International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. Comp. Info. Vol. 12 (2026) 1–13

Paper Type: Original Article

Deep Learning for Primary Bone Tumor Analysis on Radiographs: A Narrative Review of Classification, Detection, Localization, and Explainability

Doha Khaled [1,*], Soha Safwat [1], Lamiaa S. Fayed [2], and Soaad M. Naguib [2]

[1] Faculty of Computers and Information Systems Egyptian Chinese university, Egypt; Doha.Khaled@ecu.edu.eg; ssafwat@ecu.edu.eg.

[2] Department of Information Systems; Faculty of Computers and Informatics, Zagazig University, Egypt; lsfayed@fci.zu.edu.eg; smnagieb@fci.zu.edu.eg.

Received: 03 Mar 2026 **Revised:** 26 May 2026 **Accepted:** 23 Aug 2026 **Published:** 24 Aug 2026

Abstract

Primary bone tumors are uncommon but clinically important lesions with heterogeneous radiographic appearances, making accurate diagnosis challenging. Conventional radiography remains an important first-line imaging modality, while artificial intelligence (AI), particularly deep learning (DL), has increasingly been investigated for computer-aided analysis of musculoskeletal abnormalities. This narrative review synthesizes recent applications of DL for primary bone tumor analysis on conventional radiographs, with emphasis on classification, detection and localization, segmentation, multitask learning, transformer-based approaches, and explainable artificial intelligence (XAI). The reviewed studies are compared in terms of model architectures, datasets, task definitions, performance measures, external validation, comparison with human readers, and explainability strategies. The literature indicates that DL models can achieve promising performance in selected bone tumor classification and detection tasks, with some studies reporting results comparable to experienced radiologists. However, direct comparison across studies remains difficult because of differences in datasets, tumor categories, imaging settings, evaluation protocols, and validation strategies. Important challenges include the rarity of primary bone tumors, class imbalance, limited multicenter and external validation, heterogeneous radiographic data, and inconsistent evaluation of model explanations. Recent developments in publicly available datasets, lesion localization, transformer-based architectures, multimodal learning, and XAI provide opportunities for more robust and clinically relevant systems. Future research should prioritize larger multicenter datasets, patient-level validation, clinically meaningful lesion localization and explainability, integration of complementary clinical information, and prospective evaluation of AI-assisted diagnostic workflows.

Keywords: Artificial Intelligence; Deep Learning; Bone Tumors; Bone Neoplasms; Radiography; Convolutional Neural Networks; Object Detection.

1 | Introduction

Primary bone tumors represent a heterogeneous group of neoplasms with substantially different biological behavior, imaging characteristics, and clinical management. Their radiographic appearance may vary considerably according to tumor type, anatomical location, growth pattern, and biological aggressiveness. Consequently, accurate interpretation of bone radiographs may be challenging, particularly when lesions are subtle or when the interpreting physician has limited experience in musculoskeletal oncology [1, 13].

Corresponding Author: Doha.Khaled@ecu.edu.eg

Conventional radiography remains an important component of the initial evaluation of suspected bone tumors. Radiographs provide information about lesion location, bone destruction, matrix mineralization, periosteal reaction, and other structural characteristics that contribute to differential diagnosis. However, interpretation is dependent on reader experience, and some lesions may have overlapping or nonspecific appearances. These challenges have encouraged the development of computer-aided diagnostic systems capable of identifying suspicious abnormalities and assisting clinicians in their interpretation [11, 13]. Artificial intelligence (AI) has increasingly been investigated in medical imaging, with DL becoming an important approach for image-based analysis. DL models can automatically learn hierarchical representations from medical images and have been applied to classification, detection, segmentation, and prediction tasks. Convolutional neural networks(CNNs) have been particularly influential because convolutional operations learn spatially localized visual features, while transformer-based architectures provide mechanisms for modeling longer-range relationships within images [7, 9, 10, 18].For primary bone tumors specifically, the literature has expanded from CNN-based classification approaches toward object detection, segmentation, transformer-based models, and multitask systems. The studies reviewed here include EfficientNet-B0, InceptionV3-based fusion, Mask R-CNN, Vision Transformer, YOLO-based detection, and newer YOLOv11-based detection. [2–5, 8, 11].

Despite these advances, the existing literature remains fragmented across different tasks, datasets, anatomical regions, and evaluation strategies. Furthermore, although some models provide visual explanations or clinically interpretable outputs, explainability has not been consistently incorporated into bone tumor AI systems. Therefore, this narrative review aims to synthesize the current evidence on DL-based primary bone tumor analysis using conventional radiographs, with emphasis on classification, detection and localization, segmentation, multitask learning, transformer-based approaches, and XAI, while critically examining dataset characteristics, external validation, clinical comparison, and persistent challenges to clinical translation.

2 | Review Methodology

2.1 | Review Scope and Objective

This narrative review was designed to synthesize the published evidence on AI and DL methods for primary bone tumor analysis on conventional radiographs. The review focuses on the clinical tasks represented in the literature, including classification, detection and localization, segmentation, multitask learning, transformer-based approaches, and XAI. The objective was to organize the available evidence thematically, compare methodological characteristics and reported performance, and identify persistent limitations and research gaps.

2.2 | Literature Identification and Scope

Relevant literature was identified through structured searches of Google Scholar, Web of Science, and Scopus, using combinations of keywords related to primary bone tumors, bone neoplasms, radiography, X-ray imaging, AI, DL, machine learning (ML), classification, detection, segmentation, localization, and explainable AI. The search covered studies published between 2020 and 2026, while earlier foundational studies were retained when necessary to provide methodological context.

A total of 8 primary AI/DL studies focusing on primary bone tumors on conventional radiographs met the scope of the review and were included in the comparative synthesis (Table 3). These were supplemented by 2 narrative or survey articles [7,13], which were used to provide background and methodological context. Two additional AI/DL studies identified during the literature search, Deng et al. (2024) [6] and Anand et al. (2022) [24], were not included in the formal radiographic comparison because they relied on histopathological rather than radiographic images and therefore did not meet the imaging-modality scope of the comparative synthesis. These studies were retained only as complementary references where relevant to the broader discussion.

Overall, 23 sources were incorporated into the review, comprising 10 primary AI/DL studies addressing bone tumor imaging (8 based on conventional radiographs and included in the comparative synthesis, and 2 using complementary histopathological data), 1 publicly available dataset resource [21] cited through its Kaggle repository and accompanying Scientific Data publication, and 12 foundational, methodological, or narrative-review references providing background and methodological context. The literature was selected to represent studies directly addressing primary bone tumors and applying AI or DL methods to clinically relevant imaging tasks. The core evidence synthesis focused on conventional radiographs, while selected studies using other data modalities were included only when they provided relevant methodological or contextual information. Particular attention was given to classification, detection and localization, segmentation, multitask learning, external validation, comparison with radiologists, explainability, and publicly available datasets

2.3 | Eligibility Framework

Studies were considered relevant when they met the following scope criteria: (1) primary bone tumors were the principal clinical focus; (2) conventional radiography was used as the primary imaging modality for the reported analysis; (3) an AI, ML, or deep-learning method was applied to classification, detection, localization, segmentation, or a closely related task; and (4) the study provided sufficient methodological or performance information to support comparison. Foundational methodological papers were retained when they were directly relevant to architectures or XAI methods discussed in the reviewed studies.

2.4 | Study Organization and Data Extraction

For each representative study, information was synthesized according to publication year, dataset characteristics, number of patients or radiographs when reported, study setting, model or architecture, primary task, localization or segmentation capability, explainability strategy, external validation, comparison with human readers, and principal reported performance measures. The extracted information was organized into comparative tables and thematic sections covering classification, detection/localization, transformer-based approaches, multitask learning and segmentation, XAI, datasets, clinical value, and research gaps.

2.5 | Narrative Synthesis

Because the reviewed studies differ substantially in datasets, target definitions, anatomical regions, validation strategies, and evaluation metrics, the evidence was synthesized narratively rather than by pooling performance estimates. Reported accuracy, AUC, sensitivity, specificity, mAP, IoU, and Dice values are therefore presented as study-specific results and should not be interpreted as a direct ranking of models. This approach also allows methodological limitations, external validation, explainability, and clinical relevance to be considered alongside predictive performance.

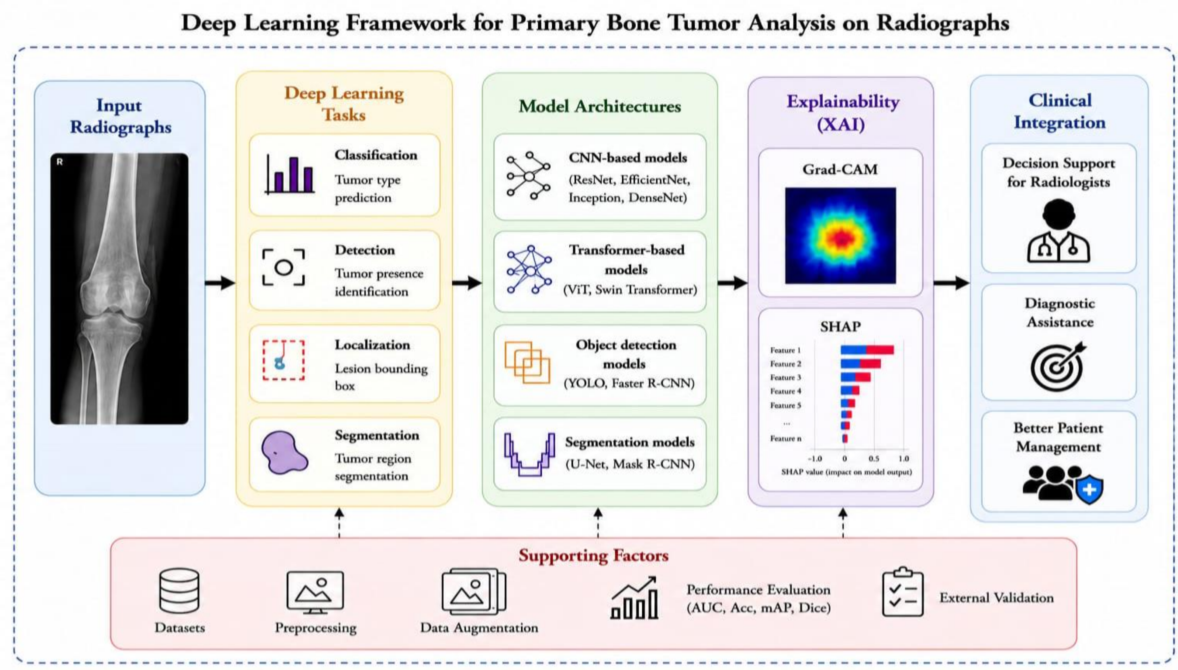

Figure 1. Conceptual framework of deep learning for primary bone tumor analysis on conventional radiographs.

3 | Artificial Intelligence in Bone Tumor Imaging

3.1 | Artificial Intelligence, Machine Learning, and Deep Learning

AI refers broadly to computational methods designed to perform tasks that traditionally require human intelligence. MML represents a subset of AI in which algorithms learn patterns from data rather than relying exclusively on explicitly programmed rules. DL is a further subset of ML based on multilayer neural networks capable of learning increasingly complex representations from input data. In medical image analysis, DL has several advantages over conventional machine-learning pipelines because feature extraction and prediction can be learned jointly from image data. CNNs became particularly influential because their convolutional operations learn spatially localized visual features. More recently, transformer-based architectures such as the Vision Transformer and Swin Transformer have been introduced to capture broader contextual relationships. [9,10,16,17]

The reviewed bone tumor literature reflects this evolution. Early studies primarily relied on CNN-based architectures, including EfficientNet-B0, while later work introduced object detectors, Vision Transformers, and multitask architectures. The methodological background is supported by foundational work on residual networks, EfficientNet, Vision Transformers, and transformer-based vision models [14–17].

4 | Deep Learning Applications in Primary Bone Tumor Analysis

4.1 | Classification of Primary Bone Tumors

Classification was among the earliest applications of DL in primary bone tumor radiography. The objective may range from binary discrimination, such as tumor versus normal or malignant versus nonmalignant, to more challenging multiclass classification involving benign, intermediate, and malignant tumors [2,5]. A landmark preliminary study published in 2020 investigated DL-based classification of primary bone tumors using 2,899 radiographs from 1,356 patients collected from five institutions. For benign versus not-benign classification, the model achieved AUC values of 0.894 during cross-validation and 0.877 on external testing. For malignant versus not-malignant classification, AUC values were 0.907 and 0.916, respectively. Three-way classification achieved an accuracy of 72.1% during cross-validation and 73.4% on external testing. The model showed performance comparable to subspecialist radiologists and better performance than several junior radiologists [2]. This study is important because it demonstrated generalization beyond a single institution through external testing. At the same time, the lower accuracy of the three-way task compared with the binary tasks illustrates the difficulty of separating biologically and radiographically overlapping categories [2]. A subsequent approach combined DL with conventional ML and clinical information. Liu et al. analyzed 643 patients and 982 radiographs. In the three-category classification task, the radiological model achieved a macro-average AUC of 0.813, whereas the fusion model achieved 0.872. The mean macro-average AUC of the radiologists was 0.819, and model assistance improved the radiologists' macro-average AUC by 0.026. SHAP analysis was used to assess feature contributions [5]. These findings suggest that image-based DL may benefit from complementary clinical information. Rather than relying exclusively on visual information, multimodal models can integrate demographic, clinical, and radiographic features when these data are available. [5]

Deng et al. (2024) provide related evidence for multimodal computational analysis, but their study used whole-slide histopathological images rather than radiographs. The study therefore should not be treated as directly comparable with the radiograph-based studies reviewed here. It is nevertheless relevant to the broader discussion of multimodal and ML-assisted bone tumor diagnosis, particularly because the reported performance differed markedly between tumor-versus-normal and benign-versus-malignant tasks. [6]

4.2 | Detection and Localization

Classification alone does not indicate where a tumor is located within a radiograph. Localization is therefore an important component of computer-aided diagnosis because a clinically useful system should ideally identify suspicious lesions and provide an interpretable spatial indication of the abnormality. Object detection explicitly addresses this limitation by producing spatial predictions such as bounding boxes [8,11]. Object-detection approaches have increasingly been applied to this problem. Li et al. developed a YOLO-based model using 1,085 bone tumor radiographs and 345 normal radiographs from two centers. The model detected bone lesions and classified radiographs into normal, benign, intermediate, or malignant categories. In validation, it correctly placed 86.36% of bounding boxes in the internal set and 85.27% in the external set at IoU > 0.5 [8,11].

The use of YOLO represents an important transition from image-level classification toward integrated localization and classification. Instead of producing only a class probability, an object detector can provide a bounding box around a suspicious region, making the output more closely aligned with radiographic interpretation [8].

A 2024 multicenter study investigated DL-based detection of primary bone tumors around the knee. The study included 687 tumor patients and 1,988 normal participants from four tertiary referral centers after image exclusions. The model correctly localized 94.5% and 92.9% of tumor lesions in the internal and external test sets, respectively. Patient-level detection accuracy was 0.964 and 0.920, with AUC values of 0.981 and 0.990. The model also showed higher agreement with the reference diagnosis than two junior radiologists in the internal test set [11].

The multicenter design and external evaluation are particularly important because generalization is a major challenge in medical AI. The Xu et al. study also assessed performance across different imaging devices and centers, providing evidence that external testing can reveal how models behave under changing acquisition conditions [11].

4.3 | Transformer -Based Approaches

While CNNs have dominated early bone tumor research, transformer-based architectures have attracted attention because of their ability to model broader contextual relationships. The Vision Transformer introduced a transformer architecture for image recognition, while Swin Transformer provides hierarchical representations through shifted-window attention [16, 17].

Breden et al. investigated a pretrained Vision Transformer for detecting pathological versus healthy knee radiographs in children. The study used extensive data augmentation to address limited data availability. In the test groups, the model achieved an accuracy of 89.1%, sensitivity of 82.2%, and specificity of 93.2%. [3,16]

An important aspect of this study was the use of Grad-CAM to investigate the plausibility of model predictions. Heatmaps highlighted image regions influencing predictions, but misleading activations were also observed in some misclassified cases [3, 20]. This observation is important because high predictive performance alone does not establish that a model has learned clinically meaningful features. Explanation methods can help investigators inspect whether the model attends to a plausible lesion region or to irrelevant image characteristics, although attribution maps should not be interpreted as proof of causal reasoning [20].

5 | Multitask Learning and Tumor Segmentation

Another important direction is the development of models that perform multiple related tasks simultaneously. In bone tumor analysis, localization, segmentation, and classification are closely related: identifying the lesion can facilitate segmentation, while the segmented lesion can provide a focused representation for classification.

von Schacky et al. proposed a multitask DL model based on Mask R-CNN-X101 for simultaneous bounding-box placement, segmentation, and classification of primary bone tumors on radiographs. The study included 934 patients in the internal dataset and 111 patients in an external test set. The model achieved 80.2% accuracy, 62.9% sensitivity, and 88.2% specificity for benign-versus-malignant classification, with an IoU of 0.52 ± 0.34 for bounding boxes and a mean Dice score of 0.60 ± 0.37 for segmentation [4]. The model's classification accuracy was higher than that of two radiologic residents (71.2% and 64.9%) and comparable with two musculoskeletal fellowship-trained radiologists (83.8% and 82.9%) [4]. Multitask learning has several potential advantages. It allows one model to produce multiple clinically useful outputs and can provide explicit lesion localization and segmentation. However, the richness of these outputs comes with greater annotation and model-development requirements [4]. Multitask learning also introduces additional annotation requirements. Bounding boxes and segmentation masks require expert annotation and are more costly to obtain than image-level labels. Public datasets containing multiple annotation types can therefore be valuable for reproducible research [4, 9].

6 | Explainable Artificial Intelligence in Bone Tumor Analysis

6.1 | Motivation for Explainability

DL models are frequently described as black-box systems because their internal representations are difficult to interpret directly. In medical imaging, interpretability is particularly important when predictions may influence clinical decision-making [10,13]. Explainable AI can provide additional information about model predictions. In radiographic bone tumor analysis, explanation approaches include visual attribution methods such as Grad-CAM and feature-attribution approaches such as SHAP, while bounding boxes and segmentation masks can provide explicit spatial outputs. [5,20]

6.2 | Grad-CAM and Visual Explanations

Gradient-weighted Class Activation Mapping (Grad-CAM) produces a heatmap indicating image regions associated with a model's prediction. In the pediatric bone tumor study, Grad-CAM was used to assess whether the Vision Transformer focused on relevant regions of the radiograph [3,20]. This demonstrates both the value and limitation of visual explanation. A heatmap can improve transparency and help investigators inspect model behavior, but an apparently reasonable heatmap does not by itself establish causal correctness. [20]

6.3 | SHAP and Feature-Based Explanations

SHAP provides a different form of interpretability by estimating the contribution of input features to a prediction. Liu et al. used SHAP-based feature importance in their imaging-plus-clinical fusion approach, providing information about which variables contributed to the model's predictions [5]. The distinction between these approaches is important. Grad-CAM primarily addresses a spatial question—where did the model focus? SHAP addresses a feature-contribution question—which input variables contributed to the prediction? [5, 20]

For clinical deployment, combining spatial explanations with structured feature explanations may provide a more comprehensive description of model behavior, but the clinical validity of these explanations requires dedicated evaluation.

6.4 | Comparison of Explainability Approaches Across Studies

Table 1 summarizes the explainability strategy, or its absence, across the reviewed studies.

Table 1. Explainability and interpretability strategies reported in representative bone tumor AI studies.

Study	Model / approach	XAI/ explanation	Type	Main finding / role	Limitation
He et al., 2020	EfficientNet-B0	No dedicated XAI reported	—	Interpretability assessed mainly through reader comparison	No saliency-based explanation
von Schacky et al., 2021	Mask R-CNN-X101	Bounding boxes + masks	Spatial output	Provides explicit lesion localization and segmentation	Does not identify visual features driving classification
Liu et al., 2022	InceptionV3 + XGBoost + clinical data	SHAP	Feature attribution	Quantifies contributions of clinical and model-derived features	Not a direct lesion-localization explanation
Breden et al., 2023	Vision Transformer	Grad-CAM	Spatial attribution	Assessed whether predictions on focused plausible radiographic regions	Misleading activations occurred in some misclassified cases
Li et al., 2023	YOLO-based model	Detection boxes	Spatial output	Shows the region selected by the detector	Does not explain why that region drove the decision
Xu et al., 2024	DL detection model	No dedicated XAI reported	Spatial localization only	Bounding boxes provide localization	No feature-level explanation
YOLOv11-MTB, 2025	YOLOv11-MTB	No explicit XAI reported	—	Attention is used architecturally for detection, not presented as a clinical explanation	Clinical interpretability not directly evaluated

7 | Datasets for Deep Learning-Based Bone Tumor Analysis

Dataset availability is one of the central limitations of primary bone tumor AI research. Primary bone tumors are uncommon, and obtaining large, diverse, well-annotated datasets is difficult. Reviews of medical image AI have emphasized the importance of dataset size, annotation quality, transfer learning, and external validation for reliable model development [7,10,13]. The early studies included in this review relied largely on institutional or multicenter datasets. For example, the 2020 classification study included 1,356 patients from five institutions, while the 2024 knee tumor detection study included data from four tertiary referral centers [2,11]. A particularly important development was the publication of the Bone Tumor X-ray Radiograph Dataset (BTXRD). BTXRD contains 3,746 radiographs: 1,879 normal, 1,525 benign tumor, and 342 malignant tumor images. Each image includes clinical metadata, while tumor images additionally include diagnosis/subtype information and manually annotated bounding boxes and segmentation masks. The dataset is publicly available and supports classification, localization, and segmentation research [21,22]. The availability of such a dataset is important because it supports multiple research tasks using the same underlying data and facilitates more reproducible benchmarking [22]. Nevertheless, BTXRD also illustrates a persistent challenge: the malignant class is substantially smaller than the normal and benign classes. The distribution is 342 malignant, 1,525 benign, and 1,879 normal images. Such imbalance can make overall accuracy an incomplete indicator of clinically important malignant-case performance [22].

7.1 | Dataset Characteristics Across Reviewed Studies

Table 2 summarizes sample size, number of centers, class/target information, annotation type, public availability, and external validation for the principal radiograph-based studies reviewed here.

Table 2. Dataset characteristics and annotation resources in representative radiograph-based studies.

Study / dataset	Sample size	Centers	Class / target	Annotations	Public availability	External validation
He et al., 2020	2,899 radiographs / 1,356 patients	5	Benign / intermediate / malignant	Image-level diagnostic labels; lesion cropping	Not public	Yes
von Schacky et al., 2021	934 internal + 111 external patients	Multicenter	Benign vs malignant	Bounding boxes + segmentation masks	Not public	Yes
Liu et al., 2022	982 radiographs / 643 patients	Not stated	Benign / malignant / intermediate	Radiographic labels + clinical variables	Not public	Test-set evaluation
Li et al., 2023	1,085 tumor + 345 normal radiographs	2	Normal / benign / intermediate / malignant	Bounding boxes + class labels	Not public	External validation set
Breden et al., 2023	176 pediatric tumor patients + 220 healthy controls	1 center	Pathological vs healthy	Image-level labels	Not public	No
Xu et al., 2024	687 tumor + 1,988 normal participants	4 tertiary centers	Tumor vs normal	Lesion localization labels	Not public	Yes
BTXRD (Yao et al., 2025)	3,746 radiographs: 1,879 normal; 1,525 benign; 342 malignant	Multiple sources / centers	Normal / benign / malignant + subtype/site metadata	Global labels + bounding boxes + masks + clinical metadata	Public	Public benchmark

8 | Comparative Summary of Existing Studies

Table 3 presents a structured comparison of the reviewed studies, including the model or approach used, imaging modality, task capabilities (detection, segmentation, and/or classification), explainability strategy, and the main reported result.

Table 3. Comparative summary of representative AI/DL studies relevant to primary bone tumor analysis on radiographs.

Study	Year	Task	Model / approach	Dataset	Localization / segmentation	External validation	Main result
He et al.	2020	Classification	EfficientNet-B0	2,899 images; 1,356 patients; 5 centers	No / No	Yes	Binary AUC 0.877–0.916; external 3-class accuracy 73.4%
von Schacky et al.	2021	Multitask	Mask R-CNN-X101	934 internal + 111 external patients	Yes / Yes	Yes	Accuracy 80.2%;IoU 0.52;Dice 0.60
Liu et al.	2022	Classification + fusion	InceptionV3 + XGBoost+ clinical data	982 images; 643 patients	No / No	Test set	3-class macro-AUC 0.872 vs 0.813;reader AUC +0.026
Li et al.	2023	Detection + classification	YOLO-based	1,085 tumor + 345 normal images; 2 centers	Yes / No	Yes	BBox placement 86.36% / 85.27% (IoU >0.5)
Breden et al.	2023	Healthy vs pathological	Vision Transformer + augmentation	176pediatric tumor patients + 220 controls	Indirect / No	No	Accuracy 89.1%; sensitivity 82.2%; specificity 93.2%
Xu et al.	2024	Detection / localization	Multicenter DL model	687 tumor + 1,988 normal; 4 centers	Yes / No	Yes	Localization 94.5% / 92.9%;AUC 0.981 / 0.990
Dalai et al.	2025	Healthy vs cancerous	SqueezeNet + GSO + LSTM	X-ray dataset; size not stated	No / No	Not reported	Test accuracy 94.79%; not primary-tumor-specific
YOLOv11-MTB	2025	Detection	YOLOv11+ attention+ multiscale fusion	BTXRD+ Munich dataset	Yes / No	Cross-center / cross-dataset	BTXRD mAP 79.6%; small 55.8%; multi-lesion 63.2%

The studies summarized in Table 3 differ substantially in datasets, tasks, class definitions, validation strategies, and evaluation metrics. Therefore, the reported performance values should be interpreted as study-specific results rather than as a direct ranking of models.

9 | Clinical Value of Deep Learning

The potential clinical value of DL in bone tumor radiography extends beyond achieving high classification accuracy. A clinically useful system should support the radiologist throughout the workflow, from identifying suspicious examinations to localizing lesions and providing information relevant to differential diagnosis. [2–5, 8, 11, 13].

Classification models can potentially assist with triage and diagnostic categorization. Detection models can direct attention toward suspicious regions, particularly in full-field radiographs. Segmentation models can provide more precise lesion boundaries, while multitask systems can combine localization, segmentation, and classification within one framework. [4, 8, 11].

The comparison with radiologists reported in several studies is encouraging. The 2020 classification study reported performance comparable to subspecialists and better performance than several junior radiologists; the 2021 multitask model outperformed two residents and was comparable with two fellowship-trained radiologists; the 2022 fusion model improved radiologists' diagnostic performance; and the 2024 multicenter detection model showed higher agreement with the reference diagnosis than two junior radiologists. [2, 4, 5, 11].

Table 4. Comparisons between AI systems and human readers in representative studies.

Study	AI vs reader comparison	Reported outcome
He et al., 2020	AI vs 2 subspecialists + 3 junior radiologists	External 3-way accuracy 73.4%; comparable to subspecialists and higher than several junior readers
von Schacky et al., 2021	AI vs 2 residents + 2 fellowship-trained radiologists	AI accuracy 80.2%; higher than residents (71.2%, 64.9%) and comparable to fellowship-trained readers (83.8%, 82.9%)
Liu et al., 2022	Fusion model support vs unassisted radiologists	Radiologists' mean macro-AUC improved by 0.026 with model support
Xu et al., 2024	AI vs 2 junior radiologists	Higher agreement with reference diagnosis; model kappa 0.927 vs 0.777 and 0.841

However, these findings should not be interpreted as evidence that AI can replace radiologists. Differences in datasets, reader expertise, tasks, and validation designs make direct comparisons difficult. Current evidence more strongly supports AI as an assistive technology that may improve detection, standardize assessment, and prioritize suspicious examinations. [2, 4, 5, 11, 13].

10 | Challenges and Limitations

10.1 Limited Dataset Size and Disease Rarity

Primary bone tumors are uncommon, making the collection of large, representative datasets difficult. Limited sample sizes can increase the risk of overfitting and reduce the reliability of performance estimates across diverse populations [7, 10, 13].

10.2 Class Imbalance

The distribution of tumor categories is often uneven. In BTXRD, malignant tumors account for 342 of 3,746 images, compared with 1,525 benign and 1,879 normal images. This imbalance is important because a model can achieve high overall accuracy while still showing weaker performance on malignant cases [22].

10.3 Heterogeneity of Imaging Data

Radiographs may differ according to acquisition equipment, imaging protocols, anatomical region, projection, image quality, and patient population. Models trained in one institution may therefore learn dataset-specific characteristics that do not generalize to other settings. Multicenter studies such as Xu et al. provide stronger evidence of robustness than single-center evaluations [11, 13].

10.4 Limited External Validation

Although some studies have performed external validation, many AI systems remain primarily internally evaluated. The studies by He et al., von Schacky et al., and Xu et al. provide examples of external evaluation, whereas other studies are more limited in this respect [2, 4, 11].

10.5 Explainability

Only a subset of the reviewed studies explicitly incorporated XAI. The reviewed literature includes Grad-CAM and SHAP, while detection and segmentation systems provide spatial outputs but do not necessarily explain the visual features driving a prediction [3–5, 20].

10.6 Clinical Translation

High performance on a retrospective dataset does not automatically demonstrate clinical effectiveness. Prospective validation, workflow integration, evaluation across institutions, and assessment of the effect of AI assistance on clinical decision-making are needed before widespread deployment [10, 13].

11. Research Gaps and Future Directions

11.1 Multicenter and Larger Datasets

Future studies should prioritize diverse multicenter datasets with standardized annotation and clearly defined patient-level splits. Public resources such as BTXRD provide an important foundation for reproducible research [11, 22].

11.2 Improved Localization of Small and Multiple Lesions

Small lesions and cases containing multiple lesions remain challenging for object detection. The 2025 YOLOv11-MTB study specifically targeted these challenges by incorporating large-kernel attention, multiscale fusion, angle-aware encoding, and adaptive focal loss. On BTXRD, it reported an overall mAP of 79.6%, small-lesion mAP of 55.8%, and multi-lesion mAP of 63.2% [19, 23].

11.3 Integration of Global and Local Information

A promising direction is the combination of full-image information with lesion-focused representations. Global features can capture anatomical context, whereas local features can emphasize tumor morphology. Detection and segmentation frameworks provide one mechanism for combining these complementary representations [4, 8, 11].

11.4 CNN–Transformer Hybrid Models

CNNs remain effective for local spatial feature extraction, whereas transformer architectures can model broader contextual relationships. Future bone tumor systems may therefore benefit from hybrid architectures that combine local and global representations [14–17].

11.5 Explainable AI

Future models should move beyond qualitative heatmaps toward clinically evaluated explainability. Explanation methods should be assessed not only for visual plausibility but also for whether they consistently identify clinically relevant regions or features [20].

11.6 Multimodal Learning

Combining radiographs with clinical characteristics or other complementary information may improve diagnostic performance. The improvement observed with the DL–ML fusion approach supports the potential value of combining radiographic and clinical information [5].

12. Conclusion

DL has emerged as a promising technology for the analysis of primary bone tumors on radiographs. Existing studies demonstrate that CNN-based classification models can achieve performance comparable to experienced radiologists in selected tasks, while object-detection models add lesion localization. Multitask architectures combine localization, segmentation, and classification, and transformer-based approaches provide alternative strategies for capturing broader image context. [2–5, 8, 11, 16, 17].

Despite these advances, the field remains constrained by limited and imbalanced datasets, heterogeneous imaging data, incomplete external validation, differences in task definitions, and inconsistent explainability. The availability of BTXRD, with classification, localization, and segmentation annotations, provides an important opportunity for reproducible benchmarking [22].

Future research should focus on multicenter validation, robust lesion localization, integration of global and local features, clinically meaningful explainability, multimodal learning where appropriate, and prospective evaluation. Rather than replacing radiologists, AI systems are more likely to provide their greatest clinical value as assistive tools that improve detection, characterization, and prioritization of suspicious bone lesions. [2, 4, 5, 11, 13].

Author Contribution

All authors contributed equally to this work.

Funding

This research received no funding.

Data Availability

The datasets referenced in this study are fully anonymized and publicly accessible. As this work constitutes a literature review, no new data were generated.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] National Cancer Institute. Tumor. NCI Dictionary of Cancer Terms. Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor (Accessed July 20, 2026).

[2] He, Y. et al. Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. EBioMedicine 62, 103121 (2020). https://doi.org/10.1016/j.ebiom.2020.103121

[3] Breden, S. et al. Deep Learning-Based Detection of Bone Tumors around the Knee in X-rays of Children. J. Clin. Med. 12(18), 5960 (2023). https://doi.org/10.3390/jcm12185960

[4] von Schacky, C. E. et al. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. Radiology 301(2), 398–406 (2021). https://doi.org/10.1148/radiol.2021204531

[5] Liu, R. et al. A deep learning-machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. Eur. Radiol. 32(2), 1371–1383 (2022). https://doi.org/10.1007/s00330-021-08195-z

[6] Deng, S. et al. Auxiliary diagnosis of primary bone tumors based on machine learning model. J. Bone Oncol. 49, 100648 (2024). https://doi.org/10.1016/j.jbo.2024.100648
[7] Atasever, S. et al. A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning. Clin. Imaging 94, 18–41 (2023). https://doi.org/10.1016/j.clinimag.2022.11.003
[8] Li, J. et al. Primary bone tumor detection and classification in full-field bone radiographs via YOLO deep learning model. Eur. Radiol. 33(6), 4237–4248 (2023). https://doi.org/10.1007/s00330-022-09289-y
[9] Shin, H. C. et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35(5), 1285–1298 (2016). https://doi.org/10.1109/TMI.2016.2528162
[10] Litjens, G. et al. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88 (2017). https://doi.org/10.1016/j.media.2017.07.005
[11] Xu, D. et al. Deep learning-based detection of primary bone tumors around the knee joint on radiographs: a multicenter study. Quant. Imaging Med. Surg. 14(8), 5420–5433 (2024). https://doi.org/10.21037/qims-23-1743
[12] Dalai, S. S. et al. Automated Bone Cancer Detection Using Deep Learning on X-Ray Images. Surg. Innov. 32(2), 94–108 (2025). https://doi.org/10.1177/15533506241299886
[13] Papageorgiou, P. S. et al. Artificial Intelligence in Primary Malignant Bone Tumor Imaging: A Narrative Review. Diagnostics 15(13), 1714 (2025). https://doi.org/10.3390/diagnostics15131714
[14] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
[15] Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proc. 36th Int. Conf. Mach. Learn. (ICML), 6105–6114 (2019).
[16] Dosovitskiy, A. et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Int. Conf. Learn. Represent. (ICLR) (2021).
[17] Liu, Z. et al. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 10012–10022 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
[18] Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning (MIT Press, Cambridge, MA, 2016).
[19] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.324
[20] Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
[21] Mgdan, M. Bone Tumor X-ray with Object Detection. Kaggle Dataset. Available at: https://www.kaggle.com/datasets/mohamedmgdan/bone-tumor-x-ray-with-object-detection (Accessed July 20, 2026).
[22] Yao, S. et al. A Radiograph Dataset for the Classification, Localization, and Segmentation of Primary Bone Tumors. Sci. Data 12, 88 (2025). https://doi.org/10.1038/s41597-024-04311-y
[23] Chen, S. et al. Enhancing YOLOv11 with Large Kernel Attention and Multi-Scale Fusion for Accurate Small and Multi-Lesion Bone Tumor Detection in Radiographs. Diagnostics 15(16), 1988 (2025). https://doi.org/10.3390/diagnostics15161988
[24] Anand, D., Arulselvi, G., Balaji, G. N. & Chandra, G. R. A deep convolutional extreme machine learning classification method to detect bone cancer from histopathological images. Int. J. Intell. Syst. Appl. Eng. 10(4), 39–47 (2022). https://ijisae.org/index.php/IJISAE/article/view/2194