POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

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

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedico

Author: Reza Gonabadi

Advisor: Prof. Luca Mainardi

Co-advisor: Jiaying Liu

Academic year: 2024-2025

1. Introduction

Computed tomography (CT) plays a vital role in early disease detection, particularly lung cancer. However, the high radiation dose from standard CT scans poses long-term health risks, especially in screening programs requiring repeated imaging. Low-dose CT (LDCT) protocols offer a safer alternative, but at the cost of image quality degradation, introducing noise, reduced contrast, and poor resolution, which hinder accurate diagnosis and machine learning performance. The scarcity of annotated LDCT datasets further limits the development of robust diagnostic models. Moreover, challenges such as class imbalance and feature sensitivity to noise complicate lung nodule classification. This thesis addresses these limitations by proposing a deep learning framework that first simulates realistic LDCT images from standard-dose data and then classifies lung nodules using a radiomics-based pipeline optimized for LDCT scenarios.

2. Methodology overview

The proposed methodology consists of multiple sequential stages, as illustrated in Figure 1. The

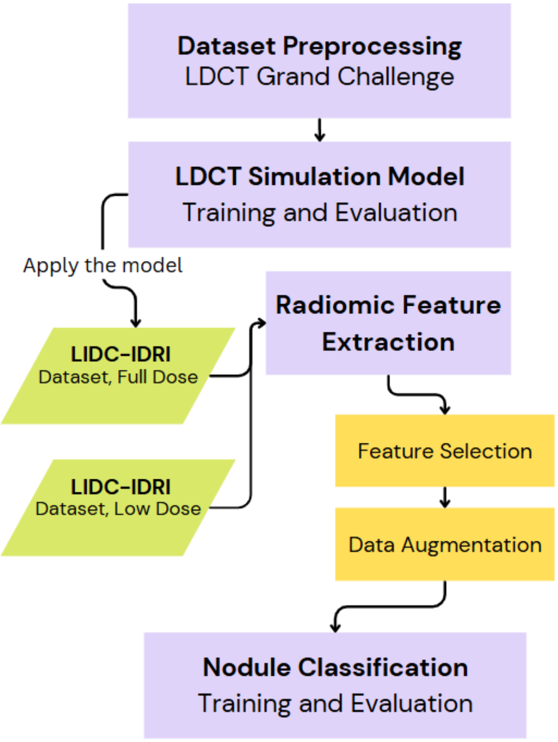

Figure 1: End-to-end pipeline overview

process begins with dataset preprocessing and LDCT simulation applied to the Grand Challenge dataset. The trained simulation model is then used to generate synthetic LDCT images from the full-dose (FD) CT scans of the LIDC-IDRI dataset. Radiomic feature extraction is performed on both the simulated LD and original LD LIDC-IDRI Slices. Subsequently, data augmentation techniques are applied to balance the dataset, followed by the classification stage to assess lung nodule malignancy.

2.1. Dataset and preprocessing

Two datasets were employed to support the simulation and classification tasks of this study. The first dataset, from the 2016 Low Dose CT Grand Challenge, originally included 299 subjects, of which 99 were available for download. Among these, 50 subjects provided paired FD and simulated LD chest CT slices. Each of these 50 subjects contributed approximately 250–350 aligned slice pairs, resulting in a total of around 16,500 paired slices used for training and evaluating the LDCT simulation model. All images were acquired using standard protocols (120 kV, 200 mAs) and cover the thoracic region [1]. A summary of this dataset is provided in Table 1. For the simulation task, 85% of the data was allocated to model training, while the remaining 15% was reserved for final test evaluation.

Table 1: Dataset overview for simulation task

Subjects (Total / Downloaded)	299 / 99
With Paired FD-LD Scans	50
Slices per Subject	250–350
Total Paired Slices	∼16,500

The second dataset, LIDC-IDRI [2], originally contained 1,018 thoracic CT scans; however, 8 were excluded due to incompleteness or duplication, resulting in 1,010 valid cases. Nodules measuring ≥3 mm with consensus annotations were selected, yielding 2,625 nodules in total. Malignancy labels were assigned using an average score threshold of ≥4. CT scans were categorized based on `XRayTubeCurrent`, with values $>$ 80 mA labeled as FD and $\leq$ 80 mA as true LD. The trained simulation model was applied to FD cases to generate synthetic LDCT images, which were then used exclusively for training and validation (80% for training, 20% for validation). The true LD scans were preserved without modification and used solely as the test set for lung nodule classification. Table 2 summarizes the composition of the LIDC-IDRI dataset, reporting the number of participants and nodules for both real FD and LD scans. It also highlights the distribution of malignancy scores and the underlying class imbalance that motivated the use of augmentation techniques during model training.

Table 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

To ensure data quality and consistency in the simulation dataset, all CT images from the Grand Challenge underwent intensity normalization. Additionally, a custom artifact removal pipeline was applied to eliminate scanner-induced linear structures near the bottom edge of some slices. Using shape descriptors such as eccentricity and area, the algorithm masked non-anatomical regions while preserving lung tissue. Figure 2 illustrates this process, showing a FDCT slice before and after artifact suppression. These preprocessing steps ensured cleaner inputs for the LDCT simulation process.

2.2. Low-dose CT simulation

To simulate realistic LDCT slices from FD scans, a progressive deep learning pipeline was developed in three stages: a simple autoencoder (AE), a U-Net architecture with skip connections, and a final attention-enhanced U-Net. This encoder–decoder design enables effective image-to-image translation, progressively improving structural fidelity and noise realism at each stage.

The initial AE learns compressed representations of FDCT images and reconstructs their LDCT counterparts by minimizing the Mean Absolute Error (MAE), offering stable training

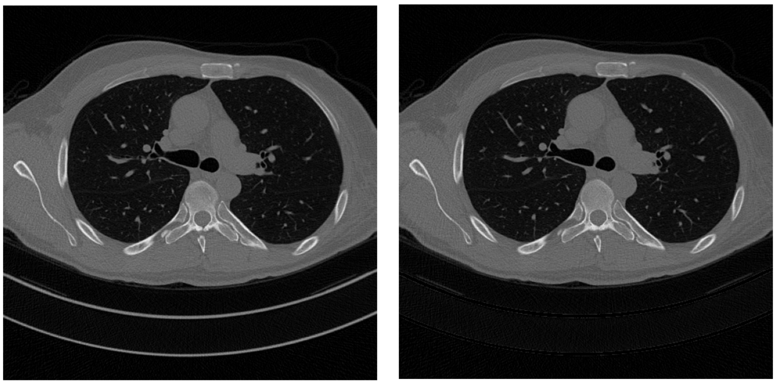

Figure 2: Artifact removal in a FDCT slice from the 2016 LDCT Grand Challenge dataset (subject 162, slice 120). Left: before preprocessing. Right: after removing scanner-induced lines.

but limited preservation of fine anatomical detail. To overcome this, a U-Net was introduced to incorporate skip connections between encoder and decoder layers, enabling high-resolution spatial features to bypass the bottleneck. This improved the reconstruction of subtle structures critical for clinical interpretation.

To further preserve anatomical textures and enhance fine structural details in reconstructed slices, attention mechanisms were integrated into the U-Net architecture. These modules improve the network's ability to retain subtle but clinically important features, particularly in noisy or low-contrast regions, while maintaining the benefits of skip connections. An attention mechanism is a learnable module that enables the network to focus more on informative regions of the feature map by assigning higher weights to relevant spatial features and suppressing less important ones. In this context, it helps highlight subtle anatomical details that might otherwise be overlooked in LD reconstructions. The architecture consists of 6 convolutional layers (2 main, 4 within attention blocks), 2 deconvolutional layers, 2 attention blocks, and a single skip connection. This compact design balances structural preservation with computational efficiency, as illustrated in Figure 3. Note that attention modules, unlike standard convolutional blocks, cannot be visually depicted in the same manner and are therefore not explicitly shown.

To assess image quality, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) were used. PSNR evaluates pixel-level accuracy, with values above 30 dB generally considered high-quality. SSIM measures perceptual similarity, where values closer to 1 indicate better structural preservation.

Initially, the autoencoder model was trained using only the MAE to minimize pixel-wise intensity differences. To improve structural preservation, the U-Net model employed a composite loss function that combined MAE with SSIM, assigning weights of 30% to MAE and 70% to SSIM. In the final attention-based model, the weight of SSIM was further increased to 84% to better capture perceptual and anatomical fidelity, while MAE was reduced to 16%. These weights were determined empirically through a trial-and-error process to balance intensity accuracy with structural coherence.

2.3. Nodule classification using radiomic features

Radiomic feature extraction was applied to both the synthetic and true LDCT images from the LIDC-IDRI dataset. In medical imaging, radiomic features quantify the shape, intensity, and texture of anatomical structures, enabling the transformation of medical scans into high-dimensional, mineable data. These features are commonly categorized into four groups: first-order statistics (e.g., energy, entropy), shape descriptors (e.g., sphericity, major axis length), texture features from gray-level matrices (e.g., GLCM), and wavelet-transformed representations.

To refine the extracted feature space, a three-stage feature selection pipeline was applied to nodules with four separate annotations from different radiologists to ensure labeling consistency. First, feature stability was assessed using the inter-class correlation coefficient (ICC), and only features with ICC values greater than 0.75 were

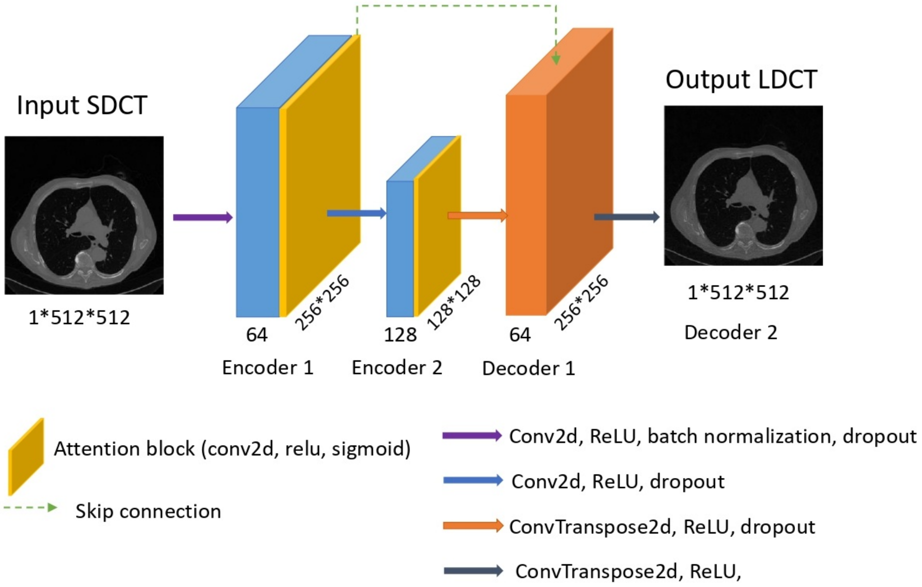

Figure 3: Final model architecture based on an attention-enhanced U-Net for LDCT simulation.

retained. Next, to evaluate feature robustness to ROI localization shifts, the same nodule mask was translated spatially, and ICC was recalculated to filter out unstable features under minor positional perturbations. Finally, Spearman's correlation was used to remove redundant features, applying a threshold of $|\rho| > 0.85$. This process produced a compact and robust set of radiomic features suitable for machine learning model training.

To address class imbalance (1,725 benign vs. 225 malignant nodules), a targeted augmentation strategy was applied exclusively to malignant cases. Three augmentation techniques were used: (1) morphological operations (dilation and erosion), (2) noise injection (salt-and-pepper and Gaussian blur), and (3) superpixel-based structural deformations. Each malignant sample produced 4–6 augmented variants, increasing the malignant pool to over 1,100 samples [3].

A range of supervised classifiers, including Random Forest, AdaBoost, XGBoost, Decision Tree, K-Nearest Neighbors, Logistic Regression, and a Multilayer Perceptron (MLP), were trained using the selected radiomic features. Evaluation was conducted using a 20-iteration bootstrapped protocol with fixed test sets. Performance metrics such as balanced accuracy, sensitivity, specificity, F1-score, and ROC AUC were computed and averaged to assess model robustness and generalization [4].

3. Results

3.1. Low-dose simulation results

Three models were progressively developed for LDCT simulation from FDCT inputs. The simple autoencoder (AE) achieved a PSNR of 23.50 dB and SSIM of 0.723 on the test set, but suffered from excessive smoothing and poor structural preservation. Incorporating skip connections via a U-Net architecture improved anatomical detail retention, increasing PSNR to 27.84 dB and SSIM to 0.8881. Finally, the addition of attention mechanisms further enhanced perceptual quality and sharpness, resulting in a PSNR of 32.02 dB and SSIM of 0.9311 on the test set, as summarized in Table 3.
Overall, the U-Net yielded an 18.5 PSNR gain over the baseline, and the attention-enhanced model added another 15.0, with SSIM improving by 28.8% from start to finish.

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

Table 3: Performance comparison across different architectures on test set.

In addition to PSNR and SSIM, the reconstructed images were evaluated based on their noise power spectrum (NPS). The proposed method achieved a normalized mean absolute difference (MAD) of **0.0672** between the radial NPS curves of reconstructed and real LDCT images. Although previous studies such as Naziroglu et al. [5] did not report PSNR or SSIM, the NPS error falls within the typical literature range (0.05–0.15), confirming strong noise texture consistency.

Figure 4 presents a sample output from the LIDC-IDRI dataset, where the final model was used to generate synthetic LD images from FD input slices.

3.2. Feature selection results

A total of 851 radiomic features were extracted and passed through a structured selection pipeline. The stepwise reduction process, summarized in Table 6, resulted in a compact set of 79 highly informative features. For classification, these final features were organized into three groups: shape and size descriptors (3 features), non-shape features (76 features), and the combined full set (79 features), enabling comparative analysis of geometry-based versus texture and intensity-based models, as shown in Table 4.

Table 4: Final feature groups for classification

Group	Count
All features	79
Shape and size only	3
Non-shape features	76

3.3. Nodule classification and comparative insights

Shape and size

Using only three features, *MajorAxisLength*, *Sphericity*, and *Elongation*, Logistic Regression achieved 81.3% balanced accuracy and 0.851 AUC before augmentation, followed by MLP. After augmentation, both models reached 80.0% sensitivity with balanced accuracies above 79%, and AUCs of 0.839 (LR) and 0.847 (MLP).

Non-shape

With 76 non-shape features, XGBoost achieved 77.5% balanced accuracy and 0.802 AUC before augmentation. After augmentation, AdaBoost performed best with 78.3% accuracy and 0.807 AUC, followed by Logistic Regression (78.1%, 0.833 AUC).

All features

Combining all 79 features, Random Forest and XGBoost reached 76.4% and 76.2% accuracy with AUCs of 0.816 and 0.797 before augmentation. After augmentation, Random Forest led with 79.2% accuracy and 0.831 AUC; AdaBoost and XGBoost followed, while simpler models remained less effective.

A comparative summary of test performance across the three feature groups, shape-only, non-shape, and all features, highlighting the best-performing model and its corresponding balanced accuracy and ROC AUC, is presented in Table 5.

Features	**Top Model**	**Bal. Acc. / AUC**
Shape-only (3)	Logistic Reg.	80.0% / 0.839
Non-shape (76)	AdaBoost	78.3% / 0.807
All (79)	Rand. Forest	79.2% / 0.831

Table 5: Test performance summary across feature groups after augmentation.

Performance comparison with prior work

Our findings are consistent with those of Liu et al. [4], whose approach informed our pipeline design. Their logistic regression model using three shape and size (SS) features achieved 80.5% balanced accuracy and 0.874 AUC, closely matching our shape-based model, which reached 80.6% accuracy and 0.847 AUC after augmentation.

Liu et al.'s best model using all features (AdaBoost) reached 76.9% accuracy and 0.865 AUC, while our Random Forest model performed slightly better in accuracy (79.2%) but slightly lower in AUC (0.831). For non-shape features, our AdaBoost model (78.3%, AUC 0.807) outperformed theirs (66.4%, AUC 0.822).

Table 6: 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

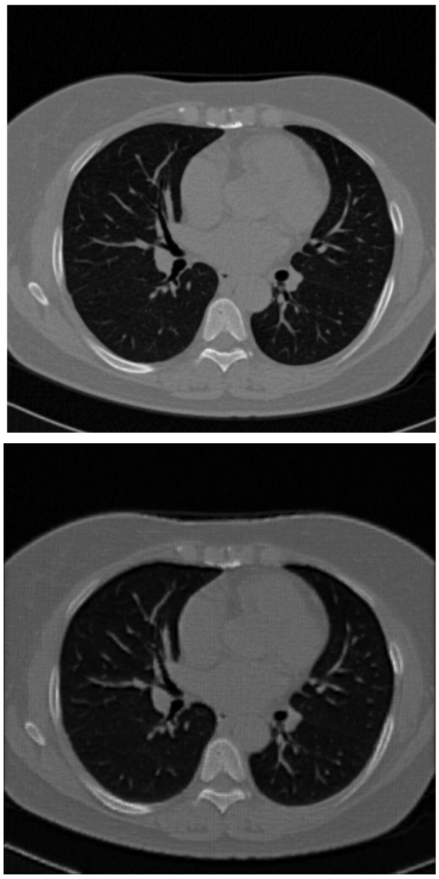

Figure 4: Examples of synthetic LD generation using the trained model on LIDC-IDRI. Top: FDCT slice. Bottom: Simulated LDCT slice.

4. Conclusions

This work introduced a unified deep learning pipeline for LDCT simulation and lung nodule malignancy classification. To address the risks of repeated FD imaging and the scarcity of annotated LDCT data, a progressive architecture culminating in an attention-enhanced U-Net was developed. This model produced realistic LD images with improved structural fidelity and was subsequently applied to the LIDC-IDRI dataset for downstream analysis. On the test set, the final model achieved a PSNR of **32.02 dB** and SSIM of **0.9311**. In addition, a normalized mean absolute difference (MAD) of **0.0672** was observed between the radial NPS curves of the reconstructed and real LDCT images, indicating close alignment in noise properties.

Despite the challenges of working with LDCT images, the classification results remained strong across all feature configurations. Shape and size descriptors proved most robust, achieving **80.0%** balanced accuracy and **0.839** AUC using only three features. Non-shape and combined feature sets also performed well, reaching **78.3%** and **79.2%** accuracy with AUCs of **0.807** and **0.831**, respectively. These outcomes confirm the effectiveness of radiomics-based classification pipelines, even under reduced-dose imaging conditions.

References

[1] https://www.aapm.org/grandchallenge/lowdosect/.

[2] https://www.cancerimagingarchive.net/collection/lidc-idri/.

[3] F. Lo 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.

[4] Jiaying Liu, Anna Corti, Valentina D.A. Corino, and Luca Mainardi. Lung nodule classification using radiomics model trained on degraded sdct images. *Computers in Biology and Medicine*, 144:105386, 2022.

[5] Ramazan E Naziroglu, Vince F van Ravesteijn, Lucas J van Vliet, Geert J Streekstra, and FM Vos. Simulation of scanner- and patient-specific low-dose ct imaging from existing ct images. *Physica Medica*, 36:12–23, 2017.