056. doi:10.1038/s41591-019-0462-y

Kawakami, H., Zaanan, A., Sinicrope, F.A., 2015. Microsatellite instability testing and its role in the management of colorectal cancer. Current treatment options in oncology 16, 1–15.
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Schirris et al.

Ke, A., Ellsworth, W., Banerjee, O., Ng, A.Y., Rajpurkar, P., 2021. Chextransfer: performance and parameter efficiency of imagenet models for chest x-ray interpretation, in: Proceedings of the Conference on Health, Inference, and Learning, pp. 116–124.

Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Churnick, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al., 2018. Genomic and molecular landscape of dna damage repair deficiency across the cancer genome atlas. Cell Reports 23, 239–254.e6. doi:10.1016/j.celrep.2018.03.076.

Lord, C.J., Ashworth, A., 2016. Brcaness revisited. Nature Reviews Cancer 16, 110–120. doi:10.1038/nrc.2015.21.

Lu, M.Y., Chen, R.J., Wang, J., Dillon, D., Mahmood, F., 2019. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. arXiv:1910.10825 [cs, q-bio] ArXiv: 1910.10825.

Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering , 1–16 doi:10.1038/s41551-020-00682-w.

Lu, M.Y., Zhao, M., Shady, M., Lipkova, J., Chen, T.Y., Williamson, D.F.K., Mahmood, F., 2020. Deep learning-based computational pathology predicts origins for cancers of unknown primary. arXiv:2006.13932 [cs, q-bio] ArXiv: 2006.13932.

Ma, N., Zhang, X., Zheng, H.T., Sun, J., 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European conference on computer vision (ECCV), pp. 116–131.

Marquard, A.M., Eklund, A.C., Joshi, T., Krzystalik, M., Favero, F., Wang, Z.C., Richardson, A.L., Silver, D.P., Szallasi, Z., Birkbak, N.J., 2015. Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. Biomarker research 3, 1–10.

Mauri, G., Arena, S., Siena, S., Bardelli, A., Sartore-Bianchi, A., 2020. The dna damage response pathway as a land of therapeutic opportunities for colorectal cancer. Annals of Oncology .

Oord, A.v.d., Li, Y., Vinyals, O., 2019. Representation learning with contrastive predictive coding. arXiv:1807.03748 [cs, stat] ArXiv: 1807.03748.

Pearl, L.H., Schierz, A.C., Ward, S.E., Al-Lazikani, B., Pearl, F.M., 2015. Therapeutic opportunities within the dna damage response. Nature Reviews Cancer 15, 166–180.

Pilié, P.G., Tang, C., Mills, G.B., Yap, T.A., 2019. State-of-the-art strategies for targeting the dna damage response in cancer. Nature reviews Clinical oncology 16, 81–104.

Pinckaers, H., Bulten, W., van der Laak, J., Litjens, G., 2020a. Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. arXiv:2006.03394 [cs, eess] ArXiv: 2006.03394.

Pinckaers, J.H.F.M., Ginneken, B.v., Litjens, G., 2020b. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–1doi:10.1109/TPAMI.2020.3019563.

Popova, T., Manié, E., Reunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., Delattre, O., Sigal-Zafrani, B., Bollet, M., Longy, M., et al., 2012. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with brca1/2 inactivation. Cancer research 72, 5454–5462.

Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. arXiv preprint arXiv:1902.07208 .

Snowsill, T., Coelho, H., Huxley, N., Jones-Hughes, T., Briscoe, S., Frayling, I.M., Hyde, C., 2017. Molecular testing for lynch syndrome in people with colorectal cancer: systematic reviews and economic evaluation. Health Technol Assess .
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

2020. Extending unsupervised neural image compression with supervised multitask learning, in: Medical Imaging with Deep Learning, PMLR. p. 770–783.

Tellez, D., Litjens, G., van der Laak, J., Ciompi, F., 2019. Neural image compression for gigapixel histopathology image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1doi:10.1109/TPAMI.2019.2936841

Timms, K.M., Abkevich, V., Hughes, E., Neff, C., Reid, J., Morris, B., Kalva, S., Potter, J., Tran, T.V., Chen, J., et al., 2014. Association of brca1/2 defects with genomic scores predictive of dna damage repair deficiency among breast cancer subtypes. Breast Cancer Research 16, 1–9.

Valieris, R., Amaro, L., Osório, C.A.B.d.T., Bueno, A.P., Rosales Mitrowsky, R.A., Carraro, D.M., Nunes, D.N., Dias-Neto, E., Silva, I.T.d., 2020. Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer. Cancers 12, 3687.

van der Velden, D.L., Hoes, L., Van Der Wijngaart, H., van Berge Henegouwen, J., Van Werkhoven, E., Roepman, P., Schilsky, R., De Leng, W., Huijtema, A., Nuijten, B., et al., 2019. The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs. Nature 574, 127–131.

Xie, C., Muhammad, H., Vanderbilt, C.M., Caso, R., Yarlagadda, D.V.K., Campanella, G., Fuchs, T.J., 2020. Beyond classification: Whole slide tissue histopathology analysis by end-to-end part learning, in: Medical Imaging with Deep Learning, PMLR. pp. 843–856.

Zhu, L., Huang, Y., Fang, X., Liu, C., Deng, W., Zhong, C., Xu, J., Xu, D., Yuan, Y., 2018. A novel and reliable method to detect microsatellite instability in colorectal cancer by next-generation sequencing. The Journal of Molecular Diagnostics 20, 225–231.
Appendix A. Decision for weighted standard deviation

We will explain this decision in more detail with a toy example below. Let’s imagine our WSIs have only 6 tiles, and that the feature describing each tile is 1-dimensional, e.g., the average nucleus size in that tile. We can imagine having two different WSIs, the first of which has most nuclei enlarged, and the other only part of the nuclei enlarged, but which are greatly enlarged. We would expect the first to be caused by, e.g., a TP53 mutation without HRD, whereas the second could be caused by HRD. The attention layer might have learned that only tiles with nuclei are important and sets the attention to all other tiles to 0, and will in-
Table B.3: Results of tile-supervision with majority vote with a variety of feature extractor initialization for TP53 mutation prediction. Each experiment was run for 4 epochs with a batch size of 512, with a learning rate of $5 \times 10^{-5}$, evaluated on the validation set every 100 steps. The model with the highest AUC on the validation set was used for prediction on the test set, for which the 5-fold mean and standard deviation is reported.

| pre-training method | Batch size (# GPUs) | Training time (hours) per epoch | # Epochs | Downstream task AUC% |
|---------------------|---------------------|---------------------------------|-----------|----------------------|
| BYOL                | 400 (1)             | 2.08                            | 5         | 69.87 ± 5.66         |
|                     |                     |                                 | 35        | 72.12 ± 5.44         |
|                     | 4000 (1)            | 2                               | 5         | 66.35 ± 4.90         |
|                     |                     |                                 | 35        | 69.15 ± 4.56         |
| SimCLR              | 500 (1)             | 1.4                             | 5         | 71.18 ± 3.77         |
|                     |                     |                                 | 35        | 70.54 ± 6.26         |
| ImageNet            | -                   | -                               | -         | 68.65 ± 7.48         |
| Random init         | -                   | -                               | -         | 60.34 ± 7.11         |

Figure B.6: Loss over time for SimCLR. SimCLR-500 is trained with a batch size of 500, whereas SimCLR-950 uses a batch size of 950.

Figure B.7: Loss over time for BYOL. BYOL-400 uses a batch size of 400, while BYOL-4000 uses a virtual batch size of 4000.

Appendix B.2. Hyperparameter sensitivity of DeepMIL

The experiments below are performed on extracted features using a feature extractor pre-trained with SimCLR-950 for 45 epochs. DeepMIL is generally trained with a mini batch size of 1 since the bag contains a variable number of instances. We use a batch size larger than one, and solve the variable size issue by padding each patient’s feature matrix with zero-vectors up to a predefined size. When using all available tiles, we pad up to 12,000 instances, since the largest WSI has 11,000 tiles, when subsampling 500 tiles we pad up to 550 instances, and when we subsample fewer than 500 tiles, we pad the patients with fewer tiles than this up to the number of subsampled tiles. The base setting is subsampling 500 tiles. To test the hyperparameter settings with DeepMIL we do a grid search and sample the batch size from $[1, 4, 8, 16]$ the learning rate and regularization from $[10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-3}, 10^{-3}]$, and for the learning rate we additionally tested the effect of increasing it further to $5 \times 10^{-3}, 10^{-2}$, and $5 \times 10^{-2}$. For the subsampling experiments, we use all tiles or randomly sample 500 or 50 tiles from a patient. We compare performance on TP53 mutation prediction on the TCGA-BC dataset.

Effect of subsampling In Table B.4 we see that subsampling 50-500 tiles results in a similar performance as not subsampling any tiles. To enjoy the performance increase, faster training, and to allow inspection of a larger
DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images

Table B.4: Results of DeepMIL on SimCLR-950 extracted features on TCGA-BC for a variety of subsampling sizes. Experiments are run for 8 epochs with a batch size of 16, learning rate of $5 \times 10^{-4}$, regularization of $10^{-4}$, evaluating every 10 steps and testing the model with the highest AUC on the validation set on the test set.

| # tiles for classifier training | Performance (AUC%) |
|--------------------------------|--------------------|
|                                | TP53               | mHRD               |
| All                            | 70.49 ± 5.06       | 67.42 ± 2.81       |
| 500                            | 73.42 ± 4.13       | 71.38 ± 4.14       |
| 50                             | 73.00 ± 4.45       | 74.04 ± 2.68       |

Table B.5: Results of DeepMIL on SimCLR-950 extracted features on TCGA-BC trained with a variety of batch sizes. Experiments are run on a subsample of 500 tiles for each patient for 16 epochs with varying batch size, learning rate of $5 \times 10^{-4}$ and regularization of $10^{-4}$, evaluating every 10 steps and testing the model with the highest AUC on the validation set on the test set.

| Batch size | Performance (AUC%) |
|------------|--------------------|
|            | TP53               |
| 1          | 72.47 ± 4.97       |
| 4          | 74.22 ± 5.15       |
| 8          | 73.74 ± 4.08       |
| 16         | 74.16 ± 4.54       |

variety of tiles, we perform the main experiments with a subsample of 500 tiles.

**Effect of minibatch training** Next, Table B.5 shows that the DeepMIL model is invariant to varying batch size on the large TCGA-BC dataset. To speed up training, we use a batch size of 16 for all our following experiments.

**Effect of learning rates and regularization** An overview of final TP53 classification performance of DeepMIL with a variety of training scheme parameters can be found in Table B.6 and shows us that the base parameters as previously published give us the best results.
### Table B.6: Results of hyperparameter grid search using DeepMIL for TP53 mutation prediction on the same feature vectors on TCGA-BC. Trained for 8 epochs with a batch size of 16, evaluating every 10 steps, with varying learning rate and regularization, trained on a subsample of 500 tiles.

| Learning rate | Regularization | $10^{-5}$ | $5 \times 10^{-5}$ | $10^{-4}$ | $5 \times 10^{-4}$ | $10^{-3}$ |
|---------------|----------------|-----------|-------------------|-----------|-------------------|-----------|
| $10^{-5}$     |                | 59.20 ± 7.10 | 59.21 ± 7.11 | 59.22 ± 7.10 | 59.20 ± 7.10 | 59.20 ± 7.08 |
| $5 \times 10^{-5}$ |                | 66.13 ± 8.17 | 66.13 ± 8.17 | 66.12 ± 8.18 | 66.11 ± 8.14 | 66.16 ± 8.17 |
| $10^{-4}$     |                | 67.42 ± 8.00 | 67.42 ± 7.99 | 67.41 ± 7.99 | 67.41 ± 8.01 | 67.44 ± 8.07 |
| $5 \times 10^{-4}$ |                | 72.33 ± 4.68 | 72.33 ± 4.72 | 72.34 ± 4.74 | 72.25 ± 4.85 | 72.12 ± 4.94 |
| $10^{-3}$     |                | 74.25 ± 4.62 | 74.17 ± 4.60 | 74.06 ± 4.65 | 74.00 ± 4.64 | 73.80 ± 4.55 |
| $5 \times 10^{-3}$ |                |              |               |          | 66.10 ± 1.03 |          |
| $10^{-2}$     |                |              |               |          | 62.19 ± 6.26 |          |
| $5 \times 10^{-2}$ |                |              |               |          | 61.95 ± 4.30 |          |