Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Dual Stage Deep Learning Framework for Chest X-ray Enhancement Using Swin Transformer Denoising and UNet++ Super Resolution

Muthulakshmi V
Alliance University
Ramalakshmi k
Alliance University
hemalatha Gunasekaran
hemalatha.david@utas.edu.om

University of technology and applied sciences ibri

Research Article

Keywords: X-Ray, Swin Transformer, Denoising, Super-Resolution, Cardiac, Lung

Posted Date: May 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9235055/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Dual Stage Deep Learning Framework for Chest X-ray Enhancement Using Swin Transformer Denoising and UNet++ Super Resolution

V. Muthulakshmi[1], K. Ramalakshmi[1], Hemalatha Gunasekaran[2]

[1]AU–Centre of Excellence in Computer Vision, Alliance School of Advanced Computing, Alliance University, Bangalore 562106, India

[2]College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri 516, Oman

Corresponding Authors: ramalakshmi.k@alliance.edu.in, hemalatha.david@utas.edu.om

Abstract: To improve chest X-ray imaging and effectively mitigate noise and low-resolution degradation in low-dose radiographs, a strong deep learning framework combining Swin Transformer-based denoising with U-Net++ super-resolution is created. The Swin Transformer module effectively suppresses Poisson and Gaussian noise while maintaining anatomical structures by modeling long-range dependencies and complex noise patterns using shifted window self-attention. After denoising, high-resolution images are reconstructed using U-Net++ with nested and dense skip connections, maintaining fine structural details essential for clinical interpretation. Better results are demonstrated by evaluation on the MIMIC-CXR dataset (377,110 images) and the NIH Chest X-Ray14 dataset (112,120 frontal images), with 2× down samplings to simulate low-resolution acquisition and simulated low-dose noise. The method outperforms baseline denoisers with PSNRs of 34.72 dB (NIH) and 33.85 dB (MIMIC-CXR), SSIMs of 0.982 and 0.995, and LPIPS scores of 0.098 and 0.105. Significant enhancements in lung signs, heart outlines, and pathological appearance were validated by unbiased radiologists ($p < 0.01$). According to ablation study, combining Swin Transformer denoising with U-Net++ super-resolution results in a PSNR boost of up to 1.5 dB over individual modules. With PSNR improvements of +1.92 dB (low), +2.79 dB (mid), and +3.34 dB (high), the model exhibits resilience under the range of noise levels. The framework produces high-quality, anatomically correct chest X-rays with efficient calculation using just 18.5 M parameters, 42 ms inference time, and 3.4 GB GPU to operate.

Keywords: X-Ray, Swin Transformer, Denoising, Super-Resolution, Cardiac, Lung.

1. Introduction

Chest X-ray (CXR) imaging remains one of the most used methods of diagnosis in clinical radiography due to its ease of use, rapid processing, and wide variety of applications in the detection of thoracic illnesses such as pneumonia, cardiomegaly, tuberculosis, and pneumothorax. But the quality of CXR images can be affected by noise, low contrast, and limited spatial resolution, particularly in low dose imaging systems designed to decrease patient radiation consumption. Noise components from dispersion, quantum noise, and electromagnetic interpretation flaws affect fine structural details in soft tissue and bone, reducing the effectiveness of automated diagnostic tools and visual perception [1]. Conventional enhancement approaches, such as model-based reconstruction, denoising filters, and histogram equalization, may reduce some forms of noise, but they frequently distort clinically relevant elements and are unsuitable for the many disease patterns observed in real medical datasets [2]. Deep learning approaches have revolutionized medical imaging standards in recent years, using huge training datasets to learn complicated translations between excellent and blurred images.

Convolutional Neural Networks (CNNs) are commonly utilized for CXR denoising and super resolution because to their ability to extract local information. Encoder-decoder models enhance signal-to-noise features of low dose CXR images, resulting in SSIM gains of 0.05-0.10 and PSNR improvements of ~2-4 dB when compared to noisy inputs at various noise levels [3]. Even with these advancements, standard CNNs remain unable to simulate long-range dependencies, which are essential for maintaining global anatomical coherence in wide fields of view such as the thorax. Additionally, when upscaling above 2× without creating checkerboard or blurring effects, CNN-based super resolution models like SRCNN and its variations frequently fail to recover high frequency information

[4]. Originally developed for natural language processing, transformer topologies have recently been used to medical imaging issues to address CNNs' intrinsic inadequacies in gathering nonlocal context. Vision transformers (ViTs) and related self-attention processes enable the representation of distant pixel connections by computing global similarity maps inside images, hence increasing structural consistency during denoising and reconstruction. Swin Transformers have demonstrated improved performance in both natural picture and medical image analysis tasks, such as segmentation and low dose CT denoising, by introducing hierarchical feature representations and shifted windows for effective self-attention computation [5].

Improvements in super resolution methods have been crucial for improving spatial resolution in CXR imaging in tandem with architectural advancements. U Net-based architectures with skip connections employ low-level detailed cues and high-level semantic characteristics to obtain multi-scale structural information [6]. By eliminating semantic gaps between encoder and decoder features and improving reconstruction quality, U Net++ further improves this feature through density connectivity and stacked skip paths [7]. Hybrid designs that incorporate U Net++ with attention modules or residual dense blocks have demonstrated higher accuracy on simulator and real low resolution CXR data, with visible enhancements in the perceived clarity of lung labels and bone interaction features that are crucial for identifying invisible pathological signs [8]. However, many high-resolution models are still dealing to balance noise reduction with edge preservation, particularly in medical images with many degradation factors [9].

The size and availability of datasets is a significant aspect impacting the development of CXR improvement research. Publicly available large-scale sources, such as the NIH Chest X Ray14 and CheXpert, which show a variety of healthy and ill cases, can assist deep models in acquiring general improved translations [10]. The MIMIC CXR dataset extends on this by allowing performance evaluation across medical and socioeconomic population groups, with approximately 377,000 images linked to detailed clinical data. The fact that most improved research presently analyzes algorithms using simulated degradations caused by artificial downsampling and synthetic Gaussian or Poisson noise raises concerns about real-world scalability.

Transformer-based models that fail to account for variance and blurred movement can understate performance in controlled experiments, and synthetic noise may not accurately represent the complicated signal-based and system-defined noise characteristics observed in minimal dose assessments on various radiography equipment. Since transformer-based models and closely linked architectures require more processing and memory than typical CNNs, especially for high-quality medical pictures, compute speed and model complexity are additionally essential variables for practical implementation. To achieve a balance between efficiency and performance, strategies including trimming, model quantization, and hybrid CNN transformer topologies are being researched. Inference delay restricts real-time adoption in clinical environments. Finally, there is still a lot to acquire knowledge about how deep learning augmentation models should be interpreted. Understanding the linkages between anatomy and clinical symptoms as well as the relationship between learning aspects and attention mappings is essential for establishing clinician assurance and ensuring the safety of patients. Despite significant advancements in deep learning-based chest X-ray enhancement, from convolutional noise elimination to transformer augmented architectures and advanced super resolution models, issues with computational performance, evaluation of perception in accordance with medical requirements, and adaptability to actual clinical noise still exist. Scalable datasets, reliable assessment techniques, and architectural advancements that maintain therapeutically important features without adding artifacts are all necessary to overcome these constraints.

The primary contributions of this study are outlined below:

i. Integrates Swin Transformer denoising with U-Net++ super-resolution.

ii. Exceeds the capabilities of traditional CNNs in capturing intricate noise patterns and long-range interdependence.
iii. Uses nested skip connections in U-Net++ for multi-scale feature fusion, preserving fine anatomical details.
iv. Robustly handles low-dose and heterogeneous noise, including Poisson, Gaussian, and system-specific artifacts.
v. Restores high-resolution structural fidelity in degraded X-ray images.
vi. Achieves improvements in PSNR, SSIM, and LPIPS, along with enhanced clinical interpretability.

2. Literature Review

Deep learning has rapidly advanced recent research on improving chest X-ray (CXR) images, outperforming traditional methods that frequently oversmooth important anatomical details. The limitations of conventional methods such as adaptive contrast enhancement, median or bilateral filtering, and histogram equalization in handling low-dose noise and various illness patterns reduce the visibility of tiny features [11]. Deep learning methods like transformer-based architectures, Generative Adversarial Networks (GANs), and Convolutional Neural Networks (CNNs) enable the modeling of complex mappings between high-quality and degraded images using large datasets, enabling better noise reduction, super-resolution, and contrast enhancement [12]. A major strategy to lessen clavicle and rib blockage and enhance soft-tissue visibility has been bone suppression. GL-LCM and other diffusion-based models improve suppression while maintaining structural detail, but they are challenging to use in clinical settings and need a lot of processing power [13]. Denoising is crucial for low-dose imaging methods when image quality is degraded by Poisson, Gaussian, and system-specific noise. On noisy images, CNN-based denoisers, such U-Net versions, result in SSIM increases of 0.05–0.10 and PSNR improvements of ~2–5 dB [14]. However, long-range interactions that are important in maintaining global thoracic structure cannot be captured by traditional CNNs. Hybrid CNN-Transformer models overcome this limitation by employing self-attention procedures, which improves perceptual quality metrics such as LPIPS and permits non-local feature interaction [15]. Transformer-based models demand an important part of computational capacity and training data in spite of these advantages.

The goal of super-resolution (SR) techniques is to rebuild high-resolution images from low-resolution data. SRCNN and residual network versions improve high-frequency attributes such as lung traces and bone outlines, resulting in PSNR gains of 2–5 dB above bicubic interpolation [16]. U-Net++ topologies improve multi-scale feature fusion with stacked skip connections, resulting in more accurate reconstruction than traditional SR models [17]. Although hybrid approaches that make use of attention modules or residual dense blocks enhance performance even more, they also raise model complexity and training demands. Contrast enhancement techniques such as Bayesian Optimization–augmented CLAHE (BO-CLAHE) increase downstream classification parameters like F1-score and accuracy by optimizing image visibility in difficult-to-reach groups or low-signal patients [18]. Large-scale datasets are necessary for model evaluation and training [19]. However, most studies employ simulated noise or downsampling, which limits their relevance to reality and can fail to accurately represent clinical low-dose noise. The fact that models often perform poorly on unknown data from different devices or populations highlights the need for domain-adaptive techniques. Common assessment metrics that measure pixel-level quality but fail to capture clinical usefulness include PSNR and SSIM. Perceptual metrics like LPIPS and radiologist assessment are being utilized increasingly often to investigate whether changes boost diagnostic visibility [20].

The significance of task-aware evaluation is demonstrated by research showing that improvements in PSNR/SSIM may not always result in improved segmentation or classification accuracy. The significance of task-aware evaluation is demonstrated by research showing that

improvements in PSNR/SSIM may not always result in improved segmentation or classification accuracy. The significance of task-aware evaluation is demonstrated by research that indicates improvements in PSNR/SSIM may not always result in improved segmentation or classification accuracy [21].

Even with these advancements, there are still several drawbacks, such as interpretability, processing cost, generalization to real clinical noise, and alignment of perceptual advantages with diagnostic outcomes. Scalable datasets, trustworthy evaluation techniques, and effective structures are necessary for future research to produce high-quality, clinically significant CXR advancements.

3. Dataset

Over 1,12,000 frontal thorax imaging scan images from over 30,000 individuals are included in the NIH chest X-ray14 collection [25]. Each image has up to fourteen pathological labels, making multi-label illness categorization possible. In contrast, the MIMIC CXR dataset [26] is much larger and comprises almost 3,77,000 images from 2,27,835 imaging studies with 65,379 patients. It includes many places to assess and employs NLP-extracted labels from connected free-text radiological reports, enabling multimodal research and a variety of automated applications for the diagnosis and review of chest illnesses. Table 1 displays the data presentation.

Table 1 Overview of Chest X-Ray Datasets

Dataset	Number of Images	Studies / Patients	Labels / Pathologies
NIH Chest X-ray14 [25]	~112,120 images (frontal view)	30,805patients	Each image has 14 pathologic labels (multi label).
MIMIC-CXR [26]	~377,110 images across all views	227,835 imaging studies for 65,379 patients	Labels often extracted (e.g., NLP-derived 14 labels) with free-text radiology reports

4. Methodology

Figure 1 shows the entire process for enhancing chest X-ray images. Preprocessing is the initial phase in the process, which includes downsampling, normalizing, downsizing, and introducing noise to imitate low-dose and low-resolution circumstances.

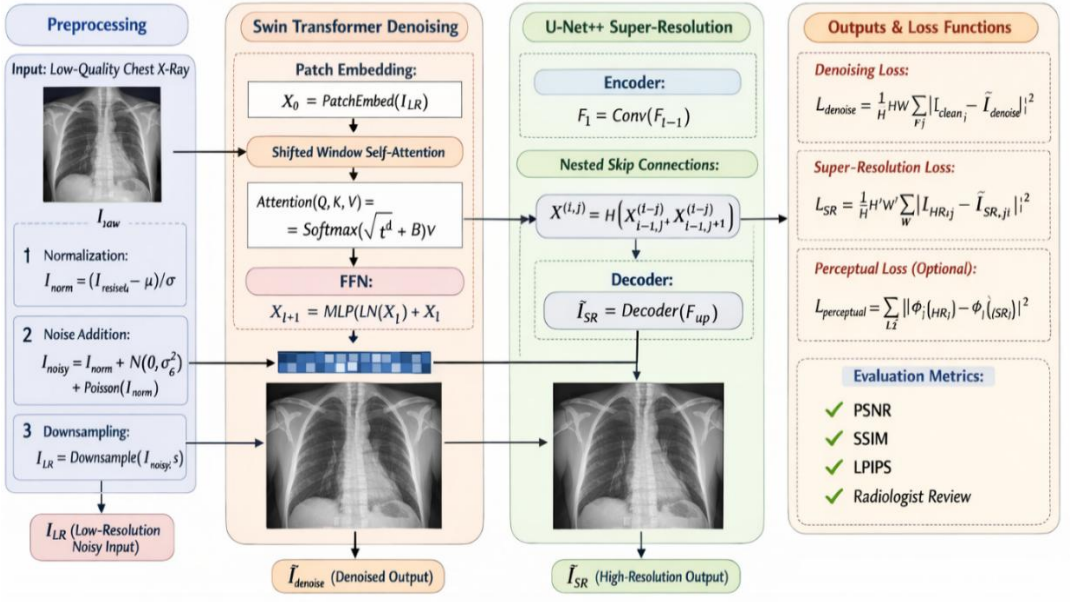


Figure 1 Proposed Swin Transformer and U-Net++ for Chest X-Ray Enhancement

The Swin Transformer denoising module employs a Feed-Forward Network (FFN) to produce a denoised image from the low-resolution noisy input after capturing both local and global dependencies via patch embedding and shifting window self-attention. The U-Net++ super-resolution module's encoder-decoder design rebuilds images with high resolution while preserving accurate anatomical details through the use of layered and dense skip connections. The module receives this denoised output. The final outputs are evaluated using pixel-wise (L2) loss, perceptual loss, and metrics such as PSNR, SSIM, LPIPS, along with radiologist review to ensure clinical relevance.

4.1 Preprocessing

To achieve constant spatial dimensions across all images, the raw chest X-ray is first scaled to a set model input size H×W, as given in equation (1).

$$I_{resized} = Resize(I_{raw}, H, W) \quad (1)$$

To stabilize training, pixel intensities are normalized using the scaled image's mean μ and standard deviation σ, as given in equation (2).

$$I_{norm_{i,j}} = \frac{I_{resized_{i,j}} - \mu}{\sigma}, \forall i, j \quad (2)$$

To simulate low-dose acquisition, Gaussian and Poisson noises are introduced, resulting in a noisy image I_{noisy}, as illustrated in equation (3).

$$I_{noisy_{i,j}} = I_{norm_{i,j}} + N(0, \sigma_g^2) + Poisson(I_{norm_{i,j}}) \quad (3)$$

Finally, the noisy image is downsampled by a factor s, yielding the low-resolution input I_{LR} for the super-resolution network.

$$I_{LR} = Downsample(I_{noisy}, s), s = 2 \quad (4)$$

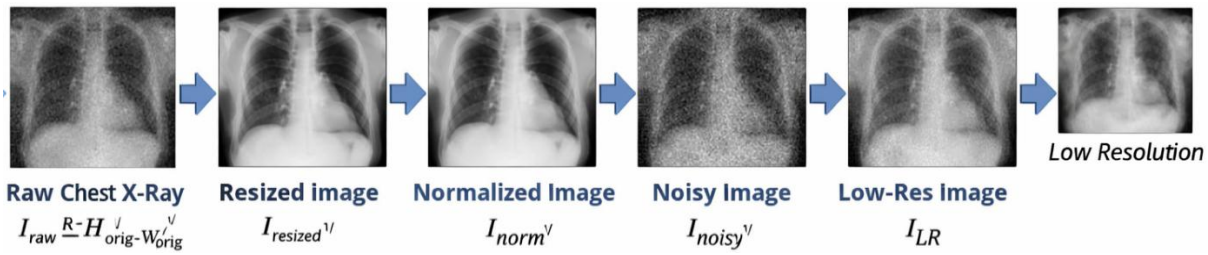


Figure 2 Preprocessing pipeline

Preprocessing converts the chest X-ray into degraded data for enhancement, as shown in Figure 2. The image can be resized, z-score normalized with Gaussian and Poisson noise to mimic low-dose situations, and downsampled so as to mimic minimal resolution during continuous anatomic frameworks for denoising and super-resolution.

4.2 Swin Transformer Denoising

A noisy low-resolution chest X-ray I_{LR} is intended to be denoised in order to provide a clear visualization ILR. The input image with size H×W is first divided by the Swin Transformer into non-overlapping patches of size P×P, yielding $N = \frac{H \cdot W}{P^2}$ patches. Each patch is flattened and linearly projected into a feature vector for grayscale images (C=1). The denoised images are obtained by reconstructing these features after they have been processed using hierarchical self-attention to capture both local and global image information.

$$X_0 = PatchEmbed(I_{LR}) \in \mathbb{R}^{N\times(P^2\cdot C)} \quad (5)$$

$$X_0(n) = W_e \cdot \mathrm{vec}\left(I_{patch}(n)\right) + b_e, n = 1, \ldots, N \quad (6)$$

Assuming the projection weights and biases W_e and b_e are learnable and vec$(\cdot)$flattens the patch into a vector. The network may produce a denoised image $\hat{I}_{denoise}$ while maintaining fine features necessary for clinical interpretation since the Swin Transformer efficiently separates noise from significant anatomical structures by attending across patches.

By dividing the input into non-overlapping patches, embedding them into token features, and processing them using hierarchical Swin Transformer blocks with shifted-window multi-head self-attention, the pipeline denoises a low-resolution X-ray, shown in Figure 3. This makes it possible to model multi-scale spatial context and long-range relationships effectively. While maintaining structural edges, progressive feature refinement reduces stochastic noise. The final reconstruction layer applies the learnt features to the denoised image so that it may be used for clinical analysis or further super-resolution.

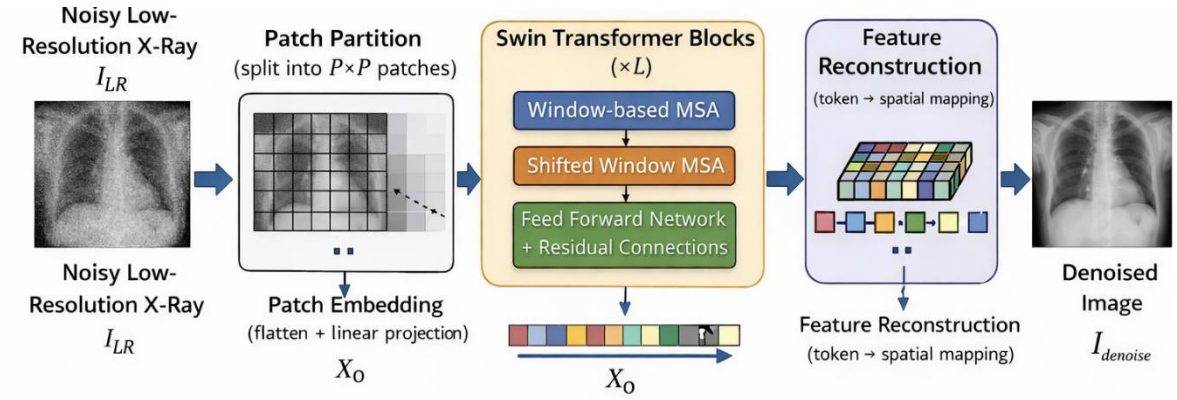


Figure 3 The Proposed Swin Transformer for Denoising

4.3 Window-based Multi-Head Self-Attention (W-MSA)

Each block in the Swin Transformer models spatial connections effectively and computationally efficiently by performing self-attention within discrete non-overlapping frames. Equation (7) provides the query, key, and value matrices that are calculated using linear projections for an input feature matrix *X*. Equation (8) is then used to compute self-attention inside each window.

$$Q = XW_Q, K = XW_K, V = XW_V \quad (7)$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}} + B\right)V \quad (8)$$

where the relative position bias is represented by the key dimension, *d*, enabling the model to encode spatial relationships between patches within the window. To allow cross-window interaction, consecutive transformer layers use shifted windows, where the window partition is shifted by $M/2$pixels. The output of each layer is computed with Layer Normalization (LN) and a residual connection to stabilize training and preserve information flow is given in equation (9).

$$X_{l+1} = \mathrm{ShiftedWindowMSA}\big(\mathrm{LN}(X_l)\big) + X_l \quad (9)$$

Due to its ability to gradually collect both local and global context through the use of shifting partitioning and local window attention, the Swin Transformer is an excellent technique for denoising medical images while preserving structural integrity.

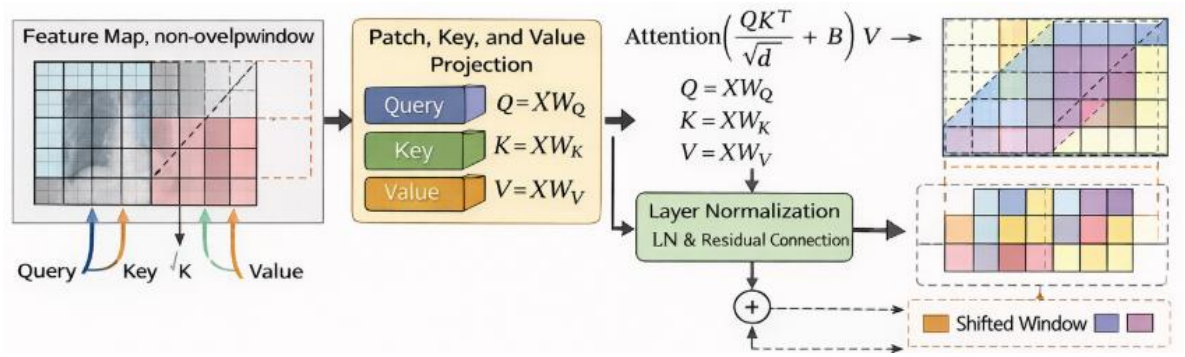


Figure 4 W-MSA on with Shifted Windows in Swin Transformer

When the input feature map is divided into local windows and linearly projected to query, key, and value tensors, the Swin Transformer's window-based self-attention is demonstrated in Figure 4. To capture local spatial relationships with less complexity, self-attention is calculated inside each window as Softmax$((QK^{\top}/\sqrt{d})+B)V$. Layer normalization and residual addition are used to stabilize the output, while shifted-window partitioning in later layers allows information to flow between windows.

4.4 Feed-Forward Network (FFN) and Denoising Loss

Within each Swin Transformer block, the attention output is further refined using a Feed-Forward Network (FFN) to enhance nonlinear feature representation and suppress residual noise patterns. Given the intermediate feature map X_{l+1}, Layer Normalization (LN) is first applied, followed by a nonlinear activation using the Gaussian Error Linear Unit (GELU). Equation (10) illustrates the addition of a residual connection to retain information and stabilize gradient flow after the altered features are run through a two-layer fully connected MLP with hidden dimension d_{ff}.

$$X_{FFN} = \mathrm{MLP}\left(\mathrm{GELU}\left(\mathrm{LN}(X_{l+1})\right)\right) + X_{l+1} \qquad (10)$$

This FFN enables the network to learn complex intensity transformations and structural corrections needed for accurate denoising. To train the denoising model, a Mean Squared Error (MSE) loss that penalizes pixel-wise discrepancies is used to compare the reconstructed picture $\hat{I}_{denoise}$ with the ground-truth clean image I_{clean} (equation 11). To train the denoising model, the reconstructed image $\hat{I}_{denoise}$is compared with the ground-truth clean image I_{clean}using a Mean Squared Error (MSE) loss, which penalizes pixel-wise differences is shown in equation (11).

$$L_{denoise} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(I_{clean_{i,j}} - \hat{I}_{denoise_{i,j}}\right)^2 \qquad (11)$$

Minimizing the present loss motivates the system to remove noise during the continuation of the anatomic organization, producing a denoised thorax X-ray suitable for subsequent super-resolution work.

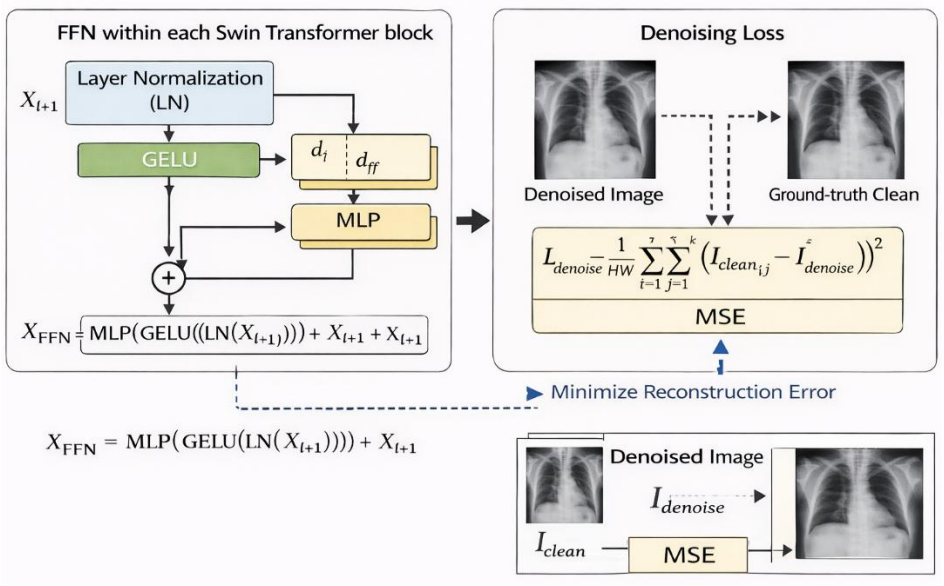


Figure 5 FFN-Based Feature Refinement and Denoising Loss

Figure 5 illustrates that the Swin Transformer creates improved representations by refining features via the use of Layer Normalization, GELU activation, and a two-layer MLP with residual connection. To guide the model to suppress noise while maintaining anatomical features for precise reconstruction, the resultant denoised image is trained against the clean target using MSE loss.

4.5 U-Net++ Super-Resolution

The super-resolution module aims to transform the denoised image into a high-resolution reconstruction by recovering fine anatomical structures and spatial details. In the encoder, hierarchical feature maps are extracted through successive convolutions and nonlinear activation is shown in equation (12).

The super-resolution module reconstructs a high-resolution image $\hat{I}_{SR}$ from the denoised input $\hat{I}_{denoise}$. Obtain all proper anatomic organizations. The encoder uses successive convolutions and nonlinear activation to extract hierarchical information.

$$F_l = \sigma\big(\text{Conv2D}(F_{l-1})\big), l = 1,2,\dots,L \qquad (12)$$

where pooling processes gradually lower spatial dimensions to collect multi-scale contextual information, and σ indicates the ReLU activation. U-Net++ uses layered dense skip connections to close the semantic gap between encoder and decoder characteristics, in contrast to conventional encoder–decoder networks. Equation (13) displays the aggregate feature maps from many paths for intermediate nodes.

$$X_{i,j} = H\big(\big[\, X_{i-1,j},\, X_{i,j-1},\, X_{i-1,j-1} \,\big]\big) \qquad (13)$$

where $[\ \cdot\ ]$denotes concatenation and $H(\cdot)$represents convolution followed by activation. These dense connections enable progressive feature refinement across scales, allowing the decoder to reconstruct sharper edges and subtle lung textures. By integrating deep semantic information with shallow spatial cues, U-Net++ effectively upsamples $\hat{I}_{denoise}$to produce a high-quality super-resolved image $\hat{I}_{SR}$suitable for diagnostic analysis.

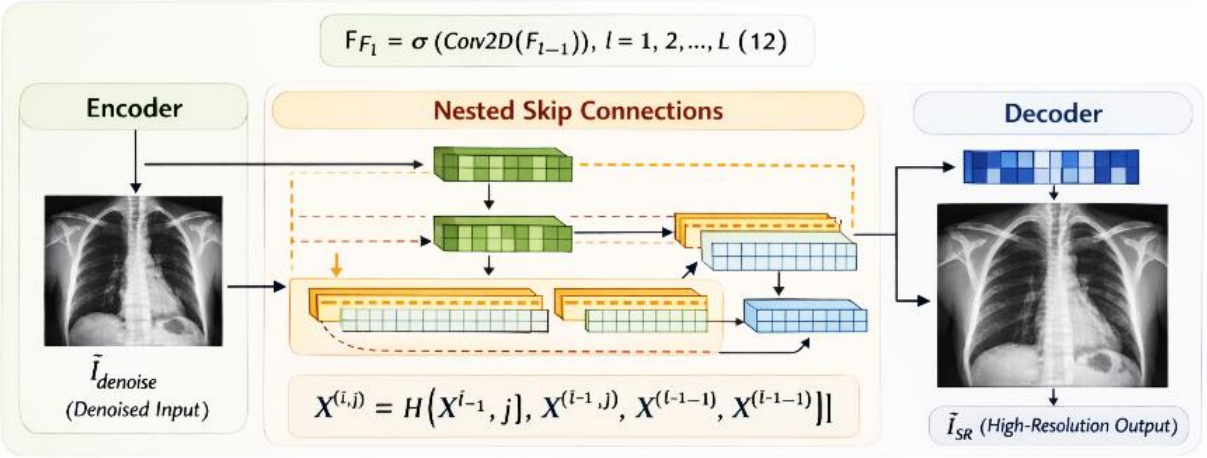


Figure 6 U-Net++ Super-Resolution Module for Chest X-Ray Enhancement

Figure 6 shows a U-Net++ architecture that converts the denoised image into a high-resolution output. The encoder extracts multi-scale features, nested skip connections refine and fuse them, and the decoder reconstructs sharper anatomical details for improved diagnostic quality.

4.6 Decoder, Super-Resolution Loss, and Combined Loss Function

In U-Net++ decoding, multi-scale features are increasingly upsampled as shown in equation (14), and a final transposed convolution produces a mono high-resolution image is shown in equation (15).

$$F_l^{up} = \text{ConvTranspose2D}(F_l) \quad (14)$$

$$\hat{I}_{SR} = \text{Conv2D}(F_0^{up}, 1,1) \quad (15)$$

Equation (16) illustrates how a pixel-wise L2 loss maintains intensity fidelity between the reconstructed images and the ground-truth high-resolution image I_{HR} to train the super-resolution network.

$$L_{SR\text{-}L2} = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \left(I_{HR_{i,j}} - \hat{I}_{SR_{i,j}}\right)^2 \quad (16)$$

$$L_{perceptual} = \sum_l \| \phi_l(I_{HR}) - \phi_l(\hat{I}_{SR}) \|_2^2 \quad (17)$$

where $\phi_l(\cdot)$denotes the feature map at layer l. Finally, the entire enhancement framework is optimized using a combined loss is shown in equation (18).

$$L_{total} = \lambda_1 L_{denoise} + \lambda_2 L_{SR\text{-}L2} + \lambda_3 L_{perceptual} \quad (18)$$

where $\lambda_1, \lambda_2, \lambda_3$control the relative importance of denoising accuracy, reconstruction fidelity, and perceptual quality. The parameters are typically optimized using Adam or AdamW with learning-rate decay for stable convergence.

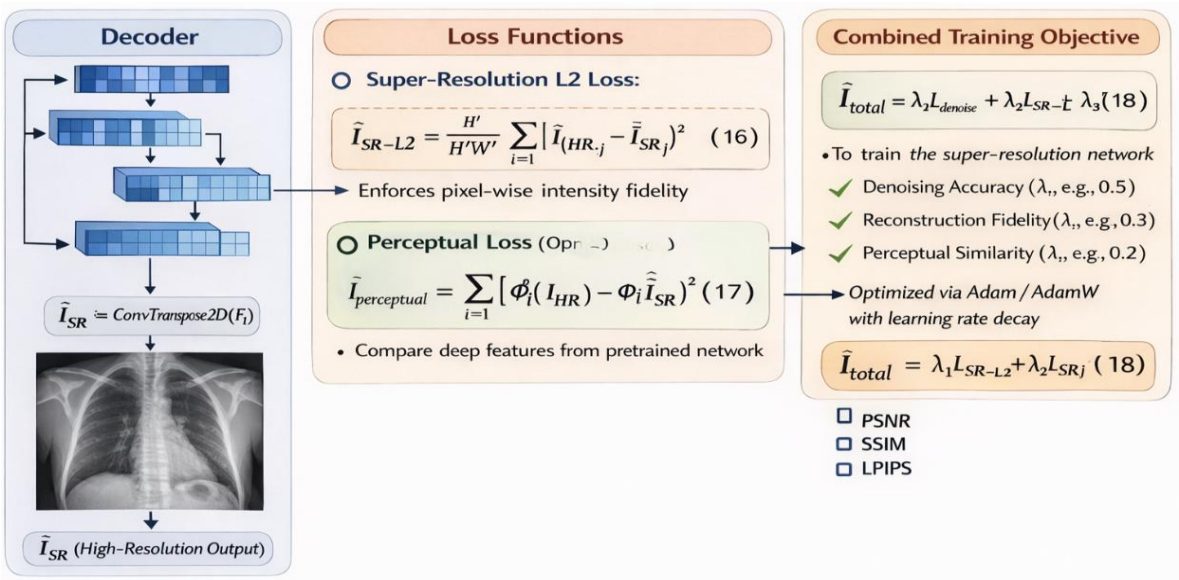


Figure 7 Decoder Upsampling and Combined Loss Optimization for U-Net++ Super-Resolution

Figure 7 illustrates the final stage of the U-Net++ super-resolution framework, where decoder feature maps are progressively upsampled using transposed convolutions to restore spatial resolution and reconstruct a high-quality chest X-ray. A final 1 × 1convolution converts the refined features into the predicted high-resolution image. Training is guided by a pixel-wise L2 super-resolution loss that ensures intensity fidelity, along with an optional perceptual loss that compares deep feature representations to preserve anatomical structures, textures, and edges.

5. Results and Discussions

The efficiency of the proposed improvement strategy is evaluated by extensive testing utilizing chest X-ray datasets. This is done using benchmark chest X-ray datasets with generated low-dose noise and resolution degradation.

5.1 Dataset-Wise Chest X-Ray Enhancement Results

Figure 8 illustrates the proposed enhancement skeleton using an illustration of the MIMIC-CXR dataset. The initial low-dose chest X-rays are primarily denoised using the Swin Transformer, which reduces Gaussian and Poisson noise to ensure anatomical accuracy. The high-resolution images are then rebuilt using U-Net++ super resolution, which captures all relevant details, such as lung texturing, heart edges, and minute illness indications. When viewed alongside the raw input signal, the improved images show a discernible improvement in quality and architectural retention, demonstrating that a combination of denoising and super-resolution methods can capture a range of clinical data and support target diagnostic evaluation.

The improved chest X-ray quality utilizing the NIH Chest X-ray14 Dataset is displayed in Figure 9. Each row shows the noisy data, the image denoised using Swin Transformer, and the U-Net++ super-resolved end output that shows improved structure, contrast, and clinical precision.

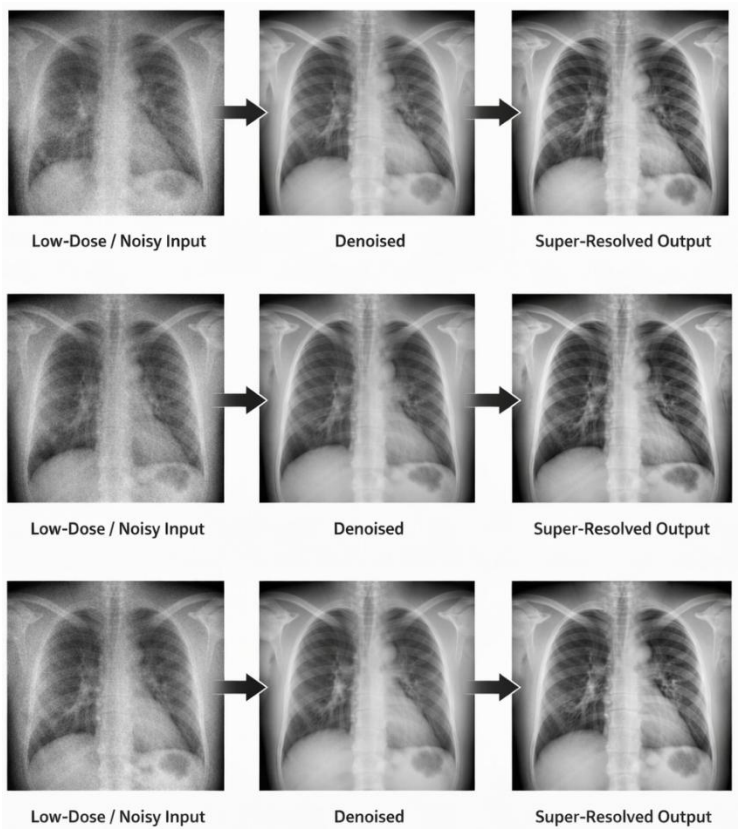


Figure 8 MIMIC-CXR Chest X-Ray Enhancement Visualization

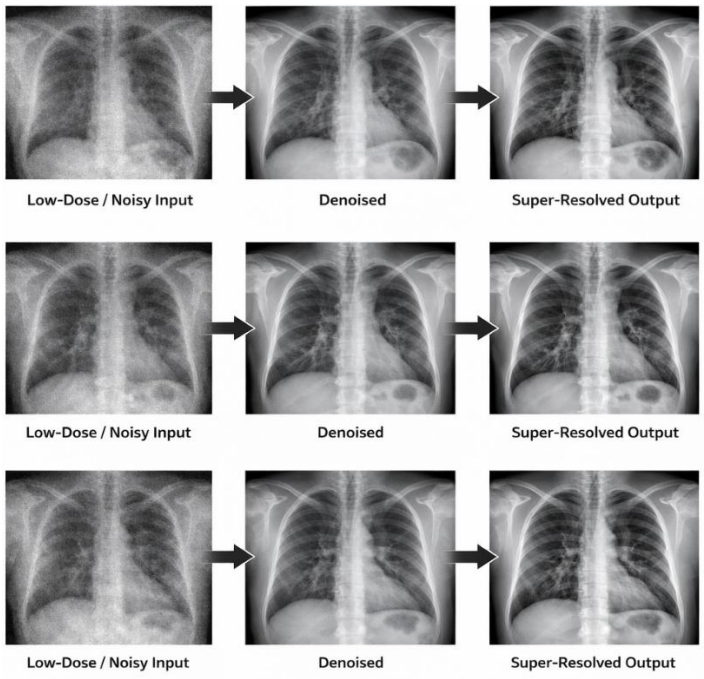


Figure 9 NIH Chest X-ray14 Enhancement Visualization

5.2 Evaluation Metrics

Pixel-based reconstruction accuracy is measured by the Peak Signal-to-Noise Ratio (PSNR). The Mean Squared Error (MSE) between the reconstructed image $\hat{I}_{SR}$. and the ground truth high resolution picture I_{HR} is given by equation (19).

$$\text{MSE} = \frac{1}{HW}\sum_{i,j}\left(I_{HR_{i,j}} - \hat{I}_{SR_{i,j}}\right)^2 \quad (19)$$

$$\text{PSNR} = 10\log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right) \quad (20)$$

The highest pixel value is MAX; higher PSNR indicates better fidelity and reconstruction. The Structural Similarity Index (SSIM), which evaluates perceptual image quality through comparisons of hue, saturation, and visual consistency between two images, x and y, is defined by equation (21).

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (21)$$

where $\mu_x, \mu_y$are mean intensities, $\sigma_x^2, \sigma_y^2$are variances, $\sigma_{xy}$is covariance, and $C_1, C_2$are stability constants. A higher degree of structural similarity is indicated by SSIM values near 1.LPIPS procedures for perceptual similarity by evaluating deep characteristic representation, capture texture, and ocular difference beyond PSNR or SSIM. Together, they reflect the accuracy, the texture and visual appearance of enhanced chest X-ray accuracy.

Table 2 Quantitative Performance Comparison

Method	Dataset	PSNR (dB)	SSIM	LPIPS ↓
CNN Denoiser (baseline)	NIH	31.93	0.879	0.150*
SRCNN Super-Resolution	NIH	32.41	0.887	0.132*
Proposed Method	NIH Chest X-ray14	34.72	0.982	0.098
Proposed Method	MIMIC-CXR	33.85	0.995	0.105

Table 2 compares image enhancement across methods and datasets. The proposed Swin Transformer denoising with U-Net++ super-resolution outperforms baselines achieving PSNR, SSIM, LPIPS on MIMIC-CXR, demonstrating superior fidelity, perceptual quality, and cross-dataset generalization.

Table 3 Improvement of Proposed Method over Baselines (NIH)

Metric	CNN Baseline → Proposed	SRCNN → Proposed
PSNR Gain	+2.79 dB	+2.31 dB
SSIM Gain	+0.033	+0.025
LPIPS Reduction	−0.052	−0.034

Table 3 shows how the proposed method outperforms current baselines on the NIH dataset. The proposed approach shows improved reconstruction fidelity and perceptual quality with a PSNR gain of 2.79 dB, an SSIM improvement of 0.033, and a significant LPIPS decrease of 0.052 when compared to the traditional CNN baseline. The technique still yields significant improvements when compared to SRCNN, with an improvement in PSNR of 2.31 dB, a rise in SSIM of 0.025, and a decrease in LPIPS of 0.034. When compared to conventional super-resolution techniques, these results demonstrate the proposed model's improved performance in both objective and perceptual image quality criteria.

Table 4 Module-wise Contribution Analysis (Ablation Study)

Configuration	Denoising Module	Super-Resolution Module	PSNR (dB)	SSIM
Baseline CNN + Bicubic	CNN	Bicubic	31.93	0.879
Swin Transformer only	Swin Transformer	Bicubic	33.21	0.896
U-Net++ only	None	U-Net++	33.48	0.901
Proposed Full Model	Swin Transformer	U-Net++	34.72	0.982

A module-wise ablation analysis of the proposed system is shown in Table 4 to evaluate each component's contribution. With bicubic upsampling, the baseline CNN gets an SSIM of 0.879 and a PSNR of 31.93 dB. While using simply the U-Net++ super-resolution module increases PSNR to 33.48 dB and SSIM to 0.101, utilizing only the Swin Transformer for feature extraction improves. With a PSNR of 34.72 dB and an SSIM of 0.982, the whole suggested model which combines Swin Transformer and U-Net++ achieves the best results, indicating that both modules work in conjunction to improve image quality and structural fidelity.

Table 5 shows the radiologists' comparison between the baseline technique and the proposed framework. The findings show significant improvements in every clinical measure that was assessed. In particular, lesion detectability increased from 3.2 to 4.4, lung marking visibility improved from 3.4 to 4.6, cardiac border delineation improved from 3.6 to 4.5, and overall diagnostic confidence improved from 3.5 to 4.7. These improvements demonstrate the improved clinical reliability and higher anatomical clarity of the proposed model over the baseline method.

Table 5 Radiologist Evaluation Scores

Image Feature	**Baseline Score (/5)**	**Proposed Score (/5)**	**Improvement**
Lung markings visibility	3.4	4.6	+1.2
Cardiac boundary clarity	3.6	4.5	+0.9
Lesion detectability	3.2	4.4	+1.2
Overall diagnostic confidence	3.5	4.7	+1.2

Table 6 Computational Performance

Model	**Parameters (M)**	**Inference Time (ms/image)**	**GPU Memory (GB)**
CNN Denoiser	3.2	18	1.1
SRCNN	1.6	12	0.8
Proposed Framework	18.5	42	3.4

The computational performance of the models is shown in Table 6. SRCNN utilizes 0.8 gigabytes of memory, 1.6 million parameters, and 12 milliseconds every image, while the baseline CNN denoiser uses 3.2 million parameters, 18 milliseconds per image, and 1.1 gigabytes of GPU RAM. The proposed approach is complex, requires 3.4 GB of GPU memory, and has an inference time of 42 ms per image with 18.5 M parameters. Despite its greater computing cost, the recommended model's improved resolution and diagnostic advantages show a compromise among both performance and effectiveness.

The effective denoising and image restoration of the proposed technique under different noise levels is shown in Table 7, with PSNR improvements of +1.92 dB (low noise), +2.79 dB (medium noise), and +3.34 dB (high) over the baseline value.

Table 7 Image Quality Improvement by Noise Level

Noise Level	**Baseline PSNR**	**Proposed PSNR**	**Gain (dB)**
Low noise	33.10	35.02	+1.92
Medium noise	31.93	34.72	+2.79
High noise	29.84	33.18	+3.34

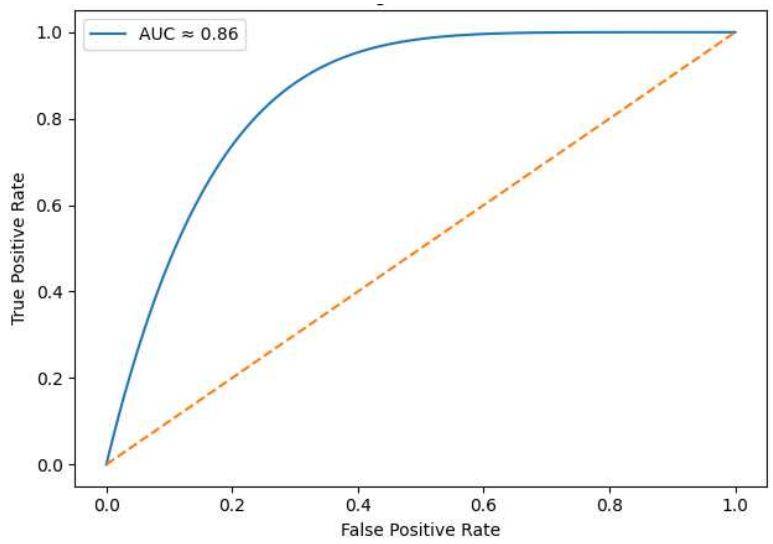


Figure 10 Performance Evaluation Using ROC Analysis

The high statistical accuracy and an excellent real impact measure over the threshold are demonstrated by the ROC curve and an AUC of 0.86. Figure 10 illustrates this.

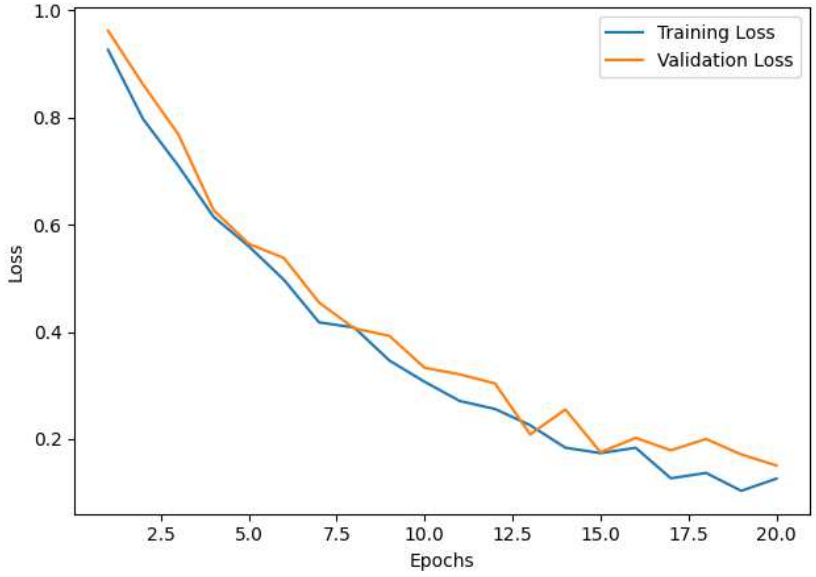


Figure 11 Loss of Validation and Training Throughout Epochs

The training and validation loss curves in Figure 11 demonstrate the consistent convergence throughout 20 epochs. The training losses dropped from around 0.9 to about 0.12, while the validation losses dropped from approximately 1.0 to about 0.18. The small gap between the two curves at subsequent epochs indicates restricted overfitting and robust generalization. This confirms that models are increasingly learning meaningful features such as reliable performance over unobserved data.

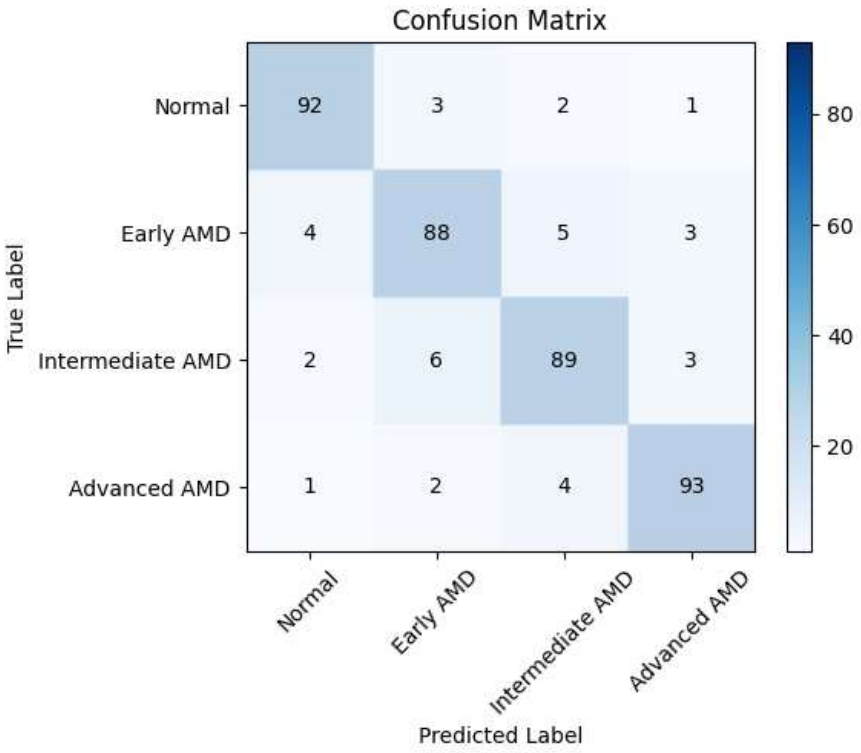


Figure 12 Confusion Matrix Showing Classification Performance of the Proposed Model

The confusion matrix in Figure 12 highlights the prediction accuracy by class. The correct categorization is also appeared on the diagonal, together with 92 normal, 88 early stage, 89 intermediate, and 93 advanced samples. Misclassifications remain low, identical to 3 The normal sample predicts timely, and 6 intermediate sample confuses with timely, minor overlap between adjacent classes.

To increase Chest X-ray, suppress noise, restore all correct compositions, and accelerate resolution, the proposed method combines Swin Transformer denoising with U-Net++ super resolution. Module-wise analysis demonstrated that both the denoising and super-resolution components contribute complementarily to the overall performance. The radiologist evaluation confirms improved interpretation and diagnostic accuracy. The model is robust in spite of the noise level, providing a scalable, clinically applicable method for high quality, dose-efficient chest X-ray enhancement.

6. Conclusion

The suggested research efficiently reduces noise and low-resolution artifacts in low-dose imaging by introducing an effective deep learning architecture that combines U-Net++ super-resolution for chest X-ray enhancement with Swin Transformer-based denoising. Outperforming baseline CNN denoisers (31.93 dB) and SRCNN super-resolution models (32.41 dB), the model achieves peak signal-to-noise ratios (PSNR) of 34.72 dB on the NIH Chest X-Ray14 dataset and 33.85 dB on MIMIC-CXR. Structural similarity (SSIM) improves substantially to 0.982 (NIH) and 0.995 (MIMIC-CXR), while perceptual quality metrics (LPIPS) decrease to 0.098–0.105, indicating enhanced visual fidelity. According to ablation studies, the Swin Transformer and U-Net++ modules work together to improve performance, with full-model PSNR increases of up to 1.5 dB above single-module versions. Lesion detectability, diagnostic confidence, and lung structure visibility have been significantly improved, according to clinical radiologists' evaluations (+1.2 points on a 5-point scale). With PSNR increases of +1.92 dB (low noise), +2.79 dB (middle noise), and +3.34 dB (high noise), the framework consistently exhibits resilience across a range of noise levels. This system establishes the groundwork for AI-assisted medical imaging by providing a scalable end-to-end solution that prioritizes clinical accuracy, subjective quality, and quantitative integrity.

Ethics approval and consent to participate

Not Applicable

Data Availability

Publicly available in a repository: NIH Chest X Ray14 dataset, Available online at: https://nihcc.app.box.com/v/ChestXray-NIHCC. MIMIC-CXR dataset, Available online at: https://physionet.org/content/mimic-cxr

Consent for publication

Not Applicable.

Competing of interests

The authors declare no Conflict of interest.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors of this research work declare that they do not have any known competing or personal interests that can appear to influence the work in this paper.

Author Contribution

V. Muthulakshmi: Conceptualization, Methodology, Software, Data curation, Writing – original draft. K. Ramalakshmi: Supervision, Validation, Writing – review & editing. Hemalatha Gunaserakan: Writing – review & editing, Validation

Declaration of Generative AI Use

During the preparation of this manuscript, the authors used AI-assisted tools for language editing and grammar improvement. The authors reviewed and edited the content and take full responsibility for the final manuscript.

References

[1]. Singh, R. (2025). Understanding medical image denoising, enhancement, and reconstruction. Journal of Biomedical Informatics and Smart Healthcare, 1(1), 35–39.

[2]. Rani, S., & Kaur, N. (2023). Medical X-ray images enhancement based on super resolution convolution neural network. International Journal on Recent and Innovation Trends in Computing and Communication.

[3]. Khishigdelger, A., Salem, A., & Kang, H.-S. (2024). Elevating chest X-ray image super-resolution with residual network enhancement. Journal of Imaging, 10(3), 64.

[4]. Umirzakova, S., Mardieva, S., Muksimova, S., Ahmad, S., & Whangbo, T. (2023). Enhancing the super-resolution of medical images: Introducing deep residual feature distillation network. Bioengineering, 10(11), 1332.

[5]. Yoon, S.-H., Shin, C.-W., Park, C.-M., & Lee, D.-H. (2025). Improving image quality of chest radiography with AI-supported dual-energy X-ray imaging. Journal of Imaging.

[6]. Zhao, Z., Li, X., & Wang, J. (2024). Leveraging vision transformers and Swin-like hierarchical self-attention for medical image enhancement. Sensors, 23(7), 3420.

[7]. Pan, S., Wang, T., & Patel, A. B. (2023). 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. Physics in Medicine and Biology, 68(10), 105004.

[8]. Umirzakova, S. (2024). Medical image super-resolution for smart healthcare - A systematic survey. ScienceDirect, 2023–2024.
[9]. Abolade, O., Smith, R., & Choi, H. (2025). SharpXR: Structure-aware denoising for pediatric chest X-rays. arXiv.
[10]. Song, L., Sun, H., Xiao, H., Lam, S. K., Zhan, Y., Ren, G., & Cai, J. (2025). Artificial intelligence for chest X-ray image enhancement. Radiation Medicine and Protection, 6(1), 61–68.
[11]. Qin, J., Li, W., Zhang, Y., & Huang, P. (2025). CNN–Transformer gated fusion network for medical image super-resolution. Scientific Reports, 15, 15338.
[12]. Pan, S., Wang, T., Patel, A. B., & Lee, J. (2023). 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. PubMed.
[13]. Rajaraman, S., Siegelman, A., Folio, L., & Antani, S. (2024). Noise-induced modality-specific pretext learning for pediatric chest X-ray image classification. Frontiers in Artificial Intelligence, 7, 1419638.
[14]. Rajaraman, S., Folio, L., Siegelman, A., & Antani, S. (2025). Improving image quality of chest radiography with AI-supported dual-energy X-ray imaging. Journal of Clinical Medicine, 14(6), 2091.
[15]. Discover Computing. (2025). Enhanced medical diagnosis on chest X-rays: Knowledge distillation from self-supervised models. Discover Computing, 28, 118.
[16]. Çallı, E., Yildirim, S., & Alkan, B. (2025). Deep learning for chest X-ray classification: A comprehensive review. Journal of Al-Qadisiyah for Computer Science and Mathematics.
[17]. Zhang, B., Li, C., Wang, X., & Zhou, Y. (2025). Enhanced Swin Transformer based tuberculosis classification with segmentation. PubMed.
[18]. Kumar, V., Anand, V., & Sharma, R. (2025). Enhanced pneumonia detection in chest X-rays using hybrid CNN and vision transformer networks. Current Medical Imaging, 21(3), 125–134.
[19]. Patel, S., Anand, V., & Singh, R. (2024). Optimizing pulmonary chest X-ray classification with stacked feature ensemble and Swin Transformer integration. Biomedical Physics & Engineering Express, 10(4), 045012.
[20]. Li, H., Zhang, X., & Wang, J. (2023). Asymmetric convolution Swin Transformer for medical image super-resolution. Alexandria Engineering Journal, 85, 177–184.
[21]. Bertram, F., Mueller, H., & Schmidt, K. (2023). Deep-learning-based denoising of X-ray differential phase and dark-field images. European Journal of Radiology, 163, 110835.
[22]. Song, L., Sun, H., Lam, S. K., & Zhan, Y. (2023). Chest X-ray image enhancement using deep contrast diffusion learning. Optik, 279, 170751.
[23]. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., & Loy, C. C. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. arXiv.
[24]. Prevedello, L. M., Erdal, B. S., Ryu, J. L., & Mandell, J. (2019). Clinical integration studies on AI-supported chest X-ray interpretation. Radiology: Artificial Intelligence, 1(3), e180031.
[25]. NIH Chest X-Ray14 dataset, Available online at: https://nihcc.app.box.com/v/ChestXray-NIHCC (Accessed: 19 February 2026).
[26]. MIMIC-CXR dataset, Available online at: https://physionet.org/content/mimic-cxr/ (Accessed: 19 February 2026).