PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Evaluating reproducibility of AI algorithms in digital pathology with DAPPER

Andrea Bizzego[1,2☯], Nicole Bussola[1,3☯], Marco Chierici[1], Valerio Maggio[1], Margherita Francescatto[1], Luca Cima[4], Marco Cristoforetti[1], Giuseppe Jurman[1]*, Cesare Furlanello[1]

1 Fondazione Bruno Kessler, Trento, Italy, **2** DIPSCO, University of Trento, Trento, Italy, **3** Department CIBIO, University of Trento, Trento, Italy, **4** Pathology Unit, Santa Chiara Hospital, Trento, Italy

☯ These authors contributed equally to this work.
* jurman@fbk.eu

Abstract

Artificial Intelligence is exponentially increasing its impact on healthcare. As deep learning is mastering computer vision tasks, its application to digital pathology is natural, with the promise of aiding in routine reporting and standardizing results across trials. Deep learning features inferred from digital pathology scans can improve validity and robustness of current clinico-pathological features, up to identifying novel histological patterns, *e.g.*, from tumor infiltrating lymphocytes. In this study, we examine the issue of evaluating accuracy of predictive models from deep learning features in digital pathology, as an hallmark of reproducibility. We introduce the DAPPER framework for validation based on a rigorous Data Analysis Plan derived from the FDA's MAQC project, designed to analyze causes of variability in predictive biomarkers. We apply the framework on models that identify tissue of origin on 787 Whole Slide Images from the Genotype-Tissue Expression (GTEx) project. We test three different deep learning architectures (VGG, ResNet, Inception) as feature extractors and three classifiers (a fully connected multilayer, Support Vector Machine and Random Forests) and work with four datasets (5, 10, 20 or 30 classes), for a total of 53, 000 tiles at 512 × 512 resolution. We analyze accuracy and feature stability of the machine learning classifiers, also demonstrating the need for diagnostic tests (*e.g.*, random labels) to identify selection bias and risks for reproducibility. Further, we use the deep features from the VGG model from GTEx on the KIMIA24 dataset for identification of slide of origin (24 classes) to train a classifier on 1, 060 annotated tiles and validated on 265 unseen ones. The DAPPER software, including its deep learning pipeline and the Histological Imaging—Newsy Tiles (HINT) benchmark dataset derived from GTEx, is released as a basis for standardization and validation initiatives in AI for digital pathology.

Author summary

In this study, we examine the issue of evaluating accuracy of predictive models from deep learning features in digital pathology, as an hallmark of reproducibility. It is indeed a top priority that reproducibility-by-design gets adopted as standard practice in building and

OPEN ACCESS

Citation: Bizzego A, Bussola N, Chierici M, Maggio V, Francescatto M, Cima L, et al. (2019) Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. PLoS Comput Biol 15(3): e1006269. https://doi.org/10.1371/journal.pcbi.1006269

Editor: Gustavo Rohde, University of Virginia, UNITED STATES

Received: June 3, 2018

Accepted: February 20, 2019

Published: March 27, 2019

Data Availability Statement: All relevant data are within the paper, its supporting information file or otherwise available from the software gitlab repositories: https://gitlab.fbk.eu/mpba-histology/dapper https://gitlab.fbk.eu/mpba-histology/histolib.

Funding: Microsoft Azure Research grant for Deep Learning for Precision Medicine assigned to CF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

validating AI methods in the healthcare domain. Here we introduce DAPPER, a first framework to evaluate deep features and classifiers in digital pathology, based on a rigorous data analysis plan originally developed in the FDA's MAQC initiative for predictive biomarkers from massive omics data. We apply DAPPER on models trained to identify tissue of origin from the HINT benchmark dataset of 53, 000 tiles from 787 Whole Slide Images in the Genotype-Tissue Expression (GTEx) project, available at the web address https://gtexportal.org. We analyze accuracy and feature stability of different deep learning architectures (VGG, ResNet and Inception) as feature extractors and classifiers (a fully connected multilayer, Support Vector Machine and Random Forests) on up to 20 classes. Further, we use the deep features from the VGG model (trained on HINT) on the 1, 300 annotated tiles of the KIMIA24 dataset for identification of slide of origin (24 classes). The DAPPER software is available together with the scripts to generate the HINT benchmark dataset.

Introduction

Artificial Intelligence (AI) methods for health data hold great promise but still have to deal with disease complexity: patient cohorts are most frequently an heterogeneous group of subtypes diverse for disease trajectories, with highly variable characteristics in terms of phenotypes (*e.g*. bioimages in radiology or pathology), response to therapy, clinical course, thus a challenge for machine-learning based prognoses. Nevertheless, the increased availability of massive annotated medical data from health systems and a rapid progress of machine learning frameworks has led to high expectations about the impact of AI on challenging biomedical problems [1]. In particular, Deep Learning (DL) is now surpassing pattern recognition methods in the most complex medical images challenges such as those proposed by the Medical Image Computing & Computer Assisted Intervention conferences (MICCAI, https://www.miccai2018.org/en/WORKSHOP---CHALLENGE---TUTORIAL.html), and it is comparable to expert accuracy in the diagnosis of skin lesions [2], classification of colon polyps [3, 4], ophthalmology [5], radiomics [6] and other areas [7]. However, the reliable comparison of DL with other baseline ML models and human experts is not a diffuse practice yet [8], and also the reproducibility and interpretation of the challenges' outcome have been recently criticized [9, 10]. DL refers to a class of machine learning methods that model high-level abstractions in data through the use of modular architectures, typically composed by multiple nonlinear transformations estimated by training procedures. Notably, deep learning architectures based on Convolutional Neural Networks (CNNs) hold state-of-the-art accuracy in numerous image classification tasks without prior feature selection. Further, intermediate steps in the pipeline of transformations implemented by CNNs or other deep learning architectures can provide a mapping (*embedding*) from the original feature space into a *deep feature* space. Of interest for medical diagnosis, deep features can be used for interpretation of the model and can be directly employed as inputs to other machine learning models.

Deep learning methods have been applied to analysis of histological images for diagnosis and prognosis. Mobadersany and colleagues [11] combine in the Survival Convolutional Neural Network (SCNN) architecture a CNN with traditional survival models to learn survival-related patterns from histology images, predicting overall survival of patients diagnosed with gliomas. Predictive accuracy of SCNN is comparable with manual histologic grading by neuropathologists. Further, by incorporation of genomic variables for gliomas in the model, the extended model significantly outperforms the WHO paradigm based on genomic subtype and histologic grading. Similarly, deep learning models have been successfully applied to histology for colorectal cancer [12], gastric cancer [13], breast cancer [14] and lung cancer [15, 16].

As human assessments of histology are subjective and hard to repeat, computational analysis of histology imaging within the information environment generated from a digital slide (*digital pathology*) and advances in scanning microscopes have already allowed pathologists to gain a much more effective diagnosis capability and dramatically reduce time for information sharing. Starting from the principle that underlying differences in the molecular expressions of the disease may manifest as tissue architecture and nuclear morphological alterations [17], it is clear that automatic evaluation of disease aggressiveness level and patient subtyping has a key role aiding therapy in cancer and other diseases. Digital pathology is in particular a key tool for the immunotherapy approach, which stands on the characterization of tumor-infiltrating lymphocytes (TILs) [18]. Indeed, quantitative analysis of the immune microenvironment by histology is crucial for personalized treatment of cancer [19, 20], with high clinical utility of TILs assessment for risk prediction models, adjuvant, and neoadjuvant chemotherapy decisions, and for developing the potential of immunotherapy [21, 22]. Digital pathology is thus a natural application domain for machine learning, with the promise of accelerating routine reporting and standardizing results across trials. Notably, deep learning features learned from digital pathology scans can improve validity and robustness of current clinico-pathological features, up to identifying novel histological patterns, *e.g.* from TILs.

On the technical side, usually deep learning models for digital pathology are built upon imaging architectures originally aimed at tasks in other domains and trained on non-medical datasets. This is a foundational approach in machine learning, known as *transfer learning*. Given domain data and a network pretrained to classify on huge generic databases (*e.g.* ImageNet, with over 14 million items and 20 thousand categories [23]), there are three basic options for transfer learning, *i.e.* to adapt the classifier to the new domain: a) train a new machine learning model on the features preprocessed by the pretrained network from the domain data; b) retrain only the deeper final layers (the *domain layers*) of the pretrained network; c) retrain the whole network starting from the pretrained state. A consensus about the best strategy to use for medical images is still missing [24, 25].

In this study we aim to address the issue of reproducibility and validation of machine learning models for digital pathology. Reproducibility is a paramount concern in biomarker research [26], and in science in general [27], with scientific communities, institutions, industry, and publishers struggling to foster adoption of best practices, with initiatives ranging from enhancing reproducibility of high-throughput technologies [28] to improving the overall reuse of scholarly data and analytics solutions (*e.g.* the FAIR Data Principles [29]). As an example, the MAQC initiatives [30, 31], led by the US FDA, investigate best practices and causes of variability in the development of biomarkers and predictive classifiers from massive omics data (*e.g.* microarrays, RNA-Seq or DNA-Seq data) for precision medicine. The MAQC projects adopt a Data Analysis Plan (DAP) that forces bioinformatics teams to submit classification models, top features ranked for importance and performance estimates all built on training data only, before testing on unseen external validation data. The DAP approach is methodologically more robust than a simple cross validation (CV) [30] as the internal CV and model selection phase is replicated multiple times (*e.g.*, 10 times) to smooth the impact of a single training/test split; the performance metrics is thus evaluated on a much larger statistics. Also, features are analyzed and ranked multiple times, averaging the impact of a small round of partitions. The ranked feature lists are fused in a single ranked list using the Borda method [32] and the bootstrap method is applied to compute the confidence intervals. This approach helps mitigating the risk of selection bias in complex learning pipelines [33], where the bias can stem in one of many preprocessing steps as well as in the downstream machine learning model.

Further, it clarifies that increasing task difficulty is often linked to a decrease not only in accuracy measures but also of stability of the biomarker lists [32], *i.e.* the consistency in the selection of the top discriminating features across all repeated cross validation runs.

Although openness in sharing algorithms and benchmark data is a solid attitude of the machine learning community, the reliable estimation on a given training dataset of predictive accuracy and stability of deep learning models (in terms of performance range as a function of variations of training data) and the stability of deep features used by external models (as the limited difference of top ranking variables selected by different models) is still a gray area. The underlying risk is that of overfitting the training data, or worse to overfit the validation data if the labels are visible, which is typical when datasets are fully released at the end of a data science challenge on medical image data. As the number of DL-based studies in digital pathology is exponentially growing, we suggest that the progress of this field needs environments (*e.g.*, DAPs) to prevent such pitfalls, especially if features distilled by the network are used as radiomics biomarkers to inform medical decision. Further, given an appropriate DAP, alternative model choices should be benchmarked on publicly available datasets, as usual in the general computer vision domain (*e.g.*, ImageNet [23] or COCO [34]).

This study provides three main practical contributions to controlling for algorithmic bias and improving reproducibility of machine learning algorithms for digital pathology:

1. A Data Analysis Plan (DAP) specialized for digital pathology, tuned on the predictive evaluation of deep features, extracted by a network and used by alternative classification heads. To the best of our knowledge, this is the first study where a robust model validation method (the DAP) is applied in combination with the deep learning approach. We highlight that the approach can be adopted in other medical/biology domains in which Artificial Intelligence is rapidly emerging, *e.g.*, in the analysis of radiological images.

2. A benchmark dataset (HINT) of 53, 727 tiles of histological images from 30 tissue types, derived from GTEx [35] for the recognition of tissue of origin of up to 30 classes. The HINT dataset can be used by other researchers to pretrain the weights of DL architectures that shall be applied on digital pathology tasks (*e.g.*, detection of TILs) thus accelerating the training of application-specific models. In the past 5 years, having a shared image dataset (*e.g.*, the ImageNet) allowed the development of a number of deep learning models for general image classification (*e.g.* VGG, ResNet, AlexNet). Such pretrained networks have then been effectively applied on a variety of different tasks. With the HINT dataset we aim at favouring a similar boost on digital pathology.

3. An end-to-end machine learning framework (DAPPER) as a baseline environment for predictive models in digital pathology, where end-to-end indicates that the DAPPER framework is directly applied to the digital pathology images, with the deep learning component producing features for the machine learning head, without an external procedure (*e.g.*, a handcrafted feature extraction) to preprocess the features. To the best of our knowledge, this is the first example of a DL approach for the classification of up to 30 different tissues, all with the same staining, which represents, *per se*, a valuable contribution to the digital pathology community.

We first apply DAPPER to a set of classification experiments on 787 Whole Slide Images (WSIs) from GTEx. The framework (see Fig 1) is composed by (A) a preprocessing component to derive patches from WSIs; (B) a 3-step machine learning pipeline with a data augmentation preprocessor, a backend deep learning model, and an adapter extracting the deep features; (C) a downstream machine learning/deep learning head, *i.e.* the task specific predictor. In our experiments, we evaluate the accuracy and the feature stability in a multiclass setting for the

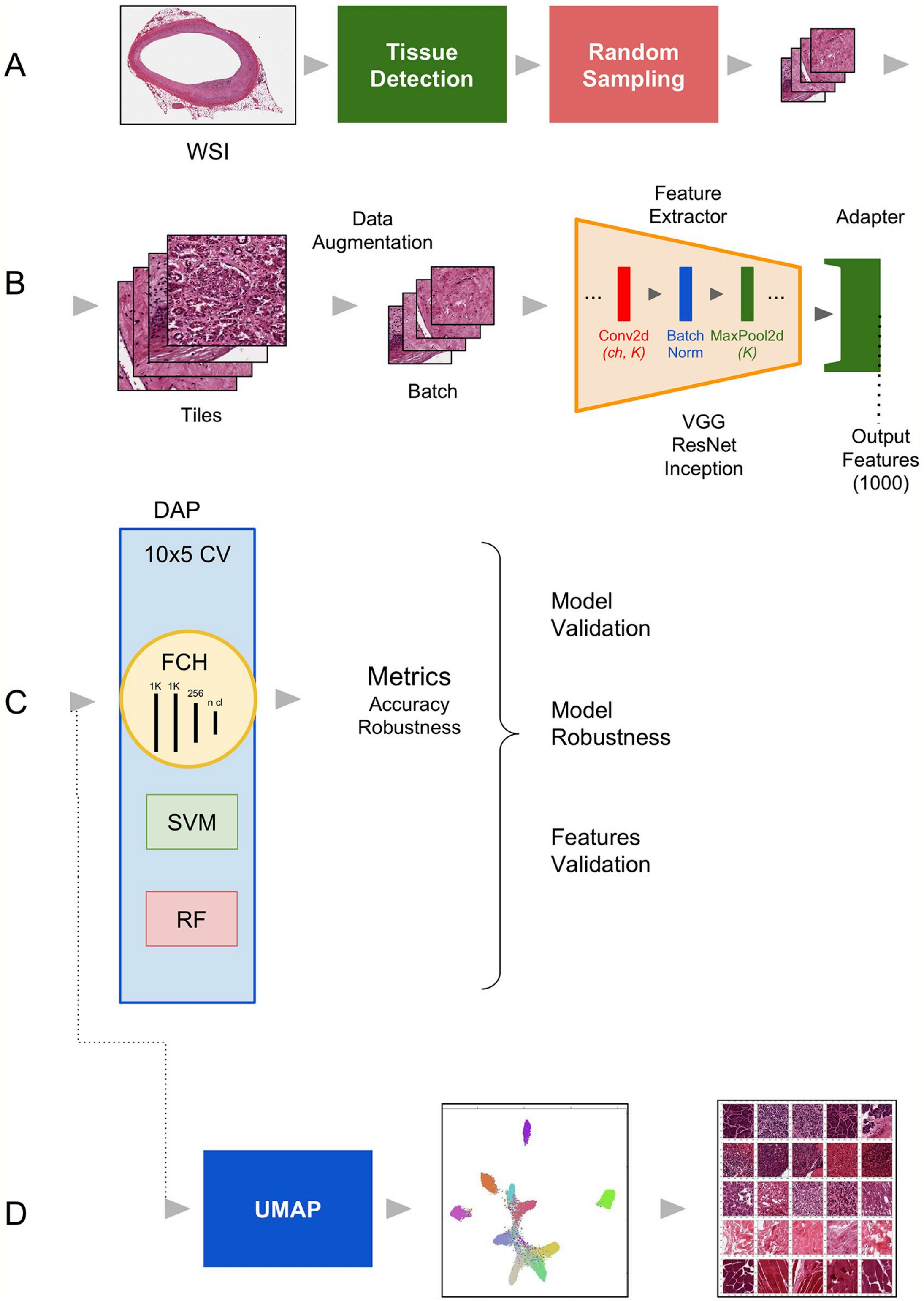


Fig 1. The DAPPER environment. Components: A) The WSI preprocessing pipeline; B) the deep learning backend to extract deep features; C) the Data Analysis Plan (DAP) for the machine learning models; and D) the UMAP module and other modules for unsupervised analysis.

https://doi.org/10.1371/journal.pcbi.1006269.g001

combination of three different deep learning architectures, namely VGG, ResNet and Inception, used as feature extractors, and three classifiers, a fully connected multilayer network, Support Vector Machine (SVM) [36] and Random Forest (RF) [37]. This component is endowed with the DAP, *i.e.*, a 10 × 5 CV (5-fold cross validation iterated 10 times). The 50 internal validation sets are used to estimate a vector of metrics (with confidence intervals) that are then used for model selection. In the fourth component (D) we finally provide an unsupervised data analysis based on the UMAP projection method, and methods for feature exploration. The DAPPER software is available together with the Python scripts and the instructions to generate the HINT benchmark dataset as a collection of Jupyter notebooks at `gitlab.fbk.eu/mpba-histology/dapper`, released under the GNU General Public License v3. Notably, the DAP estimates are provided in this paper only for the downstream machine learning/deep learning head in component (C); whenever computational resources are available, the DAP can be expanded also to component (B). Here we kept as a separate problem the model selection exercise on the backend deep learning architecture in order to clarify the change of perspective with respect to optimization of machine learning models in the usual training-validation setting.

As a second experiment, in order to study the DAPPER framework in a transfer learning condition, we use the deep features from the VGG model trained on a subset of HINT on the 1, 300 annotated tiles of the KIMIA Path24 dataset [38] to identify in this case the slide of origin (24 classes).

Previous work on classifying WSIs by means of neural networks was introduced by [38, 39], also with the purpose of distributing the two original datasets KIMIA Path960 (KIMIA960) and KIMIA Path24 (KIMIA24). KIMIA24 consists of 24 WSIs chosen on purely visual distinctions. Babaie and coauthors [38] manually selected a total of 1, 325 binary patches with 40% overlap. On this dataset, in addition to two models based on Local Binary Patterns (LBP) and Bag-of-Visual-Words (BoVW), they applied two shallow CNNs, achieving at most 41.8% accuracy. On the other hand, KIMIA960 contains 960 histopathological images belonging to 20 different WSIs that, again on visual clues, were used to represent different texture/pattern/staining types. The very same experimental settings as the one for KIMIA24, *i.e.*, LBP, BoVW and CNN, has been replicated on this dataset by Kumar and coauthors [39]. In particular, the authors applied AlexNet or VGG16, both pretrained on ImageNet, to extract deep features; instead of a classifier, accuracy was established by computing similarity distances between the 4, 096 features extracted. Also, Kieffer and coauthors in [25] explored the use of deep features from several pretrained structures on KIMIA24, controlling for the impact of transfer learning and finding an advantage of pretrained networks against training from scratch. Conversely, Alhindi and coworkers [40] analyzed KIMIA960 for slide of origin (20 slides preselected by visual inspection), and similarly to our study they compared alternative classifiers as well as feature extraction models in a 3-fold CV setup. Considering the importance of clinical validation of predictive results [8], we finally compared the performance of the DAPPER framework with an expert pathologist. DAPPER outperforms the pathologist in classifying tissues at tile level, while at WSI level performance are similar.

DAPPER represents an advancement over previous studies, due to the DAP structure and its application to the large HINT dataset free of any visual preselection.

Materials and methods

Dataset

The images used to train the models were derived from the Genotype-Tissue Expression (GTEx) Study [35]. The study collects gene expression profiles and whole-slide images (WSIs)

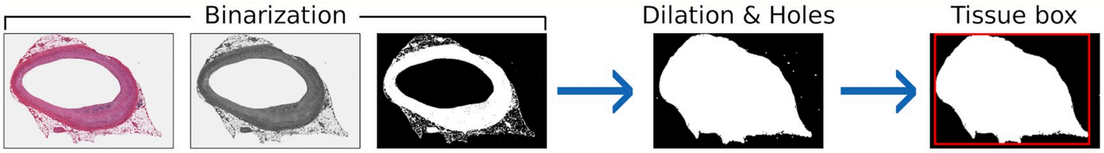


Fig 2. The tissue detection pipeline. The identification of the tissue bounding box is performed on the WSI thumbnail in three steps: a) Binarization of the grayscale image by applying Otsu's thresholding; b) Binary dilation and filling of the holes; c) Selection of the biggest connected region as tissue region and computation of the vertex of the containing rectangle.

https://doi.org/10.1371/journal.pcbi.1006269.g002

of 53 human tissues histologies used to investigate the relationship between genetic variation and tissue-specific gene expression in healthy individuals. To ensure that the collected tissues meet prescribed standard criteria, a Pathology Resource Center validated each sample origin, content, integrity and target tissue (https://biospecimens.cancer.gov/resources/sops/). After sectioning and Haemotoxylin and Eosin staining (H&E), tissue samples were scanned using a digital whole slide imaging system (Aperio) and stored in *.svs* format [41].

A custom Python script was used to download 787 WSIs through the Biospecimen Research Database (total size: 192 GB, average 22 WSIs for each tissue). The list of the downloaded WSIs is available in S1 Table.

A data preprocessing pipeline was developed to prepare the WSIs as training data (see Fig 2). The WSIs have a resolution of 0.275 μm/pixel (Magnification 40X) and variable dimensions. Further, the region interested by the tissue is only a portion of the WSI and it varies across the samples. Hence first we identified the region of the tissue in the image (see Fig 2), then we extracted at most 100 tiles (512 × 512 pixel) from the WSIs, by randomly sampling the tissue region. We applied the algorithm for the detection of the tissue region (see Fig 2) on each tile and rejected those where the portion of the tissue was below 85%. A total number of 53, 727 tiles was extracted, with a number of tiles per tissue varying between 59 (for Adipose—Visceral (Omentum)) and 2, 689 (for Heart—Left Ventricle). Four datasets (HINT5, HINT10, HINT20, HINT30) have been derived with increasing number of tissues for a total of 52, 991 tiles; the full number of tiles per anatomical zone, for each dataset, is available in S2 Table and summarized in Table 1. We refer to the four sets as the HINT collection, or the HINT dataset in brief. We choose the five tissues composing HINT5 based on exploratory experiments, while the other three datasets were composed including the tissues with higher number of tiles. The class imbalance is accounted for by weighting the error on predictions. In detail, the weight w of the class i used in the cross entropy function is computed as: $w_i = n_{\max}/n_i$, where $n_{\max}$ is the number of tiles in the class with more tiles and n_i is the number of tiles in the class i.

Since image orientation should not be relevant for the tissue recognition, the tiles are randomly flipped (horizontally and vertically) and scaled, following a common practice in deep learning known as *data augmentation*. Data augmentation consists of different techniques

Table 1. Summary of the HINT datasets. Total: total number of tiles composing the dataset; Min: number of tiles in the class with less samples; Max: number of tiles in the class with more samples; Average; average number of tiles for each class.

Name	# tissues	Total	Min	Max	Average
HINT5	5	8, 218	1, 009	2, 424	1, 643.6
HINT10	10	22, 885	1, 890	2, 689	2, 288.5
HINT20	20	40, 516	1, 574	2, 689	2, 025.8
HINT30	30	52, 991	957	2, 689	1, 766.4

https://doi.org/10.1371/journal.pcbi.1006269.t001

(such as cropping, flipping, rotating images) performed each time a sample is loaded, so that the resulting input image is different at each epoch. Augmentation has proven effective in multiple problems, increasing the generalization capabilities of the network, preventing overfitting and improving models performance [42–44].

Such randomized transformations were found to provide more comparable performance between the prognostic accuracy of the deep learning SCNN architecture and that of standard models (*i.e*., Support Vector Machine, Random Forest) based on combined molecular subtype and histologic grade [11]. In addition, each tile is cropped to a fixed size, which is dependent on the type of network used to extract the deep features.

Deep learning architectures and training strategies

We exploited three backend architectures commonly used in computer vision tasks:

1. VGG, Net-E version (19 layers) with Batch Normalization (BN) layers [45];
2. ResNet, 152-layer model [46];
3. Inception, version 3 [47].

These architectures have reached highest accuracy in multiclass classification problems over the last 4 years [48] and differ in resource utilization (see Table 2). The feature extraction layer of each backend network is obtained as the output of an end-to-end pipeline composed of the following main blocks (see panel B in Fig 1):

1. Data augmentation: the input tiles are processed and assembled into batches of size 32;
2. Feature Extractor: series of convolutional layers (Conv2d: with different number of channels and kernel size), normalization layers (Batch Norm) and pooling layers (MaxPool2d: with different kernel size) designed to fit with the considered backend architecture (VGG, ResNet, Inception). The number of output features of the Feature Extractor depends on the structure of the backend architecture used;
3. Adapter: as the backend networks have output features of different sizes, we add a linear layer at the end of the Feature Extractor, in order to make the pipeline uniform. The Adapter takes the features of the backend network as input and output a fixed number of features (1, 000).

The 1, 000 Adapter features are then used as input for a classifier providing predicted tissue labels as output. As predictive models, we used a linear SVM with regularization parameter C set to 1, a RF classifier with 500 trees (both implemented in *scikit-learn*, `v0.19.1`) and a fully connected head (FCH), namely a series of fully connected layers (see panel C in Fig 1). Inspired by [11] and [49], our FCH consists of four dense layers with 1, 000, 1, 000, 256 and *number of tissue classes* nodes, respectively. The feature extraction block was initialized with the weights already trained on the ImageNet dataset [23], provided by *PyTorch* (`v0.4.0`) and frozen. The Adapter block is trained together with the FCH as a one network. Training also

Table 2. Backend architectures statistics.

Name	Output features	#Parameters	Layers
VGG	25, 088	155×10^6	19
ResNet	2, 048	95×10^6	152
Inception	2, 048	35×10^6	42

https://doi.org/10.1371/journal.pcbi.1006269.t002

the weights of the feature extraction block improves accuracy (see S3 Table). However, these results were not validated rigorously within the DAP and therefore they not are not claimed as generalized in this study.

For the optimization of the other weights (Adapter and FCH) we used the Adam algorithm [50] with the learning rate set to 10^{-5} and fixed for the whole training. We used the cross entropy as the loss function, which is appropriate for multiclass models.

The strategy to optimize the learning rate was selected based on results of a preparatory study with the VGG network and HINT5. The strategy approach with fixed learning rate achieved the best results (see S4 Table) and was therefore adopted in the rest of the study.

Data analysis plan

Following the rigorous model validation techniques proposed by the MAQC projects [30, 31], we adopted a DAP to assess the validity of the features extracted by the networks, namely a 10 × 5-fold cross validation (CV) schema. The input dataset is first partitioned in two separate datasets, the *training set* and the *test set*, also referred as *external validation set* as reported in [30, 31]. The external validation set will be kept completely unseen to the model, and it will be only used in the very last step of the DAP for the final model evaluation. In our experimental settings, we used 80% of the total samples for the training set, and the remaining 20% for the external validation set. A stratification strategy upon the classes of tiles, *i.e.*, 5, 10, or 20, has been adopted in the partitioning. The training set further undergoes a 5-fold CV iterated 10 times, resulting in 50 separated *internal validation sets* used for model evaluation within the DAP. The same stratification strategy is used in the creation of the folds.

At each CV iteration, features are ranked by KBest, with ANOVA F-score as the scoring function [51], and four separate models are trained on sets of increasing number of ranked features (namely: 10%, 25%, 50%, 100% of the total number of features). A list of top-ranked features is obtained by Borda aggregation of the ranked lists, for which we also compute the Canberra stability with a computational framework designed for sets of ranked biomarker lists [32].

As for model evaluation, we considered the accuracy (ACC), and the Matthews Correlation Coefficient (MCC) in their multiclass generalization [52–54]:

$$\mathrm{ACC} = \frac{\sum_{k=1}^{N} C_{kk}}{\sum_{i,j=1}^{N} C_{ij}}, \qquad 0 \leq \mathrm{ACC} \leq 1 \tag{1}$$

$$\mathrm{MCC} = \frac{\sum_{k,l,m=1}^{N} \left(C_{kk}C_{ml} - C_{lk}C_{km}\right)}{\sqrt{\sum_{k=1}^{N}\left[\sum_{l=1}^{N} C_{lk} \sum_{\substack{f,g=1\\ f\neq k}}^{N} C_{gf}\right]}\sqrt{\sum_{k=1}^{N}\left[\sum_{l=1}^{N} C_{kl} \sum_{\substack{f,g=1\\ f\neq k}}^{N} C_{fg}\right]}}, \qquad -1 \leq \mathrm{MCC} \leq 1 \tag{2}$$

where N is the number of classes and C_{st} is the number of elements of true class s that have been predicted as class t.

MCC is widely used in Machine Learning as a performance metric, especially for unbalanced sets, for which ACC can be misleading [55]. In particular, MCC gives an indication of prediction robustness among classes: MCC = 1 is perfect classification, MCC = −1 is extreme misclassification, and MCC = 0 corresponds to random prediction.

Finally, the overall performance of the model is evaluated across all the iterations (*i.e.*, internal validation sets), in terms of average MCC and ACC with 95% Studentized bootstrap confidence intervals (CI) [56], and then on the external validation set.

As a sanity check to avoid unwanted selection bias effects, the DAP is repeated stochastically scrambling the training set labels (*random labels* mode) or by randomly ranking features before building models (*random ranking* mode: in presence of pools of highly correlated variables, top features can be interchanged with others, possibly of higher biological interest). In both modes, a procedure unaffected by selection bias should achieve an average MCC close to 0.

Experiments on HINT

We designed a set of experiments reported in Table 3 to provide indications about the optimal architecture for deep feature extraction, while keeping fixed the other hyper-parameters. In particular we set batch size (32) and number of epochs (50), large enough to let the network converge: we explored increasing numbers of epochs (10, 30, 50, 100) and, since the loss stabilizes after about 35 epochs, we set the number of epochs to 50. First, we compared the three backend architectures on the smallest dataset HINT5, with fixed learning rate. Both VGG and ResNet architectures achieved good results, outperforming Inception as shown in Tables 4 and 5. In successive analyses we thus restricted to use VGG and ResNet as feature extractors and validated performance and features with the DAP. The same process was adopted on HINT10 and HINT20. An experiment with 30 tissues has also been performed. Results are listed in S5 Table.

Experiments on KIMIA24

In the second experiment, we used VGG on the KIMIA24 dataset with the deep features extracted by VGG on GTEx; the task is the identification of the slide of origin (24 classes). In the DAPPER framework, classifiers were trained on 1, 060 annotated tiles and validated on 265 unseen ones.

UMAP analysis

In order to perform an unsupervised exploration of the features extracted by the Feature Extractor module, we projected the deep features onto a bi-dimensional space by using the Uniform Manifold Approximation and Projection (UMAP) multidimensional projection method. This dimension reduction technique, which relies on topological descriptors, has proven competitive with state-of-the-art visualization algorithms such as t-SNE [57], preserving both global and local structure of the data [58, 59]. We used the *R umap* package with

Table 3. Summary of experiments with the backend architectures.

Experiment	Dataset	Feature extractor	Version/Model
VGG-5	HINT5	VGG	Net-E+BN
ResNet-5	HINT5	ResNet	152-layer
Inception-5	HINT5	Inception	3
VGG-10	HINT10	VGG	Net-E+BN
ResNet-10	HINT10	ResNet	152-layer
VGG-20	HINT20	VGG	Net-E+BN
ResNet-20	HINT20	ResNet	152-layer

https://doi.org/10.1371/journal.pcbi.1006269.t003

Table 4. Matthew correlation coefficient values for each experiment, and classifier head pairs on HINT dataset. The average cross validation MCC with 95% CI (**H-MCCt**), and MCC on the external validation set (**H-MCCv**) are reported. Best-performing backend network, and classifier head combination on each dataset are reported in bold.

	FCH		SVM		RF	
Experiment	**H-MCCt**	**H-MCCv**	**H-MCCt**	**H-MCCv**	**H-MCCt**	**H-MCCv**
VGG-5	0.841 (0.838, 0.843)	0.820	0.786 (0.783, 0.789)	0.777	0.750 (0.748, 0.753)	0.747
ResNet-5	**0.879 (0.877, 0.881)**	**0.883**	0.852 (0.850, 0.854)	0.840	0.829 (0.827, 0.832)	0.849
Inception-5			0.747 (0.744, 0.750)	0.734	0.703 (0.699, 0.706)	0.701
VGG-10	**0.896 (0.894, 0.897)**	**0.894**	0.861 (0.859, 0.862)	0.866	0.889 (0.888, 0.891)	0.886
ResNet-10	0.857 (0.856, 0.859)	0.860	0.825 (0.824, 0.827)	0.832	0.845 (0.843, 0.847)	0.850
VGG-20	0.771 (0.770, 0.772)	0.774	0.729 (0.727, 0.730)	0.731	0.761 (0.760, 0.762)	0.766
ResNet-20	0.756 (0.754, 0.757)	0.757	**0.788 (0.787, 0.789)**	**0.792**	0.738 (0.737, 0.739)	0.738

https://doi.org/10.1371/journal.pcbi.1006269.t004

following parameters: `n_neighbors` = 40, `min_dist` = 0.01, `n_components` = 2, and Euclidean `metric`.

Implementation

All the code of the DAPPER framework is written in *Python* (`v3.6`) and *R* (`v3.4.4`). In addition to the general scientific libraries for Python, the scripts for the creation and training of the networks are based on *PyTorch*; the backend networks are implemented in *torchvision*. The library for processing histological images (available at `gitlab.fbk.eu/mpba-histology/histolib`) is based on *OpenSlide* and *scikit-image*.

The computations were performed on Microsoft Azure Virtual Machines with 4 NVIDIA K80 GPUs, 24 Intel Xeon E5-2690 cores and 256 GB RAM.

Results

Results of the tissue classification tasks in the DAPPER framework are listed in Table 4 for Matthews Correlation Coefficient (MCC) and Table 5 for Accuracy (ACC), respectively. See also Fig 3 for a comparison of MCC in internal cross validation with external validation.

All backend network-head pairs on HINT have MCC$>$ 0.7 with narrow CIs, with estimates from internal validation close to performance on the external validation set (Fig 3). Agreement of internal estimates with values on external validation set is a good indicator of generalization and potential for reproducibility. All models reached their top MCC accuracy with 1, 000 features. On HINT5 and HINT10, the FCH neural network performs better than SVMs and RF.

Table 5. Accuracy values for each experiment, and classifier head pairs on HINT dataset. The average cross validation ACC with 95% CI and ACC on the external validation set are reported. Best-performing backend network, and classifier head combination on each dataset are reported in bold.

	FCH		SVM		RF	
Experiment	**H-ACCt**	**H-ACCv**	**H-ACCt**	**H-ACCv**	**H-ACCt**	**H-ACCv**
VGG-5	87.2 (87.0, 87.5)	85.6	82.9 (82.7, 83.1)	82.1	79.9 (79.7, 80.1)	79.7
ResNet-5	**90.3 (90.1, 90.5)**	**90.7**	88.1 (88.0, 88.3)	87.2	86.3 (86.1, 86.5)	87.9
Inception-5			79.8 (79.5, 80.0)	78.7	76.2 (75.9, 76.4)	75.9
VGG-10	**90.6 (90.5, 90.7)**	**90.5**	87.5 (87.3, 87.6)	88.0	90.0 (89.9, 90.2)	89.7
ResNet-10	87.2 (87.0, 87.3)	87.4	84.3 (84.1, 84.4)	84.9	86.1 (85.9, 86.2)	86.5
VGG-20	78.2 (78.1, 78.4)	78.5	74.1 (74.0, 74.2)	74.4	77.3 (77.2, 77.4)	77.7
ResNet-20	76.7 (76.6, 76.9)	76.9	**79.9 (79.8, 80.0)**	**80.3**	75.1 (75.0, 75.2)	75.2

https://doi.org/10.1371/journal.pcbi.1006269.t005

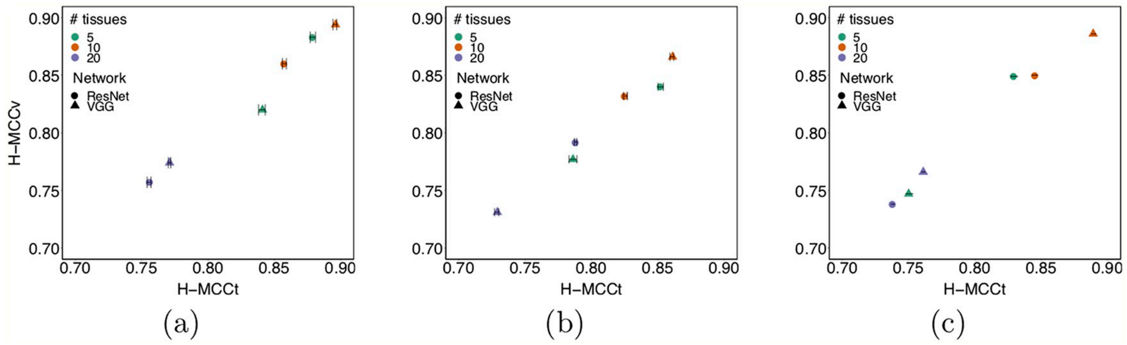


Fig 3. Comparison of DAPPER cross validation MCC (H-MCCt), vs MCC on external validation (H-MCCv) performance for each classifier. (a) FCH; (b) SVM; (c) RF.

https://doi.org/10.1371/journal.pcbi.1006269.g003

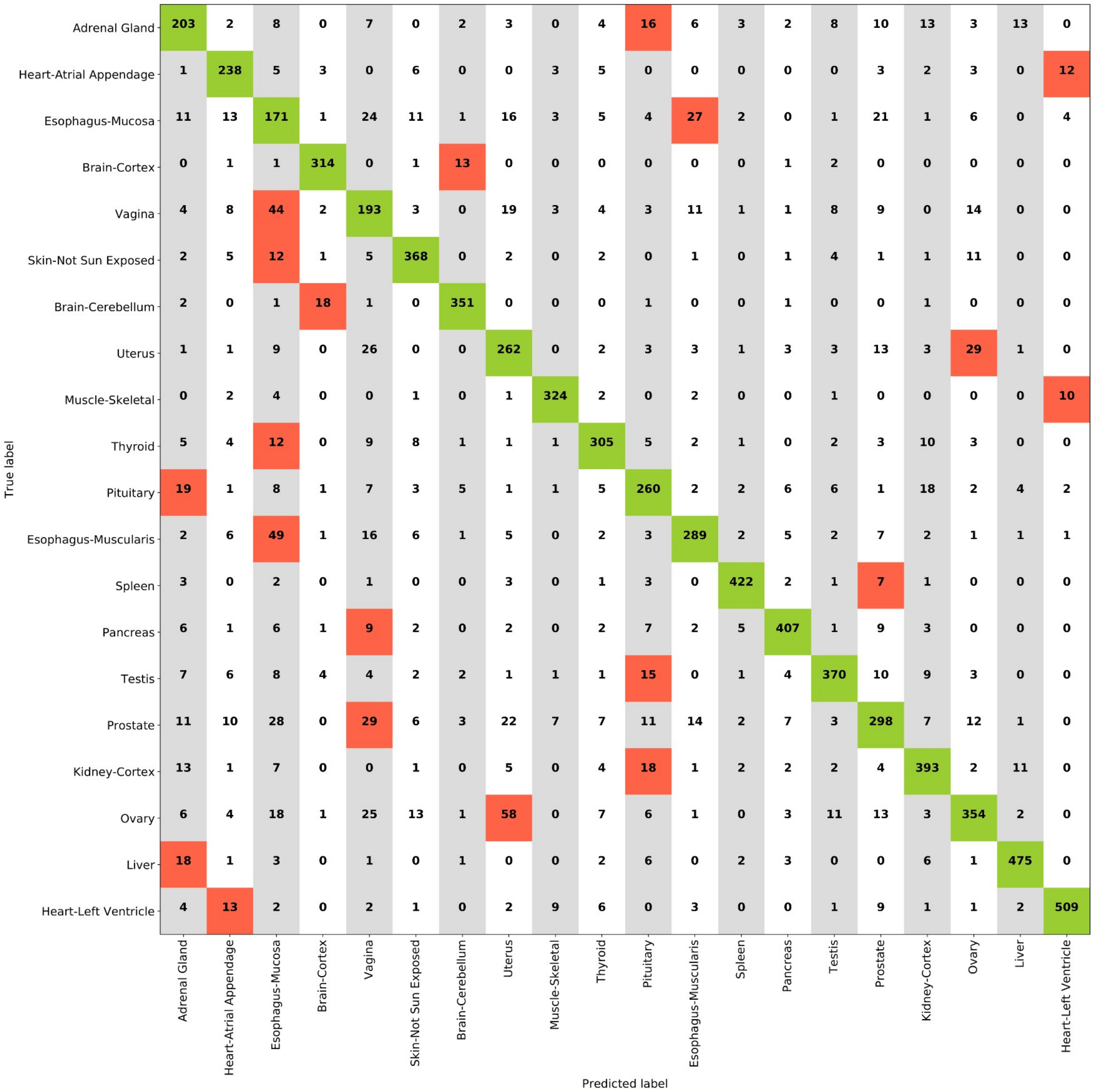
As expected, MCC ranged close to 0 for random labels; random ranking for increasing feature set sizes reached top MCC only for all features (tested for SVMs).

The most accurate models both for internal and external validation estimates were the ResNet+FCH model with MCC = 0.883 on HINT5, the VGG+FCH model on HINT10, and the ResNet+SVM model on HINT20. In S6 Table we show the results with a lower number of dense layers in the FCH, which are comparable with the FCH with 4 dense layers. Results on HINT30 are detailed in S5 Table; on external validation set, the VGG model reaches accuracy ACC = 61.8% and MCC = 0.61. Performance decreases for more complex multiclass problems. Notably the difficulty of the task is also complicated by tissue classes that are likely to have similar histological patterns, such as misclassification of Esophagus-Muscularis (ACC: 72.1%) with Esophagus-Mucosa (ACC: 53.2%), or the two Heart tissue subtypes or the 58 Ovary (ACC: 68.3%) tiles predicted as Uterus (ACC: 72.8%). The full confusion matrix for ResNet with SVMs on HINT20 is reported in Fig 4. In this paper we establish a methodology to evaluate reproducibility and predictive accuracy of machine learning models, in particular of the model selection phase. This is obtained by moving from single training-test split procedure to an evaluation environment that uses data replicates and averaged statistical indicators, thus enabling to select a model on the basis of statistical indicators derived from the internal validation loop. In this framework, we can honestly evaluate model performance differences along a set of experiments on a group of tasks. The DAPPER framework cannot by itself identify the reason of such difference, and indeed the emergence of optimal architectures for a specific task may be due to different factors, as revealed by appropriate experimental design. In terms of the experimental design described in this paper, for any model type we expect and find a decrease in accuracy for increasing number of classes, which requires learning more decision surfaces with less data per class. Notably, the best model in the internal DAPPER validation is confirmed to be the best also on the unseen test data, with a value within the confidence interval or immediately close.

Results on KIMIA24

Regardless of difference in image types, VGG-KIMIA24 with both RF and SVM heads with ACC = 43.4% (see Table 6), improving on published results (ACC = 41.8% [38]).

It is worth noting that transfer learning from ImageNet to HINT restricts training to the Adapter and Fully Connected Head blocks. In one-shot experiments, MCC further improves when the whole feature extraction block is retrained (see S3 Table). However, the result still needs to be consolidated by extending the DAP also to the training or retraining of the deep



	Adrenal Gland	Heart-Atrial Appendage	Esophagus-Mucosa	Brain-Cortex	Vagina	Skin-Not Sun Exposed	Brain-Cerebellum	Uterus	Muscle-Skeletal	Thyroid	Pituitary	Esophagus-Muscularis	Spleen	Pancreas	Testis	Prostate	Kidney-Cortex	Ovary	Liver	Heart-Left Ventricle
Adrenal Gland	203	2	8	0	7	0	2	3	0	4	16	6	3	2	8	10	13	3	13	0
Heart-Atrial Appendage	1	238	5	3	0	6	0	0	3	5	0	0	0	0	0	3	2	3	0	12
Esophagus-Mucosa	11	13	171	1	24	11	1	16	3	5	4	27	2	0	1	21	1	6	0	4
Brain-Cortex	0	1	1	314	0	1	13	0	0	0	0	0	0	1	2	0	0	0	0	0
Vagina	4	8	44	2	193	3	0	19	3	4	3	11	1	1	8	9	0	14	0	0
Skin-Not Sun Exposed	2	5	12	1	5	368	0	2	0	2	0	1	0	1	4	1	1	11	0	0
Brain-Cerebellum	2	0	1	18	1	0	351	0	0	0	1	0	0	1	0	0	1	0	0	0
Uterus	1	1	9	0	26	0	0	262	0	2	3	3	1	3	3	13	3	29	1	0
Muscle-Skeletal	0	2	4	0	0	1	0	1	324	2	0	2	0	0	1	0	0	0	0	10
Thyroid	5	4	12	0	9	8	1	1	1	305	5	2	1	0	2	3	10	3	0	0
Pituitary	19	1	8	1	7	3	5	1	1	5	260	2	2	6	6	1	18	2	4	2
Esophagus-Muscularis	2	6	49	1	16	6	1	5	0	2	3	289	2	5	2	7	2	1	1	1
Spleen	3	0	2	0	1	0	0	3	0	1	3	0	422	2	1	7	1	0	0	0
Pancreas	6	1	6	1	9	2	0	2	0	2	7	2	5	407	1	9	3	0	0	0
Testis	7	6	8	4	4	2	2	1	1	1	15	0	1	4	370	10	9	3	0	0
Prostate	11	10	28	0	29	6	3	22	7	7	11	14	2	7	3	298	7	12	1	0
Kidney-Cortex	13	1	7	0	0	1	0	5	0	4	18	1	2	2	2	4	393	2	11	0
Ovary	6	4	18	1	25	13	1	58	0	7	6	1	0	3	11	13	3	354	2	0
Liver	18	1	3	0	1	0	1	0	0	2	6	0	2	3	0	0	6	1	475	0
Heart-Left Ventricle	4	13	2	0	2	1	0	2	9	6	0	3	0	0	1	9	1	1	2	509

Fig 4. Confusion matrix for ResNet+SVM model on HINT20. Red shaded cells indicate the most confused classes.

https://doi.org/10.1371/journal.pcbi.1006269.g004

learning backend networks to check for actual generalization. The Canberra stability indicator was also computed for all the experiments, with minimal median stability for ResNet-20 (Fig 5).

Results at WSI-level

We evaluated the performance of DAPPER at WSI-level on the HINT20 external validation set, with the ResNet+SVM model. In particular, all the predictions for the tiles are aggregated

Table 6. Performance of DAPPER framework for VGG backend network, and classifier heads (FCH, SVM, RF) on KIMIA24 dataset. The average cross validation MCC (**K24-MCCt**), and ACC (**K24-ACCt**) with 95% CI, as well as MCC (**K24-MCCv**), and ACC (**K24-ACCv**) on external validation set are reported.

Model	K24-MCCt	K24-MCCv	K24-ACCt	K24-ACCv
VGG+FCH	0.317 (0.306, 0.327)	0.207	34.4 (33.2, 35.2)	23.8
VGG+SVM	0.446 (0.439, 0.454)	0.409	47.1 (46.4, 47.8)	43.4
VGG+RF	**0.457 (0.449, 0.465)**	**0.409**	**48.0 (47.3, 48.8)**	**43.4**

https://doi.org/10.1371/journal.pcbi.1006269.t006

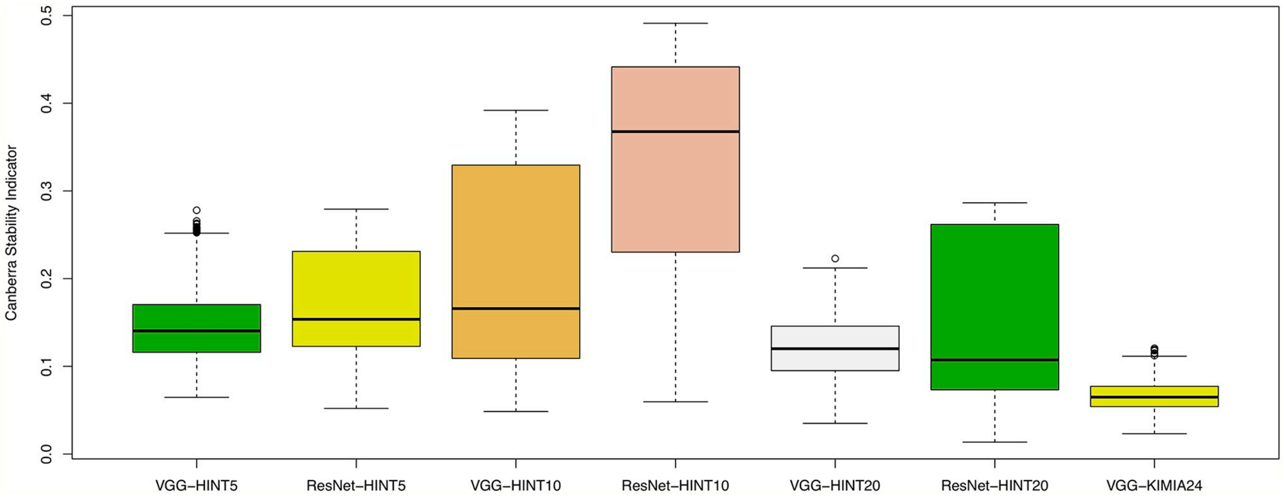


Fig 5. Canberra stability indicator on HINT and KIMIA datasets. For each architecture, a set of deep feature lists is generated, one list for each internal run of training in the nested cross validation schema, each ranked with KBest. Canberra stability is computed as in [32]: lower stability is better.

https://doi.org/10.1371/journal.pcbi.1006269.g005

by WSI, and the resulting tissue will be the most common one among those predicted on the corresponding tiles. However, it is worth noting that the number of tiles per WSI in the HINT20 external validation set varies (min 1, max 31) due to a stratification strategy only considering the tissues-per-sample distribution (see Section *Data Analysis Plan*). Therefore, we restricted our evaluation to a subset of 15 WSI per class (300 WSI in total), each of which associated to 10 tiles randomly selected. This value represents a reasonable number of Regions of Interest (ROIs) a human pathologist would likely consider in his/her evaluations. In this regard, we further investigate how the DAPPER framework performs on an increasing number of tiles per WSI, namely 3, 5, 7, and 10. As expected, the overall accuracy improves as the number of tiles per WSI increases, reaching 98.3% when considering all 10 tiles per WSI. Notably, the accuracy is high even when reducing to 3 tiles per WSI (see Table 7).

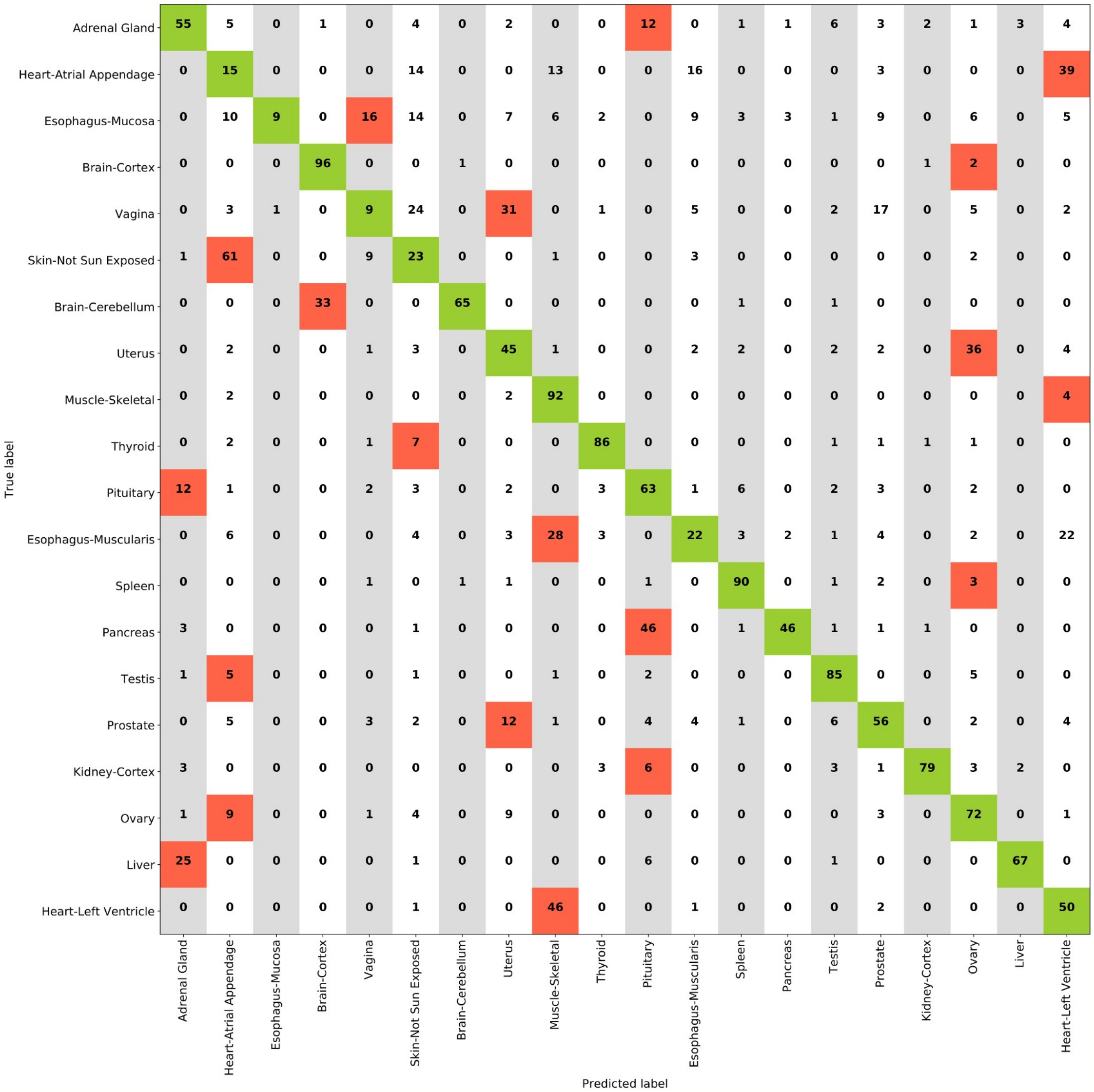
Comparison with pathologist

We tested the performance of DAPPER against an expert pathologist on about 25% of the HINT20 external validation set, 2, 000 tiles out of 8, 103, with 100 randomly selected tiles for each class. We asked the pathologist to classify each tile by choosing among the 20 classes of the HINT20 dataset, without imposing any time constraint. The confusion matrix resulting from the evaluation of tiles as produced by the pathologist is shown in Fig 6. Predictions produced by the DAPPER framework for comparative results are then collected on the same data. The best-performing model on the HINT20 dataset, namely the ResNet+SVM model, has

Table 7. Metrics at WSI-level for increasing number of tiles per WSI. Metrics are computed on a subset of HINT20 external validation set, consisting of 15 WSI per class (300 WSI in total). The WSI class is determined by the most frequent predicted class by the *ResNet+SVM* model for the considered tiles.

# Tiles per WSI	MCC	ACC (%)
3	0.86	86.3
5	0.93	93.7
7	0.96	96.0
10	0.98	98.3

https://doi.org/10.1371/journal.pcbi.1006269.t007



	Adrenal Gland	Heart-Atrial Appendage	Esophagus-Mucosa	Brain-Cortex	Vagina	Skin-Not Sun Exposed	Brain-Cerebellum	Uterus	Muscle-Skeletal	Thyroid	Pituitary	Esophagus-Muscularis	Spleen	Pancreas	Testis	Prostate	Kidney-Cortex	Ovary	Liver	Heart-Left Ventricle
Adrenal Gland	55	5	0	1	0	4	0	2	0	0	12	0	1	1	6	3	2	1	3	4
Heart-Atrial Appendage	0	15	0	0	0	14	0	0	13	0	0	16	0	0	0	3	0	0	0	39
Esophagus-Mucosa	0	10	9	0	16	14	0	7	6	2	0	9	3	3	1	9	0	6	0	5
Brain-Cortex	0	0	0	96	0	0	1	0	0	0	0	0	0	0	0	0	1	2	0	0
Vagina	0	3	1	0	9	24	0	31	0	1	0	5	0	0	2	17	0	5	0	2
Skin-Not Sun Exposed	1	61	0	0	9	23	0	0	1	0	0	3	0	0	0	0	0	2	0	0
Brain-Cerebellum	0	0	0	33	0	0	65	0	0	0	0	0	1	0	1	0	0	0	0	0
Uterus	0	2	0	0	1	3	0	45	1	0	0	2	2	0	2	2	0	36	0	4
Muscle-Skeletal	0	2	0	0	0	0	0	2	92	0	0	0	0	0	0	0	0	0	0	4
Thyroid	0	2	0	0	1	7	0	0	0	86	0	0	0	0	1	1	1	1	0	0
Pituitary	12	1	0	0	2	3	0	2	0	3	63	1	6	0	2	3	0	2	0	0
Esophagus-Muscularis	0	6	0	0	0	4	0	3	28	3	0	22	3	2	1	4	0	2	0	22
Spleen	0	0	0	0	1	0	1	1	0	0	1	0	90	0	1	2	0	3	0	0
Pancreas	3	0	0	0	0	1	0	0	0	0	46	0	1	46	1	1	1	0	0	0
Testis	1	5	0	0	0	1	0	0	1	0	2	0	0	0	85	0	0	5	0	0
Prostate	0	5	0	0	3	2	0	12	1	0	4	4	1	0	6	56	0	2	0	4
Kidney-Cortex	3	0	0	0	0	0	0	0	0	3	6	0	0	0	3	1	79	3	2	0
Ovary	1	9	0	0	1	4	0	9	0	0	0	0	0	0	0	3	0	72	0	1
Liver	25	0	0	0	0	1	0	0	0	0	6	0	0	0	1	0	0	0	67	0
Heart-Left Ventricle	0	0	0	0	0	1	0	0	46	0	0	1	0	0	0	2	0	0	0	50

Fig 6. Confusion matrix for pathologist classification on a subset of HINT20 external validation set. Red shaded cells indicate the most confused classes.

https://doi.org/10.1371/journal.pcbi.1006269.g006

been considered for this experiment. As reported in Table 8, DAPPER outperforms the pathologist in the prediction of tissues at a tile-level.

To provide an unbiased estimation of the performance of DAPPER, we repeated the same evaluation on 10 other randomly generated subsets of 2, 000 tiles extracted from the HINT20 external validation set. The obtained average MCC and ACC with 95% CI are 0.786 (0.783, 0.789), and 79.6 (79.3, 79.9), respectively.

Table 8. Tissue classification performance of DAPPER vs pathologist. DAPPER with *ResNet+SVM* model outperforms the pathologist at tile-level. Metrics are computed on a subset of HINT20 external validation set (2, 000 tiles).

Classifier	MCC	ACC (%)
Pathologist	0.542	56.3
DAPPER	0.786	79.6

https://doi.org/10.1371/journal.pcbi.1006269.t008

Finally, since the classification at tile-level is an unusual task for a pathologist, who is instead trained on examining the whole context of a tissue scan, as a second task we asked the pathologist to classify 200 randomly chosen WSIs (10 for each class of HINT20). As expected, the results in this case are better than those at tile-level, *i.e.*, MCC = 0.788, and ACC = 79.5%, to be compared with the DAPPER performances reported in Table 7.

The HINT benchmark dataset

As a second contribution of this study, we are making available the HINT dataset, generated by the first component of tools in the DAPPER framework, as a benchmark dataset for

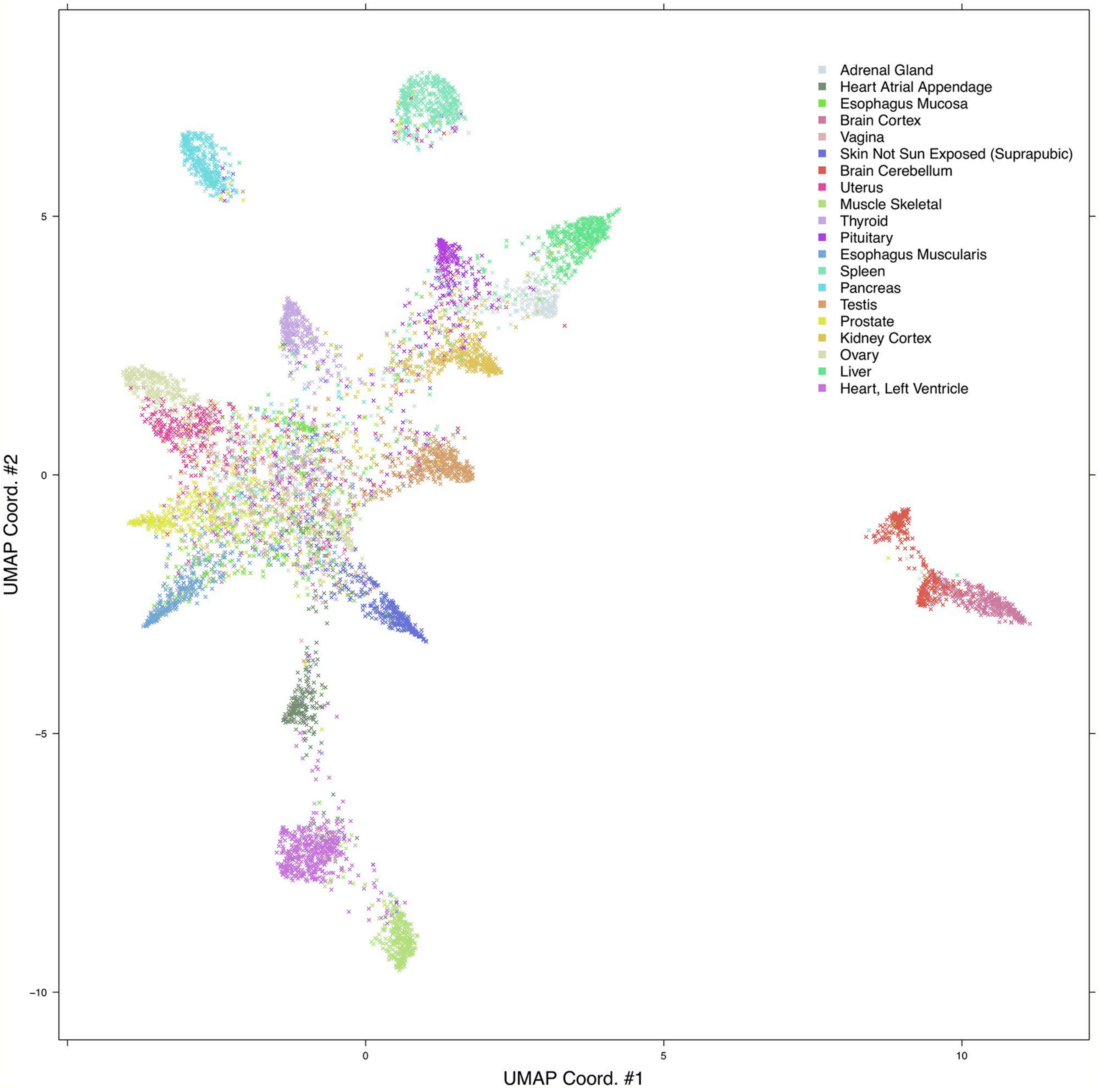


Fig 7. UMAP projection of external validation set for VGG-20 experiment.

https://doi.org/10.1371/journal.pcbi.1006269.g007

validating machine learning models in digital pathology. The HINT dataset is currently composed of 53, 727 tiles at 512 × 512 resolution, based on histology from GTEx. HINT can be easily expanded to over 78, 000 tiles, as for this study we used a fraction of the GTEx images and at most 100 tiles from each WSI were extracted. Digital pathology still misses a universally adopted dataset to compare deep learning models as already established in vision (*e.g.*, ImageNet for image classification, COCO for image and instance segmentation). Several initiatives for a "BioImageNet" will eventually improve this scenario. Histology data are available in the generalist repository Image Data Resource (IDR) [60, 61]. Further, the International Immuno-Oncology Biomarker Working Group in Breast Cancer and the MAQC Society have launched a collaborative project to develop data resources and quality control schemes on Machine Learning algorithms to assess TILs in Breast Cancer.

HINT is conceptually similar to KIMIA24. However, HINT inherits from GTEx more variability in terms of sample characteristics, validation of donors and additional access to molecular data. Further, we used a random sampling approach to process tiles excluding background and minimize human intervention in the choice and preparation of the images.

Deep features

We applied an unsupervised projection on all the features extracted by VGG and ResNet networks on all tissues tasks. In the following, we discuss an example for features extracted by VGG on the HINT20 task, displayed as UMAP projection (Fig 7), points are coloured for 20 tissue labels. The UMAP displays for the other tasks are available in S1–S4 Figs. The UMAP display is in agreement with the count distributions in the confusion matrix (Fig 4). The deep learning embedding separates well a set of histology types, including Muscle-Skeletal, Spleen, Pancreas, Brain-Cortex and Cerebellum, Heart-Left Ventricle and Atrial Appendage which group into distinct clusters (See Fig 7 and Table 9). The distributions of the activations for the top-3 deep features of the VGG backend network on the HINT10 dataset are displayed in S5 Fig; the top ranked deep feature (`#668`) is clearly selective for Spleen. The UMAP projection also shows an overlapping for tissues such as Ovary and Uterus, or Vagina and Esophagus-Mucosa, or the two Esophagus histotypes, consistently with the confusion matrix (Fig 4).

Examples of five tiles from two well separated clusters, Muscle-Skeletal (ACC: 93.4%) and Spleen (ACC: 94.6%), are displayed in panel A of Fig 8. Tiles from three clusters partially overlapping in the neural embedding and mislabeled in both the VGG-20 and ResNet-20 embeddings with SVM (Esophagus- Mucosa ACC = 53.2%, Esophagus-Muscularis ACC = 72.1%, Vagina ACC = 59.0%) are similarly visualized in Fig 8B. While the aim of this paper is to introduce a framework for honest comparison of models that will be used for clinical purposes rather than fine-tuning accuracy in this experiment, it is evident that these tiles have

Table 9. Histology types well separated by SVM+ResNet model for HINT20. Accuracy is computed with respect to the confusion matrix in Fig 4 and expressed in percentage, together with the total number of samples for each class.

Histology type	ACC(%)	#samples
Spleen	94.6	446
Brain—Cortex	94.3	333
Muscle—Skeletal	93.4	347
Brain—Cerebellum	93.4	376
Heart—Left Ventricle	90.1	565
Pancreas	87.9	463
Heart—Atrial Appendage	84.7	317

https://doi.org/10.1371/journal.pcbi.1006269.t009

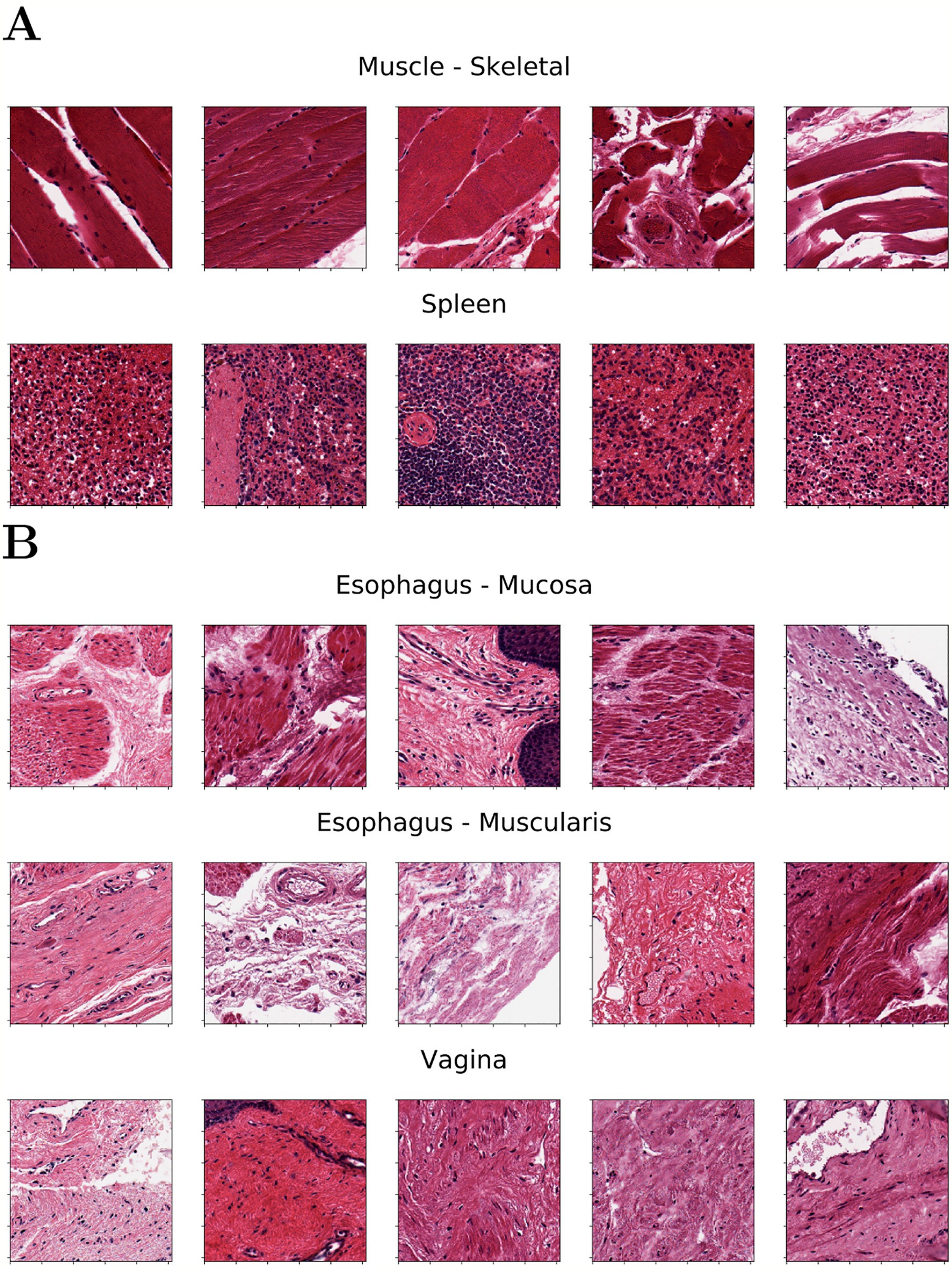


Fig 8. Representative tiles predicted from VGG-20 experiment. A) Examples from two well-separated clusters observed in the UMAP embedding. B) Samples of mislabeled tiles from tissues partially overlapping in the UMAP embedding.

https://doi.org/10.1371/journal.pcbi.1006269.g008

morphologies that are hard to classify. This challenge requires more complex models (*e.g.* ensembles) and a structured output labeling, already applied in dermatology [2]. Further, we are exploring the combination of DAPPER with image analysis packages, such as HistomicsTK (https://digitalslidearchive.github.io/HistomicsTK/) or CellProfiler [62], to extract features useful for interpretation and feedback from pathologists.

Discussion

Digital pathology would greatly benefit from the adoption of machine learning, shifting human assessment of histology to higher quality, non-repetitive tasks. Unfortunately, there is no fast, easy route to improve reproducibility of automated analysis. The adoption of the DAP clearly sets in a computational aggravation not usually considered for image processing exercises. However, this is an established practice with massive omics data [28], and reproducibility by design can handle secondary results useful for diagnostics and for interpretation.

We designed the DAPPER framework as a tool for evaluating accuracy and stability of deep learning models, currently only backend elements in a sequence of processing steps, and possibly in the future end-to-end solutions. We choose as test domain H&E stained WSIs for prediction of tissue of origin, which is not a primary task for trained pathologists, but a reasonable benchmark for machine learning methods. Also, we are aware that tissue classification is only a step in real digital pathology applications. Mobadersany and colleagues [11] used a deep learning classifier to score and visualize risk on the WSIs. Similarly, deep learning tile classification may be applied to quantify histological differences in association to a genomic pattern, *e.g.*, a specific mutation or a high-dimensional protein expression signature. In this vision, the attention to model selection supported by our framework is a prerequisite for developing novel AI algorithms for digital pathology, *e.g.*, for analytics over TILs.

Although we are building on deep learning architectures known for applications on generic images, they adapted well to WSIs in combination with established machine learning models (SVM, RF); we expect that large scale bioimaging resources will give the chance of improving the characterization of deep features, as already emerged with the HINT dataset that we are providing as public resource. In this direction, we plan to release the network weights of the backend DAPPER models that are optimized for histopathology as alternative pretrained weights for digital pathology, similarly to those for the ImageNet dataset and available in `torchvision`.

Supporting information

S1 Table. Summary of available samples, downloaded WSIs and extracted tiles.
(PDF)

S2 Table. Summary of the datasets.
(PDF)

S3 Table. Impact of retraining the backend network. Accuracy and Matthews Correlation Coefficient improve when retraining also the feature extraction block (VGG backend network, not in DAP). We observe an improvement of the accuracy from 5.5% to 24.8% for the four chosen experiments. Possibly the neural network benefits from adjusting also the initial weights because the layers learn characteristics of the images diverse from the ImageNet dataset.
(PDF)

S4 Table. Comparison of the three optimization methods to set the learning rate. The best method for setting the learning rate was assessed using the VGG as backend network on the 5

tissues dataset HINT5. Three methods were tested: Fixed (FIX): the learning rate is set to 10^{-5} for the whole training; Step-wise (STEP): the learning rate is initialized at $\lambda_{init} = 10^{-3}$ and updated every 10 epochs with the following rule: $\lambda_{new} = \lambda_{old}/10$; Polynomial (POLY): the learning rate is initialized at 10^{-3} and updated every 10 iterations with a polynomial law: $\lambda_{\text{new}} = \lambda_{\text{init}}\left(1 - \frac{i}{I_{\max}}\right)^{0.9}$, where i is the index of the iteration and I_{max} is the total number of iterations.
(PDF)

S5 Table. Impact of task complexity (VGG backend network). Performance decreases when the number of tissues increases. Adding more classes to the task is possibly complicated by the introduction of tissues with similar histological patterns.
(PDF)

S6 Table. Impact (MCC) of number of internal layers on FCH ($<$ 4 dense layers) on HINT dataset. FCH3: three dense layers with 1000, 256 and # tissue classes nodes, respectively; FCH2: two dense layers with 256 and # tissue classes nodes, respectively. The average cross validation MCC with 95% CI (H-MCCt), and MCC on the external validation set (H-MCCv) are reported. In bold: MCC (bold) values of Table 4 of the main text.
(PDF)

S1 Fig. UMAP projection on training (circles) and external validation (crosses) set for VGG-5 experiment.
(PNG)

S2 Fig. UMAP projection on training (circles) and external validation (crosses) set for ResNet-5 experiment.
(PNG)

S3 Fig. UMAP projection on training (circles) and external validation (crosses) set for VGG-10 experiment.
(PNG)

S4 Fig. UMAP projection on training (circles) and external validation (crosses) set for ResNet-10 experiment.
(PNG)

S5 Fig. Deep features and tissue of origin. Distributions of the values of the top-3 deep features computed with the VGG backend architecture for the 10 classes of the HINT10 dataset.
(PDF)

Acknowledgments

We derived the HINT dataset from the Genotype-Tissue Expression (GTEx) Project, supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS (data downloaded from the GTEx Portal on *05/10/18*).

The authors thank Intel Italy for technical support and availability of high performance computing resources. We also thank H. Tizhoosh for availability of KIMIA Path24 dataset, L. Coviello for his help in the networks' optimization and G. Franch for the realization of the striking image.

Author Contributions

Conceptualization: Margherita Francescatto, Cesare Furlanello.

Data curation: Andrea Bizzego, Nicole Bussola.

Funding acquisition: Cesare Furlanello.

Investigation: Margherita Francescatto, Luca Cima, Marco Cristoforetti.

Methodology: Andrea Bizzego, Nicole Bussola, Marco Chierici, Valerio Maggio, Margherita Francescatto, Marco Cristoforetti, Giuseppe Jurman, Cesare Furlanello.

Project administration: Cesare Furlanello.

Software: Andrea Bizzego, Nicole Bussola, Marco Chierici, Valerio Maggio, Marco Cristoforetti, Giuseppe Jurman.

Supervision: Cesare Furlanello.

Visualization: Nicole Bussola, Marco Chierici, Giuseppe Jurman.

Writing – original draft: Andrea Bizzego, Nicole Bussola, Luca Cima, Cesare Furlanello.

Writing – review & editing: Luca Cima, Cesare Furlanello.

References

1. Lu L, Zheng Y, Carneiro G, Yang L. Deep Learning and Convolutional Neural Networks for Medical Image Computing. Springer; 2017.
2. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542(7639):115–118. https://doi.org/10.1038/nature21056 PMID: 28117445
3. Komeda Y, Handa H, Watanabe T, Nomura T, Kitahashi M, Sakurai T, et al. Computer-Aided Diagnosis Based on Convolutional Neural Network System for Colorectal Polyp Classification: Preliminary Experience. Oncology. 2017; 93(Suppl. 1):30–34. https://doi.org/10.1159/000481227 PMID: 29258081
4. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep Learning for Classification of Colorectal Polyps on Whole-Slide Images. Journal of Pathology Informatics. 2017; 8:30. https://doi.org/10.4103/jpi.jpi_34_17 PMID: 28828201
5. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine. 2018; 24(9):1342–1350. https://doi.org/10.1038/s41591-018-0107-6 PMID: 30104768
6. Ciompi F, Chung K, Van Riel SJ, Setio AAA, Gerke PK, Jacobs C, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Scientific Reports. 2017; 7:46479. https://doi.org/10.1038/srep46479 PMID: 28422152
7. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017; 42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
8. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digital Medicine. 2018; 1(1):40. https://doi.org/10.1038/s41746-018-0048-y
9. Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, Scholz P, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Communications. 2018; 9(1):5217. https://doi.org/10.1038/s41467-018-07619-7 PMID: 30523263
10. Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, Scholz P, et al. Author Correction: Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Communications. 2019; 10(1):588. https://doi.org/10.1038/s41467-019-08563-w PMID: 30700735
11. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences. 2018; 115(13):E2970–E2979. https://doi.org/10.1073/pnas.1717139115

12. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. Scientific Reports. 2018; 8(1):3395. https://doi.org/10.1038/s41598-018-21758-3 PMID: 29467373

13. Sharma H, Zerbe N, Klempert I, Hellwich O, Hufnagl P. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. Computerized Medical Imaging and Graphics. 2017; 61:2–13. https://doi.org/10.1016/j.compmedimag.2017.06.001 PMID: 28676295

14. Paeng K, Hwang S, Park S, Kim M. A unified framework for tumor proliferation score prediction in breast histopathology. In: Proceedings of the Third International Workshop on Deep Learning in Medical Image Analysis (DLMIA 2017) and the Sixth International Workshop on Multimodal Learning for Clinical Decision Support (ML-CDS 2017), held in conjunction with the Twentieth International Conference on Medical Imaging and Computer-Assisted Intervention (MICCAI 2017). Springer; 2017. p. 231–239.

15. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification And Mutation Prediction From Non-Small Cell Lung Cancer Histopathology Images Using Deep Learning. Nature Medicine. 2018; 24(10):1559–1567. https://doi.org/10.1038/s41591-018-0177-5 PMID: 30224757

16. Coudray N, Moreira AL, Sakellaropoulos T, Fenyö D, Razavian N, Tsirigos A. Determining EGFR and STK11 mutational status in lung adenocarcinoma histopathology images using deep learning. Cancer Research. 2018; 78(Supp.13):5309–5309. https://doi.org/10.1158/1538-7445.AM2018-5309

17. Basavanhally A, Ganesan S, Feldman M, Shih N, Mies C, Tomaszewski J, et al. Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides. IEEE Transactions on Biomedical Engineering. 2013; 60:2089–2099. https://doi.org/10.1109/TBME.2013.2245129 PMID: 23392336

18. Denkert C, Wienert S, Poterie A, Loibl S, Budczies J, Badve S, et al. Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group. Modern Pathology. 2016; 29(10):1155–1164. https://doi.org/10.1038/modpathol.2016.109 PMID: 27363491

19. Mina M, Boldrini R, Citti A, Romania P, D'Alicandro V, De Ioris M, et al. Tumor-infiltrating T lymphocytes improve clinical outcome of therapy-resistant neuroblastoma. Oncoimmunology. 2015; 4(9):e1019981. https://doi.org/10.1080/2162402X.2015.1019981 PMID: 26405592

20. Salgado R, Sherene L. Tumour infiltrating lymphocytes in breast cancer: increasing clinical relevance. The Lancet Oncology. 2018; 19(1):3–5. https://doi.org/10.1016/S1470-2045(17)30905-1 PMID: 29233560

21. Stovgaard ES, Nielsen D, Hogdall E, Balslev E. Triple negative breast cancer–prognostic role of immune-related factors: a systematic review. Acta Oncologica. 2018; 57(1):74–82. https://doi.org/10.1080/0284186X.2017.1400180 PMID: 29168430

22. Shibutani M, Maeda K, Nagahara H, Fukuoka T, Iseki Y, Matsutani S, et al. Tumor-infiltrating Lymphocytes Predict the Chemotherapeutic Outcomes in Patients with Stage IV Colorectal Cancer. In Vivo. 2018; 32(1):151–158. https://doi.org/10.21873/invivo.11218 PMID: 29275313

23. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. Imagenet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2009. p. 248–255.

24. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging. 2016; 35(5):1299–1312. https://doi.org/10.1109/TMI.2016.2535302 PMID: 26978662

25. Kieffer B, Babaie M, Kalra S, Tizhoosh HR. Convolutional Neural Networks for Histopathology Image Classification: Training vs. Using Pre-Trained Networks. In: Proceedings of the Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA 2017). IEEE; 2017. p. 1–6.

26. Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nature Genetics. 2009; 41(2):149–155. https://doi.org/10.1038/ng.295 PMID: 19174838

27. Baker M. 1,500 scientists lift the lid on reproducibility. Nature News. 2016; 533(7604):452. https://doi.org/10.1038/533452a

28. Shi L, Kusko R, Wolfinger RD, Haibe-Kains B, Fischer M, Sansone SA, et al. The international MAQC Society launches to enhance reproducibility of high-throughput technologies. Nature Biotechnology. 2017; 35(12):1127–1128. https://doi.org/10.1038/nbt.4029 PMID: 29220036

29. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016; 3:160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244

30. The MAQC Consortium. The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology. 2010; 28 (8):827–838. https://doi.org/10.1038/nbt.1665 PMID: 20676074

31. The SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. Nature Biotechnology. 2014; 32:903–914. https://doi.org/10.1038/nbt.2957 PMID: 25150838

32. Jurman G, Merler S, Barla A, Paoli S, Galea A, Furlanello C. Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics. 2008; 24(2):258–264. https://doi.org/10.1093/bioinformatics/btm550 PMID: 18024475

33. Furlanello C, Serafini M, Merler S, Jurman G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinformatics. 2003; 4(1):54. https://doi.org/10.1186/1471-2105-4-54 PMID: 14604446

34. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. In: Proceedings of the Thirteenth European Conference on Computer Vision (ECCV 2014). Springer; 2014. p. 740–755.

35. The GTEx Consortium. The genotype-tissue expression (GTEx) project. Nature Genetics. 2013; 45 (6):580–585. https://doi.org/10.1038/ng.2653 PMID: 23715323

36. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20(3):273–297. https://doi.org/10.1007/BF00994018

37. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition (ICDAR 1995). IEEE; 1995. p. 278–282).

38. Babaie M, Kalra S, Sriram A, Mitcheltree C, Zhu S, Khatami A, et al. Classification and Retrieval of Digital Pathology Scans: A New Dataset. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE; 2017. p. 8–16.

39. Kumar MD, Babaie M, Zhu S, Kalra S, Tizhoosh HR. A Comparative Study of CNN, BoVW and LBP for Classification of Histopathological Images. In: Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE; 2017. p. 1–7.

40. Alhindi TJ, Kalra S, Ng KH, Afrin A, Tizhoosh HR. Comparing LBP, HOG and Deep Features for Classification of Histopathology Images. In: Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN). IEEE; 2018. p. 1–7.

41. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreservation and Biobanking. 2015; 13 (5):311–319. https://doi.org/10.1089/bio.2015.0032 PMID: 26484571

42. Nader Vasconcelos C, Nader Vasconcelos B. Increasing deep learning melanoma classification by classical and expert knowledge based image transforms. arXiv. 2017;1702.07025.

43. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. AutoAugment: Learning Augmentation Policies from Data. arXiv. 2018;1805.09501.

44. Wang J, Perez L. The effectiveness of data augmentation in image classification using deep learning. arXiv. 2017;1712.04621.

45. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the Third International Conference on Learning Representations (ICLR 2015). arXiv:1409.1556; 2015. p. 1–14.

46. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. p. 770–778.

47. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. p. 2818–2826.

48. Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. arXiv. 2017;1605.076784:1–7.

49. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv. 2012;1207.0580:1–18.

50. Kinga D, Adam JB. Adam: A Method for Stochastic Optimization. In: Proceedings of the Third International Conference on Learning Representations (ICLR 2015). arXiv:1412.6980; 2014. p. 1–15.

51. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; 2009.

52. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta. 1975; 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9 PMID: 1180967

53. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000; 16(5):412–424. https://doi.org/10.1093/bioinformatics/16.5.412 PMID: 10871264

54. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. PLOS ONE. 2012; 7(8):e41882. https://doi.org/10.1371/journal.pone.0041882 PMID: 22905111

55. Chicco D. Ten quick tips for machine learning in computational biology. BioData Mining. 2017; 10:35. https://doi.org/10.1186/s13040-017-0155-3 PMID: 29234465

56. Di Ciccio TJ, Efron B. Bootstrap confidence intervals (with Discussion). Statistical Science. 1996; 11:189–228. https://doi.org/10.1214/ss/1032280214

57. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008; 9:2579–2605.

58. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018; 3(29):861. https://doi.org/10.21105/joss.00861

59. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology. 2018;Online:2018/12/03. https://doi.org/10.1038/nbt.4314 PMID: 30531897

60. Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, et al. Image Data Resource: a bioimage data integration and publication platform. Nature Methods. 2017; 14(8):775. https://doi.org/10.1038/nmeth.4326 PMID: 28775673

61. Nirschl JJ, Janowczyk A, Peyster EG, Frank R, Margulies KB, Feldman MD, et al. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. PLOS ONE. 2018; 13(4):e0192726. https://doi.org/10.1371/journal.pone.0192726 PMID: 29614076

62. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology. 2006; 7(10):R100. https://doi.org/10.1186/gb-2006-7-10-r100 PMID: 17076895