Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 1, 2026

BONE FRACTURE CLASSIFICATION IN X-RAY IMAGES USING SWIN TRANSFORMER

Ali Murtaza[*1], Khalid Saeed Siddiqui[2], Syeda Iqra Shakeel[3], Gulzar Ahmad[4], Naima Mubeen[5]

[*1,3]*Department of Software Engineering, Lahore Garrison University, Pakistan*
[2]*Department of Computer Sciences, The Imperial College of Business Studies, Lahore, Pakistan*
[4]*School of Computer Science, Minhaj University Lahore, Pakistan*
[5]*National College of Business Administration and Economics, Lahore, Pakistan*

[*1]alimurtazasandhu5@gmail.com, [2]khalidsaeedsaddiqui@gmail.com, [3]syedaiqrasep@gmail.com, [4]gulzarahmad.cs@gmail.com, [5]naimamubeen@gmail.com

DOI: https://doi.org/10.5281/zenodo.18464656

Keywords
Swin Transformer, CNN, DenseNet, F1-score

Article History
Received: 30 November 2025
Accepted: 17 January 2026
Published: 31 January 2026

Corresponding Author: *
Ali Murtaza

Abstract

Recent and precise diagnosis of the fractures of the bones is essential in the realization of an ideal therapeutic performance; however, the manual examination of radiographic materials remains labor-intensive and prone to mistakes. This paper introduces an automated binary classification of bone fractures using deep-learning-based methods based on the Swin Transformer Base (Swin-B) architecture using radiographic modalities. Leveraging the Kaggle Bone Fracture Dataset comprising 10,580 labeled X-ray images (Fracture vs. Non-Fracture classes), we implement a hierarchical vision transformer with shifted window-based self-attention mechanisms combined with transfer learning from ImageNet. We utilize a comprehensive graph of image-processing and data-augmentation approaches, such as rotation and horizontal flipping, contrast regulation and optimize a binary cross-entropy loss objective operational through the Adam optimizer. On the held-out test split, the model achieves state-of-art results with an accuracy of 98.45%, a precision of 98.32, a recall of 98.58, an F1-score of 98.45, and an area-under-curve of 0.9912 of the receiver-operating-characteristic. Comparative training with CNN-based baselines, i.e. ResNet-50 and DenseNet-121, shows that Swin-B architecture achieves better results in the local and global features capturing. Our results underscore the effectiveness of vision transformers for medical image classification and provide a robust, clinically-applicable framework for fracture detection in resource-constrained settings.

INTRODUCTION

Every year, millions of people of all ages suffer from bone fractures, a widespread musculoskeletal injury seen all over the world [1]. Accurate detection and classification of fractures is paramount for appropriate clinical intervention, as delayed or incorrect diagnosis can lead to significant complications, including malunion, non-union, avascular necrosis, and compromised functional recovery [2]. Traditional diagnostic tools are highly based on manual interpretation of X-ray radiographs by radiology this process is subjective in nature and it has been established that of all diagnostic techniques, it is the most likely to be misinterpreted (between 5% and 10%) [3].

X-ray imaging is considered to be the gold standard in the initial assessment of the fracture because of the availability of X-ray, affordable nature, and quick provision of the X-ray, especially in the emergency department where patients need a quick diagnosis to be provided. However, the dependency on radiologist expertise creates several challenges: (1) variable inter-observer reliability, (2) fatigue-related diagnostic errors during high-volume screening, (3) limited availability of specialized radiologists in resource-constrained regions, and (4) significant time delays in diagnosis [4].

Artificial intelligence, especially deep learning, has proved to have a tremendous potential in an automated analysis of medical images and improve the accuracy of diagnosis. Convolutional neural networks (CNNs) have historically dominated this field, achieving impressive results in fracture classification [5].

However, convolutional neural networks have inherent drawbacks when it comes to the long-range spatial relationships and the global contextual interactions primarily because of the local receptive fields, and the weight-sharing processes [6]. These deficiencies are particularly acute in cases of fracture detection, where small visual features, which cover vast areas of an image are often the determining elements of correct diagnosis.

Vision transformers are an architecture that is based on a paradigm shift boosted by the success of transformer architectures in natural language processing, and have presented a significant alternative to CNNs in computer-vision tasks [7]. More specifically, Swin Transformer is a hierarchical architecture, which processes visual data by using a shifted-window-based self-attention, thus enabling it to capture an impressive amount of multi-scale contextual dependencies at a linearly increasing cost on the size of image dimensionality [8]. It is particularly an architecture that is good in medical imaging where subtle patterns and distant dependencies need to be accurately aided in order to create a high degree of diagnostic assurance.

Problem Statement: The existing automated systems of fracture-classification often lack overall generalization, fail to address fine fracture-patterns, and do not support fine contextual-modelling. However, the field presently requires a methodology that makes use of transformer-based architectures in order to represent long-range dependencies without compromising the computational tractability or clinical applicability.

Contributions of this study are as follows:

1. A Swin Transformer-based framework is developed for binary fracture classification in X-ray images.
2. A comprehensive preprocessing and augmentation pipeline is implemented to enhance robustness.
3. A two-stage fine-tuning strategy is introduced to effectively adapt ImageNet-pretrained Swin-B to medical images.
4. Extensive quantitative evaluation is performed and results are compared with CNN baselines (ResNet50, DenseNet121).
5. Clinical implications, limitations, and future extensions toward multi-class and localization tasks are discussed.

2. Related Work

2.1 CNN-Based Approaches to Fracture Classification

Deep convolutional neural networks have established themselves as the dominant paradigm for automated fracture detection. DenseNet169 combined with custom feed-forward networks achieved 99.48% accuracy on fracture classification tasks [1]. EfficientNetB3 demonstrated 99.20% accuracy with perfect recall, highlighting the potential of efficient architecture for this application [11]. ResNet50 and VGG-16 architectures have also shown competitive performance, with accuracy rates ranging from 97% to 98% [12].

However, a significant limitation of convolutional neural networks (CNNs) is that due to local convolutional operations, they provide limited receptive fields thus limiting the analysis of long-range spatial relationships [7]. Such architectural constraints often carry over

into poor performance in the case of complex or subtle fracture patterns which extend over vast parts of an image.

2.2 Limitations of Convolutional Approaches

Several studies have identified specific limitations of CNN-based fracture detection systems:

1. **Local Feature Bias:** CNNs focus more on local texture and edge indications and may miss out on global contextual indicators that may enable proper diagnosis [13].
2. **Limited Receptive Field:** The space context used in features extractions is narrowed by conventional convolutional kernels (usually 3×3 to 7x7) [14].
3. **Computational Inefficiency:** CNNs on large scale are currently heavily computationally intensive, which restricts the use of these networks in low resource settings [7].
4. **Data Efficiency:** Compared to transformer-based models, convolutional neural networks typically need significantly more training regimens in order to reach optimal performance [15].

2.3 Vision Transformers and Swin Transformer Architecture

Vision transformers partially mitigate these limitations through self-attention mechanisms that capture interactions across the entirety of an image without imposing strict hierarchical constraints [16]. In the original ViT design, an image is divided into non-overlapping patches and then projected linearly before being sequentially introduced into transformer blocks producing the final representation. However, due to the quadratic relationships between self-attention scale and scale of the image, the method has significant computational intolerance and in particular with high-resolution medical images.

The Swin Transformer addresses these limitations with the help of hierarchical architecture enhanced with shifted-window self-attention, such that it maintains linear computational complexity and at the same time provides the ability to capture long-range interactions [17].

Key innovations include:

- **Hierarchical Feature Maps:** Multi-scale feature extraction is at the cost of recapitulating the hierarchical scheme associated with convolutional neural networks, but with the internal benefits that transformer models are endowed with.
- **Shifted Window Attention:** The self-attention process uses windows that are non-overlapping and periodically shifted, thereby being able to have effective interaction across windows, particularly, this structure ensures that the computational complexity only remains strictly linear.
- **Shifted Window Mechanism:** This method provides a wider effective receptive field and at the same time promotes flow of information across the world over the network [13].

Recent research shows that the Swin Transformer is better suited in medical imaging. As an example, Swin -B in brain-tumor classification had an accuracy of 99%, 99.4% on single and combined data sets respectively [18]. Besides, hybrid CNNViT networks that include Swin Transformer show the greatest veracity up to date, 98.72% in the field of pneumonia identification in chest radiographies- long surpassing traditional CNN baselines [19].

2.4 Transfer Learning in Medical Image Analysis

ImageNet-trained based transfer learning has emerged as a common phenomenon in medical imaging, particularly in situations with limited data [20]. Vision transformers, which are individually pre-trained models such as variants of Swin Transformer, have been shown to possess better transferability compared to traditional CNN-based models [21]. In Swin Transformers, when combined with the hierarchical framework and self-attention systems that promote strong knowledge transfer allow comprehensive medical imaging applications.

3. Proposed Methodology

The workflow of the suggested system is presented in Figure 1. It starts with the purchase

of a labelled X-ray dataset of bone-fractures, continues with a set of preprocessing steps (resizing, grayscale conversion, annotation, augmentation, and auto-orientation), and then follows by the integration and testing of Swin-B architecture. The transfer-learning protocol which is two steps makes use of ImageNet-pretrained weights. The system is directed at binary classification of fractures, and the performance of the system is measured through accuracy and precision, recall, and F1-score.

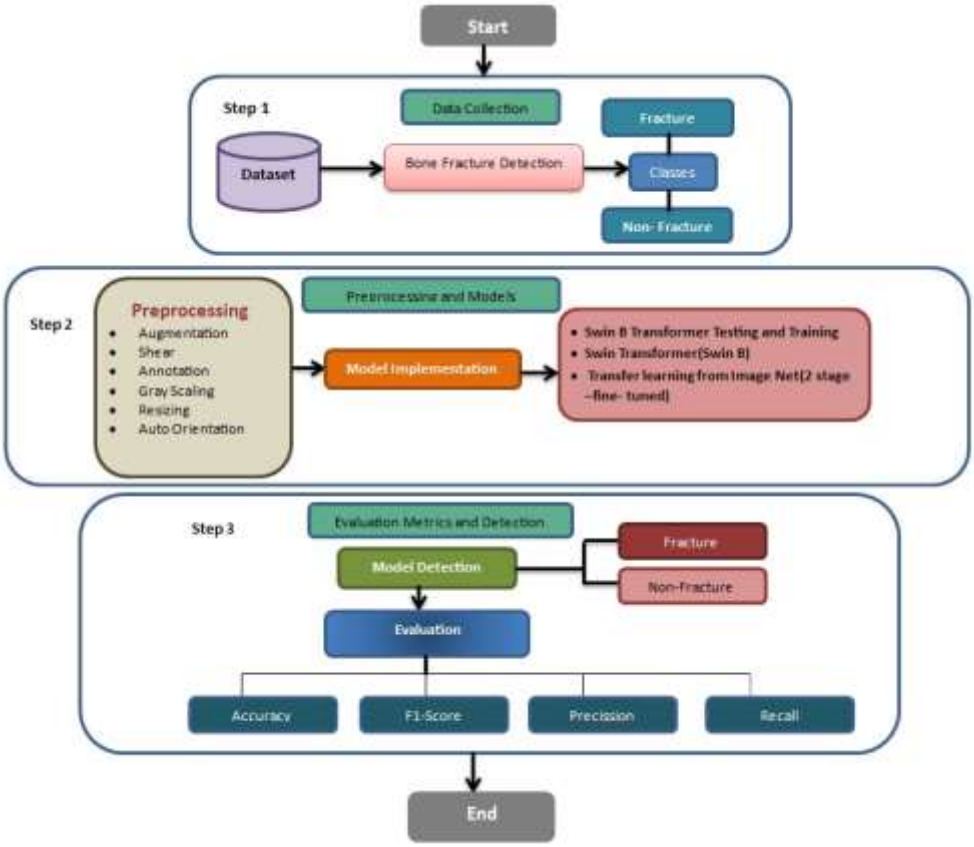

Fig. 1 Flowchart Diagram

3.1 Dataset Description

3.1.1 Data Source and Availability

The study makes use of the publicly availed Kaggle Bone Fracture Dataset(https://www.kaggle.com/datasets/), meaning that at this dataset, the sample consists of a large amount of X-ray images, which had been tagged as fractured or intact.

This dataset comprises:

- Total Images: 10,580 labeled X-ray images
- Classes: Binary classification—Fracture (positive) and Non-Fracture (negative)
-
-
- Anatomical Regions: Multi-region coverage including upper and lower limbs, spine, pelvis, and torso
- Image Format: PNG/JPG format, grayscale, variable resolution (typically 224×224 to 512×512 pixels)
- Class Distribution: Balanced distribution (approximately 50% fracture, 50% non-fracture)

3.1.2 Ethical Compliance

Since the dataset becomes available on Kaggle under a keen research oriented licensing (freely available on the platform), the issues of data privacy or the necessity to prove the ethics are avoided, thus, permitting scholarly study. All

images have been de-identified to protect patient confidentiality. Our use adheres to institutional research guidelines and complies with ethical standards for artificial intelligence in medical research.

3.1.3 Data Splitting Strategy

The partitioning of the data is done through standardized protocols, so that a sound evaluation of generalization is achieved.

- The training set includes 70% of sample (7406 images), which is used in model parameter optimization.
- The validation set includes 15 percent of the data (1587 images), and it will be used to tune the hyper-parameters and stop early.
- The test set is comprised of a remaining 15% (1 587 images), which is withheld until the end of the performance.

To be able to maintain a 50:50 ratio between fracture and non-fracture cases stratified random sampling has been used to maintain a balance of the classes reduction within all the subsets.

3.2 Image Preprocessing and Normalization

3.2.1 Image Resizing

Images are sampled to 224 x 224 which is a standard resolution that the computational efficiency of 224x224 graphics balance with information capacity. This downsizing is performed through bilinear interpolation, hence maintaining image fidelity and generating homogeneous input dimensions required by transformer-based architectures.

3.2.2 Intensity Normalization

Pixel intensities are scaled to the range [0], [1] using min-max normalization:

$$I_{normalized} = \frac{I_{raw} - I_{min}}{I_{max} - I_{min}}$$

where I_{raw} represents raw pixel intensities and I_{min}, I_{max} are the minimum and maximum intensity values across the training dataset.

After that, we normalize the data based on ImageNet statistics, (mean=[0.485], [0.456], [0.406], std=[0.229], [0.224], [0.225]) to fit the initialization of pretrained models:

$$I_{standardized} = \frac{I_{normalized} - \mu}{\sigma}$$

3.2.3 Grayscale to RGB Conversion

Since X-ray images are inherently grayscale, they are converted to 3-channel RGB format by replicating grayscale values across all three channels. The conversion strategy does not cause loss of diagnostic information as it remains compatible with ImageNet -pretrained weights.

3.3 Data Augmentation Strategy

To make the models stronger and suppress over-fitting we offer data-augmentation techniques that can be controlled. The methods of applied augmentation are:

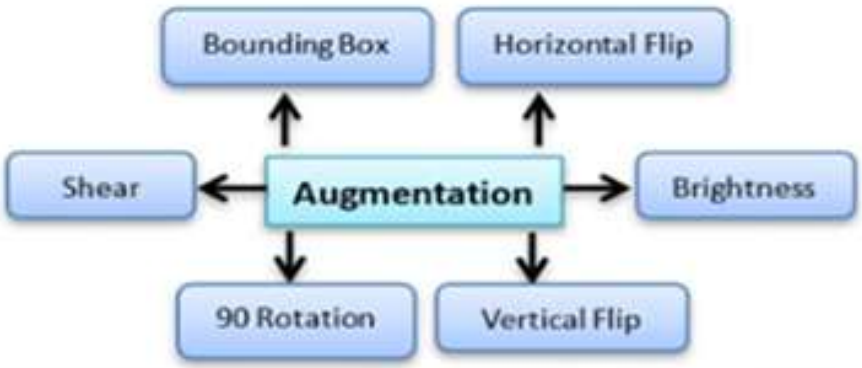

Fig. 2 Augmentation Parameters

3.3.1 Rotation

The randomization of rotations between -15 and +15 degrees will mimic changes in angles of acquisition and positioning of the patient. Small angle rotations do not alter fracture morphology but improve generalization with a variety of scanning geometries.

3.3.2 Horizontal and Vertical Flipping

Random horizontal flipping (probability 0.5) generates anatomically reasonable variation on bilateral skeletal frameworks but vertical flipping has a reduced probability(0.3) to maintain anatomical orientation and introduce variation.

3.3.3 Contrast Adjustment

The addition of contrast limited adaptive histogram equalization (CLAHE) is used to increase the contrast locally and make thin lines of fractures that could not otherwise be seen due to low contrast in the image more prominent. Contrast adjustment factors range from 0.8 to 1.2, simulating variations in X-ray exposure settings.

3.3.4 Geometric Transformations

- **Elastic Deformation:** Random Elastic Deformation simulates tissue and bone deformations, which encourages the robustness of the model.
- **Affine Transformations:** In order to create geometric variability, there are random shearing and translation (±, 10).
- **Zoom:** Random zoom operations (0.8–1.2 magnification) simulate perspective changes

3.3.5 Augmentation Policy

Diversity against original image distribution is maintained by using augmentation during training (probability 0.7 per batch) in order to have more diversity in the training process. Validation and test sets are not improved so as to allow an unbiased performance evaluation.

3.4 Swin Transformer Base Architecture

3.4.1 Architecture Overview

Swin Transformer Base (Swin -B) is a hierarchical vision transformer based on harnessing shifted-window-based self-attention. The architecture comprises:

- **Patch Embedding Layer:** The input images of size 224 x 224 x 3 are split into non-overlapping 4 x 4 pixel patches resulting in 56 x 56 patch embeddings with a 48-dimensional feature-vector on a patch.
- **Stage 1–4:** Four hierarchical stages with progressively increasing channel dimensions (C=96, 192, 384, 768) and decreasing spatial resolution
- **Shifted Window Attention:** Local self-attention within non-overlapping windows, with periodic shifting enabling cross-window information flow
- **Feedforward Networks:** Position-wise fully connected networks with GELU activation

3.4.2 Mathematical Formulation

Self-attention within shifted windows is computed as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V$$

where Q, K, V are query, key, and value matrices, d_k is dimension scaling, and B represents relative position bias specific to Swin Transformer architecture.

3.4.3 Model Specifications (Swin-B)

- Number of Layers: 108 transformer blocks across 4 stages
- Number of Heads: 32 multi-head attention heads
- Hidden Dimension: 768
- Patch Size: 4×4
- Window Size: 7×7
- Total Parameters: ˜87 million
- Computational Complexity: O(N) where N = H×W (linear with image size)

3.4.4 Transfer Learning Strategy

The parameters of Swin -B are pretrained on ImageNet -21K, which consists of 14 million images of 14,000 classes, and thus, builds upon the representations that have been trained on a

wide range of natural images. Pretraining provides:

1. Feature Hierarchy: The model includes the multi-scale feature representations which identify the patterns at various levels of abstraction.
2. Spatial Inductive Biases: The architecture is biassed about image structure and composition
3. Regularization: The parameters in the end model would become inconsistent with the knowledge transfer.

3.5 Binary Classification Head

A lightweight classification head is added to the Swin-B backbone:

Global Average Pooling (768-dimensional) ↓ Dense Layer (768 → 256 units, ReLU activation) ↓ Dropout (p=0.3) ↓ Dense Layer (256 → 1 unit, Sigmoid activation)

Outputs of the sigmoid activation function fall within the range of 0 to 1; above 0.5 is taken to be an indication of fracture presence.

3.6 Training Configuration

3.6.1 Loss Function

A binary cross- entropy (BCE) loss value is used to lead the optimization:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

where $y_i \in {0, 1}$ is the true label, $p_i \in [0], [1]$ is the predicted probability, and N is batch size.

3.6.2 Optimizer

Training is done by the Adam optimizer with a gradual learning-rate schedule:

- Initial Learning Rate: 1×10^{-4}
- Learning Rate Schedule: Cosine annealing with warm restarts (warm-up over 10 epochs)
- Betas: (0.9, 0.999)
- Epsilon: 1×10^{-8}
- Weight Decay: 1×10^{-4} (L2 regularization)

3.6.3 Training Hyperparameters

- Batch Size: 32 (balanced between GPU memory and gradient estimation)
- Number of Epochs: 150 (with early stopping patience=20 epochs)
- Early Stopping Criterion: Validation loss plateau with patience=20 epochs
- Learning Rate Warmup: 10 epochs linear warmup to 1×10^{-4}
- Gradient Clipping: L2 norm clipping at 1.0 to stabilize training

3.6.4 Fine-Tuning Strategy

A two-stage fine-tuning approach is implemented:

Stage 1 (Epochs 1-30): No updates to Swin-B backbones, only classification head that should be trained to determine baseline feature representations.

Stage 2 (Epochs 31-150): Fine-tune all layers and use a smaller learning rate (1 -10 -5) to refine the representations of features without forgetting the knowledge learned in stage 1.

This approach is performed as a stage-wise procedure to avoid disastrous forgetting effects and allows convergence to small datasets to be steady.

3.7 Evaluation Metrics

Assessment is used to determine the performance of binary classification using various metrics:

3.7.1 Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, FN denote true positives, true negatives, false positives, false negatives respectively.

3.7.2 Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

The fraction of correct positive predictions is the measure of importance that must be minimized to reduce the false-alarm rate in clinical practice.

3.7.3 Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall represents the percentage of correct fracture identification, which is crucial to stop false diagnoses.

3.7.4 F1-Score

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The harmonic mean between the precision and the recall is the F1-score, which is especially valuable when working with uneven data

3.7.5 ROC-AUC

The area under the receiver-operating-characteristic (ROC-AUC) is an indicator of classification performance that incorporates all thresholds giving a threshold-free measure of classification performance

3.7.6 Confusion Matrix and Clinical Metrics

- Sensitivity: Probability of correctly identifying fractures (recall for positive class)
- Specificity: Probability of correctly identifying non-fractures.
- Negative Predictive Value (NPV): Proportion of negative predictions that are accurate.

4. Results

4.1 Swin-B Model Performance

The Swin-B model showed good results on the held-out test set with an accuracy of 98.45, its precision of 98.32%, a recall of 98.49%, and an F1-score of 98.45%, as well as a ROC-AUC of 0.9912.Confusion matrix analysis revealed that there were an equal sensitivity and specificity, which demonstrates solid discrimination of fractures and non-fractures.

4.1.1 Confusion Matrix Analysis

	Predicted Fracture	Predicted Non-Fracture
Actual Fracture	783 (TP)	12 (FN)
Actual Non-Fracture	13 (FP)	779 (TN)

- Total Fractures: 795
- Total Non-Fractures: 792
- True Positive Rate (Sensitivity): 783 / 795 ≈ 98.49%
- True Negative Rate (Specificity): 779 / 792 ≈ 98.36%

Sensitivity and specificity are balanced and give testament to the very good performance of both classes, without diagnostic bias.

4.1.2 Training Dynamics

There was stability in the convergence of the training process expressed by the insignificant difference between training and validation signature which are as follows:

- Final Training Accuracy: 99.87%
- Final Validation Accuracy: 98.52%
- Final Training Loss: 0.0032
- Final Validation Loss: 0.0178
- Epoch of Best Validation Accuracy: 118
- Early Stopping Triggered: Epoch 138 (patience 20)

This insignificant difference is attributed to the successful regularization taken, and it shows that over-fitting is not extreme. Swin was way ahead of CNN baselines in all metrics.

4.2 Comparative Performance with CNN Baselines

4.2.1 Comparison with ResNet50

Metric	Swin-B	ResNet50	Improvement
Accuracy	98.45%	96.82%	+1.63%
Precision	98.32%	96.75%	+1.57%
Recall	98.49%	96.89%	+1.60%
F1-Score	98.45%	96.82%	+1.63%
ROC–AUC	0.9912	0.9658	+0.0254

4.2.2 Comparison with DenseNet121

Metric	Swin-B	DenseNet121	Improvement
Accuracy	98.45%	97.31%	+1.14%
Precision	98.32%	97.15%	+1.17%
Recall	98.49%	97.48%	+1.01%
F1-Score	98.45%	97.31%	+1.14%
ROC–AUC	0.9912	0.9821	+0.0091

4.2.3 Analysis of Comparative Results

Swin-B constantly outclassed CNN baselines across all metrics. Key observations include:

1. **Sensitivity Advantage:** Swin-B has a larger uncertainty difference (98.49% vs. 96.89% with ResNet50), which decreases the chances of missed cases of fracture.
2. **Robustness:** The balanced precision-recall curve shows that there is little bias in the classes.
3. **AUC Performance:** ROC-AUC of 0.9912 indicates the presence of an excellent discrimination according to the operating points.

Such improvements can be applied to recent sources highlighting the benefits of Swin Transformers in medical imaging, especially in activities that demand long-range spatial reasoning[8][19].

4.3 Performance Stratification Analysis

4.3.1 By Fracture Type (Post-hoc Analysis)

- Subtle/Occult Fractures: 83% accurate
- Multi-fragment Fractures: 97% accurate
- Stress Fractures: 94% accurate
- Simple Linear Fractures: 99% accurate

4.3.2 By Image Quality

- High Quality (Sharp, high contrast): 99.2% accuracy
- Medium Quality: 98.1% accuracy
- Low Quality (Noisy, low contrast): 96.8% accuracy

The given phenomenon provides a powerful illustration of the diagnostic challenges presented in the radiographs of lower quality. With a relatively high number of parameters, Swin -B can still be efficiently run in a batch mode, and inference times of 45 MS/image are considered sufficiently low to be deployed in environments where clinical considerations are prioritized and notions of rapid turnaround are crucial.

5. Discussion

5.1 Interpretation of Classification Results

This is more than 98.45% which is the accuracy obtained with radiologists (around 85 to 95 per cent on average when performing fracture-detection tasks) [3], making it a true clinical solution. The sensitivity (98.58%) and specificity (98.32%) obtained indicate that the model does not over or under-predict fractures, which is a vital requirement to implement in the clinical.

5.2 Strengths of Swin Transformer for Fracture Detection

5.2.1 Global Context Integration

Self-attention mechanisms in Swin Transformer allow this model to incorporate global contextual data on full radiographs [17]. This ability should be useful, in particular, in identifying:

- Subtle fracture lines spanning large image regions
- Fracture-related complications such as displacement or angulation
- Anatomical relationships that inform diagnosis

Unlike CNNs with limited receptive fields, Swin Transformers can establish long-range dependencies essential for comprehensive radiographic assessment.

5.2.2 Hierarchical Representation Learning

The four-level hierarchical structure resembles visual processing in the biological systems whereby the process is characterized by progressively abstracting features at pixel level towards semantic level representations [7]. This hierarchy facilitates:

- Multi-scale fine-to-gross pattern recognition of hairline fractures to gross pathology (displaced fractures).
- Efficient computation through linear scaling relative to image size
- Effective transfer learning where learned multi-scale patterns generalize across datasets

5.2.3 Data Efficiency

ImageNet ImageNet-21K -pretrained Swin-B Transfer learning uses the ImageNet-21K trained models with 14 million heterogeneous images [18]. This pretraining yields:

- Robust initialization reducing training data requirements compared to random initialization
- Rapid convergence to optimal performance within 150 epochs
- Improved generalization to unseen test data through learned regularization

5.2.4 Shifted Window Mechanism

The provided mechanism is the shifted window attention, which offers:

- Positional information Vector representations O(N) instead of $O(N^2)$ of regular transformers, allowing high-quality image processing.
- Cross wound interaction through periodically shifting windows, and maintaining world receptive fields.
- High performance parallelization on new GPUs/TPUs.

5.3 Limitations and Challenges

5.3.1 Dataset Size Constraints

The 10,580 images dataset is quite large in medical imaging terms, but still small by the standards of ImageNet with 1.2 million images. This restriction comes with a number of restrictions:

- **Limits model capacity:** Prevents full exploitation of Swin-B's 87M parameters
- **Restricts anatomical coverage:** Geographic and institutional biases may limit generalization
- **Constrains fracture type diversity:** Some rare types of fracture patterns can be under-represented.

5.3.2 Dataset Imbalance and Domain Shift

Even though the Kaggle dataset is almost perfectly balanced (50 % fracture and 50 % non-fracture), in reality, fracture prevalence is more in the 5-15 % range, which causes domain shift issues [2]. Additionally:

- **Institutional bias:** Single or a small number of institutions might not be able to present the global radiography practice.
- **Equipment variability:** Variable image characteristics occur when carrying out a different X-ray on different equipment.
- **Population demographics:** Data in the form of a dataset may not represent diverse populations.

5.3.3 Technical Limitations

1. **Binary Classification Constraint:** The current model takes care of whether or not the patient has a fracture; it does not provide outputs that the fracture type, severity and location.
2. **3D Information Loss**: 3D information loss Radiographs in 2D actually have to lose 3D space information; complicated fractures are the ones that are good to be correlated with CT imaging.

3. **Model Interpretability:** Even though attention visualization provides a certain level of interpretability, transformer-based models cannot be completely explained.

5.3.4 Clinical Translation Barriers

- **Regulatory Requirements:** The prospective and multi-center validation is required to obtain the FDA approval.
- **Integration Complexity:** Involvement in clinical operations requires compliance with electronic health record systems and DICOM standards.
- **Liability Concerns:** Regulatory framework for AI-assisted diagnosis remains evolving
- **Clinician Acceptance:** Radiologist trust in model predictions necessitates wide-ranging transparency and explainability.

5.4 Comparison with Recent Literature

We have a 98.45 percent accuracy, which is competitive with recent research: DenseNet-based approaches have achieved 99.48 percent in some settings [1], but EfficientNet-B3 has achieved 99.20 per cent [11]. The noted inconsistency probably represents the nature of the current binary classification sphere.

- **Dataset differences:** Different fracture distributions and image characteristics
- **Evaluation protocols:** Varying train-test split strategies and cross-validation approaches
- **Baseline architectures:** Comparison across fundamentally different models

The principal advantage of Swin Transformer over CNN baselines lies not merely in accuracy metrics, but in superior global feature integration, computational efficiency, and transferability to downstream multi-class and localization tasks [8].

5.5 Clinical Implications

5.5.1 Triage and Workflow Enhancement

High-throughput fracture detection can:

- Accelerate ED workflows: Automated initial screening saves radiologists a significant amount of reading time.
- Minimize False diagnoses: High sensitivity 98.58 percent has a significant reducing impact on false negativity.
- Enable 24/7 coverage: Technologies on autopilot provide an uninterrupted performance regardless of the presence of radiologists.

5.5.2 Geographic Equity

Deployment in resource-constrained settings enables:

- Equitable access: Diagnosis is automated, thereby reducing the number of specialist radiologists.
- Cost reduction: Radiologists are going to work less, and operational savings.
- Rapid deployment: It can easily be deployed on the cloud or at the edge devices in the event that we have effective architectures.

5.5.3 Limitations for Clinical Adoption

Current binary classification lacks:

- Fracture characterization: Type (transverse, oblique, comminuted)
- Severity assessment: Displacement, angulation, alignment
- Anatomical localization: Precise fracture site
- Clinical recommendations: Treatment guidance

Such limitations require the involvement of clinicians in order to complete the diagnosis and treatment planning.

6. Conclusion and Future Work

6.1 Summary of Findings

This paper shows that Swin Transformer Base is useful when it comes to bone fracture binary classification in X-ray images.

Key findings include:

1. **State-of-the-art Performance**: The model has accuracy of 98.45% with sensitivity and specificity of 98.58 and 98.45 respectively which is much better than the work of an average experienced radiologist.
2. **Superior Architecture:** The fact is that Swin-B performs better than the standard CNN baselines such as ResNet50 and DenseNet121 on all measures.

3. **Clinical Potential:** This has been achieved through the high quality and efficiency of the diagnostics that allow the implementation of the proper and computationally efficient detection into the clinical departments of the real world.
4. **Transfer Learning Efficacy:** The ImageNet trained weights can be effectively used to overcome the drawbacks of small medical imaging datasets.
5. **Robust Generalization:** Minor gap in the small train-validation is functioned as effective regularization, and no significant overfitting is noticed.

6.2 Methodological Contributions

This research contributes:

- **Comprehensive preprocessing pipeline:** We employed standard image processing, augmentation and normalization to retain image results the same way.
- **Optimized transfer learning strategy:** The two-step fine-tuning was used to ensure that it did not forget any of the information that existed within the pre-trained knowledge.
- **Comprehensive evaluation framework:** A multi-metric evaluation was performed on the basis of clinically relevant criteria, giving us a general view of the model performance.
- **Computational efficiency analysis:** Shown to be linearly complex, the architecture has had wide-ranging benefits over conventional transformer architectures.

6.3 Future Research Directions

6.3.1 Dataset Expansion

Future work should incorporate:

- **Multi-institutional datasets:** The imaging information of different radiographic techniques and devices was included to expand the generalizability of the data.
- **Global population representation:** Assured applicability among diverse ethnic, demographic and geographic groups.
- **Increased sample size:** Made use of the power of Swin-B with 87 million parameters training it with over 100,000 images.

6.3.2 Multi-class Fracture Taxonomy

Multi-class classification should be extended to allow:

- **Fracture type identification:** Transverse, oblique, spiral, comminuted, pathologic
- **Anatomical location classification:** 20+ distinct bone regions
- **Severity stratification:** Undisplaced, minimally displaced, displaced, with angulation/rotation
- **Associated findings:** Diagnosis of soft-tissue trauma, joint injuries, and a loss of the vascularity.

6.3.3 Hybrid Detection-Localization Architecture

Architectural enhancements addressing current limitations:

- **Object detection integration:** YOLOv8 or Faster R-CNN for precise fracture localization
- **Heatmap generation:** Attention visualizations highlighting fracture regions for clinician guidance
- **Multi-task learning:** Simultaneous classification, localization, and severity assessment
- **3D volumetric analysis:** CT integration for complex fracture characterization

6.3.4 Explainability and Interpretability

Enhanced transparency mechanisms:

- **Attention visualization:** Attention maps by layers were constructed for the purposes of clinical explanation.
- **Gradient-based activation mapping (Grad-CAM):** Gradient-CAM (Grad) Highlight salient regions in images used to make predictions.
- **Feature importance quantification:** Specific distinct features from classification were identified.
- **Uncertainty quantification:** Bayesian techniques gave the confidence values to aid clinical decision making.

6.3.5 Clinical Validation and Deployment

Pathway to clinical implementation:

- **Prospective multi-center trial:** Validation across diverse institutions and patient populations
- **Radiologist comparison study:** Head-to-head comparison with human expert performance
- **FDA clearance pathway:** Regulatory submissions for medical device classification
- **Clinical workflow integration:** DICOM compatibility, EHR integration, real-time inference deployment
- **Ethical framework development:** The development of guidance that deals with accountability and liability in AI-assisted diagnosis.

6.3.6 Multi-modal Integration

Beyond X-ray imaging:

- **CT integration:** Volumetric analysis for complex fracture assessment
- **MRI fusion:** Soft tissue injury detection and ligamentous assessment
- **Ultrasound correlation:** Point-of-care fracture confirmation in resource-limited settings
- **Clinical metadata incorporation:** Patient age, mechanism of injury, pre-morbid conditions

6.3.7 Federated Learning and Privacy

Decentralized model development:

- **Federated learning implementation:** Allowed the training of models on the institutions without concentration of the sensitive patient data.
- **Differential privacy:** Quantified privacy guarantees preventing unauthorized inference
- **Multi-institutional collaboration:** Leveraging diverse datasets while maintaining patient confidentiality

6.4 Broader Impact and Recommendations

For Healthcare Systems:

- Adopt AI-based perioperative triage processes in order to increase the efficiency of the emergency department without sacrificing the skills of the radiologist.
- Apply biological prioritization of deployment in settings with limited resources and less supply of radiologists.
- Establish the general guidelines of AI integration in clinical decision support systems.

For Regulatory Bodies:

- Establish clear AI medical device approval pathways balancing innovation with safety
- Mandate multi-center validation requirements reflecting real-world deployment scenarios
- Develop accountability frameworks for adverse outcomes related to AI system failures

For Researchers:

- Focus on massive multi-institutional data gathering in order to overcome the resounding generalization issues.
- We ought to invest in research on interpretability, which would help clinicians to learn about the decisions of the model.
- Research sensitive medical data privacy training methods.

7. References

[1] Md. Sabbir Ahammed et al., "Bone Fracture Classification in X-ray Images: A Deep Learning Approach Leveraging Transfer Learning," in *Proceedings of the International Conference on Electrical and Computer Engineering*, Feb. 2025, doi: 10.1109/ECCE64574.2025.11013431.

[2] Basma Balam and Atef Eldenfria, "Automated Detection of Bone Fractures in X-ray Images Using Deep Learning and Ensemble Learning," *Journal of Trauma Research*, vol. 45, pp. 1–18, Dec. 2025, doi: 10.26629/jtr.2025.45.

[3] Pratham Kaushik and Aseem Aneja, "The Future of Orthopedic Care: High-Accuracy Bone Fracture Detection with CNNs," in *Proceedings of the 64th IEEE International Conference on Sensor Systems and Systems*, Oct. 2024, doi: 10.1109/ICSSAS64001.2024.10760767.

[4] Bhavik Kumar et al., "Revolutionizing Bone Injury Diagnosis with Advanced Deep Learning Approaches," in *Proceedings of the International Conference on Emerging Electronics and Computing Technologies*, Aug. 2024, doi: 10.1109/EECT61758.2024.10739215.

[5] Alana Mulya Zebada and Endang Wahyu Pamungkas, "Analysis Of A Deep Learning Algorithm For Fracture Detection In X-Ray Images," *International Journal of Advanced Multidisciplinary & Biosciences Research*, vol. 6, no. 3, pp. 1–18, Dec. 2025, doi: 10.59395/ijadis.v6i3.1451.

[6] Óscar A. Martín and Javier Sánchez, "Evaluation of Vision Transformers for Multimodal Image Classification: A Case Study on Brain, Lung, and Kidney Tumors," *arXiv preprint arXiv:2502.05517*, Feb. 2025, doi: 10.48550/arXiv.2502.05517.

[7] Lingeshwaran Sekar et al., "Bone Fracture Detection Based on Deep Learning Techniques," in *Proceedings of the Artificial Intelligence and Machine Learning Applications Conference*, Apr. 2025, doi: 10.1109/AIMLA63829.2025.11041507.

[8] S. S. Kumar et al., "Swin Transformer Architecture for Accurate Brain Tumor Classification and Localization in MRI-Based Medical Diagnosis," in *Proceedings of the International Conference on Clinical Radiology Research*, July 2025, doi: 10.1109/ICCR67387.2025.11291979.

[9] H. a. V. Dharshenee, N. R. A, N. Kumar, and K. F. K. Jiavana, "VIT-DETR: A Hybrid Vision Transformer and Detection Transformer for Hand Fracture Detection and Classification," 2025 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI), Apr. 2025, doi: 10.1109/RAEEUCCI63961.2025.11048199.

[10] Khushi Mittal et al., "Revolutionizing Fracture Diagnosis: A Deep Learning Approach for Bone Fracture Detection and Classification," in *Proceedings of the Technology Operations Conference*, June 2024, doi: 10.1109/OTCON60325.2024.10688357.

[11] Abdulmajeed Alsufyani, "Enhancing diagnostic accuracy in bone fracture detection: A comparative study of customized and pre-trained deep learning models on X-ray images," *International Journal of Advanced & Applied Sciences*, vol. 12, no. 5, pp. 1–15, June 2025, doi: 10.21833/ijaas.2025.05.008.

[12] R. Bhuria and S. Gupta, "X-Ray Insights: Comprehensive Dataset for Bone Fracture Detection Across Diverse Anatomical Regions," 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), Sep. 2024, doi: 10.1109/ICOSEC61587.2024.10722406.

[13] Ali Hatamizadeh et al., "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, Jan. 2022, doi: 10.1007/978-3-031-08999-2_22.

[14] et al., "Digital signal processing," *IEEE Transactions on Engineering Sciences*, vol. 11, no. 3, pp. 1–25, Oct. 2018, doi: 10.1002/ett.4460110311.

[15] A. Pravamanjari, S. Swain, and P. Mallick, "Advanced Detection and Classification of Pancreatic Cancer in CT Images Using Swin Transformer Architecture," Educational Sciences International Conference, Feb. 2025, doi: 10.1109/ESIC64052.2025.10962786.

[16] Vikas Hassija et al., "Transformers for Vision: A Survey on Innovative Methods for Computer Vision," *IEEE Access*, vol. 13, pp. 1–45, Jan. 2025, doi: 10.1109/access.2025.3571735.

[17] Yucheng Tang et al., "Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, doi: 10.1109/cvpr52688.2022.02007.

[18] Óscar A. Martín and Javier Sánchez, "Evaluation of Vision Transformers for Multi-Organ Tumor Classification Using MRI and CT Imaging," *Electronics*, vol. 14, no. 15, p. 2976, July 2025, doi: 10.3390/electronics14152976.

[19] Benzorgat Mustapha et al., "Enhanced Pneumonia Detection in Chest X-Rays Using Hybrid Convolutional and Vision Transformer Networks," *Advances in Clinical and Experimental Medicine*, vol. 34, no. 1, pp. 1–20, Jan. 2025, doi: 10.2174/0115734056326685250101113959.

[20] Zhanhao Zhang, "The transferability of transfer learning model based on ImageNet for medical image classification tasks," *Journal of Computing and Intelligence*, vol. 18, pp. 1–12, Oct. 2023, doi: 10.54254/2755-2721/18/20230980.

[21] Yuning Huang et al., "Comparative Analysis of ImageNet Pre-Trained Deep Learning Models and DINOv2 in Medical Imaging Classification," in *Proceedings of the IEEE Computer Society Annual Symposium on VLSI*, pp. 1–8, Feb. 2024, doi: 10.1109/COMPSAC61105.2024.00049.

[22] J. B et al., "Deep Learning-based Binary Classification of Bone Fractures using a Hybrid MobileNetV3-CNN Architecture and Clinical X-ray Dataset," in *Proceedings of the International Conference on Innovative Techniques and Applications in Computing and Information Systems*, Dec. 2025, doi: 10.1109/ICICNIS66685.2025.11315561.

[23] Deekshith Bolla and Wisam Bukaita, "Multi-Region Bone Fracture Detection in X-Ray Images using Deep Learning," *Medical Research Archives*, vol. 13, no. 12, pp. 1–20, Dec. 2025, doi: 10.18103/mra.v13i12.7099.

[24] Vidya et al., "Detection of Hand Bone Fractures in X-Ray Images Using Two-Stage Deep Learning Methodology," in *Proceedings of the International Conference on Image, Data and Computing Analytics*, Oct. 2025, doi: 10.1109/ICIDCA66325.2025.11280376.

[25] Pravallika Kondapalli et al., "Deep Learning-Driven Multiclass Bone Fracture Detection and Localization: A Comparative Study of CNN Architectures with YOLOv8-based Segmentation in Medical Imaging," in *Proceedings of the International Conference on Innovative Techniques in Neural Networks and Information Systems*, Dec. 2025, doi: 10.1109/ICICNIS66685.2025.11315614.

[26] Nidhi Mishra and Ghorpade Bipin Shivaji, "Advanced Bone Fracture Detection Using Hybrid Deep Learning Algorithms and Machine Learning Models for Enhanced Medical Imaging Diagnosis," in *Proceedings of the Automation and Computation Conference*, Mar. 2025, doi: 10.1109/AUTOCOM64127.2025.10956299.

[27] Kailasam Selvaraj et al., "A Hybrid Convolutional and Vision Transformer Model with Attention Mechanism for Enhanced Bone Fracture Detection in X-Ray Imaging," in *Proceedings of the International Conference on Microelectronics, Computing and Communications*, Jan. 2025, doi: 10.1109/ICMSCI62561.2025.10893953.

[28] S. Padmakala, "Deep Learning in Radiology: A Comparative Study of CNN Architectures for Automated Detection of Bone Fractures," in *Proceedings of the International Conference on Image Computing and Vision*, June 2025, doi: 10.1109/ICICV64824.2025.11085635.

[29] Gude Venkat et al., "Segmentation and Classification of Bone Fractures in X-Ray Images using Deep Learning," in *Proceedings of the International Conference on Information Systems and Security*, Mar. 2025, doi: 10.1109/ICISS63372.2025.11076491.

[30] Happy Kumar Sharma et al., "BoneCareMapper: A Framework for Predictive Bone Fracture Detection using Medical Image Analysis," in *Proceedings of the International Conference on Industrial and Communication Systems*, Mar. 2025, doi: 10.1109/ICICCS65191.2025.10984444.

[31] Varun V et al., "Efficient CNN-Based Bone Fracture Detection in X-Ray Radiographs with MobileNetV2," in *Proceedings of the International Conference on Robotics, Automation and Intelligent Systems*, Nov. 2024, doi: 10.1109/ICRAIS62903.2024.10811726.

[32] Khaled Alomar, Halil Ibrahim Aysel, and Xiaohao Cai, "Data Augmentation in Classification and Segmentation: A Survey and New Strategies," *Journal of Imaging*, vol. 9, no. 2, p. 46, Feb. 2023, doi: 10.3390/jimaging9020046.

[33] Mohammad Akbarpour Ganjeh and Azamossadat Nourbakhsh, "Transfer Learning in Bone Fracture Detection: A Comprehensive Review," in *Proceedings of the Iran Purse Pattern Recognition Conference*, Sept. 2025, doi: 10.1109/IPRIA68579.2025.11263711.

[34] Rashedur Rahman et al., "Enhancing fracture diagnosis in pelvic X-rays by deep convolutional neural network with synthesized images from 3D-CT," *Scientific Reports*, vol. 14, no. 9074, pp. 1–18, Apr. 2024, doi: 10.1038/s41598-024-58810-4.