Check for updates

Veterinary Pathology
2025, Vol. 62(2) 152–160
© The Author(s) 2024

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/03009858241286806
journals.sagepub.com/home/vet
Sage

Artificial intelligence can be trained to predict *c-KIT*-11 mutational status of canine mast cell tumors from hematoxylin and eosin-stained histological slides

Chloé Puget[1*], Jonathan Ganz[2*], Julian Ostermaier[2], Thomas Conrad[1], Eda Parlak[3], Christof A. Bertram[3], Matti Kiupel[4], Katharina Breininger[5], Marc Aubreville[2], and Robert Klopfleisch[1]

Abstract

Numerous prognostic factors are currently assessed histologically and immunohistochemically in canine mast cell tumors (MCTs) to evaluate clinical behavior. In addition, polymerase chain reaction (PCR) is often performed to detect internal tandem duplication (ITD) mutations in exon 11 of the *c-KIT* gene (*c-KIT*-11-ITD) to predict the therapeutic response to tyrosine kinase inhibitors. This project aimed at training deep learning models (DLMs) to identify MCTs with *c-KIT*-11-ITD solely based on morphology. Hematoxylin and eosin (HE) stained slides of 368 cutaneous, subcutaneous, and mucocutaneous MCTs (195 with ITD and 173 without) were stained consecutively in 2 different laboratories and scanned with 3 different slide scanners. This resulted in 6 data sets (stain-scanner variations representing diagnostic institutions) of whole-slide images. DLMs were trained with single and mixed data sets and their performances were assessed under stain-scanner variations (domain shifts). The DLM correctly classified HE slides according to their *c-KIT*-11-ITD status in up to 87% of cases with a 0.90 sensitivity and a 0.83 specificity. A relevant performance drop could be observed when the stain-scanner combination of training and test data set differed. Multi-institutional data sets improved the average accuracy but did not reach the maximum accuracy of algorithms trained and tested on the same stain-scanner variant (ie, intra-institutional). In summary, DLM-based morphological examination can predict *c-KIT*-11-ITD with high accuracy in canine MCTs in HE slides. However, staining protocol and scanner type influence accuracy. Larger data sets of scans from different laboratories and scanners may lead to more robust DLMs to identify c-*KIT* mutations in HE slides.

Keywords

artificial intelligence, *c-KIT* mutation, canine cutaneous mast cell tumor, convolutional neural network, deep learning, digital pathology, exon 11, genotype prediction, machine learning

Currently, hematoxylin and eosin (HE) stained slides are the gold standard to diagnose and prognosticate canine mast cell tumors (MCTs). After identification of a canine cutaneous MCT, tumor grading according to either Patnaik (3 tiers: grade 1, 2, or 3), Kiupel (2 tiers: high grade or low grade), or both are performed.[13,18,22,33] In addition, prognostic evaluation of MCTs often includes the detection of *c-KIT* gene mutations, specifically in exons 8 and 11.[26]

The most commonly analyzed *c-KIT* mutation of canine cutaneous MCTs is the internal tandem duplication (ITD) in exon 11 (c-*KIT-11-ITD*), which results in a constitutive *c-KIT* activation leading to uncontrolled cell growth and more aggressive behavior.[27,29,35] Besides the general prognostic value of information on the mutational status, confirmation of a mutation of the *c-KIT* gene is required by the European Medicine Agency before treatment of dogs with non-resectable MCTs with the tyrosine kinase inhibitor Masitininb.[12] Polymerase chain reaction (PCR)-based *c-KIT* mutation analysis is, however, not performed routinely as it represents a supplementary cost and time factor for the dog owners. In addition, mutation analysis is often performed on formalin-fixed and paraffin-embedded material, which is usually affected by DNA

[1]Freie Universität Berlin, Berlin, Germany
[2]Technische Hochschule Ingolstadt, Ingolstadt, Germany
[3]University of Veterinary Medicine, Vienna, Austria
[4]Michigan State University, East Lansing, MI
[5]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

*Equal contribution.

Supplemental material for this article is available online.

Corresponding Author:
Robert Klopfleisch, Institut für Tierpathologie, Freie Universität Berlin, Robert-von-Ostertag-Straße 15, Berlin 14163, Germany.
Email: Robert.klopfleisch@fu-berlin.de

Table 1. PCR status of internal tandem duplication (ITD) in exon 11 of *c-KIT*, and 2-tier and 3-tier grades of 368 canine mast cell tumors.

c-KIT-11 ITD	Positive				Negative			
n	195 (53%)				173 (47%)			
Kiupel grade	High		Low		High		Low	
N	121 (62.1%)		74 (37.9%)		97 (56.1%)		76 (43.9%)	
Patnaik grade	3	2	2	1	3	2	2	1
N	55 (28.2%)	66 (33.8%)	69 (35.4%)	5 (2.6%)	35 (20.2%)	62 (35.8%)	70 (40.5%)	6 (3.5%)

fragmentation. This can lead to decreased sensitivity of the PCR, the current gold standard method for *c-KIT*-11-ITD detection.[5,17,25]

Recent studies in human oncology have shown that deep learning models (DLMs) are able to predict the genotype of tumors solely based on their microscopic morphology in HE slides. For instance, a machine-learning algorithm was trained to predict 10 specific driver mutations in human non-small cell lung carcinoma,[8,21] as well as mutations in human gliomas,[19] colorectal cancer,[15] and liver cancer[6] among others obtaining classification accuracies ranging from 0.64 to 0.93.[6,8,19,21]

This study evaluated the use of DLMs for the prediction of *c*-KIT-*11-ITD* mutations in MCTs based on HE slides. For this study, we curated a data set of digitized HE slides of canine MCTs with PCR-confirmed *c*-KIT-*11-ITD* genotype. To reflect the diversity of HE-staining variants and scanner types in diagnostic institutions, each slide was stained in 2 different laboratories and scanned with 3 different scanning devices, resulting in 6 different data sets (stain-scanner variants). DLMs were either trained and evaluated on the same data set they were trained on (same stain-scanner variant, in-domain) or trained on 1 data set and tested on unknown data sets (different stain-scanner variants, out-of-domain).

Material and Methods

Tumors

The study set was composed of 368 HE-stained histologic slides of MCT biopsies that had been submitted to the Veterinary Diagnostic Laboratory of Michigan State University between 2018 and 2022 for routine diagnostics with histologic grading as well as a prognostic MCT panel including molecular evaluation of *c-KIT-11-ITD* and *c-KIT* exon 8. The 2-tier[18] and 3-tier[22] histological grades were gathered from the histopathology reports (274 cases) or attributed retrospectively (94 cases) by 3 veterinary pathologists (2 board-certified and 1 resident). In the event of non-unanimity, the majority grade was attributed. For the 3-tier grading system, if all pathologists assigned different grades, grade 2 was attributed. In this study, all MCTs were graded equivalently to cutaneous MCTs, even though studies mention applying the 2-tier grading scheme to subcutaneous MCTs with modified cut-off values.[7,24,30] There are no specific recommendations regarding the grading of mucocutaneous MCTs. Further information, such as breed, age, sex, localization of the MCT, involvement of the different skin layers (dermal, subcutaneous, or mucocutaneous), and surface condition (ulcerated or intact) could be collected in most of the cases (Supplemental Table S1). Thick tissue sections (several cell layers in fine focus), tumors with extensive necrosis, specimens in bad conservation state, damaged slide glass, presence of artifacts under the coverslip (air bubbles, tissue folds), and missing information about the *c-KIT-11-ITD* status were exclusion criteria. One hundred ninety-five MCTs with ITD in *c-KIT* exon 11 and 173 *c-KIT* non-mutated MCTs (negative in exons 11 and 8) were selected (Table 1).

PCR-Based Identification of ITD Mutation in the c-KIT Exon 11

The PCR-based identification of *c-KIT-11-ITD* was done at the Michigan State University as part of the MCT panel for routine diagnostics, as previously described.[32] Exon 11 of the c-KIT gene was amplified by PCR using the primer pair at the 5' end of the exon 11 (PE1: 5'-CCATGTATGAAGTACAGTGGAAG-3' sense, bp 1657–1680 of exon 11) and the 5' end of intron 11 (PE2: 5'GTTCCCTAAAGTCATTGTTACACG-3' anti-sense, nucleotides 43–66 of intron 11).[4] A Mutation Detection PCR Kit (Qiagen) was used for the amplification.[3] PCRs were prepared in a 25 µl total reaction volume, with 10 to 25 ng of extracted DNA; 5 pmol of each primer; 0.5 U of Taq polymerase (Invitrogen, Carlsbad, California); and final concentrations of 80 µm deoxynucleoside triphosphate, 2 mM $MgCl_2$, 20 mM Tris-HCl, and 50 µl of KCl.[32] Cycling conditions were as follows: 95°C for 5 minutes; 35 cycles at 95°C for 30 seconds, 58°C for 90 seconds, 72°C for 30 seconds; followed by 68°C for 10 minutes.[3] PCR products were visualized on the QIAxcel Capillary Electrophoresis System (Qiagen).[3]

Staining Variants

To investigate the dependency of the mutational status prediction on the staining protocol, we applied 2 stains (stains A and B) to the same tissue slice, which was de-stained in-between the staining steps, using the following protocol:

HE—stain A—Paraffin sections of all tumors were originally stained at Michigan State University Veterinary Diagnostic Laboratory using the Ventana HE600 system from

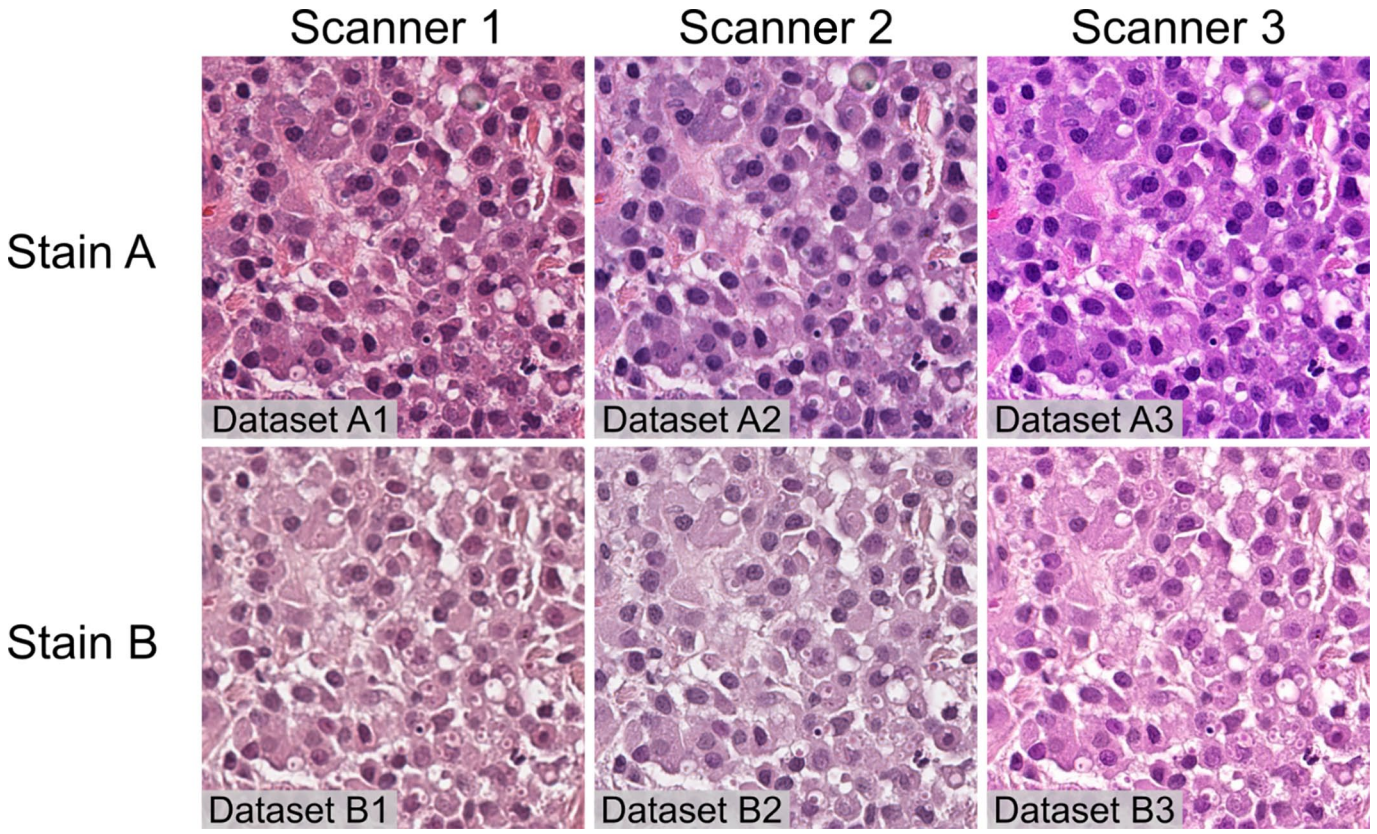

Figure 1. Impact of stain and scanner type on whole-slide image quality. Images of the same tumor sample with 6 stain-scanner variants. Mast cell tumor, skin, canine, and hematoxylin and eosin. Each image represents the same portion of the same tumor in a different stain-scanner variant. Images produced by the different scanners (comparison in the same row) show differences in color calibration and a slight difference in sharpness/depth of field. The staining differs in colors shades/intensity and in the amount of visible detail (comparison in the same column). Scanner 1, Aperio CS2; Scanner 2, Aperio AT2; Scanner 3, 3DHistech Pannoramic II.

Roche. Dry deparaffinization, rehydration, and HE staining were run according to a standard protocol[11] and using Ventana HE 600 Hematoxylin (Roche, 07024282001) and Ventana HE 600 Eosin (Roche, 06544304001).

De—staining—After scanning with all 3 scanners (see below), all slides were manually de-stained by immersion in xylol for at least 2 days before carefully manually removing the coverslip. They were then transferred into a 1% hydrochloric acid alcohol solution until all color was washed away and subsequently rinsed with distilled water.

HE—stain B—De-stained slides were then manually dipped in Mayer's hemalaun (Carl Roth, T865.3) for 8 minutes, rinsed in tap water, dipped in eosin (Diagonal, 2C140.01000) for 30 seconds, and shortly rinsed in tap water before going through an ascending alcohol series (1 minute in 70%, 80%, 96%, and 100% alcohol, respectively).

Scanners and Stain-Scanner Variant Data Sets

All 368 slides were scanned twice (first with stain A and consecutively with stain B) on 3 different scanners, Aperio CS2 (scanner 1), Aperio AT2 scanner (scanner 2), and 3DHistech Pannoramic II scanner (scanner 3). This resulted in 6 distinct stain-scanner variant data sets of whole-slide images (WSIs) (Fig. 1): data set A1 (stain A + scanner 1), data set A2 (stain A + scanner 2), data set A3 (stain A + scanner 3), data set B1 (stain B + scanner 1), data set B2 (stain B + scanner 2), and data set B3 (stain B + scanner 3).

Deep Learning Model

The DLMs were trained to predict the *c-KIT*-11 mutational status of canine MCTs based on their respective WSI. The aim was to discriminate between "*c-KIT*-11 mutated" and "non-mutated" tumors.

In this work, this was done by formulating the problem as a multiple-instance learning task. In multiple-instance learning, the WSI is split into image patches (instances), which together are commonly referred to as a *bag of instances*.[16] The information from all patches is extracted and combined to form a prediction for the entire bag, rather than making predictions for each individual patch.

The clustering-constrained attention multiple-instance learning (CLAM) algorithm by Liu et al[20] was used for multiple-instance learning (Fig. 2). To optimize computational efficiency in the first step, the WSIs were segmented into tissue and background, using the segmentation approach described by Liu et al.[20] Afterwards, the area of the slide with sufficient

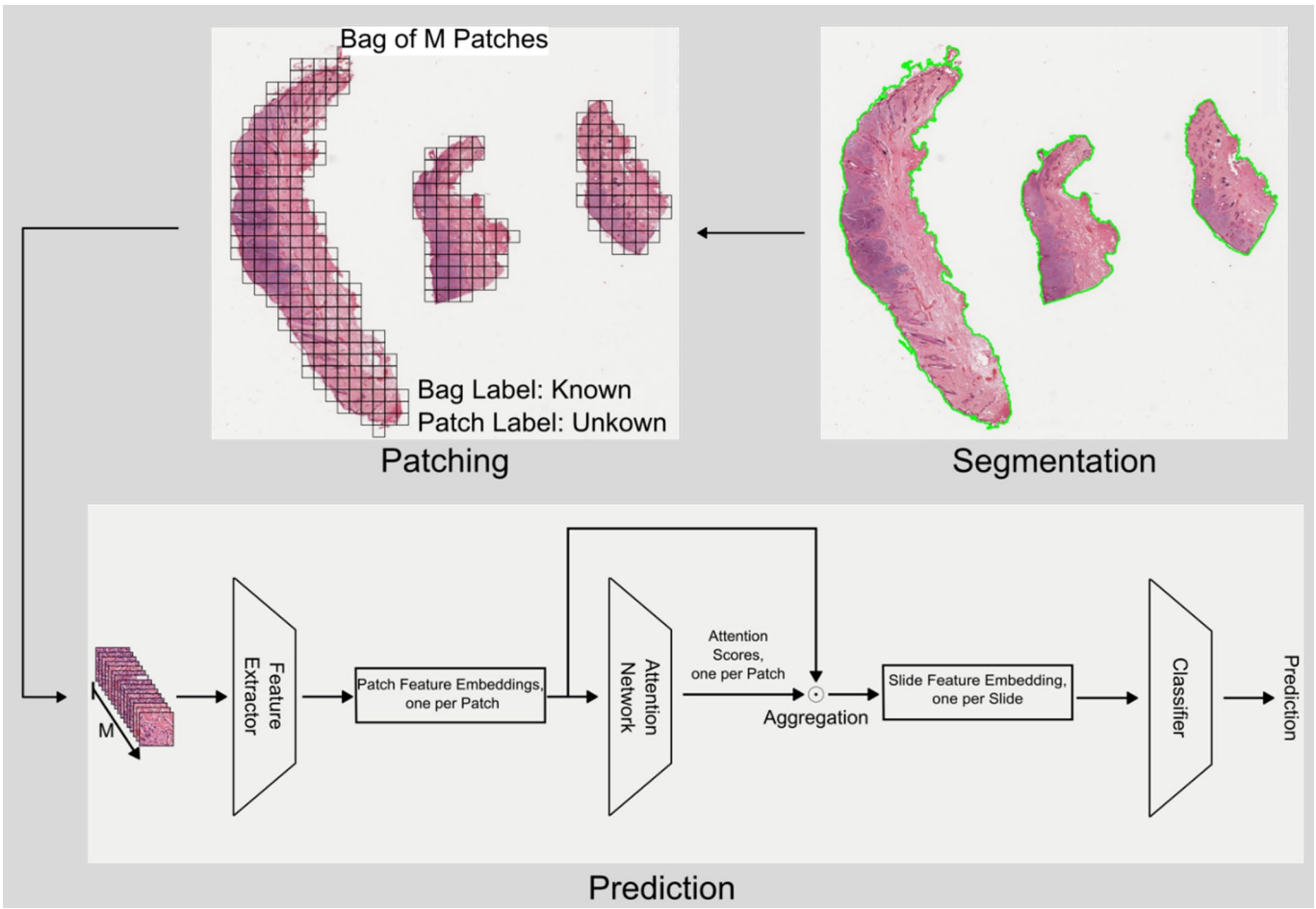

Figure 2. Schematic representation of the clustering-constrained attention multiple-instance learning (CLAM) algorithm that was used for multiple-instance learning. Whole-slide image classification is performed in several steps within the CLAM pipeline used in this study. First, the tissue is segmented from the background and the tissue region is separated into M patches. For each patch, a patch feature embedding is computed, which encodes the visual information contained in the respective patch. To aggregate these patch embeddings into a slide embedding, the deep learning model computes a single attention score for each patch that is directly related to the importance of the respective patch for the model's prediction. By combining the patch embeddings with their respective attention scores, the overall slide feature embedding is computed. This feature embedding is representative for the whole slide and is used for the final prediction of the mutational status.

tissue coverage was divided into a set of M patches. Then, a feature embedding was computed for each patch, encoding the visual information of the respective patch into a 1-dimensional vector. To combine the information from all patch embeddings into a slide embedding, the DLM calculated an attention score for each patch embedding, which directly indicated the patch's importance to the model's prediction. If a patch did not contain meaningful information, it received an attention score close to zero and was discarded from the aggregation of the overall slide embedding. The slide embedding was a 1-dimensional vector that contained characteristic information of the entire slide. Finally, this embedding was used to predict the mutational status (Supplemental Materials).

Model Training Scheme

To increase the statistical informativeness of the results, the DLMs were trained using a 10-fold Monte-Carlo cross-validation. This means that for each of 10 consecutive runs, the data set was randomly split into disjoint training, validation, and test splits. Within each split, 85% of the slides were used as training and validation data, and 15% were used as test data. In each run, a model was trained on the training data. To ensure comparability across experiments, the same 10 trainings, validations, and test splits were used for the experiments across all data set variants.

To assess the DLM performances, the following metrics were used: classification accuracy, mean classification accuracy (MCA), and area under the receiver-operating characteristic curve (AUC). The classification accuracy measures the number of correct predictions divided by the total number of predictions made in 1 test run. MCA is the mean accuracy achieved across 10 different test folds. AUC is used to measure the DLM's ability to correctly distinguish the labels across all possible probability thresholds. Furthermore, the sensitivity and specificity of the models were assessed. All metrics were

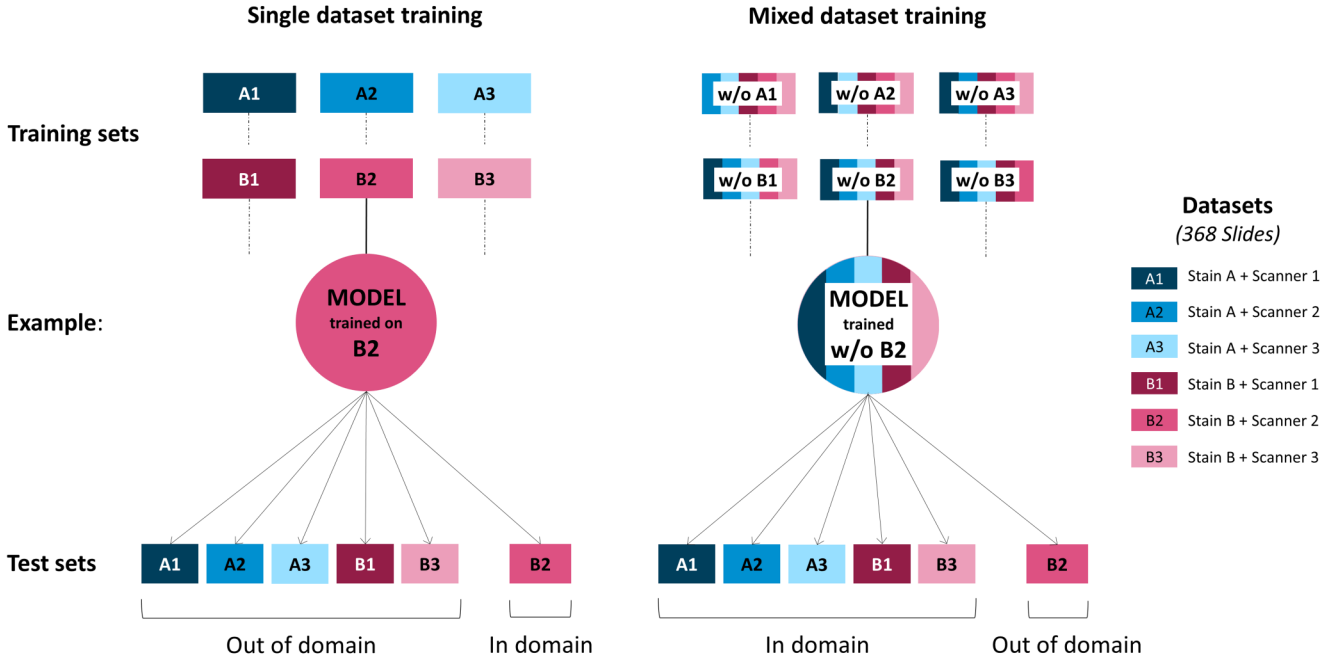

Figure 3. Schematic representation of the model training approaches for the detection of internal tandem duplication mutation in exon 11 of *c-KIT*. Two general training strategies were applied for deep learning model (DLM) training. Training sets composed of single data sets (left side) and training sets composed of equal shares of 5 of the 6 data sets (right side) were used. The resulting trained DLMs were consecutively tested on all 6 data sets to assess their performances in-domain (tested on a data set from the same stain-scanner variant) and out-of-domain (tested on a stain-scanner variant that was not included in the training). w/o, without; Scanner 1, Aperio CS2; Scanner 2, Aperio AT2; Scanner 3, 3DHistech Pannoramic II.

computed using the scikit-learn python package (v. 1.3.0).[23] All DLMs were trained until convergence as observed by the validation loss. Implementation details of the used model architecture and training parameters can be found in the supplementary materials.

Training Approach—Single Data Set

In a first approach, DLMs were trained using the training fraction of the 6 data sets. Each DLM was then tested on the test fraction of all data sets, resulting in 1 in-domain and 5 out-of-domain tests (Fig. 3, left side). Subsequently, this approach was referred to as single data set training. This simulated the following scenario: a pathology laboratory trains a model on its own archive cases, uses it to test its own cases (in-domain), and shares that model with other facilities to test their cases (out-of-domain).

Training Approach—Mixed Data sets ("Leave-One-Out Approach")

To assess whether the training with a more diverse data set leads to a more robust prediction of the *c*-KIT-*11-ITD* status, DLMs were trained with combined training data sets composed of 5 of the 6 original data sets, that is, leaving 1 data set out of the training. The DLMs were then tested on the test fraction of all data sets, resulting in 5 in-domain and 1 out-of-domain training (Fig. 3, right side). In the following sections, we will refer to this approach as mixed data set training. This simulated the following scenario: several pathology laboratories contribute to a collaborative data set to train a model on variably produced archive cases, use it to test their own cases (in-domain), and share that model with further facilities to test their cases (out-of-domain).

Test for Independence of Diagnostic Variables

To assess the association between the model's predictions and clinical-pathologic variables like the 2-tier or 3-tier tumor grade, we used the chi-square test of independence. This is a statistical method for assessing the associations between variables of different scales. To lower the chances of misinterpreting 2 variables as associated, we applied Bonferroni correction since we conducted multiple statistical tests simultaneously. Bonferroni correction was applied separately for each of the clinical-pathologic variables. Given that 7 tests were conducted for each variable, the resulting *P*-values were divided by this number. The assessment only considered models trained on a

Table 2. Mean classification accuracies and area under the curve with respective standard deviations of *c-KIT*-11 internal tandem duplication mutation predictions with deep learning models (DLMs) trained on single data sets.

Tested on Data Set		Trained on Data Set					
		A1	A2	A3	B1	B2	B3
A1	MCA	**0.75**[a] ± 0.08	0.63[b] ± 0.04	0.68 ± 0.08	0.59 ± 0.04	0.61 ± 0.04	0.62 ± 0.07
	AUC	0.83 ± 0.04	0.77 ± 0.07	0.76 ± 0.07	0.71 ± 0.07	0.75 ± 0.06	0.70 ± 0.04
A2	MCA	0.74 ± 0.05	**0.75** ± 0.07	0.68 ± 0.05	0.71 ± 0.05	0.70 ± 0.04	0.68 ± 0.04
	AUC	0.80 ± 0.05	0.84 ± 0.06	0.82 ± 0.04	0.79 ± 0.05	0.81 ± 0.05	0.77 ± 0.4
A3	MCA	0.73 ± 0.06	0.66 ± 0.03	**0.80** ± 0.04	0.72 ± 0.06	0.57 ± 0.04	0.58 ± 0.05
	AUC	0.71 ± 0.07	0.84 ± 0.05	**0.89** ± 0.03	0.82 ± 0.06	0.84 ± 0.04	0.76 ± 0.06
B1	MCA	0.63 ± 0.08	0.61 ± 0.06	0.61 ± 0.06	**0.81** ± 0.05	0.76 ± 0.06	0.61 ± 0.08
	AUC	0.69 ± 0.07	0.75 ± 0.06	0.74 ± 0.07	**0.89** ± 0.05	0.85 ± 0.05	0.80 ± 0.06
B2	MCA	0.63 ± 0.06	0.72 ± 0.07	0.61 ± 0.06	0.74 ± 0.05	**0.78** ± 0.05	0.67 ± 0.09
	AUC	0.70 ± 0.07	0.79 ± 0.06	0.77 ± 0.08	0.84 ± 0.05	**0.87** ± 0.04	0.82 ± 0.05
B3	MCA	0.65 ± 0.07	0.58 ± 0.04	0.54 ± 0.08	0.75 ± 0.05	0.59 ± 0.03	**0.87** ± 0.05
	AUC	0.75 ± 0.06	0.77 ± 0.07	0.59 ± 0.08	0.84 ± 0.04	0.83 ± 0.04	**0.94** ± 0.03

Abbreviations: MCA, mean classification accuracy; AUC, area under the receiving-operating characteristic curve; A, stain A; B, stain B; 1, scanner 1 (Aperio CS2); 2, scanner 2 (Aperio AT2); 3, scanner 3 (3DHistech Pannoramic II).
[a]Bold font = MCA of DLMs trained and tested on the same data set (in-domain).
[b]Regular font = MCA of DLMs tested on a different training data set from the one they were trained on (out-of-domain).

single data set to establish a direct link between training data and DLM results. The pingouin python package (v. 0.5.3),[31] a statistical analysis package, was used to compute the chi-square test.

Results

Data Set Composition

In total, 50 different dog breeds were represented in the data set. The 10 most common breeds, accounting for 198 cases (53.8%), were: mixed breed (20.1%), Labrador retriever (12.8%), golden retriever (4.6%), boxer (3.3%), American pit bull terrier (3.0%), Maltese dog (2.4%), Pug (2.4%), French bulldog (1.9%), Boston terrier (1.9%), and Shih Tzu (1.4%) (Supplemental Table S2). Breed information was unavailable for 96 (26.1%) cases.

Of the selected 195 *c-KIT-11-ITD* positive MCTs, 121 (62.1%) were high grade and 74 (37.9%) were low grade (Table 1). One hundred sixty-four MCTs (84.1%) were cutaneous, 18 (9.2%) were subcutaneous, 6 (3.1%) were mucocutaneous, and 7 (3.6%) could not be reliably located due to lacking histologic reports and absence of orientation criteria on the slide. The 173 *c-KIT-11-ITD* negative MCTs were chosen in the aim of mirroring the repartition of the *c-KIT-11-ITD* positive group, with 97 (56.1%) being classified as high grade and 76 (43.9%) as low grade. Moreover, 145 (83.8%) were cutaneous, 26 (15.0%) subcutaneous, and 2 (1.2%) were mucocutaneous (Supplemental Table S1).

Prediction of the c-KIT 11 ITD Status Using Single Data Set Training

To assess the capacity of a DLM trained in a single stain-scanner environment to predict the *c-KIT-11-ITD* status in HE-stained WSIs, 6 DLMs were trained on a single training set (stain-scanner variant) and tested on a test set of the same stain-scanner variant (in-domain) and on different stain-scanner variants (out-of-domain) (Table 2). Across all DLMs, an in-domain average MCA of 0.79 (range 0.75–0.87) was found. The highest MCA of 0.87 was reached by the DLM trained on the data set 6 (stain B, scanner 3) with a sensitivity of 0.90 and a specificity of 0.83. In other words, this DLM predicted the correct *c-KIT-11-ITD* status in 87% of the cases with 90% probability of true positive prediction and 83% of true negative prediction. In general, DLMs trained and tested on stain B WSIs allowed for a higher MCAs (average MCA 0.82) than DLMs trained on stain A WSIs (average MCA 0.76), pointing toward a general influence of HE-staining protocol for accurate *c-KIT-11-ITD* prediction. Furthermore, scanner 3 seemed to be slightly more suitable (average MCA 0.84) than scanner 1 (average MCA 0.78) and scanner 2 (average MCA 0.77). However, when the models were tested out-of-domain, they only reached an average MCA of 0.65 (range 0.54–0.76). Detailed performance metrics of all single data set trained DLMs can be found in Supplemental Table S4.

Prediction of the c-KIT-11-ITD Status Using Mixed Data Set Training (Leave-One-Out Approach)

To test whether more diverse data sets from multiple institutions may lead to more robust *c-KIT-11-ITD* predictions in HE-stained WSIs from unknown stain-scanner variants, a second set of 6 DLMs was trained on a data set that contained equal fractions (20%) of 5 of the 6 data sets and tested on the sixth, unknown, data set (leave-one-out-approach, Table 3). The average in-domain MCA was 0.76 (range 0.72–0.85), which is lower than the in-domain MCA of the DLMs trained and tested on a single data set (Table 2). However, with an

Table 3. Mean classification accuracies and area under the curve with respective standard deviations of *c-KIT*-11 internal tandem duplication mutation predictions with deep learning models (DLMs) trained on mixed data sets (leave-one-out-approach).

		Trained on Data Set					
Tested on Data Set		w/o A1	w/o A2	w/o A3	w/o B1	w/o B2	w/o B3
A1	MCA	**0.69**[a] ± 0.04	0.74[b] ± 0.05	0.73 ± 0.06	0.73 ± 0.07	0.72 ± 0.06	0.73 ± 0.07
	AUC	**0.80** ± 0.06	0.82 ± 0.05	0.81 ± 0.05	0.82 ± 0.05	0.82 ± 0.05	0.82 ± 0.07
A2	MCA	0.76 ± 0.06	**0.75** ± 0.06	0.76 ± 0.06	0.76 ± 0.07	0.75 ± 0.07	0.75 ± 0.08
	AUC	0.84 ± 0.05	**0.84** ± 0.05	0.83 ± 0.06	0.84 ± 0.06	0.84 ± 0.05	0.84 ± 0.06
A3	MCA	0.77 ± 0.06	0.76 ± 0.05	**0.71** ± 0.05	0.77 ± 0.06	0.76 ± 0.04	0.79 ± 0.06
	AUC	0.84 ± 0.05	0.86 ± 0.06	**0.83** ± 0.05	0.86 ± 0.05	0.86 ± 0.05	0.86 ± 0.05
B1	MCA	0.80 ± 0.05	0.77 ± 0.05	0.75 ± 0.05	**0.72** ± 0.05	0.75 ± 0.06	0.76 ± 0.06
	AUC	0.87 ± 0.05	0.85 ± 0.05	0.84 ± 0.05	**0.80** ± 0.06	0.85 ± 0.05	0.84 ± 0.06
B2	MCA	0.79 ± 0.06	0.77 ± 0.05	0.77 ± 0.05	0.77 ± 0.06	**0.77** ± 0.05	0.77 ± 0.06
	AUC	0.86 ± 0.05	0.85 ± 0.05	0.84 ± 0.05	0.83 ± 0.06	**0.83** ± 0.06	0.84 ± 0.06
B3	MCA	0.85 ± 0.04	0.82 ± 0.04	0.83 ± 0.04	0.82 ± 0.05	0.83 ± 0.04	**0.73** ± 0.05
	AUC	0.93 ± 0.03	0.91 ± 0.03	0.92 ± 0.04	0.91 ± 0.04	0.91 ± 0.04	**0.80** ± 0.05

Abbreviations: MCA, mean classification accuracy; AUC, area under the receiving-operating characteristic curve; w/o, without; A, stain A; B, stain B; 1, scanner 1 (Aperio CS2); 2, scanner 2 (Aperio AT2); 3, scanner 3 (3DHistech Pannoramic II).
[a]Bold font = MCA of DLMs tested on a data set different from the one they were trained on (out-of-domain).
[b]Regular font = MCA of DLMs trained and tested on the same data set (in-domain).

Table 4. Associations between the deep learning model (DLM) prediction and known clinical-pathologic parameters of the tumor.

	PCR	DLM Trained on Data Set					
Parameter		A1	A2	A3	B1	B2	B3
Location	0.47	1	1	1	0.82	1	1
Ulcerated surface	5.87E–09	9.72E–08	1.03E–08	9.72E–08	9.79E–04	1.12E–04	1.91E–07
Kiupel grade	0.02	0.01	0.07	0.13	0.09	0.10	0.27
Patnaik grade	0.74	0.01	9.97E–04	4.94E–04	1	1	0.58
Gender	1	1	0.11	0.46	1	1	1

Abbreviations: A, stain A; B, stain B; 1, scanner 1 (Aperio CS2); 2, scanner 2 (Aperio AT2); 3, scanner 3 (3DHistech Pannoramic II); PCR, polymerase chain reaction.
P-values of the chi-square test of independence with Bonferroni correction. *P*-values < 0.05 are considered significant.

average MCA of 0.73 (range 0.69–0.77), the models trained on mixed data sets showed a better out-of-domain performance than the models trained on a single data set, which achieved an average out-of-domain MCA of 0.65 (range 0.54–0.76) (Table 2). Detailed performance metrics of all mixed data set trained DLMs can be found in Supplemental Table S5.

Analysis of c-KIT Mutation Prediction, c-KIT-11-ITD Status, and Other Clinical-Pathologic Parameters

Chi-square analysis was used to test for independence between the DLM-based mutation prediction and the following parameters: tumor grade (2-tier and 3-tier grading system), location of the tumor (subcutaneous, cutaneous, and mucocutaneous), skin ulceration, and sex (Table 4 and Supplemental Table S5). Analysis confirmed that the DLM predictions were not significantly independent with either the skin layer involvement or the sex of the affected dog. A significant association was found between the 3-tier grade and the predictive output of all models trained on stain A (DLMs trained on data sets A1–3) as well as an association of *c-KIT-11-ITD* mutation prediction by the DLM trained on data sets A1 and B1, and the 2-tier grade. Furthermore, chi-square analysis indicated that the ulceration status of the overlying epidermis of the skin is significantly associated with the mutation prediction by all DLMs.

Discussion

This study was conducted to assess the efficacy of artificial intelligence-based DLMs in predicting *c-KIT-11-ITD* of canine MCTs solely based on HE-stained WSIs. A total of 368 MCT cases were gathered and digitized, creating 6 data sets corresponding to various stain-scanner combinations. The most successful DLM achieved an MCA of 0.87, with a sensitivity of 0.90, specificity of 0.83, and an AUC of 0.94, underscoring the effectiveness of DLMs in this predictive task.

Given the variability in fixation and staining protocols across pathology laboratories, as well as differences in WSIs from various image acquisition devices, domain shifts (ie, variations between training and test situations) are bound to affect

a DLM's performance.[1] The study addressed this by consecutively staining slides with 2 HE-staining protocols (A and B) from different laboratories and digitalizing them with 3 slide scanners (1–3). DLMs were trained on individual scanner-stain combinations (single data set) or mixtures thereof (mixed data set), simulating institution-specific versus collaborative training scenarios. Cross-validation results for DLMs trained on a single data set and tested in-domain (ie, on WSIs produced in the same training institution) showed an average MCA of 0.79. However, testing on out-of-domain WSIs (ie, from different institutions) led to an important MCA drop (0.65), underlining the domain-dependency of DLMs. A transfer of a locally trained DLM to another institution is thus not possible without tailored mitigation of the domain gap.

Mixed data set-trained DLMs exhibited comparable in-domain performance (average MCA of 0.76) and slightly improved out-of-domain performance (average MCA of 0.73) compared with single data set-trained DLMs, suggesting enhanced transferability through collaborative training. However, training with multiple data sets also increases the demands on the hardware used, especially in terms of memory space, as the calculated features of each data set must be stored in addition to the actual WSI. This high domain-dependency of DLMs has been described for HE-stained WSIs before,[2] and domain generalization is currently a major research focus in computer science to reduce data set variability required for robust real-life artificial intelligence applications.

Although scanner 1 (Aperio CS2) and scanner 2 (Aperio AT2) use the same optical system and produce images with the same resolution, they differ in their loading capacity and scanning speed. The observed performance differences between DLMs trained and tested on these respective devices might be due to processing-induced image variations.

The utilization of a weakly supervised training approach, relying solely on mutational status information without explicit annotations or tumor segmentations, posed a potential bias risk (eg, random overrepresentation of necrosis in one of the groups) but has proven to be an efficient method to train the DLMs. Histologic review of false-positive and false-negative cases did not reveal any evident sources of error. The morphologic causes of wrong predictions are thus unclear, as are the morphologic structures that allow correct predictions by DLMs. Moreover, the varying associations found between DLM-based predictions and clinical-pathologic variables through chi-square analyses emphasize the significance of considering stain-scanner combinations as integral factors influencing predictive modeling outcomes. Hence, a model trained on 1 stain-scanner combination may not generalize to other combinations without specific adjustments. For instance, tumors with ulcerated surfaces, as well as high-grade tumors, were more likely to be classified as *c*-KIT-*11-ITD* positive, as indicated by a *P* value largely inferior to 0.05 across all DLMs. Of the 119 cases exhibiting an ulcerated epidermis, 86 (72%) are *c*-KIT-*11-ITD* positive against 33 (28%) *c*-KIT-*11-ITD* negative. This association might just be due to the percentage of high-grade MCTs carrying a *c*-KIT-*11-ITD* being fundamentally more elevated, and high-grade MCTs being more prone to ulcerate.[4,9,14,28,34] Interestingly, our study identified stain B as allowing better prediction of the *c*-KIT-*11-ITD* status across all scanners. This suggests that stain B might highlight a morphological feature linked with the *c*-KIT-*11-ITD* mutation that might not be visible in stain A and highlights the importance of stain selection in enhancing model performance. Minimal morphological changes resulting from the de-staining and re-staining processes might also be a potential source of performance variations.

This study was inspired by similar work done in the human medicine field with comparably sized data sets (between 62 and 433 samples).[8,15,19,21] It is worth mentioning that these human data sets often exhibit imbalances, with notably more non-mutated samples. The AUC metric had to be used to assess the performance of DLMs trained on unbalanced data sets, whereas MCA can be used for balanced data sets. These studies reported AUCs ranging from 0.64 to 0.86. Noteworthy comparisons include a study by Elche et al, which worked with a very large data set of 8343 cases compiled from 9 distinct data sets with the aim of predicting microsatellite instabilities, a highly relevant prognostic factor in colorectal cancers, from HE slides. AUCs ranging from 0.74 to 0.96 were reported.[10] The data set generated for the present study represents, to this date, the largest digital *c-KIT-11-ITD* mutated canine MCT collection in veterinary medicine and achieved comparable or even superior performances, with AUCs ranging from 0.80 to 0.94.

These promising outcomes led to the planning of several follow-up projects to refine the model's performance, ensure robustness in diverse real-life scenarios, and explore practical applications of DLMs in routine diagnostics (eg, low-cost and fast screening tool).

Authors' Note

CP and JG contributed equally to the study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Jonathan Ganz https://orcid.org/0009-0008-1299-8716
Christof A. Bertram https://orcid.org/0000-0002-2402-9997
Marc Aubreville https://orcid.org/0000-0002-5294-5247
Robert Klopfleisch https://orcid.org/0000-0002-6308-0568

References

1. Aubreville M, Bertram C, Veta M, et al. Quantifying the scanner-induced domain gap in mitosis detection. arXiv. 2021.
2. Aubreville M, Stathonikos N, Bertram CA, et al. Mitosis domain generalization in histopathology images—the MIDOG challenge. *Med Image Anal*. 2023;**84**:102699.

3. Brocks BAW, Bertram CA, Bartel A, et al. Internal tandem duplication of exon 8 of c-KIT is associated with longer total survival in canine cutaneous mast cell tumors. *Vet Pathol*. 2021;**58**(2):315–324.
4. Cameron LRJ, Grahn RA, Chien MB, et al. Detection of c-KIT mutations in canine mast cell tumors using fluorescent polyacrylamide gel electrophoresis. *J Vet Diagn Invest*. 2004;**16**(2):95–100.
5. Cao W, Hashibe M, Rao J-Y, et al. Comparison of methods for DNA extraction from paraffin-embedded tissues and buccal cells. *Cancer Detect Prev*. 2003;**27**(5):397–404.
6. 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 Precis Oncol*. 2020;**4**(1):1–7.
7. Chen P, Marconato L, Sabattini S, et al. Mutations in exons 8 and 11 of c-KIT gene in canine subcutaneous mast cell tumors and their association with cell proliferation. *Vet Sci*. 2022;**9**:493.
8. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;**24**(10):1559–1567.
9. Downing S, Chien MB, Kass PH, et al. Prevalence and importance of internal tandem duplications in exons 11 and 12 of c-KIT in mast cell tumors of dogs. *Am J Vet Res*. 2002;**63**(12):1718–1723.
10. Echle A, Ghaffari Laleh N, Quirke P, et al. Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a pre-screening tool for clinical application. *ESMO Open*. 2022;**7**(2):100400.
11. Roche Diagnostics. VENTANA HE 600 System User Guide, Software Version 1.9.5, Version 2.0. Roche Inc. January 1, 2021. Accessed October 10, 2024. https://elabdoc-prod.roche.com/eLD/api/downloads/9bcdf633-637b-ea11-fc90-005056a71a5d?countryIsoCode=XG
12. European Medicines Agency. Masivet: European Public Assessment Report (EPAR) for the public. European Medicines Agency, 2009. https://www.ema.europa.eu/en/documents/overview/masivet-epar-summary-public_en.pdf
13. Freytag JO, Queiroz MR, Govoni VM, et al. Prognostic value of immunohistochemical markers in canine cutaneous mast cell tumours: a systematic review and meta-analysis. *Vet Comp Oncol*. 2021;**19**(3):529–540.
14. Gil da Costa RM. C-KIT as a prognostic and therapeutic marker in canine cutaneous mast cell tumours: from laboratory to clinic. *Vet J*. 2015;**205**(1):5–10.
15. Hildebrand LA, Pierce CJ, Dennis M, et al. Artificial intelligence for histology-based detection of microsatellite instability and prediction of response to immunotherapy in colorectal cancer. *Cancers*. 2021;**13**(3):391.
16. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. *arXiv*. 2018.
17. Janecka A, Adamczyk A, Gasińska A. Comparison of eight commercially available kits for DNA extraction from formalin-fixed paraffin-embedded tissues. *Anal Biochem*. 2015;**476**:8–10.
18. Kiupel M, Webster JD, Bailey KL, et al. Proposal of a 2-tier histologic grading system for canine cutaneous mast cell tumors to more accurately predict biological behavior. *Vet Pathol*. 2011;**48**(1):147–155.
19. Liechty B, Xu Z, Zhang Z, et al. Machine learning can aid in prediction of IDH mutation from H&E-stained histology slides in infiltrating gliomas. *Sci Rep*. 2022;**12**(1):22623.
20. Liu Z, Mao H, Wu C-Y, et al. A ConvNet for the 2020s. arXiv. 2022.
21. Montalto MC, Edwards R. And they said it couldn't be done: predicting known driver mutations from H&E slides. *J Pathol Inform*. 2019;**10**:17.
22. Patnaik AK, Ehler WJ, MacEwen EG. Canine cutaneous mast cell tumor: morphologic grading and survival time in 83 dogs. *Vet Pathol*. 1984;**21**(5):469–474.
23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *arXiv*. 2018.
24. Sabattini S, Brocanelli A, Zaccone R, et al. The 2-tier grading system identifies canine cutaneous and/or subcutaneous mast cell tumors with aggressive biological behavior regardless of growth model [published online ahead of print April 5, 2024]. *Vet Pathol*. 2024;0(0). doi:10.1177/03009858241240443
25. Sarnecka AK, Nawrat D, Piwowar M, et al. DNA extraction from FFPE tissue samples—a comparison of three procedures. *Contemp Oncol*. 2019;**23**(1):52–58.
26. Sledge DG, Webster J, Kiupel M. Canine cutaneous mast cell tumors: a combined clinical and pathologic approach to diagnosis, prognosis, and treatment selection. *Vet J*. 2016;**215**:43–54.
27. Takeuchi Y, Fujino Y, Watanabe M, et al. Validation of the prognostic value of histopathological grading or c-KIT mutation in canine cutaneous mast cell tumours: a retrospective cohort study. *Vet J*. 2013;**196**(3):492–498.
28. Tamlin V, Kessell A, Mccoy R, et al. Prevalence of exon 11 internal tandem duplications in the C-KIT proto-oncogene in Australian canine mast cell tumours. *Aust Vet J*. 2017;**95**(10):386–391.
29. Thamm DH, Avery AC, Berlato D, et al. Prognostic and predictive significance of KIT protein expression and c-KIT gene mutation in canine cutaneous mast cell tumours: a consensus of the Oncology-Pathology Working Group. *Vet Comp Oncol*. 2019;**17**(4):451–455.
30. Thompson JJ, Yager JA, Best SJ, et al. Canine subcutaneous mast cell tumors: cellular proliferation and KIT expression as prognostic indices. *Vet Pathol*. 2011;**48**(1):169–181.
31. Vallat R. Pingouin: statistics in Python. *J Open Source Softw*. 2018;**3**(31):1026.
32. Webster JD, Yuzbasiyan-Gurkan V, Kaneene JB, et al. The role of c-KIT in tumorigenesis: evaluation in canine cutaneous mast cell tumors. *Neoplasia N Y N*. 2006;**8**(2):104–111.
33. Webster JD, Yuzbasiyan-Gurkan V, Thamm DH, et al. Evaluation of prognostic markers for canine mast cell tumors treated with vinblastine and prednisone. *BMC Vet Res*. 2008;**4**:32.
34. Weishaar KM, Ehrhart EJ, Avery AC, et al. C-KIT mutation and localization status as response predictors in mast cell tumors in dogs treated with prednisone and toceranib or vinblastine. *J Vet Intern Med*. 2018;**32**(1):394–405.
35. Zemke D, Yamini B, Yuzbasiyan-Gurkan V. Mutations in the juxtamembrane domain of c-KIT are associated with higher grade mast cell tumors in dogs. *Vet Pathol*. 2002;**39**(5):529–535.