The American Journal of Pathology, Vol. 193, No. 1, January 2023

The American Journal of
PATHOLOGY
ajp.amjpathol.org

MACHINE LEARNING, COMPUTATIONAL PATHOLOGY, AND BIOPHYSICAL IMAGING

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. Thomas Brown,* and Amy LeBlanc†

From the Artificial Intelligence Resource and the Comparative Oncology Program,† Molecular Imaging Branch, National Cancer Institute, NIH, Bethesda, Maryland*

Accepted for publication September 28, 2022.

Address correspondence to Sushant Patkar, Ph.D., National Cancer Institute, 5413 W. Cedar Ln., Ste. 102-C, Bethesda, MD 20814; or Amy Leblanc, D.V.M., National Cancer Institute, Bldg. 10, Room 1B53, Bethesda, MD 20892.
E-mail: patkar.sushant@nih.gov or amy.leblanc@nih.gov.

Osteosarcomas (OSs) are aggressive bone tumors with many divergent histologic patterns. During pathology review, OSs are subtyped based on the predominant histologic pattern; however, tumors often demonstrate multiple patterns. This high tumor heterogeneity coupled with scarcity of samples compared with other tumor types render histology-based prognosis of OSs challenging. To combat lower case numbers in humans, dogs with spontaneous OSs have been suggested as a model species. Herein, a convolutional neural network was adversarially trained to classify distinct histologic patterns of OS in humans using mostly canine OS data during training. Adversarial training improved domain adaption of a histologic subtype classifier from canines to humans, achieving an average multiclass F1 score of 0.77 (95% CI, 0.74—0.79) and 0.80 (95% CI, 0.78—0.81) when compared with the ground truth in canines and humans, respectively. Finally, this trained model, when used to characterize the histologic landscape of 306 canine OSs, uncovered distinct clusters with markedly different clinical responses to standard-of-care therapy. *(Am J Pathol 2023, 193: 60—72; https://doi.org/10.1016/j.ajpath.2022.09.009)*

Osteosarcoma (OS) is a rare but aggressive pediatric malignancy with approximately 800 cases reported annually in the United States.[1] Patients with metastatic or relapsed disease have dismal outcomes, with survival rates of <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 on review of tumor sections harvested after local tumor control via surgery. Despite this, a subset of patients with high necrosis still develop metastatic disease after completion of frontline therapy. Hence, additional prognostic biomarkers are needed for accurate prognosis prediction. Because 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, occurs in >90% of canine patients within several months of diagnosis.[10–14] In contrast to humans, the clinical workflow in

Supported by the Intramural Program of the National Cancer Institute, NIH ZIH BC 012032 (A.L.).
Disclosures: None declared.

https://doi.org/10.1016/j.ajpath.2022.09.009

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 (ie, *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, alkaline phosphatase status, and age/sex/breed.[10,12,14] Studies examining the prognostic significance of histologic subtype have identified conflicting findings in different data sets.[11,13] This study took advantage of a larger patient cohort accumulated during a prospective randomized clinical trial conducted in >300 canine patients.[12] This yielded a well-annotated canine OS data set in which to examine osteosarcoma histology and explore the potential of artificial intelligence (AI)—derived biomarkers. Specifically, the study investigated whether techniques in AI using adversarial learning could support the development of a histologic subtype classifier for osteosarcomas that adapts from dogs to humans and a prognostic signature in dogs based on digital pathology whole slide images.[15–17]

Materials and Methods

Curation of Hematoxylin and Eosin—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 Oncology Trials Consortium (COTC) institutions (*https://ccr.cancer.gov/comparative-oncology-program/consortium*, last accessed May 13, 2022). 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 before any treatment. Dogs were randomized to receive either standard of care 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 disease-free interval or survival between the two arms; thus, cases were included together in the analysis presented herein. In addition 39 human osteosarcoma samples were obtained from an in-house pathology residency training cohort. Of these 39 samples, 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 ×40 magnification or 0.23 μm per pixel. No additional manual quality control of surgical tumor specimen size or percentage tumor tissue was completed before 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 were derived.

Annotation and Preprocessing of Whole Slide Image Data

Pathologist annotations for 95 dog slides and 11 human slides were obtained in xml format using HALO (Albuquerque, NM). Each annotation file contained coordinates of roughly marked region boundaries for each histologic subtype within each slide. Because 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 necrotic, vessel rich (VR), chondroblastic, fibroblastic, or giant cell rich.[3,5] In canine tumors annotated as VR, CD31 immunohistochemistry was used to confirm the presence of tumor cell (CD31-) lined vascular spaces (Supplemental Figure S1).[18,19] Any unmarked regions falling outside main tumor areas were classified as other and consisted primarily of nontumor tissue, osteoid formations, and, in some cases, slide preparation artifacts, such as folded tissue and slide debris.

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 histologic features of interest are open ended.[20–23] However, in this study, we had region-level pathologist annotations that were based on previously defined histologic subtypes of osteosarcoma that are distinguishable at ×10 magnification level.[24] The smallest regions of interest annotated by the pathologist have an area of approximately 25,000 μm^2 and are represented by at least one tile of size 256 × 256 at ×10 magnification. A larger tile size would have resulted in fewer training tiles per histologic 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 (eg, tumor cells surrounding blood-filled spaces, which are a characteristic feature of telangiectatic osteosarcoma). Hence, to train our image classification model, each whole slide image was scanned at ×10 magnification level and broken down into 256 × 256 pixel tiles.

Tiles containing >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 (ie, a region exceeding 15% of the tile area), the tile was assigned the

histologic class of the most dominant lesion (ie, the divergent lesion covering the highest percentage area). Otherwise, the tile was assigned label osteoblastic. For example, if a tile had 35% of its area marked as fibroblastic, then the tile gets assigned the label fibroblastic. If a tile is dominated by nontumor tissue or hemorrhage, it was assigned the label other. All other tiles from unmarked slides were regarded as unlabeled.

For training, 80% of all labeled tiles from dogs (source domain) and additional 2000 labeled tiles from humans (target domain) were randomly selected. Of the remaining 20% labeled tiles from dogs, half were randomly selected for validation and hyperparameter 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, the random seed in the codes generating the train, validation, and test splits was fixed. The distribution of tiles by histologic subtype and train, validation, and test split is shown in Figure 1 and Supplemental Table S1.

Before feeding a tile as input to the classification model, each tile was rescaled to 224 × 224 pixels, and its per-channel pixel intensities (ranging from 0 to 1) were normalized to follow a standard normal distribution using the following per-channel mean intensity and SDs estimated from the dog training data: mean ($r = 0.8938$, G = 0.5708, B = 0.7944) and SD ($r = 0.1163$, 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.

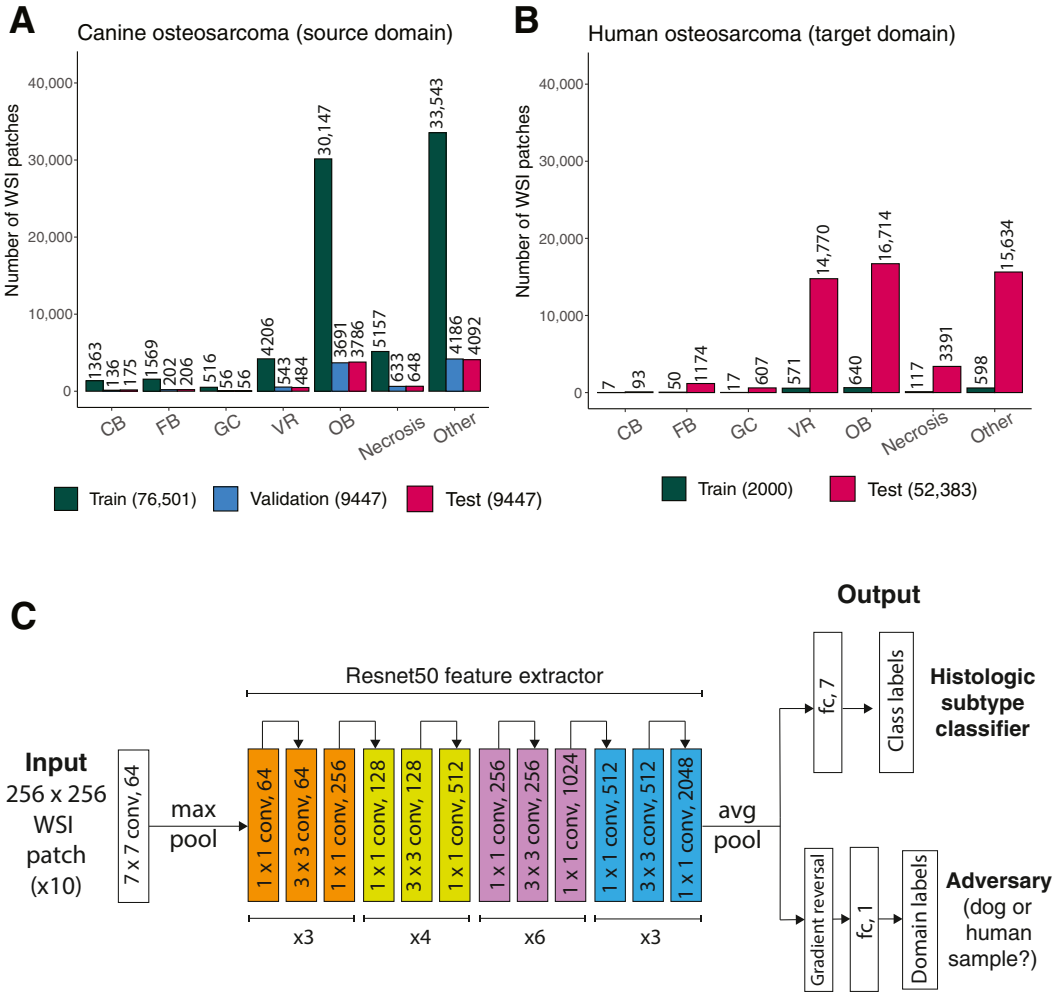

Figure 1 Overview of the training data and adversarial learning approach. **A:** Nonoverlapping whole slide image (WSI) patches from 95 canine whole slide images were extracted at ×10 base magnification and split at random into 80% train, 10% validation, and 10% test. The distribution of patches by each class is shown. **B:** Nonoverlapping whole slide image patches from 11 human whole slide images were extracted at ×10 base magnification. A total of 2000 patches (approximately 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 *Materials and 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*. Avg, average; CB, chondroblastic; conv, convolution; FB, fibroblastic; GC, giant cell rich; max, maximum; OB, osteoblastic; VR, vessel rich.

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 ($d = dogs$) and $(X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)$ be examples from a target domain ($d = humans$) where the number of examples available is 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 Ganin and Lempitsky[25] to the supervised setting. Specifically, let θ_f be the parameter of the feature extraction backbone $G_f(.;\theta_f)$, (ie, the function that takes as input an example X_i and maps it to a set of features), let θ_y be the parameter of the subtype classifier $G_y(.;\theta_y)$, (ie, the function that receives input from the feature extractor and predicts class label Y_i), and let θ_d be the parameter of the domain classifier $G_d(.;\theta_d)$ (ie, the function that receives input from the feature extractor and predicts the domain label d_i). Furthermore, let:

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

$$= \sum_{i=1}^{m+n} L_y^i(\theta_f, \theta_y) - \lambda \sum_{i=1}^{m+n} L_d^i(\theta_f, \theta_d) \quad (1)$$

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 λ controls the trade-off between the two errors. The goal of a domain adaption algorithm is then to find the saddle point of E:

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg\min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d) \quad (2)$$

$$\hat{\theta}_d = \arg\max_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d) \quad (3)$$

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 θ_f at the saddle point minimize the subtype classification error (ie, the learned features are discriminative) while maximizing the domain classification error (ie, 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 (Figure 1). The update rule for the parameters after incorporating the gradient reversal layer is given by Equations 4, 5, and 6:

$$\theta_f \leftarrow \theta_f - \mu\left(\frac{\partial L_y^i}{\partial \theta_f} - \lambda \frac{\partial L_d^i}{\partial \theta_f}\right) \quad (4)$$

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y} \quad (5)$$

$$\theta_d \leftarrow \theta_d - \mu \frac{\partial L_d^i}{\partial \theta_f} \quad (6)$$

The hyperparameter μ represents the learning rate. To obtain a head start during training, we initialize the parameters of the feature extraction portion of the resnet50 convolutional neural network (θ_f) to the values obtained from pretraining resnet50 on the ImageNet data set.[26] Initializing convolutional neural networks with pretrained weights from ImageNet has previously demonstrated success in transfer learning on many digital pathology applications.[27,28] 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 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 Ganin and Lempitsky[25]:

$$\mu(p) = \frac{\mu_0}{(1 + \alpha p)^{\beta}} \quad (7)$$

Similarly, the hyperparameter λ is increased over each epoch, following Ganin and Lempitsky,[25] while periodically setting it to 0 every three epochs.

$$\lambda(p) = \frac{2}{(1 + e^{-\alpha p})} - 1 \quad (8)$$

Such hyperparameter annealing is commonly practiced, achieving better convergence during training.[29] 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 $\beta = 0.75$ are following Ganin and Lempitsky.[25] 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 data set 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 (*https://pytorch.org*) on an in-house dedicated server using a single Nvidia Ray Tracing Texel eXtreme A6000 Graphics Processing Unit with 48 GB of video RAM (Nvidia, Santa Clara, CA).

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 manner with a window size of 256 × 256 pixels and an overlap of 64 pixels. The resulting probability maps generated were further down sampled to ×5 base magnification via local average pooling of tile probabilities. We eventually generate six 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 gray scale or color images and visualized as shown in Figures 2, A–C 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:

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

$P_{ij}(subtype)$ 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. Absolute burden of each subtype, instead of relative burden, was quantified because each tumor was scanned at the same base magnification. Additionally, multiple slides scanned for each tumor in our cohort were available, including slides with tissue artifacts, such as folded tissue, and osteoid formations. See Supplemental Table S2 for the estimated absolute burden of each subtype for all 306 canine cases analyzed in this study.

Data Preprocessing for K-Means Clustering Analysis

Given the estimated burden of each histologic subtype in each dog sample, the study first centered and scaled the data

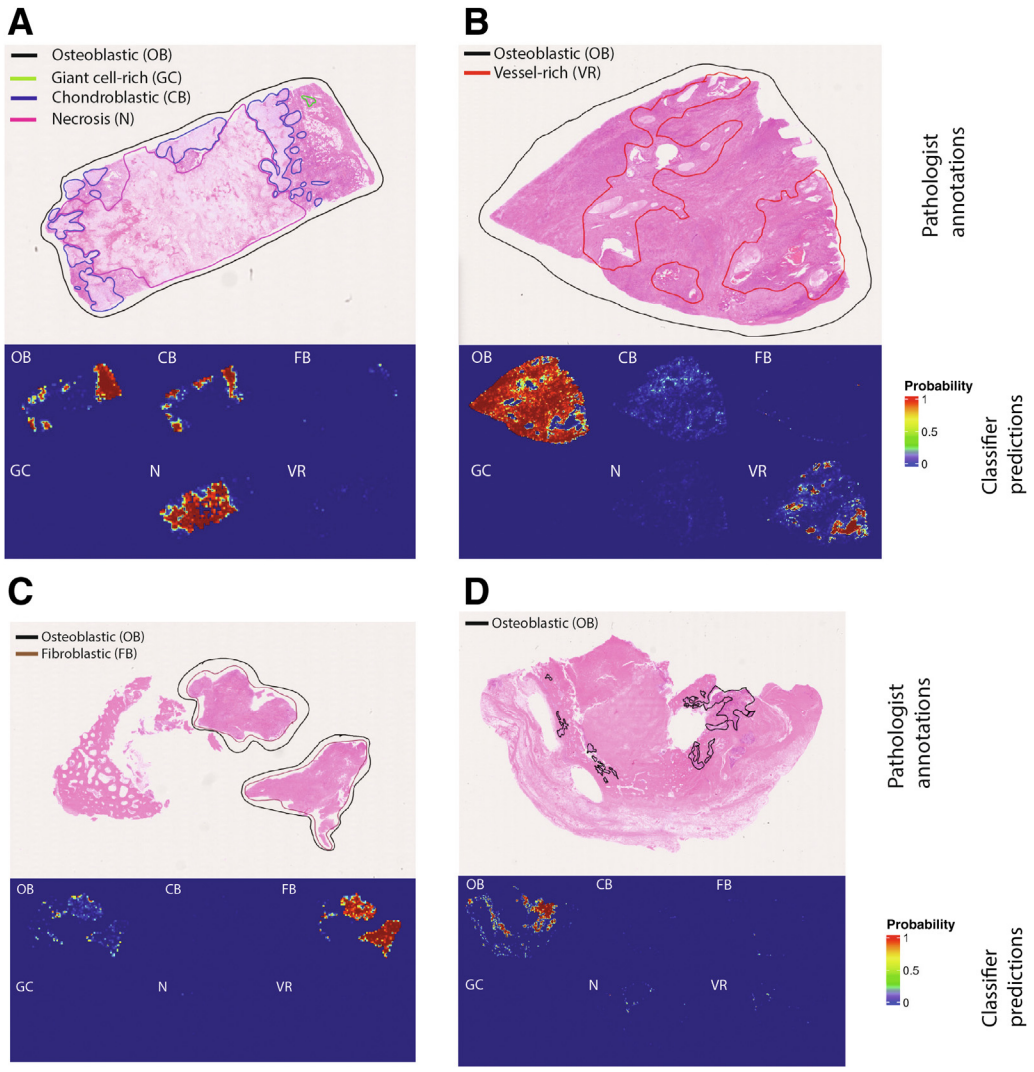

Figure 2 **A–D:** Pathologist-marked regions versus classifier-generated spatial probability maps for each osteosarcoma subtype over whole slide images of tumor samples from dogs. The probability maps (depicted below each whole slide image) are generated by applying the trained patch-level subtype classifier in a sliding window manner over the whole slide image using a window size of 256 × 256 pixels. For more details, see *Materials and Methods*. Original magnification, ×10 (**A–D**). CB, chondroblastic; FB, fibroblastic; GC, giant cell rich; N, necrosis; OB, osteoblastic; VR, vessel rich.

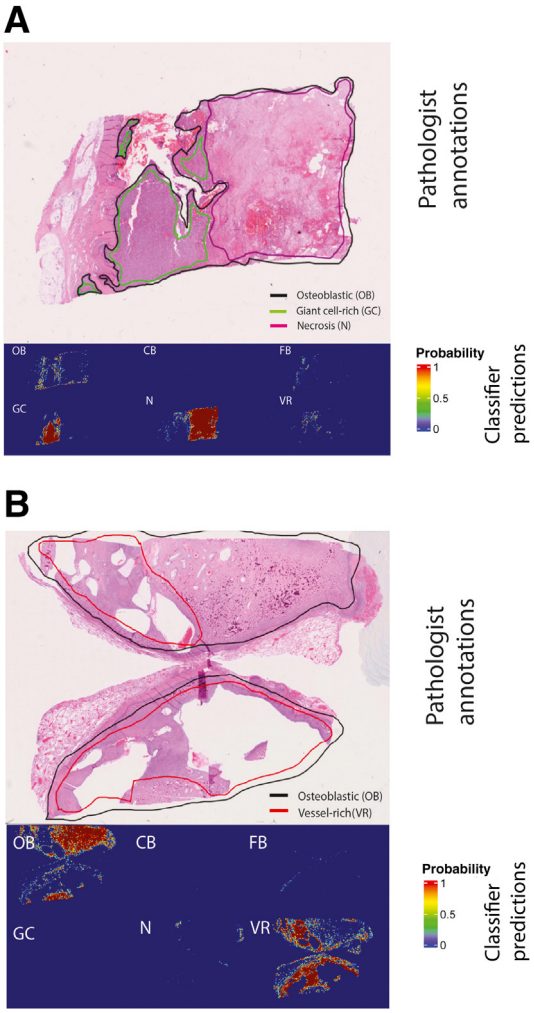

Figure 3 **A** and **B:** Pathologist-marked regions versus classifier-generated spatial probability maps for each osteosarcoma subtype over whole slide images of tumor samples from humans. The probability maps (depicted below each whole slide image) are generated by applying the trained patch-level subtype classifier in a sliding window manner over the whole slide image using a window size of 256 × 256 pixels. Original magnification, ×10 (**A** and **B**). CB, chondroblastic; FB, fibroblastic; GC, giant cell rich; N, necrosis; OB, osteoblastic; VR, vessel rich.

and then performed a principal component analysis. The projections of each sample along the first two principal components, which capture most of the variability in the data, were then used for K-means clustering.

Implementation Details of K-Means Clustering and Survival Analysis

To perform K-means clustering, the kmeans() utility function implemented in R stats package (*https://cran.r-project.org*) was used with the following options set: maximum iterations = 500, and nstart (number of random initializations of cluster centers) = 100. For performing Kaplan-Meier and Cox proportional hazards regression analysis of the clinical data, the survfit() and cph() utility functions from the R survival package were used. 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 (*https://github.com/spatkar94/adversarialdogs.git*, last accessed September 30, 2022).

Results

Overview of Whole Slide Imaging Cohorts Analyzed in this Study and the Adversarial Learning Approach

To precisely characterize the morphologic heterogeneity of osteosarcomas, 600 hematoxylin and eosin−stained slides of treatment-naïve primary tumors were systematically collected and scanned from a diverse collection of 306 dogs enrolled in a two-armed National Cancer Institute COTC clinical trial.[12] The distribution of dogs analyzed in this study by geographic location and breed is summarized in Supplemental Tables S3 and S4. In addition, 39 de-identified hematoxylin and eosin 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 or tumor-specific histologic patterns,[3−5] including osteoblastic, chondroblastic, fibroblastic, giant cell−rich, and VR regions. Unannotated regions were classified as other.

A resnet50 convolutional neural network was trained on whole slide image patches of osteosarcoma to classify them into different histologic subtypes, necrosis, or nontumor areas in both dogs (source domain) and humans (target domain). Figure 1, A and B, and Supplemental Table S1 depict the distribution of whole slide image patches corresponding to each class in training, validation, and test data sets generated for dogs and humans, respectively. Patches from both the dog and human training set were simultaneously fed to a resnet50 convolutional neural network 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.[25] 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 predefined classes, whereas the other classifier aimed to distinguish the domain of each patch (ie, whether the patch comes from a dog or human sample). During training, the weights of the shared feature extraction backbone are updated to 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 *Materials and 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 manner, the study 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, the study 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 with the ground-truth labels obtained from overlapping pathologist annotations (see *Materials and 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, A−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 that 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), three control experiments were performed: 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, and 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

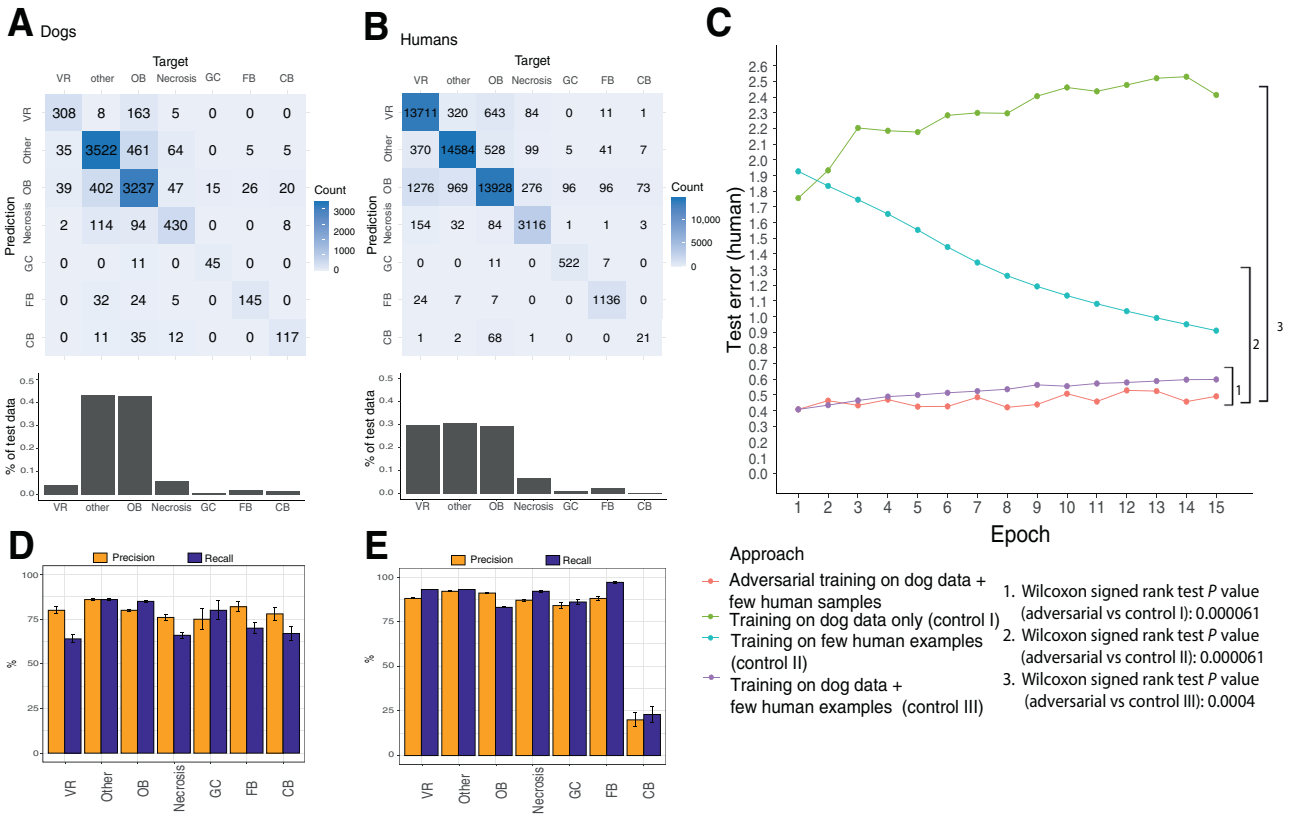

Figure 4 Performance evaluation on held-out whole slide image patches from the test set in both dogs and humans. **A** and **B:** Confusion matrices generated after evaluating model predictions on dog and human whole slide image patches from the held-out test set. The rows represent the predicted class of each whole slide patch (ie, the class achieving the highest probability based on the classification model). The columns represent the ground truth (ie, 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:** The evolution of the test error achieved by the classification model on human whole slide image patches (target domain) while progressing 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 multiclass 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. The error bars were approximately determined by a bootstrap analysis, where the test data sets were repeatedly down-sampled to 50% original size and the precision and recall on each down-sampled version was recomputed. CB, chondroblastic; FB, fibroblastic; GC, giant cell rich; OB, osteoblastic; VR, vessel rich.

and hyperparameters. Overall, the domain adversarial learning approach achieved significantly lower test error per epoch compared with 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, spatial probability maps were generated. They depicted regions of high versus 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 manner (see *Materials and Methods* for details). As a qualitative validation, Figures 2 and 3 depict pathologist-marked region boundaries within four dog and two 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, the study next estimated the absolute burden of each subtype in each canine sample and applied the K-means clustering algorithm to identify clusters of dogs with similar whole

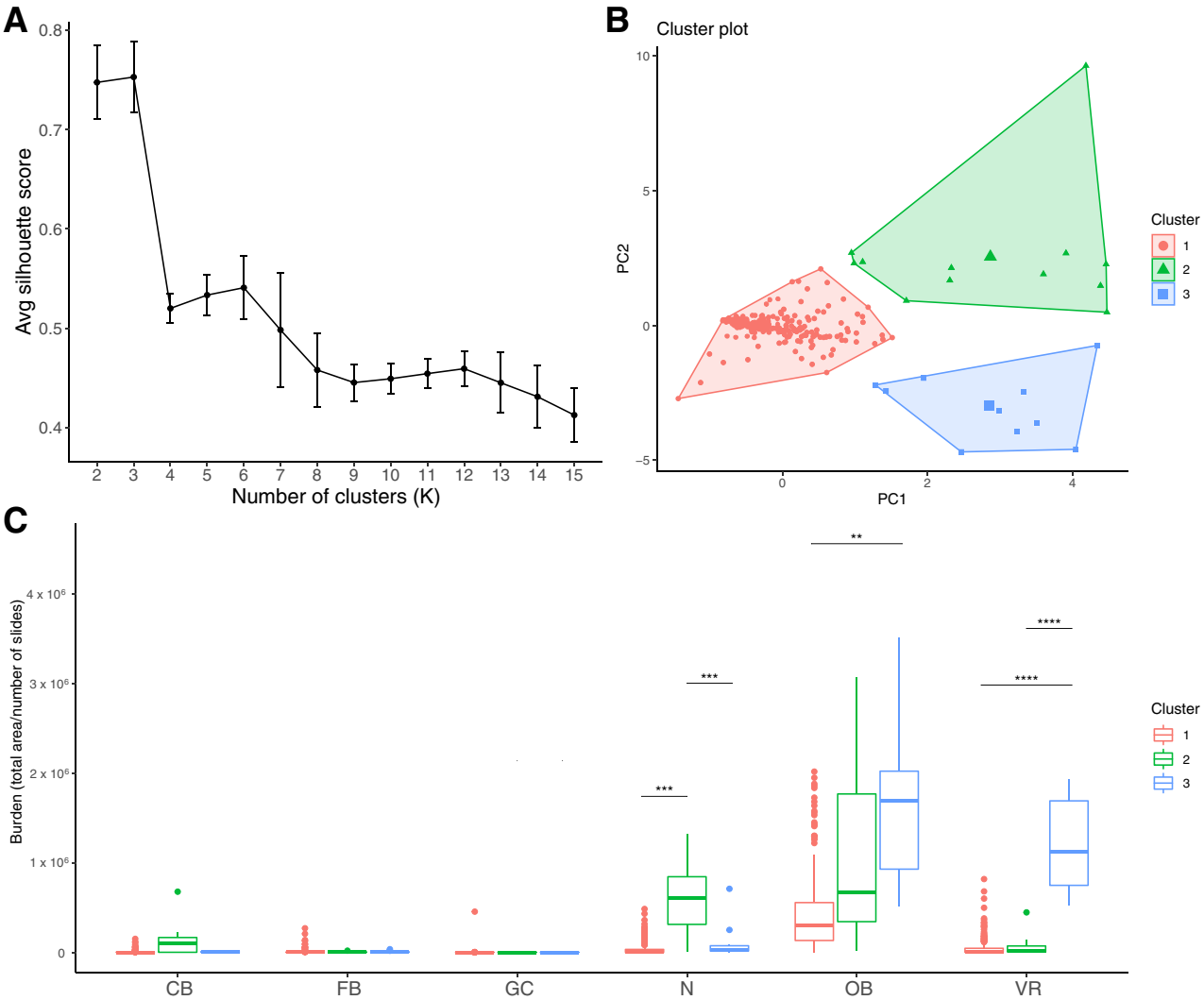

Figure 5 K-means clustering analysis of 306 canine osteosarcoma tumors based on estimated burden of histologic subtypes. **A:** The average silhouette score as a function of the number of clusters used by the K-means algorithm to cluster the data. The higher the average silhouette score, the better the clustering. The smallest value of K achieving the highest silhouette score represents the best possible clustering of the data. **B:** Principal component (PC) 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 red, points belonging to cluster 2 are green, and points belonging to cluster 3 are blue. **C:** Distribution of the burden of each histologic subtype in each cluster. $^{**}P < 0.01$, $^{***}P < 0.001$, and $^{****}P < 0.0001$ (*U*-test). Avg, average; CB, chondroblastic; FB, fibroblastic; GC, giant cell rich; N, necrosis; OB, osteoblastic; VR, vessel rich.

slide tumor histology (Supplemental Table S2) (see *Materials and Methods*). Figure 5A indicates the average silhouette score of inferred clusters for different values of K.[30] The higher the average silhouette score, the more compact and well separated were the clusters (maximum score = 1). The error bars indicate the CI estimated by repeatedly performing K-means clustering on randomly down-sampled versions of the original cohort (down-sampling to approximately 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.

The distribution of the estimated burden of each subtype in each cluster and the clinical outcomes were examined next. The clinical characteristics of the cases analyzed in this study are provided in Table 1. See Supplemental Table S5 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, dogs belonging to cluster 3 had significantly worse clinical outcomes compared with 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 $P = 0.038$), 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 $P = 0.0071$). 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 versus non-proximal humerus), alkaline phosphatase levels (elevated versus normal), age, weight, sex, and adjuvant treatment type in a multivariable Cox proportional hazards regression model.

Finally, subgroup analysis was performed to ensure that prognostic signatures remained significant in unlabeled data not used in training. The first subgroup consisted of 55 reviewed cases ($n = 95$ pathologist-annotated slides). The second subgroup consisted of the remaining 251 unreviewed cases. In each subgroup, the survival association remained consistent, thus demonstrating the clinical utility of model predictions beyond cases previously annotated by the pathologist (Supplemental Figure S2).

Table 1 Clinical Characteristics of the Dog Osteosarcoma Cohort ($N = 306$)

Clinical characteristics	Value
Age, years	8.1 (1.4–15.6)
Weight, kg	38.8 (21.2–94.5)
Tumor location	
Proximal humerus	64 (21)
Non-proximal humerus	242 (79)
ALP levels	
Elevated	74 (24)
Normal	232 (76)
Sex	
Castrated male	171 (56)
Intact male	13 (4)
Spayed female	118 (39)
Intact female	4 (1)
Disease-free interval, time from surgery, days	157 (3–1127)
Overall survival, time from surgery, days	235 (3–1652)
Treatment	
Standard of care	155 (51)
Standard of care + sirolimus (rapamycin)	151 (49)

For continuous variables, values in parentheses represent the minimum and maximum range, and values outside the parentheses represent the median over the entire cohort. All other data are given as number (percentage).

ALP, alkaline phosphatase.

Discussion

Through the activities of the National Cancer Institute COTC, this study examined the largest data set of canine osteosarcomas to date for which complete clinical outcome data were available and standardized therapy was applied ($n = 306$). This large resource was used to 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 small fraction of human data for training. Although this is not the first application of deep learning in osteosarcomas,[31–33] to the best of our knowledge, it is the first attempt to identify histologic features of osteosarcoma that transfer from canine to human samples.

The trained species-agnostic histologic subtype classifier was used to perform an unsupervised exploratory analysis of whole slide imaging data of 306 dogs and identify distinct clusters that respond differently to standardized chemotherapy based on the classifier-estimated burden of histologic subtypes. These results are consistent with 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 because of the difficulty in accumulating a large enough data set 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,

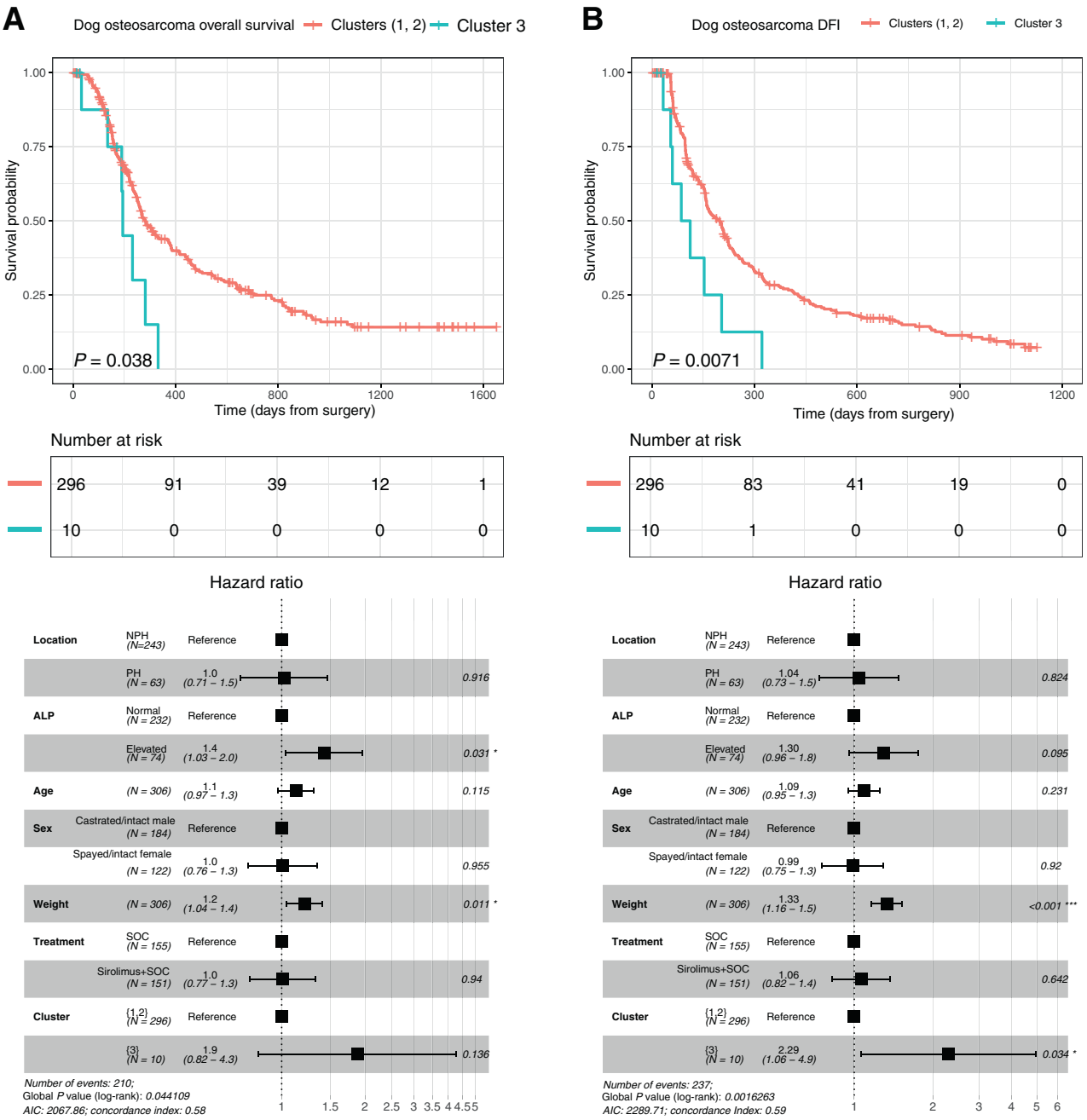

Figure 6 Survival outcomes of cluster 3 versus clusters 1 and 2. **A: Top:** Kaplan-Meier plot depicting the overall survival rates of cases belonging to cluster 3 versus rest (cluster 1 or cluster 2). **Bottom:** Estimated hazard ratio of each factor. **B: Top:** Kaplan-Meier plot depicting the disease-free survival rates of cases belonging to cluster 3 versus rest (cluster 1 or cluster 2). **Bottom:** Estimated hazard ratio of each factor. The log-rank test *P* value was estimated to determine the significance of the differences in survival rates. $^{*}P < 0.05$, $^{***}P < 0.001$ (log-rank test). AIC, Akaike information criterion; ALP, alkaline phosphatase; DFI, disease-free interval; NPH, non-proximal humerus; PH, proximal humerus; SOC, standard of care.

information gleaned from our approach is of substantial clinical value to clinicians treating dogs with OS.

This study refrains from quantifying overlap between pathologist annotations and AI predictions using Dice or Intersection over Union (IoU) metrics. These metrics are preferable in segmentation applications, where the ground truth segmentation boundaries are precisely defined.[34] However, because of intratumor heterogeneity, osteoblastic tumor cells are frequently observed intermixed with other histologic subtypes.[3,5,24] 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 most 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 (Supplemental 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 disease-free interval and overall survival. 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.[3–5] Although an early study[35] suggested that telangiectatic OS carries a poor prognosis in human patients, others suggest that although there may be a correlation with clinical features, such as pathologic fracture, an association with prognosis is less clear.[36] In dogs, the telangiectatic subtype has been associated with poor prognosis in studies of OS originating in the ulna[37] ($n = 30$) or flat and irregular bones[38] ($n = 45$). In our case set, we defined VR regions as containing blood-filled spaces lined by tumor cells. On hematoxylin and eosin staining, 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 staining confirmed the presence of blood-filled spaces lined by tumor cells in VR-annotated canine osteosarcomas (Supplemental Figure S2). Some VR regions also contained cellular debris, which has been described in human OS.[39–41] Although VR morphology was uncommon in our data set, 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 (ie, telangiectatic or osteoblastic-telangiectatic[37]). In fact, up to 65% of canine osteosarcomas are reported to demonstrate multiple histologic subtypes.[13] This underlines 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, there was no 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 method 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 (intraclass correlation coefficient = 0.626), whereas agreement on grade was fair using two different classification systems.[42] In the future, we aim to convene a comparative pathology board of M.D. and D.V.M. pathologists to review canine and human osteosarcoma histology with the goal of assessing the impact of our model on interobserver variability, and identifying additional features, such as immune cell infiltration, that may be incorporated into our prognostic model alongside ongoing genomic work. Third, the data are 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 nonoverlapping patches scanned at high magnification and split at random into train validation and test sets (see *Materials and 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 because of thousands of image patches that can be extracted during training at high magnifications.[22,43] However, neural networks trained this way are prone to overfitting to slide, staining, or scanner-specific properties.[44] In this work, an adversarial learning approach was used to help neural networks overcome the bias that present in domain-specific training paradigms. Adversarial training can, however, be complex in practice compared with 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.[25] 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 *Materials and Methods*. Lastly, no additional manual quality control of surgical tumor specimens was completed before data collection from different sites. Instead, our model was adversarially trained to classify nontumor regions in addition to the six different histologic subtypes of osteosarcoma based on pathologist annotations. The robustness and accuracy of the classification model is expected to improve as additional data are 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 are scarce.

Acknowledgments

We thank the Comparative Oncology Clinical Trials Consortium (COTC) members for execution of the COTC-21/022 trials, which provided the clinical outcome data that were analyzed herein; and Dr. Markku Miettinen for granting access to 39 human osteosarcoma slides from his residency training materials.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.ajpath.2022.09.009*.

References

1. Ottaviani G, Jaffe N: The epidemiology of osteosarcoma. Cancer Treat Res 2009, 152:3−13
2. Misaghi A, Goldin A, Awad M, Kulidjian AA: Osteosarcoma: a comprehensive review. SICOT J 2018, 4:12
3. Beck J, Ren L, Huang S, Berger E, Bardales K, Mannheimer J, Mazcko C, LeBlanc A: Canine and murine models of osteosarcoma. Vet Pathol 2022, 59:399−414
4. Maxie GG: Jubb, Kennedy & Palmer's Pathology of Domestic Animals, vol 2. Amsterdam, the Netherlands: Elsevier Health Sciences, 2015
5. Meuten DJ: Tumors in Domestic Animals. Hoboken, NJ: John Wiley & Sons, 2020
6. Gorlick R, Meyers PA: Osteosarcoma necrosis following chemotherapy: innate biology versus treatment-specific. J Pediatr Hematol Oncol 2003, 25:840−841
7. LeBlanc AK, Breen M, Choyke P, Dewhirst M, Fan TM, Gustafson DL, Helman LJ, Kastan MB, Knapp DW, Levin WJ, London C, Mason N, Mazcko C, Olson PN, Page R, Teicher BA, Thamm DH, Trent JM, Vail DM, Khanna C: Perspectives from man's best friend: National Academy of Medicine's Workshop on Comparative Oncology. Sci Transl Med 2016, 8:324ps5
8. LeBlanc AK, Mazcko CN: Improving human cancer therapy through the evaluation of pet dogs. Nat Rev Cancer 2020, 20:727−742
9. LeBlanc AK, Mazcko CN, Khanna C: Defining the value of a comparative approach to cancer drug development. Clin Cancer Res 2016, 22:2133−2138
10. Selmic LE, Burton JH, Thamm DH, Withrow SJ, Lana SE: Comparison of carboplatin and doxorubicin-based chemotherapy protocols in 470 dogs after amputation for treatment of appendicular osteosarcoma. J Vet Intern Med 2014, 28:554−563
11. Al-Khan AA, Nimmo JS, Day MJ, Tayebi M, Ryan SD, Kuntz CA, Simcock JO, Tarzi R, Saad ES, Richardson SJ, Danks JA: Fibroblastic subtype has a favourable prognosis in appendicular osteosarcoma of dogs. J Comp Pathol 2020, 176:133−144
12. LeBlanc AK, Mazcko CN, Cherukuri A, Berger EP, Kisseberth WC, Brown ME, et al: Adjuvant sirolimus does not improve outcome in pet dogs receiving standard-of-care therapy for appendicular osteosarcoma: a prospective, randomized trial of 324 dogs. Clin Cancer Res 2021, 27:3005−3016
13. Nagamine E, Hirayama K, Matsuda K, Okamoto M, Ohmachi T, Kadosawa T, Taniyama H: Diversity of histologic patterns and expression of cytoskeletal proteins in canine skeletal osteosarcoma. Vet Pathol 2015, 52:977−984
14. 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 canine appendicular osteosarcoma: a randomized, phase III trial. Vet Comp Oncol 2016, 14:81−87
15. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A: Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019, 16:703−715
16. Harmon SA, Sanford TH, Brown GT, Yang C, Mehralivand S, Jacob JM, Valera VA, Shih JH, Agarwal PK, Choyke PL, Turkbey B: Multiresolution application of artificial intelligence in digital pathology for prediction of positive lymph nodes from primary tumors in bladder cancer. JCO Clin Cancer Inform 2020, 4:367−382
17. Harmon SA, Tuncer S, Sanford T, Choyke PL, Turkbey B: Artificial intelligence at the intersection of pathology and radiology in prostate cancer. Diagn Interv Radiol 2019, 25:183−188
18. Ferrer L, Fondevila D, Rabanal RM, Vilafranca M: Immunohistochemical detection of CD31 antigen in normal and neoplastic canine endothelial cells. J Comp Pathol 1995, 112:319−326
19. Giuffrida MA, Bacon NJ, Kamstock DA: Use of routine histopathology and factor VIII-related antigen/von Willebrand factor immunohistochemistry to differentiate primary hemangiosarcoma of bone from telangiectatic osteosarcoma in 54 dogs. Vet Comp Oncol 2017, 15:1232−1239
20. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019, 25:1301−1309
21. D'Amato M, Szostak P, Torben-Nielsen B: A comparison between single- and multi-scale approaches for classification of histopathology images. Front Public Health 2022, 10:892658
22. Kuklyte J, Fitzgerald J, Nelissen S, Wei H, Whelan A, Power A, Ahmad A, Miarka M, Gregson M, Maxwell M, Raji R, Lenihan J, Finn-Moloney E, Rafferty M, Cary M, Barale-Thomas E, O'Shea D: Evaluation of the use of single- and multi-magnification convolutional neural networks for the determination and quantitation of lesions in nonclinical pathology studies. Toxicol Pathol 2021, 49:815−842
23. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F: Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021, 5:555−570
24. Dahlin DC: Pathology of osteosarcoma. Clin Orthop Relat Res 1975: 23−32
25. Ganin Y, Lempitsky V: Unsupervised domain adaptation by backpropagation International Conference on Machine Learning, 37; 2015. pp. 1180−1189
26. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF: ImageNet: a large-scale hierarchical image database Cvpr: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 1-4; 2009. pp. 248−255
27. Ahmed S, Shaikh A, Alshahrani H, Alghamdi A, Alrizq M, Baber J, Bakhtyar M: Transfer learning approach for classification of histopathology whole slide images. Sensors 2021, 21:5361
28. Sharmay Y, Ehsany L, Syed S, Brown DE: HistoTransfer: understanding transfer learning for histopathology. Edited by 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2021. pp. 1−4
29. Loshchilov I, Hutter F: SGDR: stochastic gradient descent with warm restarts. arXiv 2016 [Preprint], doi:10.48550/arXiv.1608.03983
30. Rousseeuw PJ: Silhouettes - a graphical aid to the interpretation and validation of cluster-analysis. J Comput Appl Math 1987, 20:53−65
31. Mishra R, Daescu O, Leavey P, Rakheja D, Sengupta A: Convolutional neural network for histopathological analysis of osteosarcoma. J Comput Biol 2018, 25:313−325
32. D'Acunto M, Martinelli M, Moroni D: Deep learning approach to human osteosarcoma cell detection and classification Multimedia and Network Information Systems, 833; 2019. pp. 353−361
33. Fu Y, Xue P, Ji HZ, Cui WT, Dong EQ: Deep model with Siamese network for viable and necrotic tumor regions assessment in osteosarcoma. Med Phys 2020, 47:4895−4905
34. Taha AA, Hanbury A: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015, 15:29

35. Matsuno T, Unni KK, McLeod RA, Dahlin DC: Telangiectatic osteogenic sarcoma. Cancer 1976, 38:2538–2547
36. Huvos AG, Rosen G, Bretsky SS, Butler A: Telangiectatic osteogenic sarcoma: a clinicopathologic study of 124 patients. Cancer 1982, 49: 1679–1689
37. Sivacolundhu RK, Runge JJ, Donovan TA, Barber LG, Saba CF, Clifford CA, de Lorimier LP, Atwater SW, DiBernardi L, Freeman KP, Bergman PJ: Ulnar osteosarcoma in dogs: 30 cases (1992-2008). J Am Vet Med Assoc 2013, 243:96–101
38. Hammer AS, Weeren FR, Weisbrode SE, Padgett SL: Prognostic factors in dogs with osteosarcomas of the flat or irregular bones. J Am Anim Hosp Assoc 1995, 31:321–326
39. Bacci G, Ferrari S, Ruggieri P, Biagini R, Fabbri N, Campanacci L, Bacchini P, Longhi A, Forni C, Bertoni F: Telangiectatic osteosarcoma of the extremity: neoadjuvant chemotherapy in 24 cases. Acta Orthop Scand 2001, 72:167–172
40. Liu JJ, Liu S, Wang JG, Zhu W, Hua YQ, Sun W, Cai ZD: Telangiectatic osteosarcoma: a review of literature. Onco Targets Ther 2013, 6:593–602
41. Sangle NA, Layfield LJ: Telangiectatic osteosarcoma. Arch Pathol Lab Med 2012, 136:572–576
42. Schott CR, Tatiersky LJ, Foster RA, Wood GA: Histologic grade does not predict outcome in dogs with appendicular osteosarcoma receiving the standard of care. Vet Pathol 2018, 55:202–211
43. Salvi M, Acharya UR, Molinari F, Meiburger KM: The impact of pre- and post-image processing techniques on deep learning frameworks: a comprehensive review for digital pathology image analysis. Comput Biol Med 2021, 128:104129
44. Schomig-Markiefka B, Pryalukhin A, Hulla W, Bychkov A, Fukuoka J, Madabhushi A, Achter V, Nieroda L, Buttner R, Quaas A, Tolkach Y: Quality control stress test for deep learning-based diagnostic model in digital pathology. Mod Pathol 2021, 34:2098–2108