RESEARCH PAPER

Detecting bone lesions in X-ray under diverse acquisition conditions

Tal Zimbalist [ID],[a],* Ronnie Rosen,[a] Keren Peri-Hanania,[a] Yaron Caspi,[a] Bar Rinott,[b] Carmel Zeltser-Dekel,[b] Eyal Bercovich,[b] Yonina C. Eldar,[a] and Shai Bagon [ID][a]

[a]Weizmann Institute of Science, Department of Computer Science & Applied Mathematics, Rehovot, Israel
[b]Rambam Health Care Campus, Medical Imaging Division, Haifa, Israel

ABSTRACT. **Purpose:** The diagnosis of primary bone tumors is challenging as the initial complaints are often non-specific. The early detection of bone cancer is crucial for a favorable prognosis. Incidentally, lesions may be found on radiographs obtained for other reasons. However, these early indications are often missed. We propose an automatic algorithm to detect bone lesions in conventional radiographs to facilitate early diagnosis. Detecting lesions in such radiographs is challenging. First, the prevalence of bone cancer is very low; any method must show high precision to avoid a prohibitive number of false alarms. Second, radiographs taken in health maintenance organizations (HMOs) or emergency departments (EDs) suffer from inherent diversity due to different X-ray machines, technicians, and imaging protocols. This diversity poses a major challenge to any automatic analysis method.

Approach: We propose training an off-the-shelf object detection algorithm to detect lesions in radiographs. The novelty of our approach stems from a dedicated preprocessing stage that directly addresses the diversity of the data. The preprocessing consists of self-supervised region-of-interest detection using vision transformer (ViT), and a foreground-based histogram equalization for contrast enhancement to relevant regions only.

Results: We evaluate our method via a retrospective study that analyzes bone tumors on radiographs acquired from January 2003 to December 2018 under diverse acquisition protocols. Our method obtains 82.43% sensitivity at a 1.5% false-positive rate and surpasses existing preprocessing methods. For lesion detection, our method achieves 82.5% accuracy and an IoU of 0.69.

Conclusions: The proposed preprocessing method enables effectively coping with the inherent diversity of radiographs acquired in HMOs and EDs.

© 2024 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.11.2.024502]

Keywords: bone lesions; deep learning; histogram equalization; object detection; vision transformer

Paper 23135GR received Jul. 3, 2023; revised Feb. 11, 2024; accepted Mar. 4, 2024; published Mar. 19, 2024.

1 Introduction

Malignant bone tumors are relatively rare and account for ∼3% of tumors in children and adolescents.[1] Osteosarcoma (OS) and the Ewing sarcoma family of tumors (ESFT) are the most common malignant bone tumors in children and adolescents.[1] Bone sarcomas is ranked as the third leading cause of cancer death among humans up to 20 years old in the United States.[2] The early diagnosis of a malignant bone tumor is of vital importance because it may increase not only

*Address all correspondence to Tal Zimbalist, tal.zimbalist@weizmann.ac.il

the chance of survival but also the possibility of performing a limb-sparing resection. [3] However, the nonspecific signs and symptoms in patients with primary bone tumors often lead to a delay in the diagnosis and initiation of treatment.

X-ray is the imaging modality frequently used for the initial evaluation of bone lesions[4] because it is highly available and relatively inexpensive. However, many radiologists are not able to develop sufficient expertise to reliably identify and assess these lesions on radiographs due to their low prevalence in the population. Additionally, early indications of bone lesions may be incidentally missed, as radiographs taken in health maintenance organizations (HMOs) or emergency departments (EDs) are usually obtained for other clinical reasons, and the radiologists examining the scan simply do not look for lesions in the scans.

In recent years, AI has gained popularity in diagnostic imaging to automatically recognizing complex patterns in imaging data and providing quantitative assessments of radiographic characteristics,[5] though its clinical applicability remains challenging. There are two main challenges that such automatic methods face. The first is the low prevalence of lesions in scans, which dictates very high precision to avoid prohibitive false alarm rates. The second challenge is caused by the diverse data acquisition conditions. Different acquisition conditions, such as different hardware and imaging protocols, results in images with different bit depths and a variety of artifacts.[6] These differences, although insignificant for the human expert, pose a great challenge for any computerized analysis methods which are in general very sensitive to these changes.[7] In addition, each radiograph consists of a large background area that corrupts the image statistics. Therefore, classic contrast enhancement methods fail in producing high-contrast detailed images.

In this work, we overcome these challenges by directly addressing the diversity resulting from the acquisition conditions; we explicitly enhance important anatomical features while eliminating background artifacts, thus extracting more knowledge from the limited dataset.

To summarize, in this work we notice the diversity of clinical radiographs, which is exhibited in narrow histograms and a poorly allocated dynamic range, thus having low contrast. We introduce a simple foreground-aware histogram equalization (HE) to overcome this challenge and enhance the important signal in the scanned image. We further showcase the effectiveness of our method by developing a deep learning model for the classification and detection of bone lesions, as an example of such analysis methods. We show how our preprocessing method significantly boosts the performance of our lesion detection model, given the same training and evaluation data.

2 Related Work

In this section, we discuss related work. First, we review previous studies of bone lesion analysis via deep learning algorithms. Next, we provide background on classic contrast enhancement methods that are commonly used in medical imaging. Finally, we provide a brief overview of vision transformers (ViTs) related to the way we utilize them in this work.

2.1 Bone Cancer Analysis on Radiographs

Several studies have used deep learning models to analyze bone tumors on radiographs. Yu He et al.[8] used a deep learning model to classify primary bone tumors in a multi-institutional dataset; however, the images were cropped by musculoskeletal radiologists to highlight the tumor before being inputted into the network, thus limiting the applicability of their method to be used for scanning arbitrary radiographs obtained for other clinical reasons. von Schacky et al.[9] used a multi-task deep learning model for simultaneous bounding box placement, segmentation, and classification of primary bone tumors on radiographs. However, they only considered radiographs of patients that had received a diagnosis of either benign tumors or malignant tumors, thus limiting the clinical applicability in HMOs and EDs. Neither of these studies coped with the challenge of obtaining high precision while avoiding false alarms. In addition, von Schacky et al.[9] ignored the diversity of the data, whereas our method explicitly accounts for at least some aspects of the diversity, resulting in two benefits: reducing the diversity that the model sees and better utilization and efficiency of the training data.

2.2 Contrast Enhancement Methods

The information inherent in the image histogram can be quite useful for different image processing applications as it provides important image statistics. Existing contrast enhancement methods, such as HE and contrast limited adaptive histogram equalization (CLAHE), use a transformation function based only on information available in the histogram of the input image, to achieve an effect of high contrast detailed image. HE modifies the pixels values in a way that the intensity level histogram of the equalized image spans a wider range of the intensity scale, thus resulting in enhanced contrast.[10] HE is useful for increasing the image's global contrast, though it has limitations as it does not consider any high-level information in the image during the equalization process. Adaptive histogram equalization is a contrast enhancement method that demonstrated excellent results on both natural and medical images.[11] It differs from HE in that it computes several histograms corresponding to different sections of the image, which makes it suitable for improving the local contrast. CLAHE[12] is a variant of adaptive histogram equalization in which the contrast amplification is limited to reduce the problem of noise over-enhancement, which is typical in adaptive histogram equalization.[11] However, clipping level must vary with the imaging modality, body region imaged, and imaging variables, thus limiting its robustness. We use these methods as preprocessing baselines in our experiments, and we show that using, our diversity-driven preprocessing method, the performance of an off-the-shelf lesion detector improves substantially.

2.3 Vision Transformers

ViTs[13] are recent popular deep learning architectures for image analysis that often outperform state-of-the-art convolutional neural networks in terms of accuracy and computational efficiency. In ViTs, an image is processed as a sequence of non-overlapping patches and a class token that serves as the global representation of the image. The tokens are passed through L transformer layers in which each layer is based on a self-attention mechanism[14] that computes the attention between patches.

DINO-ViT[15] is a ViT that has been trained in a self-supervised manner using a self-distillation approach with no labels. It learns powerful and semantically meaningful representations. The class token serves as the global representation of the image, and the attention maps illustrate that the model automatically learns class-specific features that lead to self-supervised foreground segmentation. It has been shown in several applications that the class token attention map of trained DINO-ViT models provides good results in foreground detection. See, for example, Ref. 16. In this work, we take advantage of this property and combine it with an existing contrast enhancement method to obtain enhanced images while avoiding noise amplification, thus improving the performance of a simple object detector.

3 Method

This retrospective study was approved by the institutional review boards (IRB) of both the Rambam Health Care Campus (HCC study approval number RMB-0663-18) and the Weizmann Institute of Science IRB committee.

The inherent diversity of radiographs degrades the performance of lesion detection algorithms. Our method directly addresses the inherent diversity of radiographs caused by acquisition conditions as exhibited by a poorly allocated dynamic range. We apply foreground-based histogram equalization (HE) in a way that only relevant pixels are considered in the enhancement process. In this way, we emphasize the important signal only, which is important when dealing with limited training data. Foreground masks are generated in an automatic manner that demonstrates robustness to different image intensity acquisition characteristics and imaged organs. Figure 1 shows an overview of our method. We propose a method that takes a diverse dataset as an input [Figs. 1(a) and 1(b)] and extracts the foreground masks in a self-supervised manner via self-attention [Fig. 1(c)]. We use the masks to perform HE based on the distribution of relevant pixels only, resulting in enhanced images as shown in Figs. 1(d) and 1(e). The enhanced images are used as the input of an off-the-shelf object detection algorithm that outputs lesion detection, as seen in Fig. 1(f). We perform the enhancing process to both the training and test datasets. Using the diversity-driven method, we benefit twice; we explicitly address the diversity resulting
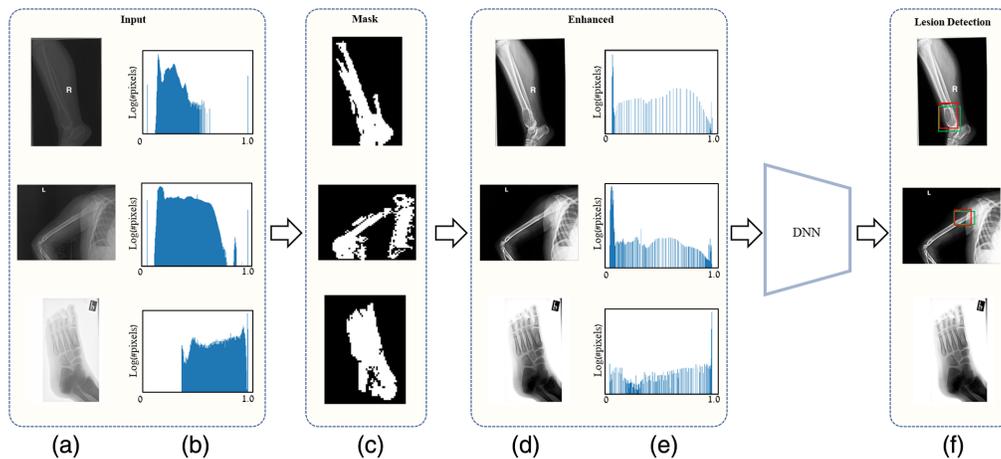
Fig. 1 Overview: our method takes in low contrast images (a) with a poorly allocated dynamic range (b) due to diverse acquisition conditions and extracts ROI segmentation (c) using self-attention maps. We use the ROI masks to generate enhanced images (d) and (e) that highlight the ROI and eliminate background artifacts. This process is applied to every image during training and testing. The enhanced images are used as input for the deep neural network (DNN), that outputs bone lesions detection (f).

from acquisition conditions, thus extracting more knowledge from the limited training data to achieve a high recall at a low false-positive rate.

In this section, we first provide background on self-supervised foreground detection using ViT's. Next, we describe our proposed enhancement method, which is applied to both the train and test sets. Finally, we detail the automatic lesion detection algorithm and evaluation metrics.

3.1 Foreground Detection Using DINO-ViT

As can be seen in Fig. 2, each radiograph consists of a large background area that corrupts the image statistics. Thus, we use the meaningful representations DINO-ViT to separate the object from the background. We show that, surprisingly, simply using ImageNet pretrained DINO-ViT to extract the attention maps leads to successful foreground detection on a completely different domain such as medical imaging; see Figs. 4 and 5. Figure 3 illustrates the attention heads and the cls attention of a 19-year-old male with a benign tumor at the shoulder. The cls attention detects the foreground using an off-the-shelf model and pretrained weights, even though no training or finetuning was involved in the process.

Figure 3 also demonstrates that it is possible to detect specific organs by applying mathematical manipulations to different heads instead of the cls attention, as each head corresponds to a different body part of the patient.
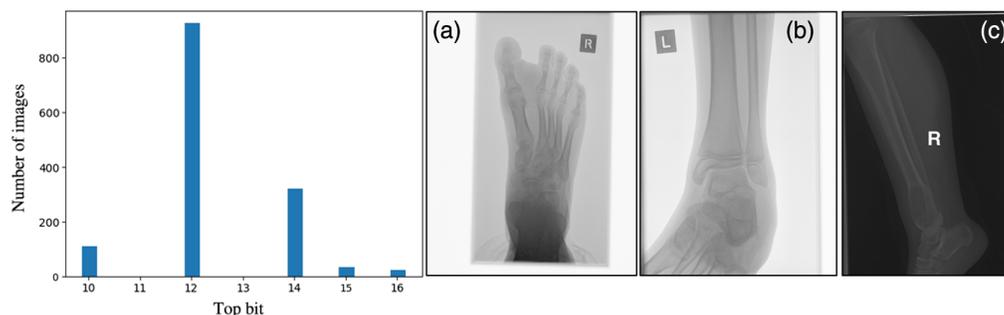

Fig. 2 Dataset Characteristics. Left: data distribution. Right: low contrast images. (a) 39-year-old male with a malignant tumor on the foot. (b) 12-year-old male with a malignant tumor on the ankle. (c) 13-year-old female with a malignant tumor on the tibia.
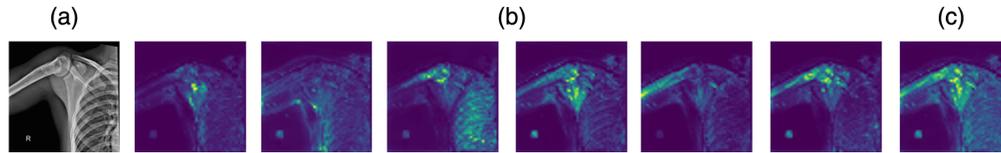
(a)   (b)   (c)

Fig. 3 Attention heads and cls attention map extracted from the last layer of DINO-ViT/S-8. We show that, even though DINO-ViT/S-8 was trained on natural images, it impressively detects the different organs in X-ray images. (Radiograph of a 19-year-old male with a benign tumor at the shoulder). (a) Input, (b) attention heads 0 to 5, and (c) avg [cls] att.

3.2 Enhancing Process

Contrast enhancement methods adjust the image contrast while considering the intensity level distribution of the entire image. Contrast enhancement based on only the region of interest (ROI) intensity levels, especially in images in which the ROI is relatively small, can improve the enhancement process while avoiding noise amplification. Figure 4 illustrates three input images and their corresponding poorly allocated dynamic range resulting from acquisition conditions. For each image in Figs. 4(a)–4(c), the image in the top row is the original radiograph given by the clinical collaborator, and the other rows artificially demonstrate non-ideal dynamic range conditions. We show that the class token attention map not only detects the object successfully in standard radiographs (as discussed in Sec. 3.1) but also demonstrates robustness to diverse imaging conditions including contrast inversion, shifting, and scaling (Fig. 4 rows 2 to 8). It is also robust to the scanned organs as each input image [Figs. 4(a)–4(c)] represents different organ. Finally, we take the foreground mask and apply HE based on the distribution of the ROI intensity levels. The full enhancing process is illustrated in Fig. 5. This process is applied to each input X-ray scan (namely, both the train and test sets), resulting in 16-bit enhanced images.

The proposed enhancing method is quite general and can be applied for different medical imaging applications and tasks. We demonstrate our method on two tasks – abnormality classification and detection of bone tumors on radiographs.
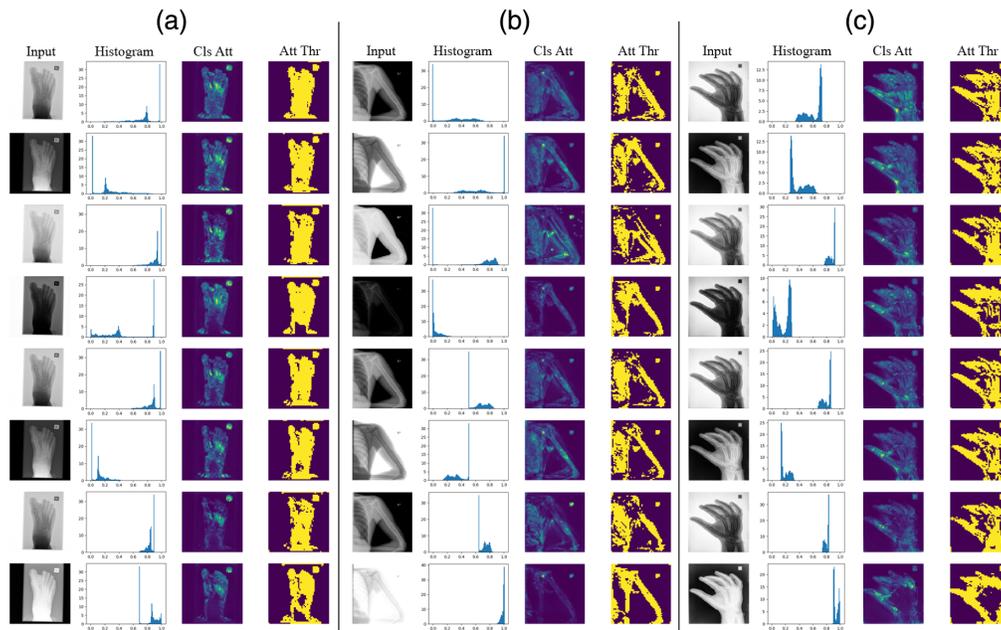

Fig. 4 ROI segmentation. From left to right: input image, histogram, class token attention map, thresholded attention map (70th percentile). We artificially demonstrate the robustness of DINO-ViT to diverse input images (first row) using different augmentations-contrast inversion, scaling, shifting, etc. (rows 2–8). (a)–(c) DINO-ViT robustness to different scanned organs: (a) 39-year-old male with a malignant tumor in his foot. (b) 26-year-old male with a benign lesion in his shoulder. (c) 23-year-old male with a benign lesion in his hand.
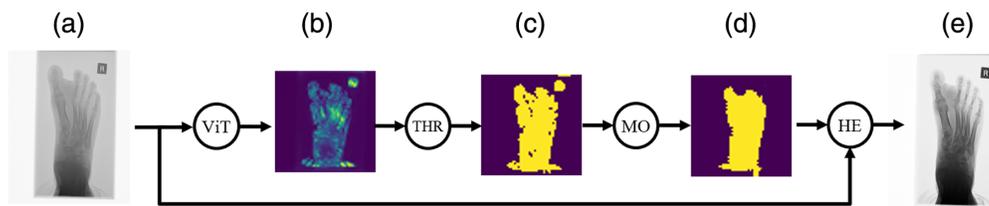
Fig. 5 Full contrast enhancement process: Input images (a) are fed to DINO ViT-S/8 to obtain attention maps (b). Foreground binary masks (c) are generated by employing 70th percentile thresholding on the attention maps. Blobs are removed (d) by applying morphological operations (MO). Finally, HE is applied on the input images based on the distribution of the foreground intensity levels only to yield enhanced images (e). The process is applied to each input X-ray scan.

4 Implementation Details

The following sections describe the general characteristics of our clinical dataset and the implementation details of an off-the-shelf algorithm used for both bone tumor classification and detection tasks.

4.1 Data Characteristics

The clinical dataset was acquired at the Medical Imaging Division, Rambam Health Care Campus, Haifa, Israel. As shown in Table 1, this retrospective study analyzed bone tumors on 1421 radiographs (mean age 23 years ± 8 [standard deviation]; 1001 men) obtained between January 2003 and December 2018, 973 of which are normal images with no pathology or any clinical finding and 448 abnormal images that include benign or malignant lesions. Radiographs included in this study were acquired from patients aged 5 to 40 years. Radiographs with poor image quality were excluded. In addition, postoperative scans, chest and abdominal scans, and scans that were non-relevant for diagnosis were also excluded. Participants were identified based on International Classification of Diseases code in the electronic medical record. Imaging data was reviewed by two senior radiologists and one radiology resident to establish the ground truth.

The data was acquired over a decade with a variety of X-ray machines and imaging protocols, resulting in a diverse dataset in terms of image characteristics. Image quality is determined by several factors including spatial resolution, patient movement, and the technician, which have an important role in maintaining good image quality.[17] Figure 2 (left) shows that the dataset consists of images with different bit depths – representing significant variations in the captured dynamic range. Moreover, in some cases, the dynamic range is not fully utilized, which means that the dataset consists of low-contrast images, as shown in Fig. 2 (right). In addition, the dataset not only varies in terms of intensity characteristics but also heterogeneously in terms of the scanned organs.

4.2 Data Preparation

The radiology team classified each tumor-containing radiograph as depicting benign or malignant tumors and additionally performed bounding box placement of the tumors. Images with no

Table 1 Radiographs Characteristics: age expressed as mean mean $\pm$ standard deviation.

Characteristic	Overall ($n = 1421$)	Training set ($n = 1148$)	Test set ($n = 273$)
Age (y)	23 ± 8	23 ± 9	24 ± 8
Male	1001	812	189
Normal	973	774	199
Abnormal	448	374	74
Benign	137	122	15
Malignant	311	252	59

pathology were classified as normal images. The dataset was split into training and testing sets, composed of 80% (1148/1421) and 20% (273/1421), respectively. The division was carried out to avoid any overlap of patients between the training and testing datasets, a crucial step in maintaining the validity of our study's findings.

4.3 DINO-ViT for ROI Segmentation

We use ImageNet pre-trained DINO ViT-S/8 with overlapping patches (stride = 4)[16] to extract the self-attention maps from the last transformer layer. We employ 70th percentile thresholding to the average attention map to generate an ROI binary mask. Morphological operations were used to clean up blobs in the binary mask, and the largest object, which represents the scanned organ, was extracted using connected components.

4.4 Lesion Detection

We train an off-the-shelf object detection algorithm; we use Detectron2's[18] implementation of Faster R-CNN[19] model using R101-FPN as a base model. Both pixel-level and spatial-level augmentations were applied using Albumentations library.[20] The loss functions used for bounding box placement and classification are regression loss and cross-entropy loss, respectively. The model outputs one or more bounding boxes, and each bounding box has a class label and a confidence score. Output with no bounding boxes is considered to be a normal image with no bone lesions.

5 Results

5.1 Statistical Analysis

Because some benign bone lesions may require treatment (complete local excision/curettage) and should be observed by a specialist, we evaluate the model by binary classification into abnormal (malignant or benign) and normal radiographs. For model evaluation, we used the receiver operating characteristic (ROC) curve and the area under the curve (AUC). For bounding box placement, we used accuracy (detected instances out of total instances) and intersection over union (IoU) similar to Ref. 9. The correct placement of the bounding box was assumed for $IoU > 0.5$. We evaluated on a set of 273 radiographs including benign and malignant tumors and normal images with no lesions or any other abnormality. All statistical analyses were performed using scikit-learn, version 0.22.1.[21]

5.2 Baselines

We compared the performance of an off-the-shelf object detection algorithm using different preprocessing methods. The first one is a naïve conversion of our diverse dataset to 8-bit images by simply dividing each image by its bit depth and multiplying by 255, similar to a prior study.[9] Other preprocessing methods that we examined included applying 16-bit HE and contrast limited adaptive histogram equalization (CLAHE) on the 8-bit images. For each of these baseline preprocessing methods, we trained a lesion detection network using the configuration described in Sec. 4 and compared the performance of the resulting detectors.

5.3 Abnormality Classification Results

The ROC curve is shown in Fig. 6; it gives the trade-off between the false positive rate (FPR) and the true positive rate (TPR) for different classification threshold values. Because bone lesions incidence among the population is low,[2] we focus on the low FPR area of the ROC curve (in this case 1.5%) to avoid prohibitive false alarm rates. Table 2 gives an overview of the performance of the off-the-shelf object detector on the test set at 1.5% FPR. Data in brackets represents 95% confidence interval (CI). We allow only four false alarms across the entire test set; under this constraint, the deep learning model obtains 82.43% sensitivity with our diversity-aware preprocessing method, compared with 68.91% and 55.4% sensitivity with HE and CLAHE, respectively. The deep learning model also obtains 55.4% sensitivity at zero FPR with naïve conversion to 8-bit images. Both naïve conversion and CLAHE include int8 quantization, meaning that information loss due to data quantization degrades the model performance. The AUC for our method is 0.98, which is the highest AUC out of all preprocessing baselines that were considered
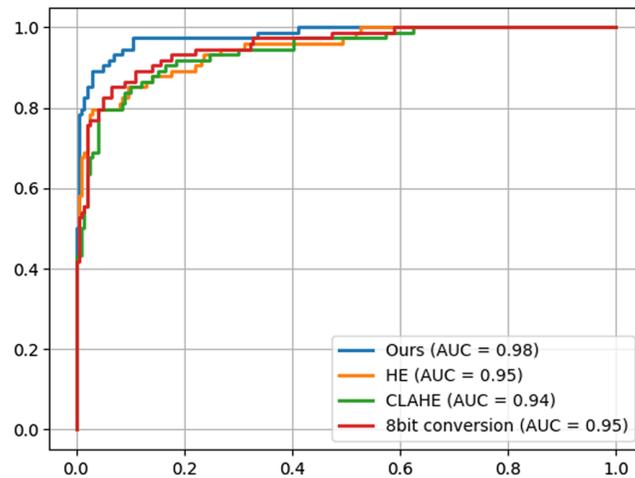
Fig. 6 ROC curve for normal/abnormal classification.

Table 2 Classification results: metrics comparison of Faster R-CNN trained with various preprocessing methods. Data in brackets are 95% CIs.

Preprocessing method	AUC	Sensitivity @1.5% FPR
Ours	0.98	82.43% [77.9, 86.9]
HE	0.95	68.91% [63.4, 74.4]
CLAHE	0.94	55.40% [49.5, 61.3]
8 bit Conversion	0.95	55.40% [49.5, 61.3]

in this study. Our method surpasses all others across the board as our method efficiently overcomes the diversity in the data using the proposed preprocessing phase.

5.4 Tumor Detection Results

Table 3 gives an overview of the detection performance comparison. Our model placed 82.5% (66 out of 80 instances; 95% CI: 74.17, 90.83) of the bounding boxes correctly (IoU > 0.5) and demonstrates an IoU of 0.69 ± 0.11. Our method surpassed the baselines in detection accuracy as both HE and CLAHE obtained 81.25% accuracy, whereas naive 8-bit conversion obtained only 77.5% accuracy. In terms of IoU, naive 8-bit conversion obtained the highest result (0.7 ± 0.10) out of all of the preprocessing methods.

Figure 7 demonstrates detection results of a 13-year-old female with a malignant tumor of the tibia. The model placed the bounding box correctly (IoU > 0.5) and classified the tumor

Table 3 Lesion detection results: note the correct placement of the bounding box assumed for IoU > 0.5. Data for accuracy is expressed as detected instances out of total instances. Data listed in brackets are 95% confidence intervals (CIs). IoU expressed as mean ± standard deviation.

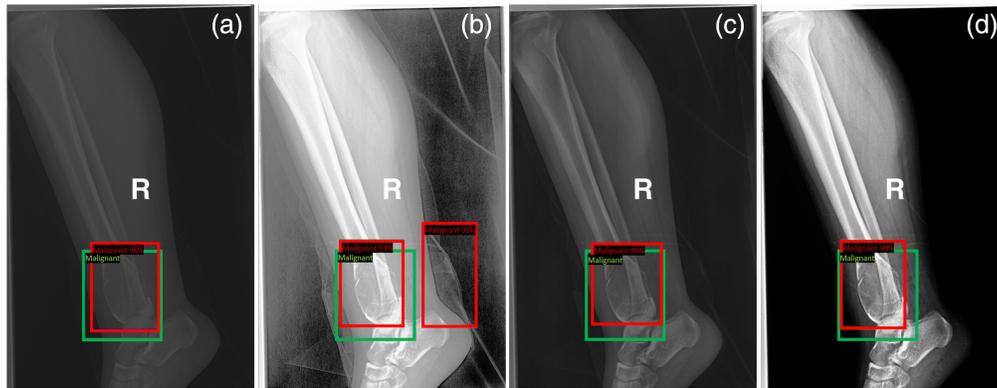| Method	Ours	HE	CLAHE	8 bit	v. Schacky et al.[9] Internal test set	v. Schacky et al.[9] External test set
Accuracy	82.50%	81.25%	81.25%	77.50%	65.0%	59.50%
	(66/80)	(65/80)	(65/80)	(62/80)	(91/140)	(66/111)
	[74.1, 90.8]	[72.7, 89.8]	[72.7, 89.8]	[68.3, 86.6]	[57.1, 72.9]	[50.3, 68.6]
IoU	0.69 ± 0.11	0.69 ± 0.09	0.68 ± 0.10	0.70 ± 0.10	0.54 ± 0.32	0.52 ± 0.34

Fig. 7 Detection examples of a 13-year-old girl with a malignant tumor with different preprocessing methods. (a)–(d) Eight-bit naïve conversion, HE, CLAHE, our diversity aware method. Using HE, the model detects a background artifact as a malignant tumor with high confidence. Our method (d) highlights the ROI while eliminating those background artifacts. Note that the ground truth (GT) is presented as a green bounding box with the class name at its top left corner and prediction is presented as a red bounding box with the class name and confidence score at the top left corner.

correctly as a malignant tumor in all preprocessing methods. However, using HE, the model detects background artifacts as a malignant tumor with a 90% confidence, meaning that in some cases HE can overamplify noise by enhancing unwanted artifacts instead of enhancing the region of interest. Our method eliminates background artifacts caused by objects external to the patient (e.g., clothing) in addition to region of interest contrast enhancement. Using our method, we lose some of the patient's soft tissue, but it does not affect bone tumor detection. Figure 8 demonstrates the detection results of an 8-year-old girl with a malignant tumor of the humerus. As can
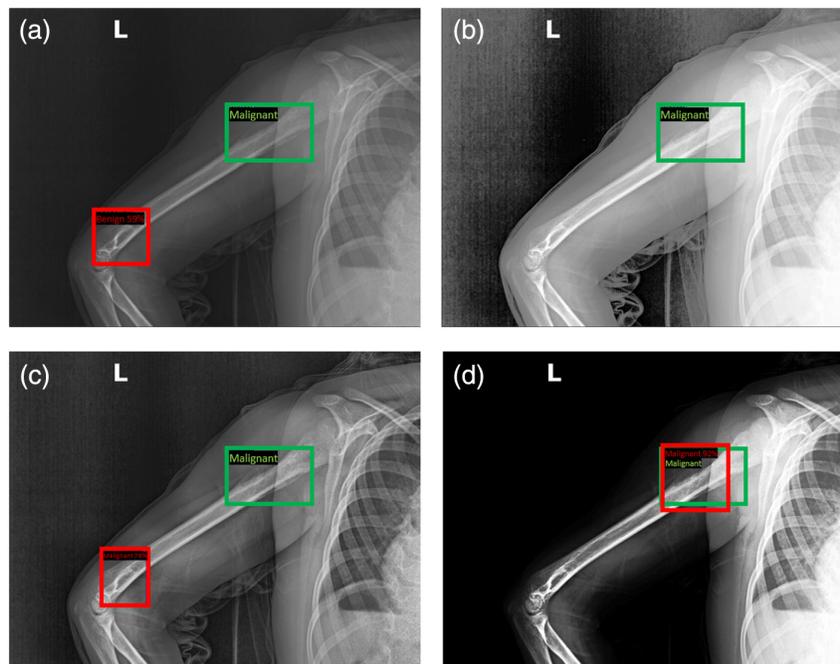

Fig. 8 Detection examples of an 8-year-old girl with a malignant tumor with different preprocessing methods. (a)–(d) Eight-bit naïve conversion, HE, CLAHE, our diversity aware method. Using 8-bit naïve conversion and CLAHE, the model incorrectly detects benign and malignant tumors respectively with high confidence; using our method, the model managed to successfully detect (IoU > 0.5) the tumor. Note that the GT is presented as a green bounding box with the class name at its top left corner and prediction is presented as a red bounding box with the class name and confidence score at the top left corner.

be seen in Fig. 8, the model successfully detects the tumor only using our diversity-aware pre-processing method. Using naïve 8 bit conversion and CLAHE, the model detects benign and malignant tumors, respectively, with confidence >50% incorrectly, as the bounding box placement is completely off. Our method can be adapted across a spectrum of clinical applications, by setting different kernels for the morphological operations or using another attention map to extract the ROI segmentation; averaging specific attention heads, exploring the attention heads in the different transformer layers.

By averaging and analyzing specific attention heads across different layers, we enhance the model's ability to focus on clinically relevant features, thereby improving the accuracy and relevance of the ROI segmentation. This granular control over both the spatial processing and attention-guided feature extraction aspects of our method ensures a tailored approach to image analysis, enabling the extraction of clinically significant insights with high precision.

A previous study by von Schacky et al.[9] ignored the diversity of the data and trained a deep learning model using the same used backbone (ResNet101) only on 8-bit images. They used a different dataset, which included both the internal and external datasets (data from a different university hospital in the same country) and obtained 65% and 59.50% detection accuracy and IoU of 0.54 ± 0.32 and 0.52 ± 0.34, respectively. In contrast to Ref. 9, our method enables training on 16-bit images and therefore avoids the quantization error resulting from conversion to 8-bit.

5.5 Runtime Analysis

We conducted a comprehensive assessment of the computational demands of each stage in our enhancing process. The analysis results in Table 4 reveal the model's runtime performance for key preprocessing steps – cls token attention map extraction, cleaned mask extraction, and contrast enhancement – across different GPUs. These metrics are critical, considering the necessity for rapid processing in a clinical context. The Nvidia A100 GPU, representing the pinnacle of

Table 4 Runtime performance comparison for various enhancement process stages-cls token attention extraction, mask extraction, and contrast enhancement-across different GPU models and configurations. The reported runtimes are in seconds for an input image size of 2000×1000 pixels.

GPU	ViT type	Attn map extraction	Mask extraction	Contrast enhancement
		Runtime (s)	Runtime (s)	Runtime (s)
Nvidia A100	Small	1.8667	0.0375	0.0284
	Base	2.1323	0.0918	0.0284
Tesla V100	Small	2.1397	0.0382	0.0364
	Base	2.1745	0.1142	0.0336
RTX2080	Small	1.7547	0.0597	0.0285
	Base	2.8636	0.1548	0.0279

Table 5 Detectron2 inference runtime performance across different GPU nodes. Detectron2 configuration is described in Sec. 4.4. The reported runtimes are in seconds for an input image size of 2000×1000 pixels.

GPU	Detectron2 inference time runtime (s)
Nvidia A100	2.9473
Tesla V100	3.9367
RTX2080	4.2338

current technology, processes feature extraction in 1.8667 s for the "small" configuration of the ViT and 2.1323 s for the "base" configuration. This demonstrates the model's capacity to integrate seamlessly into clinical workflows without substantial delays. The Tesla V100 and RTX2080 GPUs also deliver competitive runtimes, ensuring that the model's deployment is feasible even in clinical settings with varying levels of computational resources.

Detectron2 inference times range from 2.9473 to 4.2338 s for an input image size of 2000×1000 pixels, as shown in Table 5. The overall runtime of our method showcases its adaptability to varying hardware configurations while maintaining rapid processing speeds. This versatility ensures that our approach remains applicable across a spectrum of clinical settings, for which computational resources may vary, without necessitating extensive infrastructure upgrades. Together, these results affirm the feasibility of integrating our methodology into real-world clinical workflows.

6 Conclusion

"It turns out [that when] you take [...] that same AI system, [...] and the technician uses a slightly different imaging protocol, that data drifts to cause the performance of AI system to degrade significantly"

Andrew Ng[7]

In this work, we directly addressed the issue of a data drift caused by different acquisition conditions (e.g., imaging protocols and scanning machines). We introduced a method that explicitly address this issue and gain a robust framework capable of processing data from different clinical sites and imaging equipment to achieve high sensitivity while maintaining low false-positive rate, thus making it suitable for clinical setups. Our method allows for treating images with different bit depths in a unified manner while utilizing the full dynamic range. Explicitly enhancing the ROI makes the finetuning process of the deep learning model efficient and robust to acquisition parameters. We showcase the effectiveness of our method by developing a deep learning model for the classification and detection of primary bone tumors. Using the same analysis algorithm, we show how our preprocessing method significantly boosts performance compared with existing enhancement methods.

Our study has several limitations. First, the model was not trained to detect other diseases such as fractures or osteoporosis. To make it clinically applicable, it should be trained on a dataset that consists of radiographs with other abnormalities in addition to bone lesions. Second, in the general population, benign tumors are more common than malignant ones, yet our dataset contained a smaller number of benign bone tumors than malignant bone tumors, indicating potential bias. Third, our study did not consider patient age. Considering age may provide additional information to the model and improve its performance.

In conclusion, a deep learning model for detecting bone lesions can be beneficial in HMOs and EDs for analyzing radiographs that are not usually assessed by radiologists. The detection performance of the model surpassed the previous study by von Schacky et al.[9] The proposed method improves the robustness of the deep learning model and its ability to generalize across medical images acquired under diverse conditions and protocols as the classification performance using this method surpassed that of other preprocessing methods. A framework that extracts more knowledge from the data is extremely important, especially in domains that lack training data such as medical imaging. In a future study, the proposed method should be demonstrated on different tasks to inspect its impact on deep learning algorithm performance.

Code and Data Availability

Acknowledgments

References

1. T. M. Jackson, M. Bittman, and L. Granowetter, "Pediatric malignant bone tumors: a review and update on current challenges, and emerging drug targets," *Curr. Probl. Pediatr. Adolesc. Health Care* **46**, 213–228 (2016).
2. R. L. Siegel et al., "Cancer statistics, 2022," *CA: Cancer J. Clin.* **72**, 7–33 (2022).
3. M. Salom et al., "Diagnosis and staging of malignant bone tumours in children: What is due and what is new?" *J. Children's Orthopaedics* **15**, 312–321 (2021).
4. D. N. Mintz and S. Hwang, "Bone tumor imaging, then and now," *HSS J.* **10**, 230–239 (2014).
5. X. Tang, "The role of artificial intelligence in medical imaging research," *BJR Open* **2**, 20190031 (2020).
6. A. Shetty et al., "X-ray artifacts," 2014, Radiopaedia.org (accessed 13 March 2024).
7. T. S. Perry, "Andrew Ng x-rays the AI hype," *IEEE Spectrum: Technology, Engineering, and Science News* (2021).
8. Y. He et al., "Deep learning-based classification of primary bone tumors on radiographs: a preliminary study," *eBioMedicine* **62**, 103121 (2020).
9. C. E. von Schacky et al., "Multitask deep learning for segmentation and classification of primary bone tumors on radiographs," *Radiology* **301**(2), 398–406 (2021).
10. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed., Prentice-Hall, Inc. (2008).
11. S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Comput. Vis. Graph. Image Process.* **39**, 355–368 (1987).
12. K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graph. Gems* **0**, 474–485 (1994).
13. A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," CoRR abs/2010.11929 (2020).
14. A. Vaswani et al., "Attention is all you need," CoRR abs/1706.03762 (2017).
15. M. Caron et al., "Emerging properties in self-supervised vision transformers," CoRR abs/2104.14294 (2021).
16. S. Amir et al., "Deep vit features as dense visual descriptors," arXiv:2112.05814 (2021).
17. A. Tafti and D. W. Byerly, *X-ray Image Acquisition*, StatPearls (2022).
18. Y. Wu et al., "Detectron2," https://github.com/facebookresearch/detectron2 (2019).
19. S. Ren et al., "Faster r-CNN: towards real-time object detection with region proposal networks," CoRR abs/1506.01497 (2015).
20. A. Buslaev et al., "Albumentations: fast and flexible image augmentations," *Information* **11**, 125 (2020).
21. L. Buitinck et al., "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Lang. for Data Mining and Mach. Learn.*, pp. 108–122 (2013).

Biographies of the authors are not available.