Computational Pathology for Tumor Histopathology
Version: CPATH 1.0
Guideline date: May 2021

Authors: Christof A. Bertram*, Marc Aubreville*, Taryn A. Donovan, Frances M. Moore, Robert Klopfleisch

**Denotes communication authors; all other contributing authors are listed alphabetically;* contact communication authors to suggest updates, provide edits and comments: Christof Bertram (christof.bertram@vetmeduni.ac.at) or Marc Aubreville (marc.aubreville@thi.de)

Recommended Citation: Bertram CA et al. Computational Pathology for Tumor Histopathology Guideline, version 1.0. Veterinary Cancer Guidelines and Protocols. http://vetcancerprotocols.org

Accessed on (date).

Contents

1. Introduction

Computational pathology (CPATH) is a branch of pathology using computerized methods to gather information on diseases in patients.[1] This Guideline gives guidance on methods for automated image analysis (AIA) of microscopic tumor images. A precondition of automated analysis is digitization of the tissue section (or parts thereof) and availability of such digital images is increasingly facilitated by the integration of whole slide image (WSI) scanners into routine laboratory workflows.[2] Compared to pathologists, computerized methods – if well-implemented – have the potential to improve reproducibility and accuracy of obtained data.[3] Therefore, computer-assisted analysis of prognostic parameters might lead to more meaningful prognostic information (see discussion).

Current state-of-the art methods are rapidly changing with ongoing research. As it is beyond the scope of this Guideline to address all available methods, we focus on important aspects for tumor prognostication that have been well-studied in the current literature. With further research, this document will be complemented with additional automated image analysis (AIA) approaches/prognostic parameters in future versions. First, we address some general considerations, then give recommendations for relevant broad categories and specific prognostic tasks. A list with definitions of relevant terms and important evaluation matrices is given in section 5 (Table 3) and section 6 (Table 4).

2. General considerations

There is a wide range of methods that can be applied for development of image analysis algorithms and depending on the methods used, there are general and specific aspects that need to be considered during development, performance evaluation, and application/implementation. Before starting to develop an automated image analysis solution, it has to be determined which of the available methods, or combination thereof, is best suited for the specific prognostication task and its intended use. Besides achieving the highest possible test performance, methods may be selected based on availability to labeled data, model flexibility, available computational power, tolerable processing time, ability to explain results ('black box' vs. 'glass box'), and algorithmic robustness (for definitions of terms see Table 3).

For decision of detailed methods, it may be helpful to categorize the prognostically relevant microscopic feature into the following main aspects (there may be overlap):

1) Algorithmic approach:
    a) Predefined algorithms with manual adjustment: most commonly using thresholding-based approaches
    b) Data-driven algorithms: mostly supervised learning, rarely unsupervised learning
        i) Traditional machine learning (handcrafted feature-based)
        ii) Deep learning (learned feature-based)
2) Pattern recognition (PR) task:
    a) Classification
    b) Object detection
    c) Segmentation
    d) Rarely others (such as regression, clustering etc.)
3) Dimension/level of pattern:
    a) Pixel-color value (pixel level)
    b) Cellular pattern
    c) Structural pattern (tissue architecture level)
    d) Global pattern (overall tumor/case/image level)

While algorithms with manual adjustment are a fast and user-friendly tool for scoring immunohistochemical staining intensity, machine learning-based algorithms have also been used to quantify immunopositive tumor cells.[4] Complex data-driven methods are suitable for qualitative or quantitative assessment of a very wide range of prognostically relevant features. Whereas some research is available for AIA of some prognostic aspects, especially for human tumors (see Table 1 and discussion), state-of-the art methods are likely to change significantly with ongoing research. Regardless, already available studies propose a vast variety of methods, even for the same prognostic aspect/parameter (Table 1). For example, algorithms for assessment of metastasis in H&E stained lymph node sections have been phrased as: 1) a classification task, i.e. is metastasis present in the lymph node section?,[5] 2) an object detection task, i.e. where is the metastatic foci in the lymph node section?,[5] or 3) a segmentation task, i.e. what is the area of the metastatic foci?.[6] The decision on which pattern recognition task is used does not only affect development of

algorithms, but also evaluation methods. Development and evaluation of the same algorithms does not necessarily need to be considered as the same pattern recognition tasks depending on the intended interpretation of the algorithmic predictions, for example a classification network can be evaluated with evaluation metrics for segmentation[7] or vice-versa.[6]

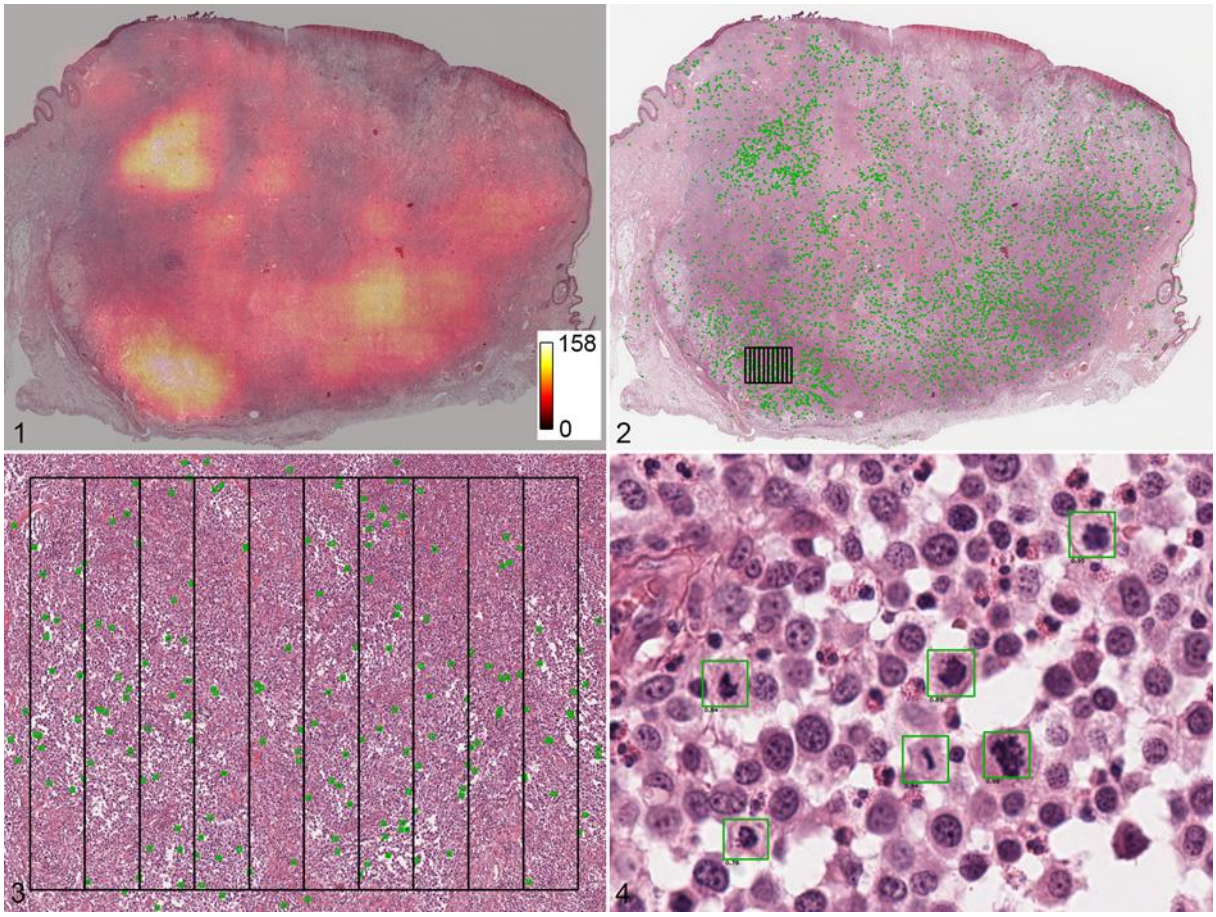
Table 1. Examples of automated image analysis solutions for tumor prognostication of microscopic images (H&E stained images in blue font color; immunohistochemical staining in brown font color) using thresholding-based (TH), traditional machine learning-based (ML) and deep learning-based (DL) algorithms.

Dimension of pattern	Pattern recognition task			
	Object detection	Classification	Segmentation	Other
Pixel-color value		Ki67 scoring, TH [8]		
Cellular pattern	Mitotic figures, DL [9]	Mitotic figures, DL [10]	Mitotic figures, DL [11,12]	Mitotic Count: regression DL [11]
	Mitotic figures, DL [11]			
		IHC scoring, DL [13], ML [4]		
Structural pattern	Identification of metastatic foci in lymph nodes, DL [5]		Metastatic area, DL [6]	
	Presence of tumor, DL [7]		Extent of tumor, DL [7]	
			Necrotic tumor area	
Global pattern		Presence of metastasis in lymph node, DL [5]		
		Tumor grade, DL [14]		
		Tumor type, DL [15]		
		Tumor malignancy and differentiation, DL [16-18]		
		Patient outcome, DL [19]		

		Molecular profile, DL[18,20]		

2.1. AIA implementation

Computer-assisted prognostication (CAP) with review by trained pathologists, as opposed to fully computerized prognostication, is currently recommended, if applicable. Image analysis solutions that allow visualization of algorithmic output (result) as an overlay on the digital image should be favored as this allows confirmation of algorithmic performance of each detected/classified/segmented object by pathologists. Visualization of algorithmic object detection results (overlay on the WSIs; Fig. 4), display of model scores along with detections (Fig. 4), heat maps (Fig 1), segmentation maps and/or preselection of regions of interest (Figs. 1-4) may be integral parts of those CAP systems. Manual correction by pathologists is required if algorithms fail to produce accurate results for cases not covered by the data variability of the training material. These CPATH applications will aid pathologist efficiency, reproducibility and accuracy by reducing repetitive tasks or simplify difficult tasks and are not meant to replace pathologists. This will increase reliability and ability to explain the prognostication approach and allow the reviewing pathologist to retain responsibility in making final decisions (liability). Algorithms that merely classify prognostic aspects from images, are often intransparent (‘black box’) and a more thorough initial performance evaluation and possibly ongoing monitoring (in a certain subset of diagnostic cases) is necessary for those solutions. With considerable changes of the image acquisition / preanalytic workflow (different scanner type, scanning resolution etc.) performance of all types of algorithms have to be reevaluated and – based on results of reevaluation – algorithms may require modification.



Figures 1-4. Example of a computer-assisted prognostication (CAP) approach for counting mitotic figures in H&E stained tumor sections of a canine cutaneous mast cell tumor based on a previously published deep-learning based algorithm.[11] **Figure 1**: Visualization of a mitotic density heat map (hot spots) as an overlay on the original WSI. **Figure 2**: Visualization of the mitotic figure detections (small green squares) and preselection of the tumor area (2.37 mm², black box) with highest mitotic density (hot spot). **Figure 3**: Higher magnification of the preselected mitotic hot spot tumor area. **Figure 4**: Visualization of individual mitotic figure detections (green squares) at high magnification.

3. Specific Recommendations

3.1. Algorithms with Manual Adjustment

Algorithms for manual adjustment are validated image analysis tools that are applicable for a wider range of applications (for example different immunohistochemical stains and tumor types). Handcrafted parameters of those generic algorithms can be adjusted to a specific use/case by manual optimization of certain parameters (e.g., thresholds). Algorithms typically consist of programmed sets

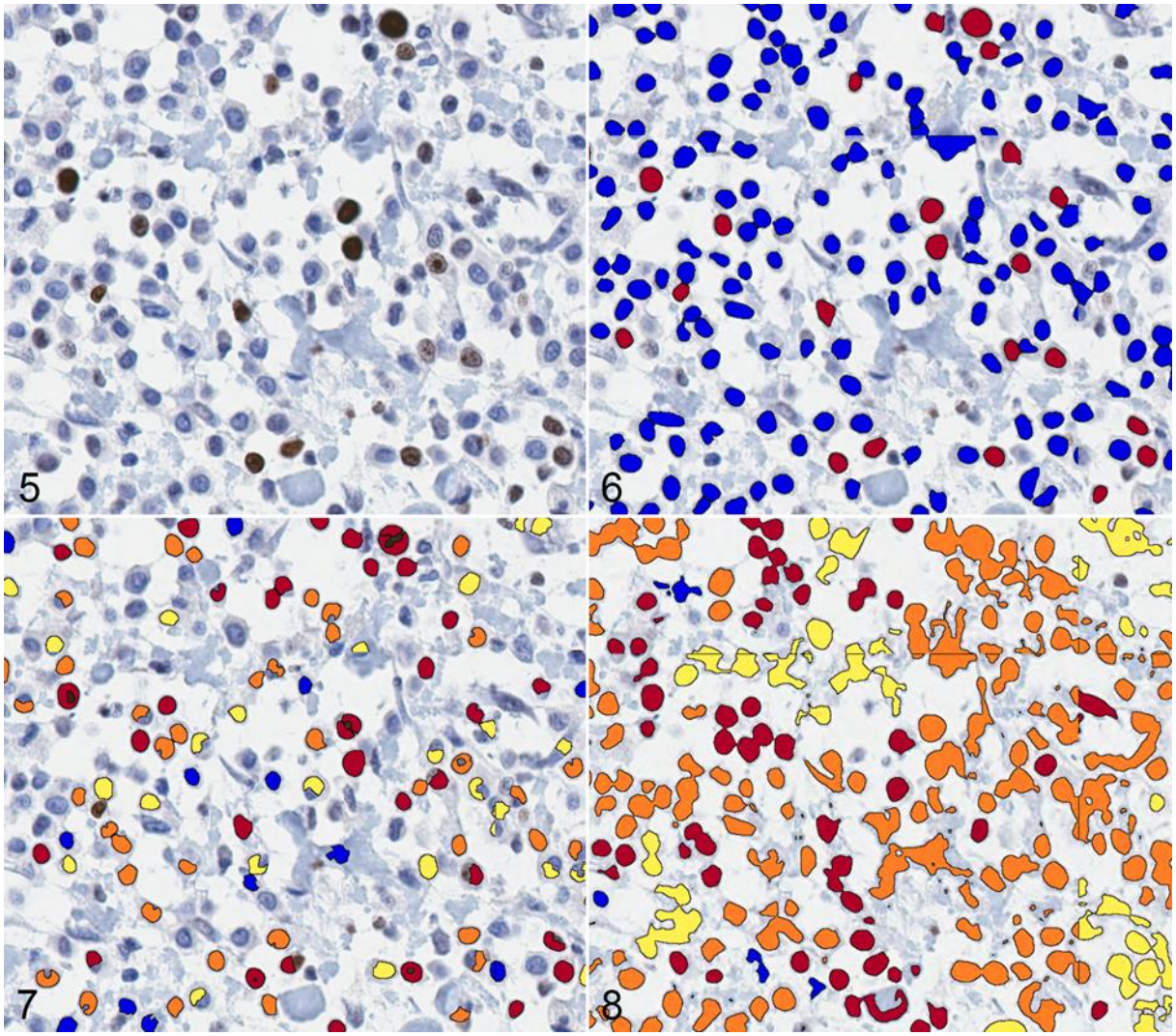
of filters and are not derived from autonomous learning with recorded data. In contrast to data-driven algorithms, performance is manually optimized by adjustment of parameters such as color or size threshold (thresholding). Algorithmic performance assessment is usually obtained by visual assessment of the output (e.g., as overlay on the image). A relevant example of this is automated scoring of immunohistochemical staining intensity of (prognostic) biomarkers. For this, several commercial or open-source, ready-to-use software solutions are available.

3.1.1. Scoring Immunohistochemical Tumor Markers with Segmentation-based Thresholding

While some immunohistochemical stains may be used for classification of the tumor type (presence of signal in a minimal percentage of tumor cells), other tumor markers (such as the proliferation marker Ki-67) require quantification as the proportion of positive cells and signal intensity per tumor cell correlates with patient outcome. It is the goal to measure the biological variability between tumors and reproducible as well as accurate measurements may be facilitated by automated scoring.

1. Ensure minimal pre-analytic variability (see Ramos-Vara and Miller [21]) by consistent tissue handling and staining procedure
    a. Controls for pre-analytic variability are difficult
2. Segmentation of cellular compartments (membrane, cytoplasmic, nuclear) that are labeled by the biomarker through thresholding or other methods. Review the following aspects:
    a. Sensitivity and specificity of segmentation
    b. Contour accuracy: avoid undersegmentation and oversegmentation (see Fig. 5-8)
3. Tumor cell classification through thresholding, based on color intensity of chromogen (hue range, saturation range). Four grades are commonly used: 0 (negative), 1+ (lightly stained), 2+ (intermediately stained), 3+ (darkly stained)
4. Image analysis and interpretation (Cave: inter-platform and inter-operator variability)
    a. Selection (manual or automated) of region of interest on WSI with useful prognostic information for analysis
    b. Review output visually (overlay on WSI) and possibly fine-tune parameter thresholds

c. For validation compare with manual scores by pathologist and/or patient outcome
d. Report weighted scoring system (such as the histochemical score, H-score) and possibly other measurements



Figures 5-8. Microscopic image of a canine cutaneous mast cell tumor with immunohistochemical labeling for the proliferative marker Ki-67. **Figure 5**: without overlay of automated analysis. **Figure 6**: With overlay of automated analysis from a thresholding-based algorithm (Aperio Nuclear Algorithm, Leica Biosystems, Wetzlar, Germany) using appropriate segmentation thresholds. **Figure 7**: Thresholding-based analysis with oversegmentation. **Figure 8**: Thresholding-based analysis with undersegmentation.

3.2. Data-driven algorithms

Data-driven algorithms are developed by training a machine learning model with input data (example images) and – in the case of supervised learning – with output data (associated labels). With increasing computational power, the models can have higher complexity and benefit from higher quantity of labeled data. In contrast to manually adjusted algorithms, fully data-driven algorithms can show a strong dependency on the distribution of the data, hindering adaptation to different domains. If data used to generate the algorithm is, however, representative, the natural image variability can be learned during training and performance can be outstanding on unseen data of the same domain. For application of algorithms to other domains (different tumor types, different WSI scanner etc.) not initially used for testing, a review of the performance (robustness) is required. A great advantage is that algorithmic performance can be accurately evaluated using various metrics, but particular attention must be paid to an unbiased and representative test dataset. For histological tumor prognostication, supervised learning with data from local morphological patterns, such as mitotic figures, has been applied frequently in current literature and recommendations are given in the following section.

3.2.1. Supervised Deep Learning for local morphological patterns

1. Development of a ground truth dataset with local (object-level) annotations (pixel-wise or centroid/bounding box) in digital histological images
    a. Sufficient quantity of cases and annotations per label class(es) (Note A)
    b. Highest possible dataset quality, i.e. consistency and accuracy (Note B)
    c. Labeling method:
        i. Complete labeling of all patterns of interest present in entire WSIs or subparts thereof
        ii. Manual confirmation of all labels by experts recommended. Labels (object-level) may be initially generated by:
            1. Manual screening of images (gold standard, Note B)
            2. Algorithmic-assisted labeling with human expert assessment (Note C)
            3. Combination of 1. manual and 2. algorithmic-augmented labeling (Note B and C)

2. Split of the dataset into independent subsets on case level based on “random” (considering density of pattern of interest per case) or “systematic” (for cross-validation) case selection: training set (training and validation subset, for step 3) and hold-out test set (for step 4). The test set shall be chosen based on a maximum representativeness for the problem, e.g. comprising the complete range of tumor subtypes and grades. While selection of cases according to prevalence can lead to more representative statistical measures, it may underrepresent rare cases and should be treated with caution.
3. Training of deep learning models with supervised learning
    a. The most appropriate artificial / convolutional neuronal network (“state-of-the-art models”) is selected based on the chosen pattern recognition task: classification, segmentation, object detection, regression
    b. Training data subset (with data augmentation) is used for training the model
    c. The validation data subset is used for evaluation of in-dataset generalization performance, selection of networks (recommendation: use ‘early stopping’ or other model state selection methods to prevent under- or overfitting), and calculation of appropriate thresholds
4. Final evaluation of deep learning models:
    a. Test of algorithmic performance on the ground truth test set (Note B). If applicable, repeat training and testing for several (3-5) independent runs (report average, standard deviation and/or range) or perform K-fold cross validation. Possible evaluation metrics and performance visualization (Table 4) include:
        i. Classification task: Accuracy, Error Rate, Precision, Recall (Sensitivity), Specificity, F1-score, Receiver operating characteristic (ROC) curve, Area under the ROC curve (AUC), Confusion matrix for multi-class classification
        ii. Semantic segmentation task: Intersection over union (IOU, also known as Jaccard index), Dice coefficient
        iii. Object detection task: Mean average precision (mAP), Precision, Recall, F1 score
    b. It is not acceptable to modify algorithms based on test results (test dataset)

c. Possibly compare algorithmic performance to performance of multiple pathologists on object-level and/or scores-level (compare pathologists vs. pathologists with algorithm vs. pathologists)
d. Possibly correlate algorithmic results (in comparison to manual scores) to patient outcome

Notes:

A. Sufficient dataset quantity is important for training and testing of suitable deep learning-based algorithms. As deep learning networks learn by iterative training with progressive improvement of performance (up to a certain point), high amounts of data are necessary to ensure that important image features can be extracted by the network. The minimal number of labels and cases required is impossible to generalize as it depends very much on the degree of morphological variability of the searched pattern and other patterns present in the images. For mitotic figures, a few thousand labels was able to yield an algorithmic performance compatible with trained pathologists.[9] However, higher number of annotations can still improve deep learning-based algorithms (for mitotic figures) as long as data quality (see Note C) is also high.[22] Besides sheer numbers of annotations and cases, it is absolutely essential that the images contain a representative degree of variability of the pattern expected to be present on the analyzed images (intended use). In contrast to trained pathologists,[23] image analysis software cannot inherently adapt to biological variability ("normal" variability of morphological pattern, tissue types etc.) and pre-analytic image variability derived from staining protocols, type of WSI scanner and many others. Therefore, it is important that algorithms have been trained with and tested against images with a realistic degree of variability. However, it might be worthwhile to consider excluding images with inappropriate tissue quality for which pathologist-defined labels of the training and test dataset would have unacceptable reproducibility/consistency (dataset quality, see Note B); however, this would require to exclude poor quality cases from analysis as well. While algorithms that have only been trained with the perfect examples of the present pattern will not cope with imperfect examples, algorithms that have been trained mostly with

inconsistent data might generalize poorly. If datasets are created from small subparts of WSI (region of interest) and algorithms are later to be used for analysis of entire WSIs, it is necessary that the image sections selected are representative for the WSI, i.e. that all present tissue patterns and morphological variants are included.

B. Highest possible dataset quality leads to optimal algorithmic performance but is a trade off with time investment for dataset development and expertise of annotators. Trained and experienced pathologists are defined as the gold standard for labels of morphological patterns, however even experts have visual and cognitive limitations[3] that lead to inter- and intra-observer variability, especially with ambiguous patterns. Inconsistency of decision criteria used for generating labels in the training data set may lead to a less optimal learning process for the algorithm (errors in learning because the label was wrong). Inconsistency in the pathologist-derived test data will result in miscalculation of the true algorithmic performance. In contrast, poor label accuracy, i.e. whether the annotations actually represent the pattern of interest, does not necessarily influence evaluation metrics, but will diminish the prognostic value of the derived deep learning model.

Despite high measures of precaution, pathologist-derived datasets will inevitably have less than perfect label quality. For example, for the AMIDA13 dataset two pathologists independently labeled human breast cancer images for mitotic figures and identified 1,088 and 1,599 mitotic figures, respectively, in 23 images, of which 649 (concordance: 31.8%) had a positive label by both experts and 1,389 pattern had positive labels only by one pathologist (discordance: 68.2%). Of the 1,389 singly annotated patterns, a final decision was made by two further pathologists who agreed upon 434 labels being a mitotic figure (31.2%) and disagreed in 955 instances (68.8%).[24] For a later challenge (TUPAC16) further cases were added to this dataset resulting in a total number of 1,552 mitotic figures in the training set (agreement of pathologists for individual labels not published).[23] Another research group relabeled those 73 images and identified 1,239 mitotic figure labels in the same location as

in the original dataset – that is 79.8% of the original labels – and identified 760 additional mitotic figure labels (+ 49 %) previously not labeled in the original dataset.[25] From this we draw two conclusions: 1) pathologists may misclassify objects and 2) pathologists may overlook objects which are actually present, both inevitably leading to less than perfect dataset quality. Reduction of pattern misclassification may be achieved by: well-defined decision criteria, diligent annotations, multiple blinded expert annotators with agreement on disagreed pattern (both, majority vote by an additional pathologists or consensus by the initial annotators seem to be acceptable[26]), reassessment of labels (unaided or by visual clustering based on feature vectors), and inclusion of cases with sufficiently high quality (quality must be representative for intended use). Reduction of missing relevant patterns may be achieved by repeated screening, usage of annotation software with guided screening, and algorithmic missed candidate screening (Note C).

C. Pathologist-derived labels are frequently the gold standard for local labels, however fully manual dataset development has some degree of inconsistency and is very time consuming. Different computer-assisted labeling methods have been developed to a) reduce human error arising from inter- and intra-observer variability, mislabeling and overlooking objects (dataset quality, see Note B) or b) to increase dataset size with reduced time-investment. Just like fully manual approaches, computer-assisted approaches may have biases that have to be carefully weighed against the benefits. We consider it essential and of utmost importance that computer-assisted label creation or changes are always reviewed by pathologists in order to reduce the bias introduced by these methods. Fully algorithmic generation of labels without review by a pathologist (pseudo-labels) is not acceptable. Computer-assisted labeling methods include:

a. Identification of mislabeled objects by review of the initial classification can be facilitated by visual clustering based on feature vectors, for example obtained by representation learning.[26]

b. Identification of overlooked objects can be done by analyzing the images with image analysis methods, which have been trained by a preliminary, pathologist-defined dataset for these images or with different datasets from a similar domain. This approach is best applied with high sensitivity / low specificity (high number of true and false positives and low number of false negatives) algorithmic detection of potential candidates that require review by the annotator(s).[26,22,25] The high number of false positive detections aims to provide a high detection rate, while additionally reducing the confirmation bias of pathologists.
c. Computerized creation of labels in additional images with subsequent review by pathologists (expert-algorithm-collaboration) requires machine learning-based algorithms with high specificity to ensure that mostly true positive labels are generated. While this approach significantly reduces the time required for labeling, it is known that pathologists may miss algorithmically induced errors.[27] Therefore, high diligence is necessary for the review by an expert.

3.2.1.1. Mitotic figures in HE-stained images

1. Dataset: (some open datasets are available, see Table 2)
    a. Quantity: >>1500 mitotic figure annotations from several cases with fully labeled images. If algorithm is intended to analyze entire WSI, images need to be representative for WSI
    b. Quality:
        i. reduce misclassification by multi-expert annotations (see Wilm et al.)[28] : ground truth by at least two pathologists with final agreement (either consensus by reviewing pathologists or additional evaluation by a third pathologist of disagreed labels, Note B)
        ii. reduce overlooked candidates: screen images at least twice, subsequently computer-augmented methods with object-level expert assessment may be used (see Bertram et al.[22,25] or Aubreville et al.)[26]
2. Training: Supervised learning with state-of-the-art object detection networks, possibly a second stage with a classification network

3. Evaluation: at least F1-score (see Table 4), ideally from 3-5 runs to report performance variability
4. Implementation as a CAP system with preselection of mitotic 'hot spots' (Figs. 1-4) and visualization of algorithmic detections (Fig. 9) as an overlay on WSI that can be reviewed by a pathologist

Table 2. Relevant open datasets of histological tumor images with mitotic figure annotations (MF).

Dataset	Tumor type	MF	Tumor cases	Published F1 scores
MITOS-ATYPIA 2014[29]	Human breast cancer	749	11	0.356
TUPAC 2016[23]	Human breast cancer	1,552	73	0.652
Alternative TUPAC16[25]	Human breast cancer	1,999	73	0.735
MITOS_WSI_CCMCT_ODAEL[22]	Canine mast cell tumor	44,800	32	0.820
MITOS_WSI_CMC_CODAEL[26]	Canine breast cancer	13,937	21	0.791

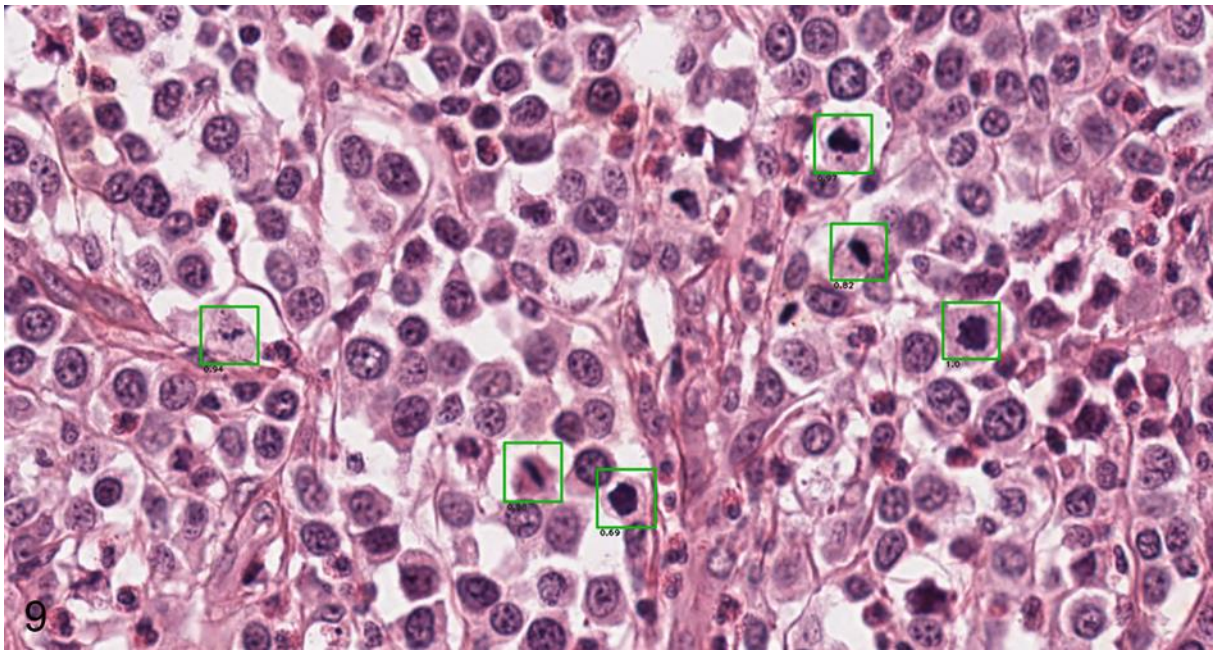


Figure 9. Mitotic figure detections of a deep-learning based algorithm (published by Aubreville et al. [11]) from a histological section (H&E stain) of a canine cutaneous mast cell tumor

4. Discussion

Histologic and immunohistochemical tumor prognostication is traditionally performed by manual assessment of glass slides by trained pathologists. However, high inter- and intra-observer variability is well-known for visual assessment of numerous prognostic parameters by pathologists, such as for mitotic figures,[30] regardless of several attempts of standardization. Therefore, automated image analysis (AIA) has been proposed as a potential method to reduce human bias of visual assessment by pathologists.[3,2] Besides higher reproducibility, AIA of microscopy images can have higher accuracy and can analyze vast amounts of data, i.e. large WSIs, in much shorter time and thereby increase diagnostic efficiency. Although AIA is currently almost exclusively used for research purposes, availability for diagnostic services will be facilitated by implementation of digital microscopy workflows in laboratories, increasing computational power and advancing AIA methods. As computerized approaches have the potential to improve tumor prognostication, future studies that develop those tools are indicated which should identify the benefits and limitations through comparison to manual assessment by pathologists. Due to higher reproducibility and possibly higher accuracy, computerized or computer-assisted methods likely will have different predictive ranges and cut-offs compared to the established manual approaches, which need to be elucidated in future studies.

Various methods for creation of algorithms are available, ranging from simple thresholding-based and more sophisticated deep learning, which have different sources of error and all require careful development. While thresholding-based algorithms are mostly restricted to immunohistochemical images, data-driven methods (especially deep learning) are suitable for a wide range of morphological patterns. For example, immunohistochemical staining intensity of tumor cells can be automatically quantified using thresholding-based or data-driven algorithms.[13] Using those methods, AIA has been shown to have higher reproducibility and higher prognostic value compared to the manual approach by pathologists for Ki-67 index in human breast cancer,[31,32] and various membrane-binding biomarkers in human esophageal adenocarcinomas.[33] In contrast, high performance for analysis of complex morphological patterns, such as detection of mitotic figures in H&E stained tumor sections, is mostly achieved with advanced deep learning networks and

sufficient training datasets (quality and quantity). The most common application of deep learning in pathology on the cellular-level is the detection of mitotic figures in tumor sections. Current studies on human[25] and canine[26] breast cancer as well as canine mast cell tumor[11,22] reported quite high performances metrics. While it has been shown that those algorithms are on par with pathologists for detecting individual mitotic figures in images [34,9] and outperform pathologists in detecting the 'hot spot' regions (mitotically most active tumor region) in WSI,[11] correlation to patient outcome have not yet been investigated for those object detection tasks.

Besides evaluating patterns on the cellular level, deep learning can solve problems at lower magnification, such as detecting metastasis in lymph node sections.[5] For metastasis identification, algorithms can be used for prescreening of images and a computer-assisted approach has been shown to have higher sensitivity, higher diagnostic speed and reduction of the perceived difficulty compared to the unassisted approach.[35] A rather recent field of research of human pathology is the recognition of global patterns representing the entire H&E stained WSIs such as classification of tumor malignancy (normal tissue, benign tumor, in situ carcinoma or invasive carcinoma) [17] or classification of images according to their molecular features (genetic alterations and gene expression)[20] and even direct estimation of patient outcome.[19] Although this approach reduces the time-investment for manual labels on the morphological pattern-level (strong label level), deep learning networks often require thousands of weakly labeled WSI (one WSI = one label) in order to be able to extract relevant morphological features that correlate with tumor (sub)type, molecular features, or outcome and have not yet been investigated in veterinary pathology.

There are multiple open-source as well as commercial software solutions available that allow pathologists to develop image analysis tools by themselves. However, data-driven approaches are particularly challenging and prone to bias. Involvement of experienced pathologists is essential for dataset development, model evaluation and software implementation into diagnostic workflows[36,37] and pathologists need to get familiar with terminology (Table 3), available CPATH methods and evaluation matrices (Table 4). For algorithm development, cooperation with pattern recognition scientists, who have high expertise on CPATH methods and can build customized software solutions, can be highly beneficial for more difficult

projects. While thresholding-based approaches have high explainability and can be easily adapted by pathologists based on visualization of results, data-driven approaches are often considered a ‘black box’ as decision criteria of the algorithms are usually too complex to understand. While the lack of decision-making criteria limits identification of possible sources of error, it is absolutely essential that test datasets are created in an unbiased way and fully encompass the intended use. Although algorithms are 100% reproducible (same result for the same image), they are not necessarily robust, meaning that they cannot always cope with biological and pre-analytic variability on which pathologists can adapt more easily. For example, a deep learning-based algorithm for mitotic figures may cope only poorly with domain shifts introduced by different WSI scanners.[26] While for thresholding-based solutions it is necessary to use slide standardization and color normalization to be able to work with similar cutoff values, data-driven algorithms are capable of learning a high degree of image variability and training datasets should include realistic variability that will likely be encountered during the intended use. If algorithms are to be used for a new, but related domain (characterized by biological or preanalytic features being not within the variability of the training set), performance needs to be reevaluated and transfer of machine learning algorithms to those related domains may be improved by methods like threshold optimization, transfer learning and domain adaptation.[26] Besides proper performance evaluation, there are approaches that can convert that ‘black box’ into a more transparent ‘glass box’ that are likely to have higher acceptance among pathologists. For example, some algorithms can be implemented as computer-assisted prognosis (CAP) systems (as opposed to fully computerized decisions) which are geared toward aiding, not replacing pathologists. Image analysis solutions that allow visualization of algorithmic output (result) as an overlay on the digital image should be favored as this allows confirmation of algorithmic performance of each case by pathologists. These CAP approaches will improve the reliability of the AIA system and allow the reviewing pathologist to retain responsibility in making final decisions with regards to these prognostic parameters. In contrast, WSI classification tasks, such as classification of H&E-stained images into associated molecular features, are much more difficult to convert in to a ‘glass box’ and an appropriate performance evaluation with unbiased and representative test datasets is absolutely essential. Some studies have used heat maps to localize the tumor regions that were relevant for the classifier and pathologists can try to

interpret the underlying morphological features, which is not necessarily consistent with known prognostic parameters and may lead to identification of new, prognostically relevant features.[19]

AIA of microscopic tumor images has many potential advantages over visual assessment by pathologists and might lead to more reproducible and accurate prognostic information, especially if pathologists review predictions for correctness (CAP). Nevertheless, there are numerous challenges that still hinder use of AIA for routine tumor evaluation. Future research is highly encouraged in order to develop open access datasets, improve deep learning methods, find ways to cope with domain shifts, investigate usefulness of software solutions, prove prognostic relevance of algorithms and many more. These issues will likely be resolved as more progress is made in this growing field of veterinary and human pathology. It is not only important that pathologists are willing to become users of AIA software but pathologists are also indispensable for developing and scientifically evaluating these tools in cooperation with pattern recognition scientists.

5. Definitions of terms

Table 3. Definitions of CPATH terms relevant for this Guideline

Term	Definition
Algorithm:	An image analysis algorithm is the ready-to-use software that is used for computerized assessment of images (input data), such as histological WSI, that result in predictions about the respective image analysis task. Using machine learning, the algorithm is commonly built around a machine learning model.
Annotation:	Annotations are individual events in the dataset with the position and/or outline for patterns of interest in the image and associated label classes. They are mostly created manually by annotators (i.e. pathologists, gold standard) with annotation software or created using computer-assisted labeling. Annotations are used as target data to an image (input data) for supervised machine learning.
Annotation software:	Annotations software enables an annotator to generate annotations for (whole slide) images, which can be saved in a database, a text file corresponding to the image, or even within the image file itself. Besides simple viewing of the images, a number of annotation tools (single and multi-coordinate annotations) and tools that facilitate the annotation process (such as guided screening, blind multi-label annotations etc.) are available.
Artificial intelligence (AI):	AI is a branch in computer science commonly used for computerized gathering of information from raw data, such as

	histological images, by simulating intelligent behavior in computers (as opposed to natural intelligence displayed by humans and animals). Machine learning is a subset of artificial intelligence.
Artificial neuronal networks (ANN):	ANNs are machine learning models that are inspired by natural neuronal systems of humans and animals. Artificial neurons are organized in layers (input, hidden, output layer) that are connected to each other and can receive (with a specific weight), process and transmit signals. By modifying/adjusting the weights of neurons and their connections during training, a network is trained to predict a certain task. A highly relevant type of an ANN for deep learning is a convolutional neural network (CNN).
Augmentation:	Data augmentation is a method applied to data within the training process of a machine learning model. Annotated data gets altered (rotation, zooming, cropping, color alterations) in order to increase variability of the available data and thereby improve ability to generalize the most important image features during training.
Automated image analysis (AIA) / digital image analysis (DIA):	AIA is the process of extracting information from a digital image (typically whole slide image) by computerized methods. Some algorithms for automated image analysis are based on AI (including traditional machine learning and deep learning), others on simple image processing steps (including thresholding-based approaches).
Black box algorithm:	A black box (opposed to glass box) algorithm is characterized by the lack of an understandable relationship of input data and algorithmic output, i.e. the used decision criteria are too complex to understand. Deep learning-based algorithms, which extract relevant features of the pattern of interest by themselves, are often considered to be a ‘black box’. Display of intermediate results (e.g., object detections as an overlay in the digital images) can increase comprehensibility of these algorithms.
Classification:	Classification is the task of image analysis that assigns categories of patterns (label classes) to an input image or image patch. Binary classification assigns one out of two label classes and multi-class classification assigns one out of several label classes to an image.
Color normalization:	Color normalization is the transformation of color properties of a (whole slide) image to align to a single standard. Color values of a WSI (or subset thereof) may vary based on tissue processing protocols, method of digitization of the glass slide, or due to other factors. WSI obtained from different WSI scanner types commonly have different color representations. This color variation of images can significantly influence image analysis and therefore compensation by color normalization may be beneficial.
Computational pathology (CPATH):	A branch in pathology using computerized methods to gather relevant information on a disease in a patient from one or multiple sources of raw data such as histology images, macroscopic images and gene sequences. In this Guideline,

	CPATH especially refers to the field in pathology using automated image analysis (AIA) methods for digitized microscopic tumor sections (especially whole slide images, WSI). Methods commonly used for AIA come from the field of artificial intelligence (AI), more specifically deep learning. A broader definition of CPATH is the extraction of relevant information from any source of raw data including clinical electronic medical records, laboratory data, diagnostic imaging, genomics and others. CPATH uses computerized methods (AI and others for histological images) analogous to molecular pathology using molecular methods (PCR and others for detection of mutations).
Computer-assisted/aided prognosis (CAP):	CAP is a diagnostic workflow for tumor prognostication using automated image analysis software to support the decision making by a pathologist. In contrast to fully computerized workflows, for CAP, pathologists always review algorithmic detections and use the information obtained by algorithms as a guide for the final diagnosis (algorithm/software-assisted decision support). CAP systems may support the pathologist best in critical steps that are known to have high inter- and intra-rater variability due to visual or cognitive limitations of humans. One example would be preselection of an area of interest (such as the mitotically most active region for performing the mitotic count). While algorithmic predictions intend to improve reproducibility and accuracy as well as reduce pathologists' time investment, the pathologists review intends to ensure reliability of the algorithm for each analyzed case.
Convolutional neural network (CNN):	A specific type of an artificial neural network (ANN) commonly used within deep learning in microscopic images. CNNs contain one or more layers that contain convolutional operators. These are specifically designed to extract spatial patterns from within the image, such as edges (input layers) or more complex patterns descriptive of an object (later layers).
Cross validation:	For cross validation the dataset is divided systematically into a specific number (K-folds, at least three) of subsets (on the case level) that are alternately used as the validation or test subset, while the remainder sets are commonly used for training. With each iteration (fold) of the training-test-process a different part of the dataset is used for training or testing, respectively. This allows a more thorough evaluation of the generalization of the method on the whole dataset and allows training with a larger proportion of the dataset.
Database:	In the context of this Guideline, a database is a collection of labels assigned to pathological, especially microscopic, (whole slide) images. Databases usually contain the image names, the identification number of the label, a label class, the location (x, y coordinates) of the label in the image, the annotators name for multi-expert annotations and possibly other information. Together with the annotated images, databases are part of

	datasets, which are a precondition for training and testing supervised machine learning algorithms.
Dataset:	In the context of this Guideline, datasets are collections of pathological, especially microscopic, (whole slide) images and the associated database of annotations. Datasets are a precondition for development of supervised machine learning algorithms and performance testing.
Deep learning (DL):	DL is a subset of machine learning (ML). In contrast to ML, DL uses artificial neuronal networks with multiple (“deeper”) hidden layers and is capable of extracting (learning) relevant image features of pattern by itself directly from the raw image data. Deep learning systems are capable of providing unprecedentedly accurate solutions, yet require high-quality and high-quantity datasets (‘big data’).
Domain:	In machine learning, a domain defines the set of factors that influence the feature distribution of a dataset. For microscopy images, common domain-defining factors include the tissue type present, type of (neoplastic) disorder, slide preparation, staining method, image capturing and processing (WSI scanner hardware and software). Variabilities of domains between two datasets cause a domain shift, which can result in reduced algorithmic performance, if an algorithm is not “robust”. Multiple methods for reducing the domain shift exist, such as transfer learning, color space adaptation, and unsupervised domain adaptation of a model.
Generalization performance:	Data-driven algorithms are trained in such a way that they learn relevant features that distinguish the pattern of interest from the background class(es). Networks learn from training examples, but extracted features must be applicable to unseen (out-of-sample) data. Generalization performance describes how well algorithms perform on unseen data, i.e. if the extracted features are generally descriptive. Due to morphological variability (biological and preanalytic) of the patterns of interest, training data is prone to a sampling error and the provided input information might not be predictive for other samples. Therefore, representativeness and high-quantity of training data, i.e. containing relevant degree of variability, is important. Further generalization errors may occur from overfitting or underfitting models (see definitions below).
Gold standard:	The gold standard is the practical method that is well-established and most suitable for development of the ground truth labels. For histological and immunohistochemical specimens, trained pathologists are most commonly used as gold standard, regardless of their visual and cognitive limitations leading to some degree of inter- and intra-observer variability. As automated image analysis is designed to overcome human limitations, this approach seems to be somewhat paradoxical. However, a true gold standard is often lacking for most morphological patterns, whereas global features may have a more objective gold standard, such as presence/absence of genetic mutation.

Ground truth:	Ground truth is the information of the 'true' label class derived by the defined gold-standard method. These ground truth labels are critical as they represent the reference during model training. They are also the reference for testing the algorithmic model's performance (see Table 3). As manual assessment by human experts (pathologists) with well-known inter- and intra-rater variability are the gold standard for most histological patterns of interest, the ground truth can be subject to various biases and can include annotation errors. It is the aim of a highly diligent dataset creation to limit these errors.
Histochemical (H)-score:	Scoring system commonly used to quantify of tumor cell immunopositivity (immunohistochemistry). Tumor cells are graded based on their staining intensity into four grades: 0 (negative), 1 (lightly stained), 2 (intermediately stained), 3 (darkly stained). The weighted score is relative to the tumor cells enumerated and range between 0 - 300. $$H = (\%\ of\ "1") + 2 \times (\%\ of\ "2") + 3 \times (\%\ of\ "3")$$
Label:	A label is an assigned label class to a morphological pattern or image. An annotation is a label in a specific location (centroid x,y position or demarcated area) of an image.
Label class:	Different categories of labels used for distinguishable morphological patterns/image features of interest.
Machine learning (ML):	ML is a subset of artificial intelligence in which an algorithm learns from representative data in order to create a model that can make decisions on new data without human interaction. ML can be categorized into "traditional" ML and deep learning (DL) methods. While the relevant features of the patterns of interest are given to the model with traditional ML (also called "hand-crafted"), DL networks are capable of extracting these features by themselves. Compared to traditional ML, DL systems are generally more powerful but often require more data ('big data') for training.
Model:	A machine learning model is the product of training a machine learning algorithm for solving a specific task. Models created with the same inner structure and datasets may vary somewhat, as the learning process involves random sampling of and random variations within the data, and is thus not deterministic.
Object detection:	Object detection is a task in pattern recognition with localization (x and y coordinates) of a pattern of interest in the image and categorization of the label class of the pattern (classification). Errors of object detection may occur by wrong localization or class-assignment of the object.
Overfitting:	Machine learning aims to generate algorithms that can generalize features of the pattern of interest and thereby accurately predict output in unknown images. During training, networks gradually (over consecutive iterations) reduce the training error while learning from the training data. If the network has learned features specific only for the training data too well, i.e. was overfitted (usually at a late stage of the training), then the algorithms cannot extrapolate ("generalize") to the validation

	and test data. To evaluate this effect, the training process is regularly validated against an independent dataset (validation set) and training can be terminated (early stop) at the best validation performance. The opposite of overfitting is underfitting, which results from insufficient training iterations or training data.
Over-segmentation:	Oversegmentation is an error in object segmentation with the compartment being smaller than the objects of interest (such as a cell or nucleus). This may lead to single objects being segmented into more than one compartment.
Patching:	Training of artificial neuronal networks with entire WSI is hampered by limited computational power. Patching is the process of producing smaller image sections (“patches”) from WSI that contain training examples (input data) for the training process. The appropriate patch size depends on various factors such as on the pattern of interest, available and required (downscaling) resolution of the images and available computational power.
Regression:	Regression is a task in pattern recognition that predicts a continuous/numerical output variable to an image.
Robustness:	Robustness is the reproducibility of an algorithm under variable image conditions (i.e. domain shift including variable staining, different scanners etc.). While all algorithms have 100% reproducibility when analyzing the same image, algorithms cannot necessarily adapt to pre-analytic and biological variability (generalization performance) in such a way like pathologists, which might significantly reduce algorithmic performance.
Segmentation:	Segmentation is the demarcation of compartments comprising all pixels which represent a morphological pattern (at the cellular or spatial-arrangement level). Imperfect segmentation results in under- or oversegmentation. Segmentation as a pattern recognition task describes the classification of each pixel of the image.
Supervised learning:	Supervised learning is a specific form of machine learning (as opposed to unsupervised learning) that results in algorithmic predictions based on both input and output data. Labelled data (output data) assigned the training image patches (input data) is required for training the artificial networks. It is the most commonly used form of artificial learning for CPATH in tumor histology. Pattern recognition tasks of supervised learning can be image classification, object detection, segmentation or regression.
Thresholding (algorithm):	Thresholding is a relevant way how parameters of algorithms for manual adjustment can be optimized to individual images. Thresholding-based algorithms use a set of simple image processing methods (filters) for segmentation of images, one of which is categorizing pixel-color values of compartments based on human-defined lower and upper limits of pixel values in these filters. Such algorithms commonly do not involve machine learning and therefore does not learn from data.

Under-segmentation:	Undersegmentation is an error in object segmentation with compartments being larger than the objects of interest (such as a cell or nucleus). This may lead to individual compartments containing more than one object.
Unsupervised learning:	A specific form of machine learning (opposed to supervised learning) that discovers patterns in data based only on input data (images). This form of learning does not use labeled data (output) for training of networks. A common pattern recognition task of unsupervised learning is clustering, where the dissimilarity and similarity of data is exploited to form categories (clusters).
Whole slide image (WSI):	WSI are “digital slides” with high microscopic resolution and containing all relevant tissue sections of a glass slide. WSI are generated (“scanned”) with WSI scanners (hardware), a process called whole slide imaging, and modified (for example stitching, image compression etc.) by associated software. WSI can be viewed by pathologists on computer monitors using a viewing software (digital microscopy) or analyzed by means of automated image analysis (AIA).

6. Evaluation matrices and performance visualization

Table 4. Important evaluation metrics for test performance of data-driven algorithms of object detection (OD), image classification (IC) and semantic image segmentation (IS) algorithms.

Evaluation matrices	**Definition and Formula**	**Useful for…**		
		OD	IC	IS
Accuracy (Acc):	$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$ The Acc is suitable for image classification, which states how many images were correctly classified. If classes are independent, accuracy can be given for each class separately. For image segmentation, accuracy can theoretically be calculated on the pixel-level (pixel-accuracy), however, this metric is inappropriate if the pattern of interest does not have a similar proportion of area as compared to the background class (which is common for histologic patterns).	No	Yes	No
Area under the ROC curve (AUC):	The AUC is the area underneath the ROC curve (see below) and is suitable for classification experiments. This metric aggregates measures for all possible thresholds and ranges from 0 to 1 with better performing algorithms having higher scores. The advantage of the ROC AUC value is that it is representative of the raw	No	Yes	No

	prediction performance, independent of the precision-recall trade-off (which is determined by the threshold).			
Average precision (AP)	Average Precision (AP) is the average of several precision measurements for different recall values (recall-precision graph). Most commonly, the AP is determined at 11 different recall values (0.0, 0.1, … up to 1.0). The AP is determined for a binary object detection tasks and ranges from 0 (poor performance) to 1 (perfect performance).	Yes	No	No
Confusion matrix / error matrix	The confusion matrix is a table layout for performance evaluation of multi-class classification tasks with rows for each predicted class and columns for each label class. This table separates the predicted classes into the actual (ground truth) class and it can be visualized which classes lead to most "confusion".	No	Yes	No
Dice coefficient:	The dice coefficient is a metric to evaluate the overlap between a predicted segmentation mask and a ground truth mask. It ranges between 0 (no overlap) to 1 (all pixel overlap).	No	No	Yes
Error rate (Err):	$$Err = \frac{FP + FN}{TP + FP + TN + FN}$$ The Err is suitable for image classification experiments, which states how many images were misclassified. The type of error (false positive or false negative) is, however, not distinguished.	No	Yes	No
False negative (FN):	A FN is an image analysis error with discrepancy between negative predictions and ground truth positive condition. *Binary Image classification task*: Algorithmic classification determined absence of the searched class while the ground truth indicates presence. *Object detection task:* a pattern of interest is not detected within the predefined distance (in pixel) to or overlap (IOU) with a ground truth annotation.	Yes	Yes	No
False positive (FP):	A FP is an image analysis error with discrepancy between positive predictions (with a model score above defined threshold) and ground truth negative condition. *Binary Image classification task:* Algorithmic classification determined presence of the searched class while the ground truth indicates absence.	Yes	Yes	No

	Object detection task: a pattern of interest is detected outside the vicinity (outside maximal distance in pixel or with too small IOU overlap) of an ground truth annotation.			
F_1 score	$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$ Or $$F1 = \frac{2TP}{2TP + FP + FN}$$ The F_1 score is suitable for object detection and image classification experiments. It states the harmonic mean of recall and precision. The score ranges from 0 (poor performance) to 1 (perfect performance).	Yes	Yes	No
Mean average precision (mAP)	$$mAP = \frac{AP_{Q1} + AP_{Q2} + AP_{Q3} + \ldots + AP_{QN}}{N}$$ The mAP is the mean of the AP values determined for all (N) different queries (Q1,Q2,…QN). Queries are mostly defined as different label classes for the object detection task (multi-class task). The mAP evaluates the overall performance under different queries and values range from 0 (poor performance) to 1 (perfect performance).	Yes	No	No
Intersection over unit (IOU) / Jaccard index	$$IOU = \frac{Area\ of\ overlap\ (intersection)}{Area\ of\ union}$$ IoU measures the overlap between areas of the annotated (ground truth) and algorithmically predicted pattern. For a segmentation task, it measures to which extent the ground truth and prediction overlap (ranging from 0 -1). In object detection tasks, we typically require a minimum IOU for two objects (ground truth and prediction) to be considered as matching.	No	No	Yes
Precision (positive predictive value)	$$Precision = \frac{TP}{TP + FP}$$ Precision is suitable for object detection and image classification experiments. It states how many of the positive predictions are relevant, or, in other words, how many of the detected patterns (object detection task) or positive classifications (image classification task) are in agreement with the ground truth positive labels.	Yes	Yes	No
Recall (sensitivity,	$$Recall = \frac{TP}{TP + FN}$$	Yes	Yes	No

true positive rate)	Recall is suitable for object detection and image classification experiments. In an object detection task it states how many of the relevant patterns, i.e. ground truth positive labels, have been detected in the image. In an image classification task it states how many of the ground truth positive images have been classified as positive.			
Receiver operating characteristics (ROC) curve:	A graph plotting true positive rate (TPR, Recall; y-axis: 0 - 1) against False positive rate (FPR; x-axis: 0 - 1) at numerous classification thresholds of a binary classification experiment. Commonly multiple classification algorithms (with different performances) are compared in these graphs. $$TPR = Recall = \frac{TP}{TP + FN}$$ $$\text{FPR} = \frac{FP}{FP + TN} = 1 - Specificity$$ An orthogonal line in the graph (line of no-discrimination) indicates random classification at the respective threshold (balance points with TPR = FPR). A classification result that is better than guessing gives points above the line and classification results worse than random guessing gives points below the line. The FPR=0 and TPR=1 point (error-free point) would indicate perfect classification. If comparing multiple algorithms, the curve closest to the left upper corner is interpreted as having best performance. For individual algorithms, the point in the curve closest to the error-free point (point of the curve with a tangent parallel to the line of non-discrimination) indicates the highest performance. However, depending on the desired test, it may be preferable to minimize one type of classification error over the other (sensitivity vs. specificity). Lower thresholds will result in more positive classifications (more true positives and more false positives), while increasing the threshold will result in more negative classifications (more true and false negatives).	No	Yes	No
Specificity (True negative rate):	$$Specificity = \frac{TN}{TN + FP}$$ Specificity can be used for image classification experiments and states how many of the ground	No	Yes	No

	truth negative images have been classified as negative.			
True negative (TN):	TN classifications are defined if algorithmic classification and ground truth annotation both determined absence of the searched class. TNs are not reasonable for object detection tasks.	No	Yes	No
True positive (TP):	A TP is a positive prediction and ground truth positive condition. Image classification task: Algorithmic classification and ground truth annotation determined presence of the searched class in the image. *Object detection task:* a pattern of interest is detected within a maximum distance (in pixel) or IOU threshold to the ground truth annotation. The maximum distance should be adapted to the specific task, for example be equivalent to the median radius of a tumor cell for detection of mitotic figures.	Yes	Yes	No

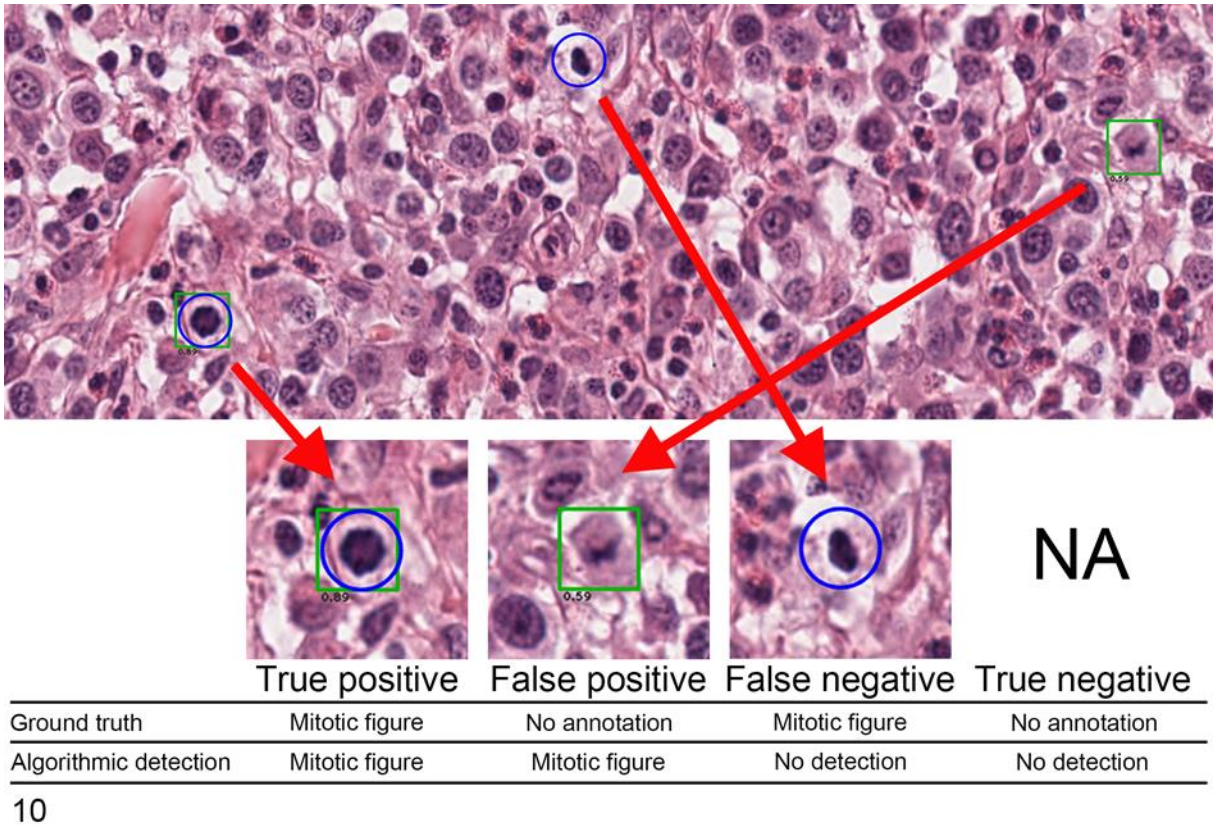


Figure 10. Examples of a true positive, false positive and false negative detection of a deep learning-based algorithms that was trained to detect mitotic figures in canine cutaneous mast cell tumors (see Aubreville et al.) [11] True negatives are not available (NA) for object detections tasks. Green squares represent algorithmic detections

(positive detection with a model score above 0.5, indicated below the box) and blue circles represent ground truth annotations by pathologists.

7. References:

1. Abels E, Pantanowitz L, Aeffner F, et al. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *J Pathol*. 2019;249**:** 286-294.
2. Bertram CA, Klopfleisch R. The Pathologist 2.0: An Update on Digital Pathology in Veterinary Medicine. *Vet Pathol*. 2017;54**:** 756-766.
3. Aeffner F, Wilson K, Martin NT, et al. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. *Arch Pathol Lab Med*. 2017;141**:** 1267-1275.
4. Acs B, Pelekanou V, Bai Y, et al. Ki67 reproducibility using digital image analysis: an inter-platform and inter-operator study. *Lab Invest*. 2019;99**:** 107-117.
5. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*. 2017;318**:** 2199-2210.
6. Pan Y, Sun Z, Wang W, et al. Automatic detection of squamous cell carcinoma metastasis in esophageal lymph nodes using semantic segmentation. *Clinical and Translational Medicine*. 2020**:** e129.
7. Cruz-Roa A, Gilmore H, Basavanhally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Scientific reports*. 2017;7**:** 46450.
8. Klauschen F, Wienert S, Schmitt WD, et al. Standardized Ki67 diagnostics using automated scoring—clinical validation in the GeparTrio breast cancer study. *Clinical Cancer Research*. 2015;21**:** 3651-3657.
9. Veta M, van Diest PJ, Jiwa M, Al-Janabi S, Pluim JP. Mitosis Counting in Breast Cancer: Object-Level Interobserver Agreement and Comparison to an Automatic Method. *PLoS One*. 2016;11**:** e0161286.
10. Aubreville M, Krappmann M, Bertram C, Klopfleisch R, Maier A. A guided spatial transformer network for histology cell differentiation. *arXiv preprint arXiv:170708525*. 2017.
11. Aubreville M, Bertram CA, Marzahl C, et al. Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region. *Sci Rep*. 2020**:** 1-11.
12. Li C, Wang X, Liu W, Latecki LJ, Wang B, Huang J. Weakly supervised mitosis detection in breast histopathology images using concentric loss. *Medical image analysis*. 2019;53**:** 165-178.
13. Saha M, Chakraborty C, Arun I, Ahmed R, Chatterjee S. An advanced deep learning approach for Ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer. *Sci Rep*. 2017;7**:** 1-14.
14. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*. 2020;21**:** 233-241.
15. Kiani A, Uyumazturk B, Rajpurkar P, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine*. 2020;3**:** 1-8.
16. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using convolutional neural networks. *PloS one*. 2017;12**:** e0177544.
17. Aresta G, Araújo T, Kwok S, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*. 2019;56**:** 122-139.
18. Chen M, Zhang B, Topatana W, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *npj Precision Oncology*. 2020;4**:** 1-7.
19. Courtiol P, Maussion C, Moarii M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*. 2019;25**:** 1519-1525.
20. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*. 2020;1**:** 789-799.
21. Ramos-Vara J, Miller M. When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry—the red, brown, and blue technique. *Vet Pathol*. 2014;51**:** 42-87.

22. Bertram CA, Aubreville M, Marzahl C, Maier A, Klopfleisch R. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. *Scientific Data*. 2019;6**:** 1-9.
23. Veta M, Heng YJ, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Medical image analysis*. 2019;54**:** 111-121.
24. Veta M, van Diest PJ, Willems SM, et al. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med Image Anal*. 2015;20**:** 237-248.
25. Bertram CA, Veta M, Marzahl C, et al. *Are pathologist-defined labels reproducible? Comparison of the tupac16 mitotic figure dataset with an alternative set of labels*Springer, Cham; 2020.
26. Aubreville M, Bertram C, Donovan T, Marzahl C, Maier A, Klopfeisch R. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Sci Data*. 2020**:** 1-10.
27. Marzahl C, Bertram CA, Aubreville M, et al. Are fast labeling methods reliable? A case study of computer-aided expert annotations on microscopy slides. *arXiv preprint arXiv:200405838*. 2020.
28. Wilm F, Bertram CA, Marzahl C, et al. How Many Annotators Do We Need?--A Study on the Influence of Inter-Observer Variability on the Reliability of Automatic Mitotic Figure Assessment. *arXiv preprint arXiv:201202495*. 2020.
29. Roux L, Racoceanu D, Capron F, et al. MITOS & ATYPIA-Detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. IPAL, Agency Sci, Technol Res Inst Infocom Res. *Technol Res Inst Infocom Res, Singapore, Tech Rep*. 2014.
30. Bertram CA, Aubreville M, Gurtner C, et al. Computerized calculation of mitotic count distribution in canine cutaneous mast cell tumor sections: Mitotic count is area dependent. *Vet Pathol*. 2020;57**:** 214-226.
31. Gudlaugsson E, Skaland I, Janssen EA, et al. Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer. *Histopathology*. 2012;61**:** 1134-1144.
32. Stålhammar G, Robertson S, Wedlund L, et al. Digital image analysis of Ki67 in hot spots is superior to both manual Ki67 and mitotic counts in breast cancer. *Histopathology*. 2018;72**:** 974-989.
33. Feuchtinger A, Stiehler T, Jutting U, et al. Image analysis of immunohistochemistry is superior to visual scoring as shown for patient outcome of esophageal adenocarcinoma. *Histochem Cell Biol*. 2015;143**:** 1-9.
34. Pantanowitz L, Hartman D, Qi Y, et al. Accuracy and efficiency of an artificial intelligence tool when counting breast mitoses. *Diagn Pathol*. 2020;15**:** 1-10.
35. Steiner DF, MacDonald R, Liu Y, et al. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am J Surg Pathol*. 2018;42**:** 1636-1646.
36. Aeffner F, Wilson K, Bolon B, et al. Commentary: Roles for Pathologists in a High-throughput Image Analysis Team. *Toxicol Pathol*. 2016;44**:** 825-834.
37. McAlpine ED, Michelow P. The cytopathologist's role in developing and evaluating artificial intelligence in cytopathology practice. *Cytopathology*. 2020.