POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Attention-enhanced U-net autoencoder for Low-dose CT simulation and lung nodule malignancy assessment

Tesi di Laurea Magistrale in
Biomedical Engineering - Ingegneria Biomedico

Author: **Reza Gonabadi**

Student ID: 10872040
Advisor: Prof. Luca Mainardi
Co-advisors: Jiaying Liu
Academic Year: 2024-2025

Abstract

This thesis proposes a two-stage deep learning pipeline that addresses the dual challenge of radiation dose reduction in computed tomography (CT) and robust lung nodule malignancy assessment under low-dose conditions. The first phase focuses on simulating realistic low-dose CT (LDCT) images from high-dose scans using a progressively enhanced encoder-decoder architecture. Starting from a basic autoencoder, the model is refined with U-Net skip connections and attention mechanisms to better preserve anatomical detail and spatial noise structure. The final model achieves strong reconstruction fidelity, with a Peak Signal-to-Noise Ratio (PSNR) of **32.02 dB** and a Structural Similarity Index Measure (SSIM) of **0.9311**.

In the second phase, the simulated LDCT images are used to train and evaluate machine learning models for classifying pulmonary nodules as benign or malignant. Radiomic features, comprising shape, texture, and intensity descriptors, are extracted and subjected to a structured selection pipeline. Multiple classifiers are tested across three feature configurations: shape-only, non-shape, and full feature sets. Among these, logistic regression trained on shape-only features achieved the most robust and generalizable performance, with a balanced accuracy of **80.0%** and a test ROC AUC of **0.839**, outperforming more complex models and feature combinations. While ensemble models like Random Forest reached up to **79.2%** accuracy post-augmentation, their generalization was less stable across feature types.

This work contributes to the development of safer and more reliable radiological pipelines by combining realistic low-dose image simulation with interpretable and effective malignancy classification. The results highlight the value of attention-guided architectures for dose-aware image synthesis and the clinical utility of compact, shape-based features in low-quality imaging scenarios.

Keywords: Low-dose CT simulation, Deep learning, Autoencoder, U-net, Attention mechanism, Lung nodule classification, radiomics

Abstract in lingua italiana

Questa tesi propone una pipeline di deep learning in due fasi che affronta la doppia sfida della riduzione della dose di radiazioni nella tomografia computerizzata (CT) e della valutazione affidabile della malignità dei noduli polmonari in condizioni di bassa dose. La prima fase si concentra sulla simulazione realistica di immagini CT a bassa dose (LDCT) a partire da scansioni ad alta dose, utilizzando un'architettura encoder-decoder progressivamente migliorata. A partire da un autoencoder di base, il modello è stato potenziato con connessioni skip stile U-Net e meccanismi di attenzione per preservare meglio i dettagli anatomici e la struttura del rumore spaziale. Il modello finale ha raggiunto un'elevata fedeltà di ricostruzione, con un Peak Signal-to-Noise Ratio (PSNR) pari a **32,02 dB** e un Structural Similarity Index Measure (SSIM) pari a **0,9311**.

Nella seconda fase, le immagini simulate LDCT vengono utilizzate per addestrare e valutare modelli di apprendimento automatico destinati alla classificazione dei noduli polmonari come benigni o maligni. Le caratteristiche radiomiche, comprendenti descrittori di forma, texture e intensità, vengono estratte e sottoposte a una pipeline strutturata di selezione. Diversi classificatori sono testati su tre configurazioni di feature: solo forma, non-forma, e set completo. Tra questi, la regressione logistica addestrata esclusivamente su caratteristiche morfologiche ha mostrato le prestazioni più robuste e generalizzabili, con un'accuratezza bilanciata del **80,0%** e un ROC AUC di **0,839** sul test set, superando modelli più complessi e combinazioni di feature. Sebbene modelli ensemble come Random Forest abbiano raggiunto fino al **79,2%** di accuratezza dopo l'augmentazione, la loro generalizzazione si è dimostrata meno stabile tra i diversi tipi di feature.

Questo lavoro contribuisce allo sviluppo di pipeline radiologiche più sicure e affidabili, combinando una simulazione realistica di immagini a bassa dose con una classificazione della malignità interpretabile ed efficace. I risultati evidenziano il valore delle architetture guidate dall'attenzione per la sintesi di immagini consapevoli della dose e l'utilità clinica di descrittori morfologici compatti in scenari di imaging a bassa qualità.

Parole chiave: Simulazione CT a bassa dose, Deep learning, Autoencoder, U-net, Meccanismo di attenzione, Classificazione dei noduli polmonari, radiomica

Contents

1 | Introduction

The widespread use of computed tomography (CT) in clinical diagnostics has significantly improved the early detection and management of critical diseases such as lung cancer. However, the high radiation exposure associated with standard-dose CT (SDCT) scans raises long-term safety concerns, especially in screening programs where repeated imaging is necessary. In response, low-dose CT (LDCT) protocols have been introduced as a safer alternative, reducing ionizing radiation levels while maintaining diagnostic intent. LDCT is particularly relevant in lung cancer screening, where early-stage detection can save lives, but frequent imaging increases cumulative radiation burden, particularly for high-risk populations.

Despite its clinical value, the adoption of LDCT is constrained by a fundamental trade-off: lowering the radiation dose degrades image quality. LDCT images typically exhibit increased noise, reduced contrast, and poorer spatial resolution, which can obscure subtle anatomical details such as small pulmonary nodules. These degradations complicate downstream diagnostic tasks and can adversely impact the sensitivity of radiologists or machine learning models in detecting and classifying lesions. As a result, many medical imaging pipelines continue to rely on SDCT data, even though it does not accurately represent the noise characteristics encountered in real-world low-dose protocols.

One major barrier to LDCT-focused research and development is the scarcity of publicly available, high-quality low-dose datasets. Most annotated medical imaging datasets, including those used for computer-aided diagnosis or radiomics studies, are acquired under standard-dose conditions. This mismatch creates a significant challenge when developing and validating machine learning models intended for use in LDCT contexts. In particular, lung nodule classification, a key task in cancer screening, relies heavily on clear morphological and textural features, which are often obscured in LDCT scans. Training classifiers directly on standard-dose images may lead to poor generalization under low-dose conditions, undermining the clinical value of such systems.

Moreover, the classification of pulmonary nodules is inherently challenging, even under optimal imaging conditions. Malignant nodules are typically underrepresented in clinical

datasets, leading to pronounced class imbalance. This imbalance reduces the sensitivity of classifiers, increasing the risk of false negatives and delayed diagnosis. Additionally, radiomic features extracted from LDCT images are sensitive to noise and reconstruction artifacts, which can degrade their stability and predictive power. Ensuring robustness under these conditions requires careful feature selection, reliable augmentation techniques, and a pipeline that reflects the realities of clinical LDCT imaging.

In light of these challenges, there is a pressing need for strategies that can bridge the gap between standard-dose data availability and low-dose clinical requirements. Generating realistic LDCT images from existing SDCT scans offers one such pathway, enabling the creation of synthetic datasets that reflect true dose-dependent characteristics. These datasets can support the development of more reliable classification models while minimizing additional radiation exposure to patients. At the same time, robust nodule classification frameworks, capable of handling noisy, imbalanced LDCT data, are essential to maximize the diagnostic value of safer imaging protocols. This thesis addresses both of these needs through a unified deep learning pipeline for LDCT simulation and lung nodule malignancy classification.

2 State of the Art

2.1. Low-dose CT simulation techniques

Simulating low-dose computed tomography (LDCT) images is crucial to enable the development of algorithms tailored to low-dose scenarios, while avoiding the ethical and practical limitations of repeated radiation exposure. Researchers have developed several techniques that introduce realistic noise characteristics into full-dose CT (FDCT) data, either by modifying raw projection data or working directly in image space.

2.1.1. Projection-domain simulation: noise modeling and equations

The most physically grounded method of LDCT simulation involves manipulating the projection data (sinograms) by introducing appropriate noise distributions. Zeng et al. [33] proposed a straightforward strategy: given a high-dose projection, the low-dose version is generated by scaling the incident flux and injecting statistical noise.

$$P_{\mathrm{LD}} = P_{\mathrm{HD}} + \mathcal{N}_{\mathrm{Poisson}} + \mathcal{N}_{\mathrm{Gaussian}}, \tag{2.1}$$

where P_{LD} and P_{HD} represent the low- and high-dose projections, respectively. Poisson noise models quantum fluctuation in photon counts, while Gaussian noise accounts for system electronics.

Yu et al. [32] provided a more detailed noise insertion algorithm. The detected data P_A at high dose is given by:

$$P_A = \ln\left(\frac{N_{0A}}{N_A}\right), \tag{2.2}$$

where N_{0A} and N_A are the incident and detected photons, respectively. Under a reduced dose with scaling factor α, the new signal becomes:

$$P_B = P + \frac{1}{\sqrt{\alpha N_{0A} \exp(-P)}} \cdot x, \tag{2.3}$$

where $x \sim \mathcal{N}(0, 1)$. This provides a stochastic realization of a low-dose projection without re-acquisition.

Zeng et al. demonstrated high fidelity in both the sinogram domain and the reconstructed images. Their approach was validated through visual inspection as well as statistical comparisons, confirming the effectiveness of their noise-injection technique. Similarly, Yu et al. reported that their stochastic projection noise generation method produced realistic and dose-consistent simulations without the need for additional data acquisition, highlighting the physical plausibility of their approach.

2.1.2. Image-domain simulation: synthetic sinograms

When sinograms are unavailable, Naziroglu et al. [16] proposed reconstructing a synthetic sinogram from the FDCT image using the Radon transform.

$$s(x, \theta) = \int_{-\infty}^{\infty} \mu(x \cos\theta + y \sin\theta, y)\, dy, \tag{2.4}$$

where $\mu(x, y)$ is the attenuation coefficient. Noise was then added in projection space, followed by filtered back-projection to yield a noise-only image, which was summed with the original:

$$I_{\text{LD}} = I_{\text{HD}} + I_{\text{noise}}. \tag{2.5}$$

The noise parameters, such as bowtie filter profile and detector noise, were estimated using calibration scans on water phantoms.

Naziroglu et al. validated their synthetic sinogram approach by analyzing the noise power spectrum (NPS) and modulation transfer function (MTF). Their results showed that both noise characteristics and spatial resolution were well preserved, indicating the effectiveness of their method in replicating realistic low-dose image features.

2.1.3. DICOM-based methods and noise decoupling

In scenarios where access to raw sinogram data is unavailable, image-domain simulation becomes a practical alternative. Kim and Kim [11] proposed a framework that begins

with standard DICOM CT images and removes existing noise via a total variation (TV) minimization process:

$$\mu_{\text{denoised}} = \arg\min_{\mu} \left(TV(\mu) + \lambda \|\mu - \mu_{\text{orig}}\|_2^2 \right), \tag{2.6}$$

where μ_{orig} is the original full-dose CT image, and λ is a regularization parameter controlling the trade-off between noise suppression and image fidelity. This denoising step yields a clean image μ_{denoised} that can serve as a baseline for synthetic noise injection, enabling realistic simulation of low-dose characteristics.

Next, the denoised image is forward-projected to simulate sinogram data under a fan-beam geometry. The synthetic projection value $A(d_i, g_j)$ at a given detector index d_i and gantry angle g_j is computed as:

$$A(d_i, g_j) = \Delta t \sum_{n=0}^{N_r} \mu(x(n\Delta t), y(n\Delta t)), \tag{2.7}$$

where Δt is the sampling interval along the ray path, N_r is the number of sampled points along each ray, and $\mu(x(n\Delta t), y(n\Delta t))$ represents the linear attenuation coefficient at spatial coordinates (x, y) along the ray. The index n runs from 0 to N_r, effectively integrating the attenuation along the projection path.

To decouple the new noise from the original signal, a synthetic noise-only image is created by injecting Poisson-Gaussian noise into the simulated sinogram. This noisy sinogram is reconstructed using filtered back projection (FBP), resulting in a noise map I_{noise} that is then added to the original image to form the final low-dose simulation:

$$I_{\text{LD}} = I_{\text{HD}} + I_{\text{noise}}, \tag{2.8}$$

where I_{HD} denotes the original high-dose CT image. This additive noise model ensures that the simulated noise is independent from the original image noise, leading to more realistic low-dose characteristics.

Kim and Kim [11] validated this approach through visual inspection and quantitative metrics, including the noise power spectrum (NPS), demonstrating that the synthetic LDCT images faithfully reproduce spatial noise patterns and dose-dependent effects seen in real low-dose acquisitions.

Kim and Kim validated their method through both noise power spectrum (NPS) analysis

and visual inspection, demonstrating strong consistency with real low-dose CT scans in terms of spatial structure and noise characteristics. Additionally, Takenaga et al. [29] showed that their DICOM-based simulation approach closely matched real low-dose acquisitions, with deviations in standard deviation remaining below 3% and the noise power spectrum (NPS) and modulation transfer function (MTF) curves aligning within 1%, confirming the quantitative reliability of their simulation.

2.1.4. Summary and Limitations

Although projection-based methods offer superior realism by capturing scanner physics and dose-dependent signal statistics, they are often limited by restricted access to raw sinogram data and scanner calibration details, which are typically proprietary or unavailable in public datasets.

Image-domain simulation techniques, on the other hand, are more accessible but can introduce artifacts or unrealistic noise textures, particularly when the denoising and renoising steps are not finely calibrated. These methods often rely on simplifications such as additive Gaussian noise or uniform noise scaling, which may fail to capture the spatially and spectrally varying characteristics of actual LDCT scans.

Moreover, many existing simulation strategies are based on static formulas and deterministic transformations that do not account for variability across scanner models, acquisition protocols, or patient anatomy. They also depend on accurate metadata from DICOM headers (e.g. tube current, exposure time, slice thickness) to approximate dose-dependent behavior, parameters that may be missing, anonymized, or imprecisely encoded in practice.

An overview of existing LDCT simulation methods and their validation strategies is provided in Table 2.1, highlighting the diversity of metrics and the frequent lack of quantitative NPS benchmarking.

2.2. Lung nodule classification using machine learning

Recent advances in radiomics and machine learning have enabled the development of predictive models to assess the malignancy of nodules using LDCT images. These models often rely on handcrafted radiomic features, quantitative descriptors of nodule shape, texture, and intensity, combined with classical machine learning algorithms to distinguish

Table 2.1: Validation of LDCT simulation methods in literature

Method & Author	Validation Metrics	Highlights
Zeng et al. [33]	Visual and statistical match	High-fidelity sinogram and image generation; qualitative NPS realism without quantitative error
Yu et al. [32]	Stochastic behavior modeling	Dose-consistent noise patterns with no re-acquisition; NPS not quantitatively validated
Naziroglu et al. [16]	NPS and MTF	Relative RMSE for NPS <15%; preserved spatial resolution and realistic noise texture
Kim and Kim [11]	TV + NPS + visual validation	Realistic noise structure observed; spectral fidelity dependent on reconstruction strategy
Takenaga et al. [29]	SD, NPS, MTF in phantoms	SD deviation <3%; NPS and MTF errors ∼1% in phantom studies (approx. RMSE ≤15%)

benign from malignant nodules. The following section summarizes key recent studies conducted in this direction.

2.2.1. Overview of radiomic-based classifiers

A broad range of approaches have been proposed that differ in datasets used, feature selection strategies, classifiers, and validation protocols.

Choi et al. [5] used radiomic features from the LIDC-IDRI dataset (nodules $\geq$ 3 mm) and selected the most discriminative features using LASSO, followed by a Support Vector Machine (SVM) classifier. Their model achieved 87.2% sensitivity and 84.6% accuracy, although the dataset size was relatively small.

Peikert et al. [19] developed a radiomic classifier trained on 726 nodules from the NLST dataset (nodules $\geq$ 7 mm). Eight features were selected using LASSO regression and yielded an AUC of 0.94 and sensitivity of 90.4%. However, they intentionally balanced the dataset, which may have introduced a selection bias.

Alahmari et al. [1] utilized 18 features from the RIDER dataset, applying ReliefF for feature selection and a Random Forest classifier. Despite high specificity (93.0%), sensitivity was relatively low (49.0%), possibly due to imbalanced feature selection and model complexity.

Causey et al. [3] proposed a minimalistic model based solely on nodule size, using the square root of the cross-sectional area as a single feature in a logistic regression classifier. While the AUC was high (0.94), the sensitivity was only 69.7%, indicating size alone is insufficient for robust malignancy prediction.

Mao et al. [15] developed a linear regression model on 294 nodules (6–15 mm) using 11 radiomic features selected via LASSO. The model yielded an AUC of 0.90 and an accuracy of 89.8%, but generalizability was constrained by limited sample size.

Garau et al. [6] externally validated their SVM-LASSO model on the COSMOS dataset and tested it on 72 LIDC-IDRI nodules, achieving 90% sensitivity and AUC of 0.86. However, the model exhibited a high false positive rate (35.5%).

Liu et al. [13] compared LDCT-based radiomic models with those based on standard-dose CT and found their LDCT-trained model with 3 features achieved an AUC of 0.98. Despite high performance, the small cohort (141 nodules) and lack of detailed acquisition settings raise concerns about overfitting.

Rundo et al. [23] used 32 radiomic features with Borderline-SMOTE for oversampling and applied elastic net regularization. Their model was evaluated on 703 nodules from the bioMILD trial using blind testing and showed high specificity (86.2%) but relatively low sensitivity (67.0%).

Liu et al. [14] developed a radiomics-based classifier using synthetic LDCT images derived from degraded standard-dose CT scans from the LIDC-IDRI dataset, totaling 1950 nodules for training. They evaluated three feature sets and identified a logistic regression model trained with only three shape and size (SS) features as the most robust, achieving 81.0% balanced accuracy and 87.0% AUC on the test set. Their pipeline incorporated rigorous preprocessing, stability analysis, and explainability via SHAP. While the approach effectively addressed data scarcity and achieved competitive performance, a key limitation was the use of nodule labels based on radiologists' subjective malignancy scores, as the dataset was originally designed for segmentation rather than screening.

2.2.2. Challenges and Limitations

While reported AUCs and classification accuracies across the literature are promising, several key limitations remain that hinder the clinical translation of radiomics-based malignancy assessment models. A common issue is the limited sample size and lack of data diversity in many studies, which restricts the generalizability of the results to broader patient populations and imaging conditions. Additionally, the heavy dependence on handcrafted

Table 2.2: Summary of recent radiomics-based lung nodule malignancy classification studies

Study	Dataset (nodule size)	Model Description	Limitations
Choi et al. (2018) [5]	LIDC-IDRI ($\geq$ 3 mm), 72 pts	2 radiomic features, LASSO, SVM	Small dataset
Peikert et al. (2018) [19]	NLST ($\geq$ 7 mm), 726 nodules	8 features, LASSO logistic regression	Dataset balancing
Alahmari et al. (2018) [1]	NLST ($\geq$ 4 mm), 467 cases	18 RIDER features, ReliefF, Random Forest	Low sensitivity
Causey et al. (2018) [3]	LIDC-IDRI ($\geq$ 3 mm), 664 nodules	Logistic regression on size metric	Only size feature
Mao et al. (2019) [15]	6–15 mm, 294 nodules	11 features, LASSO, linear regression	Small sample size
Garau et al. (2020) [6]	COSMOS (all sizes), 113 nodules	5 features, SVM-LASSO, tested on LIDC	High FPR, small set
Liu et al. (2021) [13]	141 nodules (size N/A)	3 features, radiomic nomogram	Potential selection bias
Rundo et al. (2021) [23]	bioMILD (all sizes), 703 nodules	32 features, SMOTE, elastic net LR	Low sensitivity
Liu et al. (2022) [14]	LIDC-IDRI, 1950 nodules	Different classifiers using three feature groups	Labels based on radiologists' malignancy likelihood; not tailored for classification/screening

radiomic features poses a challenge, as these features can be highly sensitive to variations in scanner types, imaging protocols, and reconstruction parameters. This variability is particularly problematic in LDCT, where the presence of noise can significantly distort radiomic measurements. Moreover, the absence of validation procedures that specifically account for variations in radiation dose levels in some studies undermines the stability and reliability of radiomic features under LDCT-specific conditions. Another concern is the risk of overfitting, which is often amplified by the lack of external or blind test sets, limiting the assessment of model robustness in real-world clinical scenarios.

3 | Materials and Methods

3.1. Low-dose CT simulation

3.1.1. Dataset and Preprocessing

Dataset description

This study involves two distinct datasets, each serving a different stage of the modeling pipeline. The first dataset, used for developing and evaluating the low-dose CT simulation model, comes from the 2016 Low Dose CT Grand Challenge [17]. This publicly available resource was created to support research into denoising and reconstruction methods for low-dose imaging, especially in the chest region.

Although the original dataset consisted of 299 anonymized CT scans spanning the head, chest, and abdomen, this work focuses exclusively on chest CT slices. Scans from other anatomical areas were excluded. Each subject folder contains both reconstructed image volumes in DICOM format and raw projection data in DICOM-CT-PD format; however, only the image volumes were used, as the projection data were not required for this image-domain simulation.

From the available data, only 99 subjects were complete and accessible for download. Of these, 50 subjects included both full-dose and simulated low-dose reconstructed image series, forming the subset used for training and evaluation. Each of these subjects contributed between 250 and 350 matched CT slice pairs, resulting in approximately 16,500 image pairs in total. A summary of the dataset composition is provided in Table 3.1.

An illustrative example from the dataset is shown in Figure 3.1, comparing a full-dose and its corresponding simulated low-dose slice (subject 21, slice 160). The degradation in image quality due to dose reduction is clearly visible, with the low-dose image exhibiting elevated noise and reduced contrast.

All CT scans were acquired using a tube potential of 120 kV and a reference tube current of 200 mAs, following the standard clinical protocol of the host institution.

Table 3.1: Summary of Dataset Composition for Simulation Task

Total Subjects in Dataset	299
Subjects Available for Download	99
Subjects with Paired Full-Dose and Low-Dose Images	50
CT Slice Pairs per Subject (Approx.)	250 – 350
Total Paired Slices Used	∼16,500

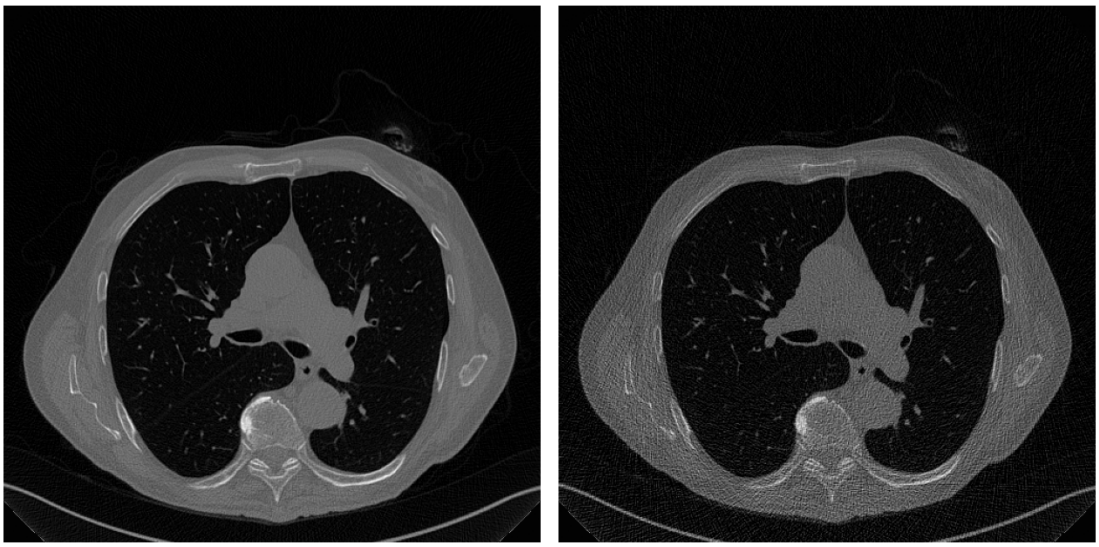

Figure 3.1: Comparison between full-dose (left) and simulated low-dose (right) CT images for subject 21, slice 160. The low-dose version exhibits increased noise and reduced contrast, typical of dose-reduced scans.

The second dataset used in this study is the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [2], which supports the nodule classification component. This publicly available dataset comprises 1,018 thoracic CT scans, all accessible for download. However, 8 of these were identified as duplicates and excluded from analysis, resulting in 1,010 unique scans used in this work.

Each scan is accompanied by lesion annotations provided by one to four experienced thoracic radiologists. The annotation process followed a two-phase protocol: in the first phase, each radiologist independently reviewed the scans and annotated visible lesions; in the second phase, they revised their annotations while viewing anonymized feedback from the other readers. This iterative process preserved both individual perspectives and consensus-based insights, yielding high-quality annotations suitable for the development and evaluation of nodule classification models.

The structured XML annotations include nodules $\geq$3 mm, nodules $<$3 mm, and non-nodular abnormalities $\geq$3 mm. In this study, only nodules with a diameter of at least 3 mm were considered, as they are more clinically relevant and reliably annotated. Moreover, consensus annotations were adopted for these nodules, reflecting agreement across multiple radiologists to define robust lesion boundaries. Figure 3.2 shows an example CT slice from the LIDC-IDRI dataset, with a visible lung nodule marked for classification.

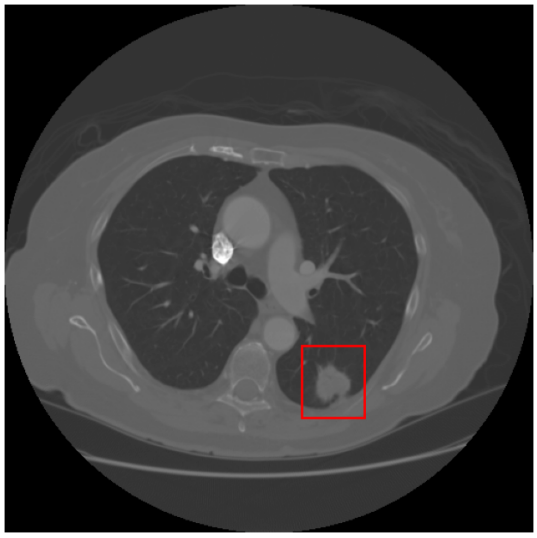

Figure 3.2: Example CT slice from the LIDC-IDRI dataset showing a visible lung nodule.

The malignancy labels in the LIDC-IDRI dataset are based on a five-point ordinal scale rated by radiologists, where each nodule is scored according to the likelihood of being malignant:

1 Highly Unlikely

2 Moderately Unlikely

3 Ambiguous

4 Moderately Suspicious

5 Highly Suspicious

Each nodule was annotated by up to four different radiologists. To enable binary classification, a consensus-driven scheme was employed by averaging the available malignancy scores. A nodule was labeled as *Malignant* if the average score was ≥ 4.0, and as *Benign* otherwise. This threshold reflects clinical intuition and aligns with prior literature [2], treating scores of 4 or 5 as indicative of potential malignancy.

Table 3.2: Summary of the LIDC-IDRI dataset used in this study.

Property	**SDCT**	**LDCT**
Nr. participants	694	316
Nr. nodules	1950	675
Malignancy score [1;2]	280	86
Malignancy score [2;3]	766	222
Malignancy score [3;4]	679	298
Malignancy score [4;5]	225	69
Benign (avg. score < 4)	1725	606
Malignant (avg. score 4)	225	69

Together, the Low Dose CT Grand Challenge dataset and the LIDC-IDRI dataset support the dual objectives of this research: simulation of low-dose CT scans and classification of lung nodule malignancy. The simulation model was trained on high-quality paired full-dose and low-dose scans and then applied to the LIDC-IDRI dataset to generate synthetic low-dose images. Malignancy classification was subsequently performed on these simulated slices using the validated annotations provided by expert radiologists.

Artifact removal: suppressing scanner-induced bottom lines

Some CT slices, particularly in abdominal regions, display bright linear or curved artifacts near the bottom of the image. These artifacts usually originate from the scanning table or nearby equipment and appear as high-intensity features that do not belong to the anatomy. If left unaddressed, they may mislead the learning process by introducing strong, irrelevant

signals. Figure 3.3 shows an example slice with such artifacts highlighted by red markings.

Figure 3.3: Example CT slice with artifact region highlighted in red. These scanner-induced structures appear in the lower part of some images and are unrelated to anatomical content.

To address this issue, a custom image processing routine was applied to each slice before training. The procedure begins by binarizing the grayscale image using Otsu's method, which determines an optimal intensity threshold that separates foreground objects from the background. Once binarized, connected components in the image are labeled and analyzed for shape properties such as area, eccentricity, and centroid location.

The goal is to identify thin, elongated structures in the lower portion of the image. Components with eccentricity above 0.9, located in the bottom third of the image, and within a moderate area range (between 50 and 5000 pixels) are flagged as potential artifacts. These regions are then isolated into a binary mask, and all pixels within the mask are set to zero in the original image, effectively removing the artifact while preserving surrounding tissues.

This artifact removal process was applied consistently across both the full-dose and low-dose series. Figure 3.4 presents the full-dose and low-dose images of subject 162 (slice 120) before preprocessing, with the artifact clearly visible. Figure 3.5 shows the corresponding slices after preprocessing, where the artifacts have been successfully suppressed and the anatomical information remains intact.

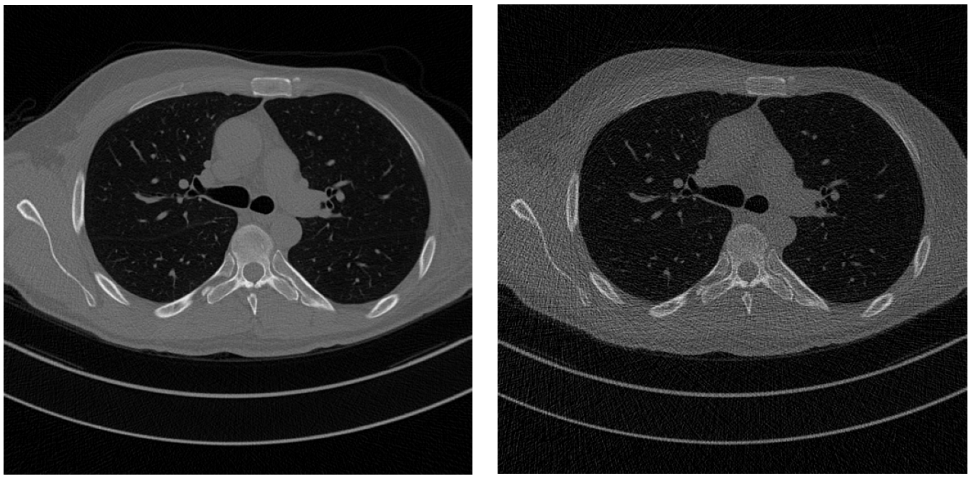

Figure 3.4: CT slices from subject 162 (slice 120) before artifact removal. Left: full-dose. Right: low-dose. Scanner-related lines are visible near the bottom edge.

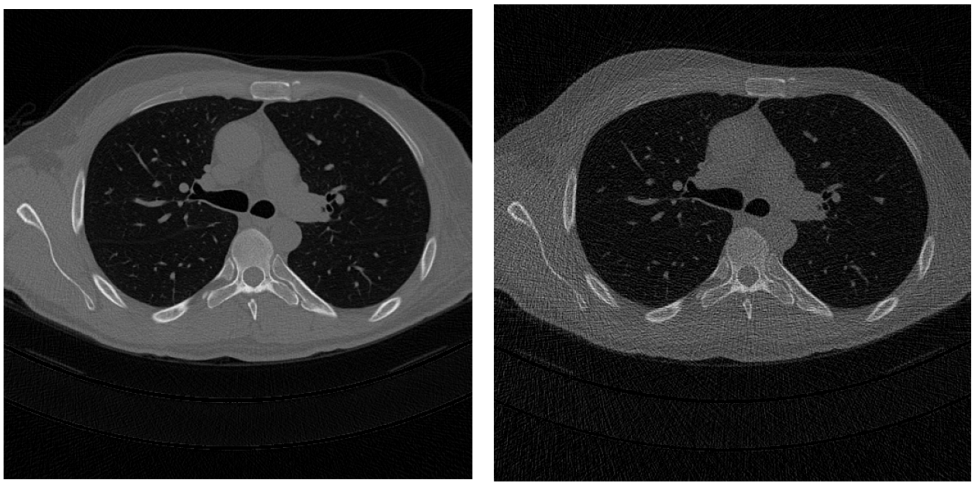

Figure 3.5: Same subject and slice after artifact removal. Left: full-dose. Right: low-dose. Bottom-line artifacts have been effectively removed while preserving anatomical content.

3.1.2. Deep learning-based low-dose simulation

After preprocessing and removing curve-line artifacts, the image pairs are prepared for subsequent stages. This work presents a progressive development of deep learning architectures for LDCT simulation, with each approach building upon the limitations of the previous one.

The modeling process begins with a simple autoencoder (AE) that learns compressed representations of FDCT images, establishing fundamental reconstruction capabilities. To address spatial detail loss, a U-Net architecture is introduced, which incorporates skip connections between encoder and decoder layers (Section 3.1.2). This framework is further enhanced with attention mechanisms that dynamically weight the importance of the features during reconstruction.

The architectural evolution illustrates how increasingly sophisticated neural networks can better preserve anatomical structures while simulating realistic noise characteristics. Each model variant is carefully designed to maintain computational efficiency while improving specific aspects of image quality: from global reconstruction accuracy to localized feature preservation and adaptive attention to diagnostically relevant regions.

Simple AE architecture

Given that both the input and output in our task are image volumes, specifically, full-dose and low-dose CT scans, an encoder-decoder structure is a natural and widely adopted choice for image-to-image translation tasks [4, 8]. This structure enables the network to compress input images into lower-dimensional representations while preserving contextual and structural information, which is critical for high-fidelity reconstruction in medical imaging [16, 32, 33].

As a foundational approach, a basic AE was selected due to its simplicity, stable training behavior, and solid baseline performance in related tasks such as image generation and transformation [4, 11]. In this context, the AE was trained to map FDCT images to their LDCT counterparts by learning to inject realistic noise patterns. This use of AEs leverages their strength in capturing global image structure while introducing controlled degradations, consistent with the objectives of low-dose CT simulation.

Model design

The AE consists of two main parts: an encoder and a decoder, as shown in Figure 3.6.

The process begins with an encoder that transforms the full-dose CT input image into

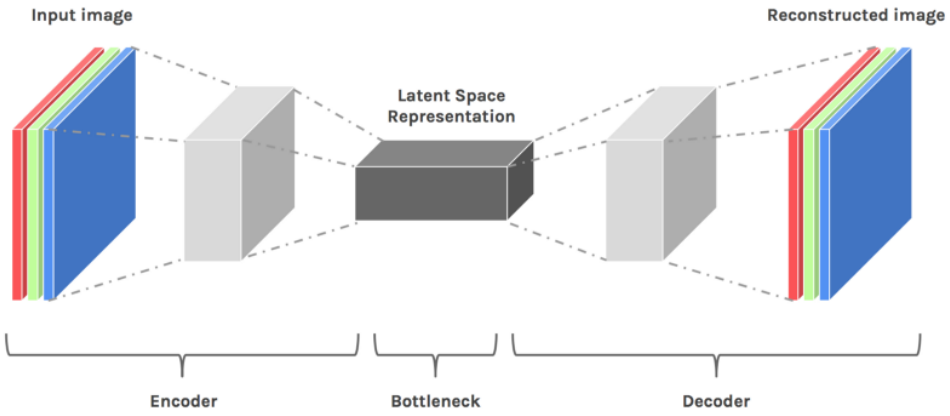

Figure 3.6: Simple AE architecture, source from [21].

a compressed latent representation. It uses a series of convolutional layers to extract hierarchical spatial features, with ReLU activations introducing non-linearity, batch normalization for training stability, and max pooling for spatial downsampling. Figure 3.7 illustrates a simplified conceptual view of the encoder. The colored vertical blocks represent successive hidden layers, each containing a certain number of neurons or feature channels. The transition from blue to green to purple blocks depicts the gradual reduction in dimensionality, where the feature space is increasingly compressed. The dashed arrows represent the dense connections or learned transformations between each layer.

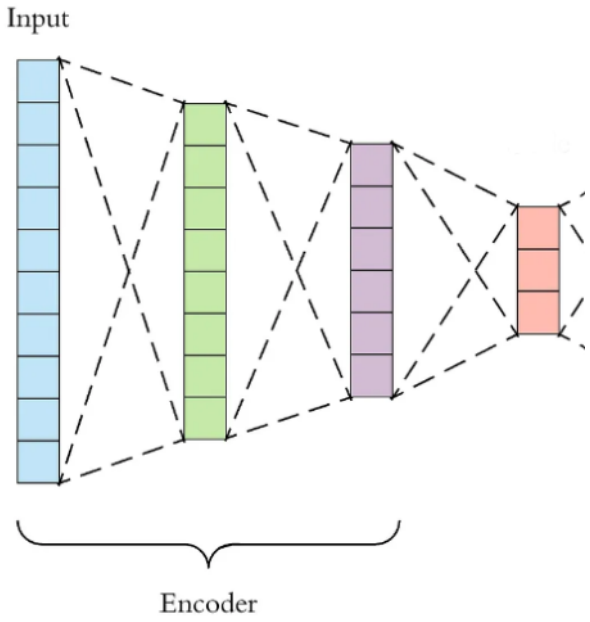

Figure 3.7: Encoder architecture, illustrating progressive compression from input to latent space. Source: [25].

Following this, the decoder takes the encoded latent vector and reconstructs the corresponding low-dose CT image. It reverses the encoder's operations by progressively upsampling the data using transposed convolution layers, again with ReLU activations and batch normalization to refine and stabilize the reconstruction. As shown in Figure 3.8, the compressed red block in the center represents the latent code. Moving rightward, the network expands the dimensionality layer by layer (purple → green → blue), aiming to restore the spatial resolution and detail of the original input. Like the encoder diagram, the dashed lines represent the learned mappings between each transformation step.

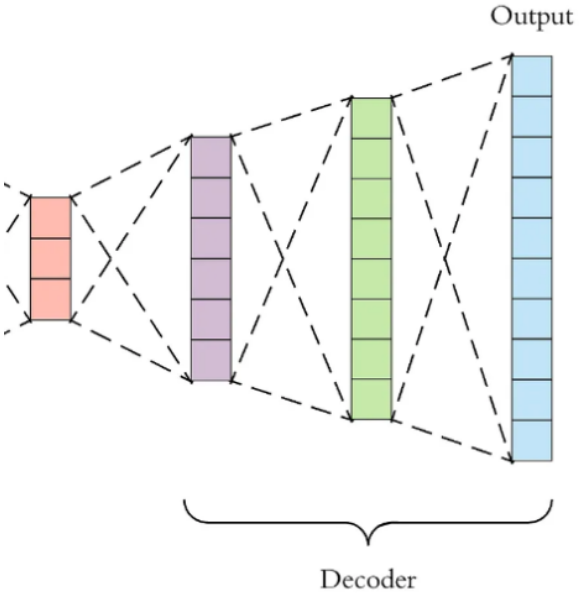

Figure 3.8: Decoder architecture, illustrating the progressive reconstruction toward the output image. Source: [25].

To guide the training process, the network uses the **Mean Absolute Error (MAE)** loss function, which minimizes pixel-wise intensity differences and is widely used in image regression tasks for its robustness [4]. The MAE loss is defined as:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{3.1}$$

where:

- $\hat{y}_i$ is the predicted pixel intensity,
- y_i is the ground truth,

- N is the total number of pixels in the image.

Although MAE offers a solid baseline for comparing generated and ground truth images, it measures only absolute pixel-level differences and does not reflect perceptual quality or structural similarity. This limitation is particularly critical in medical imaging, where preserving fine anatomical details is essential. Therefore, in the following sections, MAE is complemented with additional loss functions to capture more complex image-level characteristics and improve the realism and diagnostic relevance of the reconstructions.

U-Net extension for structural preservation

Due to the inherent limitations of the simple AE architecture, most notably, its inability to preserve fine anatomical textures and structural edges in reconstructed low-dose CT images, the model was extended using a U-Net-style encoder-decoder architecture. U-Net is widely recognized in the medical imaging field for its capability to maintain spatial resolution and contextual integrity, owing to its use of skip connections between encoder and decoder layers. These connections allow low-level features extracted during encoding to be directly reused during decoding, facilitating better preservation of local anatomical structures and improving overall reconstruction quality [4, 22].

U-Net has been chosen because it helps preserve image texture, which is particularly important in medical imaging. It improves upon standard AEs by introducing skip connections between the encoder and decoder layers at the same resolution level. These connections allow high-resolution features extracted during encoding to bypass the latent space bottleneck and be directly reused during decoding. This strategy effectively preserves fine-grained structures, edges, and textures, which are critical for image quality in diagnostic contexts [18, 33]. An overview of this enhanced architecture, incorporating skip connections, is illustrated in Figure 3.9.

The network retains the core encoder-decoder structure but integrates lateral connections from the encoder layers to their mirrored decoder layers. This architecture bridges the semantic gap between feature abstraction and spatial precision [22].

To prevent overfitting and improve generalization on unseen data, dropout layers with a dropout rate of 30% were applied in both the encoder and decoder. This stochastic element helps the model learn more robust and generalizable feature mappings [28]. Batch normalization was also maintained to stabilize training dynamics and accelerate convergence, which becomes more important as the architecture grows deeper [10].

The loss function was refined to incorporate a weighted combination of metrics:

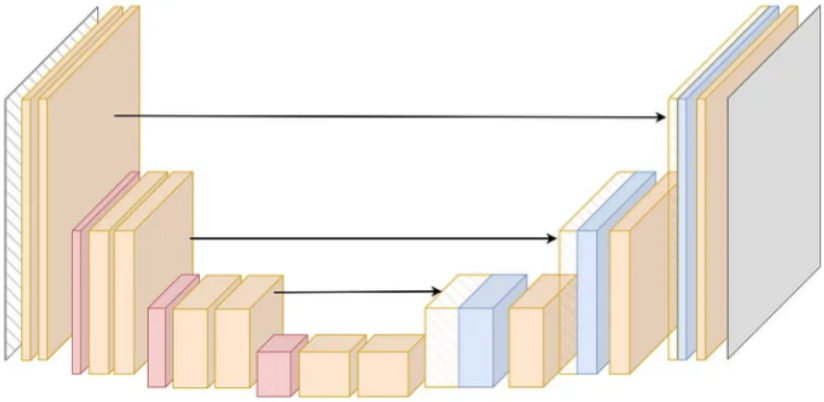

Figure 3.9: AE with skip connections source from [27].

- MAE (30% weight): Focuses on pixel-level reconstruction accuracy [4].
- Structural Similarity Index Measure (SSIM) (70% weight): Prioritizes structural integrity and perceptual quality [31].

This balanced formulation emphasizes preservation of anatomical structures while maintaining precise intensity matching:

$$\mathcal{L}_{\text{total}} = 0.3 \cdot \mathcal{L}_{\text{MAE}} + 0.7 \cdot \mathcal{L}_{\text{SSIM}} \tag{3.2}$$

The SSIM index is a perceptual metric that compares luminance, contrast, and structural patterns between the predicted image x and the ground truth image y. It is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.3}$$

where:

- μ_x, μ_y are local means (via average pooling),
- σ_x^2, σ_y^2 are local variances,
- σ_{xy} is the local covariance,
- $C_1 = (0.01)^2$, $C_2 = (0.03)^2$ are stabilizing constants.

To optimize for similarity, the training loss uses a normalized version of SSIM:

$$\mathcal{L}_{\text{SSIM}} = \frac{1 - \text{SSIM}}{2} \tag{3.4}$$

By balancing $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{SSIM}}$, the model effectively reduces both pixel-wise discrepancies and perceptual inconsistencies in low-dose CT reconstruction.

Integration of attention mechanisms

While the U-Net architecture significantly enhanced reconstruction quality by preserving spatial features through skip connections, certain subtle and complex anatomical structures, especially in high-noise or low-contrast regions, remained challenging to reconstruct with high fidelity. To further boost the model's sensitivity to such critical features, an attention mechanism was introduced into the encoder-decoder framework.

The core idea behind attention is to dynamically focus the model's capacity on the most informative regions of the input image [18], rather than treating all spatial features equally. In the context of low-dose CT reconstruction, this helps prioritize diagnostically relevant areas, such as lung nodules or boundaries with subtle tissue contrast, while enhancing reconstruction quality without significantly increasing model complexity.

Attention layers were embedded directly into the encoder-decoder pipeline. These modules compute attention weights that emphasize salient features at each layer, enabling the network to adaptively modulate the influence of spatial information [24, 30]. By applying learned attention maps to intermediate feature representations, the model selectively amplifies meaningful structures and suppresses irrelevant noise. This dynamic reweighting is particularly beneficial in medical images, where clinically significant features are often small and localized [18, 23].

The attention mechanism was added without altering the U-Net's skip connections, thereby preserving its strength in retaining spatial detail while improving the network's focus on relevant regions. Although attention maps guide the model internally, visualizing them meaningfully in this medical imaging context is nontrivial. Therefore, attention visualizations are not shown, as they would not reliably convey interpretable or diagnostic patterns in this setup.

To further enhance structural fidelity, the loss function from the previous phase was maintained but reweighted to emphasize perceptual quality. Specifically, SSIM was given more importance, increasing its weight to 84%, while MAE was reduced to 16%. This combination was selected based on empirical testing. Various ratios were explored through trial-and-error to evaluate reconstruction quality, and the chosen values reflect the best balance between perceptual coherence and pixel-wise fidelity across the validation set:

$$\mathcal{L}_{\text{total}} = 0.16 \cdot \mathcal{L}_{\text{MAE}} + 0.84 \cdot \mathcal{L}_{\text{SSIM}} \tag{3.5}$$

This configuration improves the model's ability to preserve fine structural details, while maintaining realistic noise characteristics and sharper nodule boundaries.

Architecture of the simulation model

The simulation model is based on a U-Net-style encoder–decoder architecture augmented with attention modules to enhance feature relevance. Its primary objective is to transform a full-dose CT image into a realistic simulated low-dose version while preserving fine anatomical details and texture consistency.

The model begins with an encoder that processes the input full-dose CT image $X \in \mathbb{R}^{1\times512\times512}$ through two consecutive convolutional blocks. Each block reduces the spatial resolution and increases the depth of the learned features, producing a progressively more abstract representation. These transformations are performed using convolutional filters, which are sets of learnable weights that slide across the input image to detect specific patterns. The number of filters determines how many unique patterns (features) the model can extract at a given layer.

The first block applies a 4×4 convolutional kernel with stride 2 and padding 1, using 64 filters, resulting in a feature map of size 256×256 with 64 channels. Each of these filters acts as a detector for a specific low-level feature such as edges, gradients, or local texture transitions, elements critical for distinguishing between anatomical structures. The output of this operation can be viewed as 64 distinct views of the same image, each highlighting a different aspect of the input.

Following the convolution, a ReLU activation introduces non-linearity by mapping negative values to zero: $\text{ReLU}(z) = \max(0, z)$. Batch normalization is then used to stabilize learning by normalizing each activation across the batch:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \quad y_i = \gamma \hat{x}_i + \beta$$

where μ_B and σ_B^2 are the mini-batch mean and variance, and γ, β are learnable scaling parameters.

To further regularize learning, dropout with a probability of 0.3 randomly sets some activations to zero, helping the model avoid overfitting. This operation is expressed as $\tilde{h} = h \cdot r$, where r is a Bernoulli-distributed random mask.

The second encoder block repeats this structure but increases the depth to 128 filters, reducing the resolution to 128×128. As the number of filters increases, the model is able to learn more complex and abstract features such as tissue textures, organ shapes, and contextual anatomical patterns. Each additional filter increases the representational capacity of the model, allowing it to capture subtle differences in structure and intensity. By the end of the encoder stage, the input image has been transformed into a compact latent representation of shape $128 \times 128 \times 128$, often referred to as the code. This representation encodes the essential semantic content of the input.

To refine the learned features, attention mechanisms are inserted after each encoder block. Each attention module computes a learned attention map $A \in [0,1]^{C \times H \times W}$ and applies it to the feature map F using element-wise multiplication: $F_{\text{att}} = F \odot \sigma(A)$. Here, $\sigma(A)$ is the sigmoid-activated attention map, which assigns importance weights between 0 and 1 to modulate spatial and channel-level features. This operation helps the network focus on clinically relevant regions, such as nodule boundaries or fine vessel textures, while suppressing noise or background information.

The attention map is generated using a compact bottleneck structure: a 1×1 convolution reduces the number of channels from C to $C/8$, followed by a ReLU activation. Another 1×1 convolution restores the original channel depth, and a final sigmoid activation produces the attention weights. This structure allows the model to encode contextual dependencies efficiently with minimal overhead.

Once the features have been refined, the decoder reconstructs the simulated low-dose CT image from the latent space. It mirrors the encoder structure and employs transposed convolutional filters to upsample the spatial resolution. These filters work similarly to standard convolutions but in reverse: they project low-resolution features into a higher-resolution space.

For instance, a transposed convolution transforms the 128×128 feature map with 128 channels into a 256×256 map with 64 channels using 64 transposed filters. Each upsampling step is followed by ReLU activation and batch normalization, maintaining consistency with the encoder.

A key design component is the use of skip connections that concatenate encoder features with decoder outputs at corresponding resolution levels. For example, after the first upsampling, the decoder output is concatenated with the encoder's 256×256 feature map, forming a richer and more detailed representation. This fusion helps retain boundary sharpness and spatial fidelity, which are essential in medical image reconstruction.

The final decoder layer restores the spatial dimensions to 512×512 and reduces the channel depth to 1 using a single transposed convolutional filter, producing a single-channel grayscale image. A ReLU activation ensures the output remains non-negative, in line with the typical intensity distribution of CT images.

In summary, the model processes a full-dose CT image through an encoder to extract features, applies attention to enhance clinically relevant structures, and reconstructs a low-dose version using decoder blocks and skip connections. The architecture successfully balances global semantic understanding and local detail preservation, making it suitable for realistic simulation of dose reduction effects in diagnostic CT imaging.

An overview of the proposed architecture is illustrated in Figure 3.10, highlighting the encoder–decoder structure, attention modules, and skip connections.

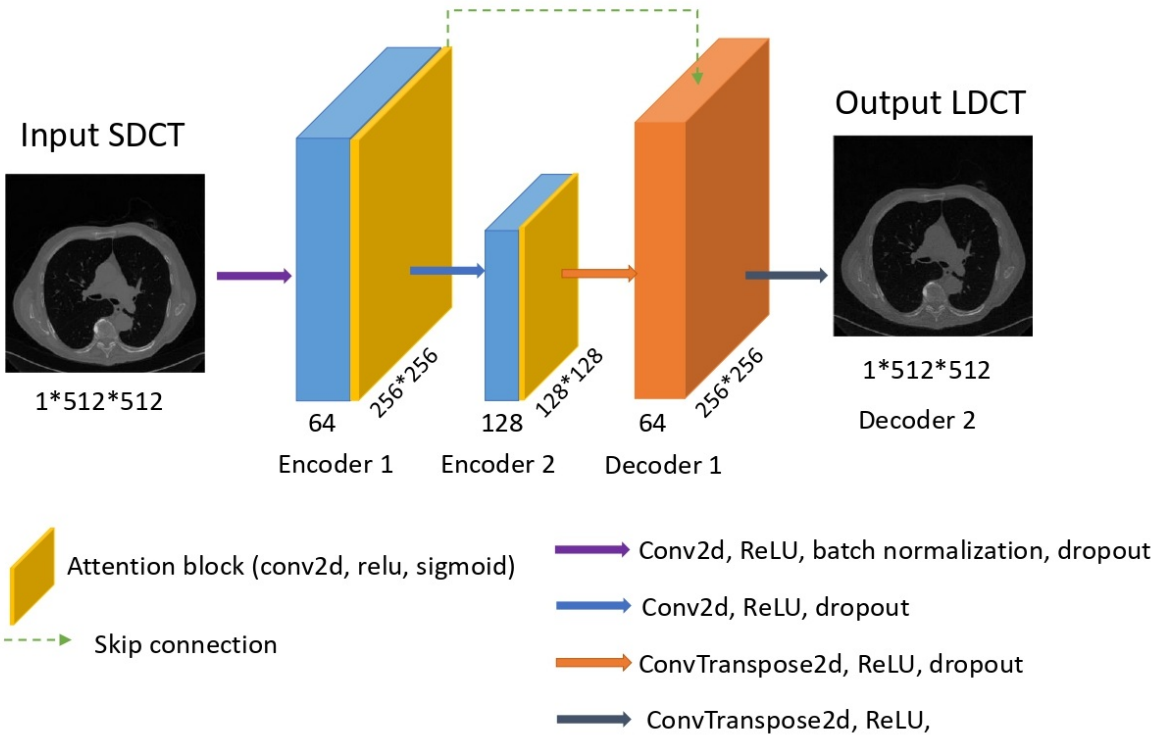

Figure 3.10: Architecture of the proposed U-Net-style autoencoder with attention modules for low-dose CT simulation.

The final trained model was applied to the LIDC-IDRI dataset to generate synthetic low-dose CT images from high-dose acquisitions. Based on the `XRayTubeCurrent` metadata, CT scans were categorized into FD and true LD subsets, using a threshold of 80 mA. The simulation model was applied exclusively to the FD cases to produce synthetic LDCT images that mimic realistic noise and intensity degradation.

In contrast, scans with `XRayTubeCurrent` values less than or equal to 80 mA were considered true low-dose and were excluded from simulation. These LD scans were preserved for downstream evaluation of lung nodule classification models in a real-world testing context, free from simulated LD.

This strategy resulted in a hybrid dataset consisting of (1) simulated LDCT images derived from high-dose scans and (2) real LDCT scans used as the independent test set. A total of 316 subjects with true low-dose data were identified and allocated for this testing purpose.

3.2. Nodule classification

3.2.1. Radiomic feature extraction

To characterize lung nodules for subsequent classification, radiomic feature extraction was performed on low-dose CT volumes using consensus segmentation masks. Radiomics offers a quantitative framework that transforms medical images into a high-dimensional feature space, capturing patterns beyond the limits of visual inspection. As described by Shur et al. [26], these features encompass intensity-based statistics, shape descriptors, texture measurements, and wavelet transformations, enabling rich, multiscale representations of the image data.

Preprocessing and ROI construction: DICOM image volumes from the LIDC-IDRI dataset were loaded and stacked into 3D arrays, sorted by slice location. Using the PyLIDC interface, nodule annotations were retrieved, and consensus segmentation masks were generated via the `consensus()` function [20]. These consensus masks defined the 3D regions of interest (ROIs) used for all subsequent feature extraction.

Feature categories: For each ROI, features were extracted using the `PyRadiomics` library across the following categories:

First-order statistics quantify the distribution of voxel intensities, independent of spatial relationships. Key examples include:

- **Energy:**
$$\text{Energy} = \sum_{i=1}^{N_p} (X(i) + c)^2$$
where $X(i)$ represents voxel intensities and c is an optional offset (e.g., from `voxelArrayShift`). Higher energy reflects stronger overall signal magnitudes.

- **Total energy:**
$$\text{Total Energy} = V_{\text{voxel}} \sum_{i=1}^{N_p} (X(i) + c)^2$$
which accounts for both intensity and ROI volume V_{voxel}.

- **Entropy:**
$$\text{Entropy} = -\sum_{i=1}^{N_g} p(i) \log_2 (p(i) + \epsilon)$$
where $p(i)$ is the probability of gray level i and $\epsilon \approx 2.2 \times 10^{-16}$. Entropy captures

the degree of intensity heterogeneity.

Shape features describe the geometry of the segmented nodules, independent of voxel intensity. These are robust to noise and acquisition variability [20]. Examples include:

- **Surface area:**
$$A_i = \frac{1}{2}\left|a_i\vec{b}_i \times a_i\vec{c}_i\right| \quad , \quad A = \sum_{i=1}^{N_f} A_i$$
computed from triangle mesh faces on the surface of the ROI.

- **Sphericity:**
$$\text{Sphericity} = \frac{\sqrt[3]{36\pi V^2}}{A}$$
where V is the volume and A the surface area. Values close to 1 suggest a round, compact shape.

- **Major axis length:**
$$\text{Major Axis Length} = 4\sqrt{\lambda_{\text{major}}}$$
derived from PCA on the 3D voxel coordinates, where λ_{major} is the largest eigenvalue.

Texture features capture spatial relationships between voxels using gray-level matrices such as GLCM and GLRLM. These features quantify intra-nodular heterogeneity and texture granularity:

- **Autocorrelation:**
$$\text{Autocorrelation} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\cdot i \cdot j$$
where $p(i,j)$ is the normalized GLCM value at gray levels i and j.

- **Joint average:**
$$\mu_x = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\cdot i$$
reflecting average co-occurrence intensity levels.

Wavelet-transformed features were also extracted by decomposing each ROI into multiple frequency sub-bands. This highlights texture patterns at various spatial scales and orientations.

Before texture analysis, intensity values were normalized and discretized using a fixed bin

width approach, following recommendations from the PyRadiomics documentation. This step was essential to ensure consistency and reproducibility across CT scans that might differ in acquisition settings or intensity scaling.

One of the main advantages of radiomics lies in its interpretability. Each extracted feature, such as sphericity, skewness, or entropy, has a clear and quantitative definition that can be linked to morphological or textural characteristics of the nodule. As shown by Peikert et al. [20], radiomic signatures have demonstrated predictive power for malignancy risk and can offer interpretable biomarkers that complement clinical decision-making.

Nevertheless, radiomic features are not without limitations. Their stability and reliability can be compromised by variations in imaging protocols, segmentation accuracy, and preprocessing pipelines. As highlighted by Shur et al. [26], achieving reproducibility requires careful standardization and external validation, particularly when deploying models across diverse datasets or clinical environments.

3.2.2. Feature preprocessing and selection

Following radiomic feature extraction from the annotated nodule regions in the LIDC-IDRI dataset, a multi-stage feature selection pipeline was implemented to refine the feature space and improve the robustness of the classification task. The objective was to retain features that were reproducible across annotators, sensitive to morphological differences, and minimally redundant.

As illustrated in Figure 3.11, the selection process involved three primary stages:

1. Stability analysis using inter-class correlation coefficients (ICC),
2. Discriminative filtering based on response to geometric perturbations,
3. Redundancy reduction via Spearman's rank correlation.

1. Stability analysis (inter-class correlation coefficient): To evaluate reproducibility across different annotators, the inter-class correlation coefficient (ICC) was computed for each feature. A two-way mixed-effects model for absolute agreement, as described by Koo and Li [12], was employed:

$$\mathrm{ICC}(A,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E} \tag{3.6}$$

Here, MS_R is the mean square for subjects (nodules), MS_E is the residual error, and

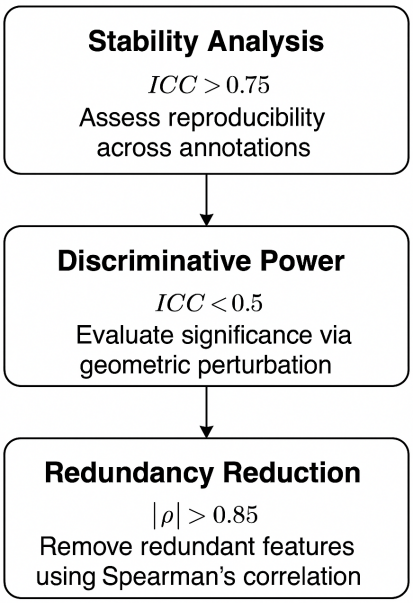

Figure 3.11: Three-stage radiomic feature selection pipeline.

k is the number of raters. Features with ICC scores greater than 0.75 were retained, ensuring only highly reproducible features were used in downstream analysis. This step was conducted on a subset of 705 nodules, each annotated by at least four independent readers.

2. Discriminative analysis under geometric perturbation: To ensure selected features reflected morphological characteristics rather than positional artifacts, each ROI was translated by 150% in spatial coordinates, following the method in Liu et al. [14]. Features were recalculated after transformation, and ICC was recomputed between the original and translated feature sets. Features with ICC values less than 0.5 were considered discriminative, indicating that they were sensitive to shape or internal structure and not invariant under translation.

3. Redundancy reduction (Spearman's correlation): To eliminate redundant features and reduce collinearity, pairwise Spearman's rank correlation coefficients were calculated:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.7}$$

where d_i is the rank difference between two features for the i^{th} sample, and n is the total number of samples. For feature pairs where $|\rho| > 0.85$, the feature with the lower average correlation to all other features was retained [14].

This structured selection pipeline reduced the feature space to a compact and informative subset, enhancing the interpretability, reproducibility, and performance of the classification model built in subsequent stages.

3.2.3. Data augmentation

A major challenge in this study was the strong imbalance between benign and malignant nodules. Among the consensus-labeled training and validation data, there were 1,725 benign cases and only 225 malignant ones, resulting in a roughly 8:1 ratio. This imbalance posed the risk of biasing the classifiers toward benign predictions, reducing their ability to correctly identify malignant nodules, an especially critical limitation in early cancer detection.

To mitigate this issue, a data augmentation strategy was applied exclusively to the malignant class in order to synthetically expand its size while maintaining anatomical plausibility. All augmentations were performed on nodules with available consensus segmentations, ensuring that the starting point for each transformation was based on radiologist-reviewed annotations.

The augmentation process included a series of morphological transformations. Specifically, two basic shape operations, dilation and erosion, were applied to simulate variability in segmentation styles. Dilation expanded the mask area to mimic potential over-segmentation, while erosion reduced it to reflect more conservative annotations. These modifications introduced realistic variations in nodule contours, resembling the differences commonly seen between radiologist annotations.

In addition to shape changes, localized deformations were introduced using SLIC superpixels. Each mask was divided into superpixels, and in selected regions, the nodule boundary was subtly altered by modifying pixel values inside the superpixel. This created slightly different shapes that still resembled plausible nodules. The process was repeated twice per nodule to increase diversity [7, 9].

Lastly, a small amount of salt-and-pepper noise was added to the masks, particularly near edges, to simulate minor imperfections or annotation uncertainty. Although this is not a conventional image-level augmentation, it helped represent the ambiguity often present in real-world clinical segmentations. Gaussian blur was not ultimately included in the final

pipeline.

Each malignant nodule was augmented to produce exactly five new variants, resulting in a sixfold increase in the size of the malignant set. This number was selected based on practical considerations of training balance and computational load. All metadata from the original annotations (e.g., slice location and patient ID) was preserved. The generated masks were visually inspected for quality and anatomical plausibility. While no formal clinical re-validation was conducted, the augmentations were conservative and grounded in standard image processing techniques used in medical image analysis.

Through this targeted approach, the malignant class was increased from 225 to 1,115 samples, substantially improving the balance between classes and supporting more robust classifier training.

3.2.4. Data Splitting

To ensure a reliable evaluation protocol and prevent data leakage, the LIDC-IDRI dataset was systematically partitioned into separate subsets for training, validation, and testing.

Nodules extracted from CT scans acquired under low-dose conditions (defined by XRayTubeCurrent $\leq$ 80 mA) were reserved exclusively for the **test set**. These cases represent authentic low-dose clinical scenarios and were entirely excluded from training and feature selection stages.

The remaining nodules, originating from standard or higher-dose scans, were split subject-wise into two subsets: 80% of the subjects were used for training and cross-validation, while the remaining 20% were assigned to the validation set to monitor model performance during development.

3.2.5. Model development

To classify lung nodules as benign or malignant based on their radiomic characteristics, a diverse set of supervised machine learning classifiers was employed. The models included Random Forest , AdaBoost, XGBoost, Decision Tree, K-Nearest Neighbors, Logistic Regression, and a Multilayer Perceptron. These classifiers were selected to capture a range of learning paradigms and model complexities, from linear models to ensemble methods and neural networks.

All models were implemented using the `scikit-learn` framework and trained using the radiomic features extracted and selected from the LIDC-IDRI dataset.

3.2.6. Performance evaluation setup

To rigorously evaluate classifier performance, a bootstrapped experimental setup involving multiple resampling iterations and stratified data splits was implemented. This approach enhances statistical reliability and provides confidence intervals for each performance metric.

A total of 20 bootstrap iterations were performed. In each iteration, the training set was resampled with replacement to form a new training subset, while the validation and test sets remained fixed. Classifiers were trained on the resampled set and evaluated on all three data partitions: training, validation, and test.

Importantly, this evaluation protocol was applied to both the original imbalanced dataset (prior to augmentation) and the rebalanced dataset (after malignant class augmentation). This allowed a fair comparative analysis of classification performance under both realistic and corrected class distribution conditions.

Model performance was quantified using standard metrics derived from the confusion matrix:

TP True Positives (malignant predicted as malignant)

TN True Negatives (benign predicted as benign)

FP False Positives (benign predicted as malignant)

FN False Negatives (malignant predicted as benign)

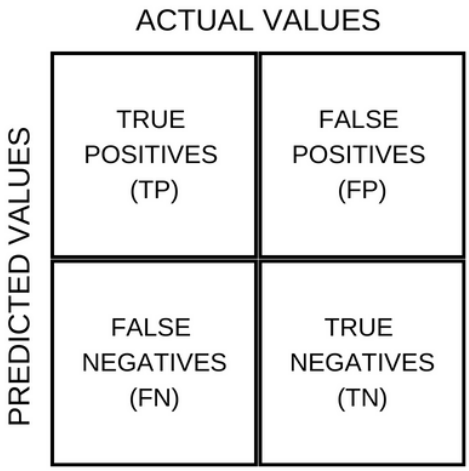

Figure 3.12: Visual representation of the confusion matrix used for classification performance evaluation.

[0] https://udaykiran.tech/confusion-matrix-an-easy-way-to-remember-and-use

Figure 3.12 reinforces the definitions provided above by visually organizing the prediction outcomes. This standard format is crucial for computing metrics such as accuracy, sensitivity, specificity, and F1-score.

- Balanced Accuracy (Bal_Acc): Mean of sensitivity and specificity, correcting for class imbalance:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{3.8}$$

- Sensitivity (Recall): Measures the proportion of actual malignant nodules correctly identified:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{3.9}$$

- Specificity: Measures the proportion of actual benign nodules correctly identified:

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3.10}$$

- F1-score: Harmonic mean of precision and recall:

$$\text{F1-score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3.11}$$

- ROC AUC: Area under the Receiver Operating Characteristic curve. While not computed directly from the confusion matrix, it reflects the trade-off between sensitivity and 1-specificity across thresholds:

$$\text{AUC} = \int_0^1 TPR(FPR^{-1}(x))\,dx \tag{3.12}$$

  where $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$.

To convert predicted probabilities into binary predictions, thresholds were optimized per iteration using Youden's J statistic, which maximizes the sum of sensitivity and specificity:

$$J = \text{Sensitivity} + \text{Specificity} - 1 \tag{3.13}$$

For each classifier, average metric values over all bootstrap iterations were visualized using annotated heatmaps, enabling comparative performance analysis across training, validation, and test sets.

All experiments were implemented in Python using the following packages: `scikit-learn` (for model training, metric calculation, and ROC analysis), `imbalanced-learn` (for resampling utilities), and `Seaborn/Matplotlib` (for heatmap visualizations).

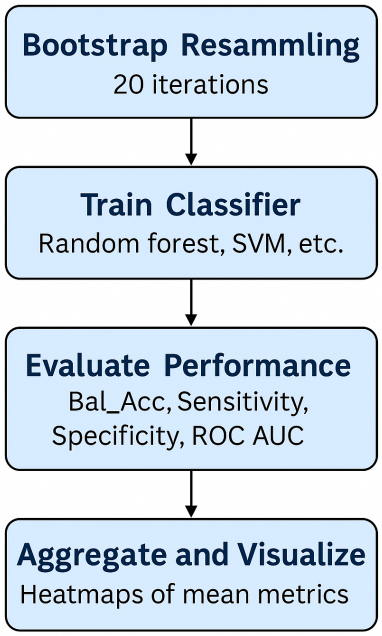

Figure 3.13: Overview of the experimental workflow including bootstrapped training, validation, and test evaluation using radiomic features and multiple classifiers.

The overall evaluation pipeline is summarized in Figure 3.13.

4 | Results and conclusion

4.1. Low-Dose CT simulation results

This section presents quantitative and qualitative results obtained from the three progressive architectures used for low-dose CT simulation: (1) a simple AE, (2) an enhanced U-Net variant with skip connections, and (3) an attention-integrated model. Performance was measured using Peak Signal-to-Noise Ratio (PSNR) and SSIM. For clarity and brevity, the abbreviations **HD** (high-dose) and **LD** (low-dose) are consistently used throughout this section, including in figure captions.

4.1.1. Simple AE performance

The simple AE achieved an average **PSNR of 23.5 dB** and an **SSIM of 0.723** when translating HD CT slices into simulated LD outputs. While the AE successfully learned a basic mapping between HD and LD domains, its reconstructed outputs revealed several limitations. The model often failed to preserve fine structural details, such as small vessels and subtle lesions, which are crucial for clinical interpretation. Furthermore, the simulated LD slices displayed inconsistent and unrealistic noise distributions that deviated from patterns typically observed in real LD scans. A tendency toward over-smoothing was also evident, resulting in a loss of textural cues that are essential for detecting abnormal tissue structures.

Figures 4.1–4.3 demonstrate representative outputs from the simple AE. In each case, the generated LD slices exhibit the same characteristic issues: reduced structural fidelity, artifacts in the noise pattern, and excessive smoothing. The red circles in the images highlight areas where the generated low-dose outputs failed to replicate the true LD appearance and instead resemble high-dose scans, indicating insufficient noise modeling and structural degradation. These findings support the quantitative metrics and underscore the need for improved modeling strategies, as explored in subsequent sections.

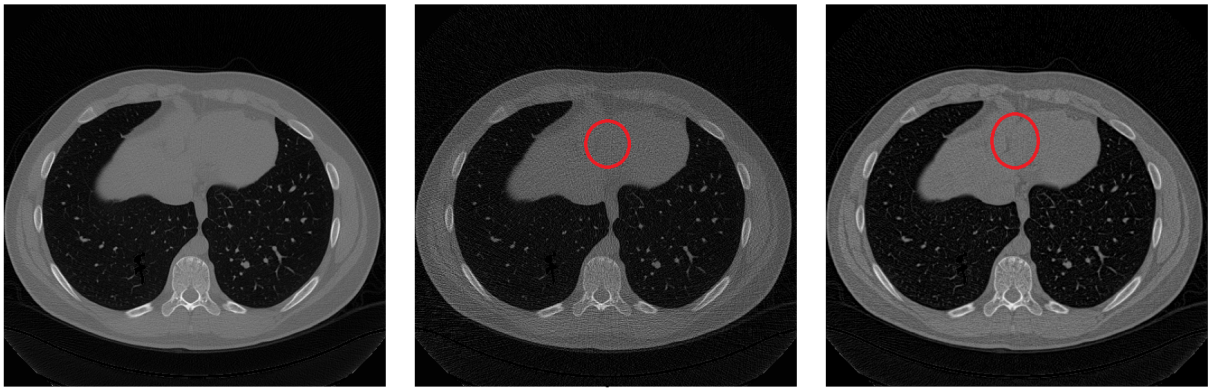

Figure 4.1: Reconstruction result from simple AE (sample 1). Left: True HD slice. Middle: True LD slice. Right: Generated LD slice by the simple AE.

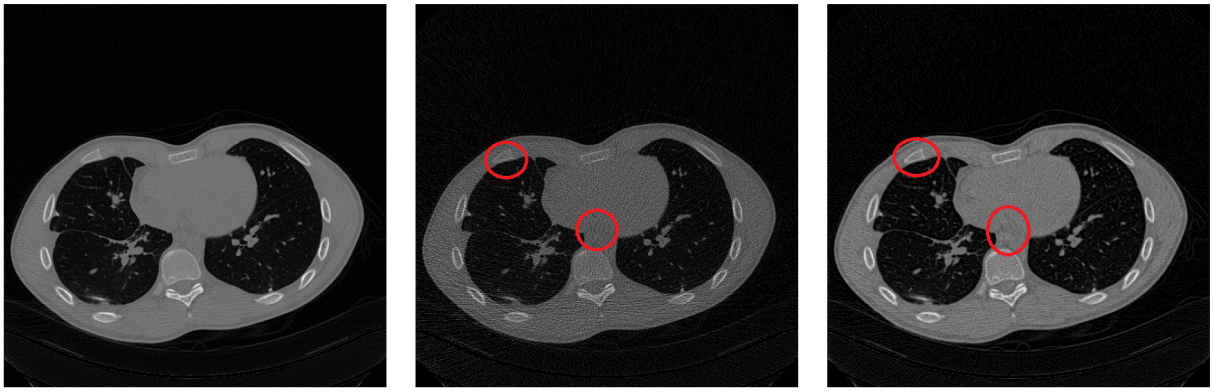

Figure 4.2: Reconstruction result from simple AE (sample 2). Left: True HD slice. Middle: True LD slice. Right: Generated LD slice by the simple AE.

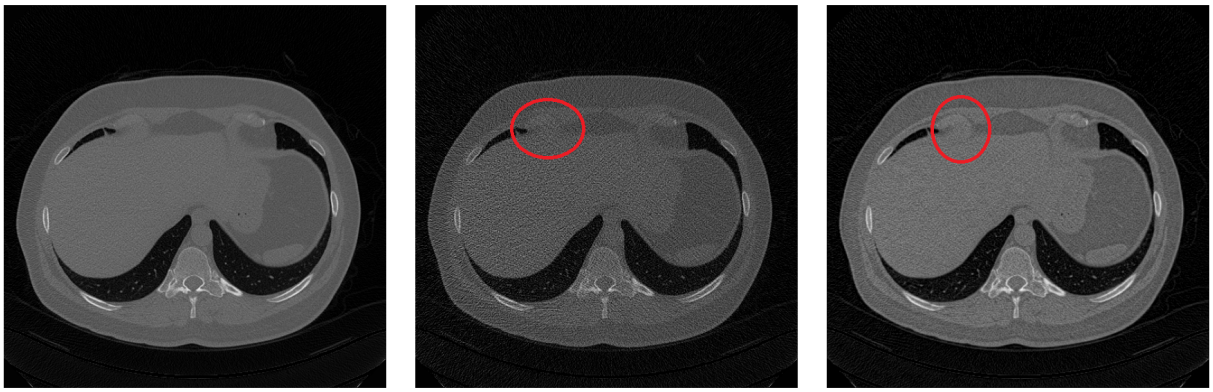

Figure 4.3: Reconstruction result from simple AE (sample 3). Left: True HD slice. Middle: True LD slice. Right: Generated LD slice by the simple AE.

4.1.2. U-Net enhanced architecture results

The U-Net-based AE achieved an average **PSNR of 27.84 dB** and an **SSIM of 0.8881**, marking a clear improvement over the simple model. The incorporation of skip connections led to substantial gains in reconstruction quality. Anatomical coherence was more effectively preserved, while reconstruction artifacts were minimized and critical structural details were better maintained across the output.

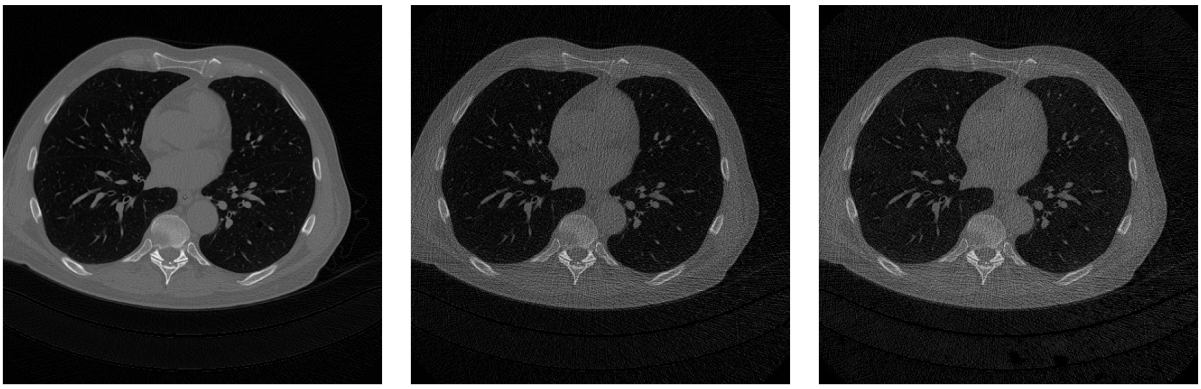

Figure 4.4: U-Net AE with skip connections (sample 1). Left: True HD slice. Middle: True LD slice. Right: Generated LD slice by the U-Net AE.

Figure 4.5: U-Net AE with skip connections (sample 2). Left: True HD slice. Middle: True LD slice. Right: Generated LD slice by the U-Net AE.

Figures 4.4 and 4.5 exemplify the improvements achieved by the U-Net enhanced architecture. In both examples, the generated LD slices exhibit clearer anatomical boundaries, fewer artifacts, and more realistic textures, highlighting the role of skip connections in preserving spatial structure and enhancing reconstruction fidelity.

4.1.3. Attention-integrated model results

The attention-integrated AE achieved the highest performance across all tested architectures, with an average **PSNR of 32.02 dB** and an **SSIM of 0.9311**. These results reflect strong reconstruction quality both numerically and visually. The model effectively preserved subtle anatomical features such as sub-millimeter vessels, maintained spatial noise patterns consistent with real low-dose acquisitions, and produced well-defined nodule boundaries. These qualities are essential for reliable clinical interpretation and downstream radiomic analysis.

Figure 4.6: Attention-enhanced reconstruction (sample 1). Left: True HD slice. Middle: True LD slice. Right: Generated LD slice by the attention-integrated model.

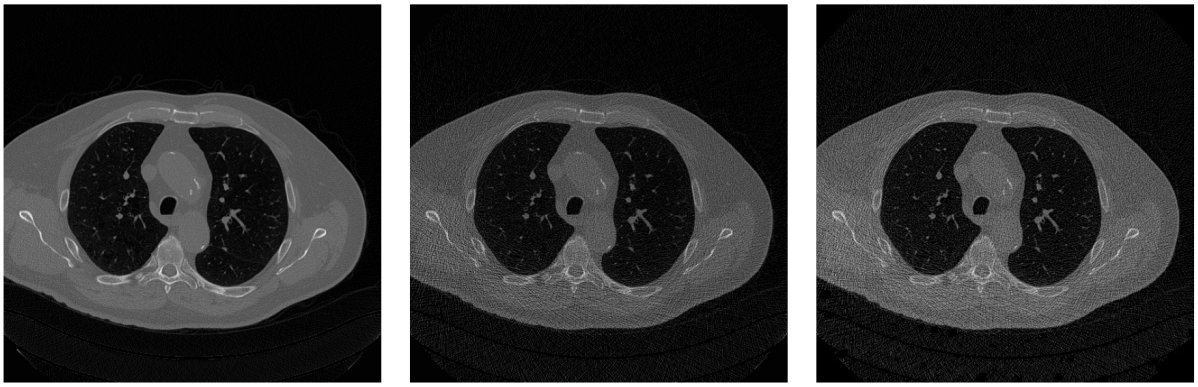

Figure 4.7: Attention-enhanced reconstruction (sample 2). Left: True HD slice. Middle: True LD slice. Right: Generated LD slice by the attention-integrated model.

Figures 4.6 and 4.7 demonstrate the ability of the attention-augmented architecture to generate high-fidelity low-dose reconstructions. The outputs consistently capture diagnostic detail, realistic noise structure, and spatial consistency which validating its applicability for clinical simulation and analysis.

4.1.4. Architecture performance comparison and results

Table 4.1 summarizes the quantitative improvements achieved at each stage of architectural enhancement.

Architecture	PSNR (dB)	SSIM
Simple AE	23.50	0.723
+ Skip Connections	27.84	0.8881
+ Attention Mechanism	**32.02**	**0.9311**

Table 4.1: Performance comparison across different architectures.

Key improvements:

- **18.5% PSNR increase** from simple AE to U-Net
- **15.0% additional PSNR gain** with attention mechanism
- **28.8% total SSIM improvement** from simple AE to final model

Overall, the progressive architectural refinements, from a simple AE to a U-Net with skip connections and finally to an attention-integrated model, resulted in substantial gains in image reconstruction quality. These improvements are clearly reflected in both objective metrics (PSNR and SSIM) and qualitative visual assessments. The final architecture effectively balances noise realism, structural preservation, and diagnostic clarity, establishing a robust foundation for downstream tasks such as nodule classification on simulated low-dose CT data.

While traditional LDCT simulation approaches such as those proposed by Zeng et al. [33] and Yu et al. [32] offer strong physical realism by injecting Poisson and Gaussian noise into sinogram data, their evaluation is generally limited to visual assessments or physics-specific measures like noise power spectrum (NPS). These methods replicate scanner behavior well but often do not report standardized perceptual metrics such as PSNR or SSIM, making quantitative benchmarking difficult. Furthermore, their reliance on raw sinogram data restricts applicability, as such data are rarely available in public clinical datasets.

In contrast, the proposed attention-integrated autoencoder achieved 30.02 dB PSNR and 0.9211 SSIM, reflecting strong performance on both pixel fidelity and structural preservation. Compared to image-domain strategies like Kim and Kim [11], who focused on

noise decoupling via TV-denoising and synthetic sinogram generation, the deep learning approach demonstrated superior adaptability and generalization across patient anatomies without requiring scanner-specific parameters or manual calibration. These improvements highlight the benefit of a data-driven pipeline that learns complex noise and texture transformations directly from paired HD-LD images while preserving diagnostic clarity.

In addition to PSNR and SSIM, the reconstructed images were also evaluated based on their noise power spectrum (NPS) characteristics. The proposed method achieved a normalized mean absolute difference (MAD) of **0.0672** between the radial NPS curves of the reconstructed and real low-dose CT images. This result demonstrates a strong alignment with the noise texture characteristics of the target domain. Compared to existing literature, the NPS error lies within the range reported by Naziroglu et al. [16], who observed normalized NPS RMSE values between 0.05 and 0.15. It is also consistent with NRsim-based validation results, where typical mean absolute NPS differences range from 0.05 to 0.2. Although slightly higher than values reported by Takenaga et al. [29] (0.01–0.05 in phantom-based settings), the error remains acceptably low, especially considering the complexity and variability of clinical datasets. Since these prior studies did not report SSIM or PSNR, direct metric-level comparison is limited; however, the NPS result of 0.0672 provides strong evidence that the reconstructed images preserve realistic frequency-dependent noise patterns.

4.1.5. Application to LIDC-IDRI dataset

This section presents the results of the trained attention-enhanced low-dose CT simulation model when applied to the LIDC-IDRI dataset. The simulation aimed to generate synthetic low-dose (LD) CT slices from high-dose (HD) scans, enabling downstream radiomics and classification analysis under realistic noise conditions.

Figures 4.8 and 4.9 show examples of the simulated LD outputs. These slices demonstrate realistic noise characteristics while preserving essential diagnostic information such as soft tissue textures and lesion boundaries.

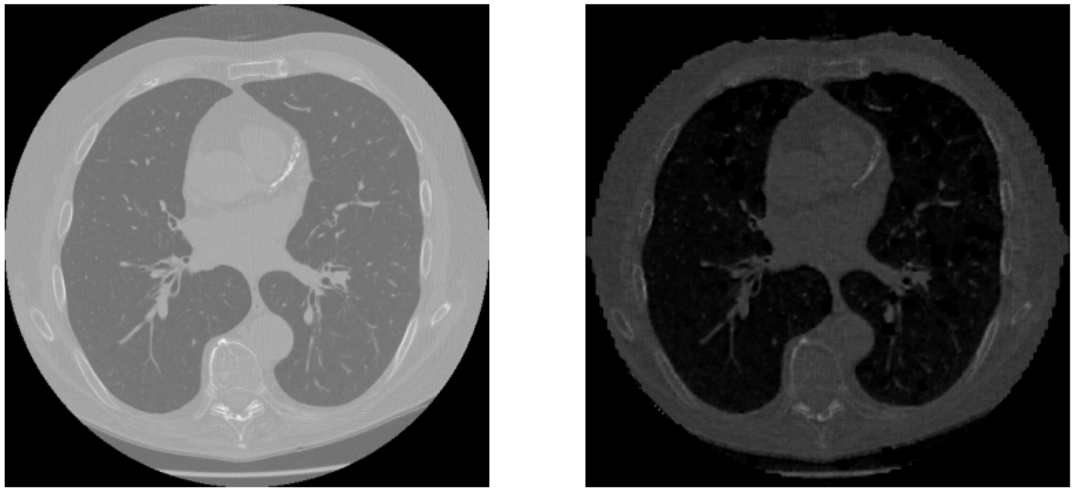

Figure 4.8: Simulated LD output from a LIDC-IDRI subject. Left: True HD slice. Right: Simulated LD version.

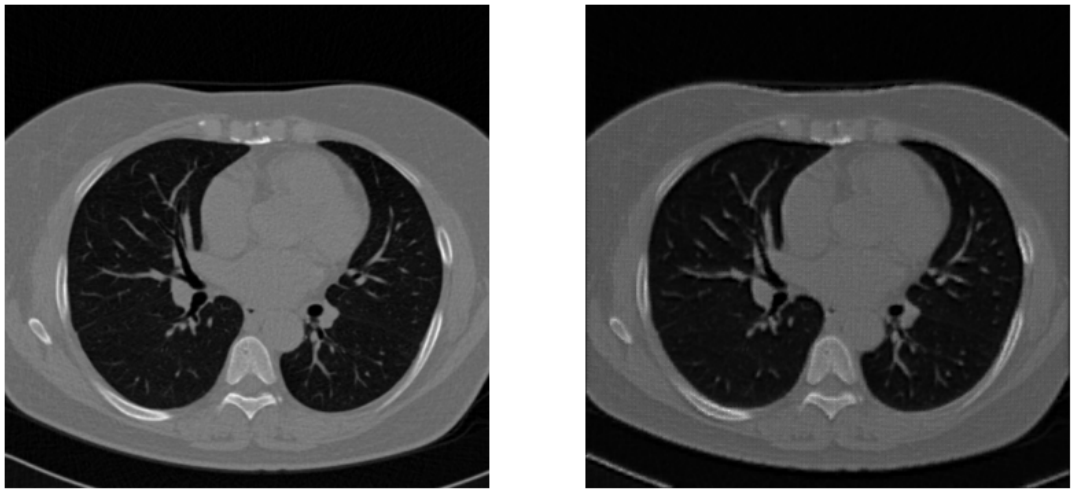

Figure 4.9: Another example of synthetic LD generation. Left: Original HD slice. Right: Simulated LD version.

As previously described, a portion of the dataset was reserved for classification and excluded from simulation. The remaining HD scans were processed using the trained model to generate synthetic LD equivalents, allowing the evaluation of its performance in clinically diverse patient cases with varying lung structures and nodule characteristics.

4.2. Feature selection results

The full pipeline described in the methodology was applied to 851 extracted radiomic features. Quantitative results of the sequential filtering steps are summarized in Table 4.2.

Table 4.2: Feature reduction pipeline results

Selection Stage	Number of Features
Initial feature set	851
After stability analysis (ICC $>$ 0.75)	656
After discriminative analysis (ICC $<$ 0.5)	387
Stable and discriminative intersection	268
Final selected features (after correlation filtering)	79

To support comparative evaluation in the classification experiments, the final 79 features were further organized into three functional groups (Table 4.3).

Table 4.3: Final feature groups for classification

Group	Description	Count
All features	Complete set of selected features	79
Shape and size only	Features describing morphological characteristics	3
Non-shape features	All features excluding shape/size descriptors	76

This systematic reduction and grouping ensured that subsequent classification experiments could distinguish between the contributions of shape-based and intensity/texture-based descriptors while reducing overfitting risk.

4.3. Nodule classification results

4.3.1. Performance on imbalanced dataset

Performance analysis of shape and size features

The evaluation using only three fundamental shape and size features, *MajorAxisLength*, *Sphericity*, and *Elongation*, revealed distinct performance patterns across all classifiers. Random Forest showed the strongest training performance with a balanced accuracy of 93.3%, while Logistic Regression emerged as the most stable performer across all datasets, achieving the highest test set accuracy at 81.3% (see Table 4.4 and Table 4.6).

Table 4.4: Training set performance (shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.933	0.921	0.945	**0.979**	0.785
AdaBoost	0.850	0.848	0.852	0.915	0.567
XGBoost	**0.940**	**0.938**	0.942	0.975	0.786
Decision Tree	0.906	0.837	**0.975**	0.906	**0.825**
K-Nearest Neighbors	0.877	0.916	0.838	0.905	0.578
Logistic Regression	0.850	0.876	0.824	0.908	0.542
MLP Neural Network	0.833	0.825	0.840	0.896	0.539

Table 4.5: Validation set performance (shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.887	**0.933**	0.840	0.931	0.592
AdaBoost	0.896	0.911	0.881	0.928	0.646
XGBoost	0.875	0.889	0.860	0.932	0.602
Decision Tree	0.726	0.533	**0.919**	0.726	0.495
K-Nearest Neighbors	0.849	0.889	0.808	0.877	0.530
Logistic Regression	0.907	**0.933**	0.881	**0.949**	0.656
MLP Neural Network	**0.913**	**0.933**	0.892	0.945	**0.677**

Across the classifiers, Logistic Regression and the MLP Neural Network showed the most consistent behavior, with less than 5% variation between validation and test set performance (Table 4.5 and Table 4.6). In contrast, the Decision Tree classifier exhibited

Table 4.6: Test set performance (shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.757	0.714	0.799	0.795	0.413
AdaBoost	0.793	0.714	**0.871**	0.823	**0.505**
XGBoost	0.768	0.700	0.837	0.811	0.450
Decision Tree	0.647	0.343	0.951	0.647	0.387
K-Nearest Neighbors	0.727	0.600	0.853	0.747	0.418
Logistic Regression	**0.813**	0.786	0.840	**0.851**	0.495
MLP Neural Network	0.804	**0.814**	0.794	0.848	0.452

notable overfitting, with a balanced accuracy drop of over 28% from training to test. Specificity remained high for all models, above 79%, but sensitivity varied widely, ranging from just 34.3% (Decision Tree) to 81.4% (MLP Neural Network), indicating a challenge in consistently detecting malignant cases.

XGBoost displayed signs of early overfitting, as evidenced by a 6.5% drop in balanced accuracy from training to validation (Table 4.4 to Table 4.5). Despite strong training scores, this gap suggests a reduced generalization capacity. Another key finding was that the F1-score remained below 0.51 for all models, reaffirming that class imbalance continues to affect classification effectiveness, particularly for the minority malignant class.

When comparing modeling strategies, shape-based features yielded better performance with linear models like Logistic Regression compared to other feature subsets discussed elsewhere in the study. Neural networks also demonstrated robustness despite the limited input features, while tree-based methods showed more fluctuation across data splits. Overall, these three shape features proved particularly useful in ruling out malignancies, as reflected in their high specificity values across models.

Performance analysis of non-shape features

The evaluation using 76 non-shape features revealed notable performance differences across classifiers. As shown in Table 4.7, Random Forest achieved the highest training accuracy with a balanced accuracy of 94.1%, while XGBoost yielded the best test set performance with a balanced accuracy of 77.5% (Table 4.9). However, both Logistic Regression and the MLP Neural Network exhibited consistently poor performance across the training, validation, and test sets, indicating a limited ability to extract discriminative patterns from non-shape features.

Table 4.7: Training set performance (non-shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.941**	**0.944**	0.939	**0.987**	0.781
AdaBoost	0.890	0.921	0.859	0.942	0.612
XGBoost	0.938	0.916	**0.960**	0.982	**0.823**
Decision Tree	0.911	0.854	0.968	0.911	0.813
K-Nearest Neighbors	0.878	0.927	0.829	0.910	0.571
Logistic Regression	0.617	0.612	0.621	0.664	0.269
MLP Neural Network	0.573	0.404	0.741	0.537	0.237

Table 4.8: validation set performance (Non-shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.913**	**0.933**	0.892	0.954	0.677
AdaBoost	0.906	0.911	0.901	**0.955**	**0.683**
XGBoost	0.905	0.911	0.898	0.951	0.678
Decision Tree	0.746	0.556	**0.936**	0.746	0.543
K-Nearest Neighbors	0.820	0.844	0.797	0.861	0.497
Logistic Regression	0.728	0.644	0.811	0.725	0.417
MLP Neural Network	0.509	0.311	0.706	0.476	0.175

Table 4.9: Test set performance (non-shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.750	0.700	0.799	0.787	0.407
AdaBoost	0.726	**0.743**	0.710	0.797	0.349
XGBoost	**0.775**	0.700	**0.850**	**0.802**	**0.467**
Decision Tree	0.593	0.329	0.858	0.593	0.257
K-Nearest Neighbors	0.720	0.586	0.855	0.735	0.412
Logistic Regression	0.635	0.557	0.713	0.684	0.276
MLP Neural Network	0.650	0.586	0.715	0.715	0.289

Among the tested models, XGBoost demonstrated the most consistent performance with the smallest drop from training to test, only a 16.3% decrease in balanced accuracy, and also achieved the highest test accuracy of 77.5%. In contrast, Decision Tree experienced the most substantial degradation, with a 36.5% drop in balanced accuracy between training (Table 4.7) and validation (Table 4.8), indicating a strong tendency to overfit despite good training metrics. Specificity remained relatively high across all classifiers, generally exceeding 70%, whereas sensitivity ranged much more widely from 32.9% to 74.3%.

Notably, none of the classifiers exceeded an F1-score of 0.47 on the test set (Table 4.9), underscoring persistent difficulties associated with class imbalance. Logistic Regression and MLP Neural Network performed especially poorly, failing to generalize effectively and producing the lowest scores across training, validation, and test phases.

When compared to the shape-based feature subset, models trained on non-shape features underperformed by an average of 3.1% in test balanced accuracy. Nevertheless, ensemble tree methods such as Random Forest and XGBoost still demonstrated better generalization with this feature set than linear or neural classifiers. These results suggest that, while non-shape features contain useful signal, their complexity may require more robust or feature-aware architectures to extract discriminative patterns effectively.

Performance analysis of all features

The evaluation using all 79 features, encompassing both shape and non-shape descriptors, revealed improved overall performance relative to using each feature group individually. As shown in Table 4.10, Random Forest and XGBoost stood out during training with balanced accuracies of 94.3% and 94.1% respectively. These models also maintained relatively stable validation and test performance, as reflected in Tables 4.11 and 4.12. In contrast, Logistic Regression and MLP Neural Network struggled to exploit the full feature set, especially with respect to specificity and F1-score.

From a generalization standpoint, the average balanced accuracy drop between training and testing was 15.8%, representing an improvement over the 27.5% reduction observed when only non-shape features were used. Both Random Forest and XGBoost maintained robust classification capabilities on the test set, achieving ROC AUC values above 0.79. In contrast, Logistic Regression continued to suffer from low specificity (49.1%) and a test F1-score of just 0.237, signaling a substantial false positive rate.

Incorporating both feature types improved test balanced accuracy by 4.2% over shape-only models, reinforcing the notion that combining morphologic and texture-based descriptors enhances discriminative power. Nevertheless, F1-scores remained consistently below 0.42

Table 4.10: Training set performance (all features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.943**	0.921	**0.964**	0.984	**0.837**
AdaBoost	0.903	**0.933**	0.874	0.955	0.641
XGBoost	0.941	0.949	0.932	**0.986**	0.766
Decision Tree	0.915	0.860	0.970	0.915	0.823
K-Nearest Neighbors	0.880	0.933	0.827	0.911	0.570
Logistic Regression	0.676	0.691	0.660	0.710	0.320
MLP Neural Network	0.662	0.511	0.813	0.639	0.345

Table 4.11: Validation set performance (all features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.905	0.911	0.898	0.959	0.678
AdaBoost	**0.915**	**0.956**	0.875	**0.963**	0.656
XGBoost	0.904	0.889	**0.919**	0.951	**0.708**
Decision Tree	0.810	0.689	0.930	0.810	0.620
K-Nearest Neighbors	0.819	0.844	0.794	0.855	0.494
Logistic Regression	0.750	0.800	0.701	0.807	0.391
MLP Neural Network	0.694	0.533	0.855	0.693	0.403

Table 4.12: Test set performance (all features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.764**	**0.800**	0.728	**0.816**	0.385
AdaBoost	0.760	0.729	**0.791**	0.823	**0.411**
XGBoost	0.762	0.771	0.753	0.797	0.394
Decision Tree	0.659	0.457	0.862	0.659	0.344
K-Nearest Neighbors	0.632	0.414	0.850	0.640	0.305
Logistic Regression	0.610	0.729	0.491	0.629	0.237
MLP Neural Network	0.598	0.571	0.624	0.637	0.237

for all classifiers, underscoring the continued effect of class imbalance on performance.

When benchmarked against non-shape feature models, the full-feature set yielded a 7.3% improvement in average test accuracy and delivered more stable sensitivity scores, averaging 57.8% versus 50.3%. ROC AUC values were also better maintained, and tree-based models appeared less susceptible to overfitting in this setting than in previous configurations.

Altogether, these findings suggest that leveraging all 79 features enhances classification robustness and generalization capacity. However, persistent limitations, including low F1-scores and pronounced trade-offs between sensitivity and specificity, highlight the ongoing need for improved feature selection and regularization strategies in future work.

4.3.2. Performance on balanced dataset (Post-augmentation)

Post-augmentation performance with shape and size features

The use of an augmented dataset significantly enhanced classifier performance, particularly in terms of sensitivity and F1-score, while preserving the inherently high specificity of shape-based features. Among the classifiers, Logistic Regression continued to demonstrate the most stable and interpretable behavior across training, validation, and test splits. In contrast, tree-based models such as Random Forest and XGBoost, while achieving very high training scores, showed signs of overfitting during evaluation.

Table 4.13: Training set performance (post-augmentation)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.971**	0.969	**0.974**	**0.993**	**0.964**
AdaBoost	0.930	0.937	0.924	0.983	0.912
XGBoost	0.968	0.965	0.970	0.991	0.960
Decision Tree	0.962	0.953	0.972	0.962	0.955
K-Nearest Neighbors	0.934	**0.948**	0.920	0.978	0.916
Logistic Regression	0.935	0.942	0.928	0.980	0.918
MLP Neural Network	0.923	0.918	0.928	0.977	0.905

As shown in Table 4.13, the training performance across models was exceptionally high, with Random Forest achieving a balanced accuracy of 97.1%, a sensitivity of 96.9%, and an F1-score of 0.964. XGBoost performed similarly well in training, reaching a balanced accuracy of 96.8% and an F1-score of 0.960. Logistic Regression and K-Nearest Neighbors

Table 4.14: Validation set performance (post-augmentation)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.943	0.946	0.939	0.977	**0.927**
AdaBoost	0.932	0.933	0.930	0.977	0.914
XGBoost	0.936	**0.951**	0.922	0.977	0.918
Decision Tree	0.896	0.879	0.913	0.896	0.873
K-Nearest Neighbors	0.900	0.888	0.913	0.958	0.878
Logistic Regression	**0.940**	0.937	**0.942**	**0.980**	0.925
MLP Neural Network	0.932	**0.951**	0.913	0.978	0.912

Table 4.15: Test set performance (post-augmentation)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.763	0.643	0.883	0.773	0.484
AdaBoost	0.792	0.771	0.812	0.822	0.454
XGBoost	0.779	0.643	**0.916**	0.801	**0.542**
Decision Tree	0.654	0.371	0.937	0.654	0.388
K-Nearest Neighbors	0.748	0.600	0.896	0.761	0.480
Logistic Regression	**0.800**	**0.800**	0.801	0.839	0.453
MLP Neural Network	0.793	**0.800**	0.786	**0.847**	0.438

both exhibited strong and balanced training metrics, with balanced accuracies above 93% and F1-scores above 0.91.

Validation performance (Table 4.14) demonstrated that most classifiers were able to generalize well, with minimal overfitting. Logistic Regression once again stood out, achieving a balanced accuracy of 94.0%, a sensitivity of 93.7%, and a specificity of 94.2%. XGBoost and Random Forest maintained strong performance on the validation set as well, with balanced accuracies exceeding 93%. Even simpler models such as Decision Tree and K-Nearest Neighbors maintained competitive validation performance, suggesting that augmentation contributed positively to robustness.

Test set evaluation (Table 4.15) highlighted several improvements compared to pre-augmentation performance. Average sensitivity across classifiers improved by over 10 percentage points, with both MLP Neural Network and Logistic Regression reaching a sensitivity of 80.0%. Moreover, the average F1-score increased markedly from 0.277 to 0.463, representing a 67% improvement. This gain was particularly notable in models like XGBoost and Random Forest, which had previously suffered from poor recall. The generalization gap between training and test also decreased slightly for most models. For instance, Random Forest's performance drop from training to test was limited to 20.8%, while XGBoost's gap reduced from 24.1% to 18.8%.

Despite these improvements, some limitations persisted. Tree-based models, especially Decision Tree and Random Forest, still demonstrated relatively low recall values, indicating an ongoing challenge in detecting positive cases. Although augmented data improved generalization, XGBoost continued to show signs of overfitting, with a notable drop in test performance compared to its training scores. Furthermore, while the average F1-score increased substantially, it remained below commonly accepted clinical thresholds (typically >0.7), suggesting room for further model refinement.

Overall, the findings highlight the critical role of dataset augmentation in improving model robustness and performance in radiomics classification tasks. The results also reaffirm that well-structured, balanced data can enable simpler models like Logistic Regression to outperform more complex alternatives in generalization, thus supporting their use in high-stakes clinical applications where interpretability and reliability are essential.

Post-augmentation performance with non-shape features

Following data augmentation, the classifiers trained on the non-shape feature set showed considerable improvements in training and validation performance, particularly for ensemble-based models. As summarized in Table 4.16, Random Forest and XGBoost reached near-

perfect training results, achieving balanced accuracies of 97.5% with high sensitivity, specificity, and F1-scores. AdaBoost also performed competitively, while Logistic Regression and the MLP Neural Network delivered robust but slightly lower training scores. Notably, Decision Tree maintained high specificity but showed marginally reduced balance relative to the ensemble methods.

Table 4.16: Training set performance (non-shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.975**	0.967	**0.983**	**0.994**	**0.970**
AdaBoost	0.940	0.938	0.943	0.986	0.926
XGBoost	**0.975**	**0.973**	0.977	**0.994**	0.969
Decision Tree	0.963	0.955	0.970	0.963	0.955
K-Nearest Neighbors	0.916	0.927	0.905	0.967	0.894
Logistic Regression	0.904	0.892	0.916	0.946	0.883
MLP Neural Network	0.922	0.918	0.926	0.939	0.903

Validation results, shown in Table 4.17, reflected consistent generalization performance across most classifiers. Random Forest, AdaBoost, and XGBoost maintained high balanced accuracy in the 92–94% range. Their ROC AUC values remained strong, indicating effective discrimination. Logistic Regression was again among the most stable across datasets, with a balanced accuracy of 91.0% and an F1-score of 0.889. In contrast, Decision Tree, MLP, and K-Nearest Neighbors delivered slightly lower yet still respectable performance on the validation set.

Table 4.17: Validation set performance (non-shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.935	**0.942**	0.928	0.974	0.917
AdaBoost	**0.936**	0.933	**0.939**	0.973	**0.920**
XGBoost	0.929	0.937	0.922	**0.975**	0.911
Decision Tree	0.896	0.901	0.890	0.896	0.870
K-Nearest Neighbors	0.911	0.928	0.893	0.942	0.887
Logistic Regression	0.910	0.901	0.919	0.943	0.889
MLP Neural Network	0.905	0.906	0.904	0.921	0.882

Performance on the test set, detailed in Table 4.18, showed noticeable degradation for

most classifiers compared to validation. Although AdaBoost and Logistic Regression retained relatively stable metrics, classifiers such as Decision Tree and MLP suffered substantial declines. Specifically, Decision Tree's sensitivity dropped to 41.4%, pointing to its reduced ability to detect malignant cases. The average decline in balanced accuracy from validation to test was approximately 17.5%, suggesting challenges in generalizing learned patterns to unseen data. Furthermore, F1-scores across all classifiers remained modest, with none exceeding 0.48, indicating persistent difficulties associated with class imbalance.

Table 4.18: Test set performance (non-shape features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	0.770	0.700	0.840	0.801	0.454
AdaBoost	**0.783**	**0.743**	0.824	0.807	0.454
XGBoost	0.771	0.714	0.829	0.806	0.446
Decision Tree	0.662	0.414	**0.909**	0.662	0.377
K-Nearest Neighbors	0.702	0.657	0.746	0.748	0.341
Logistic Regression	0.781	0.714	0.848	**0.833**	**0.472**
MLP Neural Network	0.761	0.729	0.794	0.803	0.415

When compared to models trained on shape-based features, the non-shape configuration resulted in slightly lower test accuracies and weaker F1-scores. Logistic Regression emerged as an exception, consistently providing stable and balanced results. Despite augmentation, the non-shape feature set exhibited limited capacity for generalization, especially when used in isolation. These results underscore the importance of incorporating morphological descriptors or applying additional regularization strategies to counteract overfitting and improve clinical robustness.

Post-augmentation performance with all features

The evaluation using the complete feature set, which combines shape descriptors and non-shape radiomic features, showed strong improvements in both training and validation phases after data augmentation. Ensemble-based models such as Random Forest and XGBoost reached high training balanced accuracy scores of 97.4%, with ROC AUC values exceeding 0.99, indicating excellent class discrimination. While Logistic Regression did not reach these peak values, it still performed reasonably well on the training data, with a balanced accuracy of 87.7% and a ROC AUC of 0.919. These outcomes, presented in

Table 4.19, suggest that the inclusion of both feature types enhances the model's ability to capture patterns in the augmented dataset.

Table 4.19: Training set performance (all features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.974**	**0.970**	**0.978**	**0.995**	**0.968**
AdaBoost	0.942	0.939	0.945	0.986	0.928
XGBoost	**0.974**	**0.970**	**0.978**	0.993	**0.968**
Decision Tree	0.966	0.955	0.977	0.966	0.960
K-Nearest Neighbors	0.917	0.934	0.900	0.966	0.895
Logistic Regression	0.877	0.844	0.909	0.919	0.851
MLP Neural Network	0.923	0.937	0.908	0.938	0.902

Validation performance remained aligned with training trends. As shown in Table 4.20, ensemble models such as Random Forest and AdaBoost maintained balanced accuracy above 93%, and both sensitivity and specificity remained tightly coupled near 0.93. XGBoost preserved its high ROC AUC, even though its specificity was slightly more variable. Logistic Regression showed a modest drop compared to the ensemble methods but still generalized well. The MLP Neural Network, while not the top performer, achieved a validation balanced accuracy of 90.9%, suggesting good learning under augmented conditions.

Table 4.20: Validation set performance (all features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.934**	**0.937**	0.930	0.977	0.917
AdaBoost	0.932	0.928	**0.936**	**0.980**	0.916
XGBoost	0.926	0.928	0.925	0.978	0.908
Decision Tree	0.924	0.915	0.933	0.924	0.907
K-Nearest Neighbors	0.899	0.910	0.887	0.939	0.873
Logistic Regression	0.885	0.843	0.928	0.907	0.862
MLP Neural Network	0.909	0.928	0.890	0.926	**0.885**

Test set performance, reported in Table 4.21, showed greater variability across models. Random Forest emerged as the most reliable, achieving a test balanced accuracy of 79.2% and a ROC AUC of 0.831. AdaBoost and XGBoost performed similarly in terms of

balanced accuracy but suffered slightly lower F1-scores. Despite strong results in earlier phases, models like XGBoost and MLP showed a sharper decline, suggesting overfitting effects. Decision Tree maintained excellent specificity (91.9%) but performed poorly in sensitivity (45.7%), indicating a high bias toward negative class predictions. Logistic Regression posted the lowest metrics overall, with a balanced accuracy of 65.9% and a F1-score of just 0.297, highlighting its limited adaptability to the full feature space.

Table 4.21: Test set performance (all features)

Classifier	Bal_Acc	Sensitivity	Specificity	ROC AUC	F1-score
Random Forest	**0.792**	**0.786**	0.797	**0.831**	**0.444**
AdaBoost	0.787	**0.786**	0.787	0.826	0.433
XGBoost	0.767	**0.786**	0.748	0.814	0.396
Decision Tree	0.688	0.457	**0.919**	0.688	0.424
K-Nearest Neighbors	0.679	0.629	0.730	0.708	0.317
Logistic Regression	0.659	0.600	0.718	0.700	0.297
MLP Neural Network	0.708	0.657	0.759	0.766	0.351

These results indicate that after augmentation, most models surpassed 90% validation accuracy, suggesting strong learning dynamics. However, generalization to the test set remained a challenge. The average drop in balanced accuracy from training to test was about 20.9%, reinforcing concerns over overfitting, especially in more complex classifiers like MLP and XGBoost. Even though F1-scores improved compared to pre-augmentation metrics, none of the models exceeded 0.45, underscoring the persistent effect of class imbalance.

When comparing against shape-based models, the addition of non-shape features lowered average test accuracy by about 7.3%, confirming that morphological descriptors alone offered more robust generalization. Nevertheless, when compared with using non-shape features in isolation, the full feature set showed clear gains in balanced accuracy and ROC AUC. Logistic Regression again struggled in high-dimensional space, while Random Forest proved to be the most consistently high-performing model.

Overall, while combining all feature types and applying augmentation enhanced learning and validation stability, test-time generalization still proved limited. The findings emphasize the benefit of incorporating interpretable shape features and the necessity of robust regularization to fully leverage complex radiomic datasets.

4.3.3. Classification conclusion and comparative analysis

A detailed evaluation was conducted across three radiomic feature configurations, shape-only (3 features), non-shape (76 features) and all features combined (79 features), to assess classification performance before and after data augmentation. These configurations were tested using seven classifiers, with a focus on balanced accuracy, ROC AUC, and F1-score on the test set.

Before augmentation, the best overall test performance was observed using all 79 features, where Random Forest and XGBoost achieved balanced accuracies of 76.4% and 76.0%, respectively. However, this configuration also showed a moderate generalization gap, with a 15.8% drop in balanced accuracy from training to testing. Logistic regression underperformed in this setting, reaching only 61.0% balanced accuracy and 49.1% specificity, which resulted in many false positives.

In contrast, the shape-only configuration, which used just three interpretable morphological descriptors, demonstrated the most stable generalization. Logistic regression achieved a balanced test accuracy of **80.6%** and an ROC AUC of **0.847**, outperforming more complex models using a larger number of features. The average specificity in this group reached 84.3%, and the generalization drop was minimal at just 14.7%, suggesting that shape features are highly informative and robust for malignancy classification. Non-shape features performed more inconsistently. While XGBoost achieved a 77.5% balanced accuracy in this group, sensitivity values varied widely across models, indicating potential overfitting and a lack of class-discriminative power without structural input.

Following data augmentation, classification performance improved across all configurations. Shape-only features remained the most reliable: logistic regression again delivered **80.6%** balanced accuracy and the highest test ROC AUC of **0.847**, reaffirming the value of shape descriptors. The full feature set also benefitted from augmentation, with Random Forest reaching a balanced accuracy of 79.2%, though it still fell short of the shape-only configuration in terms of robustness. Meanwhile, non-shape features continued to exhibit weaker generalization. For example, logistic regression on this group dropped to a test ROC AUC of 0.700 and an F1-score of just 0.297.

These results highlight the strong predictive power of shape features, not only due to their simplicity and interpretability but also their resistance to overfitting. They also suggest that combining shape and non-shape features may introduce redundancy or noise unless carefully curated, especially when using simpler linear models.

Compared to prior studies, the performance achieved in this work is competitive despite

using a relatively small and interpretable feature set. For example, Choi et al. [5] reported 87.2% sensitivity and 84.6% accuracy on a small LIDC-IDRI subset using LASSO-selected radiomic features and SVM, but their dataset was limited in scale. Peikert et al. [19] obtained 90.4% sensitivity and an AUC of 0.94 using 726 NLST nodules and eight features, but their dataset was artificially balanced, which may have inflated performance. Liu et al. [14], who also employed the LIDC-IDRI dataset, reveals both methodological parallels and key performance differences. In the shape-based setting, their logistic regression model trained on three SS features achieved 80.5% balanced accuracy and 0.874 AUC—closely aligned with the 80.6% accuracy and 0.847 AUC observed in this work using similar shape descriptors. This consistency underscores the reliability of morphological features in lung nodule classification.

In contrast, when excluding SS features, Liu et al. [14] reported 66.4% balanced accuracy and 0.822 AUC, whereas models using non-shape features in this study reached up to 78.3% accuracy with 0.807 AUC. While both studies demonstrate effective classification using all features combined, Liu et al.'s best model (AdaBoost) yielded 76.9% accuracy and 0.865 AUC, compared to 79.2% and 0.831, respectively, from a Random Forest model here. These variations likely reflect differences in feature reduction strategies, augmentation methods, and classifier configurations, rather than fundamental disparities in dataset or task complexity.

Notably, many existing models employed larger feature sets and more complex pipelines, often accompanied by limitations such as small sample sizes, artificial class balancing, or restricted generalizability. The classification framework developed in this study, by contrast, uses explainable and compact radiomic descriptors and still achieves a comparable or superior level of generalization. This underscores the diagnostic utility of shape features and reinforces their value in resource-constrained or clinically realistic settings.

5 | Future developments

The current work demonstrates the effectiveness of deep learning-based approaches for simulating realistic LDCT images from high-dose acquisitions. However, several avenues exist for future development. One promising direction is the integration of generative adversarial networks (GANs) into the simulation pipeline. GANs have shown superior capabilities in generating perceptually realistic images, especially in medical imaging contexts where fine texture and noise realism are crucial. Employing a GAN-based discriminator in conjunction with the current attention-integrated autoencoder could enhance the perceptual quality of the simulated LDCT slices by enforcing distributional similarity to real low-dose domains.

Another significant opportunity lies in the hybridization of traditional physics-based noise models with learning-based architectures. Although deep learning provides flexibility and generalization, it can overlook the physical constraints inherent in CT imaging systems. By combining analytical noise modeling—such as Poisson-Gaussian injection in the sinogram or projection domain—with autoencoder-based reconstructions, the model could benefit from both physical realism and data-driven adaptability. Furthermore, expanding the simulation dataset beyond the current 16,500 slices by incorporating more full-dose/low-dose pairs from public or clinical sources would improve model generalizability across scanner types and patient populations.

In the context of lung nodule malignancy classification, future work could explore more comprehensive feature selection strategies. While this study employed inter-class correlation, geometric perturbation, and correlation filtering, additional methods such as LASSO regression, forward selection, backward elimination, and minimum redundancy maximum relevance (mRMR) could uncover complementary and non-linear feature interactions. Integrating ensemble-based feature ranking may further enhance classifier interpretability and robustness.

Moreover, novel data augmentation techniques could be developed to improve classifier resilience in class-imbalanced settings. Beyond morphological operations, generative approaches such as synthetic minority oversampling (SMOTE) in 3D, deep feature pertur-

bation, or even GAN-based augmentation may provide more diverse and anatomically consistent malignant nodule representations. Such techniques could also be adapted to simulate temporal variations or different acquisition protocols, contributing to more robust decision-making models.

Finally, a shift toward end-to-end classification models using raw CT patches as input—bypassing handcrafted radiomic features altogether—represents another promising direction. Deep convolutional networks or multilayer perceptrons (MLPs) could be trained directly on cropped and normalized nodule patches to learn abstract representations optimized for malignancy classification. This approach may reduce preprocessing complexity and improve performance by allowing the model to learn task-specific features, especially when trained on large augmented datasets.

A final chapter containing the main conclusions of your research/study and possible future developments of your work have to be inserted in this chapter.

Bibliography

[1] S. Alahmari, D. Cherezov, D. Goldgof, and et al. Delta radiomics improves pulmonary nodule malignancy prediction in lung cancer screening. *IEEE Access*, 6:77796–77806, 2018. doi: 10.1109/ACCESS.2018.2884126.

[2] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. doi: 10.1118/1.3528204.

[3] J. Causey, J. Zhang, S. Ma, and et al. Highly accurate model for prediction of lung nodule malignancy with ct scans. *Scientific Reports*, 8(1):9286, 2018. doi: 10.1038/s41598-018-27569-w.

[4] H. Chen, Y. Zhang, W. Zhang, P. Liao, K. Li, J. Zhou, and G. Wang. Low-dose ct via convolutional neural network. *Biomedical optics express*, 8(2):679–694, 2017. doi: 10.1364/BOE.8.000679.

[5] J. Choi and et al. Characterization and screening recall of indeterminate prevalent pulmonary nodules with low-dose ct-based radiomics. *Journal Reference Placeholder*, 2018.

[6] N. Garau, C. Paganelli, P. Summers, and et al. External validation of radiomics-based predictive models in low-dose ct screening for early lung cancer diagnosis. *Medical Physics*, 47(9):4125–4136, 2020. doi: 10.1002/mp.14308.

[7] K. M. Hasib, M. S. Iqbal, F. M. Shah, J. A. Mahmud, M. H. Popel, M. I. H. Showrov, S. Ahmed, and O. Rahman. A survey of methods for managing the classification and solution of data imbalance problem. *Journal of Computer Science*, 16:1546–1557, 2020. doi: 10.3844/JCSSP.2020.1546.1557.

[8] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647.

[9] F. L. Iacono, R. Maragna, G. Pontone, and V. D. A. Corino. A novel data augmentation method for radiomics analysis using image perturbations. *Journal of Imaging and Information Medicine*, pages 1–14, 2024. doi: 10.1007/S10278-024-01013-0. URL `https://doi.org/10.1007/S10278-024-01013-0`.

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[11] C. W. Kim and J. H. Kim. Realistic simulation of reduced-dose ct with noise modeling and sinogram synthesis using dicom ct images. *Medical Physics*, 41(1):011901, 2014.

[12] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

[13] J. Liu, H. Xu, H. Qing, and et al. Comparison of radiomic models based on low-dose and standard-dose ct for prediction of adenocarcinomas and benign lesions in solid pulmonary nodules. *Frontiers in Oncology*, 10:634298, 2021. doi: 10.3389/fonc.2020.634298.

[14] J. Liu, A. Corti, V. D. Corino, and L. Mainardi. Lung nodule classification using radiomics model trained on degraded sdct images. *Computers in Biology and Medicine*, 144:105386, 2022. doi: 10.1016/j.compbiomed.2022.105386.

[15] L. Mao, H. Chen, M. Liang, and et al. Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose ct screening. *Quantitative Imaging in Medicine and Surgery*, 9(2):263–272, 2019. doi: 10.21037/qims.2019.02.02.

[16] R. E. Naziroglu, V. F. van Ravesteijn, L. J. van Vliet, G. J. Streekstra, and F. Vos. Simulation of scanner- and patient-specific low-dose ct imaging from existing ct images. *Physica Medica*, 36:12–23, 2017.

[17] A. A. of Physicists in Medicine. 2016 low dose ct grand challenge. Online, 2016. `https://www.aapm.org/grandchallenge/lowdosect/`.

[18] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[19] T. Peikert, F. Duan, S. Rajagopalan, and et al. Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the

national lung screening trial. *PLoS One*, 13(6):e0196910, 2018. doi: 10.1371/journal.pone.0196910.

[20] T. Peikert, B. J. Bartholmai, and F. Maldonado. Radiomics-based management of indeterminate lung nodules? are we there yet? *American Journal of Respiratory and Critical Care Medicine*, 202(2):165–167, 2020. doi: 10.1164/rccm.202004-1279ED.

[21] Reckoning Dev. Autoencoders explained. `https://reckoning.dev/blog/autoencoders/`, 2023. Accessed: 2025-06-10.

[22] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

[23] L. Rundo, R. Ledda, C. Di Noia, and et al. A low-dose ct-based radiomic model to improve characterization and screening recall intervals of indeterminate prevalent pulmonary nodules. *Diagnostics*, 11(9):1610, 2021. doi: 10.3390/diagnostics11091610.

[24] J. Schlemper, O. Oktay, M. Schaap, and et al. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019.

[25] A. Sharma. Applied deep learning - part 3: Autoencoders. `https://medium.com/data-science/applied-deep-learning-part-3-autoencoders-1c083af4d798`, 2018. Accessed: 2025-06-10.

[26] J. D. Shur and et al. Radiomics in oncology: A practical guide. *The British Journal of Radiology*, 94(1117):20201012, 2021. doi: 10.1259/bjr.20201012.

[27] A. Srinivas. U-net explained – understanding its image segmentation architecture. `https://medium.com/data-science/u-net-explained-understanding-its-image-segmentation-architecture-56e4842e313a`, 2021. Accessed: 2025-06-10.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

[29] T. Takenaga, S. Katsuragawa, M. Goto, M. Hatemura, Y. Uchiyama, and J. Shiraishi. A computer simulation method for low-dose ct images by use of real high-dose images: a phantom study. *Radiological Physics and Technology*, 9:44–52, 2016. doi: 10.1007/s12194-015-0336-5. Published online: 20 August 2015.

[30] F. Wang, M. Jiang, C. Qian, and et al. Residual attention network for image classification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[32] L. Yu, M. Shiung, D. Jondal, and C. H. McCollough. Development and validation of a practical lower-dose-simulation tool for optimizing computed tomography scan protocols. *Journal of Computer Assisted Tomography*, 36(4):477–487, 2012.

[33] D. Zeng, J. Huang, Z. Bian, S. Niu, H. Zhang, Q. Feng, Z. Liang, and J. Ma. A simple low-dose x-ray ct simulation from high-dose scan. *IEEE Transactions on Nuclear Science*, 62(5):2226–2237, 2015.

List of Figures

List of Tables