scientific **data**

OPEN

DATA DESCRIPTOR

Pan-tumor CAnine cuTaneous Cancer Histology (CATCH) dataset

Frauke Wilm[1]✉, Marco Fragoso[2], Christian Marzahl[1], Jingna Qiu[3], Chloé Puget[2], Laura Diehl[2], Christof A. Bertram[4], Robert Klopfleisch[2], Andreas Maier[1], Katharina Breininger[3,6] & Marc Aubreville[5,6]

Due to morphological similarities, the differentiation of histologic sections of cutaneous tumors into individual subtypes can be challenging. Recently, deep learning-based approaches have proven their potential for supporting pathologists in this regard. However, many of these supervised algorithms require a large amount of annotated data for robust development. We present a publicly available dataset of 350 whole slide images of seven different canine cutaneous tumors complemented by 12,424 polygon annotations for 13 histologic classes, including seven cutaneous tumor subtypes. In inter-rater experiments, we show a high consistency of the provided labels, especially for tumor annotations. We further validate the dataset by training a deep neural network for the task of tissue segmentation and tumor subtype classification. We achieve a class-averaged Jaccard coefficient of 0.7047, and 0.9044 for tumor in particular. For classification, we achieve a slide-level accuracy of 0.9857. Since canine cutaneous tumors possess various histologic homologies to human tumors the added value of this dataset is not limited to veterinary pathology but extends to more general fields of application.

Background & Summary

The skin and soft tissue are the most common anatomical sites for canine neoplasms[1] and the segmentation and classification of canine cutaneous tumors are routine tasks for veterinary pathologists. Especially different types of round cell tumors, which can have similar morphologies, are oftentimes hard to distinguish on standard histologic stainings[2,3]. Tumor-specific immunohistochemical (IHC) stainings can support the pathologist in this regard but are considerably more expensive, time-consuming, and still might not provide reliable results for undifferentiated tumors[2]. Deep learning-based algorithms can assist the pathologist in segmenting and classifying cutaneous tumors on standard Hematoxylin & Eosin (HE) staining and have successfully been applied in various works[4–10]. These algorithms, however, are often criticized for requiring vast amounts of labeled training data[11]. Therefore, publicly available datasets have become increasingly popular, as they reduce annotation costs for recurring pathological research questions and improve the comparability of computer-aided systems developed on these datasets.

Most existing open access datasets for segmentation in histopathology originated from computer vision challenges. Table 1 provides a collection of recently published datasets. These datasets not only differ in the anatomical location of the tumor and thereby the annotation classes, but also in the labeling method used for annotating the image data. Datasets consisting of small image patches, with only one tissue class present, are usually labeled on image level, whereas datasets with complete whole slide images (WSIs) are typically annotated with polygon contours. The CAMELYON[12], the BACH (Grand Challenge on BreAst Cancer Histology images)[13], and the BRACS (BReAst Carcinoma Subtyping)[14] dataset addressed lesion detection and classification for breast cancer and provide a mixture of image-level and contour annotations. Whereas the CAMELYON challenge focused on the detection of metastatic regions as a binary task, the latter two were designed for the classification into normal tissue and multiple lesion subtypes. The PAIP (Pathology Artificial Intelligence Platform)[15] WSI dataset addressed the detection of neoplasms in liver tissue as a binary segmentation task. In contrast to the aforementioned datasets, which focused on lesion detection and classification in a specific tumor region, the ADP (Atlas

[1]Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. [2]Institute of Veterinary Pathology, Freie Universität Berlin, Berlin, Germany. [3]Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. [4]Institute of Pathology, University of Veterinary Medicine, Vienna, Austria. [5]Technische Hochschule Ingolstadt, Ingolstadt, Germany. [6]These authors jointly supervised this work: Katharina Breininger, Marc Aubreville. ✉e-mail: frauke.wilm@fau.de

Dataset	Year	Organ	Images	Image Type	Classes	Class Details	Annotation Type
CAMELYON16[12]	2016	breast	399	whole slide images	2	tumor types	contours
CAMELYON17[12]	2017	breast	1000	whole slide images	5	tumor stagings	image-level (all images)
					3	tumor types	contours (50 images)
BACH[13]	2018	breast	500	patches	4	tumor types	image-level
			40	whole slide images	4	tumor types	contours (10 images)
PAIP[15]	2019	liver	100	whole slide images	2	whole tumor + viable tumor	contours
ADP[16]	2019	multi-organ	17668	patches	57	benign tissues	image-level
BRACS[14]	2021	breast	4537	patches	7	tumor types	image-level
		breast	574	whole slide images	7	tumor types	image-level
DROID[17]	2021	ovarian	193	whole slide images	11	ovarian tissue + tumor types	contours
		breast	361	whole slide images	4	tumor types	contours
		skin	99	whole slide images	32	benign + abnormal tissues	contours
		colon	101	whole slide images	38	benign + abnormal tissues	contours
CATCH (ours)	2022	skin	350	whole slide images	13	skin tissue + tumor types	contours

Table 1. Publicly available datasets for segmentation tasks on histological specimens.

of Digital Pathology)[16] and the DROID (Diagnostic Reference Oncology Imaging Database)[17] include images from multiple organs. Furthermore, they significantly exceed comparable datasets in terms of annotation classes. Whereas the ADP provides small tissue-specific patches labeled on image level, the DROID provides extensive polygon annotations on WSIs.

In this work, we present a dataset of 350 WSIs of seven canine cutaneous tumor subtypes, which we have named Canine cuTaneous Cancer Histology (CATCH) dataset. As opposed to human samples, veterinary datasets are less affected by data-privacy concerns, which makes them more suited for public access. Furthermore, previous work has demonstrated homologies between canine and human cutaneous tumors[18–20] which supports the relevance of publicly available databases for both species. We provide contour annotations for six tissue classes and seven tumor subtypes. With 12,424 annotations and 13 classes, this dataset exceeds most publicly available datasets in annotation extent and label diversity. We validated annotation quality by evaluating the inter-observer variability of three pathologists on a subset of the presented dataset with high concordance for most annotation classes. Furthermore, we present results for two computer vision tasks on the presented dataset. We first segmented the WSI into background, tumor, and the four most prominent tissue classes (epidermis, dermis, subcutis, and a joint class of inflammation and necrosis). We evaluated the segmentation result using the class-wise Jaccard coefficient, resulting in an average score of 0.7047 on our test set. Afterward, we classified the predicted tumor regions into one of seven tumor subtypes, achieving a slide-level accuracy of 98.57% on the test set. These results, achieved by standard architectures, are the first published results of computer vision algorithms trained on the CATCH dataset and can serve as a baseline for the development of more complex architectures or training strategies. Furthermore, the successful training of these architectures validates dataset consistency. The dataset, as well as the annotation database, is publicly available on The Cancer Imaging Archive (TCIA)[21]. Code examples for the methods presented in this work, along with a slide-level overview of the train-test split used for model development, can be obtained from our GitHub repository (https://github.com/DeepPathology/CanineCutaneousTumors).

Methods

Sample selection and preparation. In total, 350 cutaneous tissue samples from 282 canine patients were selected retrospectively from the biopsy archive of the Institute for Veterinary Pathology of the Freie Universität Berlin. Use of these samples was approved by the local governmental authorities (State Office of Health and Social Affairs of Berlin, approval ID: StN 011/20). All specimens were submitted by veterinary clinics or surgeries for routine diagnostic examination of neoplastic disease. As to local regulations, no ethical vote is required for these samples. No additional harm or pain was induced in the course of this study. Samples were chosen uniformly from seven of the most common canine cutaneous tumors, according to pathology reports. The case selection was guided by sufficient tissue preservation and the presence of characteristic histologic features for the corresponding tumor subtypes. Samples from the same canine patient were obtained from spatially separated sections of the same tumor or different neoplasms of the same subtype. All samples were routinely fixed in formalin, embedded in paraffin, and tissue sections were stained with H&E. 303 of the sections were digitized with the Leica ScanScope CS2 linear scanning system at a resolution of 0.2533 (40X objective lens). Due to practical feasibility, 47 slides were digitized with a different, but very similar scanning system (Leica AT2) at the same magnification and a resolution of $0.2524\frac{\mu m}{px}$ (40X objective lens).

Annotation workflow. All WSIs were annotated using the open source software SlideRunner[22]. The WSIs were predominantly (82%) annotated by the same pathologist (M.F.). The remaining annotations were gathered by three medical students in their 8th semester who were supervised by the leading pathologist (M.F.). M.F. later reviewed these annotations for correctness and completeness. Overall, annotations were gathered for seven

Class	Annotation Polygons	Annotation Area^2
epidermis	3188	2244.57
dermis	3423	16561.82
subcutis	2850	7367.62
bone	51	216.86
cartilage	16	32.15
inflammation & necrosis	719	2048.54
melanoma	379	6836.61
histiocytoma	369	2941.12
plasmacytoma	377	4750.31
trichoblastoma	423	9072.10
mast cell tumor	161	9329.85
squamous cell carcinoma	337	3513.56
peripheral nerve sheath tumor	131	11108.78
total	12424	76023.89

Table 2. Annotated polygons and area per tissue class.

canine cutaneous tumor subtypes as well as six additional tissue classes: epidermis, dermis, subcutis, bone, cartilage, and a joint class of inflammation and necrosis. The open source online platform EXACT[23] was used to monitor slide and annotation completeness.

Data Records

We provide public access to the full-resolution dataset on TCIA[21]. In total, the dataset consists of 350 WSIs – 50 each for seven cutaneous tumor subtypes: melanoma, mast cell tumor (MCT), squamous cell carcinoma (SCC), peripheral nerve sheath tumor (PNST), trichoblastoma, and histiocytoma. The WSIs are stored in the pyramidal Aperio file format (*.svs*), allowing direct access to three resolution levels ($0.25\frac{\mu m}{px}$; $1\frac{\mu m}{px}$; $4\frac{\mu m}{px}$).

In total, the 350 WSIs are accompanied by 12,424 polygon area annotations. Table 2 provides an overview of the annotated polygons and the overall annotated area per tissue class. The annotated polygons are provided in the annotation format of the Microsoft Common Objects in Context (MS COCO) dataset[24] as well as an SQLite3 database. For the MS COCO format, we have sorted the polygons in increasing order of their hierarchy level, i.e. polygons enclosed by another will be read out after their enclosing polygon. This ordering of polygons can be useful when, for instance, creating annotation masks from the annotation file. These annotation files can also be downloaded from TCIA[21].

Dataset visualization. For visualization of annotations as overlays on top of the original WSIs, we encourage researchers to use one of the following two alternatives:

SlideRunner. SlideRunner can be used to visualize the annotations collected in the SQLite3 database. Furthermore, the software allows to set up an additional annotator and extend the database with custom classes and polygon annotations. In our GitHub repository, we provide two code examples to convert SlideRunner annotations into the MS COCO format and vice versa. Figure 1a illustrates an exemplary WSI with pathologist annotations in the SlideRunner user interface.

EXACT. EXACT enables the collaborative analysis of the dataset with integrated annotation versioning. Furthermore, the REST-API of EXACT allows offline usage and direct interaction with custom machine learning frameworks. The presented dataset can be integrated as a demo dataset into EXACT which enables a direct download of the polygon annotations. Further details can be found in the documentation of EXACT. To make use of additional annotations made by the user, our GitHub repository provides a code example to convert EXACT annotations into the MS COCO format. Figure 1b shows an overview of the demo dataset in EXACT.

Technical Validation

Validation of annotations. After database collection, we ensured database consistency by using EXACT to check for and remove annotation duplicates, which occurred in rare cases due to different annotation versions. Previous work on inter-rater variability for contour delineation has demonstrated multiple influence factors on annotator disagreement for this task, such as the complexity of the medical pathology itself but also the hand-eye coordination skills of the raters[25]. Furthermore, a high level of inter-observer variability can significantly impact the performance of deep learning-based algorithms[26]. Therefore, we evaluated the inter-observer variability for the presented dataset with the help of annotations by two additional veterinary pathologists. Even though the comparison of three annotators might only provide an estimate of the full range of inter-observer variability[25], it shows the strengths and weaknesses of the provided dataset and highlights annotation classes where computer-aided systems might be of great use to pathologists. Due to the extensiveness of our dataset, we have limited the additional annotations to a 2048 μm × 2048 μm-sized region of interest (ROI) on each of the 70 test WSIs. The size of 2048 μm × 2048 μm corresponds to the patch size used for training the segmentation algorithm
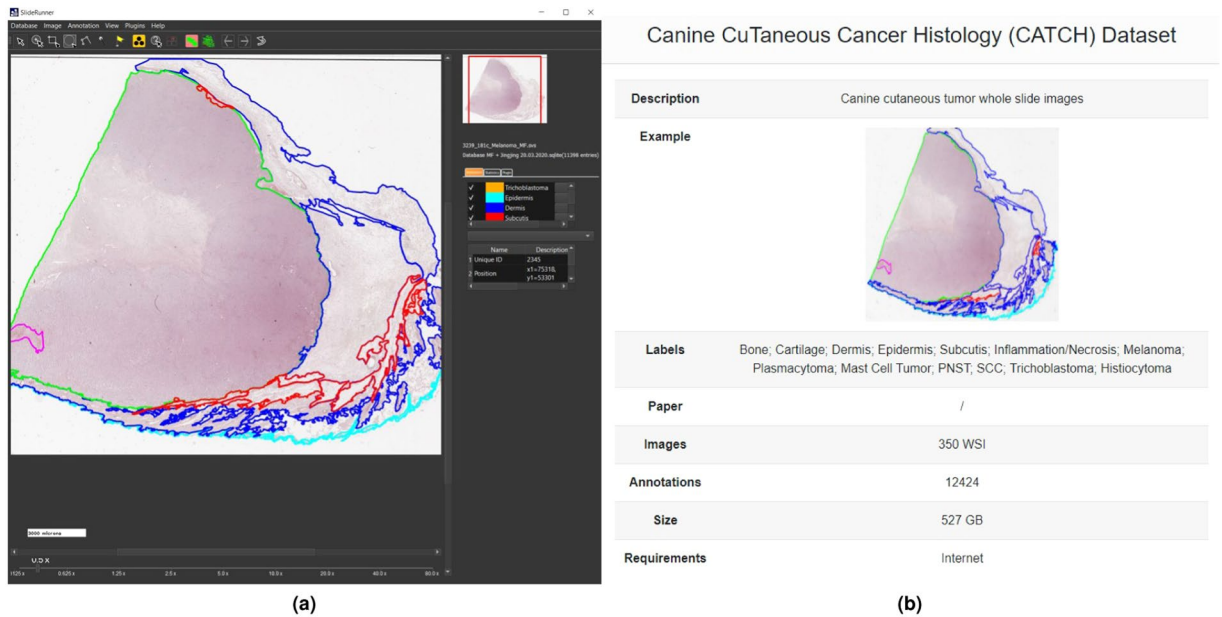
Fig. 1 User interfaces of recommended open source software tools for dataset visualization. (**a**) SlideRunner. (**b**) EXACT.

elaborated in the subsequent section. For the selection of these ROIs, we used a uniform sampling across annotation classes to counteract the class imbalance within our dataset. We then positioned the ROIs on a randomly selected vertex of a polygon of the chosen class to explicitly choose tissue boundaries where inter-annotator variability becomes most apparent. Figure 2 visualizes four exemplary patches with a high inter-rater agreement for the first two examples and a low inter-rater agreement for the second two. For quantitative evaluation of the inter-annotator variability, we computed CI_{pair}[27] as the average pair-wise Jaccard similarity coefficient for each unique pair of raters i, j and the generalized conformity index CI_{gen}[27] defined as:

$$CI_{pair,c} = \frac{2}{C(C-1)} \sum_{pairs\ ij} \frac{A_{c,i} \cap A_{c,j}}{A_{c,i} \cup A_{c,j}} \qquad CI_{gen,c} = \frac{\sum_{pairs\ ij} A_{c,i} \cap A_{c,j}}{\sum_{pairs\ ij} A_{c,i} \cup A_{c,j}},$$

(1)

where A_c are the pixels annotated as class $c \in C$. These two measures have similar values with the same mutual variability between raters, but highly differ when the delineations of one rater are considerably different from the other raters[27]. Table 3 summarizes the pair-wise Jaccard coefficients for each unique pair of raters together with CI_{pair} and CI_{gen} for all annotated tissue classes. The small differences between CI_{pair} and CI_{gen} show that deviations of rater 1, who provided the annotations for the complete dataset, fall within the mutual variability of all raters.

Tumor delineation is a routine task for all pathologists and their extensive experience in this task might be the reason for the comparably high agreement on the tumor annotations with a $CI_{gen,\ tumor}$ of 0.8514. The epidermis is the uppermost layer of the skin and therefore always located at the tissue rim. Furthermore, it is visually distinctly demarcated from the subsequent dermis tissue. These unique characteristics of the epidermis ease the annotation task and might be responsible for the comparably high inter-observer concordance indicated by a $CI_{gen,\ epidermis}$ of 0.7512. The annotators showed a higher inter-rater variability for the two subsequent layers of the healthy skin – the dermis and subcutis. A closer evaluation of the class-wise confusions showed that these lower scores mostly resulted from mix-ups between these two classes. Such an example is also illustrated in the third example in Fig. 2. When combining these two annotation classes into one, the generalized conformity index increased from 0.7169 for dermis and 0.5836 for subcutis to 0.8176 for the combined class. Whereas tumor segmentation is of high relevance for most diagnostic purposes and therefore requires precise definition criteria, we do not see the same relevance for the separation of dermis and subcutis. Thus, we argue that a high inter-rater variability for these tissue classes does not lower the diagnostic interpretability of a segmentation algorithm trained with annotations biased by how these two classes were defined.

With a generalized conformity index of 0.3302, the concordance for inflammation and necrosis was particularly low. These results, however, were not surprising as these structures are typically far less distinctly demarcated from surrounding tumor tissue due to two biological concepts that have to be considered: Firstly, necrotic areas can frequently be found within tumors where angiogenesis could not keep up with the aggressive growth of the tumor. Furthermore, secondary inflammations can be observed within tumors or at the tumor margin due to the immune system reacting to the neoplasm. Both of these biological mechanisms can result in areas that exhibit neoplastic as well as necrotic or inflammatory characteristics, which makes a precise separation from tumor tissue difficult. Such an example is shown in the last row of Fig. 2. Here, all pathologists have annotated an inflamed region located next to the outermost epidermis. Whereas pathologist 1 has annotated the adjacent region as tumor, pathologists 2 and 3 have extended the inflamed region to the rim of the tumor region in the
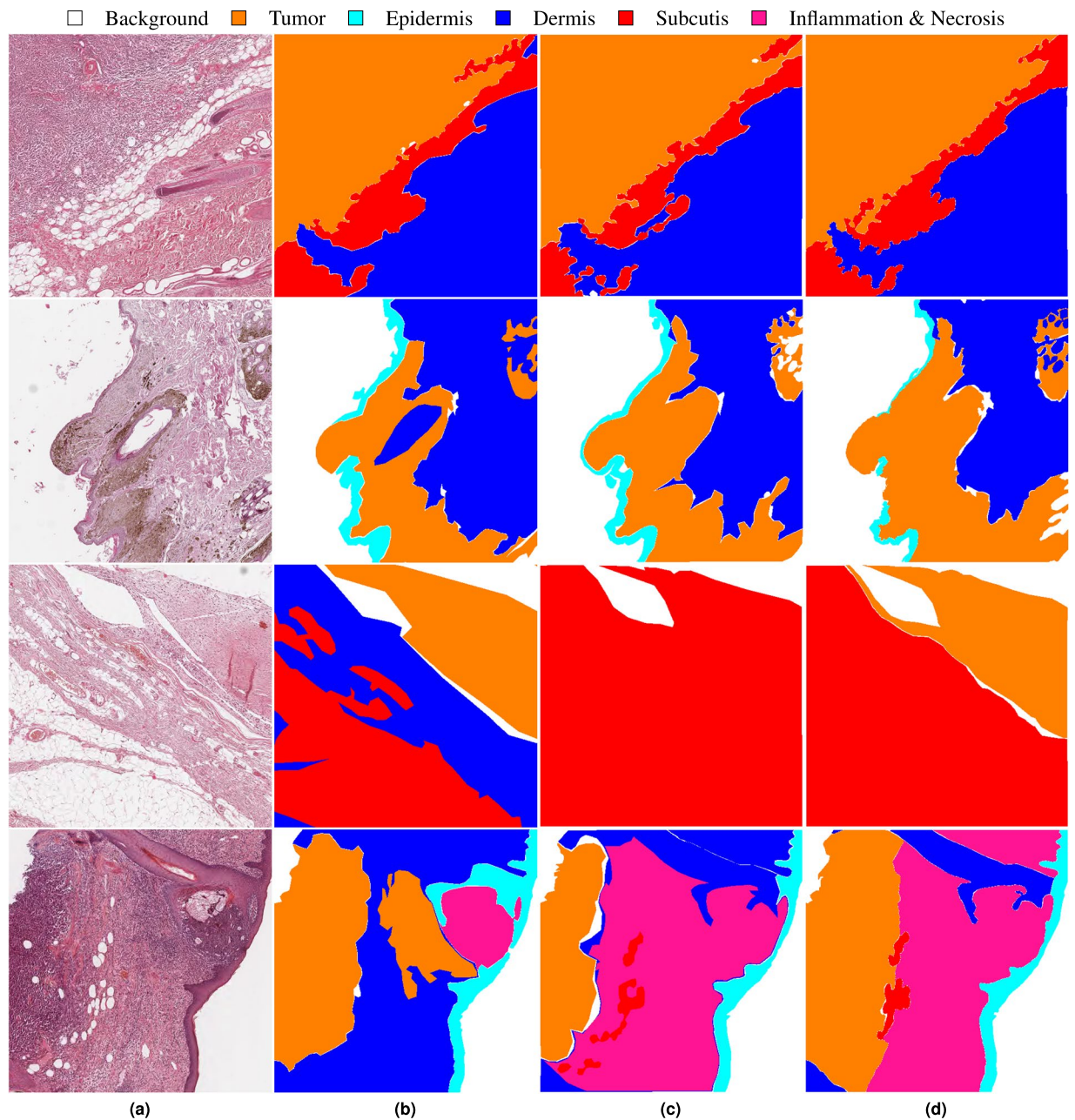
Fig. 2 Inter-rater variability for four exemplary test patches. (**a**) Original patch. (**b–d**) Annotations of pathologists. The first two rows show examples with a high inter-rater concordance and the second two rows examples with a low inter-rater concordance. The third example shows different definitions for dermis (blue) and subcutis (red) whilst the fourth example shows a high variation for inflammation & necrosis (pink).

Class	$CI_{r1,r2}$	$CI_{r1,r3}$	$CI_{r2,r3}$	CI_{pairs}	CI_{gen}
background	0.9130	0.9239	0.9319	0.9229	0.9229
tumor	0.8501	0.8506	0.8535	0.8514	0.8514
epidermis	0.8025	0.7228	0.7270	0.7508	0.7512
dermis	0.6875	0.7015	0.7659	0.7183	0.7169
subcutis	0.5254	0.5872	0.6421	0.5849	0.5836
inflammation & necrosis	0.2823	0.3040	0.3895	0.3253	0.3302

Table 3. Class-wise conformity index computed for all unique pairs of annotators. CI_{pairs} averages the pair-wise conformity indices whereas CI_{gen} is a generalized version of the Jaccard coefficient.
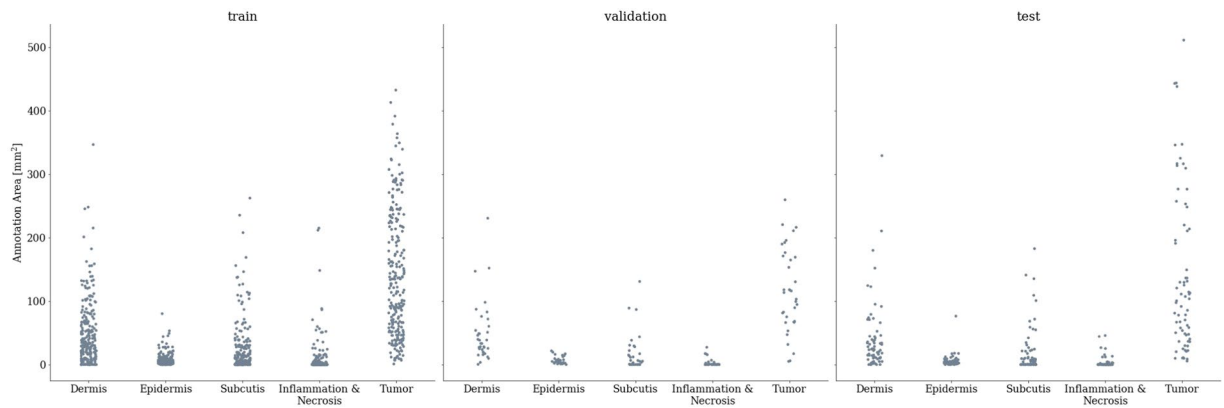
Fig. 3 Distribution of annotation area per class across dataset splits. The train split consisted of 245 whole slide images, the validation set of 35 whole slide images and the test set of 70 whole slide images.

left part of the patch. This tumor region has been delineated similarly by pathologists 1 and 3, whereas pathologist 2 has annotated this region much slimmer. The comparably low conformity index for this class shows the difficulty of clearly separating tissue areas that show a transition between two classes. This limitation of the provided annotations should be considered when evaluating the segmentation results of algorithms trained on the presented dataset.

Overall, the experiments show a high inter-observer agreement for tumor vs. non-tumor, which is highly relevant for most tasks in histopathology. However, they also highlight the difficulty of accurately demarcating necrotic or inflammatory reactions from the surrounding tumor cells. At the same time, it has to be considered that the selected regions for the inter-observer experiments were deliberately placed at tissue transitions and thereby rather over- than underestimate the inter-observer variability. Taking into account that the provided annotations mainly consisted of large connected tissue areas with little tissue interaction, we expect a considerably higher agreement for the complete dataset.

Dataset validation through algorithm development. For further validation of the dataset, we evaluated two convolutional neural network (CNN) architectures for the task of tissue segmentation and tumor subtype classification. For both tasks, we used the same dataset split: For each of the seven tumor subtypes, we randomly selected 35 WSIs for training, five for validation, and ten for testing. Thereby, we ensured equal distribution of tumor subtypes in each split. Even though WSIs from the same canine patient showed different tissue sections, we maintained a dataset split at patient level to avoid data leakage. Figure 3 visualizes the distribution of annotated area per class across the WSIs of the train, validation, and test split. For simplicity, we have combined all tumor subtypes into one class for tumor segmentation, and consider the tumor subtypes separately only during tumor classification. The visualization shows similar distributions for all splits, which ensures that our test set evaluations are representative for the complete data distribution. However, the distributions also highlight the high class imbalance within the dataset which has to be considered during the development of algorithms for computer-aided tasks. A detailed overview of the slide-level split can be obtained from the GitHub repository in form of a comma-separated value (*.csv*) table together with code for implementing the CNN architectures presented in the subsequent sections.

Tissue segmentation. For the task of tissue segmentation, we trained a UNet[28] to distinguish between four non-neoplastic tissue classes (epidermis, dermis, subcutis, and inflammation combined with necrosis) and all tumor subtypes combined into one tumor label. These five classes were accompanied by a sixth background class. For this background class, we used Otsu's adaptive thresholding[29] to compute a white threshold for each slide and assigned the background label to all non-annotated pixels that exceeded this white value. Overall, this resulted in six classes used for training the segmentation network. Figure 4 visualizes the annotation taxonomy and highlights the classes used for segmentation in green. Due to the low diagnostic significance and limited availability of bone and cartilage annotations, we excluded these classes from training and evaluating the proposed methods. Non-annotated tissue areas were also excluded from training and evaluation.

For segmentation, we used the fastai[30] UNet implementation with a ResNet18[31] backbone pre-trained on ImageNet[32]. Image patches of 512×512 pixels at a resolution of $4\frac{\mu m}{px}$ (2.5X), which corresponds to a tissue size of $2048 \times 2048\,\mu m^2$, were used as input. We decided to use this 16-fold down-sampled resolution because input patches then covered more context, which has shown to be more beneficial for segmentation results in previous work[33] and was confirmed by initial experiments on the validation dataset. To limit the number of non-informative white background patches and overcome class imbalances with random sampling, we propose an adaptive patch-sampling strategy: For each slide, we initialized the class probabilities as a uniform distribution over all annotation classes used on the respective slide. For a fixed number of training patches, we first sampled a class according to the class probabilities and then randomly selected a position within one of the polygons of this class. The final training patch was centered at this pixel location. We refer to this guided selection of a fixed number of patches as pseudo-epoch[34]. After each pseudo-epoch, the model performance was
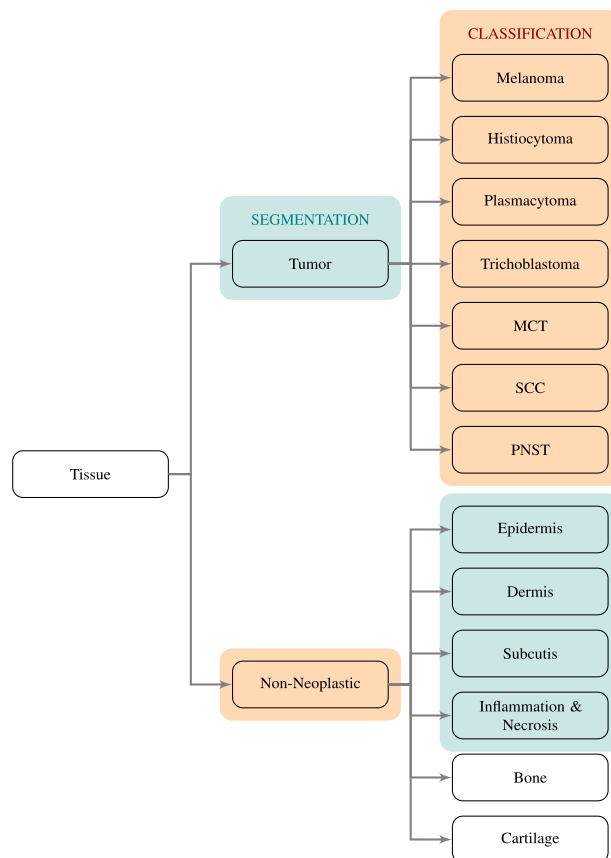
Fig. 4 Taxonomy of tissue classes. Classes highlighted in green were used for training the segmentation network and classes highlighted in orange were used to train the tumor subtype classification network. MCT: mast cell tumor, SCC:squamous cell carcinoma, PNST: peripheral nerve sheath tumor.

evaluated on a fixed number of validation patches sampled in a similar fashion. The model performance was assessed using the class-wise Jaccard similarity coefficient J_c. Prior to the next pseudo-epoch, we updated the class-wise probabilities p_c of each slide according to the complement of the corresponding class-wise Jaccard coefficient J_c:

$$p_c = 1 - J_c = 1 - \left(\frac{1}{N_v} \sum_{i=1}^{N_v} J_{c,i} \right) \qquad N_v: \text{ number of validation patches.}$$

(2)

This adaptive sampling strategy aimed for a faster convergence by over-sampling classes where the model faced difficulties. For each pseudo-epoch, we sampled ten patches per slide, resulting in 2450 training patches and 350 validation patches per pseudo-epoch. All patches were normalized using the RGB statistics of the training set, i.e. subtract the mean and divide by the standard deviation of all tissue-containing areas of the training WSIs. We trained the model for up to 100 pseudo-epochs and selected the configuration with the highest class-averaged Jaccard coefficient on the validation patches. We trained with a maximal learning rate of 10^{-4}, a batch size of four, and used discriminative fine-tuning[35] provided by the fastai package. During training, online data augmentation was used, composed of random flipping, affine transformations, and random lightning and contrast change. To meet the class imbalance within the data, the model was trained with a combination of the generalized Dice loss[36] and the categorical focal loss[37].

After model training, we computed a slide segmentation output using a moving-window patch-wise inference with an overlap of half the patch size, i.e. 256 pixels. In the overlap area, we averaged the class probabilities computed as softmax-output of the model predictions. This inference resulted in a three-dimensional output tensor with the slide dimensions in x- and y-direction and the number of segmentation classes in z-direction. The per-pixel labels were then computed as the class with the maximum entry in z-direction. Figure 5 visualizes an exemplary segmentation result with the original slide and annotated regions on the left and the predicted segmentation output on the right. For quantitative performance evaluation, we accumulated the pixel-based confusion matrices of all WSIs of the test set and then computed the class-wise Jaccard similarity coefficient. Figure 6 visualizes the row-normalized accumulated confusion matrix for a resolution of $4\frac{\mu m}{px}$. The color-coding visualizes the row-wise normalization. The first column of Table 4 summarizes the class-wise Jaccard coefficients computed from the confusion matrix. Overall, the network scored a class-averaged Jaccard coefficient of 0.7047. Due to the high class imbalance, we also computed a frequency-weighted Jaccard coefficient by multiplying the
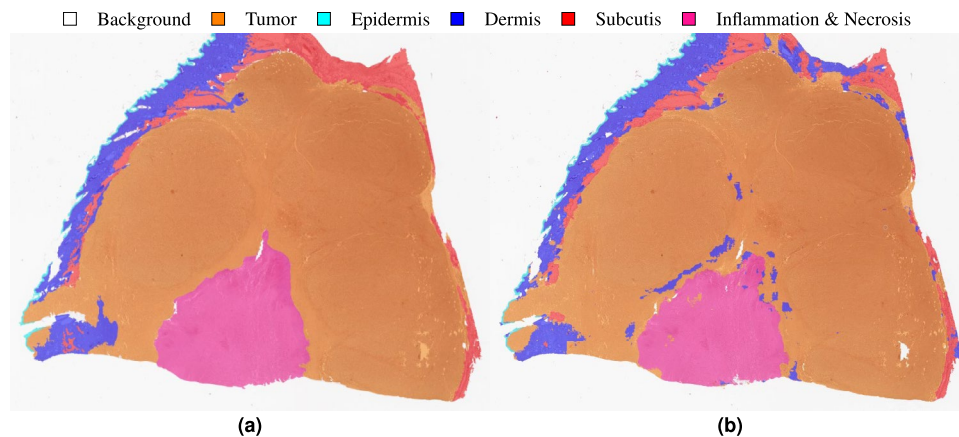
Fig. 5 Exemplary segmentation result. (**a**) Annotation. (**b**) Prediction.

		Prediction					
		Background	Tumor	Epidermis	Dermis	Subcutis	Inf. & Nec.
867,399,405	Background	0.9921	0.0018	0.0011	0.0024	0.002	0.0006
597,038,010	Tumor	0.0096	0.9503	0.0048	0.0271	0.002	0.0062
22,674,056	Epidermis	0.0424	0.035	0.8441	0.0617	0.0034	0.0134
171,757,230	Dermis	0.0233	0.113	0.0113	0.7815	0.0587	0.0123
83,540,748	Subcutis	0.0413	0.0411	0.0015	0.0799	0.8176	0.0186
13,034,954	Inf. & Nec.	0.0308	0.3883	0.0152	0.048	0.0259	0.4919

(Ground Truth)

Fig. 6 Segmentation confusion matrix (pixel-based). The numbers on the left summarize the pixel-count per class.

Class	WSI	ROI$_{r1}$	ROI$_{r2}$	ROI$_{r3}$
background	0.9757	0.9362	0.9355	0.9430
tumor	0.9044	0.8399	0.7911	0.8193
epidermis	0.6661	0.7033	0.6685	0.6167
dermis	0.6753	0.6988	0.6285	0.6740
subcutis	0.7043	0.6232	0.5101	0.6435
inflammation & necrosis	0.3023	0.2816	0.1331	0.1646

Table 4. Class-wise Jaccard similarity score for all WSIs of the test set annotated by rater 1 and the region of interests evaluated in the inter-rater experiments annotated by raters 1, 2, and 3.

class-wise coefficients with the class-wise ratio of the respective pixels in the ground truth and summing up over all values. This yielded a frequency-weighted coefficient of 0.9001. The results show that especially for the background and tumor class, the network scored high Jaccard coefficients of 0.9757 and 0.9044 respectively. This could mainly be attributed to a high sensitivity, i.e. few areas were overlooked. However, the algorithm misclassified a relatively large amount of non-neoplastic pixels as cancerous, especially inflamed and necrotic regions, yielding a comparably low Jaccard coefficient of 0.3023 for this combined class. Yet, this behavior meets clinical demands, as the costs of falsely classifying healthy tissue as tumor are far lower than overlooking neoplastic regions which could at worst lead to a false diagnosis. The high amount of neoplastic and inflammatory regions misclassified as tumor can again be ascribed to the necrotic and inflamed regions often interspersed between tumor cells, which makes a clear distinction difficult. The results of our inter-observer experiments have shown that a precise definition of these classes can be difficult even for trained pathologists. Therefore, algorithmic confusions between these classes should always be evaluated with the above-mentioned challenge in mind.

To evaluate whether training the algorithm on annotations of a single rater introduced a bias towards this rater, we additionally computed ROI Jaccard coefficients for the test patches included in the inter-rater experiments. These are summarized in Table 4. Overall, the results do not show a clear bias towards rater 1 for most classes, as the results fall within the range of the inter-annotator conformity indices. For the combined class of inflammation and necrosis, the algorithm shows a tendency towards rater 1 but still shows a very poor agreement with a Jaccard score of 0.2816. This again highlights the difficulty of accurately defining this class. When comparing the ROI Jaccard coefficients of rater 1 to the WSI Jaccard coefficients, the algorithm shows mostly lower performance, which underlines the increased complexity of the ROIs, which were deliberately placed at tissue transitions.

To evaluate whether the morphology of certain tumor subtypes within the dataset made a precise differentiation of the tissue classes more difficult, we also computed the class-wise Jaccard coefficients per tumor subtype.

Class	Tumor WSI Jaccard Score
melanoma	0.8394
histiocytoma	0.8714
plasmacytoma	0.9419
trichoblastoma	0.9650
mast cell tumor	0.8721
squamous cell carcinoma	0.7185
peripheral nerve sheath tumor	0.9166

Table 5. Tumor Jaccard similarity score computed from the confusion matrix accumulated over all ten test WSIs of the respective subtype.

These results are summarized in Table 5. The results show that the network performed exceptionally well for trichoblastoma with a Jaccard coefficient of 0.9650 but was challenged by SCC samples with a Jaccard coefficient of 0.7185. The SCC confusion matrix revealed that 60.62% of the pixels annotated as inflammation or necrosis were falsely classified as tumor. SCC, however, is known to cause severe inflammatory reactions[38]. Infiltration of these inflammatory cells in-between the nests or trabecular of neoplastic epidermal cells can make an accurate distinction of both classes difficult, which could also be seen when evaluating the inter-annotator variability on the presented dataset.

Recent work has shown that deep learning-based models face difficulties when being applied to WSIs digitized by a slide scanning system different from the one used for training the algorithm[39,40]. Due to practical feasibility, a subset of the presented dataset was digitized with a different slide-scanning system. We compensated for this by ensuring a similar distribution of scanner domains in our training and test split and observed similar performances on our test data with mean Jaccard coefficients of 0.7026 for the ScanScope CS2 and 0.6986 for the AT2. Nevertheless, we are currently creating a multi-scanner dataset of a subset of the data presented in this work and future work will evaluate the transferability of trained models to unseen scanner domains and the development of domain-invariant algorithms.

Tumor classification. Besides tissue segmentation, we trained an additional CNN for tumor subtype classification. For this, an EfficientNet-B0[41] was trained to distinguish between all seven tumor classes. We combined all non-neoplastic tissues used for training the segmentation network into one rejection class, allowing for inference on patches where no tumor was present. This resulted in eight classes used for training the classification network. Due to the high morphological resemblance of round cell tumors, where cell-level information might be required to distinguish the individual subtypes, we used the original scanning resolution of $0.25\frac{\mu m}{px}$ (40X) for classification. This corresponds to the diagnostic workflow of pathologists, who would first use a lower resolution to locate the tumor region and then use a higher resolution to classify the tumor. To retain as much context as possible, we increased the patch size to 1024×1024 pixels. We used the same train-test split as used for tissue segmentation and trained the network for 100 pseudo-epochs with ten patches per slide in each epoch. Fixing this number of sampled patches per slide ensured that each tumor was represented equally and network training was not affected by the very differently sized tumors highlighted by Table 2, where PNST annotations make up for almost 15% of the overall annotated area whereas histiocytoma annotations only for about 4%. For each slide, we set the probability of sampling a tumor patch seven times higher than the probability of sampling a non-neoplastic patch, as these were present in all slides of the seven tumor types, whereas tumor-specific patches were only available for the training slides of the respective tumor subtype. For the non-neoplastic patches, we ensured an equal sampling of the different tissue classes by first randomly sampling a class and then selecting a patch within one of the polygons of this class. We followed an area-based polygon sampling strategy to ensure an equal distribution of sampled patches across the annotated polygons of the respective class. Furthermore, a patch was only used for training the classification network if at least 90% of the pixels were annotated as the sampled class. All patches were normalized using the training set statistics. Similar to the segmentation network, we used online data augmentation. The network was trained with a batch size of four and a maximal learning rate of 10^{-3}. We used the Adam optimizer and trained the model with cross-entropy loss. We used the mean patch-level accuracy to guide the model selection process.

To combine the pixel-wise segmentation with the patch-wise tumor subtype classification, we propose the following slide inference pipeline, visualized in Fig. 7: First, a WSI is segmented into six tissue classes, using the segmentation network described in the previous subsection. The spatial resolution of the pixel-wise segmentation map corresponds to the WSI at the chosen resolution of $4\frac{\mu m}{px}$, which represents a 16-fold down-sampling in each dimension. This segmentation map is up-sampled to the original resolution and only patches that were fully segmented as tumor obtain a patch label by the tumor subtype classification network. These patch classifications are then combined into a slide label by using majority voting. By training the tumor subtype classification network on an additional rejection class comprised of non-neoplastic tissue, we aimed to compensate for false-positive tumor segmentation predictions. If the classification network assigned the rejection label for these patches, they were excluded from the subsequent majority voting. Inference time for this pipeline was measured using an NVIDIA Quadro RTX 8000 graphics processing unit. WSI segmentation took 15 ± 7 sec ($\mu \pm \sigma$) for an average of 405 ± 177 patches ($\hat{=}$ 37 msec per patch). In our two-stage inference pipeline, only patches from areas segmented as tumor were passed on to the tumor subtype classification network. This significantly reduced the
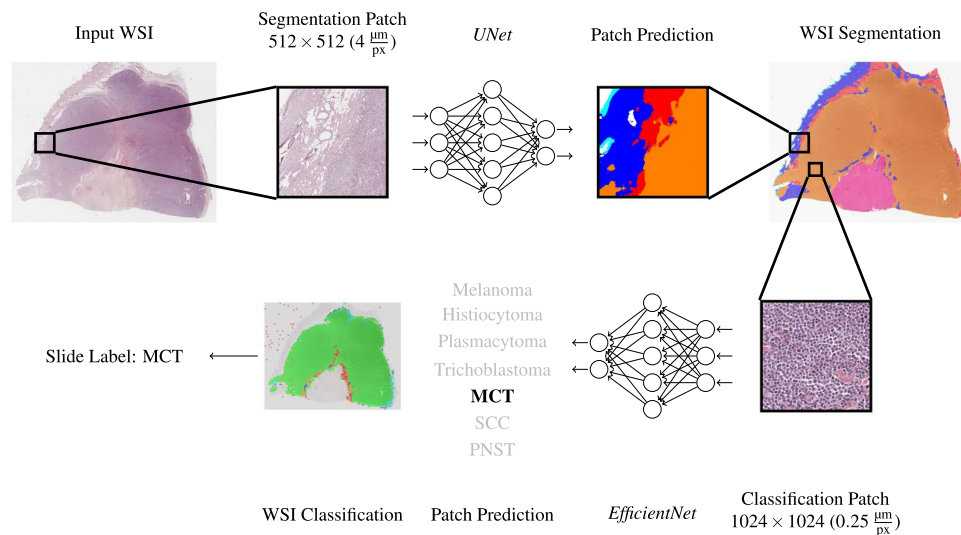
Fig. 7 Patch segmentation and classification pipeline. Due to different resolutions and patch sizes between the segmentation and the classification task, a single segmentation patch holds multiple classification patches. Only patches segmented as tumor are classified into a tumor subtype. MCT: mast cell tumor, SCC: squamous cell carcinoma, PNST: peripheral nerve sheath tumor.

number of patches to be predicted, however, due to the higher resolution of the classification network, we still measured inference times of 6 ± 5 min for classification with an average of 2472 ± 1873 patches per slide ($\widehat{=} 155$ msec per patch). The comparatively high variance within these inference times resulted from the high variance of tissue and tumor area within the test set. On average, the WSIs were sized $6.47 \pm 2.89 \times 10^9$ pixels at the original resolution.

When applying the slide inference pipeline to all 70 WSIs from the test set, we classified 69 WSIs correctly, yielding a slide classification accuracy of 98.57%. The misclassified slide is depicted in the upper example of Fig. 8. Here, the model falsely labeled a trichoblastoma slide as melanoma. A closer examination of this slide revealed a high number of undifferentiated, pleomorphic cells, i.e. tumor cells of varying shapes and sizes, visualized in the magnified tumor region on the upper right side of Fig. 8. The region shows characteristics of epithelial tumors, the superordinate tumor category of trichoblastomas, but melanomas, too, can be composed of epitheloid cells. Melanomas are typically highly pleomorphic, which might have caused the misclassification as melanoma. The upper example in Fig. 8 also shows that some misclassified patches are located on the white WSI background. A closer look at these areas revealed small parts of detached tissue or dust artifacts, which were mistaken as tumor by the segmentation network and then falsely passed on to the classification network. This could be circumvented by additionally training the classification network on background patches or applying a post-processing step such as morphological closing to the segmentation output.

To evaluate whether some tumor subtypes were more difficult for the classification network than others, i.e. the majority voting was affected by many false patch classifications, we evaluated the confusion matrix of the tumor subtype patch classification, of which a row-normalized version is shown in Fig. 9 with the color-coding again representing the normalization. This confusion matrix only includes patches that were segmented as tumor and thereby passed on to the classification network. The first row of the matrix shows that the segmentation network passed on 16238 false-positive tumor patches to the classification network of which 63.46% were recovered by the rejection class. From the remaining rows, we computed tumor-wise recalls and precisions, i.e. the ratio of all patches correctly classified as the respective subtype to all patches labeled or respectively predicted as the corresponding subtype. These metrics, summarized in Table 6, only consider confusion among tumor subtypes and not with the non-neoplastic class. The confusion matrix and the results in Table 6 show that SCC generally was the most difficult class for the network to distinguish. Looking at the results in detail, however, the comparably low F_1 score of 0.8773 can mostly be attributed to the low classification precision, meaning a lot of tumor patches were falsely classified as SCC. A closer look at the classification outputs showed that these misclassifications were mostly located at the tumor boundaries. This observation could be linked to the severe inflammatory reactions that are typically caused by SCCs[38]. During training, inflammatory reactions to tumor growth might have been more common for SCC samples than for other subtypes, which might have caused the model to mistake the interaction of tumor and inflammatory cells as SCC.

The lower example in Fig. 8 shows a melanoma sample where the majority voting yielded the correct classification label but was affected by many false patch classifications. A comparison to the ground truth annotations depicted on the lower right side of Fig. 8 reveals that this can be rooted back to a false tumor prediction of the preceding segmentation network, as only the area correctly classified as melanoma was also annotated as tumor. For this example, the rejection class could not fully recover the errors made by the segmentation network. Even though the majority voting resulted in the correct slide label for this example, one should always take into consideration some measure of confidence for the majority voting, indicating how difficult the final decision was. Determining the patch-level entropy, for instance, could highlight slides where the distribution of patch
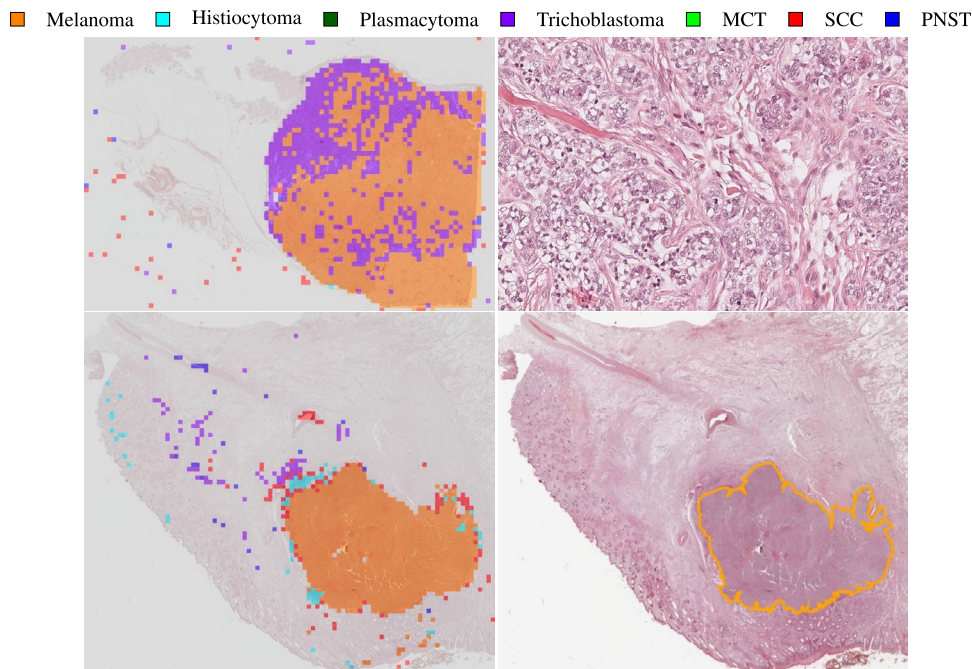
Fig. 8 Exemplary classification results. The upper example shows a trichoblastoma sample misclassified as melanoma with the classification output on the left and the magnified tumor region on the right. The tumor region shows a high number of pleomorphic tumor cells. The lower example shows a melanoma sample with the classification output on the left and the annotated sample on the right. The classification output shows a high ratio of misclassified patches caused by a false tumor prediction during segmentation. MCT: mast cell tumor, SCC: squamous cell carcinoma, PNST: peripheral nerve sheath tumor.

		Prediction							
		Non-Neoplastic	Melanoma	Histiocytoma	Plasmacytoma	Trichoblastoma	MCT	SCC	PNST
16,238	Non-Neoplastic	0.6346	0.0252	0.044	0.0559	0.0338	0.1096	0.0624	0.0344
12,083	Melanoma	0.0331	0.9354	0.0018	0.0002	0.0006	0.0	0.0098	0.019
8423	Histiocytoma	0.0274	0.0023	0.9414	0.0196	0.0024	0.0006	0.0063	0.0001
14,224	Plasmacytoma	0.0127	0.0135	0.0063	0.9502	0.0004	0.0101	0.003	0.0039
41,624	Trichoblastoma	0.0036	0.0349	0.0002	0.0004	0.9533	0.0	0.0065	0.001
17,963	MCT	0.0165	0.0035	0.005	0.0027	0.0009	0.9426	0.0041	0.0247
8035	SCC	0.0824	0.0114	0.0169	0.0157	0.0045	0.0005	0.8595	0.0091
42,378	PNST	0.0245	0.0021	0.0038	0.0019	0.0496	0.001	0.0213	0.8957

(Ground Truth on the left axis)

Fig. 9 Classification confusion matrix (patch-based). The numbers on the left summarize the patch-count per tumor class.

Class	Precision	Recall	F_1 score
melanoma	0.8555	0.9675	0.9081
histiocytoma	0.9399	0.9679	0.9537
plasmacytoma	0.9684	0.9624	0.9654
trichoblastoma	0.9478	0.9568	0.9522
mast cell tumor	0.9886	0.9584	0.9732
squamous cell carcinoma	0.8251	0.9367	0.8773
peripheral nerve sheath tumor	0.9783	0.9182	0.9473

Table 6. Patch-level tumor precision, recall and F_1 score per tumor subtype.

classifications across the tumor subtypes resembled a uniform distribution, i.e. scored a high entropy and the decision was made less confidently. This entropy could then be used for weighted voting of the slide label instead of a simple majority voting.

Dataset insights from algorithm development. Overall, the algorithm results on the provided database validate database quality as a successful training of a segmentation algorithm on the dataset proves the consistency of the annotations. When comparing the algorithm to the annotations of rater 1, the Jaccard scores fall within the

range of inter-annotator concordance, indicating that the provided annotations did not introduce a bias into algorithm development. Furthermore, the experiments highlighted strengths and weaknesses of the provided dataset, as for instance SCCs are more affected by inflammatory reactions, which makes them less suited for training an algorithm for a clear distinction of tumor and inflammation.

Usage Notes

All code examples are based on OpenSlide[42] for WSI processing and fastai[30] for network training. To apply the fastai modules to WSIs, we provide custom data loaders in our GitHub repository. The annotation and visualization tools used for this work–SlideRunner and EXACT–are both open source and can be downloaded from the respective GitHub repositories.

Code availability

Code examples for training the segmentation and classification architectures can be found in the form of Jupyter notebooks in our GitHub repository (https://github.com/DeepPathology/CanineCutaneousTumors). Furthermore, we provide exported fastai learners to reproduce the results stated in this work. The *datasets.csv* file lists the train, validation, and test split on slide level. For network inference, we provide two Jupyter notebooks for patch-level results (*segmentation_inference.ipynb* and *classification_inference.ipynb*) and one notebook for slide-level results. This *slide_inference.ipynb* notebook produces segmentation and classification outputs as compressed numpy arrays. After inference, these prediction masks can be visualized as overlays on top of the original images using our custom SlideRunner plugins *wsi_segmentation.py* and *wsi_classification.py*. To integrate these plugins into their local SlideRunner installation, users have to copy the respective plugin from our GitHub repository into their SlideRunner *plugin* directory. Additionally, the *slide_inference.ipynb* notebook provides methods to compute confusion matrices from network predictions and calculate class-wise Jaccard coefficients and the tumor classification recall. As mentioned previously, we provide six python modules to convert annotations back and forth between MS COCO and EXACT, MS COCO and SQLite, and EXACT and SQLite formats. This enables users to extend the annotations by custom classes or polygons in their preferred annotation format. These modules can be found in the *annotation_conversion* directory of our GitHub repository.

References

1. Dobson, J., Samuel, S., Milstein, H., Rogers, K. & Wood, J. Canine neoplasia in the UK: estimates of incidence rates from a population of insured dogs. *Journal of Small Animal Practice* **43**, 240–246 (2002).
2. Fernandez, N., West, K., Jackson, M. & Kidney, B. Immunohistochemical and histochemical stains for differentiating canine cutaneous round cell tumors. *Veterinary Pathology* **42**, 437–445 (2005).
3. Bertram, C. A. *et al.* Validation of digital microscopy compared with light microscopy for the diagnosis of canine cutaneous tumors. *Veterinary Pathology* **55**, 490–500 (2018).
4. Salvi, M. *et al.* Histopathological classification of canine cutaneous round cell tumors using deep learning: A multi-center study. *Frontiers in Veterinary Science* **8**, 294 (2021).
5. Thomas, S. M., Lefevre, J. G., Baxter, G. & Hamilton, N. A. Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. *Medical Image Analysis* **68**, 101915 (2021).
6. Jiang, Y. *et al.* Recognizing basal cell carcinoma on smartphone-captured digital histopathology images with a deep neural network. *British Journal of Dermatology* **182**, 754–762 (2020).
7. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* **25**, 1301–1309 (2019).
8. Halicek, M. *et al.* Detection of squamous cell carcinoma in digitized histological images from the head and neck using convolutional neural networks. In *Medical Imaging 2019: Digital Pathology*, vol. 10956, 109560K (International Society for Optics and Photonics, 2019).
9. Hekler, A. *et al.* Pathologist-level classification of histopathological melanoma images with deep neural networks. *European Journal of Cancer* **115**, 79–83 (2019).
10. Arevalo, J., Cruz-Roa, A., Arias, V., Romero, E. & González, F. A. An unsupervised feature learning framework for basal cell carcinoma image analysis. *Artificial Intelligence in Medicine* **64**, 131–145 (2015).
11. Marcus, G. Deep learning: A critical appraisal. Preprint at https://arxiv.org/abs/1801.00631 (2018).
12. Litjens, G. *et al.* 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *GigaScience* **7** (2018).
13. Aresta, G. *et al.* BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis* **56**, 122–139 (2019).
14. Pati, P. *et al.* Hierarchical graph representations in digital pathology. *Medical Image Analysis* **75**, 102264 (2022).
15. Kim, Y. J. *et al.* PAIP 2019: Liver cancer segmentation challenge. *Medical Image Analysis* **67**, 101854 (2021).
16. Hosseini, M. S. *et al.* Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11747–11756 (2019).
17. Stadler, C. B. *et al.* Proactive construction of an annotated imaging database for artificial intelligence training. *Journal of Digital Imaging* **34**, 105–115 (2021).
18. Prouteau, A. & André, C. Canine melanomas as models for human melanomas: Clinical, histological, and genetic comparison. *Genes* **10**, 501 (2019).
19. Ranieri, G. *et al.* A model of study for human cancer: Spontaneous occurring tumors in dogs. Biological features and translation for new anticancer therapies. *Critical Reviews in Oncology/Hematology* **88**, 187–197 (2013).
20. Pinho, S. S., Carvalho, S., Cabral, J., Reis, C. A. & Gärtner, F. Canine tumors: A spontaneous animal model of human carcinogenesis. *Translational Research* **159**, 165–172 (2012).
21. Wilm, F. *et al.* CAnine CuTaneous Cancer Histology dataset (version 1). *The Cancer Imaging Archive* https://doi.org/10.7937/TCIA.2M93-FX66 (2022).
22. Aubreville, M., Bertram, C., Klopfleisch, R. & Maier, A. SlideRunner. In *Bildverarbeitung für die Medizin 2018*, 309–314 (Springer, 2018).
23. Marzahl, C. *et al.* EXACT: A collaboration toolset for algorithm-aided annotation of images with annotation version control. *Scientific Reports* **11**, 1–11 (2021).
24. Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 740–755 (Springer, 2014).
25. Joskowicz, L., Cohen, D., Caplan, N. & Sosna, J. Inter-observer variability of manual contour delineation of structures in CT. *European Radiology* **29**, 1391–1399 (2019).

26. Wilm, F. *et al*. Influence of inter-annotator variability on automatic mitotic figure assessment. In *Bildverarbeitung für die Medizin 2021*, 241–246 (Springer, 2021).
27. Kouwenhoven, E., Giezen, M. & Struikmans, H. Measuring the similarity of target volume delineations independent of the number of observers. *Physics in Medicine & Biology* **54**, 2863 (2009).
28. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 234–241 (Springer, 2015).
29. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 62–66 (1979).
30. Howard, J. & Gugger, S. Fastai: A layered API for deep learning. *Information* **11**, 108 (2020).
31. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
32. Russakovsky, O. *et al*. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211–252 (2015).
33. Sirinukunwattana, K., Alham, N. K., Verrill, C. & Rittscher, J. Improving whole slide segmentation through visual context - a systematic study. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 192–200 (Springer, 2018).
34. Bertram, C. A., 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* **6**, 1–9 (2019).
35. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. Preprint at https://arxiv.org/abs/1801.06146 (2018).
36. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 240–248 (Springer, 2017).
37. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988 (2017).
38. Gasparoto, T. H. *et al*. Inflammatory events during murine squamous cell carcinoma development. *Journal of Inflammation* **9**, 1–11 (2012).
39. Moyes, A. *et al*. A novel method for unsupervised scanner-invariance with DCAE model. In *British Machine Vision Conference* (University of Leicester, 2018).
40. Aubreville, M. *et al*. Quantifying the scanner-induced domain gap in mitosis detection. In *Medical Imaging with Deep Learning* (2021).
41. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114 (PMLR, 2019).
42. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of Pathology Informatics* **4** (2013).

Acknowledgements

Author contributions

F.W. wrote the manuscript, carried out data analysis and developed the presented algorithms. F.W., K.B. and M.A. all contributed to method development. M.F. provided annotations for 82% of the data, reviewed the remaining annotations for correctness and completeness and provided medical expertise for discussion of the results. C.M. and J.Q. contributed to data collection and method development. C.P. and L.D. provided manual annotations for selected regions on the test data to evaluate inter-observer variability. C.B. and R.K. provided medical and A.M., K.B. and M.A. technical expertise. All authors contributed to the preparation of the manuscript and approved of the final manuscript for publication.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.