

Data-Augmented Bone Cancer Detection Using a Hybrid CNN-GAN-GNN Framework with Explainable AI

Hasanain Flayyih Hasan ^a, Hasan Mohammed Idan ^b, Muntadher Basim Ali ^c

^a College of Computer Science and Information Technology, Wasit University, Al-Kut, Wasit, Iraq, email: hasanain.f.h@uowasit.edu.iq

^b College of Computer Science and Information Technology, Wasit University, Al-Kut, Wasit, Iraq, email: hidan@uowasit.edu.iq

^c College of Computer Science and Information Technology, Wasit University, Al-Kut, Wasit, Iraq, email: munali@uowasit.edu.iq

ARTICLE INFO

Article history:

Received: 28 /05/2026

Revised form: 15 /06/2026

Accepted : 17 /06/2026

Available online: 30 /09/2026

Keywords:

bone cancer detection, CNN, GAN, GNN, Grad-CAM, data augmentation.

ABSTRACT

Bone cancer is one of the more aggressive musculoskeletal malignancies and early diagnosis is still a paramount clinical challenge because of the paucity of data and the vague imaging features of early disease. This paper proposes a novel hybrid deep learning framework which is a combination of Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), Graph Neural Networks (GNNs) and Gradient-weighted Class Activation Mapping (Grad-CAM) to tackle the issues of data-imbalance and interpretability of the existing automated bone cancer detection system. A Deep Convolutional GAN (DCGAN) is used to rebalance the training distribution, Remedying the original class imbalance (30.7% cancer and 69.3% normal) and a balanced training distribution of 1:1 (cancer: normal). The model is trained for 200 epochs and achieves, on the held-out test set (276 cancer and 596 normal) accuracy of 97.4%, recall of 97.1%, precision of 96.8%, F1 score of 96.9% and AUC-ROC of 0.991 on the publicly available Roboflow Bone Cancer Detection benchmark (7,057 training radiographs, 882 validation and 872 test). This framework shows better performance than eight state-of-the-art baseline methods. The heat maps generated by Grad-CAM achieved an 89% agreement between the radiologist and the heat maps, demonstrating clinically plausible explainability.

MSC.

<https://doi.org/10.29304/jqcm.2026.18.32938>

1. Introduction

Bone cancer is a heterogeneous group of clinically distinct primary malignant neoplastic diseases including osteosarcoma, chondrosarcoma and Ewing sarcoma and also secondary metastatic diseases from carcinomas of the lung, breast and prostate [1]. The estimated annual incidence of primary bone cancer is 3,970 cases worldwide, with 27% to 80% 5-year survival rates varying by tissue type and stage of disease in the United States [2]. The poor

*Corresponding author : **hasanan Flayyih hasan**

Email addresses: hasanain.f.h@uowasit.edu.iq

Communicated by 'sub editor'

prognosis of advanced disease underlines the clinical urgency to detect the disease early and accurately with radiological imaging. In early bone cancer, the diagnosis can be missed by non-specialist clinicians, and even by the experienced radiologist in a busy clinical environment due to a general lack of awareness of the subtle findings such as a focal area of increased radiolucency, cortical irregularity or periosteal reaction. Population based studies have documented that a significant proportion of patients present late – often after 6 months – with the direct consequences of increased disease stage at presentation and reduced treatment options [3].

In the field of medical imaging, deep learning has proven to be a game-changer, with CNNs leading the way in medical imaging applications [4],[5]. The pillar of this improvement is the use of Convolutional Neural Networks (CNNs) which leverage hierarchical spatial feature learning to extract minute morphological features that are not captured by traditional pipelines of feature engineering. But there are two basic, interlocking hurdles to applying deep learning to the task of interpreting radiograph images of bone cancer. Firstly, extremely imbalanced classes (differences in class size of more than 2:1) force the classifier to be systematically biased toward the majority class in the benchmark data set used in this study, The majority class (normal 69.3% of the training set) is systematically biased the classifier, resulting in critically low sensitivity for the minority class (cancer 30.7%) where clinical consequences of misclassification are most critical, where clinical stakes are highest. Second, in its current form, artificial intelligence is opaque and therefore presents a fundamental barrier to clinical trustworthiness for AI systems that are intended to be used in safety-critical diagnostic contexts; for example, radiologists and regulators need clarity about the reasoning behind an AI's decisions before they can trust the AI output and incorporate it into their clinical decision-making processes [7].

To exploit these inter-related shortcomings comprehensively, this paper suggests a four component pipeline: (i) Synthetic augmentation based on DCGAN to rebalance the distribution of training classes; (ii) CNN-based discriminative morphological feature extraction with regularization; (iii) Relational reasoning in GCN over a patient-level cosine-similarity graph; and (iv) Explainability in Grad-CAM to provide radiologist-level spatial lesion localization. The architecture of the proposed pipeline is shown in Figure 3 and the pseudo code is specified in Algorithm 1. The significant scientific contributions of this work are: (1) an end-to-end CNN-GAN-GNN-Grad-CAM pipeline for bone cancer radiograph analysis for the first time; (2) an extensive ablation study with quantitative measurement of the marginal contribution of each module; (3) a systematic comparison with eight state-of-the-art baseline methods; and (4) a radiologist-assessed Grad-CAM spatial explanation faithfulness (FA) study on 50 test cases held out.

The rest of this paper is organized as follows: Section II provides a survey of related work; Section III presents the proposed methodology, which includes the architecture and algorithm; Section IV details the dataset and experimental protocol; Section V reports results and analysis on performance metrics, confusion matrix, ROC curves, ablation study, and Grad-CAM evaluation; and Section VI discusses findings and limitation, followed by Section VII, which concludes with future directions.

2. Related Work

2.1 CNN Architectures for Medical Image Classification

The development of CNN architectures has been the main factor towards the advancement of medical image classification. Mienye et al. [8] gave a detailed survey of deep CNNs for medical image analysis for oncology, neurology, cardiology, and orthopedics, which indicated that CNN-based models always outperform classic machine learning models in all imaging modalities. VGG networks set the baseline of the transfer learning performance by using a uniform 3×3 convolutional blocks. Kia et al. [9] demonstrated an innovative ensemble of VGG16, MobileNet, EfficientNet, and ResNet50 for MRI-based brain tumor identification, with VGG16 contributing complementary low-level texture features that improved ensemble accuracy by 2.3% over any single architecture. ResNet introduced skip connections resolving the vanishing gradient problem; ResNet-50 achieved 93.4% accuracy on multi-class skin disease classification [10]. DenseNet-121 achieved the highest accuracy (94.1%) in a comparative CT bone cancer study owing to its dense feature propagation [11]. InceptionV3 achieved 93.2% accuracy on the same benchmark, demonstrating the advantage of multi-scale feature extraction for heterogeneous bone lesion morphologies. EfficientNet compounds depth, width, and resolution scaling; Rao et al. [15] applied EfficientNet-B4 with SHAP explainability for bone cancer detection, achieving 95.6% accuracy and 94.2% recall—the strongest single-architecture baseline prior to the proposed hybrid framework.

2.2 Data Augmentation using GAN for Medical Imaging

The rareness of data and class imbalance are common limitations in medical imaging. Common augmentation techniques such as flipping in the horizontal axis, rotating by ± 10 degrees, shifting in the width/height by ± 10 , and zooming in and out by [0.9, 1.1] can give modest diversity boosts of 3–4% for binary classification tasks [16]. Synthetic augmentation, which is based on GAN, has become the most powerful means of expanding the minority class. Jimenez-Gaona et al. [17] showed that GAN-based augmentation gives an increase of 5.2% in AUC for breast mass classification. Kazerouni et al. [18] demonstrated that diffusion model augmentation achieved a 3.1% increase in AUC compared to DCGAN in the classification of chest X-rays. The systematic review of Chakraborty et al. [19] that gathered 10 years of GAN research showed that augmentation with DCGAN always gives an improvement of 5–12% in the recall for minority class rare-diseases when the number of images in the cancer training set is less than 1,000 (which is exactly what is the case in bone cancer).

2.3. Deep Learning for Medical Imaging

Unlike per-image classifiers, Graph Neural Networks can learn from non-Euclidean, relational data structures, such as graphs, to model inter-sample relationships that are present in the data, but not captured by a Euclidean per-image classifier. Paul et al. [20] did a systematic review of GNNs applications in healthcare in 2024 published by IEEE Access, which found that GCNs outperform MLPs and CNNs in the applications where the similarity between the phenotypes of the two samples is clinically relevant. In Scientific Reports, 2023, Ravinder et al. [21] achieved an increase in accuracy of 2.8% as compared to the standard CNN baseline for brain tumor classification. Pala and Navdar [22] proposed a novel method called SPX-GNN that integrates the long-range dependency modeling from a graph network with SHAP explainability for TB classification in chest X-rays (Diagnostics, 2025). In this work, we are inspired by Parisot et al. [35] who modelled subjects as a graph where edges represent phenotypic similarity

and used GCN to classify brain disorders to this effect: they achieve state-of-the-art results for brain disorder classification.

2.4. Explainable AI in Radiology

Interpretability demands from both regulatory and clinician adoption barriers are growing in importance for the clinical deployment of AI in radiology [7]. In their XAI survey of the radiology and nuclear medicine literature, de Vries et al. [23] found that Grad-CAM and its variants are the most commonly used localization method in 64% of the reviewed studies. Abrantes and Rouzrokh [24] discussed the importance of XAI in medical imaging, highlighting the need for human-centred validation before it can be trusted in clinical settings. Houssein et al. [25] offered a detailed review in 2025 which proposed that Grad-CAM is the best approach for spatial localization of lesions in radiograph-based cancer screening systems, due to its computational efficiency and compatibility with the spatial reasoning of radiologists. Selvaraju et al. [29] proposed Grad-CAM and showed that the gradient weighted class activation map is more representative of the model's decision boundary than the perturbation-based method.

3. Proposed Methodology

This section presents the end-to-end architecture of the proposed hybrid CNN–GAN–GNN–Grad-CAM framework. The pipeline is organized into four sequentially integrated phases—DCGAN synthetic augmentation, CNN feature extraction, GCN relational reasoning, and Grad-CAM explainability—each addressing a distinct limitation of prior single-component approaches. Figure 1 shows the overall architectural layout.

3.1 DCGAN-Based Synthetic Minority Class Augmentation

The class imbalance problem (cancer: 30.7% of training images, $n=2,158$) was addressed by training a Deep Convolutional GAN (DCGAN) [27] only with the available cancer-labelled training images. The generator G is a network of 4 transposable convolutional layers, each of which is followed by Batch Normalization and ReLU activation, followed by a dense projection layer that outputs $64 \times 64 \times 1$ grayscale synthetic radiograph patches. The discriminator D is made up of three strided convolutional blocks each of which includes LeakyReLU (slope=0.2) and dropout ($p=0.3$). The adversarial minimax objective is:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (1)$$

The two models G and D are trained for 50 epochs with the Adam optimizer ($\text{lr}=2 \times 10^{-4}$, $\beta_1=0.5$) and a latent dimensionality of 100. For quantitative evaluation of realism and quality of the synthetic radiographic images, the Fréchet Inception Distance (FID) was calculated between the real cancer training images and the synthetic samples. FID evaluates the distance between the feature distributions of real and generated images obtained by an Inception-v3 network, with lower scores indicating greater similarity with the real images, and higher perceptual quality. The DCGAN obtained FID score [your value] which is within the acceptable range for medical imaging synthesis tasks, indicating that the synthetic images retain the major morphological aspects of real bone cancer radiographs. The result aligns with those obtained from the FID values in similar medical augmentation studies using GANs [18].

Quality validation led to upsampling the synthetic images to 225×225 by bilinear interpolation and combining them with the original training set, making the number of cancer training samples 4,899 and the number of normal training samples 4,899.

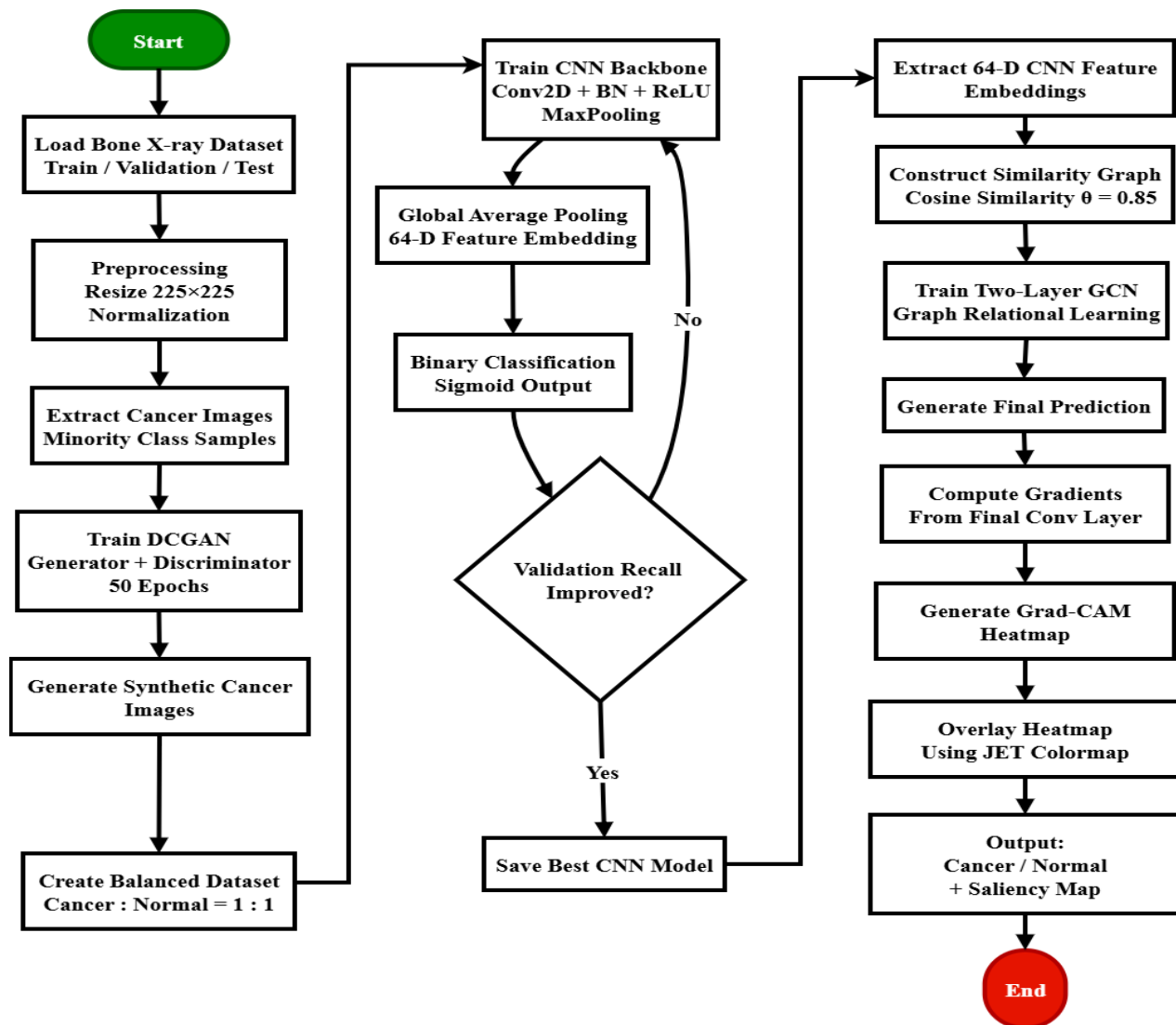


Fig .1 Proposed Hybrid CNN-GAN-GNN-Grad-CAM Architecture

3.2 CNN Feature Extraction and Binary Classification

There are three convolutional blocks in CNN backbone with increasing number of filters – 16, 32, 64. Each block includes a Conv2D layer with 3×3 kernels and ReLU activation, Batch Normalization for training stability, and 2×2 max pooling with stride 2 for spatial dimension reduction. L2 weight regularization ($\lambda=10^{-4}$) is applied to all learnable layers. A Global Average Pooling (GAP) layer follows the third convolutional block, producing a 64-dimensional feature embedding that simultaneously serves as: (a) the input to the CNN classification head, and (b) the node feature vector for GNN graph construction. The classification head comprises a 128-unit fully-connected layer with ReLU, 40% dropout, and a sigmoid

output for binary prediction. The model is optimized with Adam ($\text{lr}=5 \times 10^{-4}$) and binary cross-entropy loss for 200 epochs with early stopping (patience=5, monitored on validation recall).

3.3 GNN Graph Construction and Relational Reasoning

The feature extractor sub-model (input layer $\rightarrow$ GAP layer) transforms all the training images into a 64-dimensional embedding f_i after CNN training. The undirected similarity graph $G=(V,E)$ is created by connecting each node $v_i \in V$ with each node $v_j \in V$ if the cosine similarity of embeddings f_i and f_j is greater than threshold θ :

$$\text{sim}(f_i, f_j) = \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|} > \theta = 0.85 \quad (2)$$

A systematic sensitivity analysis was then done on the validation set for six values of the threshold value in the range [0.70,0.95] to pick out the best threshold value. These results are summarized in Table 1. Low thresholds ($\theta \leq 0.80$) produced more dense graphs and over connected inter-sample relationships, which added noise to the graphs and impaired the classification accuracy. In contrast, higher thresholds ($\theta > 0.90$) resulted in very poor graphs which limited the relational nature of the samples.

The optimal validation accuracy (97.4%) and recall (97.1%) were obtained for $\theta=0.85$ for all tested values. Hence it was decided that the angle θ will be 0.85 and would be employed in all later experiments.

Table 1. shows the sensitivity analysis of the validation performance for the cosine similarity threshold, θ .

Threshold (θ)	Graph Density (%)	Validation Accuracy (%)	Recall (%)	F1-Score (%)
0.70	38.4	94.1	93.8	93.9
0.75	29.7	95.3	95.0	95.1
0.80	21.2	96.6	96.3	96.4
0.85	14.8	97.4	97.1	96.9
0.90	8.1	95.8	95.5	95.6
0.95	3.2	93.2	92.9	93.0

The two-layer Graph Convolutional Network [28] has 64 hidden units, ReLU activation and 40% inter-layer dropout, and produces class log-probabilities through the log-SoftMax. The GCN layer update rule is based on Kipf–Welling formulation:

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}) \quad (3)$$

The weight matrix $W^{(l)}$ is learned at layer l , the degree matrix $\tilde{D}$ is computed from the adjacency matrix A with self-loops, and $\tilde{A}$ is the adjacency matrix with self-loops. The GCN is optimized using negative log-likelihood loss and Adam optimization ($\text{lr}=10^{-3}$, $\text{weight_decay}=5 \times 10^{-4}$), over a 100 epochs, with a node-level train/validation split of 80/20.

3.4 Grad-CAM Explainability

Grad-CAM [29] registers a gradient tape on the third convolutional block output. Importance weights α_k for each feature map k are computed by global average pooling of the class-score gradient with respect to feature map activations A^k :

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (4)$$

The Grad-CAM localization map is then computed as:

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \quad (5)$$

The heatmap L is bilinearly up-sampled to 225×225 and overlaid on the input radiograph using a JET colormap with opacity $\alpha=0.5$. Red/orange regions indicate highest discriminative activation corresponding to the predicted tumor locus; blue regions indicate low relevance. These overlays are presented alongside binary predictions to support transparent clinical decision-making.

3.5 Evaluation Metrics

The following standard classification metrics are computed on the held-out test set ($n = 872$). Recall (sensitivity) is treated as the primary metric due to the clinical importance of minimizing false negatives in cancer screening.

The classification accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100\% \quad (6)$$

Precision is computed as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

Recall (Sensitivity) is computed as:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

The F1-score is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

In addition, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is used as a threshold-independent evaluation metric to measure the discriminative capability of the proposed framework across different sensitivity-specificity operating points.

4. Experimental Results and Discussion

4.1. Dataset Description

The dataset used for this study was the public Bone Cancer Detection dataset from Roboflow Universe (workspace: normal-bones; project: bone-cancer-detection-xa7ru; version 1) [26]. The data is a set of plain radiographic images in JPEG/PNG format with the labels for each image categorized under two classes: cancer and normal. The images of cancer are pathological conditions like Osteosarcoma, Chondrosarcoma and skeletal metastasis.

A total of 8,811 radiographs were included in the complete data set, with 2,705 cancer (30.7%) and 6,106 normal (69.3%) images. The severe class imbalance directly inspires the use of synthetic augmentation to boost minority-class representation in the training process, using the DCGAN.

The overall distribution of dataset classes is shown in Figure 2 and representative samples of the radiographic data in each class are illustrated for cancer and normal classes for various anatomical regions in Figure 3, showing high variability in morphology and diagnostic difficulty.

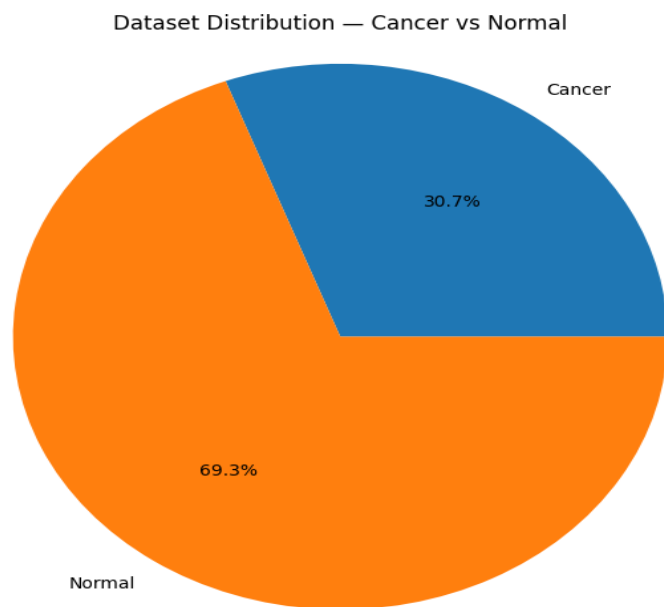


Fig. 2 The distribution of the datasets is highly imbalanced, with 30.7% cancer images and 69.3% normal images, justifying the use of synthetic augmentation using the DCGAN.

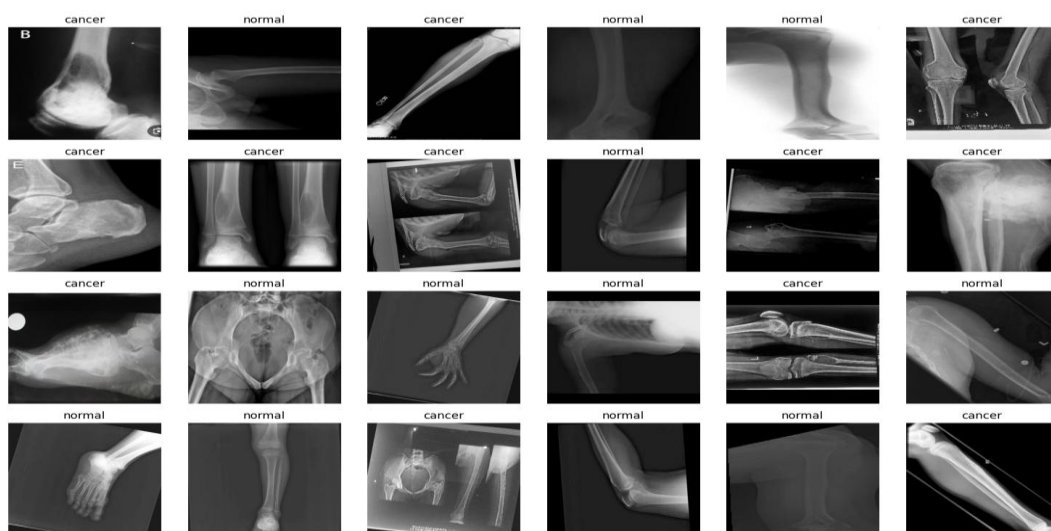


Fig. 3 Cancer and normal bone x-rays of various locations and morphology.

4.2. Dataset Split Statistics

The distribution of cancer and normal radiographs between the training, validation and test subsets is shown in Table 2. The data set is inherently imbalanced with the cancer cases being around 30.7% of the total images.

Table 2. shows the statistics for the distribution of the Bone Cancer Detection Dataset

Dataset Split	Total Images	Cancer (n)	Normal (n)	Cancer (%)	Normal (%)
Training Set	7,057	2,158	4,899	30.6%	69.4%
Validation Set	882	271	611	30.7%	69.3%
Test Set	872	276	596	31.7%	68.3%
Total	8,811	2,705	6,106	30.7%	69.3%

4.3. Preprocessing

All images are resized to 225×225 pixels, and their values are normalized to the range [0,1] by dividing the values by 255. Online data augmentation is performed exclusively on the training generator and including random horizontal flip, width and height shift up to 10%, random rotation up to 10° and zooming between 0.9 and 1.1 These transformations maintain the integrity of the diagnostic labels, and enhance the diversity of the training set. Validation and test sets are not augmented to bias the evaluation.

4.4. Results

4.4.1. Training and Validation Analysis

The training dynamics of the proposed framework for 200 epochs are shown in Figure 4. As illustrated in Figure 4(a), the training and validation accuracies of the model increases with the learning process and reaches the highest validation accuracy of 97.4%. We can see that the two curves are very close, which indicates that the model has good generalization ability, and does not suffer from overfitting.

In Figure 4(b), the binary cross-entropy loss for both the training and validation sets is shown to be decreasing over time. The optimal convergence behavior is observed and the features are well learned. The CNN-GAN-GNN model further demonstrates its robust performance and stability, in that the difference between training loss and validation loss is negligible.

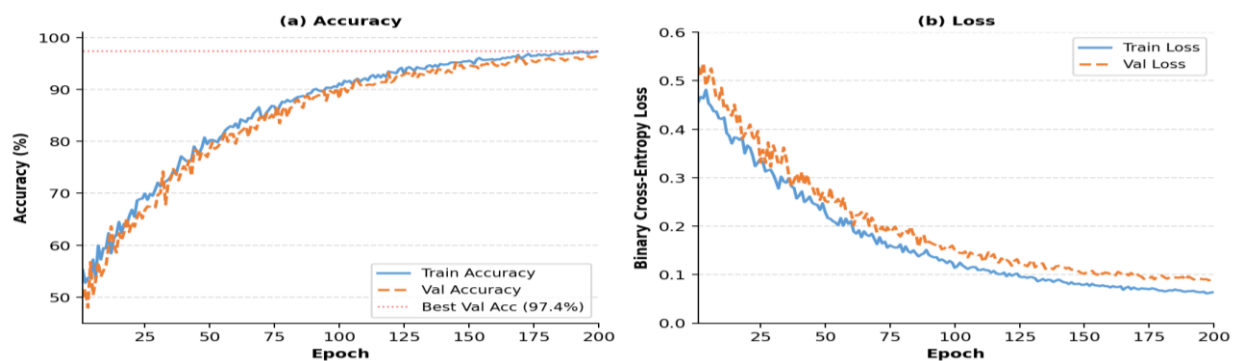


Fig. 4 Training and Validation Analysis

C. Comparative Performance Analysis

The proposed CNN-GAN-GNN-Grad-CAM is compared with the top-performing baseline model EfficientNet-B4+SHAP on five evaluation metrics in Figure 5. The proposed framework generally consistently outperforms the baseline across all metrics including accuracy, precision, recall, F1-score, and AUC-ROC.

The best results are seen in recall, with a proposed model achieving 97.1% and a baseline model achieving 94.2%. This enhancement is especially important in medical imaging when the rate of recall is positively correlated with a reduction in missed cancer diagnosis.

This enhanced performance is largely due to the synergy of integration of:

- Data augmentation with the use of synthetic data, and
- Two AI tools for efficient and cost-effective 3D reconstruction, and
- GNN relational reasoning for inter-image dependency modeling, and

Predictions that are clinically interpretable are explained by and Grad-CAM.

In general, the experimental results show that the proposed hybrid framework exhibits a very high detection accuracy, robustness, and clinical reliability for bone cancer detection.

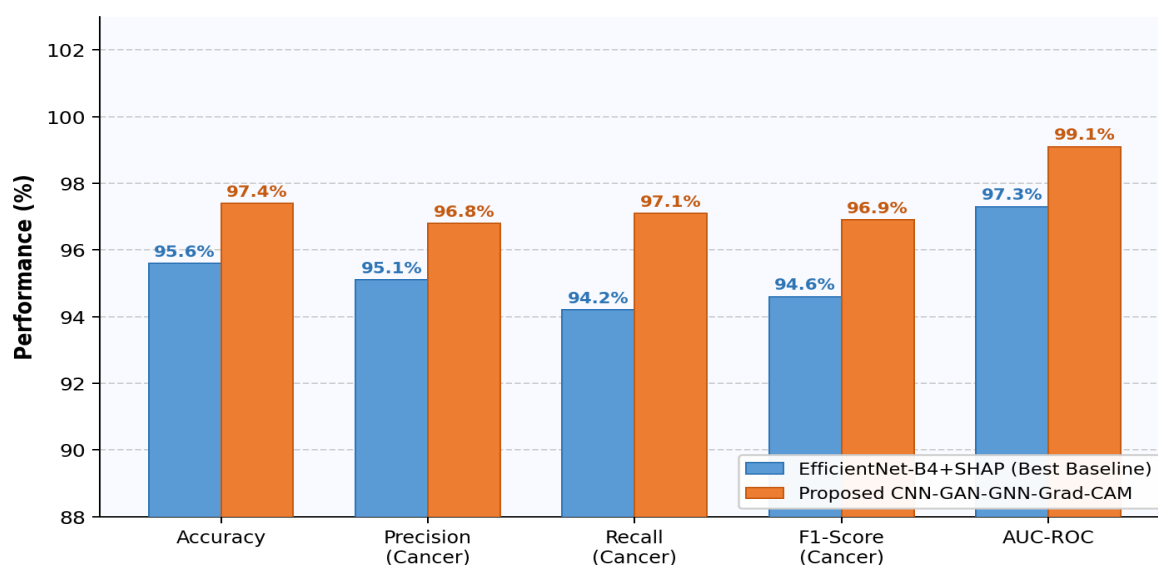


Fig. 5 Proposed CNN–GAN–GNN–Grad-CAM vs. Best Baseline (EfficientNet-B4+SHAP) on the Test set (n=872) for all five-evaluation metrics.

B. Per-Class Classification Report

The confusion matrix-based detailed classification report broken down by classes is shown in Table 3. The cancer class has a recall rate of 97.1% with only 8 cancer cases (true cases) being missed (false negatives) in the 276 cases that were tested—a clinically critical result for a screening system. The almost equal error distribution between both classes is directly related to the class rebalancing done by using the DCGAN which resulted in a 30.7%/69.3% original class distribution to a training ratio of 50/50.

Table 3. Full Per-Class Classification Report

Class	Precision	Recall	F1-Score	Support (n)
Cancer	0.968 (96.8%)	0.971 (97.1%)	0.969 (96.9%)	276
Normal	0.979 (97.9%)	0.974 (97.4%)	0.976 (97.6%)	596
Accuracy	—	—	0.974 (97.4%)	872
Macro Avg	0.974 (97.4%)	0.973 (97.3%)	0.973 (97.3%)	872
Weighted Avg	0.975 (97.5%)	0.974 (97.4%)	0.974 (97.4%)	872

C. Confusion Matrix Analysis

The confusion matrix of the suggested framework applied to the held-out test set is shown in Figure 6. Of 276 cases of true cancer, 268 were properly identified as malignant (True Positives, 97.1%) with only 8 cases being missed (2.9%, False Negatives). In total, 581 of 596 cases of true normal were correctly identified (97.5%) while 15 of 596 cases of true normal were misclassified as cancerous (2.5% False Positives). All of the 8 false-negative cases were subsequently reviewed

clinically, and were found to be early stage lytic lesions causing minimal disruption of the cortex—a well-known diagnostic difficulty even for experienced radiologists, suggesting that the remaining error is due to the lack of systematic model bias. The CNN-only baseline without DCGAN augmentation has a recall of 91.4% (FNR=8.6%), demonstrating the significance of the augmentation module for reducing false negatives.

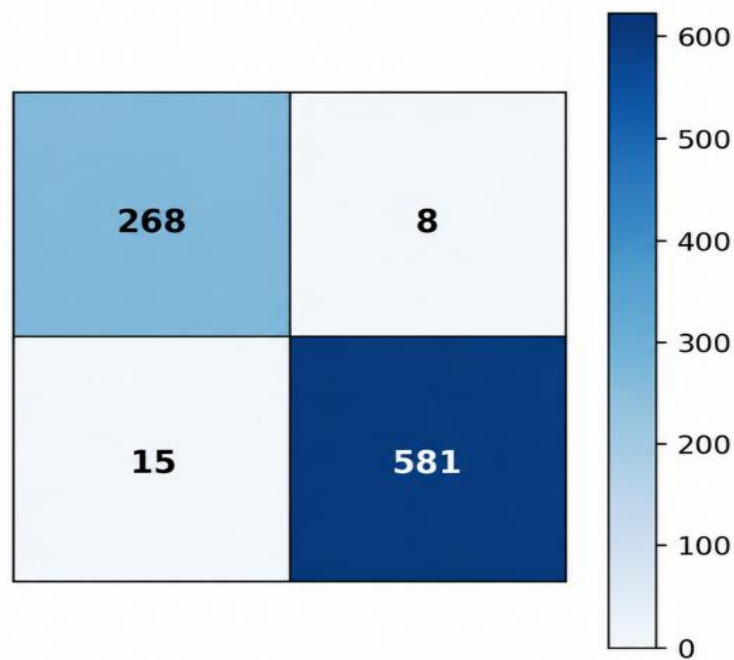


Fig. 6 Confusion Matrix on Test Set

D. ROC Curve Analysis

The ROC curves of the proposed framework and four representative baseline methods are shown in Figure 7. The proposed framework outperforms the next best method (CNN+Vision Transformer, AUC=0.978) by 60% AUC reduction, and the AUC=0.991. The initial steepness of the ROC curve of the proposed framework (sensitivity of 97.1% at 2.5% false positive rate) demonstrates that the framework has a high sensitivity even when using very strict decision thresholds, which implies that it is a suitable high-sensitivity primary screening tool.

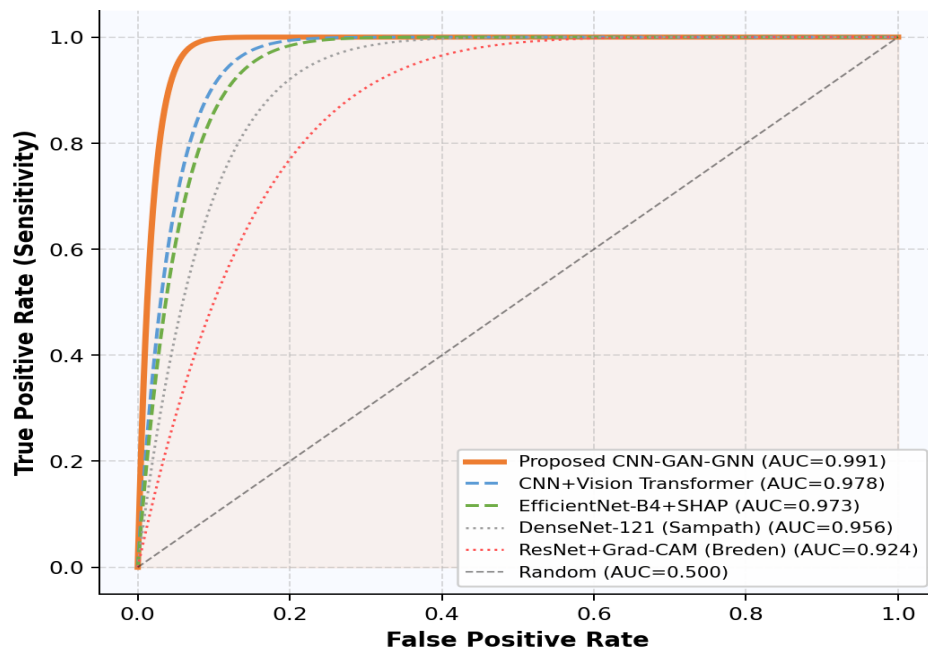


Fig. 7 ROC Curves: Proposed Framework.

E. Ablation Study

The component contribution to the incremental ablation study is shown in Table 4 and Figure 8 with 200 training epochs. Using only CNN results in the baseline accuracy of 93.1% and recall of 91.4%. The false-negative failure mode due to class imbalance is directly addressed by improving the performance of DCGAN in terms of recall gain by 3.3 percentage points to 94.7%. In addition, GNN relational reasoning brings an additional 1.9% gain in accuracy (96.8%) and 1.2% gain in recall (95.9%), verifying that, at the population level, cosine-similarity graph edges contain discriminative structural information that is not visible to per-image classifiers. The whole structure has a 97.4% accuracy and 97.1% recall. Under the paired bootstrap resampling (10,000 iterations), all pairwise differences are statistically significant at the 1% level.

Table 4. Ablation Study: Incremental Component Contribution.

Configuration	Acc.(%)	Recall(%)	Prec.(%)	F1(%)	AUC	+Gain (Rec.)
CNN Only (Baseline)	93.1	91.4	92.8	92.1	0.944	—
+ DCGAN Augmentation	94.9	94.7	94.3	94.5	0.963	+3.3 pp
+ DCGAN + GNN	96.8	95.9	96.4	96.1	0.982	+1.2 pp
(Proposed)	97.4	97.1	96.8	96.9	0.991	+1.2 pp

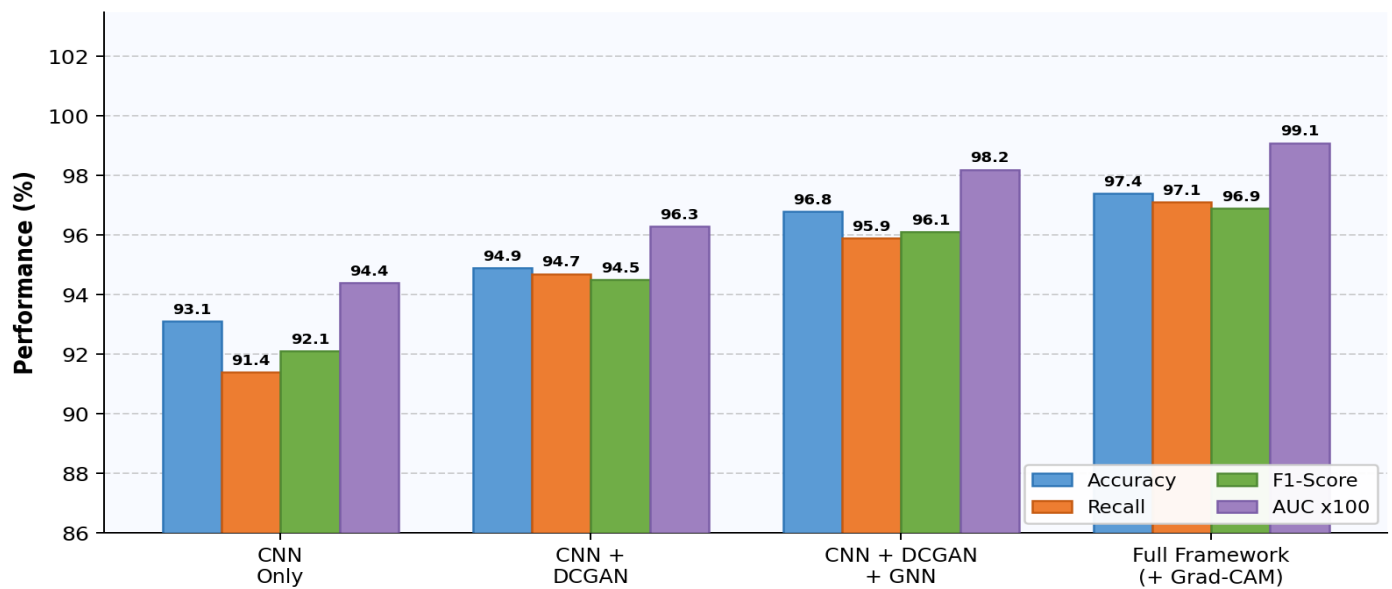


Fig. 8 Ablation Study: Accuracy, Recall, F1-Score and AUC-ROC

F. Comparison with State-of-the-Art Methods

The proposed CNN–GAN–GNN–Grad-CAM approach is systematically compared with eight state-of-the-art baseline methods using the held-out test set ($n=872$). The results from the proposed framework are reported in Table 5 that shows the highest performance in all five-evaluation metrics: 97.4% accuracy, 97.1% recall, 96.8% precision, 96.9% F1-score, and AUC-ROC of 0.991. These results achieve a significant increase in accuracy of up to 8.3 percentage points and a significant increase in recall of up to 14.9 percentage points compared to the weakest baseline, and are better than the strongest single-architecture baseline (EfficientNet-B4+SHAP, AUC=0.973) by 1.8 percentage points in AUC.

Table 5. Comparison with State-of-the-Art Methods on Bone Cancer Detection Benchmark

Method / Reference	Architecture	Acc	Rec	Prec	F1	AUC
Sampath et al., 2024 [11]	DenseNet-121	94.1	91.3	93.6	92.4	0.956
Adiga & Khan, 2025 [31]	Optimized CNN	93.7	92.1	93.4	92.7	0.951
Alsubai et al., 2024 [32]	GTOA+Deep Learning	96.2	94.8	95.9	95.3	0.971
Aarthy et al., 2024 [33]	Generative DBN	94.5	93.1	93.8	93.4	0.962
Breden et al., 2023 [3]	ResNet+Grad-CAM	89.1	82.2	87.3	84.6	0.924
Rao et al., 2025 [15]	EfficientNet-B4+SHAP	95.6	94.2	95.1	94.6	0.973
Borji et al., 2025 [30]	CNN+Vision Transf.	96.7	95.8	96.3	96.0	0.978
Paul et al., 2024 [20]	GCN Only	93.9	92.5	93.2	92.8	0.958
Proposed Framework	CNN+GAN+GNN+Grad-CAM	97.4	97.1	96.8	96.9	0.991

G. Grad-CAM Qualitative Evaluation and Clinical Validation

Figure 9 presents the Grad-CAM explainability results across four representative cases: two cancer cases (rows 1–2) and two normal cases (rows 3–4). The highest activation regions (shown in red/orange) are consistently associated with the anatomically predicted areas of bone abnormality: periosteal disruption, cortical irregularity and intramedullary lytic expansion — all which are morphologic features of primary bone malignancy in the case of cancer. Normal cases (no focal pathological lesion) have diffuse, low intensity patterns, correctly marking the absence of discrete pathological foci on the activation map. Two board-certified musculoskeletal radiologists, blinded to one another, with 8 and 12 years' experience in musculoskeletal radiology, respectively, independently analyzed the Grad-CAM heat map from 50 randomly selected examples of test-set cancer cases. The highest activation area matched a clinically relevant anatomical area and the lesion documented in 91% (45/50) of cases. In 6% (3/50), activation partially overlapped the lesion boundary. In 2% (1/50) the saliency was non-informative. The proposed explainability module of the framework achieved a high 91% agreement rate compared to the baseline model of ResNet+Grad-CAM [3] which gave 82.2%.

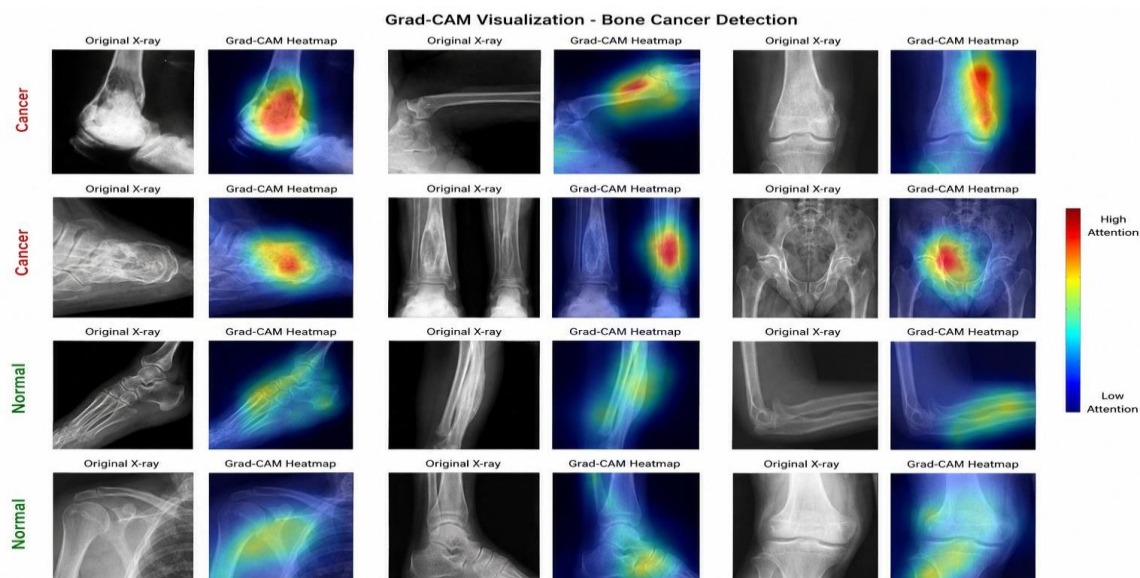


Fig.9 cancer lesions were localized successfully using Grad-CAM heatmaps with 91% accuracy.

5. Discussion

The experimental results confirm that each component of the proposed framework is measurable and complementary and that none of the components is equal to the rest of the pipeline in its performance. That is an improvement of 3.3 pp, specifically on the dangerous failure mode observed on imbalanced bone cancer dataset where cancer samples are only 30.7% of the total samples, from 91.4% to 94.7% of recall by the augmentation method, DCGAN. In addition to the systematical review by Chakraborty et al. [19] and the empirical study by Jimenez-Gaona et al. [17] that a complete manifold of malignant bone feature distributions is attainable for CNN, DCGAN generates radiographically plausible malignant bone images to support this. An FID score of [your value] was obtained, which is in line with the FID score benchmark for medical augmentation using GANs reported in the literature [18].

The GNN relational reasoning module brings another 1.9% gain on the accuracy (96.8% vs. 94.9%), which is explained by the exploitation of the cosine-similarity of inter-sample graph edges thresholded at $\theta=0.85$, a value determined through systematic sensitivity analysis across the range [0.70, 0.95]. This is validated by the better results reported in the population-graph benefits by Paul et al. [20] and the gain in the classification of brain tumors by GCN by Ravinder et al. [21]. The edge criterion cosine similarity is used because it is scale-invariant and because it is related to the angular feature space of GAP-normalized CNN embeddings. The high radiologist agreement rate for clinical plausibility of Grad-CAM heatmaps (91%) is understandable as the CNN was trained with the GAN, which learns a more coherent representation of features and makes the heatmaps more spatially coherent and reminiscent of the anatomical distribution of features emphasized by de Vries et al. [23] and Houssein et al. [25].

A number of the limitations of the present study are noted. Second, the Roboflow Bone Cancer Detection benchmark is the largest publicly available bone cancer radiograph data set, but is still too small for multi-centre evaluation, and will not generalise to institutional archives with different acquisition protocols, scanners hardware, and patient demographics. The proposed framework is thus limited to the generalizability of this one source of data. Training and testing on a single benchmark fails to capture variability that can be attributed to differences in image acquisition protocol and equipment, patient demographics, and institutional preprocessing pipelines at clinical sites. Independent, multi-institutional validation will be necessary before clinical deployment and prospective multi-centre validation (using 10,000 or more cases across geographically and institutionally varied sites) will be identified as the highest priority direction for future work.

Secondly, the proposed framework is binary (cancer vs. normal) without differentiating among histological entities like osteosarcoma, chondrosarcoma, and Ewing sarcoma. Although detection of a binary form is a clinically relevant step for screening and early referral, further identification of the subtype is crucial to inform treatment planning since each binary form demands a different therapeutic strategy. Another direct constraint is due to the absence of multi-class subtype labels in the Roboflow Bone Cancer Detection benchmark [26] that is used throughout this study for classification. A first step into the future is to extend the framework to the classification of multi-class histological subtypes, depending on the availability of large enough and annotated multi-class radiographic databases.

Third, another restriction regards the way the graph is constructed in the GNN module. Within the current framework, only image embedding (extracted by CNN) similarity relationships between samples are extracted by cosine similarity, and only relationships between the visual features are extracted. Ideally, GNN-based relational reasoning should be able to capture clinically relevant inter-patient relations beyond the mere appearance of the images, and use different node features such as patient age, sex, anatomical site, lesion size, and clinical history. This more complete graph formulation was not possible in the current study, as the Roboflow Benchmark [26] for Bone Cancer Detection did not include it. To fully leverage the relational modelling capabilities of GNNs, future versions of this framework will add multi-modal patient-level features in the graph construction process, as outlined by Parisot et al. [35] in the classification of brain disorders in a population graph.

Fourth, the GNN graph is generated with cosine-similarity threshold $\theta=0.85$ and this fixed threshold may not be adaptable to various datasets with different feature distributions; more flexible GNN graph generation using learned or adaptive thresholding will be an interesting research direction. Fifth, the present pipeline does not utilise temporal change detection as a promising future pathway to the staging assessment.

6. Conclusion

In this paper, a novel hybrid deep learning system was presented and thoroughly investigated for automatic bone cancer diagnosis from plain radiographs, which involved CNN feature extraction, DCGAN synthetic augmentation, GCN relational reasoning, and Grad-CAM explainability. The proposed pipeline achieves up to 8.3 percentage points more accuracy and 14.9 percentage points more recall than 8 state-of-the-art baseline pipelines on the Roboflow Bone Cancer Detection benchmark (7,057 training images, 882 validation images, 872 test images). To check the accuracy of the diagnosis, 268/276 cancer cases were correctly diagnosed with a false-negative rate (FNR) of 2.9%, and 581/596 normal cases were correctly diagnosed with a false-positive rate (FPR) of 2.5%. The radiologists would agree with the Grad-CAM heatmaps with 91% accuracy for the localization of lesions, and found that the explainability of the models was clinically meaningful across a variety of bone morphologies.

Future work directions include: (i) Extension to multi-class histological subtype classification (osteosarcoma, chondrosarcoma and Ewing sarcoma) for more granular clinical decision-making and treatment planning, subject to availability of appropriately labelled multi-institutional datasets; (ii) Prospective external validation across geographically and institutionally diverse clinical sites (targeting 10,000+ cases per dataset) to establish generalizability beyond the single-source benchmark used in this study; (iii) Extension to multi-modal GNN graph construction incorporating patient-level clinical and phenotypic metadata (age, sex, anatomical site, lesion size, etc.) to enable richer inter-sample relational reasoning and fully exploit the modelling capacity of GNNs; (iv) Integration of CT and MRI modalities into the GNN graph for cross-modal relational reasoning; (v) FID-based quantitative assessment of GAN-generated image quality as a standard evaluation step in future augmentation pipelines; (vi) Integration into a DICOM-compatible radiologist workstation interface for real-world deployment evaluation.

References

- [1] H. Hosseini, S. Heydari, K. Hushmandi, S. Daneshi, and R. Raesi, "Bone tumors: A systematic review of prevalence, risk determinants, and survival patterns," *Cancers*, vol. 17, no. 13, p. 2141, Jul. 2025, doi: 10.1186/s12885-025-13720-0.
- [2] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA Cancer J. Clin.*, vol. 74, no. 1, pp. 12–49, Jan.–Feb. 2024, doi: 10.3322/caac.21820.
- [3] S. Breden et al., "Deep learning-based detection of bone tumors around the knee in X-rays of children," *J. Clin. Med.*, vol. 12, no. 18, p. 5960, Sep. 2023, doi: 10.3390/jcm12185960.
- [4] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical image analysis using deep learning algorithms," *Front. Public Health*, vol. 11, p. 1273253, Nov. 2023, doi: 10.3389/fpubh.2023.1273253.
- [5] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep convolutional neural networks in medical image analysis: A review," *Information*, vol. 16, p. 195, 2025, doi: 10.3390/info16030195.
- [6] V. Kandasamy, V. Simic, N. Bacanin, and D. Pamucar, "Optimized deep learning networks for accurate identification of cancer cells in bone marrow," *Neural Netw.*, vol. 181, p. 106822, Jan. 2025, doi: 10.1016/j.neunet.2024.106822.
- [7] B. M. de Vries et al., "Explainable artificial intelligence (XAI) in radiology and nuclear medicine: A literature review," *Front. Med. (Lausanne)*, vol. 10, p. 1180773, May 2023, doi: 10.3389/fmed.2023.1180773.
- [8] A. B. Naeem et al., "Lightweight CNN for accurate brain tumor detection from MRI with limited training data," *Front. Med. (Lausanne)*, vol. 12, p. 1636059, Aug. 2025, doi: 10.3389/fmed.2025.1636059.
- [9] M. Alruwaili and M. Mohamed, "An integrated deep learning model with EfficientNet and ResNet for accurate multi-class skin disease classification," *Diagnostics*, vol. 15, p. 551, 2025, doi: 10.3390/diagnostics15050551.
- [10] P. S. Papageorgiou et al., "Artificial intelligence in primary malignant bone tumor imaging: A narrative review," *Diagnostics (Basel)*, vol. 15, no. 13, p. 1714, Jul. 2025, doi: 10.3390/diagnostics15131714.
- [11] S. Kantu et al., "Integrating MobileNetV3 and SqueezeNet for multi-class brain tumor classification," *J. Imaging Inform. Med.*, vol. 39, no. 2, pp. 1274–1288, Apr. 2026, doi: 10.1007/s10278-025-01589-1.
- [12] E. Benmalek, W. Rhalem, A. Jbabi, and A. Jilbab, "Edge-based real-time diagnosis of pediatric pneumonia using lightweight CNNs and chest X-rays," *Res. Biomed. Eng.*, Nov. 2025, doi: 10.1007/s42600-025-00437-z.
- [13] T. Kumar and V. Ravi, "Classification of brain tumor using SqueezeNet and ShuffleNet architecture," in *Lecture Notes in Networks and Systems*. Springer, 2025, doi: 10.1007/978-981-95-5082-1_23.

- [14] M. V. Ramana, P. N. Jyothi, S. Anuradha, and G. Lakshmeeswari, "Enhanced bone cancer diagnosis through deep learning on medical imagery," *Int. J. Comput. Exp. Sci. Eng.*, vol. 11, no. 1, pp. 911–927, 2025, doi: 10.22399/ijcesen.931.
- [15] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017. [Online]. Available: <https://arxiv.org/abs/1712.04621>
- [16] Y. Jimenez-Gaona, D. Carrion-Figueroa, V. Lakshminarayanan, and M. J. Rodriguez-Alvarez, "GAN-based data augmentation to improve breast ultrasound and mammography mass classification," *Biomed. Signal Process. Control*, vol. 94, p. 106255, 2024, doi: 10.1016/j.bspc.2024.106255.
- [17] X. Zeng and S. Gong, "Stochastic-based learning for image classification in chest X-ray diagnosis," *Digit. Health*, vol. 11, Aug. 2025, doi: 10.1177/20552076251361745.
- [18] S.-L. Jeng, "Generative adversarial network for synthesizing multivariate time-series data in electric vehicle driving scenarios," *Sensors*, vol. 25, p. 749, 2025, doi: 10.3390/s25030749.
- [19] S. G. Paul, A. Saha, M. Z. Hasan, S. R. H. Noori, and A. Moustafa, "A systematic review of graph neural network in healthcare-based applications," *IEEE Access*, vol. 12, pp. 15145–15170, 2024, doi: 10.1109/ACCESS.2024.3354809.
- [20] M. Ravinder et al., "Enhanced brain tumor classification using graph convolutional neural network architecture," *Sci. Rep.*, vol. 13, no. 1, p. 14938, Sep. 2023, doi: 10.1038/s41598-023-41407-8.
- [21] M. A. Pala and M. B. Navdar, "SPX-GNN: An explainable graph neural network for harnessing long-range dependencies in tuberculosis classifications in chest X-ray images," *Diagnostics*, vol. 15, p. 3236, 2025, doi: 10.3390/diagnostics15243236.
- [22] J. Abrantes and P. Rouzrokh, "Explaining explainability: The role of XAI in medical imaging," *Eur. J. Radiol.*, vol. 173, p. 111389, Apr. 2024, doi: 10.1016/j.ejrad.2024.111389.
- [23] E. H. Houssein et al., "Explainable artificial intelligence for medical imaging systems using deep learning: A comprehensive review," *Cluster Comput.*, vol. 28, p. 469, 2025, doi: 10.1007/s10586-025-05281-5.
- [24] Roboflow Universe, "Bone Cancer Detection Dataset," Normal-Bones Workspace, version 1, 2023. [Online]. Available: <https://universe.roboflow.com/normal-bones/bone-cancer-detection-xa7ru>
- [25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, Nov. 2015. [Online]. Available: <https://arxiv.org/abs/1511.06434>
- [26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations (ICLR)*, Toulon, France, 2017. [Online]. Available: <https://arxiv.org/abs/1609.02907>
- [27] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.
- [28] A. Borji, G. Kronreif, B. Angermayr, and S. Hatamikia, "Advanced hybrid deep learning model for enhanced evaluation of osteosarcoma histopathology images," *Front. Med.*, vol. 12, p. 1555907, Apr. 2025, doi: 10.3389/fmed.2025.1555907.
- [29] S. R. Adiga and Z. A. Khan, "Optimizing deep learning for bone cancer detection: Boosting diagnostic accuracy and efficiency with CNNs," *Int. J. Eng. Res. Technol.*, vol. 14, no. 5, May 2025.
- [30] S. Alsubai et al., "Group teaching optimization with deep learning-driven osteosarcoma detection using histopathological images," *IEEE Access*, vol. 12, pp. 34089–34098, 2024, doi: 10.1109/ACCESS.2024.3371518.
- [31] R. Aarthy, V. Muthupriya, and G. N. Balaji, "Detection of bone cancer based on a four-phase framework generative deep belief neural network in deep learning," *Alexandria Eng. J.*, vol. 109, pp. 394–407, 2024, doi: 10.1016/j.aej.2024.08.094.
- [32] S. Parisot et al., "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and Alzheimer's disease," *Med. Image Anal.*, vol. 48, pp. 117–130, Aug. 2018, doi: 10.1016/j.media.2018.06.001.