Bone Anomaly Identification and Localization from Musculoskeletal X-ray images using Hybrid DenseNet-Vision Transformers.

Aparna Kanakatte APARNA.KG@TCS.COM
Divya Bhatia BHATIA.DIVYA@TCS.COM
TCS Research and Innovation, Bengaluru

Rupsha Mukherjee M.RUPSHA@TCS.COM
Avik Ghose AVIK.GHOSE@TCS.COM
TCS Research and Innovation, Kolkata

Murali Poduval MURALI.PODUVAL@TCS.COM
TCS Medical Devices and Diagnostics, Mumbai

Editors: Under Review for MIDL 2026

Abstract

Musculoskeletal anomalies present a global health challenge, often requiring expert review for diagnosis. Automated detection systems can alleviate this burden and facilitate timely intervention. We introduce a hybrid framework combining DenseNet and Vision Transformers (ViT) for detecting anomalies in the MURA dataset. This model uses CNNs for local feature extraction and ViT for global context modeling, effectively capturing the subtle features of bone pathologies. Our binary model achieved 92% accuracy across seven bone types, demonstrating strong generalizability through cross-anatomy testing. To enhance fine-grained annotations, we collaborated with an orthopedic expert to label 1,379 finger radiographs into three sub-anomaly categories: arthritis, fracture, and implant, with bounding boxes. This framework reached 96% classification accuracy and 91% localization efficiency (AUC), marking the first instance of sub-anomaly classification on MURA finger data with bounding box predictions providing visual validation for clinicians. Additionally, we also analyze the importance of positional encodings in ViT, showing their necessity for localization but not for binary classification. This work offers a new annotated dataset and a robust framework, advancing automated orthopedic anomaly detection.

Keywords: Musculoskeletal anomalies, Fracture, Arthritis, Visual transformers

1. Introduction

Musculoskeletal disorders (MSD) affect bones, joints, muscles, and soft tissues and are a leading cause of disability worldwide. Common examples include hairline fractures, carpal tunnel syndrome, and ligament tears, often resulting from occupational injuries, sports trauma, or degeneration. The Global Burden of Disease study estimates that MSD impacts around 1.71 billion people, accounting for nearly 30% (Mime et al., 2024) of global disability, highlighting a critical public health challenge.

X-ray imaging is a valuable diagnostic tool for musculoskeletal disorders, but techniques like CT, MRI, and Ultrasound are used for complex cases. Early detection and accurate diagnosis from these scans are essential for treatment planning; however, this process is primarily manual and labor-intensive, requiring significant effort from trained clinicians.

With the high volume of scans performed daily, this burden delays treatment and strains healthcare resources (Yuvraj et al., 2021). Therefore, reliable automated algorithms for anomaly detection are crucial for delivering timely and effective care.

The MURA dataset (Rajpurkar et al., 2018), one of the largest public collections of musculoskeletal radiographs, has advanced research in this field. It offers binary annotations (normal/abnormal) for seven upper extremity bone regions but lacks clinically meaningful sub-anomaly labels (e.g., fracture, arthritis) and localization annotations essential for real-world diagnostics. To address this, we enhance MURA by adding expert annotations for finger radiographs, including relevant sub-anomaly categories and bounding boxes. Our automated framework enables anomaly detection and fine-grained subclassification with localization, aiming to reduce radiologists' workloads, improve diagnostic consistency, and provide interpretable results that foster clinician trust and practical use.

2. Literature review

Deep learning's application in musculoskeletal radiograph analysis is gaining traction due to its ability to reduce diagnostic time and inter-observer variability. The MURA dataset, introduced by Rajpurkar *et. al.* (Rajpurkar et al., 2018), contains over 40,000 upper extremity images annotated as normal or abnormal. Their DenseNet169 models achieved area under the ROC curve (AUROC) scores comparable to expert radiologists. Kaya *et. al.* (Kaya and Taşcı, 2023) used EfficientNet-B0 and Neighborhood Component Analysis for anomaly classification, reporting 91% accuracy. Zeng *et. al.* (Zeng et al., 2024) proposed FusionNet, combining MobileNetV2 and EfficientNet-B2 with coordinate attention, achieving 87.10% accuracy and AUC of 93.89% on elbow images. Duan *et. al.* (Duan et al., 2024) introduced HRD, a ResNet50–DenseNet121 hybrid combined with human input, achieving 88.81% accuracy. Karthik *et. al.* (Karthik and S.S.Kamath, 2023) presented MSDNet for detection and report generation, while Sultana *et. al.* (Mime et al., 2024) demonstrated the benefits of transfer learning for low-resource adaptation. Hrubý *et. al.* (Hrubý et al., 2024) explored fracture localization using YOLO with custom bounding box annotations on MURA, reporting performance across internal datasets and FracAtlas. Despite advancements, current research primarily focuses on binary classification and coarse fracture detection, lacking detailed sub-anomaly labels or localization. Transformer-based methods have recently shown promise in medical imaging for ultrasound (Valanarasu et al., 2021), histopathology (Shao et al., 2021) and MRI (A.Hatamizadeh et al., 2021). Notably, there is no existing study that integrates CNN and ViT in a hybrid framework for classifying and localizing multiple sub-anomalies in finger radiographs. The major contributions of this work includes

- **Expert-annotated dataset:** Annotated 1,379 finger x-rays for fractures, arthritis, and implants, providing the unique resources for finger orthopedic analysis.
- **Dual-level hybrid framework:** Proposed a DenseNet–ViT architecture for binary anomaly detection and fine-grained subclassification with localization.
- **Blind testing:** Achieved over 80% accuracy in cross-anatomy experiments, demonstrating the framework's robustness for real-world applications.

- **Task-specific methodological insight:** Identified that positional encodings in ViT are unnecessary for binary detection but essential for fine-grained localization.

3. Proposed method

The Vision Transformer (ViT) is an image classification model that uses a transformer-like architecture over image patches of an image. An image is divided into fixed-size patches, which are linearly embedded and combined with position embeddings before being processed by a standard transformer encoder. ViT was chosen for this study due to its superior ability to model global contextual relationships through self-attention mechanisms, crucial for medical imaging. Anomalies in radiographs are often present as irregular, non-localized patterns (as seen in arthritis), occur in low-contrast regions (such as hairline fractures), or rely on subtle spatial correlations, making global reasoning essential—something standard CNNs with local biases struggle to achieve. The proposed hybrid architecture combines a pretrained feature extractor (DenseNet) with a ViT encoder. This design retains the local inductive biases of CNN, capturing fine-grained textures and edges, while leveraging ViT to model long-range dependencies across patch embeddings. This fusion of local and global feature modeling proved advantageous for both classification and localization tasks. Our framework illustrated in Fig 1 includes two models:

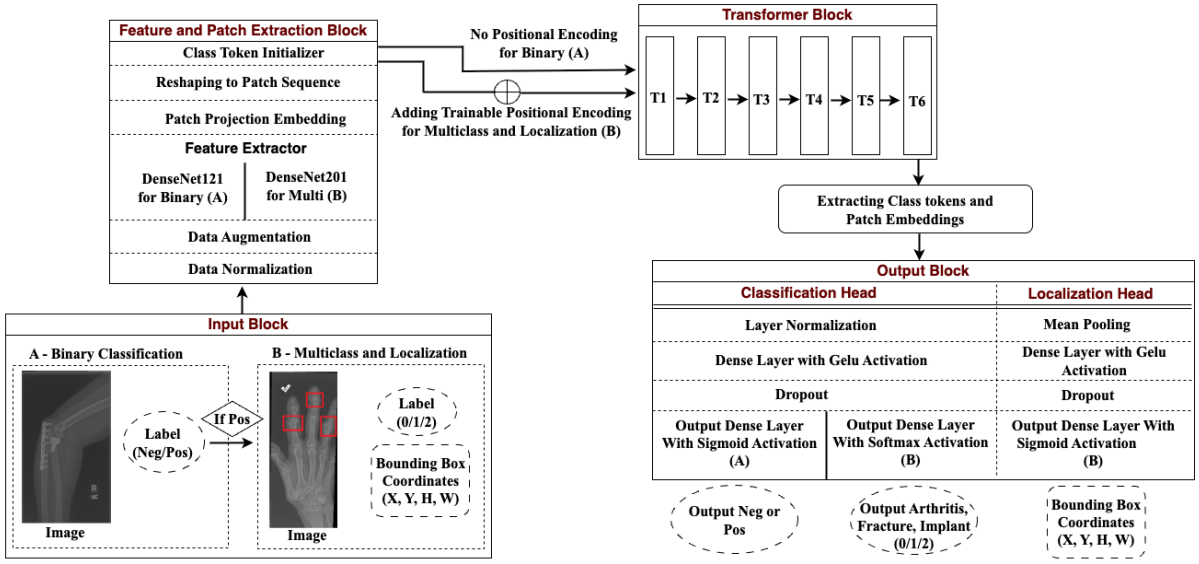


Figure 1: Block diagram of the proposed dual-structured ViT-based framework.

1. A binary classification network (A) for classifying the image into 'Neg' (no anomaly) and 'Pos' (anomaly).
2. A multiclass classification and localization model (B) to detect specific sub-anomalies (arthritis, fracture, implant) along with bounding box localization.

The proposed framework is divided into four distinct blocks, namely, input block, feature and patch extraction block, transformer block and output block.

Input block: Two distinct input blocks are used depending on the task.

- Binary Classification (A): Takes grayscale image and binary label (0 = Neg, 1 = Pos).
- Multiclass Classification and Localization (B): Takes grayscale image, a categorical label indicating anomaly type (0 = arthritis, 1 = fracture, 2 = implant), and a list of bounding box coordinates.

Feature and patch extraction block: This block utilizes a pretrained DenseNet model with imagenet weights to extract fine details from medical radiographs, such as hairline fractures. The output feature maps are then passed through a 1×1 convolution to reduce the channel dimension, which is then projected onto a fixed embedding space, thereby forming a sequence of patch embeddings. A class token is prepended to this sequence. In binary models, position is less critical, so no positional encoding (PE) is used, while multiclass localization models incorporate learnable PE to capture spatial relationships for object localization.

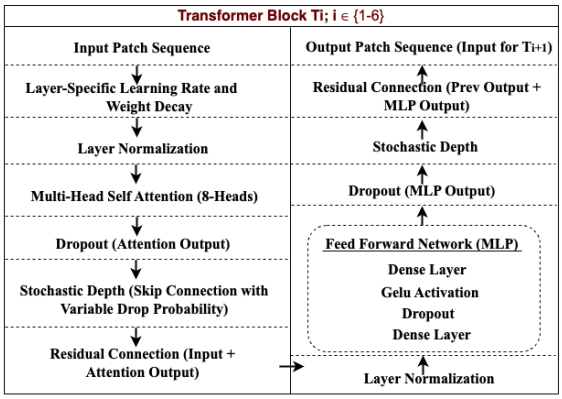


Figure 2: Block diagram of the transformer block.

Transformer block: The internal structure of each transformer is shown in Fig 2. The encoded sequence from the feature and patch extraction block passes through six transformer blocks, each composed of:

- Layer Normalization.
- Multi-Head Self-Attention (MHSA) with 8 heads, enabling global contextual modeling across patches.
- Feedforward Network (MLP) with Gaussian error linear unit (Gelu) activation and dropout.
- Stochastic Depth: Each block has a probability of being skipped during training, thereby preventing overfitting and improving generalization.
- Residual connections after both attention and MLP sub-blocks to support stable gradient flow.

A layer-wise training strategy optimizes each component with varying learning rates and weight decays: lower for the pretrained DenseNet backbone, moderate for transformer blocks, and higher for classification and detection heads. After passing through the transformer blocks, the model outputs a sequence of tokens, including a class token and transformed patch embeddings, which are routed to separate output heads. The class token, which was prepended to the input sequence at the start aggregates global information through self-attention, serving as a summary representation for the classification head's label prediction. Remaining patch tokens (*i. e.* embeddings corresponding to the individual image regions) are mean-pooled to create a compact representation for the localization head, predicting bounding box coordinates (x, y, height, width) for sub-anomalies.

Output block: Depending on the task, two output branches are generated.

- The classification head consists of layer normalization, a dense layer with Gelu activation and a dropout layer. The activation function is chosen as sigmoid for binary (A) and softmax for multiclass classification (B).
- The localization head (B) features mean-pooled patch embeddings, a dense layer with Gelu activation, and a dropout layer. It uses sigmoid activation to predict bounding box coordinates (x, y, height, width). The model predicts a fixed number of bounding boxes (k=5) to address multiple co-occurring anomalies per image, matching the dataset's maximum. During training, ground truth boxes are padded with binary masks for consistent shapes, focusing the loss on valid predictions. At inference, confidence thresholding and non-maximum suppression are applied to retain distinct anomaly regions, enabling effective detection of multiple pathologies like fractures and arthritis in a single image.

3.1. Loss function:

- Binary classification model (A) uses a binary cross-entropy loss function.
- Multiclass classification with localization (B) uses categorical cross-entropy (CCE) and intersection over union (IoU) loss functions. The final loss as shown in Equation 1 is computed as a weighted sum of CCE and IoU, allowing tuning of the relative importance between class accuracy and spatial localization precision.

$$\mathbf{L}_{total} = \alpha.L_{CCE} + \beta.L_{IoU} \tag{1}$$

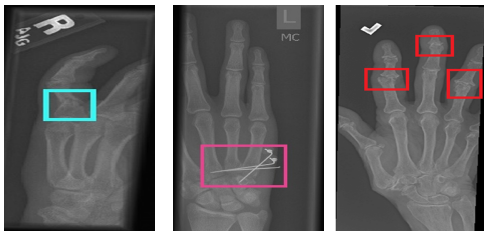
  where α=0.65 and β=0.35 are weights chosen based on validation performance.

This dual-model architecture effectively balances local feature extraction with global reasoning. Both models benefit from task-adaptive data augmentation, layer-wise optimization, and stochastic regularization, contributing to their robustness and efficiency. The binary model adopts a lightweight design with no PE, making it well-suited for rapid and efficient detection of anomaly presence. In contrast, the multiclass model incorporates learnable PE and dual-output heads to simultaneously handle complex classification and spatial localization tasks. Since both classification and localization losses are propagated jointly, the model learns to focus on discriminative anomaly regions—enhancing its accuracy, interpretability, and generalizability across varied anomaly types.

4. Results and discussion

4.1. Dataset details

The MURA dataset (Rajpurkar et al., 2018) is a comprehensive collection of musculoskeletal radiographic images, featuring 40,561 images from 14,656 studies of 12,173 patients. It includes multiview images of fingers, hands, wrists, forearms, elbows, humerus, and shoulders in the upper extremity region. This dataset was released by the Stanford group as part of the bone X-ray deep learning competition providing only binary annotations for anomalies. To incorporate localization and sub-anomalies classification, additional bounding box annotations were created, with the assistance of an orthopedic surgeon, across 1,379 finger images, categorizing them into three classes: 301 fractures, 283 arthritis, and 795 implants. This effort was driven by the complexity of the finger dataset and previous low accuracies reported in the base paper (Rajpurkar et al., 2018). Additional bounding box annotations along with the anomaly type were added using the 'labelme' annotation tool (Russel et al., 2008) as shown in Fig 3.



(*a*) Fracture (*b*) Implant (*c*) Arthritis

Figure 3: Manual annotations with sub-anomalies on finger dataset using the labelme tool.

4.2. Implementation details

Table 1 outlines the ViT parameters used in the implementation. Input images are normalized to 256×256, with zero-padding for smaller images and cropping for larger ones, ensuring that the region of interest is not affected. Pixel values are scaled between $[0, 1]$. To enhance training diversity and minimize storage needs, on-the-fly augmentations like random brightness adjustment, zoom, Gaussian noise injection, random erasing, and small-angle rotation are employed. Two additional patch-based augmentations, CutMix and MixUp, are also used: CutMix replaces a rectangular patch in one image with a patch from another, while MixUp combines two images and their labels linearly. Each augmentation is applied probabilistically to maintain a balance between diversity and realism. The framework is developed using TensorFlow (M.Abadi et al., 2015) and OpenCV.

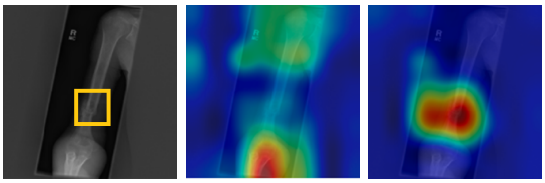
4.3. Results of binary classification

Table 2 shows that our model outperforms existing methods in binary classification accuracy and achieves high Cohen's kappa scores (Refer appendix for Kappa calculation) for all seven bone categories. Misclassified cases had low prediction confidence (probabilities between 0.52 and 0.69), prompting the implementation of a confidence-based flagging system that routes low-confidence predictions for expert review. This approach improves clinical reliability and creates a feedback loop for human validation in diagnostic workflows.

Table 1: ViT parameters with values used in implementation

Parameter name	**Parameter value**
Optimizer	Custom AdamW with weight decay
Project dimension	128
Transformer blocks	6
Patch size	4×4
Attention heads	8
MLP units (Feed forward network)	256
Dropout	0.3 for binary(A) 0.2 for multi(B)
Stochastic depth	Up to 0.1, increasing with layer depth
L2 regularization	1e-3 to 2e-3
Base learning rate (LR)	3e-4
DenseNet backbone LR	0.1 * Base LR
Transformer layer LR	0.5 * Base LR
Output head LR	1.0 * Base LR
Freezing backbone layer	For DenseNet121 (A): 100 layers For DenseNet201 (B): 300 layers
Weight decay	1e-4 (adjusted per layer)
Class weights	Inverse frequency weighting
Batch size	32
Epochs	100
Train-val-test split	60-20-20
Cross-validation	10-fold
Training strategies	Early stopping and reduce LR when metric has stopped improving

To improve interpretability and trust in clinical deployment, we applied an attention-based GradCam technique to our hybrid architecture. Fig 4 shows visualizations indicating which regions of the X-ray most influence the model's predictions, with bright/hot (red/yellow) areas signifying high relevance and dark/cool (blue/black) areas indicating minimal relevance. DenseNet-based heatmaps (Fig 4(b)) focus on high-contrast anatomical landmarks that may not indicate pathology, while transformer-derived attention maps (Fig 4(c)) emphasize clinically relevant areas, like implants and early-stage anomalies. These visualizations are not intended for direct performance comparison but to illustrate the complementary strengths of each with DenseNet providing detailed low-level features and ViT capturing global, pathology driven context, enhancing trust, transparency, and diagnostic relevance in our hybrid model.



(a) (b) (c)

Figure 4: GradCam heatmap depicting the concentration of features across the anomaly region between densenet and ViT. (a) is the input image with fracture (shown in yellow box), (b) densenet layer heatmap and (c) transformer attention layer heatmap.

Table 2: Performance comparison of the proposed binary model with other reported in literature

	Elbow		**Finger**		**Forearm**		**Hand**		**Humerus**		**Shoulder**		**Wrist**	
	Acc	κ	Acc	κ	Acc	κ	Acc	κ	Acc	κ	Acc	κ	Acc	κ
Rajpurkar (Rajpurkar et al., 2018)	-	0.71	-	0.39	-	0.74	-	**0.85**	-	0.60	-	0.73	-	0.93
Karthik (Karthik and S.S.Kamath, 2023)	83.1	0.74	79.9	0.67	83.9	0.79	78.3	0.75	85.9	0.68	79.3	0.73	84.5	0.86
Kaya (Kaya and Taşcı, 2023)	92.0	-	91.2	-	92.1	-	**91.3**	-	91.4	-	89.5	-	92.6	-
Zeng (Zeng et al., 2024)	87.0	-	81.0	-	83.0	-	81.0	-	87.0	-	81.0	-	87	-
Duan (Duan et al., 2024)	88.4	0.80	83.7	0.73	84.5	0.74	83.3	0.71	88.5	0.80	81.6	0.70	86.7	0.77
Proposed ViT No-PE	91.0	0.81	**92.0**	**0.83**	93.7	0.87	90.0	0.75	**92.1**	**0.85**	**90.2**	**0.80**	**94.0**	**0.88**
Proposed ViT with PE	**92.0**	**0.82**	91.5	0.82	**93.8**	**0.87**	90.0	0.75	91.6	0.84	90.0	0.80	93.3	0.86

4.4. Multiclass classification and localization results

We concentrated on the finger bone, using images labeled "Positive" for anomalies. Our ViT-based model, designed with parallel classification and bounding box regression heads, demonstrated strong performance in both identifying the correct sub-anomaly type and generating accurate localization coordinates. This is the first application of detailed sub-anomaly classification with bounding box localization on the MURA finger dataset, lacking previous research for comparison. To thoroughly assess the model:

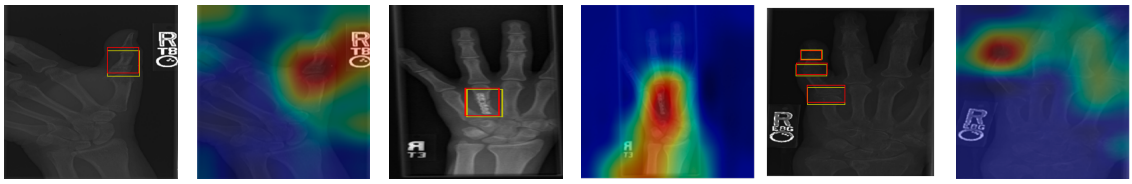
- For sub-anomaly classification, we report accuracy, precision, recall, and F1-score.
- For localization, we assess performance using AUC, IoU, precision, and recall.

The proposed method identified 264 sub-anomalies out of 276 testing samples obtained from the $60-20-20\%$ split, achieving an accuracy of 96%. Compared to established CNN models, YOLO and Faster R-CNN reached accuracies of 76% and 64%, respectively. Our hybrid model shown superior performance with high precision, recall, and F1-score as summarized in Table 3, demonstrating its effectiveness in addressing visually complex musculoskeletal anomalies.

Fig 5 illustrates the predicted bounding boxes for three anomalies identified by our algorithm, which successfully detects multiple anomalies as seen in in Fig 5(e). The GradCam images in Fig 5(b), Fig 5(d) and Fig 5(f) represent the output of attention layers from the multiclass classification and localization network. As observed, the model effectively concentrates around the regions of the bounding boxes for fracture, implant, and arthritis, demonstrating its ability to focus on clinically relevant areas.

Table 3: Performance comparison of the proposed sub-anomaly classification model with Yolo and Faster R-CNN on the finger dataset.

Class	Method	Precision	Recall	F1-score
Arthritis	Faster R-CNN	0.74	0.43	0.55
	Yolo	0.94	0.80	0.86
	Proposed	0.93	0.90	0.92
Fracture	Faster R-CNN	0.84	0.29	0.43
	Yolo	1.00	0.25	0.40
	Proposed	0.92	0.93	0.93
Implant	Faster R-CNN	0.93	0.84	0.88
	Yolo	1.00	0.93	0.96
	Proposed	0.98	0.99	0.99



(a) Fracture (b) (c) Implant (d) (e) Arthritis (f)

Figure 5: Results of the proposed bounding box anomaly localization. Note: Yellow denotes ground truth and red denotes the predicted. The corresponding feature concentration from GradCam visualization is shown in (b), (d) and (f) images.

.

The quantitative results of the anomaly localization is shown in Table 4. Our model achieved an AUC of 0.91 and IoU of 0.88. YOLO and Faster R-CNN reported IoUs of 0.66 and 0.61, respectively, demonstrating the superior localization accuracy of our approach.

Table 4: Performance comparison of the proposed anomaly localization model with Yolo and Faster R-CNN on the finger dataset.

Method	AUC	IoU	Precision	Recall
Faster R-CNN	0.63	0.61	0.72	0.48
Yolo	0.66	0.66	0.67	0.73
Proposed	0.91	0.88	0.92	0.93

4.5. Role of positional encodings

In the binary classification task, the goal was to identify anomalies (like arthritis or fractures) in radiographs without needing to pinpoint their locations. Anomalies are generally observable globally, making the exact positioning less critical. The study explored the role of positional encoding (PE) in the Vision Transformer (ViT) architecture, finding minimal differences in performance metrics of the accuracy and Cohen's Kappa with or without PE as seen in Table 2. This indicated that spatial information wasn't a limiting factor, allowing effective learning of distinguishing features without explicit patch positioning. Conse-

quently, we excluded PE in the binary model, which simplified the model and enhanced computational efficiency.

Conversely, the multiclass classification and localization task necessitated both anomaly identification and precise bounding box predictions, relying heavily on spatial reasoning. Thus, we included learnable PE to improve the model's grasp of spatial relationships, resulting in better performance in classification and localization. This highlights that the use of PE is task-specific: it can be eliminated for global classification tasks to boost efficiency, while it is crucial for spatially sensitive tasks like localization.

4.6. Blind testing on binary anomaly classification

We conducted four experiments with training and testing on non-overlapping anatomical regions. Results in Table 5 show that all setups achieved over 80% accuracy in binary anomaly classification. This supports our choice to exclude positional encoding in the ViT model, as predictions remained consistent across anatomical locations. Our findings suggest that in binary anomaly detection, the anomaly's pattern, rather than its exact location, influences model performance. This also demonstrates the robustness and adaptability of our ViT-based approach for other medical imaging tasks with minimal modifications.

Table 5: Blind testing results for various bone combinations on binary anomaly classification

	Training bones	**Tested bones**	**Accuracy (%)**	**Kappa κ**
1	Humerus, Elbow	Forearm	82	0.63
2	Finger, Hand	Wrist	86	0.72
3	Shoulder, Forearm, Hand	Elbow	88	0.75
4	Elbow, Finger, Forearm, Wrist	Humerus	89	0.77

5. Conclusion

This work enhances automated orthopedic anomaly detection by integrating clinical data with a robust AI framework. We introduced an expert-annotated subset of the MURA finger dataset, featuring fine-grained labels and bounding boxes. Our dual-level hybrid DenseNet–ViT framework enables binary anomaly detection and detailed subclassification with localization. Cross-anatomy testing confirmed its strong generalization to unseen bone regions, essential for clinical use. Attention-based visualizations support radiologist trust, while our PE analysis offers specific guidance for medical AI design. Overall, this study provides an enriched imaging resource and a clinically relevant framework, aiming for reliable, interpretable, and scalable musculoskeletal anomaly detection solutions. It could reduce radiologist workload, improve diagnostic consistency, and expedite patient triage. Future work includes extending annotations to other anatomical regions and validating against multi-institutional datasets while integrating into real-world workflows.

References

A.Hatamizadeh, Y. Tang, V. Nath, and et al. Unetr: Transformers for 3d medical image segmentation. In *https://arxiv.org/abs/2103.10504*, 2021.

G. Duan, S. Zhang, Y. Shang, and W. Kong. Research on x-ray diagnosis model of musculoskeletal diseases based on deep learning. *Applied Sciences*, 14(8), 2024.

R. Hrubý, D. Kvak, J. Dandár, A. Atakhanova A, M. Misař, and D. Dufek. Cross-center validation of deep learning model for musculoskeletal fracture detection in radiographic imaging: A feasibility study. *medRxiv*, 2024.

K. Karthik and S.S.Kamath. Msdnet: A deep neural ensemble model for abnormality detection and classification of plain radiographs. *Ambient Intelligence and Humanized Computing*, 14:16099–16113, 2023.

O. Kaya and B. Taşcı. A pyramid deep feature extraction model for the automatic classification of upper extremity fractures. *Diagnostics*, 13, 2023.

M.Abadi, A.Agarwal, P. Barham, E. Brevdo, Z. Chen, and et al. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015.

T.S. Mime, D. Bala, M. A. Hossain, M.A.Rahman, M.S.Hossain, and M.I.Abdullah. A new benchmark on musculoskeletal abnormality recognition system using deep transfer learning model. In *3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pages 1–6, 2024.

P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, and et al. Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs, 2018. URL https://arxiv.org/abs/1712.06957.

B. Russel, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77, 2008. doi: 10.1007/s11263-007-0090-8.

Z. Shao, H. Bian, Y. Chen, and et al. Transmil: Transformer-based mil for whole slide image classification. In *Proc of NIPS*, 2021.

J. M. J Valanarasu, P.Oza, I. Hacihaliloglu, and V.M. Patel. Medt: A pure transformer for medical image segmentation. In *arXiv:2102.02552*, 2021.

D. Yuvraj, B. Adikar, M. Veena, and S.S.Kamath. Deepoa: Clinical decision support system for early detection and severity grading of knee osteoarthritis. In *5th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 250–255, 2021.

Z. Zeng, C. Song, S. Yi Q. Liu, and Y. Zhu. Diagnosis of musculoskeletal abnormalities based on improved lightweight network for multiple model fusion. *Mathematical Biosciences and Engineering*, 21:582–601, 2024.

(a) Elbow (b) Finger (c) Forearm (d) Hand (e) Humerus (f) Shoulder (g) Wrist

Figure 6: Sample X-ray images in MURA dataset.

Appendix A. Kappa calculation

The Kappa coefficient κ, also known as Cohen's Kappa, is a statistical measure used to evaluate the level of agreement between two or more raters on categorical (qualitative) variables. Unlike simple percent agreement, it accounts for the agreement that could occur by chance, making it a more reliable and robust indicator of inter-rater consistency. Cohen's Kappa is particularly useful for assessing model performance on imbalanced datasets, where traditional accuracy metrics can be misleading. The formula to compute Kappa is given in eq 2.

$$K = \frac{P_o - P_e}{1 - P_e} \tag{2}$$

where P_o is the observed agreement (accuracy), *i. e.* how often the raters agree and P_e is the expected agreement, *i. e.* how often raters would agree by chance. P_o and P_e are computed from the confusion matrix using eq 3 and eq 4 respectively.

$$P_o = \frac{TP + TN}{T} \tag{3}$$

$$P_e = \{\frac{TP + FP}{T} * \frac{TP + FN}{T}\} + \{\frac{TN + FN}{T} * \frac{TN + FP}{T}\} \tag{4}$$

$$T = TP + TN + FP + FN \tag{5}$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative.