Deep domain adversarial learning for species-agnostic classification of histologic subtypes of Osteosarcoma

Sushant Patkar, Jessica Beck, Stephanie Harmon, Christina Mazcko, Baris Turkbey, Peter Choyke, G Tom Brown, Amy LeBlanc

PII: S0002-9440(22)00319-4
DOI: https://doi.org/10.1016/j.ajpath.2022.09.009
Reference: AJPA 3800

To appear in: The American Journal of Pathology

Received Date: 27 June 2022
Revised Date: 31 August 2022
Accepted Date: 28 September 2022


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Copyright © 2022 American Society for Investigative Pathology. Published by Elsevier Inc. All rights reserved
Deep domain adversarial learning for species-agnostic classification of histologic subtypes of Osteosarcoma

Sushant Patkar1,2, Jessica Beck2, Stephanie Harmon1, Christina Mazcko2, Baris Turkbey1, Peter Choyke1, G Tom Brown1, Amy LeBlanc2,3

1. Artificial Intelligence Resource, Molecular Imaging Branch, National Cancer Institute, NIH
2. Comparative Oncology Program, Molecular Imaging Branch, National Cancer Institute, NIH

# Address correspondence to:
Sushant Patkar, Ph.D., National Cancer Institute, 5413 West Cedar Lane, Suite 102-C, Bethesda, MD 20814, 240-585-3000, patkar.sushant@nih.gov

Amy Leblanc, D.V.M., National Cancer Institute, Building 10, Room 1B53, Bethesda, MD 20892, 240-760-7093, amy.leblanc@nih.gov

Competing Interests
No competing interests to be declared

Funding
This work was supported by the Intramural Program of the National Cancer Institute, NIH (ZIH BC 012032)
Abstract
Osteosarcomas are aggressive bone tumors with many divergent histologic patterns. During pathology review, osteosarcomas are subtyped based on the predominant histologic pattern; however, tumors often demonstrate multiple patterns. This high tumor heterogeneity coupled with scarcity of samples compared to other tumor types render histology-based prognosis of osteosarcomas challenging. To combat lower-case numbers in humans, dogs with spontaneous osteosarcomas have been suggested as a model species. Here, we adversarially train a convolutional neural network to classify distinct histological patterns of osteosarcoma in humans using mostly canine Osteosarcoma data during training. We show that adversarial training improves domain adaption of a histologic subtype classifier from canines to humans achieving an average multi-class F1 score of 0.77 (CI: 0.74-0.79) and 0.80 (CI: 0.78-0.81) when compared to the ground truth in canines and humans, respectively. Finally, we applied our trained model to characterize the histologic landscape of 306 canine osteosarcomas and uncovered distinct clusters with markedly different clinical responses to standard of care therapy.
Introduction

Osteosarcoma (OS) is a rare but aggressive pediatric malignancy with approximately 800 cases reported annually in the U.S.\(^1\). Patients with metastatic or relapsed disease have dismal outcomes, with survival rates less than 30% despite aggressive salvage regimens that typically include additional surgery, radiotherapy, and chemotherapy with agents such as ifosfamide, etoposide, cyclophosphamide, gemcitabine and topotecan\(^2\). Most osteosarcomas display osteoblastic differentiation, sometimes intermixed with one or more additional histologic patterns including chondroblastic, fibroblastic, giant cell-rich and vessel-rich \(^3\)-\(^5\). Currently the only reliable histologic marker for prognosis in human OS is the amount of necrosis achieved after neoadjuvant chemotherapy\(^6\). This assessment is based upon review of tumor sections harvested after local tumor control via surgery. Despite this, there is a subset of patients with high necrosis that still develop metastatic disease after completion of frontline therapy. Hence, additional prognostic biomarkers are needed for accurate prognosis prediction. Since naturally occurring canine osteosarcoma has strong biological, molecular and histologic similarities to human osteosarcoma and is at least 10 times more common than human osteosarcoma, it can serve as a powerful translational model for cancer biomarker investigation and drug development\(^7\)-\(^9\).

In dogs with OS, standard of care consists of amputation of the affected limb to achieve local tumor control, followed by systemic platinum and/or anthracycline-based chemotherapy\(^10\). However, many clinical studies demonstrate that development of metastases, most often to the lungs, occur in over 90% of canine patients within several months of diagnosis\(^11\)-\(^15\). In contrast to humans, the clinical workflow in dogs does not allow for assessment of response to neoadjuvant therapy, but rather access to the entire tumor at the time of diagnosis via limb amputation. This allows a greater area of untreated tumor for analysis and correlation with outcomes of that specific patient.

Furthermore, in canine OS, beyond tumor stage, i.e. de novo metastatic disease, there are no known consistent prognostic features either within the primary tumor histology or other patient factors such as tumor location, ALP status, age/sex/breed\(^12\), \(^14\), \(^15\). Studies examining the prognostic significance of histologic subtype have identified conflicting findings in different datasets\(^11\), \(^13\). In this study, we took advantage of a larger patient cohort accumulated during a prospective randomized clinical trial conducted in over 300 canine patients\(^12\). This yielded a well-annotated canine OS dataset in which to examine osteosarcoma histology and explore the potential of Artificial Intelligence (AI)-derived biomarkers. Specifically, we investigate whether techniques in AI using adversarial learning could support the development of i) a histologic subtype classifier for osteosarcomas that adapts from dogs to humans and ii) a prognostic signature in dogs based on digital pathology Whole Slide Images (WSI)\(^16\)-\(^18\).

Materials and Methods

Curation of H&E-stained slides of Dog and Human Osteosarcomas
Canine OS tumor samples were curated from a multisite clinical trial\(^12\). Tumors were biopsied pre-amputation and diagnosed as osteosarcoma by anatomic pathologists at Comparative
At the time of surgical limb amputation, additional tumor tissue was collected by COTC investigators as a part of the standard of care portion of the trial schema. All tumors were collected prior to any treatment. Dogs were randomized to receive either standard of care (SOC) or standard of care + adjuvant sirolimus (Rapamycin) therapy. Statistical analysis of the primary clinical outcomes of the entire cohort of dogs found no differences in DFI or survival between the two arms, thus cases were included together in the analysis presented herein. In addition, we obtained 39 human osteosarcoma samples from an in-house pathology residency training cohort. Of these 39, only 11 were utilized in our study for validation of domain-agnostic features. Tumor tissue was placed in 10% neutral buffered formalin for 24 hours and then subjected to EDTA slow decalcification. Tissue was then sectioned and stained with Hematoxylin and Eosin according to standard histopathologic practice. Three canine cases were excluded from this study as slides from these cases were not available. Slides from remaining 306 canine cases and 39 human cases were digitized using Hamamatsu S60 digital scanner (Hamamatsu Photonics, Hamamatsu, Japan) in 40X magnification or 0.23um per pixel. No additional manual quality control of surgical tumor specimen size or percent tumor tissue was completed prior to data collection. The methods were performed in accordance with relevant guidelines and regulations and approved by each participating COTC veterinary institution that enrolled canine patients onto the clinical trials from which the image data was derived.

Annotation and Pre-processing of whole slide image data
Pathologist annotations for 95 dog slides and 11 human slides were obtained in xml format using HALO (Albuquerque, New Mexico USA). Each annotation file contained coordinates of roughly marked region boundaries for each histologic subtype within each slide. Since osteoblastic subtype is the most dominant subtype in osteosarcoma, the main tumor areas were marked and annotated as osteoblastic. Any regions within this area exhibiting divergent histology were annotated as either Necrotic (N): Vessel-rich (VR), Chondroblastic (CB), Fibroblastic (FB) or Giant Cell-Rich (GC)\(^3,5\). In canine tumors annotated as VR, CD31 immunohistochemistry was used to confirm the presence of tumor cell (CD31\(^-\)) lined vascular spaces (Supplementary Figure S1)\(^19,20\). Any unmarked regions falling outside main tumor areas were classified as “other” and consisted primarily of non-tumor tissue, osteoid formations and in some cases slide preparation artefacts such as folded tissue, slide debris, etc.

Training deep learning models on whole slide image tiles extracted from multiple magnifications has proven to be effective in a weakly supervised learning setting where region level annotations by pathologists are not available and histological features of interest are open-ended\(^21-24\). However, in this study, we had region level pathologist annotations which were based on previously defined histological subtypes of Osteosarcoma that are distinguishable at 10x magnification level\(^25\). The smallest regions of interest annotated by the pathologist have an area of \(\sim 25000 \, \mu m^2\) and are represented by at least 1 tile of size 256x256 at 10x magnification. A larger tile size would have resulted in fewer training tiles per histological subtype, which would further increase class imbalance and cause overfitting, whereas a smaller tile size would have obscured...
important architectural features that go beyond cellular morphology. E.g tumor cells surrounding blood vessels, which are a characteristic feature of telangiectatic osteosarcoma. Hence, to train our image classification model, each whole slide image was scanned at 10x magnification level and broken down into 256x256 pixel tiles.

Tiles containing more than 85% of white space were filtered out. Each remaining tile was assigned a single label based on any overlapping pathologist annotations. If a tile contained one or more tumor lesions of divergent histology (i.e a region exceeding 15% of the tile area), the tile was assigned the histologic class of the most dominant lesion (i.e the divergent lesion covering the highest %area). Otherwise, the tile was assigned label “Osteoblastic”. For e.g., if a tile had 35% of its area marked as Fibroblastic, then the tile gets assigned the label “FB”. If a tile is dominated by non-tumor tissue or hemorrhage, it was assigned the label: “other”. All other tiles from unmarked slides were regarded as unlabeled.

For training, we randomly selected 80% of all labeled tiles from dogs (source domain) and additionally 2000 randomly selected labeled tiles from humans (target domain). Of the remaining 20% labeled tiles from dogs, half were randomly selected for validation and hyper-parameter tuning and the remaining half were held out for testing along with the remaining labeled human tiles that were not selected for training. For reproducibility, we fixed the random seed in our codes generating the train, validation, and test splits. The distribution of tiles by histologic subtype and train, validation and test split are shown in Figure 1 and Supplementary Table 1.

Prior to feeding a tile as input to the classification model, each tile was rescaled to 224x224 pixels and its per channel pixel intensities (ranging from 0-1) were normalized to follow a standard normal distribution using the following per channel mean intensity and standard deviations estimated from the dog training data: mean: (R=0.8938, G=0.5708, B=0.7944), standard deviation: (R= 0.1163, G=0.1528, B=0.0885). Furthermore, to artificially augment the size of the training set, each tile from a minibatch during training was flipped on one side at random.

Domain adversarial training of a histologic subtype classification model for Osteosarcomas

Let \((X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)\) be examples from a source domain \((d = \text{dogs})\) and \((X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)\) be examples from a target domain \((d = \text{humans})\) where the number of examples available are typically much less than the number of examples available from the source domain. To train a classification model that adapts from the source domain to target domain, we extend the algorithm of \(^{26}\) to the supervised setting. Specifically, let \(\theta_f\) be the parameters of the feature extraction backbone \(G_f(.; \theta_f)\), (i.e the function that takes as input an example \(X_i\) and maps it to a set of features), let \(\theta_y\) be the parameters of the subtype classifier \(G_y(.; \theta_y)\), (i.e the function that receives input from the feature extractor and predicts class label \(Y_i\)), and let \(\theta_d\) be the parameters of the domain classifier \(G_d(.; \theta_d)\) (i.e the function that receives input from the feature extractor and predicts the domain label \(d_i\)). Furthermore, let:

\[
E(\theta_f, \theta_y, \theta_d) = \sum_{i=1}^{m+n} L(G_y(G_f(X_i; \theta_f); \theta_y), Y_i) - \lambda \sum_{i=1}^{m+n} L(G_d(G_f(X_i; \theta_f); \theta_d), d_i)
\]
\[ \sum_{i=1}^{m+n} L_y^i(\theta_f, \theta_y) - \lambda \sum_{i=1}^{m+n} L_d^i(\theta_f, \theta_d) \]  

(1)

The first term in equation 1 represents the subtype classification error whereas the second term in equation (1) represents the domain classification error and the hyperparameter \( \lambda \) controls the trade-off between the two errors. The goal of a domain adaptation algorithm is then to find the saddle point of \( E \):

\[
(\hat{\theta}_f, \hat{\theta}_y) = \arg \min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d)
\]

(2)

\[
\hat{\theta}_d = \arg \max_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d)
\]

(3)

The domain classifier tries to minimize the domain classification error (because of the \(-\lambda\) term) and the subtype classifier tries to minimize the subtype classification error. To find the saddle point, the domain classifier is trained adversarially with the label classifier. Consequently, the parameters of the feature extractor \( \theta_f \) at the saddle point minimize the subtype classification error (i.e the learned features are discriminative) while maximizing the domain classification error (i.e the learned features are domain-invariant). Adversarial training is implemented in practice by simply adding a gradient reversal layer just before the domain classifier and performing standard stochastic gradient descent (See Figure 1). The update rule for the parameters after incorporating the gradient reversal layer are given by equations (4),(5) and (6):

\[
\theta_f \leftarrow \theta_f - \mu \left( \frac{\partial L_y^i}{\partial \theta_f} - \lambda \frac{\partial L_d^i}{\partial \theta_f} \right)
\]

(4)

\[
\theta_y \leftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y}
\]

(5)

\[
\theta_d \leftarrow \theta_d - \mu \frac{\partial L_d^i}{\partial \theta_f}
\]

(6)

The hyperparameter \( \mu \) represents the learning rate. To obtain a head-start during training, we initialize the parameters of the feature extraction portion of the resnet50 CNN \( \theta_f \) to the values obtained from pre-training resnet50 on the ImageNet dataset\(^{27}\). Initializing convolutional neural networks with pre-trained weights from ImageNet has previously demonstrated success in transfer learning on many digital pathology applications\(^{28,29}\). With the help of stochastic gradient descent, we then simultaneously train the histologic subtype classifier and domain classifier over several epochs using the same resnet50 backbone in order to find parameters \((\theta_f, \theta_y, \theta_d)\) that get us closest to the saddle point of \( E \). To aid in faster convergence, we decrease the learning rate hyperparameter over each epoch following\(^{26}\):

\[
\mu(p) = \frac{\mu_0}{(1 + \alpha p)^\beta}
\]

(7)

Similarly, the hyperparameter \( \lambda \) is increased over each epoch following\(^{26}\), while periodically setting it to 0 every 3 epochs

\[
\lambda(p) = \frac{2}{(1 + e^{-\alpha p})} - 1
\]

(8)
Such hyperparameter annealing is commonly practiced achieving better convergence during training\textsuperscript{30}. In equations 7 and 8, $p$ represents the training progress (fraction of total number epochs completed). The hyperparameters $\mu_0 = 0.001$, $\alpha = 10$, and $\beta = 0.75$ following\textsuperscript{26}. The training batch size was set to 256 (sampling 32 patches per whole slide image in each batch). As an early stopping criterion, model training was halted after 15 epochs as the gap between train error and validation error begins to widen after 15 epochs. Hence model training was halted after 15 epochs. The parameters achieving the best performance on the validation dataset over 15 epochs were saved and eventually used for making predictions on held out test data. The resnet50 architecture and training algorithm were implemented in python using PyTorch on an in house dedicated server using a single Nvidia RTX A6000 GPU with 48GB of VRAM.

Spatial probability map generation and burden estimation for each histologic subtype

To generate spatial probability maps, each whole slide image was processed by the trained patch-level histologic subtype classifier from left to right in a sliding window fashion with a window size of 256x256 pixels and an overlap of 64 pixels. The resulting probability maps generated were further down sampled to 5x base magnification via local average pooling of tile probabilities. We eventually generate 6 spatial probability maps: one for each class (excluding the “other” class representing normal/benign/hemorrhagic tissue). The resulting probability maps can then be converted to grey scale or color images and visualized as shown in Figures 2 and 3.

Having generated spatial probability maps for each histologic subtype, one can then estimate its absolute burden in each patient’s tumor while accounting for variable number of slides scanned per case using the following approach:

$$\Phi_{\text{case}}^{\text{subtype}} = \frac{1}{N} \sum_{ij} P_{ij}(\text{subtype}) > 0.5$$

$P_{ij}(\text{subtype})$ represents the probability of region $i,j$ being classified a particular subtype. The summation term represents the total area. The term $N$ in the denominator represents the number of slides scanned per case. We choose to quantify the absolute burden of each subtype instead of relative burden since each tumor was scanned at the same base magnification and we had access to multiple slides scanned for each tumor in our cohort including slides with tissue artefacts such as folded tissue and osteoid formations. See Supplementary table 2 for the estimated absolute burden of each subtype for all 306 canine cases analyzed in this study.

Data pre-processing for K-means clustering analysis

Given the estimated burden of each histologic subtype in each dog sample, we first center and scale the data and then perform a principal component analysis (PCA). The projections of each sample along the first two principal components, which capture most of the variability in the data, are then used for K-means clustering.

Implementation details of K-means clustering and survival analysis

To perform K-means clustering we used the kmeans() utility function implemented in R stats package with the following options set: max iterations = 500, nstart (number of random initializations of cluster centers) = 100. For performing Kaplan-Meier and cox proportional hazards regression analysis of the clinical data, we used the survfit() and cph() utility functions from the R survival package. Results of these analyses were plotted using the ggsurvplot() and ggforest utility functions from R survminer and GGally packages.
**Code availability**

The code to train a classification model using domain adversarial learning, trained model weights and scripts to reproduce the downstream results are available at: [https://github.com/spatkar94/adversarialdogs.git](https://github.com/spatkar94/adversarialdogs.git) (Date last accessed: 09/30/2022).

**Results**

**Overview of whole slide imaging cohorts analyzed in this study and the adversarial learning approach**

To precisely characterize the morphological heterogeneity of osteosarcomas, we systematically collected and scanned 600 Hematoxylin & Eosin (H&E) stained slides of treatment naïve primary tumors from a diverse collection of 306 dogs enrolled in a 2-armed NCI Comparative Oncology Trials Consortium (COTC) clinical trial. The distribution of dogs analyzed in this study by geographic location and breed are summarized in Supplementary Table 3 and 4. Additionally, 39 de-identified H&E slides of human osteosarcomas were collected to evaluate species-agnostic histologic features. A veterinary anatomic pathologist (J.B.) annotated 95 and 11 slides from canine and human samples, respectively to identify regions of necrosis (N) or tumor-specific histologic patterns including osteoblastic (OB), chondroblastic (CB), fibroblastic (FB), giant cell (GC) and vessel-rich (VR) regions. Unannotated regions were classified as “other”.

We then trained a resnet50 Convolutional Neural Network (CNN) on whole slide image patches of osteosarcoma to classify them into different histologic subtypes, necrosis or non-tumor areas in both dogs (source domain) and humans (target domain). Figure 1A, 1B and Supplementary table 1 depict the distribution of whole slide image patches corresponding to each class in training, validation, and test datasets generated for dogs and humans, respectively. Patches from both the dog and human training set were simultaneously fed to a resnet50 CNN trained using a domain adversarial approach (Figure 1C), which encourages neural networks to learn features that are important for the classification task of interest while at the same time less sensitive to domain-specific differences in the data. This was achieved by simultaneously training two classifiers that share the same feature extraction backbone. One classifier aimed to classify whole slide image patches into one of the pre-defined classes whereas the other classifier aimed to distinguish the domain of each patch (i.e., whether the patch comes from a dog or human sample). During training, the weights of the shared feature extraction backbone are updated such that we arrive at an equilibrium that minimizes classification error while maximizing domain error. Patches from the validation set were used to monitor for any signs of overfitting of the classification model (See Methods for more details). In the evaluation phase, patches from the held-out test set were evaluated using the trained histologic subtype classifier.

**Adversarial learning improves domain adaptation of the histologic subtype classifier from Dogs to Humans**

Having trained a patch-level histologic subtype classification model in a domain adversarial fashion, we next evaluated the performance of the trained model on held-out test whole slide
image patches in both dogs and humans. To evaluate the model’s performance, we computed the per class precision, recall and F1 scores obtained by comparing the model predicted class labels of each whole slide image patch in the test set to the ground-truth labels obtained from overlapping pathologist annotations (See methods). On average, the model achieved an F1-score of 0.77 (95% CI [0.74-0.79]) in dogs, and an F1-score of 0.8 (95% CI [0.79-0.81]) in humans (Figure 4, panels A, B, C and D). Overall, the histologic subtype classification model adapts from dogs (source domain) to humans (target domain) after seeing < 5% of labeled examples from the target domain. The subtype which had low precision (20%) and low recall (23%) on the target domain is the Chondroblastic subtype and was most often confused with the more dominant osteoblastic subtype.

To evaluate the effect of domain adversarial training on model generalizability from source domain (dogs) to target domain (humans), we performed three control experiments: i) train the image classification model on labeled data from the source domain only and evaluate on target domain (Transfer learning), ii) train the image classification model on labeled data from target domain only and evaluate on target domain, iii) train the image classification model on labeled data from both the source and target domain using standard supervised learning and evaluate on target domain. For each experiment, the classification model was trained starting from the same set of initialized weights and hyperparameters. Overall, we found that the domain adversarial learning approach achieved significantly lower test error per epoch compared to the other three controls when evaluated on the target domain (Figure 4E).

To visualize the predictions of the patch-level histologic subtype classification model on the whole slide image, we generated spatial probability maps depicting regions of high vs low probability for each histologic subtype based on application of the patch-level histologic subtype classification model over the whole slide image in a sliding window fashion (See Methods for details). As a qualitative validation, Figures 2 and 3 depict pathologist marked region boundaries within 4 dog and 2 human osteosarcoma surgical specimens covering each histologic subtype along with classifier-derived probability maps (one per histologic subtype) over the whole slide image.

Unsupervised exploratory analysis of whole slide imaging features uncovers distinct populations of dogs with different responses to standard of care therapy

Having generated spatial probability maps of each subtype, we next estimate the absolute burden of each subtype in each canine sample and apply the K means clustering algorithm to identify clusters of dogs with similar whole slide tumor histology (Supplementary Table 2, See methods). Figure 5A plots the average silhouette score of inferred clusters for different values of K. The higher the average silhouette score, the more compact and well separated are the clusters (maximum score = 1). The error bars indicate the confidence interval estimated by repeatedly performing K means clustering on randomly down-sampled versions of the original cohort (down-sampling to ~80% original cohort size), when keeping K fixed. The highest
silhouette score is achieved for K=3 clusters. Figure 5B depicts the data distribution along the first two principal components and corresponding cluster memberships.

We next examined the distribution of the estimated burden of each subtype in each cluster and the clinical outcomes. The clinical characteristics of the cases analyzed in this study are provided in Table 1. See Supplementary table 5 for all the clinical metadata. Cluster 3 had significantly higher levels of the vessel-rich regions, whereas cluster 2 had significantly higher tumor necrosis relative to the rest of the cohort and slightly elevated levels of the chondroblastic subtype (Figure 5C). Overall, we observe that dogs belonging to cluster 3 had significantly worse clinical outcomes compared to the other two clusters. Figure 6A shows a Kaplan Meier plot depicting differences in overall survival rates between dogs belonging to cluster 3 and rest of the cohort (log rank test p-value: 0.038), whereas Figure 6B depicts the differences in disease free interval rates between the dogs belonging to cluster 3 and rest of the cohort (log rank test p-value: 0.0071). All dogs belonging to cluster 3 relapsed within 12 months after receiving adjuvant treatment. This negative association remained significant despite adjusting for relevant clinical parameters such as tumor location (Proximal humerus vs non proximal humerus), alkaline phosphatase levels (elevated vs normal), age, weight, sex, and adjuvant treatment type in a multivariable cox proportional hazards regression model.

Finally, we performed sub-group analysis to ensure prognostic signatures remain significant in unlabeled data not used in training. The first subgroup consists of 55 reviewed cases (n=95 pathologist annotated slides). The second subgroup consists of the remaining 251 unreviewed cases. In each sub-group, the survival association remains consistent thus demonstrating the clinical utility of model predictions beyond cases previously annotated by the pathologist (See Supplementary Figure S2).

**Discussion**

Through the activities of the NCI COTC, this study examines the largest dataset of canine osteosarcomas to date for which complete clinical outcome data is available and standardized therapy was applied (n = 306). With the help of this large resource, we demonstrate how deep domain adversarial learning can be used to train a histologic subtype classifier that adapts from dog to human osteosarcoma despite utilizing a very small fraction of human data for training. Although this is not the first application of deep learning in osteosarcomas\(^{32-34}\), it is the first attempting to identify histologic features of osteosarcoma that transfer from canine to human samples to the best of our knowledge.

With the help of the trained species-agnostic histologic subtype classifier, we performed an unsupervised exploratory analysis of whole slide imaging data of 306 dogs and identified distinct clusters that respond very differently to standardized chemotherapy based on the classifier-estimated burden of histologic subtypes. Our results are consistent with some prior reports indicating that the presence of specific histologic subtypes may have prognostic value\(^{11, 13}\); however, a rigorous quantitative evaluation of OS histology that takes tumor heterogeneity into account has not been previously explored likely due to the difficulty in accumulating a large
enough dataset and the immense manual labor by the pathologist in annotating each region. This is the first exploratory study using AI to define prognostic value of variant histologic features within a large population of dogs receiving standardized care in a prescriptive clinical trial. As with the diagnostic and therapeutic approach to any cancer, many separate factors should be considered when devising a treatment and prognosis. The predictive value of our approach should be considered alongside other patient factors and not considered the sole method by which prognosis can be assigned for canine patients with OS. Nevertheless, information gleaned from our approach is of substantial clinical value to clinicians treating dogs with OS.

In this study we refrain from quantifying overlap between pathologist annotations and AI predictions using Dice or IoU metrics. These metrics are preferable in segmentation applications where the ground truth segmentation boundaries are precisely defined\textsuperscript{35}. However, due to intra-tumor heterogeneity, osteoblastic tumor cells are frequently observed intermixed with other histologic subtypes\textsuperscript{3, 5, 25}. Hence, it is not feasible for pathologists to precisely mark region boundaries of each histologic subtype at high resolution for each slide. Although the pathologist annotated the majority of tumor tissue in all annotated sections; there are examples where unannotated tumor tissue was present. Interestingly, these cases offer another example demonstrating the ability of the model to identify tumor tissue that would not be captured by Dice or IoU metrics. For example, in Figure 2D, there are several regions that were predicted to contain osteoblastic tumor cells. On review, the pathologist was able to confirm the presence of osteoblastic tumor tissue in these locations (Supplementary Figure S3). This highlights a potential utility of AI in identifying foci of tumor distal to the main tumor mass. This may be particularly important in tumors that require complete excision and could help by re-orientating the pathologist toward specific regions to review.

In this study, tumors enriched for VR regions were associated with reduced DFI and OS. These vascular structures define the rare telangiectatic subtype of osteosarcoma which is characterized by blood-filled cystic spaces surrounded by thin septa lined by tumor cells\textsuperscript{3, 5}. Although an early study\textsuperscript{36} suggested that telangiectatic OS carry a poor prognosis in human patients, others suggest that while there may be a correlation with clinical features, such as pathologic fracture, an association with prognosis is less clear\textsuperscript{37}. In dogs, the telangiectatic subtype has been associated with poor prognosis in studies of OS originating in the ulna\textsuperscript{38} (n = 30) or flat and irregular bones\textsuperscript{39} (n = 45). In our case set, we defined VR regions as containing blood-filled spaces lined by tumor cells. On H&E, these vascular spaces were multifocally lined by polygonal cells rather than flat, spindle-shaped cells, which were more likely to be interpreted as endothelium histologically. CD31 immunohistochemistry (IHC) staining confirmed the presence of vessels lined by tumor cells in VR-annotated canine osteosarcomas (See Supplementary Figure S2). Some VR regions also contained cellular debris which has been described in human OS\textsuperscript{40-42}. While VR morphology was uncommon in our dataset, the presence of tumor cell-lined vascular structures in largely solid tumors suggests that vascular differentiation can occur within a focal region of these histologically diverse tumors. Such tumors are less likely to be classified as telangiectatic OS.
which may inhibit the prognostication of histologic subtype in OS. This is emphasized by a study of OS originating in the ulna (n=30) that identified reduced survival in dogs with either pure or mixed telangiectatic morphology (i.e., telangiectatic or osteoblastic-telangiectatic). In fact, up to 65% of canine osteosarcomas are reported to demonstrate multiple histologic subtypes. This underscores the utility of AI which allows pathologists to rapidly quantify the abundance of major and minor histologic patterns within heterogeneous tumors.

Despite the merits of this study, there are still a few notable limitations that should be considered. First, we did not have access to human clinical outcome data to assess the prognostic value added by our approach over what is currently clinically practiced for humans. A future direction will be to apply this methodology to a larger set of human OS images with matched clinical outcomes to determine algorithm performance in a translational setting. Second, our study is based on annotations from a single anatomic pathologist. Agreement between pathologists can vary based on the feature of interest. This may be greater in cases where pathologists must consider an aggregate of histologic features to assign a tumor grade. For example, in one veterinary study of osteosarcomas, agreement was considered moderate for necrosis (ICC = 0.626) while agreement on grade was fair using 3 different classification systems. In the future, we aim to convene a comparative pathology board of MD and DVM pathologists to review canine and human osteosarcoma histology with the goal of assessing the impact of our model on interobserver variability, identifying additional features, such as immune cell infiltration, that may be incorporated into our prognostic model alongside ongoing genomic work. Third, the data is severely imbalanced with only a handful of canine and human tumor cases exhibiting uncommon histologic subtypes. To ensure that there exist enough training examples of each class for the patch-level classifier, pathologist annotated whole slide images were broken into non-overlapping patches scanned at high magnification and split at random into train validation and test sets (See methods). Patch-based training of neural networks in digital pathology has enabled accurate detection and quantification of complex histologic features on few whole slide images due to thousands of image patches that can be extracted during training at high magnifications. However, neural networks trained this way are prone to overfitting to slide, staining or scanner specific properties. In this work, we reasoned that an adversarial learning approach could help neural networks overcome the bias that would be present in domain-specific training paradigms. Adversarial training can however be complex in practice compared to standard supervised learning approaches. This is especially relevant during initial phases of training where noisy signals from the domain classifier can derail the learning algorithm. This issue is mitigated by having a good initialization of model parameters and by gradually increasing the influence of domain classifier in the learning process as defined in detail in the methods section. Lastly, no additional manual quality control of surgical tumor specimens was completed prior to data collection from different sites. Instead, our model was adversarially trained to classify non-tumor regions in addition to the six different histologic subtypes of osteosarcoma based on pathologist annotations. We expect the robustness and accuracy of the classification model to improve as additional data is collected.
In summary, deep domain adversarial learning could be a powerful addition to the modern pathologist's toolbox for identification of domain-agnostic histologic and molecular features of tumors and is likely to be useful for many other comparative oncology applications, especially where human data is scarce.

Acknowledgements
The authors would like to thank the Comparative Oncology Clinical Trials Consortium (COTC) members for their execution of the COTC-21/022 trials which provided the clinical outcome data that was analyzed here. We would additionally like to thank Dr. Markku Miettinen for granting us access to 39 Human OS slides from his residency training materials.
References

[15] Skorupski KA, Uhl JM, Szivek A, Allstadt Frazier SD, Rebhun RB, Rodriguez CO, Jr.: Carboplatin versus alternating carboplatin and doxorubicin for the adjuvant treatment of
Figure Legends

Figure 1. Overview of the training data and adversarial learning approach. A: Non-overlapping whole slide image patches from 95 canine whole slide images were extracted at 10x base magnification and split at random into 80% train, 10% validation and 10% test. The distribution of patches by each class are shown. B: Non-overlapping whole slide image patches from 11 human whole slide images were extracted at 10x base magnification. 2000 patches (~3% of all labeled human patches) were reserved for domain adversarial training of the histologic subtype classifier. The rest were held out for testing. See methods for details on how each whole slide image patch was assigned a class. C: Overview of the supervised domain adversarial learning approach. The domain classifier is made to work against the histologic subtype classifier by introducing a gradient reversal layer just before the domain classifier. For more details on the algorithm, see Materials and Methods.

Figure 2. A-D: Pathologist marked regions vs classifier generated spatial probability maps for each OS subtype over whole slide images of tumor samples from dogs. The probability maps (depicted below each whole slide image) are generated by applying the trained patch-level subtype classifier in a sliding window fashion over the whole slide image (scanned at 10x base magnification) using a window size of 256x256 pixels. For more details, see Materials and Methods. OB: Osteoblastic, CB: Chondroblastic, FB: Fibroblastic, GC: Giant cell-rich, VR: Vessel-rich, N: Necrosis

Figure 3 A-B: Pathologist marked regions vs classifier generated spatial probability maps for each OS subtype over whole slide images of tumor samples from humans. The probability maps (depicted below each whole slide image) are generated by applying the trained patch-level subtype classifier in a sliding window fashion over the whole slide image (scanned at 10x base magnification) using a window size of 256x256 pixels. OB: Osteoblastic, CB: Chondroblastic, FB: Fibroblastic, GC: Giant cell-rich, VR: Vessel-rich, N: Necrosis

Figure 4. Performance evaluation on held out whole slide image patches from the test set in both dogs and humans. A and B: Confusion matrices generated after evaluating model predictions on dog and human whole slide image patches from the held-out test set, respectively. The rows represent the predicted class of each whole slide patch (i.e., the class achieving the highest probability based on the classification model). The columns represent the ground truth (i.e. pathologist assigned class). Below each confusion matrix is a histogram depicting the distribution of ground truth class labels in the held-out test set. C: Depicts the evolution of the test error achieved by the classification model on human whole slide image patches (target domain) as we progress through each training epoch. The red points represent the test error trajectory achieved through adversarial learning. The rest represent the test error trajectories of the remaining control methods. The test error is defined as the average multi-class cross-entropy loss over the entire epoch. D and E: Estimated per-class precision and recall of the classification model on held out test patches in dogs and humans, respectively. The error bars were approximately determined by a bootstrap analysis where we repeatedly down sampled the test datasets to 50% original size and recomputed precision and recall on each down sampled version.

Figure 5. K-means clustering analysis of 306 Canine OS tumors based on estimated burden of histologic subtypes. A: Depicts the average silhouette score as a function of the number of clusters used by the K-means algorithm to cluster the data. The higher the average silhouette score, the better the clustering. The smallest value of K achieving the highest silhouette score represents the best possible clustering of the data. B: Principal component analysis plot depicting the distribution of all canine osteosarcoma cases based on the estimated burden of each histologic subtype. Points belonging to cluster 1 are colored red,
points belonging to cluster 2 are colored green and points belonging to cluster 3 are colored blue. C: distribution of the burden of each histologic subtype in each cluster. OB: Osteoblastic, CB: Chondroblastic, FB: Fibroblastic, GC: Giant cell-rich, VR: Vessel-rich, N: Necrosis (**: Mann Whitney U test p-value < 0.01, ***: Mann Whitney U test p-value < 0.001, ****: Mann Whitney U test p-value < 0.0001)

**Figure 6. Survival outcomes of cluster 3 vs clusters 1,2. A:** (Above) Kaplan-Meier plot depicting the overall survival rates of cases belonging to cluster 3 vs rest (cluster 1 or cluster 2). (Below) Estimated hazards ratio of each factor. **B:** (Above) Kaplan-Meier plot depicting the disease-free survival rates of cases belonging to cluster 3 vs rest (cluster 1 or cluster 2). (Below) Estimated hazards ratio of each factor. The log-rank test p-value was estimated to determine the significance of the differences in survival rates (* log-rank test p-value < 0.05, ***: log-rank test p-value < 0.001). SOC: Standard of Care, DFI: Disease Free Interval, PH: Proximal Humerus, NPH: Non-proximal Humerus, ALP: Alkaline Phosphatase, AIC: Akaike Information Criterion.
Table 1: Clinical Characteristics of the Dog Osteosarcoma cohort (N=306). For continuous variables, values in parentheses represent the minimum and maximum range, values outside the parentheses represent the median over the entire cohort.

<table>
<thead>
<tr>
<th>Clinical characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>8.1 (1.4-15.6) years</td>
</tr>
<tr>
<td>Weight</td>
<td>38.8 (21.2-94.5) kg</td>
</tr>
<tr>
<td>Tumor Location</td>
<td></td>
</tr>
<tr>
<td>Proximal Humerus (PH)</td>
<td>N=64 (21%)</td>
</tr>
<tr>
<td>Non-Proximal Humerus (NPH)</td>
<td>N=242 (79%)</td>
</tr>
<tr>
<td>ALP levels</td>
<td></td>
</tr>
<tr>
<td>Elevated</td>
<td>N=74 (24%)</td>
</tr>
<tr>
<td>Normal</td>
<td>N=232 (76%)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Castrated Male</td>
<td>N=171 (56%)</td>
</tr>
<tr>
<td>Intact Male</td>
<td>N=13 (4%)</td>
</tr>
<tr>
<td>Spayed Female</td>
<td>N=118 (39%)</td>
</tr>
<tr>
<td>Intact Female</td>
<td>N=4 (1%)</td>
</tr>
<tr>
<td>Disease-free interval</td>
<td>157 (3-1127) days from surgery</td>
</tr>
<tr>
<td>Overall survival</td>
<td>235 (3-1652) days from surgery</td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
</tr>
<tr>
<td>Standard of Care (SOC)</td>
<td>155 (51%)</td>
</tr>
<tr>
<td>Standard of Care + Sirolimus (Rapamycin)</td>
<td>151 (49%)</td>
</tr>
</tbody>
</table>
Canine Osteosarcoma (Source Domain)

Human Osteosarcoma (Target Domain)

### #WSI patches

- **Train (76501)**
- **Validation (9447)**
- **Test (9447)**

### Domain labels

- CB
- FB
- GC
- VR
- OB
- Necrosis
- other

For Canine Osteosarcoma (Source Domain):

- Train: 1363, 136, 1569, 516, 4206, 3691, 3766, 5157, 633, 648, 4186, 4092
- Validation: 7, 93
- Test: 1174, 607, 571, 14770, 640, 117, 3391, 598

For Human Osteosarcoma (Target Domain):

- Train: 136, 175, 202, 206, 516, 56, 4206, 543, 484, 543, 56, 56, 202, 206, 516, 56, 4206, 543, 484
- Test: 13, 6, 17, 607, 571, 640, 117, 3391, 598, 1174, 640, 16714, 117, 3391, 598

### Output

- **Resnet50 Feature Extractor**
- **Histologic subtype classifier**
- **Adversary (Dog or human sample??)**
A) Dogs

<table>
<thead>
<tr>
<th>Target</th>
<th>VR</th>
<th>other</th>
<th>OB</th>
<th>Necrosis</th>
<th>GC</th>
<th>FB</th>
<th>CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>VR</td>
<td>308</td>
<td>8</td>
<td>163</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>other</td>
<td>35</td>
<td>3522</td>
<td>461</td>
<td>64</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>OB</td>
<td>39</td>
<td>402</td>
<td>3237</td>
<td>47</td>
<td>15</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>Necrosis</td>
<td>2</td>
<td>114</td>
<td>94</td>
<td>430</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>GC</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>45</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FB</td>
<td>0</td>
<td>32</td>
<td>24</td>
<td>5</td>
<td>0</td>
<td>145</td>
<td>0</td>
</tr>
<tr>
<td>CB</td>
<td>0</td>
<td>11</td>
<td>35</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>117</td>
</tr>
</tbody>
</table>

B) Humans

<table>
<thead>
<tr>
<th>Target</th>
<th>VR</th>
<th>other</th>
<th>OB</th>
<th>Necrosis</th>
<th>GC</th>
<th>FB</th>
<th>CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>VR</td>
<td>1371</td>
<td>320</td>
<td>643</td>
<td>84</td>
<td>0</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>other</td>
<td>370</td>
<td>14584</td>
<td>528</td>
<td>99</td>
<td>5</td>
<td>41</td>
<td>7</td>
</tr>
<tr>
<td>OB</td>
<td>1276</td>
<td>969</td>
<td>13928</td>
<td>276</td>
<td>96</td>
<td>96</td>
<td>73</td>
</tr>
<tr>
<td>Necrosis</td>
<td>154</td>
<td>32</td>
<td>84</td>
<td>3116</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>GC</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>522</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>FB</td>
<td>24</td>
<td>7</td>
<td>7</td>
<td>1136</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>CB</td>
<td>1</td>
<td>2</td>
<td>68</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

C) Test error (human)

- 1. Wilcoxon signed rank test p-value (adversarial vs control I): 6.1e-5***
- 2. Wilcoxon signed rank test p-value (adversarial vs control II): 6.1e-5***
- 3. Wilcoxon signed rank test p-value (adversarial vs control III): 0.0004***

D) Precision and Recall

E) Precision and Recall

- Adversarial training on dog data + few human samples
- Training on dog data only (control I)
- Training on few human examples (control II)
- Training on dog data + few human examples (control III)
Dog osteosarcoma overall survival

Clusters {1, 2} vs. cluster 3

Survival probability

Number at risk

296 91 39 12 1
10 0 0 0 0

Hazard ratio

<table>
<thead>
<tr>
<th>Location</th>
<th>NPH (N=243)</th>
<th>reference</th>
<th>PH (N=63) 1.0 (0.71 - 1.5)</th>
<th>0.916</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALP</td>
<td>Normal (N=232)</td>
<td>reference</td>
<td>Elevated (N=74) 1.4 (1.03 - 2.0)</td>
<td>0.031 *</td>
</tr>
<tr>
<td>Age</td>
<td>(N=306) 1.1 (0.97 - 1.3)</td>
<td>0.115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>Spayed/Intact Female (N=122) 1.0 (0.76 - 1.3)</td>
<td>0.955</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Castrated/Intact Male (N=184)</td>
<td>reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>SOC (N=155)</td>
<td>reference</td>
<td>Sirolimus+SOC (N=151) 1.0 (0.77 - 1.3)</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cluster 1, 2 (N=296) 1.9 (0.82 - 4.3)</td>
<td>0.138</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cluster 3 (N=10)</td>
<td></td>
</tr>
</tbody>
</table>

Number of Events: 210; Global p-value (Log-Rank): 0.044109
AIC: 2067.86; Concordance Index: 0.58

Dog osteosarcoma DFI

Clusters {1, 2} vs. cluster 3

Survival probability

Number at risk

296 83 41 19 0
10 1 0 0 0

Hazard ratio

<table>
<thead>
<tr>
<th>Location</th>
<th>NPH (N=243)</th>
<th>reference</th>
<th>PH (N=63) 1.04 (0.73 - 1.5)</th>
<th>0.824</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALP</td>
<td>Normal (N=232)</td>
<td>reference</td>
<td>Elevated (N=74) 1.30 (0.96 - 1.8)</td>
<td>0.095</td>
</tr>
<tr>
<td>Age</td>
<td>(N=306) 1.09 (0.99 - 1.3)</td>
<td>0.231</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>Spayed/Intact Female (N=122) 0.99 (0.75 - 1.3)</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Castrated/Intact Male (N=184)</td>
<td>reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>SOC (N=155)</td>
<td>reference</td>
<td>Sirolimus+SOC (N=151) 1.06 (0.82 - 1.4)</td>
<td>0.642</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cluster 1, 2 (N=296) 2.29 (1.06 - 4.9)</td>
<td>0.034 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cluster 3 (N=10)</td>
<td></td>
</tr>
</tbody>
</table>

Number of Events: 237; Global p-value (Log-Rank): 0.0016263
AIC: 2289.71; Concordance Index: 0.59