Received: 29 June 2023 | Revised: 31 August 2023 | Accepted: 20 September 2023

DOI: 10.1002/jor.25693

Journal of Orthopaedic Research®

RESEARCH ARTICLE

The use of weakly supervised machine learning for necrosis assessment in patients with osteosarcoma: A pilot study

Christa L. LiBrizzi[1] | Zhenzhen Wang[2] | Jeremias Sulam[2] | Aaron W. James[3] | Adam S. Levin[1] | Carol D. Morris[4]

[1]Department of Orthopaedic Surgery, Division of Orthopaedic Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[2]Department of Biomedical Engineering, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[3]Department of Pathology, The Johns Hopkins University School of Engineering, Baltimore, Maryland, USA

[4]Orthopaedic Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York, USA

Correspondence

Carol D. Morris, Orthopaedic Surgery, Memorial Sloan Kettering Cancer-Center, 1275 York Ave., New York, NY 10065, USA.
Email: morrisc@mskcc.org

Abstract

Percent necrosis (PN) following chemotherapy is a prognostic factor for survival in osteosarcoma. Pathologists estimate PN by calculating tumor viability over an average of whole-slide images (WSIs). This non-standardized, labor-intensive process requires specialized training and has high interobserver variability. Therefore, we aimed to develop a machine-learning model capable of calculating PN in osteosarcoma with similar accuracy to that of a musculoskeletal pathologist. In this proof-of-concept study, we retrospectively obtained six WSIs from two patients with conventional osteosarcomas. A weakly supervised learning model was trained by using coarse and incomplete annotations of viable tumor, necrotic tumor, and nontumor tissue in WSIs. Weakly supervised learning refers to processes capable of creating predictive models on the basis of partially and imprecisely annotated data. Once "trained," the model segmented areas of tissue and determined PN of the same six WSIs. To assess model fidelity, the pathologist also estimated PN of each WSI, and we compared the estimates using Pearson's correlation and mean absolute error (MAE). MAE was 15% over the six samples, and 6.4% when an outlier was removed, for which the model inaccurately labeled cartilaginous tissue. The model and pathologist estimates were strongly, positively correlated ($r = 0.85$). Thus, we created and trained a weakly supervised machine learning model to segment viable tumor, necrotic tumor, and nontumor and to calculate PN with accuracy similar to that of a musculoskeletal pathologist. We expect improvement can be achieved by annotating cartilaginous and other mesenchymal tissue for better representation of the histological heterogeneity in osteosarcoma.

KEYWORDS

bone, cancer/tumors, clinical

1 | INTRODUCTION

The 5-year survival rate for patients with localized extremity osteosarcoma is 65%–70%, with cancer stage being the most important prognostic factor.[1] It is widely accepted that percent necrosis (PN) after neoadjuvant chemotherapy is a powerful prognostic factor.[2] The "gold standard" for estimation of PN is assessment of whole-slide images (WSIs) by a musculoskeletal pathologist to estimate average tumor viability.[3] WSIs are created from the resection specimen, often after two cycles of neoadjuvant

Christa L. LiBrizzi and Zhenzhen Wang contributed equally to this work.

chemotherapy. The histologic PN assessment method originally described by Huvos reflects the efficacy of systemic chemotherapy.[3] In general, patients with less than 90% necrosis have worse survival than those with 90% or more.[4,5] Further, response status can affect patient eligibility for enrollment in clinical trials. Therefore, establishing an accurate PN can carry important clinical implications.

Determination of PN is labor intensive and nonuniform, with studies reporting low interobserver reliability despite it being the standard of care.[6–8] The analysis itself requires the resection specimen to be grossly cut, mapped into segments, then sliced and processed into WSIs. Several WSIs are selected and subsequently averaged over the entire specimen to estimate PN, although the number of WSIs averaged is not standardized among institutions or pathologists. Fellowship-trained musculoskeletal pathologists, most of whom work at tertiary cancer centers, typically make PN estimates. Centers that are not staffed with musculoskeletal pathologists may lack standardized protocols for analysis, and osteosarcoma specimens obtained at such centers may require external review to estimate PN. The described method is used to assess all treated cases of osteosarcoma annually in the United States and is also used for other sarcomas, such as Ewing sarcoma and primitive neuroectodermal tumors.[9,10] Based on these observations and others, calculating PN by alternative means is a worthy exercise.

In this study, we investigated if accurate PN could be calculated using a modification of a previously described machine learning technique. Machine learning has been used successfully in histopathology for tumor detection, subtyping, and grading in several cancers.[11–13] Previous research using learning models for detection of various carcinomas determined that pathologists could exclude 65%–75% of pathology slides without any loss of diagnostic sensitivity when using a validated model, effectively reducing pathologist workload.[11] Recent studies have used supervised interactive learning to estimate PN in osteosarcoma.[14,15] These techniques require hundreds of WSIs to be partially but accurately annotated by pathologists for training, which enables the model to produce initial segmentation maps. Pathologists then assess the predicted segmentations and reannotate any misclassified areas. This process is repeated until the predictions are correct everywhere. Weak supervision covers a wide range of methods in which models are trained using partial, inexact, or otherwise inaccurate information. Wang et al.[16] developed a weakly supervised learning technique known as label cleaning multiple instance learning (LC-MIL), in which only coarse annotations of WSIs are required. Coarse annotations are loosely drawn boundaries at zero magnification of different types of cells within a WSI that is not exhaustively detailed. These coarse annotations are easy to obtain because it takes only minutes for a pathologist to annotate one slide, in stark contrast to annotations needed in traditional, fully supervised learning, or the described interactive learning. This validated method does not require a large data set, but rather can train a model with a single WSI. It leveraged the MIL technique to significantly augment training data by constructing MIL bags, as detailed in the methods section.

We aimed to develop a machine-learning model capable of (1) segmenting viable tumor, necrotic tumor, and nontumor tissue with minimal training data and, (2) calculating PN from osteosarcoma WSIs. We hypothesized that the proposed method in this proof-of-concept study would achieve similar accuracy and precision to the "gold standard" of pathologist estimates, while reducing the labor and interobserver variability associated with PN estimates on a limited series of WSIs.

2 | METHODS

After institutional review board approval, we retrospectively analyzed data from patients with high-grade intramedullary osteosarcomas who received neoadjuvant chemotherapy and underwent surgical resection at our US tertiary cancer center between 2011 and 2021. Patients were included who satisfied the above criteria and who had available resection specimen WSIs. We excluded WSIs assessing tumor margins or those obtained from biopsies. We extracted original PN estimates from patient records.

Our institution uses the method described by Huvos et al.[3] for specimen preparation and PN assessment of osteosarcoma specimens. Resection specimens were processed in a standard fashion. A photograph is taken of the specimen on its long axis, and a grid pattern is overlaid, denoting subdivisions of the specimen (Figure 1). WSIs are sectioned parallel and perpendicular to the long axis to the subdivision and stained with hematoxylin and eosin. PN is estimated from the hematoxylin and eosin WSIs at the time of resection. All WSIs and remaining resection specimens are stored in our institution's pathology archive.

We gathered the WSIs from our pathology archives and removed patient identifiers before digitization. WSIs were digitally scanned at ×40 magnification and uploaded to Proscia pathology software, version 3.8.1:6bc005c (Proscia Inc.) for digital annotation.

2.1 | WSI annotation

A senior, fellowship-trained musculoskeletal pathologist with 10 years of experience (initials removed for blinding) performed all annotations for training the machine model. The pathologist was blinded to all clinical information, including osteosarcoma subtype and original PN estimate. The pathologist coarsely and approximately annotated three types of tissue on each WSI: viable tumor, necrotic tumor, and nontumor. Annotations are created by digitally drawing boundaries on the digitized WSI in Proscia. We defined coarse annotations as approximate boundaries around populations of viable tumor, necrotic tumor, and nontumor. Thus, there is no detailed annotation of individual cells. The absence of annotation on the slide represented mixtures of patterns. WSIs were not exhaustively annotated. Instead, only partial representative regions with homotypic pattern were delineated, and the hand-drawn boundaries were

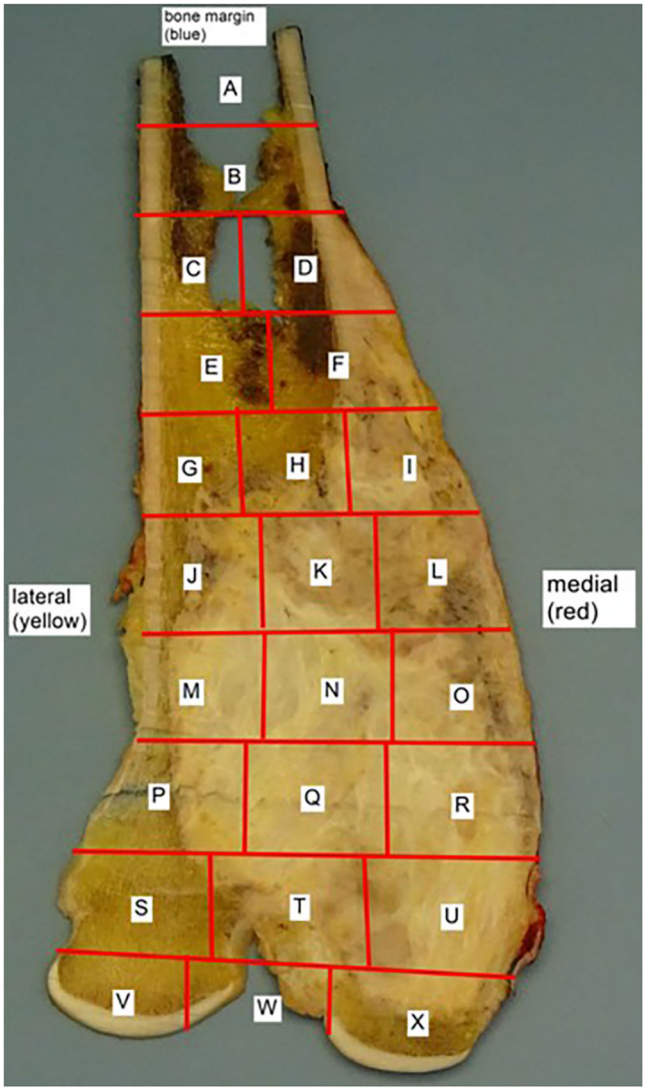

FIGURE 1 Clinical example of the grid preparation for gross specimens on a resected distal femur with osteosarcoma. A grid pattern was overlaid on the photograph of the distal femur and subsequently labeled with sequential alphabet-based annotations.

approximate. The annotation time for the pathologist was approximately 2 min per slide.

2.2 | Development of a deep learning predictor based on weakly annotated data

We developed a weakly supervised learning method that is able to use coarse annotations; does not require exhaustive, cellular-level annotation; and needs a limited amount of training data for label prediction. We adapted the method recently proposed by Wang et al.[16] that can provide accurate segmentations based on only partial and coarse annotations of tissue in WSIs, and we extended it to a multiclass setting to predict different types of tissue (i.e., normal tissue, necrotic tumor, and variable tumor) per WSI. The high-resolution WSIs were segregated into tens of thousands of small patches. The patches are 256 × 256 pixels of similar appearing tissue as determined by the pathologist's annotations. The patches with the same initially annotated label are processed into groups called "bags." Each bag consists of 10 patches. The deep learning machine is trained on the basis of these bags. This LC-MIL approach relies on learning image patterns in groups of patches, or "bags," rather than individual samples, providing a framework that is robust to potential inaccuracies in the labeling of individual image patches and relies on the label of the majority of samples in a given bag. This type of learning paradigm is especially useful in pathology because it is similar to the way in which pathologists manually annotate WSIs, with approximate assessments that are not dependent on each individual cell within the WSI. Additionally, the bags are constructed by sampling patches with replacement, such that virtually as many bags as desired can be generated from one single bag, which significantly augments the training data.

A model similar to that developed in Wang et al.[16] was thus trained but, in this case, for a multiclass classification problem. Within this multiple-instance learning framework, the model is trained to predict the class of the majority of the patches in the bag, but it is invariant to the number of patches in the bag. In this way, once trained, the model is deployed on all patches in each WSI, thus making a prediction of the type of tissue (necrotic tumor, viable tumor, or nontumor) at every location in the slide, which produces a segmentation map for downstreaming PN calculation (Figure 2).

2.3 | PN estimates

PN of each WSI was estimated by a senior pathologist (initials removed for blinding). For the machine learning model, PN was estimated from the tissue segmentation map by calculating the ratio of area of necrotic tumor to area of all tumor (viable tumor and necrotic tumor).

2.4 | Hardware, software, and statistical analysis

We used graphics processing units–intensive computing to train deep learning models with a single NVIDIA RTX 1080 Ti graphics processing unit (Nvidia Corp.). Deep learning models were implemented in PyTorch (Linux Foundation), and other postprocessing algorithms were implemented with OpenCV[17] and Scikit-Image.[18] All software implementations of the methods are publicly available at https://github.com/Sulam-Group/MIL-pathology.

We compared PN estimates of the machine learning model and the pathologist by using Pearson's correlation and mean absolute error (MAE). We used, Stata, version 17.0, software (StataCorp LLC) for all analyses.

3 | RESULTS

We included six WSIs from two patients for this proof-of-concept study. The first patient had a high-grade osteoblastic osteosarcoma of the proximal tibia with PN of 99%. The second patient had a

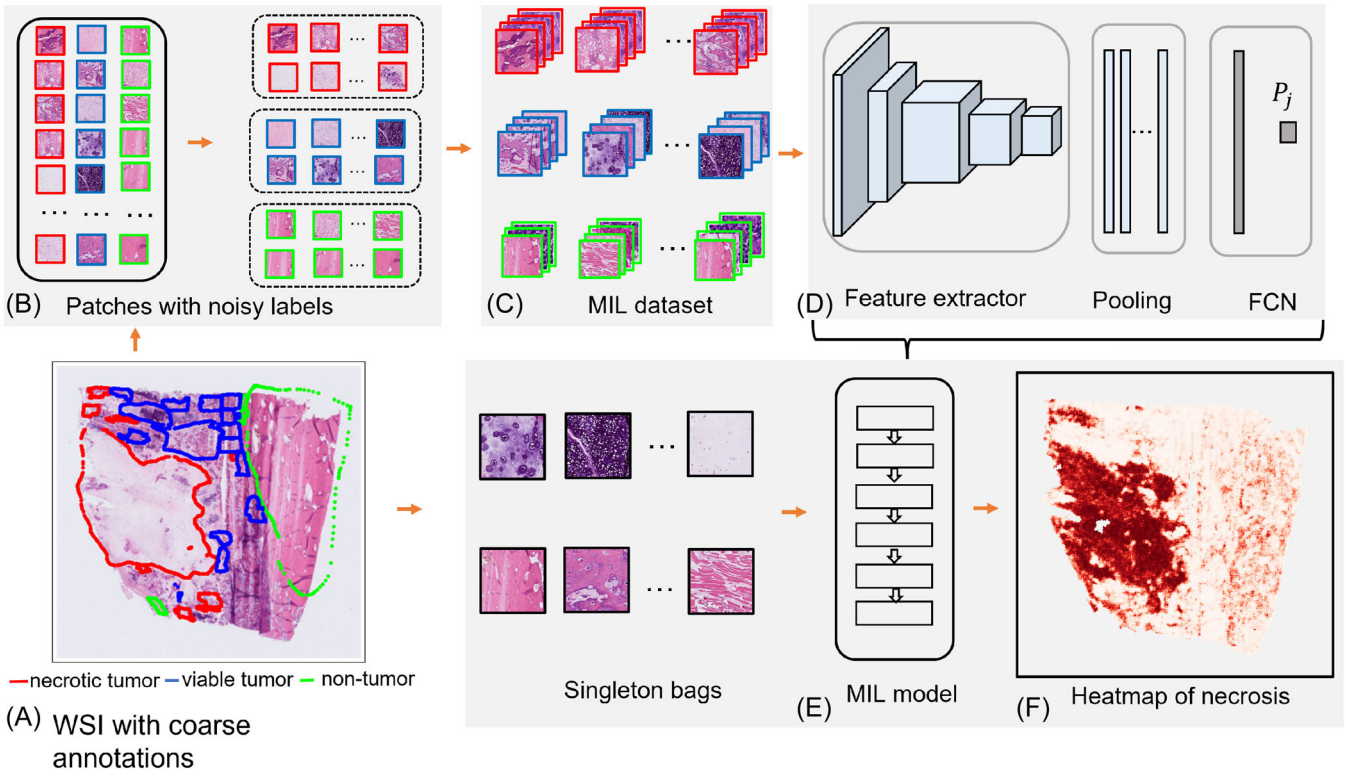

FIGURE 2 This framework uses weakly supervised techniques and multiple instance learning (MIL) methods to train a deep learning machine for assessing percent necrosis (PN) from osteosarcoma whole-slide images (WSIs). The process involves: (A) taking a WSI with coarse annotations by pathologists as input; (B) segregating the WSI into small patches with noisy labels based on the initial annotations; (C) grouping patches with the same labels into MIL bags for training; (D) training the MIL model—composed of a feature extractor, MIL pooling operator, and multilayer perceptron (MLP)—on the MIL bags; (E) using the trained MIL model to clean noisy labels and generate a segmentation map; and (F) estimating PN with the segmentation map.

TABLE 1 Percent necrosis estimates by a musculoskeletal pathologist and a machine learning model in patients with osteosarcoma.

	%	
Slide no	**Pathologist's estimate**	**Model's estimate**
1	45	58
2	20	34
3	60	60
4	5	62
5	97	94
6	99	97
Mean absolute error	15%	
Pearson's correlation	0.85	

high-grade chondroblastic osteosarcoma of the femur with a PN of 50%. Both patients underwent neoadjuvant chemotherapy with methotrexate, doxorubicin, and cisplatin and underwent surgical resection at our institution.

The pathologist's and machine learning model's PN estimates from the six WSI are shown in Table 1. There was a strong, positive correlation between the model and pathologist PN estimations ($r = 0.85$). The model lost accuracy when there was abundant cartilaginous tissue, especially noted in the outlier, slide number 4. When this outlier is removed from the data set, the model estimates have a very strong, positive correlation with the pathologist's estimates ($r = 0.99$). The MAE was 15% over the six samples but decreased to 6.4% when the outlier was removed. Figure 3 shows the pathologist's annotations and the model's segmentation maps, where areas of red represent necrosis, blue represents viable tumor, and green represents nontumor. The segmentation map demonstrates areas where our machine model had difficulty determining predictions of cartilaginous tissue. It should be emphasized that we did not specifically annotate benign or malignant cartilaginous tissue to train the model. It took 30 min per WSI to train the model.

4 | DISCUSSION

We developed a machine learning model via weakly supervised learning techniques that can segment viable tumor, necrotic tumor, and nontumor from coarse expert annotations and can calculate PN from osteosarcoma WSIs. Our model's PN estimates were strongly, positively correlated and had an MAE of 15% compared with the

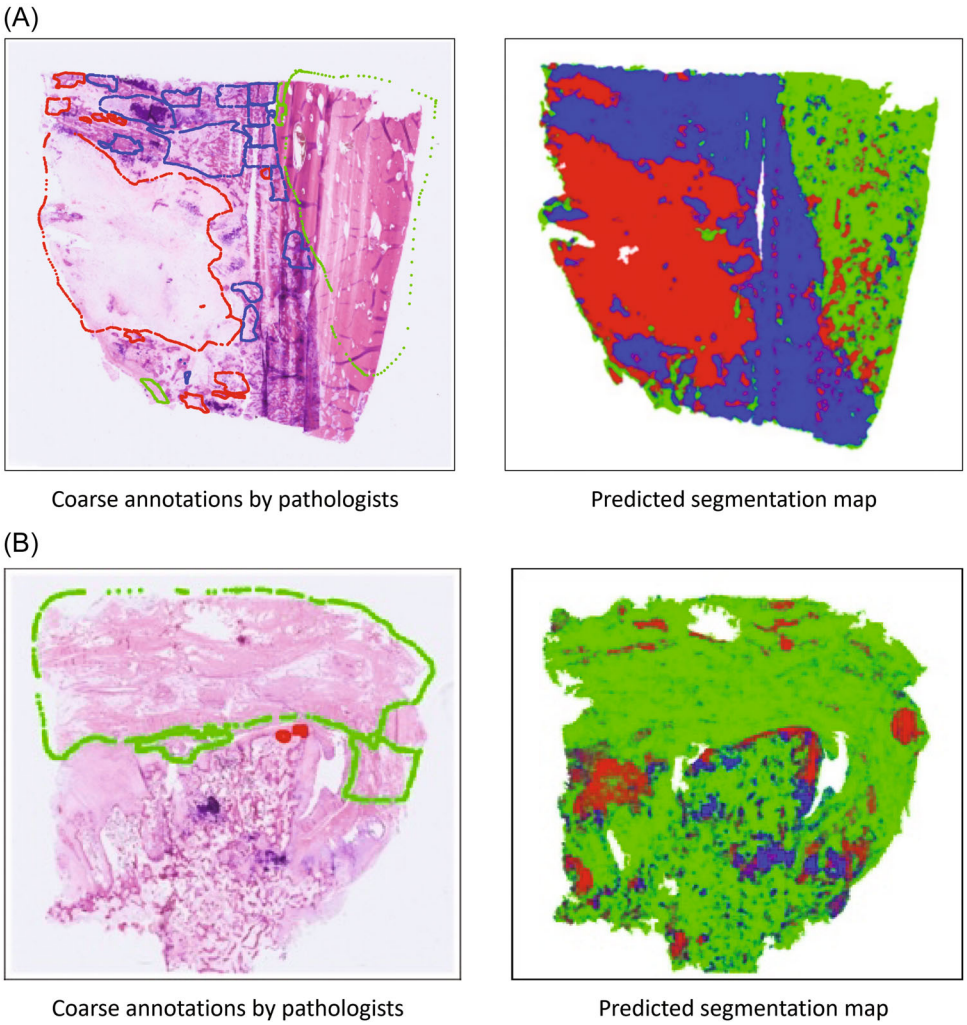

FIGURE 3 Annotation and segmentation of whole-slide images by a musculoskeletal pathologist and machine learning model in (A) osteoblastic and (B) chondroblastic osteosarcoma. Red, necrotic tumor; blue, viable tumor; green, nontumor.

pathologist's estimates. The model did not always properly label cartilage, which affected the accuracy of estimates in WSIs with abundant cartilaginous tissue. Through this proof-of-concept study, we determined that other tissues must be annotated for accurate PN estimates because of the highly heterogenous osteosarcoma histology. Future iterations should include annotations of cartilaginous tissue and other mesenchymal tissue. Next steps include diversifying the WSIs to include other histologic subtypes of osteosarcoma while increasing the number of WSIs required for training and assessment.

4.1 | Limitations

This study has several limitations. First, the machine learning model was trained on a small number of WSIs, as this was a proof-of-concept study only. However, the model correctly identified viable tumor, necrotic tumor, and nontumor with PN estimates similar to those of the pathologist. Second, we annotated only two osteosarcoma subtypes, whereas high-grade osteosarcoma has approximately 10 subtypes with varying degrees of histologic heterogeneity. We chose osteoblastic and chondroblastic subtypes because they are distinct from one another and are the most common subtypes. Finally, our model was trained from the coarse annotations of one pathologist, which may lead to bias in the model's predictions. The pathologist who completed the annotations and PN estimates is a senior attending physician with 10 years of musculoskeletal pathology expertise. Moreover, the method used for this work has been shown to be robust to variations in the coarse annotations provided for training.[16] Further validation with additional senior pathologists will enable precise quantification of bias in the resulting machine model and may lead to improved prediction performance.

4.2 | Segmentation of tissue

Our machine model was trained using LC-MIL techniques, which differ from previously published machine models for predicting

PN.[14,15] LC-MIL techniques allow the machine to learn from patterns of tissue rather than traditional supervised methods, in which the machine learns from discrete entities. We believe the LC-MIL method more closely emulates how pathologists assess WSIs and estimate PN. In contrast with prior research,[14,15] our model did not require exhaustive annotation or reannotation of incorrect labeling by pathologists for training. In fact, our model was trained using six WSIs with coarse annotation compared with traditional methods that require hundreds of WSIs with exhaustive slide-level annotations. Though we used a small number of samples, our machine model's MAE was 15% compared with 23% achieved by Ho et al.[15] in predicting PN in osteosarcoma, which demonstrates similar results with different learning techniques. The deep interactive learning approach described by Ho et al. for osteosarcoma treatment response also struggled with chondroid areas. Other concerns in machine prediction models include false negative and false positives, which may lead to different diagnoses. Campanella et al.[11] devised a weakly supervised machine model to detect various carcinoma metastases to lymph nodes and found that false negatives were frequently cause by micro-metastases, isolated tumor cells, or atypical morphologic features that were not predicted by the model to be malignant. We experienced similar challenges when our model encountered cartilaginous tissue. We did not annotate cartilaginous tissue as a separate entity and it was frequently in areas of mixed tissue, which is unlabeled. Because of the lack of discrete annotation of cartilage, the model did not accurately label chondroblastic tissue, and PN accuracy was lost in WSIs with abundant cartilage. To improve the accuracy and precision of PN estimates, we aim to add separate annotations of cartilaginous, osteoid, and fibroblastic tissues for training of our machine model.

4.3 | PN estimation

Our machine model's PN estimates from osteosarcoma WSIs had a strong, positive correlation to those of a musculoskeletal pathologist. One impetus for this study was the nonstandardized, labor-intensive nature of estimating PN, combined with the high variance of estimates among pathologists. Kang et al.[7] provided 10 WSIs from 10 osteosarcoma patients to six orthopedic pathologists to estimate PN. The interclass correlation coefficient among the six pathologists was 0.65 and did not differ according to years of expertise, suggesting high variability of estimates in the group. However, intraclass correlation was high, ranging from 0.80 to 0.95, suggesting that the pathologists were consistent in their own PN estimates. The large differences in PN estimates among pathologists were often attributable to the misidentification of atypical cells with the stroma as viable or nonviable. In this setting, a machine model would be able to consistently predict these areas of tissue with minimal variance between slides because of the training from similar bags. Thus, use of a validated machine model may solve the current challenge of high interobserver variability in PN estimates.

The theoretical production of a validated machine model would enable expeditious evaluation of chemotherapy efficacy, and hence prognosis estimates, at centers without musculoskeletal pathologists, effectively reducing healthcare costs and labor burdens on subspecialty pathologists. In addition, standardization of PN calculations may increase the fidelity of the data from which national protocols are generated, enabling more accurate risk stratification for clinical trial enrollment. Such initiatives are vital in rare diseases such as osteosarcoma, for which small numbers of patients are treated, and treatments vary among institutions and regions.

5 | CONCLUSION

In this proof-of-concept study, we created and trained a machine learning model via weakly supervised learning techniques that was able to segment viable tumor, necrotic tumor, and nontumor and to calculate PN. Our model's PN estimates were strongly, positively correlated with those of a musculoskeletal pathologist. We found that it will be necessary to annotate cartilaginous and other mesenchymal tissue to better represent the heterogeneity of osteosarcoma and to increase the number of WSIs annotated for training and predictions to improve the accuracy of the model.

AUTHOR CONTRIBUTIONS

Jeremias Sulam, Aaron W. James, and Carol D. Morris conceptualized the study. Christa L. LiBrizzi, Zhenzhen Wang, Jeremias Sulam, Aaron W. James, and Carol D. Morris curated the data. Christa L. LiBrizzi, Zhenzhen Wang, Jeremias Sulam, Aaron W. James, and Carol D. Morris analyzed the data. All authors conducted the research. Jeremias Sulam, Aaron W. James, and Carol D. Morris managed and coordinated the project. All authors provided study resources. Zhenzhen Wang and Jeremias Sulam were involved in software implementation. Jeremias Sulam, Aaron W. James, and Carol D. Morris supervised the project. Christa L. LiBrizzi, Zhenzhen Wang, Jeremias Sulam, Aaron W. James, and Carol D. Morris provided validation of results. All authors were involved in visualization/data presentation, writing the original draft, and revising the manuscript.

ACKNOWLEDGMENTS

For editorial assistance, we thank Rachel Box, MS, in the Editorial Services group of The Johns Hopkins Department of Orthopaedic Surgery.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ETHICS STATEMENT

This study was approved by IRB00323658.

ORCID

Adam S. Levin http://orcid.org/0000-0003-0450-7821

REFERENCES

1. Misaghi A, Goldin A, Awad M, Kulidjian AA. Osteosarcoma: a comprehensive review. *SICOT-J*. 2018;4:12.
2. Testa S, Hu BD, Saadeh NL, et al. A retrospective comparative analysis of outcomes and prognostic factors in adult and pediatric patients with osteosarcoma. *Curr Oncol*. 2021;28(6):5304-5317.
3. Huvos AG, Rosen G, Marcove RC. Primary osteogenic sarcoma: pathologic aspects in 20 patients after treatment with chemotherapy en bloc resection, and prosthetic bone replacement. *Arch Pathol Lab Med*. 1977;101(1):14-18.
4. Glasser DB, Lane JM, Huvos AG, Marcove RC, Rosen G. Survival, prognosis, and therapeutic response in osteogenic sarcoma. The Memorial Hospital experience. *Cancer*. 1992;69(3):698-708.
5. Richardson SM, Wurtz LD, Collier CD. Ninety percent or greater tumor necrosis is associated with survival and social determinants of health in patients with osteosarcoma in the National Cancer Database. *Clin Orthopaedics Related Res*. 2023;481(3):512-522.
6. Huvos AG. *Bone Tumors, Diagnosis, Treatment, and Prognosis*. 1st ed. Saunders; 1979:478.
7. Kang JW, Shin SH, Choi JH, et al. Inter and intra-observer reliability in histologic evaluation of necrosis rate induced by neo-adjuvant chemotherapy for osteosarcoma. *Intl J Clin Exp Pathol*. 2017;10(1):359-367.
8. Kilpatrick SE, Abdul-Karim FW, Renner JB, et al. Interobserver variability among expert orthopedic pathologists for diagnosis, histologic grade, and determination of the necessity for chemotherapy in osteosarcoma. *Pediatr Pathol Mol Med*. 2000;19(5):337-358.
9. Mirabello L, Troisi RJ, Savage SA. Osteosarcoma incidence and survival rates from 1973 to 2004: data from the surveillance, epidemiology, and end results program. *Cancer*. 2009;115(7):1531-1543.
10. Tsokos M, Alaggio RD, Dehner LP, Dickman PS. Ewing sarcoma/peripheral primitive neuroectodermal tumor and related tumors. *Pediatr Dev Pathol*. 2012;15(1 suppl):108-126.
11. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Med*. 2019;25(8):1301-1309.
12. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. *Br J Cancer*. 2021;124(4):686-696.
13. Harder N, Schönmeyer R, Nekolla K, et al. Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma. *Sci Rep*. 2019;9(1):7449.
14. Arunachalam HB, Mishra R, Daescu O, et al. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PLoS One*. 2019;14(4):e0210706.
15. Ho DJ, Agaram NP, Schueffler PJ, et al. Deep interactive learning: an efficient labeling approach for deep learning-based osteosarcoma treatment response assessment. 2020. p 540-549. Accessed May 29, 2023. http://arxiv.org/abs/2007.01383
16. Wang Z, Saoud C, Wangsiricharoen S, James AW, Popel AS, Sulam J. Label cleaning multiple instance learning: refining coarse annotations on single whole-slide images. *IEEE Trans Med Imag*. 2022;41(12):3952-3968.
17. The OpenCV library-ProQuest. Accessed May 30, 2023. https://www.proquest.com/docview/202684726
18. van der Walt S, Schönberger JL, Nunez-Iglesias J, et al. Scikit-image: image processing in Python. *PeerJ*. 2014;2:e453.

How to cite this article: LiBrizzi CL, Wang Z, Sulam J, James AW, Levin AS, Morris CD. The use of weakly supervised machine learning for necrosis assessment in patients with osteosarcoma: a pilot study. *J Orthop Res*. 2023;1-7. doi:10.1002/jor.25693