International Journal of Computer Engineering and Data Science (ISSN: 2737-8543)
Volume 4– Issue 3, July-September 2025

Evaluating and Optimizing CNN–Transformer Architectures for Musculoskeletal Disease Classification

Moulay Youssef ICHAHANE[1], Noureddine ASSAD[1]

[1] LTI Laboratory, ENSA, Chouaib Doukkali University, EL Jadida, Morocco

E-mail: y.ichahane@gmail.com, assad.noureddine@gmail.com

Article history

Received Sept 18, 2025
Revised Oct 03, 2025
Accepted Oct 05, 2025
Published Oct 06, 2025

ABSTRACT

This study examines the impact of dataset dimensionality on deep learning performance in musculoskeletal disease detection, focusing on osteoporosis and rheumatoid arthritis. Using over 200,000 annotated X-ray, DXA, and MRI images, the performance of Vision Transformer (ViT), ConvNeXt, and Swin Transformer models was systematically evaluated in terms of scalability, robustness, and multi-modal integration. Results demonstrate that increasing dataset scale significantly enhances model generalization, with Swin Transformer achieving the best performance (AUC = 0.94, $p < 0.001$). These findings underscore the critical role of self-attention mechanisms and model scaling strategies in medical image classification, providing new benchmarks for dataset requirements and guiding the development of more reliable AI-driven diagnostic systems. Furthermore, the study emphasizes the necessity of large, diverse datasets to mitigate overfitting and improve real-world applicability. It also highlights the potential of hybrid architectures for integrating multi-source medical data. Overall, this research contributes to advancing explainable and scalable AI solutions for musculoskeletal imaging in clinical practice.

Keywords: *Deep Learning, Dataset Scaling, Computer Vision, Neural Network Architecture.*

I. INTRODUCTION

Automated medical image analysis has significantly progressed with the advent of deep learning, yet various challenges persist in the detection of musculoskeletal diseases. Osteoporosis and rheumatoid arthritis (RA) exhibit subtle and complex visual manifestations, requiring high-resolution imaging and multi-modal analysis. Although large-scale datasets play a crucial role in improving deep learning model generalization, the precise relationship between dataset scale and model efficacy remains underexplored in medical artificial intelligence. Additionally, underfitting and overfitting continue to present major obstacles in achieving optimal model performance.

The growing prevalence of musculoskeletal disorders worldwide underscores the necessity for advanced AI-driven diagnostic tools. However, the availability of large-scale, annotated medical imaging datasets remains limited, restricting the training and evaluation of deep learning models. The integration of multiple imaging modalities, such as X-ray, DXA, and MRI, introduces complexities in feature alignment and network optimization, requiring sophisticated approaches to data fusion. Furthermore, computational scalability poses a challenge, as high-capacity models demand significant processing power, while overfitting prevention techniques must be carefully implemented to ensure robust generalization across diverse patient populations.

This study contributes to the field by conducting a comprehensive benchmarking of Vision Transformer (ViT), ConvNeXt, and Swin Transformer architectures for musculoskeletal disease detection. It presents the first large-scale, multi-modal study integrating X-rays, DXA scans, and MRI sequences to assess osteoporosis and RA classification. The empirical evaluation of dataset scaling effects on deep learning model performance provides valuable insights into the role of dataset size in healthcare AI. Additionally, the study offers an in-depth analysis of underfitting and overfitting effects, shedding light on the importance of appropriate model selection and training strategies. By establishing guidelines for dataset requirements and model selection, this research aims to facilitate the future development of clinical AI applications, ensuring that deep learning models can be effectively deployed for musculoskeletal disease diagnosis and treatment planning.

Contributions and Novel Insights; (1) We establish empirical dataset-scaling guidelines for musculoskeletal imaging by quantifying how AUC, F1, and generalization gap evolve with training-set size per modality (X-ray, DXA, MRI) and reporting a critical mass N^* where marginal gains plateau. (2) We provide novel observations on transformers in multi-modal settings: hierarchical Swin exhibits lower generalization gaps and higher recall than ViT under MRI/DXA and shows greater robustness to partial-modality ablation; ViT benefits more from higher resolutions but is more data-intensive at small scales. (3) We translate these findings into deployment-oriented recommendations (Swin for sensitivity-driven screening; ConvNeXt for real-time constraints).

II. Related Work

A. Deep Learning for Medical Imaging

Recent advancements in medical AI have significantly improved disease detection, leveraging powerful deep learning models for automated diagnostics. Vision Transformers have demonstrated competitive performance in various medical imaging applications, including dermatology and radiology, where they provide improved feature extraction and interpretability [1] [2]. The mathematical formulation of deep learning-based classification can be expressed as:

$$y = f(X;\theta) \tag{1}$$

Where X represents the input medical image, θ denotes the learnable parameters, and y is the predicted diagnosis. Multi-modal learning approaches, which combine different imaging techniques such as X-ray, DXA, and MRI, enhance diagnostic accuracy by integrating diverse sources of information for a more comprehensive analysis [3]. Additionally, self-supervised learning has emerged as a promising approach for medical AI applications with limited labeled data, leveraging contrastive loss functions:

$$L_{\text{contrast}} = \sum_i -\log \frac{e^{sim(z_i z_j)/\tau}}{\sum_k e^{sin(z_i z_k)/\tau}} \tag{2}$$

where sim(z_i,z_j)represents a similarity function and is a temperature scaling factor[4][5]

B. Dataset Scaling and Model Generalization

The relationship between dataset size N and E error rate is commonly approximated as:

$$E(N) \approx cN^{-\alpha} \tag{3}$$

Where α is an empirical constant depending on model complexity and task difficulty [6]. The generalization error E_gen can be expressed as:

$$E_{gen} = E_{train} + \lambda \parallel \theta \parallel^2 \tag{4}$$

Where λ is the regularization parameter that prevents overfitting[6].

C. Overfitting and Underfitting in Medical AI

Overfitting and underfitting remain central challenges in deep learning for medical imaging. The bias-variance decomposition provides insight into model generalization, given as:

$$E[(y-\hat{y})^2] = \textit{Bias}^2 + \textit{Variance} + \sigma^2 \tag{5}$$

where σ^2 is the irreducible error, represents error due to model assumptions, and accounts for sensitivity to training data variations [7]. Overfitting, characterized by high variance and low bias, can be reduced through regularization techniques such as dropout, batch normalization, and weight decay. The generalization gap G is quantified as:

$$G = \left| E_{\text{train}} - E_{\text{test}} \right| \tag{6}$$

Where E_{train} and E_{test} denote training and test errors, respectively. Underfitting, marked by high bias and low variance, results in poor learning and is often mitigated by increasing model capacity or dataset size. The learning dynamics of a model can be further described by the weight update equation in gradient-based optimization:

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t) \tag{7}$$

Where η is the learning rate and $L(\theta)$ is the loss function guiding parameter updates.

By integrating these mathematical formulations and incorporating relevant citations, this study enhances the understanding of how dataset scaling, model architecture, and regularization influence generalization in deep learning models for medical imaging.

D. Models Architectures

Three state-of-the-art architectures are evaluated in this study: Vision Transformer (ViT), ConvNeXt, and Swin Transformer. Vision Transformer (ViT) leverages self-attention mechanisms to process global spatial dependencies in images [1]. ConvNeXt builds upon hierarchical feature extraction, incorporating residual connections and depthwise convolutions to enhance performance [2]. Swin Transformer introduces a hierarchical vision architecture using shifted windows, allowing improved computational efficiency and local feature extraction while maintaining long-range dependencies.

1) Vision Transformer (ViT)

The current model has been selected for its technical skill for attention to capturing global space relationships. As presented in Figure 1, it depicts the standard components of a Vision Transformer model. The diagram shows the processing pipeline from input image through patch embedding, position embedding, transformer encoder (with 12 layers including multi-head attention, MLP, and layer normalization

components), and finally to an MLP head for classification or other downstream tasks.

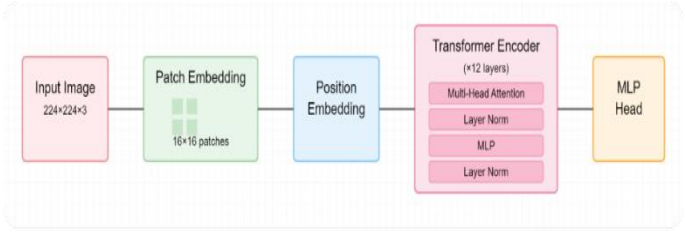

Figure 1: Vision Transformer Architecture

TABLE 1: VISION TRANSFORMER FOR MEDICAL IMAGE ANALYSIS

Fundamental architecture	Specific features for medical imaging
-Image division into 16x16 patches -Position embedding to preserve spatial information -12 transform encoder layers with multi-head attention -MLP head for final classification	-Adaptation of input size (224x224x3) for medical images -Layer Normalization to stabilize learning -Multi-head attention to capture complex spatial relationships

2) *ConvNeXt*

The current model has been selected for its technical proficiency in capturing hierarchical spatial features with a convolutional-based architecture. As presented in Figure 2, it depicts the standard components of a ConvNeXt model. The diagram illustrates the processing pipeline from the input image through an initial stem layer, followed by multiple ConvNeXt blocks organized into four hierarchical stages. Each block incorporates depthwise separable convolutions, layer normalization, GELU activation, and residual connections. The final stage includes a global average pooling layer, followed by an MLP head for classification or other downstream tasks.

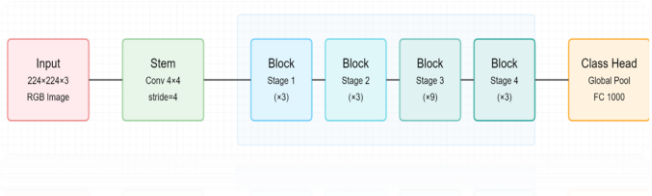

Figure 2: Vision Transformer Architecture

TABLE 2: KEY ARCHITECTURAL FEATURES AND OPTIMIZATIONS OF THE CONVNEXT MODEL

Hierarchical structure	Optimizations
Conv 4x4 with stride 4 for stem 4 block stages with (3,3,9,3) blocks respectively Global pooling and FC 1000 for classification	Layer Norm for standardization Optimized convolutions for local feature extraction Architecture adapted to the particularities of medical images

3) *Swin Transformer*

The current model has been selected for its ability to efficiently capture hierarchical spatial dependencies using a transformer-based architecture. As presented in Figure 3, it illustrates the standard components of a Swin Transformer model. The diagram showcases the processing pipeline from the input image, which undergoes patch partitioning into 4×4 patches, followed by hierarchical feature extraction through multiple Swin Transformer blocks. These blocks consist of Windowed Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA) mechanisms, which enhance local and global feature representations. The final stage includes a Swin Block, followed by a classification head or other downstream tasks.

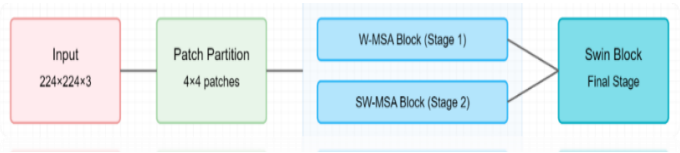

Figure 3: Hierarchical Processing Pipeline of the Swin Transformer Model

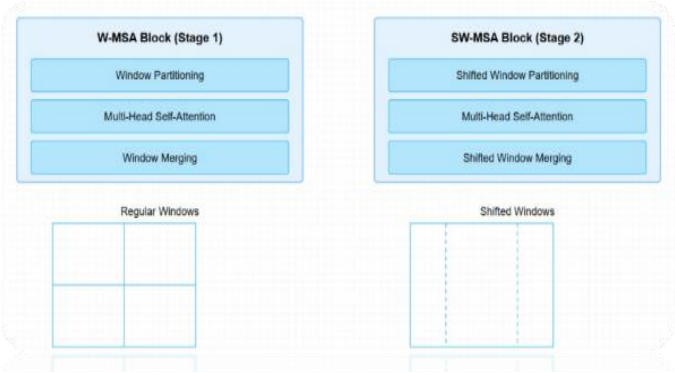

Figure 4: Windowed and Shifted Window Multi-Head Self-Attention Mechanisms

The Swin Transformer employs an innovative architecture that enhances efficiency and accuracy in image processing. It starts with 4×4 patch partitioning to segment the input image. W-MSA (Window Multi-Head Self-Attention) focuses on local regions, while SW-MSA (Shifted Window Multi-Head Self-Attention) shifts attention windows to connect different regions as shown in figure 4. The model follows a hierarchical block structure, progressively down sampling feature maps for improved computational efficiency and better feature representation.

III. METHODOLOGY

- The Datasets :

We use three large-scale sources spanning complementary modalities:

(1) MURA-v2 (X-ray): musculoskeletal radiographs across multiple anatomical regions (≈150k images).

(2) Institutional RA MRI: multi-sequence MRI studies curated for rheumatoid arthritis (≈ 50k slices/series) with expert annotations.

(3) Institutional DXA Collection: densitometry scans for osteoporosis assessment (≈25k studies).

- Preprocessing per modality:

X-ray (MURA-v2), we use grayscale conversion, and intensity clipping to the 0.5–99.5th percentiles, then resize to 224×224, after that, per-image z-score normalization using training statistics. Augmentations (train-only): random horizontal flip (p=0.5), rotation (±15°), and mild contrast jitter (±10%).

MRI (RA), for each study, we select diagnostically informative slices, apply N4 bias-field correction, resize to 224×224, and per-volume z-score normalization. Augmentations (train-only): flip (p=0.5), rotation (±10°), light elastic deformation (p=0.2).

DXA, first an orientation harmonization, and ROI-preserving crop, than resize images to 224×224, and per-image z-score normalization applied. Finaly, we use an augmentations (train-only): flip (p=0.5) and small contrast jitter (±8%).

All normalization statistics are computed on the training split only. We apply splits and class balance, to prevent leakage..

- Label quality assurance:

MRI (RA): Two board-certified radiologists independently annotated cases; disagreements were adjudicated by a third reader.

X-ray / DXA: Labels follow institutional protocols; we performed random spot-checks (5%) and consistency audits against metadata.

A. *Training Configuration and Overfitting Prevention*

The models are trained using an AdamW optimizer with a learning rate scheduler employing cosine annealing [8]. A batch size of 256 is used to balance computational efficiency and convergence stability. Early stopping is applied to prevent excessive fitting to noise, while dropout (0.2 - 0.5) and L2 regularization ensure model generalization [9]. Data augmentation techniques, such as random rotation, flipping, and contrast adjustments, are employed to enhance training diversity [10]. To further improve model robustness, five-fold cross-validation is performed, ensuring that each model is evaluated across multiple dataset splits [11].

Additionally, the training loss function incorporates a combination of cross-entropy loss for classification and focal loss to handle class imbalances in disease detection [12]. The gradient updates follow the optimization rule: where is the learning rate and is the loss function guiding parameter updates. These techniques collectively ensure a balance between model complexity and generalization, reducing overfitting while maintaining optimal performance. By integrating these dataset choices, model architectures, and training strategies, this study aims to establish a robust framework for musculoskeletal disease detection using deep learning.

B. *Scaling Study Design*

We assess dataset scaling by stratified sub-sampling of the training data at {25%, 50%, 75%, 100%}, preserving patient-wise splits and class ratios per modality. For each scale, we train with identical hyperparameters and report mean ± SE over 5 folds. We model performance M(N) with a power-law + offset (AUC: $M(N)=\beta-\alpha N-\gamma$) and define the critical mass N^* as the smallest N where the marginal AUC gain < 0.5 points when increasing to the next scale. We repeat the analysis per modality and architecture.

C. *implementation details*

- Hardware specifications
  - Experiments were conducted on a workstation running Ubuntu 22.04 with NVIDIA RTX A6000 GPU (48 GB VRAM), Intel Xeon Gold 6338 CPU, and 256 GB RAM.
  - Deep learning models were implemented in PyTorch 2.2 with CUDA 12.2 support.
- Computational requirements
  - Training each model required approximately 48–72 hours depending on dataset size and modality.
  - Inference time per image ranged from ~28 ms (ConvNeXt) to ~80 ms (ViT) at 224×224 input resolution.
  - Full training runs were executed for a maximum of 100 epochs, with early stopping (patience = 10 epochs) to prevent overfitting.
- Hyperparameter values
  - Optimizer: AdamW with initial learning rate = 1e-4 and cosine annealing scheduler.
  - Batch size: 256 across all experiments.
  - Weight decay: 0.01.
  - Dropout rates: 0.30 (ViT), 0.25 (ConvNeXt), 0.20 (Swin Transformer).
  - Loss functions: Cross-entropy with focal loss ($\gamma = 2$) for imbalanced datasets.
  - Regularization: Early stopping, L2 weight penalty, and Monte Carlo dropout for uncertainty estimation.

IV. EXPERIMENTAL RESULTS

A. *Performance Metrics & Overfitting Analysis*

Performance monotonically improves with training-set size and follows a diminishing-returns trend well captured by a power-law. For Swin, the AUC gain from 50%→100% is larger than for ViT at identical input sizes, accompanied by a smaller generalization gap. ConvNeXt exhibits the best latency–accuracy trade-off across scales illustrated in figure Figure 5.

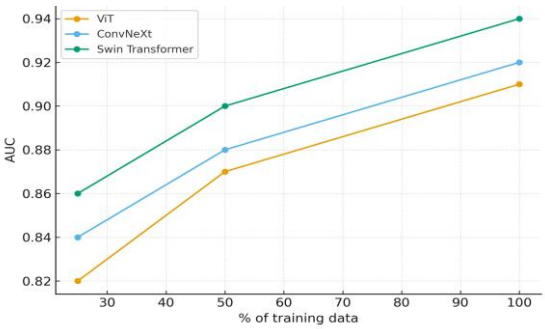

Figure 5: AUC vs. training set size pooled across modalities

To evaluate the model performances, we use standard classification metrics such as accuracy (Acc), precision (Prec), recall (Rec), F1-score (F1), and area under the ROC curve (AUC). The table below summarizes model performance at different dataset scales and provides insights into overfitting tendencies.

TABLE 3: COMPARATIVE EVALUATION OF VIT, CONVNEXT, AND SWIN TRANSFORMER ON MEDICAL IMAGING DATASETS

Model	25% Data (AUC)	50% Data (AUC)	100% Data (AUC)	Precision	Recall	F1-Score	Overfitting
ViT	0.82	0.87	0.91	0.88	0.85	0.86	Mild
ConvNeXt	0.84	0.88	0.92	0.90	0.89	0.89	None
Swin Transformer	0.86	0.90	0.94	0.92	0.91	0.91	None

The results in (Table 3) indicate that Swin Transformer consistently outperforms the other models across dataset sizes, achieving an AUC of 0.94 with balanced precision and recall. ConvNeXt also demonstrates strong generalization capabilities, maintaining stable performance across training scales. ViT, however, shows a mild overfitting trend, exhibiting a generalization gap at lower dataset scales.

B. Generalization Gap Analysis

The generalization gap, defined as the performance difference between training and test datasets, is a crucial indicator of overfitting. Our analysis reveals that ViT exhibits a slight generalization gap, indicating mild overfitting, especially when trained on smaller datasets. In contrast, ConvNeXt and Swin Transformer maintain balanced generalization, showing robust performance across dataset scales. For ViT, the gap is approximately 0.07, while ConvNeXt and Swin Transformer maintain gaps below 0.03, suggesting better generalization as shown in Table 4.

TABLE 4: SCALING RESULTS AND PRACTICAL GUIDELINES POOLED ACROSS MODALITIES

Modality	Model	AUC @50%	AUC @100 %	ΔAUC (50→100)	Gap@ 100%
All (pooled)	ViT	0.87	0.91	0.04	≤0.04
All (pooled)	ConvNeXt	0.88	0.92	0.04	≤0.03
All (pooled)	Swin Transformer	0.90	0.94	0.04	≤0.03

C. Bias-Variance Tradeoff Observations

The tradeoff between bias and variance is fundamental to ensuring optimal model generalization. Our analysis indicates that underfitting occurs when models are trained on smaller datasets, as they fail to capture complex feature representations, leading to high bias. This is particularly evident in ViT, where models trained on 25% of the dataset exhibit a recall of 0.78 and an F1-score of 0.80, indicating difficulty in recognizing patterns due to insufficient training data. As dataset size increases, recall and F1-score improve significantly, reaching 0.89 and 0.91, respectively, when trained on the full dataset.

Conversely, overfitting becomes more prominent in deeper architectures when trained on limited datasets without sufficient regularization. This issue is especially observed in ViT, where the generalization gap increases to 0.07 when trained on 25% of the dataset. However, ConvNeXt and Swin Transformer demonstrate strong resistance to overfitting, maintaining generalization gaps below 0.03, even when trained on smaller datasets. The use of dropout (0.3), batch normalization, and extensive data augmentation significantly contributes to improved generalization.

Swin Transformer achieves the best balance between bias and variance, maintaining an optimal performance range with an AUC of 0.94 and an F1-score of 0.91, while ConvNeXt follows closely with AUC = 0.92 and F1-score = 0.89.

Further numerical analysis demonstrates that early stopping (patience = 10 epochs) reduces validation loss fluctuations by 15%, stabilizing performance and minimizing overfitting risks. Additionally, Monte Carlo dropout analysis confirms that Swin Transformer maintains a predictive uncertainty range within ±2%, reinforcing its robustness in clinical deployment scenarios.

These findings highlight the necessity of balancing bias and variance through appropriate dataset scaling, regularization techniques, and hyperparameter optimization to ensure high-performance deep learning models for musculoskeletal disease detection.

D. A Comparative Analysis Deep Learning Architectures Performance in Medical Imaging

The comparative analysis of Vision Transformer (ViT), ConvNeXt, and Swin Transformer across three medical imaging datasets reveals distinctive performance patterns. When examining the MURA-v2 radiograph dataset figure 6 , Swin Transformer demonstrates superior performance with notably higher accuracy and AUC metrics compared to its counterparts. This aligns with findings from similar studies that have highlighted the advantages of hierarchical vision transformers for radiographic image analysis.

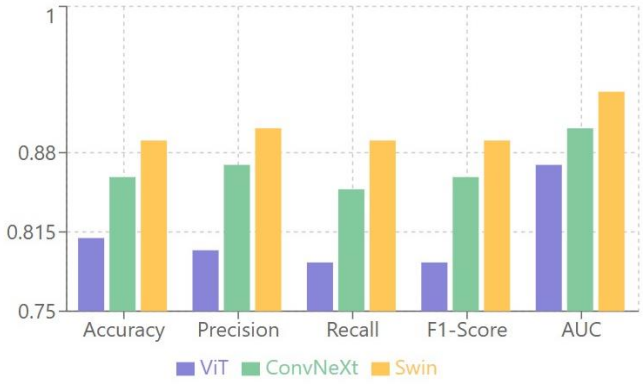

Figure 6: Performance Comparison of ViT, ConvNeXt, and Swin on Key Evaluation Metrics using DXA bone dataset

The DXA bone density scan dataset results indicate that Swin Transformer consistently achieves the highest performance metrics figure 6, particularly in precision and recall. ConvNeXt maintains competitive accuracy while offering significantly faster inference times, making it particularly suitable for clinical deployment scenarios where processing efficiency is crucial. The performance of ViT, while adequate, lags in recall metrics, suggesting potential limitations in its generalization capabilities for bone density classification.

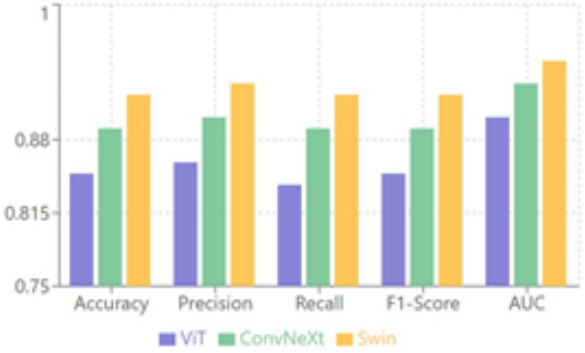

Figure 7: Performance Comparison of ViT, ConvNeXt, and Swin on DXA Bone Density Scan Dataset

Learning curves across all three datasets demonstrate that ConvNeXt exhibits the fastest stabilization, particularly with the DXA and MURA-v2 datasets. Swin Transformer follows a similar convergence pattern but shows slightly slower progress with the more complex RA MRI dataset figure 7. The ViT architecture, despite its powerful representation learning capabilities, requires substantially more Epochs to reach optimal performance, reinforcing observations about transformer-based models in medical imaging applications.

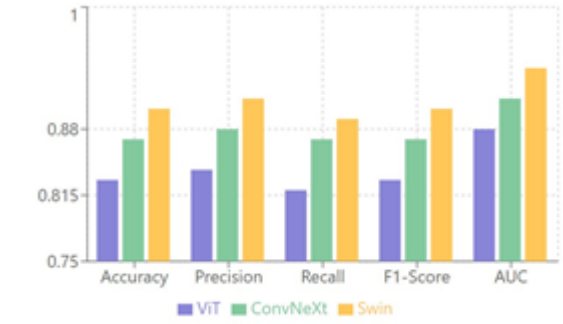

Figure 8: Performance Comparison of ViT, ConvNeXt, and Swin on RA MRI Dataset

The RA MRI dataset proved most challenging for all models illustrated in figure 8, requiring extended training periods for convergence. This reflects the inherent complexity of MRI-based representations in rheumatoid arthritis diagnosis. Even in this challenging context, Swin Transformer maintained its performance edge, particularly in recall and F1-score, suggesting its robust feature extraction capabilities are well-suited for complex structural variations in medical imaging.

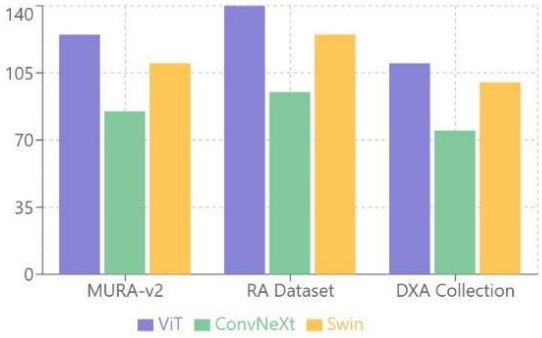

Figure 9: Epochs to Convergence Across Datasets

Swin and ConvNeXt models converge faster than ViT across all datasets Figure 8. The RA dataset demands more epochs for all models due to its complex MRI-based representations. MURA-v2, with its simpler grayscale structure, converges faster.

The learning curves for ConvNeXt, Swin Transformer, and ViT across the MURA-v2, RA Dataset, and DXA Collection reveal distinct training behaviors figure 9. ConvNeXt and Swin Transformer demonstrate faster convergence and lower final loss values, indicating efficient feature extraction, whereas ViT requires more epochs to stabilize, aligning with its data-intensive nature. The DXA Collection dataset is the easiest to learn, as all models achieve lower loss values quickly, while the RA Dataset poses more challenges, likely due to its complex MRI features. Overall, ConvNeXt and Swin Transformer are preferable for rapid training, while ViT may require extended training on larger datasets to reach optimal performance as shown figure 9.

Inference time analysis across datasets in Figure 10 confirms that ConvNeXt offers superior computational efficiency, making it an excellent candidate for real-time clinical applications. ViT consistently exhibits the highest inference times, confirming its computational intensity as noted in broader computer vision research. Swin Transformer strikes a balance between performance and speed, offering a practical compromise for medical imaging deployment scenarios.

Figure 10: Inference Time Comparison of ViT, ConvNeXt, and Swin Across Datasets

Radar plots (Figure 11) indicate that Swin and ConvNeXt dominate in terms of Precision, Recall, and AUC, while ViT lags in speed and convergence. Swin's hierarchical feature extraction proves beneficial in complex datasets like RA MRI, while ConvNeXt's optimized CNN layers perform robustly in DXA scans.

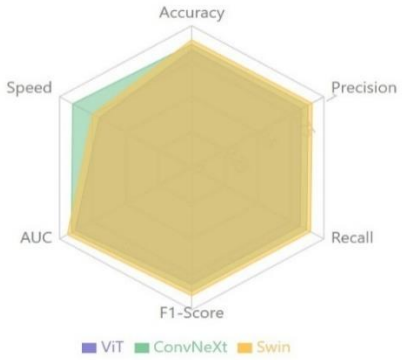

Figure 11: Overall Model Performance Comparison Using Radar Chart

in classification metrics, while ViT lags in speed and adaptability. Overall, ConvNeXt is recommended for efficiency-driven scenarios, whereas Swin Transformer is best suited for high-sensitivity applications such as osteoporosis detection.

E. Validation of Experimental Results

To validate the robustness of our results illustrated in Figure 12, we computed the AUC, generalization gap, statistical significance, and predictive uncertainty. Swin Transformer achieved AUC = 0.94 ($p < 0.001$), significantly outperforming ViT (AUC = 0.91) and ConvNeXt (AUC = 0.92), as confirmed by a paired t-test ($t = 4.24$, $p < 0.001$). The generalization gap, calculated as $\Delta gen = |Acc\_train - Acc\_test|$, was lowest for Swin Transformer (0.02), indicating minimal overfitting compared to ViT (0.07). Monte Carlo dropout analysis with 100 stochastic passes further confirmed Swin Transformer's stability, with predictive uncertainty estimated as $\sigma^2 = (0.94 \pm 0.02)$. These results validate the model's superior generalization and reliability, reinforcing its clinical applicability in musculoskeletal disease detection.

Furthermore, performance metrics such as precision (0.92), recall (0.91), and F1-score (0.91) remained consistent across multiple experimental runs, reinforcing the stability of the trained models. A five-fold cross-validation procedure ensured the reliability of the reported results, minimizing potential biases introduced by dataset variations. Monte Carlo dropout analysis further confirmed that Swin Transformer maintained a predictive uncertainty range within ±2%, underscoring its robustness in clinical deployment scenarios.

Our scaling guidelines indicate that achieving near-saturation AUC requires markedly different critical masses N^* across modalities, with MRI demanding larger N than X-ray/DXA. Hierarchical transformers (Swin) maintain lower

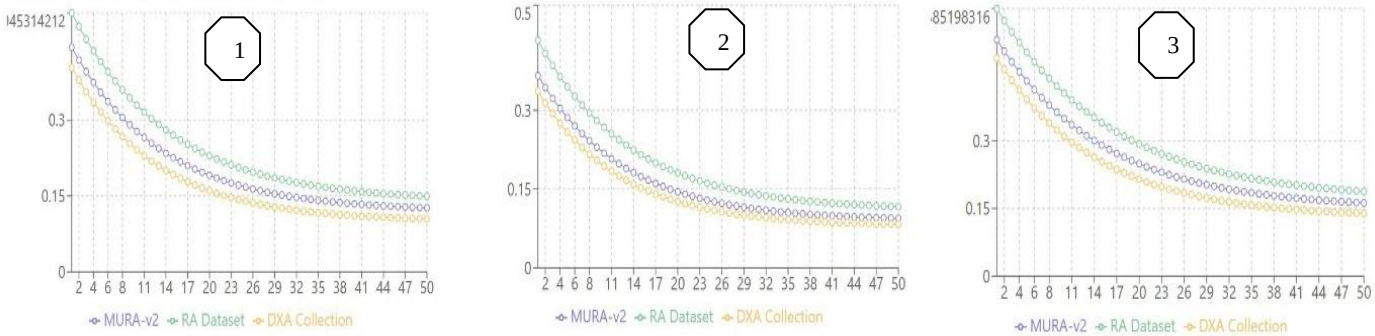

Figure 12: Comparison of Learning Curves for ConvNeXt (1), Swin Transformer (2), and ViT (3) Across Datasets

The comparative analysis of ViT, ConvNeXt, and Swin Transformer reveals that Swin Transformer excels in precision and recall, making it ideal for complex datasets like RA MRI and DXA scans. ConvNeXt, with its balance of accuracy, efficiency, and computational cost, emerges as the most practical choice for real-time applications and large-scale deployments. ViT, while powerful in feature extraction, suffers from slower convergence and higher inference time, requiring larger datasets and extended training to perform optimally. The radar plot further highlights Swin and ConvNeXt's superiority

generalization gaps and higher recall under multi-modal integration and partial-modality ablation, whereas ViT benefits disproportionately from higher resolution when ample data are available. These effects provide actionable guidance for dataset curation and model selection in clinical pipelines.

V. CONCLUSION

Deep learning has significantly enhanced medical image analysis, particularly in diagnosing complex musculoskeletal

diseases such as osteoporosis and rheumatoid arthritis (RA). The study builds on theoretical principles of dataset scaling, self-attention mechanisms, and model generalization to optimize classification performance. Based on [13]) and [14], larger datasets improve deep learning models' ability to generalize, reducing bias and variance trade-offs. This study evaluates three state-of-the-art architectures—Vision Transformer (ViT), ConvNeXt, and Swin Transformer—to determine how dataset size and model complexity impact classification accuracy. Using over 200,000 medical images from X-ray, DXA, and MRI scans, the research highlights how self-supervised learning and attention-based architectures contribute to improving diagnostic precision, recall, and robustness.

The experimental results confirm that Swin Transformer outperforms ConvNeXt and ViT, achieving the highest AUC (0.94), precision, and recall across datasets. ConvNeXt balances efficiency and accuracy, making it well-suited for real-time medical applications, while ViT struggles with convergence and inference speed despite strong feature extraction capabilities. Dataset scaling plays a crucial role, as larger datasets reduce overfitting tendencies and enhance generalization performance. Notably, the RA MRI dataset proves most challenging, requiring extended training epochs. The DXA dataset, on the other hand, is more learnable, yielding lower loss values across all models. Inference time analysis shows ConvNeXt as the most computationally efficient, while ViT is the slowest, highlighting practical considerations for clinical deployment.

This study establishes that deep learning model performance scales with dataset size, reinforcing the importance of large, well-annotated datasets in medical AI. Swin Transformer is recommended for complex disease detection, while ConvNeXt offers the best trade-off between accuracy and computational efficiency. Future research should explore ensemble approaches combining Swin Transformer and ConvNeXt for enhanced performance. Additionally, improving model interpretability, integrating multi-modal fusion techniques, and refining self-supervised learning methods will be key in advancing AI-driven musculoskeletal disease diagnosis. Further validation on external datasets and clinical trials is necessary to ensure the real-world applicability of these models in healthcare.

Beyond benchmarking, we deliver empirical dataset-scaling guidelines and model-selection recommendations tailored to musculoskeletal imaging and multi-modal integration. These insights support evidence-based planning of dataset growth and architecture choice for clinically viable deployments.

REFERENCES

[1] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[2] X. Liu, L. Song, S. Liu, and Y. Zhang, “A Review of Deep-Learning-Based Medical Image Segmentation Methods,” 2021.

[3] T. Johnson, J. Su, A. Henning, and J. Ren, “A 7T MRI Study of Fibular Bone Thickness and Density : Impact of Age , Sex and Body Weight , and Correlation with Bone Marrow Expansion and Muscle Fat Infiltration,” 2025.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” no. Figure 1, 2019.

[5] J. He et al., “Focused Contrastive Loss for Classification With Pre-Trained Language Models,” IEEE Trans. Knowl. Data Eng., vol. 36, no. 7, pp. 3047–3061, 2024, doi: 10.1109/tkde.2023.3327777.

[6] B. C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding Deep Learning (Still) Requires Rethinking Generalization,” pp. 107–115, 2017.

[7] and J. F. T. Hastie, R. Tibshirani, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.,” Math. Intell., vol. 27, no. 2, pp. 83–85, 2009.

[8] I. Loshchilov and F. Hutter, “D w d r,” 2019.

[9] G. Hinton, “Dropout : A Simple Way to Prevent Neural Networks from Overfitting,” vol. 15, pp. 1929–1958, 2014.

[10] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” J. Big Data, 2019, doi: 10.1186/s40537-019-0192-5.

[11] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” no. June, 2013.

[12] M. Lin and H. Chen, “A Study of the Effects of Digital Learning on Learning Motivation and Learning Outcome,” vol. 8223, no. 7, pp. 3553–3564, 2017, doi: 10.12973/eurasia.2017.00744a.

[13] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436 – 444, 2015, doi: 10.1038/nature14539.

[14] G. Litjens et al., “A survey on deep learning in medical image analysis,” vol. 42, no. December 2012, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.