An-Najah University Journal for Research – A
Natural Sciences

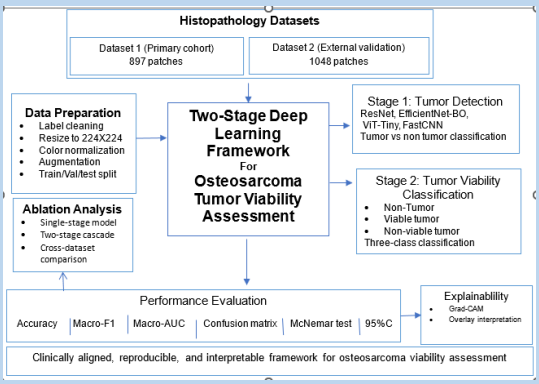
Response-gated computational pathology for viable and non-viable tissue recognition in osteosarcoma

Kapil Joshi[1], Kamal Upreti[2]*, Dyuti Banerjee[3], Vaibhav Sharma[4], Pravin Kshirsagar[5], Durgesh Mishra[6], Rituraj Jain[7]

Type: Full Article. Received 16th June. 2026, Accepted 23th Aug, 2026

Accepted Manuscript, In Press

Abstract: To design and test a patch-based deep learning model to classify H&E images of osteosarcoma into tumor vs. non-tumor and viable vs. non-viable, which is an alternative to the traditional necrosis assessment, for tumor screening and viable tissue stratification for treatment response. Eight hundred ninety-seven cleaned histopathology image patches (primary) and 1,048 public image patches (secondary) were analysed. ResNet18, EfficientNet-B0, ViT-Tiny, and FastCNN were tested to classify tumor/non-tumor. ResNet18 was then used for multiclass viability classification and compared to a two-stage gated cascade. Confidence intervals, McNemar testing, ablation studies, duplicate and case-overlap audits, confusion matrices, ROC analysis, saliency confidence-drop evaluation, and Grad-CAM visualization were all used to evaluate performance. The framework achieved an accuracy of 97.8% and a support-adjusted macro-F1 score of 0.943 on Dataset 1. On Dataset 2, it achieved an accuracy of 96.0%, a macro-F1 score of 0.958, and a macro-AUC of 0.995, demonstrating strong reproducibility and discriminative performance. The proposed framework offers clinically interpretable and reproducible computational pathology support for the treatment-response assessment in osteosarcoma at the patch level. In the future, patient-level validation and whole-slide analysis, as well as prospective expert evaluation, should be performed.



Keywords: computational pathology; digital pathology; deep learning; explainable artificial intelligence; osteosarcoma; treatment response assessment; tumor viability; tumor necrosis

Introduction

Bone cancer image analysis has recently moved away from traditional image interpretation to the use of artificial intelligence-driven decision support. There have been encouraging results obtained using deep learning algorithms for bone cancer diagnosis and classification based on the recognition of patterns associated with the diseases found in medical images [1]. Histopathological analysis is particularly important in osteosarcoma because histopathology images provide useful information on tumor presence, therapy response, and tissue viability [2, 3]. Assessment of tissue response to chemotherapy depends on the ability to distinguish between viable and necrotic tumors [4]; therefore, viability detection is a relevant clinical outcome measure. It has been established that machine and deep learning techniques are helpful tools in determining viable versus necrotic tissue areas in histopathology images of osteosarcoma [5]. However, many previous studies were limited to radiographic modalities [6], including X-rays [7], CT, or MRI scans [8]. The clinical relevance of viability assessment is supported by prognostic evidence showing a strong association between treatment response and patient outcomes in high-grade osteosarcoma [9]. As a result, in current osteosarcoma assessments [10], pathology-based artificial intelligence [11]

[1] Department of Computer Science & Engineering, Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, India
E-mail: kapiljoshi@uumail.in
[2] Department of Computer Science & Engineering, Christ University, Delhi NCR Campus, Ghaziabad, Uttar Pradesh India
E-mail: kamal.upreti@christuniversity.in
[3] Department of Artificial Intelligence and Data Science, Koneru Lakshmaiah Education Foundation, Green Fields, Bowrampet, Hyderabad-500043, Telangana, India
E-mail: dr.dyutibanerjee@klh.edu.in
[4] Department of IT, School of Engineering and Technology, Shri Guru Ram Rai University, Dehradun, Uttarakhand, India
E-mail: vaibhavsharma@sgrru.ac.in
[5] Department of Electronics and Telecommunication, Engineering, J D College of Engineering and Management, Nagpur, India
E-mail: prkshirsagar@jdcoem.ac.in
[6] School of Computer Science and Information Technology, Symbiosis University of Applied Science, Indore, Madhya Pradesh, India
E-mail: director.scsit@suas.ac.in
[7] School of Computer Science Engineering, Chandigarh University Uttar Pradesh, Unnao, U.P., India
E-mail: rituraj.l100957@culko.in
* Corresponding author: kamal.upreti@christuniversity.in, ORCID id (Kamal Upreti): 0000-0003-0665-530X

solutions have been studied more often recently [12]. Pretrained convolutional neural networks and transfer-learning approaches have been widely adopted because they can be effectively adapted to limited datasets [13]. The clinical application of artificial intelligence in medical imaging requires accuracy [14], explainability [15], rigorous validation, and practical applicability [16]. Comparative studies have also emphasized the importance of appropriate benchmarking and evaluation of AI models for bone cancer diagnosis [17, 18]. Furthermore, there have been recent advances in multi-modal diagnostic technologies integrating imaging [19], pathology [20], and biomarkers [21]. Data augmentation and effective training strategies are particularly important because medical imaging datasets are often small and susceptible to overfitting [22]. While promising results have been achieved in the context of multimodal and radiological classification of bone tumors [23], these problems differ significantly from histopathological viable versus non-viable differentiation [24, 25]. At the same time, cascade architectures have also shown promise for efficient staged inference [26]. These developments further demonstrate a trend towards disease-specific artificial intelligence models in pathology [27]. Despite these advances, several important research gaps remain. First, a majority of existing bone tumor-related research tends to rely on radiography instead of H&E-stained sections in osteosarcoma cases. Second, several pathology-based studies do not explicitly evaluate tissue viability within their broader tumor-classification frameworks. Third, direct comparisons between single-stage classifiers and clinically aligned staged workflows are often lacking. Fourth, secondary validation of the data, statistical uncertainty measures, split validation audits, and quantitative explainability are often missing. Therefore, this study aims to bridge these gaps by developing a pathology-focused AI approach to assessing the viability of osteosarcomas. The approach assigns labels at the histopathology-patch level rather than performing tumor localization or patient-level necrosis quantification. This two-stage model first separates the tumor from non-tumor tissues and then analyzes three categories – non-tumor tissue, viable tumor tissue, and non-viable tumor tissue. Rather than assuming superiority of a staged design, the framework is evaluated through ablation analysis, cross-dataset validation, statistical testing, and explainability assessment to determine whether a clinically organized workflow can provide comparable predictive performance while offering improved interpretability.

The objectives of this study are:

1. To develop and evaluate deep-learning models for classifying osteosarcoma H&E image patches into tumor and non-tumor categories using pretrained and lightweight neural architectures.
2. To classify histopathology patches into non-tumor tissue, viable tumor, and non-viable tumor categories without overstating the task as whole-slide necrosis quantification.
3. To compare a conventional single-stage multiclass classifier with a two-stage gated cascade that follows the clinical logic of screening tumor-relevant patches before viability interpretation.
4. To evaluate model reliability using confidence intervals, McNemar testing, confusion matrices, ROC analysis, secondary dataset validation, duplicate audit, and case-overlap audit.
5. To use Grad-CAM and saliency-based confidence-drop analysis to examine whether model decisions are supported by histopathologically meaningful image regions.

The novelty of this work lies in its clinically aligned formulation of osteosarcoma viability assessment and its comprehensive evaluation through cross-dataset validation, ablation studies, statistical analysis, and explainability. Rather than claiming inherent superiority, the proposed two-stage framework is rigorously compared with a conventional single-stage approach to assess its practical and interpretive value.

1. Specific Problem Reformulation: This study reformulates osteosarcoma viability assessment as a patch-level H&E tissue classification task, clearly separating it from whole-slide localization and patient-level necrosis percentage estimation.
2. Clinically Gated Architecture Testing: A two-stage gated cascade is evaluated against a conventional single-stage ResNet18 classifier to test whether staged pathological reasoning provides comparable performance and clearer workflow interpretation.
3. Two-Dataset Reproducibility Layer: The framework is validated on the primary osteosarcoma histopathology dataset and a secondary public osteosarcoma patch dataset, with duplicate and case-overlap audits used to interpret validation strength.
4. Rigorous Statistical Evidence: Performance is supported by Wilson confidence intervals, smoothed bootstrap confidence intervals, McNemar tests, confusion matrices, ROC curves, and support-adjusted reporting for small viable-tumor samples.
5. Explainability Beyond Visualization: Grad-CAM visualization is combined with saliency-based confidence-drop analysis so that interpretability is not limited to qualitative heatmap inspection.

Table 1 summarizes representative studies related to bone cancer and osteosarcoma image analysis. The comparison highlights differences in imaging modality, task definition, validation strategy, and methodological scope, illustrating the research gap addressed by the proposed patch-level viability assessment framework.

Table (1): Comparison of existing osteosarcoma and bone tumor deep learning studies.

Study	Data / Modality	Main Task	Relevance and Remaining Gap
Alabdulkreem et al. [2]	X-ray images	Bone cancer detection and classification	Demonstrates radiology-based deep learning, but does not address H&E patch-level viability classification.
Anisuzzaman et al.[3]	Osteosarcoma histological images	Osteosarcoma detection/classification	Supports histology-based deep learning, but does not provide staged response-oriented validation.
Arunachalam et al. [4]	Osteosarcoma whole-slide histopathology	Viable and necrotic tumor assessment	Closely related to viability assessment, but further statistical ablation and secondary validation are useful.
Aziz et al. [5]	Osteosarcoma image features	Hybrid classification using deep features and MLP	Focuses on classification performance, but not on clinically staged workflow comparison.

Nasir et al. [14]	Osteosarcoma histopathology images	Transfer-learning-based classification	Uses transfer learning, but does not primarily test single-stage versus two-stage viability modeling.
Vezakis et al. [22]	Osteosarcoma diagnosis studies	Comparative methodological review	Provides methodological context, but does not itself provide a two-dataset ablation-tested pipeline.
Borji et al. [8]	Osteosarcoma histopathology	Hybrid deep learning evaluation	Recent histopathology-focused work, but additional explainability and reproducibility testing remain important.
Yao et al. [26]	Osteosarcoma pathology	Multi-model cascaded diagnosis	Supports cascaded reasoning, but statistical testing and secondary dataset validation remain important.
Present study	H&E osteosarcoma image patches	Tumor/non-tumor and viable/non-viable patch classification	Adds two-dataset validation, single-stage versus two-stage ablation, confidence intervals, McNemar testing, and explainability analysis.

Materials and Methods

The proposed approach was designed as a clinically aligned computational pathology framework for osteosarcoma histopathology assessment, with emphasis on chemotherapy-induced tissue necrosis and viability interpretation. The analysis pipeline followed two sequential stages to reflect pathological reasoning. In the first stage, histopathology patches were classified as tumor-containing or non-tumor tissue. In the second stage, patches were classified into non-tumor tissue, viable tumor, and non-viable tumor categories. Transfer learning was used through pretrained neural networks implemented in PyTorch. Model explainability was incorporated to examine whether the predicted classes were supported by histopathologically meaningful image regions. The proposed patch-level approach differs from whole-slide necrosis quantification because it focuses on tissue-patch classification rather than tumor localization, slide-level necrosis estimation, or patient-level treatment-response prediction.

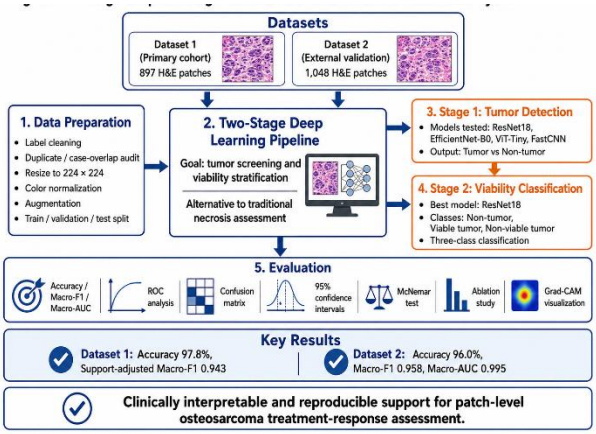


Figure (1): Clinically aligned multi-stage deep learning framework for osteosarcoma histopathological viability assessment.

Figure 1 illustrates the clinically aligned deep learning framework used for osteosarcoma histopathology assessment. The workflow includes image preprocessing, Stage 1 binary tumor detection, Stage 2 multiclass tissue viability classification, statistical validation, and explainability analysis. Pretrained and lightweight models were first compared for tumor/non-tumor classification, after which ResNet18 was used for multiclass viability assessment. Class-imbalance handling and explainability procedures were included to support robust and interpretable patch-level evaluation.

Dataset and Preprocessing

The primary dataset was the publicly available Osteosarcoma Tumor Assessment histopathology dataset containing H&E-stained image patches labeled as non-tumor, viable tumor, and non-viable tumor. After removing ambiguous labels, 897 patches were retained (468 non-tumor, 92 viable tumor, and 337 non-viable tumor). All images were resized to 224 × 224 pixels and normalized using ImageNet statistics. Data augmentation, including random cropping, flipping, rotation, and color jittering, was applied only to the training set. The dataset was divided into training, validation, and test subsets using a stratified patch-level split (70%, 15%, and 15%), and duplicate-hash and case-overlap audits were performed to assess potential data leakage. A secondary public osteosarcoma dataset containing 1,048 image patches was further used for reproducibility-oriented validation after applying the same label harmonization, image resizing, normalization, and evaluation protocol.

To improve transparency of the experimental design, the class-wise distribution of the training, validation, and test subsets is provided in Table 2. The primary dataset was divided into 627 training patches, 135 validation patches, and 135 test patches. The secondary validation dataset contained 640 training patches, 231 validation patches, and 177 test patches. In both datasets, class labels were harmonized into three categories: non-tumor tissue, viable tumor, and non-viable tumor.

Table (2): Class-wise distribution of training, validation, and test samples across datasets.

Dataset	Split	Non-Tumor	Viable Tumor	Non-Viable Tumor	Total
Dataset 1	Training	327	64	236	627
Dataset 1	Validation	70	14	51	135
Dataset 1	Test	71	14	50	135
Dataset 2	Training	340	156	144	640
Dataset 2	Validation	88	77	66	231
Dataset 2	Test	91	44	42	177

To ensure comparability between the primary and secondary datasets, a harmonization procedure was applied before model evaluation. Labels from both datasets were mapped into the same three-class scheme: non-tumor tissue, viable tumor, and non-viable tumor. Image patches from both sources were resized to 224 × 224 pixels and normalized using the same ImageNet mean and standard deviation values. The same class mapping, preprocessing pipeline, evaluation metrics, and model inference procedure were then used across both datasets. This harmonization allowed the secondary dataset to be used as a reproducibility-oriented validation set while acknowledging that both datasets remained patch-level histopathology datasets rather than patient-independent whole-slide cohorts.

Stage 1: binary tumor detection

Stage 1 classified each image patch as non-tumor or tumor-containing. The original three labels were mapped into two categories. Non-tumor patches were assigned to class 0, whereas viable tumor and non-viable tumor patches were combined as tumor-containing patches and assigned to class 1.

Four architectures were evaluated in this stage: ResNet18, EfficientNet-B0, ViT-Tiny, and FastCNN. ResNet18 and EfficientNet-B0 were used as pretrained convolutional backbones. ViT-Tiny represented the transformer-based architecture, whereas FastCNN served as the lightweight convolutional baseline. The final classification layer of each model was replaced with a two-output layer.

For model-complexity comparison, the trainable parameter counts were also recorded for all Stage 1 architectures. ResNet18 contained 11,177,538 trainable parameters, EfficientNet-B0 contained 4,010,110 trainable parameters, ViT-Tiny contained 5,524,802 trainable parameters, and FastCNN contained 23,714 trainable parameters. These values indicate that FastCNN served as the lightweight baseline, whereas ResNet18, EfficientNet-B0, and ViT-Tiny represented higher-capacity pretrained models for comparative evaluation.

For an input image patch x_i, the model generated a logit vector z_i, where the two output nodes represented non-tumor and tumor-containing tissue. The probability of class $j$was computed using the softmax function:

$$p_{ij} = \frac{\exp(z_{ij})}{\sum_{k=1}^{2} \exp(z_{ik})} \quad (1)$$

where p_{ij} is the predicted probability that image patch $i$belongs to class j, and z_{ij} is the corresponding logit value. Cross-entropy loss was used to train the binary tumor-detection model:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{2} y_{ij} \log(p_{ij}) \quad (2)$$

where N is the number of training samples, $y_{ij}$is the one-hot encoded true label for class j, and p_{ij} is the predicted probability for class j. Accuracy, precision, recall (sensitivity), F1-score, and area under the receiver operating characteristic curve (AUC) were used to assess the performance of the model. These metrics were computed using the independent test set.

Stage 2: multiclass viability classification

Stage 2 classified image patches into three categories: non-tumor tissue, viable tumor, and non-viable tumor. This stage was interpreted as patch-level tissue viability classification rather than whole-slide necrosis quantification. ResNet18 was selected as the backbone for the multiclass viability classification because it offered a practical balance between predictive performance, computational efficiency, and interpretability. Compared with larger architectures, ResNet18 is less computationally demanding while still benefiting from pretrained feature extraction. In addition, its convolutional structure is suitable for Grad-CAM-based visualization, which was required for examining whether model decisions were supported by histopathologically meaningful regions. Therefore, ResNet18 was used as the common backbone for the viability classification and explainability stages.

For Stage 2, the final output layer generated three logits corresponding to non-tumor tissue, viable tumor, and non-viable tumor. The class probabilities were computed as:

$$p_{ij} = \frac{\exp(z_{ij})}{\sum_{k=1}^{3} \exp(z_{ik})} \quad (3)$$

where p_{ij} is the predicted probability of patch $i$belonging to class j, and $j \in \{1,2,3\}$. Weighted cross-entropy loss was used to address class imbalance. The weight for class $j$was calculated as:

$$w_j = \frac{N}{n_j} \quad (4)$$

where N represents the total number of training samples and n_j is the number of samples in class j. The weighted cross-entropy loss was defined as:

$$\mathcal{L}_{WCE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{3} w_j \, y_{ij} \log(p_{ij}) \quad (5)$$

where w_j is the class weight, y_{ij} is the true class indicator, and $p_{ij}$is the predicted probability for class j. This weighting strategy increased sensitivity to the under-represented viable tumor class, which is clinically important for treatment-response assessment.

Explainability analysis

Grad-CAM was used to visualize discriminative image regions associated with model predictions. For ResNet18, the final convolutional block was used as the target layer. Heatmaps were overlaid on original H&E patches to assess whether the model emphasized histopathologically meaningful regions. Viable tumor predictions were expected to highlight cellular tumor-rich areas, whereas non-viable tumor predictions were expected to emphasize necrotic or acellular regions.

Grad-CAM generates a localization map for a specified class c:

$$L^{c}_{GradCAM} = ReLU\left(\sum_{k} \alpha_k^c A^k\right) \quad (6)$$

where A^k represents the feature maps from the final convolutional layer and α_k^c represents class-specific importance weights computed via gradient backpropagation:

$$\alpha_k^c = \frac{1}{Z} \sum_{i} \sum_{j} \frac{\partial y^c}{\partial A_{ij}^k} \quad (7)$$

To strengthen explainability beyond qualitative visualization, saliency-based confidence-drop analysis was also performed. High-saliency and low-saliency image regions were separately masked, and the change in prediction confidence was measured. A larger confidence drop after masking high-saliency regions was interpreted as supportive evidence that model predictions relied on discriminative tissue regions. This analysis was treated as computational support for interpretability, not as definitive expert pathological validation.

Training and implementation details

All experiments were implemented in Python using PyTorch, torchvision, timm, NumPy, pandas, scikit-learn, and matplotlib. ImageNet-pretrained weights were used for ResNet18, EfficientNet-B0, and ViT-Tiny when available. Adam was used as the optimizer. The learning rate was set to $1 \times 10^{-4}$for pretrained models and 1×10^{-3} for FastCNN. No explicit weight

decay was applied in the reported experiments, and the batch size was 16.

Stage 1 models were trained for up to 10 epochs with early stopping based on validation macro-F1. The Stage 2 multiclass model was trained for up to 12 epochs using the same validation-based checkpointing strategy. Validation macro-F1 was used for model selection because it is more appropriate than accuracy under class imbalance. Test data were used only for final evaluation and were not used during training or checkpoint selection.

Table 3 presents the experimental configuration used for model development, including architectures, optimization settings, training schedules, model-selection criteria, and validation procedures. These settings were applied consistently to ensure fair benchmarking and reproducible evaluation.

Table (3): Training and validation protocol.

Component	Configuration
Stage 1 models	ResNet18, EfficientNet-B0, ViT-Tiny, FastCNN
Stage 2 model	ResNet18 three-class classifier
Optimizer	Adam
Learning rate	$1 \times 10^{-4}$for pretrained models; $1 \times 10^{-3}$for FastCNN
Weight decay	Not explicitly applied
Batch size	16
Stage 1 epochs	Up to 10
Stage 2 epochs	Up to 12
Model selection	Validation macro-F1
Class imbalance handling	Weighted cross-entropy and macro-averaged metrics
Validation additions	Secondary dataset validation, CI, McNemar test, ablation, audit

Single-stage versus two-stage ablation

A single-stage versus two-stage ablation experiment was performed to evaluate whether the staged design was methodologically justified. The single-stage model directly predicted one of the three classes: non-tumor, viable tumor, or non-viable tumor. The two-stage gated cascade first applied the Stage 1 tumor/non-tumor classifier. If a patch was predicted as non-tumor, the final output was assigned as non-tumor. If a patch was predicted as tumor-containing, the viability classifier was used to assign the patch as viable tumor or non-viable tumor. This ablation was included to avoid assuming that the two-stage framework is automatically more accurate. The two-stage design was evaluated as a clinically aligned decision structure and compared directly against the conventional single-stage multiclass model.

Evaluation metrics

Model performance was evaluated using accuracy, balanced accuracy, macro-precision, macro-recall, macro-F1 score, weighted-F1 score, confusion matrices, and area under the receiver operating characteristic curve. For binary classification, AUC was computed using the tumor-class probability. For multiclass classification, macro one-vs-rest AUC was computed.

Accuracy, precision, recall, and F1-score were calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

For multiclass evaluation, macro-averaged metrics were reported to ensure balanced consideration across classes.

Clinical validation strategy

Clinical relevance was assessed through patch-level computational pathology analysis rather than formal expert pathologist validation. Previous automated medical-image analysis studies have also used expert-derived ground truth to evaluate the reliability of algorithmic image assessment [28]. Model predictions and Grad-CAM attention maps were examined to determine whether salient regions corresponded to histopathologically meaningful structures associated with viable and non-viable tumor tissue. Quantitative support was further provided through saliency-based confidence-drop testing. The framework is intended as a decision-support tool for patch-level tissue classification and does not perform whole-slide necrosis quantification, patient-level response prediction, or replacement of expert pathological assessment.

Because the available public histopathology data were organized at the image-patch level, the present evaluation was interpreted as patch-level validation rather than strict patient-independent validation. Duplicate-hash checking and case-overlap auditing were used to examine potential leakage risks across data splits. Although no duplicate image patches were intentionally used across training and testing subsets, some public split structures showed possible case-level overlap. Therefore, the reported results should be understood as evidence of patch-level discriminative performance and cross-dataset reproducibility, not as definitive proof of patient-level generalizability. A fully patient-independent split using larger multi-institutional whole-slide cohorts remains necessary before clinical deployment.

Results and Discussion

The main dataset contained 897 osteosarcoma H&E image patches in total: 468 non-tumor patches, 92 viable tumor patches, and 337 non-viable tumor patches. The dataset underwent evaluation through stratified patch-level train-validation-test splits. In addition to this, a second osteosarcoma patch dataset comprising 1,048 patches was utilized for replication-oriented validation. Because the analysis was conducted at the patch level and case-overlap auditing indicated possible overlap across public dataset splits, the findings should be interpreted as patch-level and cross-dataset validation rather than strictly patient-independent validation.

Stage 1: binary tumor detection performance

The results demonstrate that deep learning architectures can effectively distinguish tumor from non-tumor tissue.

Figure (2): Cross-dataset performance comparison of stage 1 tumor classification models using accuracy, macro-f1, and AUC metrics.

Figure 2. Cross-dataset comparison of Stage 1 tumor/non-tumor classification performance across ResNet18, EfficientNet-B0, ViT-Tiny, and FastCNN. The heatmap illustrates the measures of accuracy, macro F1, and area under the curve calculated using primary and secondary osteosarcoma datasets. The pretrained architectures consistently achieved strong tumor/non-tumor classification performance across both datasets, whereas FastCNN showed comparatively lower performance. In Dataset 1, ResNet18 and ViT-Tiny achieved the highest accuracy, while EfficientNet-B0 achieved the highest accuracy and macro-F1 in Dataset 2. All evaluated models achieved AUC values above 0.95.

Figure 3 shows the confusion matrix for the final three-class classification on Dataset 2. Most samples lie along the diagonal, indicating correct class-wise predictions. Specifically, 87 non-tumor, 43 viable-tumor, and 40 non-viable-tumor patches were correctly classified. There were few misclassifications, suggesting that the developed framework was able to differentiate between non-tumor and tumor-containing tissue as well as between viable tumor and non-viable tumor patches.

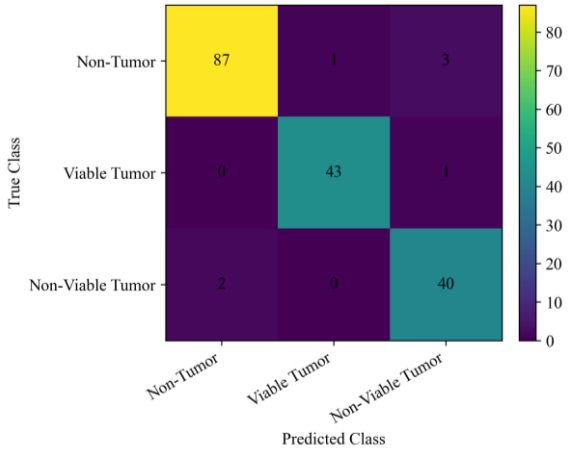


Figure (3): Confusion matrix of the final three-class tumor viability classification framework on dataset 2.

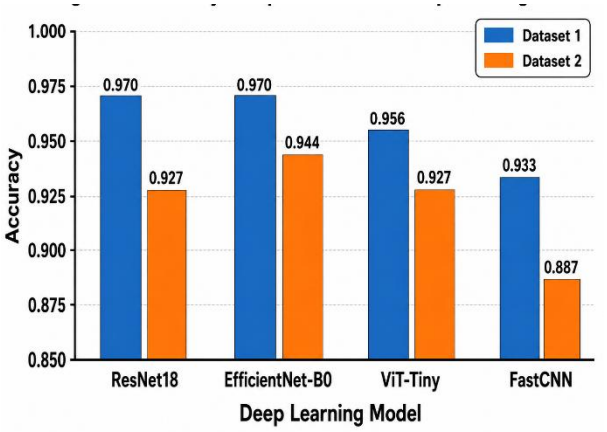


Figure (4): Stage 1 binary tumor detection accuracy across two datasets.

The comparative accuracy analysis presented in Figure 4 further confirms that pretrained deep learning architectures generalized well across both datasets. Although a performance reduction was observed on the secondary dataset, the relative ranking of the models remained largely consistent.

Table 4 shows that the pretrained architectures produced consistently strong patch-level tumor/non-tumor classification performance across both datasets. In Dataset 1, ResNet18 and ViT-Tiny achieved comparable accuracy, while ViT-Tiny produced the highest AUC. In Dataset 2, EfficientNet-B0 achieved the highest accuracy and macro-F1 score. ResNet18 was retained for the subsequent multiclass viability analysis because of its stable performance, architectural simplicity, and compatibility with Grad-CAM interpretation, not because it was uniformly superior across every metric. Pairwise McNemar testing showed that the performance difference between EfficientNet-B0 and FastCNN on Dataset 2 was statistically significant, while differences among the stronger pretrained models were not consistently significant.

Table (4): Stage 1 patch-level tumor/non-tumor classification across primary and secondary datasets.

Dataset	Model	Accuracy	95% CI	Macro-F1	AUC
Dataset 1	ResNet18	0.963	0.916–0.984	0.963	0.994
Dataset 1	EfficientNet-B0	0.933	0.878–0.965	0.933	0.991
Dataset 1	ViT-Tiny	0.963	0.916–0.984	0.963	0.996
Dataset 1	FastCNN	0.919	0.860–0.954	0.918	0.969
Dataset 2	ResNet18	0.927	0.878–0.957	0.927	0.988
Dataset 2	EfficientNet-B0	0.944	0.899–0.969	0.944	0.982
Dataset 2	ViT-Tiny	0.927	0.878–0.957	0.927	0.980
Dataset 2	FastCNN	0.887	0.832–0.926	0.887	0.950

Stage 2: multiclass viability classification

Based on the Stage 1 tumor-detection results, ResNet18 was further used for three-class tissue viability classification into non-tumor tissue, viable tumor, and non-viable tumor.

Table 5 summarizes the final three-class viability classification results. The framework achieved strong patch-level multiclass performance on the primary dataset after support-

adjusted reporting was applied to avoid overinterpreting the small viable-tumor test support. Performance was also reproducible on the secondary dataset, with an accuracy of 0.960, a macro-F1 of 0.958, and a macro-AUC of 0.995. The secondary dataset confusion matrix was [[87,1,3], [0,43,1], [2,0,40]], indicating that most errors occurred between non-tumor and non-viable tumor categories rather than within the viable tumor class.

Single-stage versus two-stage ablation

Table 6 directly compares the conventional single-stage multiclass classifier with the proposed two-stage design. In both datasets, the single-stage model achieved slightly higher numerical performance than the two-stage cascade. However, McNemar testing did not show a statistically significant paired prediction difference for Dataset 1 or Dataset 2. Therefore, the staged framework is not interpreted as more accurate than the conventional single-stage model; its principal value lies in the clinically aligned decomposition of tumor-relevant patch screening followed by tissue viability interpretation. Support-adjusted values were reported for Dataset 1 to avoid overinterpreting unstable class-level estimates arising from the small viable-tumor test subset.

Table (5): Multiclass tissue viability classification on primary and secondary datasets.

Data set	Model / Framework	Test samples	Accuracy	95% CI	Macro-F1	Macro-AUC	Confusion matrix
Data set 1	Single-stage ResNet18	135	0.978	0.937–0.992	0.943*	0.996*	[[69,1,1], [0,14,0], [1,0,49]]
Data set 2	ResNet18 Stage 2B	177	0.960	0.921–0.981	0.958	0.995	[[87,1,3], [0,43,1], [2,0,40]]

Table (6): Single-stage versus two-stage ablation across datasets.

Dataset	Framework	Accuracy	95% CI	Macro-F1	Macro-AUC	McNemar p-value
Dataset 1	Single-stage ResNet18	0.978	0.937–0.992	0.943*	0.996*	—
Dataset 1	Two-stage gated cascade	0.963	0.916–0.984	0.940*	0.994*	0.625
Dataset 2	Single-stage ResNet18	0.960	0.921–0.981	0.958	0.995	—
Dataset 2	Two-stage cascade	0.932	0.885–0.961	0.928	0.987	0.125

The ROC analysis for Dataset 1 is presented in Figure 5. Both the single-stage and two-stage frameworks achieved near-perfect discrimination among non-tumor, viable tumor, and non-viable tumor classes. The single-stage model achieved AUC values of 0.995, 0.996, and 0.996 for the respective classes, while the two-stage framework achieved corresponding AUC values of 0.991, 0.996, and 0.994.

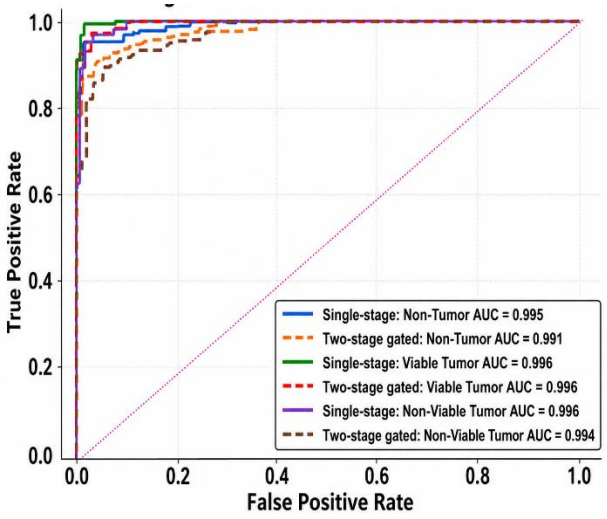


Figure (5): Dataset 1 ROC comparison of single-stage and two-stage gated cascade frameworks.

Similar findings were observed on the secondary validation dataset, as shown in Figure 6. The single-stage framework achieved AUC values of 0.994, 0.998, and 0.993 for non-tumor, viable tumor, and non-viable tumor classes, respectively, whereas the two-stage cascade achieved corresponding AUC values of 0.982, 0.991, and 0.987. These results indicate that the proposed two-stage framework maintained strong discriminative performance on the secondary validation set while providing clinically structured decision decomposition.

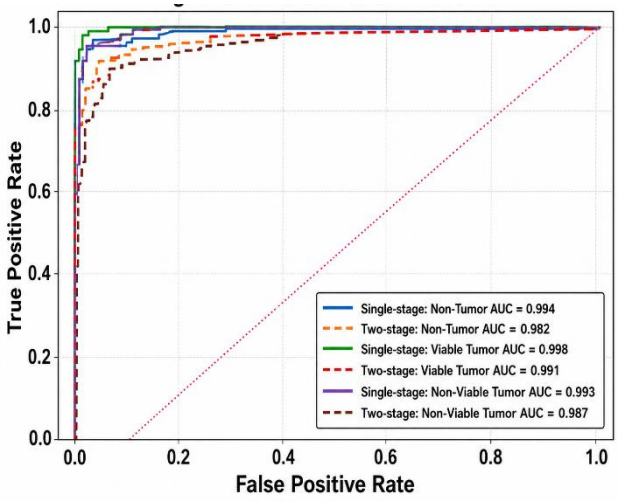


Figure (6): ROC comparison of single-stage and two-stage viability classification frameworks on the secondary validation dataset.

Figure 7. Comparative Stage 2 multiclass viability classification performance on the primary and secondary osteosarcoma datasets. Accuracy, macro-precision, macro-recall, macro-F1, and macro-AUC are shown for the final viability assessment framework. Strong performance across both datasets demonstrates reproducibility of viable tumor, non-viable tumor, and non-tumor tissue classification.

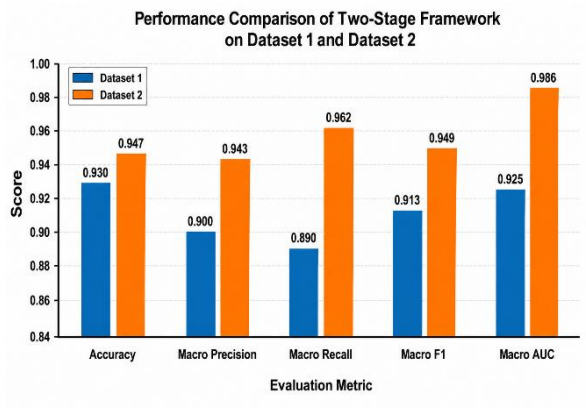


Figure (7): Stage 2 multiclass viability classification performance across two datasets.

Explainability and error analysis

Grad-CAM overlays were used to visually inspect whether the ResNet18 model attended to morphologically relevant regions of the H&E patches. Viable tumor predictions generally highlighted cellular tumor-rich areas, whereas non-viable tumor predictions emphasized necrotic or acellular tissue regions. These visual findings support the interpretability of the model but should not be considered formal expert pathological validation.

Table 7 summarizes the statistical, audit, and explainability analyses used to assess the robustness and interpretability of the framework. To complement the qualitative Grad-CAM observations, saliency-based confidence-drop analysis was performed. On Dataset 2, masking high-saliency regions produced a confidence drop of 0.315, compared with 0.303 after masking low-saliency regions; this difference was not statistically significant by the sign test.

Table (7): Statistical and explainability summary.

Analysis	Result	Interpretation
Dataset 2 Stage 1 McNemar test	EfficientNet-B0 vs FastCNN, p = 0.0129	EfficientNet-B0 significantly outperformed FastCNN.
Dataset 1 ablation McNemar test	p = 0.625	Single-stage and two-stage predictions were not significantly different.
Dataset 2 ablation McNemar test	p = 0.125	Single-stage and two-stage predictions were not significantly different.
Dataset 2 saliency confidence-drop	Top-saliency drop = 0.315; low-saliency drop = 0.303; p = 0.2005	Supportive but not statistically significant explainability evidence.
Dataset audit	Case-overlap observed in public split structures	Results reported as patch-level and secondary validation, not strict patient-independent validation.

Grad-CAM visualization suggested that model attention was associated with biologically relevant tissue structures. Viable-tumor predictions primarily focused on cellular regions with preserved tumor morphology, whereas non-viable-tumor predictions emphasized acellular, eosinophilic, or necrotic treatment-related regions. For non-tumor patches, activation was primarily observed in structurally preserved tissue regions rather than tumor-associated areas. These attention patterns support the biological plausibility and interpretability of the proposed approach. Representative Grad-CAM visualizations are shown in Figure 8.

Similar explainability behavior was observed on the secondary validation dataset. Representative Stage-2B predictions are shown in Figure 9. For non-tumor patches, attention was primarily directed toward regions with preserved tissue architecture; for viable-tumor patches, it focused on densely cellular tumor regions; and for non-viable-tumor patches, it emphasized degenerative and necrotic tissue.

Supplemental qualitative analysis

Misclassification analysis indicated that most errors occurred in morphologically heterogeneous regions containing mixtures of viable tumor cells, necrosis, stromal elements, and treatment-related changes. Such overlapping histopathological features can make patch-level distinction challenging. These findings support the use of the proposed framework as a computational decision-support approach rather than as a standalone diagnostic system.

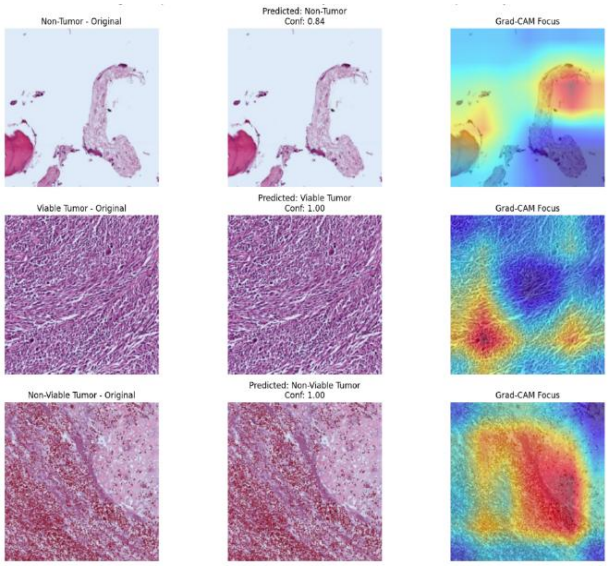


Figure (8): Dataset-1 Representative Grad-CAM overlays with predicted class and confidence scores.

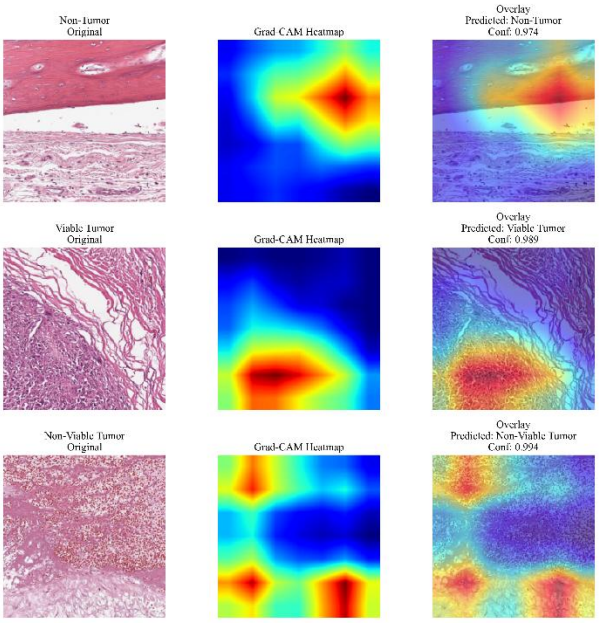


Figure (9): Stage-2B Grad-CAM explainability results on the secondary osteosarcoma dataset.

Discussion

Patch-based deep learning demonstrated strong potential for osteosarcoma tissue viability classification when the task was formulated at the image-patch level rather than as whole-slide necrosis quantification. Histological interpretation may also be influenced by complex stromal and immune-cell interactions within the tumor microenvironment, which contribute to tissue-level heterogeneity [29]. The framework showed consistent performance across both datasets, achieving an accuracy of 0.978 on Dataset 1 and 0.960 on Dataset 2, thereby supporting the reproducibility of the patch-level classification approach. The ablation analysis further showed that, although the conventional single-stage classifier achieved slightly higher numerical performance than the two-stage cascade, the difference was not statistically significant on either dataset. Therefore, the principal value of the staged framework lies in its clinically aligned decomposition of tumor screening and viability assessment rather than in superior predictive accuracy.

Several limitations should be considered when interpreting these findings. The relatively small viable-tumor class, patch-level data partitioning, possible case-level overlap in the public dataset splits, and the absence of strictly patient-independent whole-slide validation may limit the generalizability of the results. In addition, variations in H&E staining protocols, scanner characteristics, tissue preparation procedures, and institutional practices may introduce domain shifts that could affect model performance across clinical centers. Consequently, the findings should be interpreted as evidence of patch-level discriminative performance and cross-dataset reproducibility rather than as direct evidence of patient-level necrosis assessment or clinical validity. Future studies should evaluate the framework using larger multi-institutional whole-slide cohorts with patient-independent data partitioning, stain-robust preprocessing, prospective pathologist review, calibration analysis, slide-level aggregation of patch predictions, and integration with treatment-response outcomes before clinical decision-support application.

Conclusion

This study presented a deep learning framework for patch-level osteosarcoma histopathology tissue viability classification. The primary dataset contained 897 H&E patches, and a secondary dataset of 1,048 patches was used for reproducibility-oriented validation. The framework achieved strong tumor/non-tumor classification and multiclass viability performance across both datasets. On Dataset 1, the framework achieved an accuracy of 0.978 with a support-adjusted macro-F1 of 0.943, while on Dataset 2 it achieved an accuracy of 0.960, a macro-F1 of 0.958, and a macro-AUC of 0.995. The single-stage versus two-stage ablation showed that the staged design was statistically comparable, but not superior, to the direct multiclass classifier. Its main value lies in clinical workflow alignment and interpretable decision decomposition. Grad-CAM and saliency analysis supported model interpretability, although expert pathological validation remains necessary. Future work should focus on patient-independent splitting, larger multi-institutional whole-slide datasets, prospective pathologist-led validation, slide-level aggregation of patch predictions, calibration analysis, and integration with clinical outcome data.

Disclosure Statement

- **Acknowledgements:** None
- **Ethics approval and consent to participate:** Not applicable
- **Consent for publication:** Not applicable
- **Availability of data and materials:** The datasets used in this study were obtained from two publicly available osteosarcoma histopathology repositories: the Osteosarcoma Tumor Assessment dataset available through The Cancer Imaging Archive (TCIA) (https://www.cancerimagingarchive.net/collection/osteosarcoma-tumor-assessment/) and a secondary osteosarcoma histopathology dataset available on Kaggle (https://www.kaggle.com/datasets/gauravupadhyay0312/osteosarcoma).
- **Author's contribution:** The authors confirm that all authors contributed to the conceptualization, methodology, formal analysis, validation, visualization, writing of the original draft, and review and editing of the manuscript.
- **Funding:** The authors verify that no other funding was received for this study.
- **Conflicts of interest:** The authors declare that there is no conflict of interest regarding the publication of this article
- **Source of the Research**: Authors are confirming that this study is not derived from a thesis or dissertation.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated

otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc/4.0/

References

1] Aarthy R, Muthupriya V, Balaji G. Detection of bone cancer based on a four-phase framework generative deep belief neural network in deep learning. Alexandria Engineering Journal. 2024;109:394–407. doi: 10.1016/j.aej.2024.08.094.

2] Alabdulkreem E, Saeed MK, Alotaibi SS, Allafi R, Mohamed A, Hamza MA. Bone cancer detection and Classification Using Owl Search Algorithm with deep learning on X-Ray images. Ieee Access. 2023;11:109095–103. doi: 10.1109/access.2023.3319293.

3] Anisuzzaman D, Barzekar H, Tong L, Luo J, Yu Z. A deep learning study on osteosarcoma detection from histological images. Biomedical Signal Processing and Control. 2021;69:102931. doi: 10.1016/j.bspc.2021.102931.

4] Arunachalam HB, Mishra R, Daescu O, Cederberg K, Rakheja D, Sengupta A, 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. doi: 10.1371/journal.pone.0210706.

5] Aziz MT, Mahmud SH, Elahe MF, Jahan H, Rahman MH, Nandi D, et al. A novel hybrid approach for classifying osteosarcoma using deep feature extraction and multilayer perceptron. Diagnostics. 2023;13(12):2106. doi: 10.3390/diagnostics13122106.

6] Bandyopadhyay O, Biswas A, Bhattacharya BB. Bone-cancer assessment and destruction pattern analysis in long-bone X-ray image. Journal of digital imaging. 2019;32(2):300–13. doi: 10.1007/s10278-018-0145-0.

7] Bielack SS, Kempf-Bielack B, Delling Gn, Exner GU, Flege S, Helmke K, et al. Prognostic factors in high-grade osteosarcoma of the extremities or trunk: an analysis of 1,702 patients treated on neoadjuvant cooperative osteosarcoma study group protocols. Journal of clinical oncology. 2002;20(3):776–90. doi: 10.1200/jco.20.3.776.

8] Borji A, Kronreif G, Angermayr B, Hatamikia S. Advanced hybrid deep learning model for enhanced evaluation of osteosarcoma histopathology images. Frontiers in medicine. 2025;12:1555907. doi: 10.3389/fmed.2025.1555907.

9] Bourouis S, Chennoufi I, Hamrouni K, editors. Multimodal bone cancer detection using fuzzy classification and variational model. Iberoamerican Congress on Pattern Recognition; 2013: Springer. doi:10.1007/978-3-642-41822-8_22.

10] Breden S, Hinterwimmer F, Consalvo S, Neumann J, Knebel C, von Eisenhart-Rothe R, et al. Deep learning-based detection of bone tumors around the knee in X-rays of children. Journal of Clinical Medicine. 2023;12(18):5960. doi: 10.3390/jcm12185960.

11] Gawade S, Bhansali A, Patil K, Shaikh D. Application of the convolutional neural networks and supervised deep-learning methods for osteosarcoma bone cancer detection. Healthcare Analytics. 2023;3:100153. doi: 10.1016/j.health.2023.100153.

12] Meem RF, Hasan KT. Osteosarcoma tumor detection using different transfer learning models. *Clinical Imaging and Case Reports.* 2024;6(1):1–8. doi:10.52338/cicasrep.2024.3962.

13] Meyers PA, Schwartz CL, Krailo MD, Healey JH, Bernstein ML, Betcher D, et al. Osteosarcoma: the addition of muramyl tripeptide to chemotherapy improves overall survival—a report from the Children's Oncology Group. Journal of clinical oncology. 2008;26(4):633–8. doi: 10.1200/jco.2008.14.0095.

14] Nasir MU, Khan S, Mehmood S, Khan MA, Rahman A-u, Hwang SO. IoMT-based osteosarcoma cancer detection in histopathology images using transfer learning empowered with blockchain, fog computing, and edge computing. Sensors. 2022;22(14):5444. doi: 10.3390/s22145444.

15] Panayides AS, Amini A, Filipovic ND, Sharma A, Tsaftaris SA, Young A, et al. AI in medical imaging informatics: current challenges and future directions. IEEE journal of biomedical and health informatics. 2020;24(7):1837–57. doi: 10.1109/jbhi.2020.2991043.

16] Qi Y, Liu Y, Luo J. Recent application of Raman spectroscopy in tumor diagnosis: from conventional methods to artificial intelligence fusion. PhotoniX. 2023;4(1):22. doi: 10.1186/s43074-023-00098-0.

17] Ramana MV, Jyothi PN, Anuradha S, Lakshmeeswari G. Enhanced bone cancer diagnosis through deep learning on medical imagery. *International Journal of Computational and Experimental Science and Engineering.* 2025;11(1). doi:10.22399/ijcesen.931.

18] Sampath K, Rajagopal S, Chintanpalli A. A comparative analysis of CNN-based deep learning architectures for early diagnosis of bone cancer using CT images. Scientific Reports. 2024;14(1):2144. doi: 10.1038/s41598-024-52719-8.

19] Sang H, Lin T, Luo L, Liu M, Li J, Luo X, et al. Multimodal deep learning for bone tumor diagnosis with clinical imaging, pathology, and blood biomarkers. Journal of Bone Oncology. 2025:100718. doi: 10.1016/j.jbo.2025.100718.

20] Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019;6:60. https://doi.org/10.1186/s40537-019-0197-0

21] Song L, Li C, Tan L, Wang M, Chen X, Ye Q, et al. A deep learning model to enhance the classification of primary bone tumors based on incomplete multimodal images in X-ray, CT, and MRI. Cancer Imaging. 2024;24(1):135. doi: 10.1186/s40644-024-00784-7.

22] Vezakis IA, Lambrou GI, Matsopoulos GK. Deep learning approaches to osteosarcoma diagnosis and classification: A comparative methodological approach. Cancers. 2023;15(8):2290. doi: 10.3390/cancers15082290.

23] Whelan J, Jinks R, McTiernan A, Sydes M, Hook J, Trani L, et al. Survival from high-grade localised extremity osteosarcoma: combined results and prognostic factors from three European Osteosarcoma Intergroup randomised controlled trials. Annals of Oncology. 2012;23(6):1607–16. doi: 10.1093/annonc/mdr491.

24] Wu M, Hu Y, Tian F, Lin H. EfficientNetSwift: A lightweight and precise deep learning model for detecting oral squamous cell carcinoma using pathological images. Technology in cancer research & treatment. 2025;24:15330338251380966. doi: 10.1177/15330338251380966.

25] Xu Z, Niu K, Tang S, Song T, Rong Y, Guo W, et al. Bone tumor necrosis rate detection in few-shot X-rays based on deep learning. Computerized Medical Imaging and Graphics. 2022;102:102141. doi: 10.1016/j.compmedimag.2022.102141.

26] Yao H, Yang M, Jiang X, Jia H, Sun T, Li M, et al. Research on the application of a multi-model cascaded deep learning framework in the pathological diagnosis of osteosarcoma. Oncology Reviews. 2025;19:1592408. doi. https://doi.org/10.3389/or.2025.1592408

27] Zuo B, Xiong F. Enhancing Lung Cancer Diagnosis with MixMAE: Integrating Mixup and Masked Autoencoders for Superior Pathological Image Analysis. Traitement du Signal. 2024;41(2):729. doi: 10.18280/ts.410215.

28] Dwickat T, Hamad H. Detecting diabetic retinopathy exudates in fundus images using fuzzy c-means (FCM). An-Najah University Journal for Research-A (Natural Sciences). 2020;35(1):37–64. doi: 10.35552/anujr.a.35.1.1861.

29] Ahmed AM, Khaleel HK, Ibrahim TK, Kadhim AH. The Emerging Role of Stromal-Immune Cell Interactions in Tissue-Specific Immunity and Disease Progression: A Histological Perspective. An-Najah University Journal for Research-A (Natural Sciences). 2025;9999(9999):None–None. doi: 10.35552/anujr.a.40.2.2591.