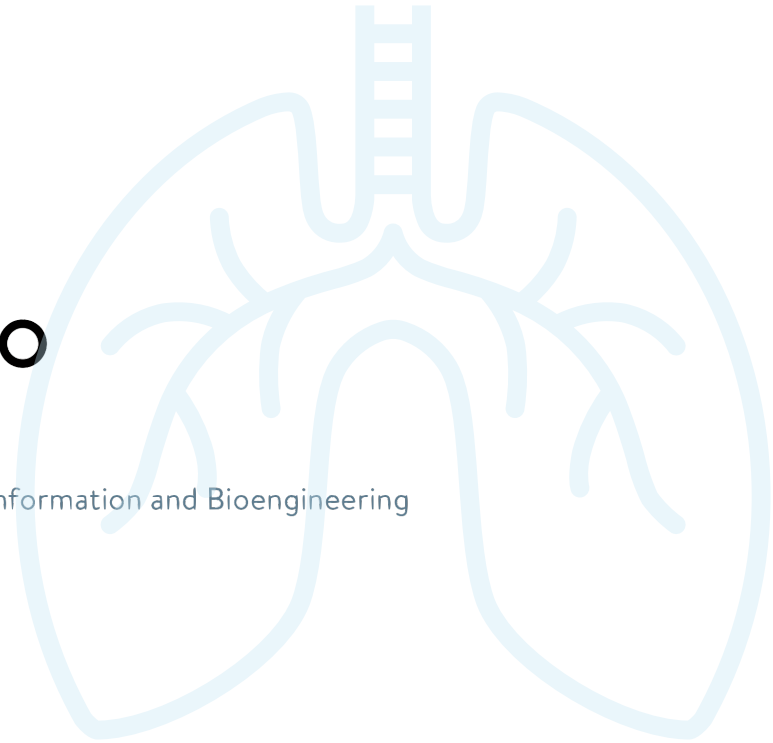




POLITECNICO
MILANO 1863

Department of Electronics, Information and Bioengineering



Advanced AI Frameworks for Lung Nodule Analysis: From Synthetic Low- Dose CT Simulation to Nodule Detection and Classification

Doctoral Dissertation of:
Jiaying Liu

Advisor:

Prof. Luca Mainardi

Prof. Valentina Corino

Dr. Anna Corti

Tutor: **Prof. Manuela Ferrario**

Coordinator of the PhD program: **Prof. Raffaele Dellacà**

XXXVIII Cycle - 2026

Acknowledgements

The journey towards completing this PhD thesis has been long, and along the way I have been fortunate to receive guidance, encouragement, and kindness from many people.

First and foremost, I would like to thank my supervisors, Professor Luca Mainardi, Professor Valentina Corino and Dr. Anna Corti, who have been invaluable to my academic growth. I deeply appreciate their support, guidance and prompt feedback on my work on these past three years. Without their dedication and thoughtful reviews, this thesis would not have been possible. I would also like to thank the members of my thesis committee, Professor João Pedrosa and Professor Laura Burattini, for their careful reading and valuable feedback.

Second, I would like to thank Professor Ender Konukoglu and Qi Ma from ETH Zurich, from whom I learned so much in just 6 months during my period abroad. I am deeply grateful for the opportunity to collaborate with them. Thanks to their mentorship, I had a wonderful experience and grew tremendously both professionally and personally.

Third, I want to thank all the friends I made during my PhD, who made this journey special and more joyful. I hope our friendships last many years after the PhD, if not forever. Special thanks to Inês and Ricardo, who were always there to listen and offer support throughout these years. Special thanks to the Camogli group of friends, they also made this journey lighter, less challenging and more enjoyable.

Finally, I would like to thank my family and Ali for their unconditional love and support. They are the pillars of my life, and I hope they are always there to witness my future adventures and successes. Above all, I thank the Lord for everything I have achieved and for all that I have been given.

Contents

GRAPHICAL ABSTRACT	1
ABSTRACT	2
LIST OF PUBLICATIONS	3
1 COVER ESSAY	4
1.1. INTRODUCTION	4
1.2. BACKGROUND AND STATE-OF-THE-ART	7
1.2.1. <i>Image degradation and restoration in computed tomography</i>	7
1.2.2. <i>Lung nodule detection</i>	23
1.2.3. <i>Lung nodule classification</i>	33
1.3. THESIS MOTIVATION AND OBJECTIVES	41
1.4. THESIS OUTLINE	44
2 A COMPARATIVE ANALYSIS OF CT DEGRADATION FOR ROBUST LDCT NODULE CLASSIFICATION USING RADIOMICS	46
3 LEVERAGING SELF-SUPERVISED PRETRAINING USING TRANSFORMERS FOR ENHANCED LUNG NODULE DETECTION IN CT SCANS	57
4 LUNG NODULE CLASSIFICATION USING RADIOMICS MODEL TRAINED ON DEGRADED SDCT IMAGES ...	69
5 DISCUSSION AND CONCLUSIONS	86
5.1. MAIN FINDINGS	86
5.2. WORK LIMITATIONS AND FUTURE DEVELOPMENTS	91
5.3. CONCLUSION	93
BIBLIOGRAPHY	94
LIST OF FIGURES	107
LIST OF TABLES	108

Graphical Abstract

Advanced AI Frameworks for Lung Nodule Analysis: From Synthetic Low-Dose CT Simulation to Nodule Detection and Classification



Main goal The goal of this thesis is to advance lung cancer screening using AI frameworks for lung nodule analysis focusing on low-dose CT images.

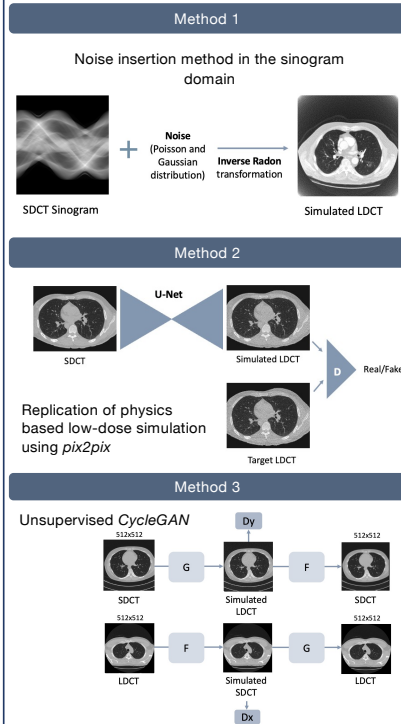
Literature gap (Chapter 1)

- Lack of reproducible CT degradation methods to simulate low-dose characteristics present in LDCT images used for lung cancer screening
- Lack of annotated LDCT images to support the development of robust AI models for alleviating radiologist workload

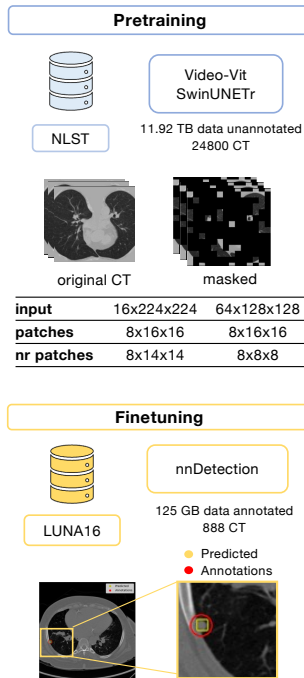
Objectives (Chapter 1)

- ❖ Investigate and simulate LDCT noise characteristics for data augmentation
- ❖ Leverage 3D transformers as self supervised pretraining method for lung nodule detection
- ❖ Develop robust and explainable radiomics-based lung nodule malignancy classifier

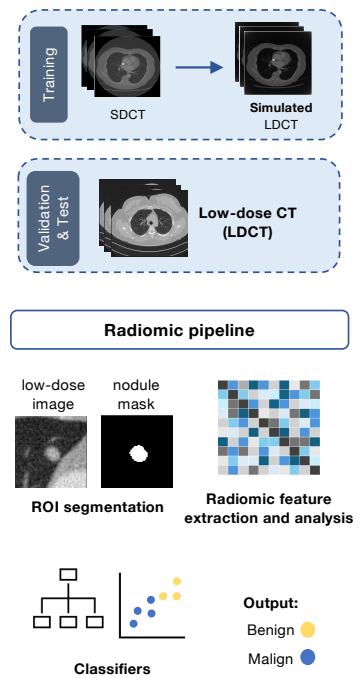
I. Degradation method comparison (Chapter 2)



II. 3D transformer self-supervised pretraining for lung nodule detection (Chapter 3)

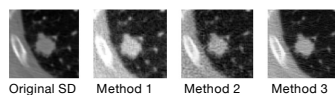


III. Lung nodule classification using radiomic model trained on degraded SDCT images (Chapter 4)



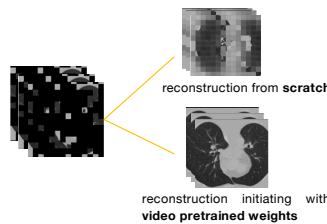
Results

I. Generating synthetic LDCT data from SDCT using unpaired CycleGAN offers a promising data augmentation strategy.



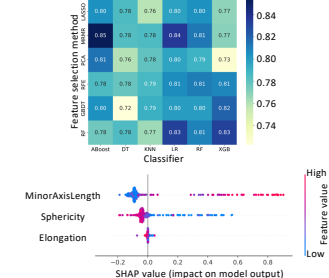
	FID	KID (mean ± std)
SDCT	1.1494	1.3695 ± 0.5763
Method 1	4.0717	3.0200 ± 0.5110
Method 2	4.6729	8.4469 ± 1.1596
Method 3	0.1734	0.0813 ± 0.1002

II. Treating 3D medical scans as video sequences is possible to adapt self-supervised learning techniques from the natural domain to the medical domain.



Leveraging vast amounts of unlabeled data has potential to improve lung nodule detection.

III. Utilizing degraded SDCT images is a viable strategy for development of robust and explainable nodule classification models for screening-detected nodules.



Abstract

Lung cancer remains the leading cause of cancer-related mortality worldwide. Screening subjects with low-dose computed tomography (LDCT) has been shown to reduce mortality. However, its implementation faces significant challenges, including the high workload for radiologists and the scarcity of large, annotated datasets required to train robust artificial intelligence (AI) support systems. This PhD thesis addresses these limitations by proposing methodologies that leverage synthetic data generation, self-supervised learning, and explainable radiomics to improve the lung cancer screening pipeline.

First, to overcome the scarcity of annotated LDCT data, this work investigated the feasibility of generating synthetic LDCT from standard-dose computed tomography (SDCT). A comprehensive comparison of image-domain and sinogram-domain degradation techniques was conducted. The results demonstrated that synthetic LDCT images generated via unpaired Generative Adversarial Networks (GANs) are statistically comparable to real clinical acquisitions, offering a scalable, cost-effective solution for data augmentation without requiring paired data that would increase patients' radiation exposure.

Second, to address nodule detection, the thesis introduced a novel approach that treats 3D volumetric computed tomography (CT) scans as video sequences. The study adapted the Video-MAE self-supervised framework from the natural video domain to the medical domain. Transformer-based backbones (Video-ViT and SwinUNETR) were pretrained on unlabeled chest LDCT images and subsequently fine-tuned for 3D nodule detection using the state-of-the-art nnDetection framework. This method demonstrated that leveraging vast amounts of unlabeled data through self-supervised learning significantly improves detection performance (average precision at IoU 0.1 82.6%), even when annotated medical data are limited.

Third, regarding nodule malignancy classification, the research developed an explainable radiomics-based framework designed for screening-detected nodules. A key contribution was the validation of a cross-domain training strategy using the LIDC-IDRI dataset: training classifier models exclusively on synthetic LDCT images and testing them on real, unseen LDCT data (tube current ≤ 80 mAs). The proposed model, utilizing interpretable radiomic features, achieved robust performance (AUC 0.874) on the test set, successfully distinguishing malignant nodules from benign ones.

This thesis confirms that utilizing degraded SDCT images is a viable and effective strategy for developing generalizable lung cancer screening tools. By combining synthetic image generation with advanced self-supervised detection and explainable classification, this work paves the way for the development of accurate, reproducible, and clinically trustworthy AI systems capable of supporting radiologists in the early detection of lung cancer.

Keywords: lung cancer screening, low-dose CT, synthetic image generation, lung nodule detection, lung nodule classification

List of Publications

The present thesis reports the following publications:

- i. **Liu J**, Corti A, Corino VD, Mainardi L. A comparative analysis of CT degradation for robust LDCT nodule classification using radiomics. IEEE Journal of Biomedical and Health Informatics. Jan 2026 (under review)
- ii. **Liu J**, Ma Q, Corti A, Corino V, Mainardi L, Konukoglu E. Leveraging self-supervised pretraining using transformers for enhanced lung nodule detection in CT scans. MICCAI 2025 MLMI Workshop. doi: 10.1007/978-3-032-09513-8_11
- iii. **Liu J**, Corti A, Corino VD, Mainardi L. Low-Dose CT-Based Radiomic Model to Identify Malignant Lung Nodules. In 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE) 2024 Oct 21 (pp. 570-575). IEEE. doi: 10.1109/MetroXRINE62247.2024.10797090
- iv. **Liu J**, Corti A, Corino VDA, Mainardi L. Lung nodule classification using radiomics model trained on degraded SDCT images. Comput Methods Programs Biomed. 2024 Dec;257:108474. doi: 10.1016/j.cmpb.2024.108474.

1 Cover Essay

1.1. Introduction

Lung cancer has been ranked first in both incidence (1.2 million cases) and mortality (1.1 million) since the first GLOBOCAN report in 2000 [1]. By 2022, lung cancer remained the most diagnosed cancer globally, accounting for approximately 2.5 million new cases, representing 12.4% of all cancers globally [2]. More significantly, it still stands as the leading cause of cancer-related death with an estimated 1.8 million deaths, representing 18.7% of total cancer mortality [2].

According to global surveillance data [3], the age-standardized 5-year net survival for lung cancer ranged from 10% to 20% in most countries during 2010-2014, including 21 European countries. Net survival measures the probability of surviving at least five years after diagnosis, assuming cancer is the only cause of death, while age standardization allows for unbiased comparisons between populations with different age demographics. This poor prognosis is primarily attributed to the asymptomatic nature of the disease in its early phases, leading to diagnosis at advanced stages, metastatic or distant, where curative treatment is no longer an option.

To address this diagnostic delay, research focused on screening high-risk individuals, such as smokers and ex-smokers. **Low-dose computed tomography (LDCT)** has emerged as the gold standard for this purpose. The efficacy of LDCT has been established through several randomized controlled trials, including the National Lung Screening Trial (NLST) in the USA and the Dutch-Belgian Randomized Lung Cancer Screening Trial (NELSON) in Europe [4], [5]. The NLST trial [4] occurred between 2002 and 2004, involved 53,454 people and was the first to demonstrate a significant mortality benefit. The trial showed a **20% reduction in lung cancer deaths** in the group that received LDCT screening compared to the group that received X-ray imaging screening. Subsequently, the NELSON study [5], conducted from 2004 to 2015, was the largest European trial with a sample size of 15,792 participants and confirmed the previous findings. At 10 years of follow-up, the study reported a cumulative rate ratio for lung cancer death of 0.76 (95% CI, 0.61 to 0.94) in the screening group compared to the control

group, corresponding to a **24% reduction in lung cancer mortality**. Notably, the study found a more pronounced benefit among women, with a reported rate ratio of 0.67 at 10 years, representing a **33% reduction in mortality**.

Besides the two aforementioned trials, other smaller European trials were also conducted to report the effect of LDCT screening on lung cancer mortality, including the Danish DLCST [6], German LUSI [7], Italian MILD [8] and ITALUNG [9], [10], for a total of 17,911 trial participants. In all studies, LDCT screening was compared to a control arm without screening intervention (total n = 8,577). Although these smaller trials individually lacked the statistical power to demonstrate significant reductions in lung cancer mortality, their collective findings support a moderate mortality reduction consistent with larger studies [7]. This accumulating evidence has driven policy changes, including the European Commission's "Europe's Beating Cancer Plan" and the introduction of national screening programs in countries such as Australia, which began in July 2025 [2]. In the US, screening is already implemented and included in the health plan as the US Preventive Services Task Force recommends annual LDCT for adults aged 50-80 with a 20-pack-year¹ smoking history who currently smoke cigarettes or who quit within the past 15 years.

Despite the proven benefit of lung cancer screening, translating trial results into large-scale population screening has been slow and difficult. First, high rates of false positives and overdiagnosis lead to unnecessary follow-up procedures and increased healthcare costs. Second, the screening process carries a significant psychological burden, specifically the waiting period associated with indeterminate results, which induce anxiety and distress [10]. Third, resource constraints in terms of costs and infrastructures are required, along with the limited number of physicians and radiologists needed to analyze the screened exams. This challenge is particularly acute in Asia, where lung cancer incidence and mortality are the highest compared with the West. Many Asian countries face additional challenges in implementing LDCT screening due to economic constraints, limited early-detection efforts, and a lack of dedicated government programs [11].

To address these challenges, **computer-aided diagnosis (CAD)** systems powered by **artificial intelligence (AI)** have become pivotal. These tools are essential not only for making screening cost-effective but also for enhancing clinical accuracy. The interpretation of scans, particularly

¹ 1 pack year is equal to smoking 1 pack per day for 1 year, or 2 packs per day for half a year.

the characterization of small indeterminate nodules, heavily depends on the radiologist's professional experience, leading to inter-observer variability and high false-positive rates.

The development of AI tools has been accelerated by large-scale community initiatives, such as the LUNA16 challenge [12] in 2016 and the Data Science Bowl in 2017², which focused on developing algorithms for nodule detection. Most recently, the LUNA25 challenge [13] in 2025 shifted the focus to malignancy classification. Particularly, this challenge was also designed to compare the performance between AI algorithms and radiologists. Results demonstrated that AI systems can achieve performance superior to human readers in classifying screened nodules (AUC 0.781³ vs 0.690) [13].

These findings suggest that AI algorithms represent a promising supportive approach to lung cancer screening, as they have the potential to (1) reduce false positives, lowering the psychological burden and financial costs associated with overdiagnosis; (2) standardize analysis, ensuring consistent detection and classification regardless of variations in image quality or radiologist experience; (3) optimize resources, alleviating the increased workload of radiologists to facilitate the implementation of population-wide screening globally.

Therefore, this thesis aims to advance the lung cancer screening field by addressing three critical pillars of the screening workflow:

- (i) mitigating low-dose image degradation** by investigating low-dose simulation techniques to leverage existing SDCT datasets, thereby overcoming the lack of annotated LDCT datasets;
- (ii) enhancing automated nodule detection** by developing a detection model with minimal supervision, i.e., minimal reliance on manually annotated labels, by adapting state-of-the-art self-supervised learning techniques to bridge the domain shift from natural video sequences to 3D medical imaging;
- (iii) improving radiomics-based malignancy classification** by building a robust lung nodule malignancy classifier specifically for screening detected nodules, aiming to improve current state-of-the-art performance metrics.

² <https://www.kaggle.com/c/data-science-bowl-2017> (accessed on 8/12/2025)

³ LUNA25 challenge leaderboard <https://luna25.grand-challenge.org/evaluation/closed-testing-phase/leaderboard/> (accessed on 8/12/2025)

1.2. Background and state-of-the-art

1.2.1. Image degradation and restoration in computed tomography

Low dose computed tomography

Computed tomography (CT) is a non-invasive image modality and it is considered the most sensitive imaging technique for detecting pulmonary nodules, widely used in screening for both primary lung cancer and metastases [14]. Despite its diagnostic efficacy, the use of ionizing radiation in CT poses health risks that could lead to the development of metabolic abnormalities and genetic disorders, which increase the risk of cancer. Consequently, the clinical application of CT is governed by the ALARA (as low as reasonably achievable) principle to minimize radiation exposure to patients [15]. This principle drives the increasing adoption of LDCT for early screening as it provides the necessary diagnostic capability while reducing the risk of cumulative radiation exposure compared to **standard-dose CT (SDCT)**.

The factors affecting dose in CT include tube voltage (kVp), tube current (mA) and tube rotation speed (s) [16]. A straightforward approach to reduce dose is by lowering the tube current (often measured as the current-time product, mAs), as it is linearly proportional to radiation dose [17]. Thus, when the tube voltage and rotation speed are constant, reducing the tube current leads to a proportional reduction in radiation exposure, making it the most effective parameter for dose reduction [18]. Supporting this observation, various LDCT protocols show that many studies maintain a fixed tube voltage (typically 120 kVp) and rotation speed (e.g., 0.5 s) while adjusting the tube current (see **Table 1**). In the context of maintaining image quality, Mayo et al. [19] investigated the minimum tube current required and reported 140 mA as the lower limit for good quality images.

The literature indicates a lack of consensus on the dose level that defines low dose. **Table 1** summarizes some LDCT protocols across studies in the literature, highlighting this lack of standardization. Gartenschläger et al. [14] compared a standard protocol (200 mAs) with a low-dose protocol (30 mAs) to evaluate nodule detection. They concluded that low-dose spiral CT maintained high sensitivity and advocated its use when dose minimization is clinically necessary. Their review of other works further demonstrated a broad range, citing 20-140 mAs for low dose and 140-400 mAs for standard dose. Karabulut et al. [18] focused on diagnostic accuracy, comparing a standard 200 mA protocol against a 50 mA low-dose protocol (both at

120 kVp). They found no statistically significant differences in the number of nodules detected, concluding that the LDCT protocol provided adequate quality for reliable nodule detection or exclusion. Karabulut et al. [18] reported values used in other LDCT studies ranging from 10 to 80 *mA*. Kube et al. [20]’s investigation compared low-dose CT (50 *mAs*) to SDCT (150 *mAs*). The reviewers assessed a wide range of findings, including nodules (>5 mm), concluding that the 50 *mAs* protocol produced results consistent with the standard dose, further supporting the routine use of low-dose chest CT. Bhalla et al. [21]’s review noted that typical LDCT is performed at 120 *kVp* and 40-80 *mAs*, but also reported ranges as low as 80-120 *kVp* and 20-40 *mAs* for specific chest indications. In summary, standard protocols typically range from 140 to 400 *mAs*, whereas LDCT protocols reported in the literature vary significantly, often falling between 20 and 140 *mAs*.

Considering the screening trials,

Table 2 summarizes the ranges used in the LDCT protocols. The reported settings typically range from 20-80 *mAs* and 80-140 *kVp*, which are adjusted according to the patient’s body mass index.

Table 1 – Comparison of SDCT and LDCT scanning protocols from selected studies in the literature.

Study	SDCT			LDCT		
	kVp	mAs	rotation (s)	kVp	mAs	rotation (s)
Gartenschläger et al. [14]	120	200 140-400	1	120	30 20-140	1
Karabulut et al. [18]	120	200 mA	-	120	50 mA 10-80	-
Kube et al. [20]	120	150	0.5	120	50	0.5
Bhalla et al. [21]	100-140	130-200	0.5	80-120	20-40	0.5

Table 2 – LDCT protocols employed in lung cancer screening trials.

Trial	kVp	mAs
NLST [22]	120-140	40-80
NELSON [23]	80-140	20
MILD [23]	120	30
ITALUNG [24]	120-140	20-43

Low-dose CT image restoration

Despite the clinical advantages of the LDCT, reducing the radiation dose compromises image quality. As the incident flux decreases, the detector experiences photo starvation, which increases noise levels and lowers the **signal-to-noise ratio (SNR)**. Compared to SDCT images, LDCT images are characterized by increased quantum noise, streak artifacts, reduced contrast, and blurring (see **Figure 1**). These degradation features can affect clinicians' diagnostic results, potentially reducing diagnostic value [17]. To mitigate these effects, research has focused on **LDCT image restoration**, aiming to recover high-quality diagnostic information without increasing radiation dose.

Before image restoration, to obtain a CT image, the process (see **Figure 2**) begins with acquisition in which the X-ray tube rotates around the patient and the detectors measure the X-ray attenuation. These raw measurements, referred to as projection data, are stored in a format called a sinogram by vertically stacking all the projection data, one for each angle. Finally, a reconstruction algorithm such as filtered back-projection transforms each slice of the sinogram into a 2D image. A stack of 2D slices then forms the final 3D volume. Based on this process, LDCT image restoration methods can be categorized into three groups: **(1) sinogram domain filtering**, applied directly to raw projection data; **(2) iterative reconstruction**, alternating

between the sinogram and image domains; and **(3) image domain restoration**, a post-process method applied directly to the image.

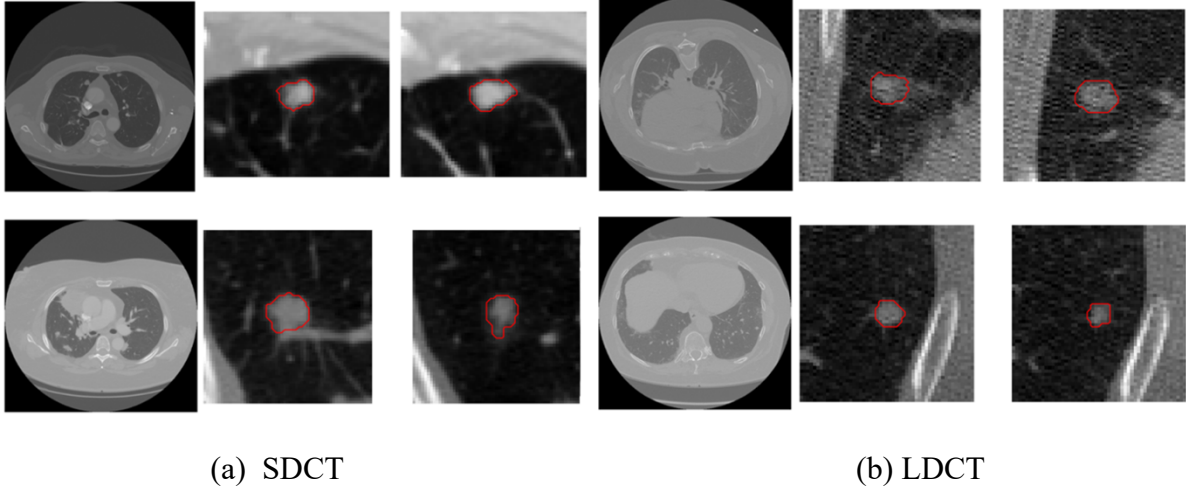


Figure 1 - Examples of SDCT (a) and LDCT (b) subjects from the LIDC-IDRI dataset and examples of the corresponding nodule annotations. LDCT images are characterized by increased quantum noise, streak artifacts, reduced contrast, and blurring. Adapted from [25].

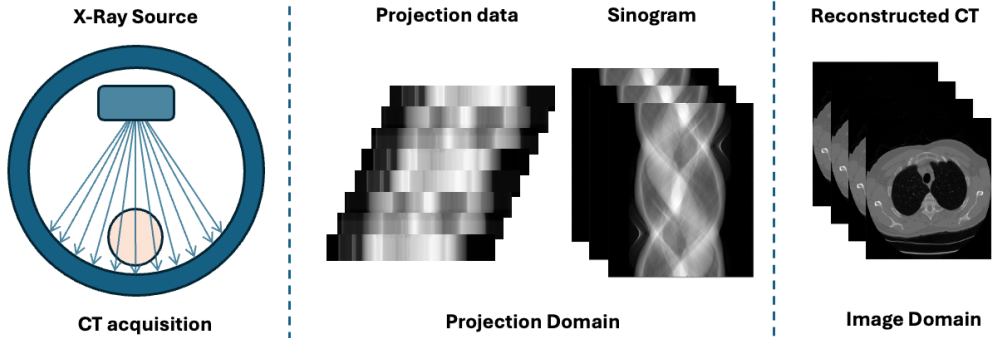


Figure 2 – Diagram of CT imaging. Adapted from [26].

Sinogram domain filtering is applied to raw projection data by filtering the noise directly in the sinogram domain before image reconstruction. Some sinogram domain filtering methods include structural adaptive filtering [27], [28], bilateral filtering [29], and penalized likelihood method [30]. However, a limitation is that projection and sinogram data are often vendor specific which is often inaccessible to many researchers.

Iterative reconstruction alternates between the sinogram and image domains, using mathematical models to correct noise and artifacts until an optimal image is formed. Examples of prior iterative reconstruction methods include non-local means [31], total variation [32], dictionary learning [33], and low-rank approximation [34], which can produce high quality but at the cost of high computational expense. Furthermore, some iterative reconstruction algorithms have been reported to cause content loss [35].

Finally, **image domain restoration** is a post-processing method that directly applies to reconstructed images to reduce noise. This domain is divided into traditional algorithms and learning-based methods. Traditional image domain methods include algorithms such as non-local means [36], [37], [38], total variation [39], block matching and 3D filtering [40], and statistics-based algorithms [41]. Even though these methods are flexible because they can be applied to any reconstructed DICOM image without requiring raw scanner data, they often struggle to distinguish between noise and anatomy, leading to edge blurring and artifacts in areas with uniform intensity, such as soft tissues. Recently, rapid advancements in deep learning (DL) have established it as the state of the art in medical imaging and LDCT restoration [17], [35], [42]. Unlike traditional filters, DL models learn complex anatomical patterns and noise characteristics from large datasets, allowing for superior noise suppression and detail preservation. However, the primary challenge in applying these data-driven DL techniques to medical imaging lies in the availability and nature of the training data. Consequently, DL models for LDCT image restoration can be categorized by learning strategy: **supervised, unsupervised and self-supervised**.

Supervised learning remains the most common strategy for LDCT denoising, where a model learns a direct mapping from a noisy LDCT input to its corresponding SDCT target. The majority of proposed supervised methods focus on three main areas of improvement: (1) altering the network structure (e.g., introducing residual or dense connections), (2) optimizing the loss function, or (3) refining the training strategy [17]. Model performance is typically evaluated using the **peak signal-to-noise ratio (PSNR)** and the **structural similarity index (SSIM)**; other less commonly used metrics include **root mean squared error (RMSE)** and **feature similarity index (FSIM)**. (For detailed metric definitions, please refer to *Metrics to evaluate denoising and degradation*.) However, training supervised models requires paired datasets for training, i.e., the same slice in both LDCT and SDCT domains, which is ethically

restricted by the ALARA principle to expose patients to repeated CT scans. To overcome this limitation, some researchers create synthetic LDCT images by degrading existing SDCT data.

In general, CT images are degraded by quantum noise and various artifacts during low-dose acquisition. Quantum noise arises from X-ray photon starvation, a direct consequence of reduced radiation dose. As mentioned before, visually, this results in the smoothing of subtle structures, loss of contrast, and the presence of striking artifacts. In mathematical modeling and simulation, the noise distribution in CT is characterized by (1) quantum noise approximated by a Poisson distribution and (2) electronic noise, best modeled using a Gaussian distribution [35]. Consequently, some applications consider the mixed Poisson-Gaussian distribution [43]. To better simulate realistic noise, several studies have injected noise into the sinogram, and then the simulated low dose image is reconstructed to create realistic training pairs. A typical simulation pipeline involves: (1) performing a Radon transform (forward projection) on an SDCT image to get the corresponding sinogram; (2) adding Poisson noise and Gaussian noise to the sinogram; and (3) applying an inverse Radon transform to reconstruct the synthetic LDCT image.

Several studies have validated the sinogram-domain approach. Huang et al. [44] utilized this pipeline to train a two-stage residual convolutional neural network (CNN), achieving a PSNR of 31.06, a SSIM of 0.8812 and a MSE of 54.15 on the *2016 NIH-AAPM-Mayo Clinic LDCT Grand Challenge*. Similarly, Chen et al. [45] utilized Siddon's ray-driven algorithm to generate simulated normal-dose sinograms, training a CNN framework that directly mapped simulated LDCT to SDCT, reporting a PSNR of 42.15 and an SSIM of 0.97. To further improve performance on ultra-low-dose images, Zhao et al. [46] used a validated pipeline [47], [48] to simulate ultra-low-dose images and trained a CNN with residual mapping, achieving an 8 dB improvement in the PSNR metric and an increase of 0.15 in the SSIM metric on the ultra-low-dose images.

Beyond standard architecture, Gholizadeh-Ansari et al. [49] generated training data similarly to that of Huang et al. [44] and introduced a network using dilated convolutions and a combination of mean-square error and perceptual loss function. While they achieved robust quantitative results across multiple datasets (e.g., PNSR 35.57 and SSIM 0.85 on the simulated lung dataset; PNSR 44.12 and SSIM 0.98 on the piglet dataset and PSNR 31.50 and SSIM 0.55 on the thoracic dataset), they highlighted a trade-off between the losses. While perceptual loss

leads to a more visually appealing outcome, it results in lower PSNR values compared to models trained solely on mean-square error. The reason is that high PSNR results from minimizing per-pixel loss. If a network is trained to minimize MSE, it will always have a higher PSNR than one trained to minimize perceptual loss. While most models were trained from scratch, Zhong et al. [50] proposed a transfer learning residual CNN to compensate for the limited availability of medical data. They first trained their model on natural images corrupted by Gaussian noise and then finetuned it on CT images simulated with Poisson and Gaussian noise in the sinogram domain [51]. They reported the following results: 48.16 ± 0.57 (PSNR), 0.9993 ± 0.0652 (RMSE), 0.9862 ± 0.0016 (SSIM), and 0.9916 ± 0.0008 (FSIM).

The introduction of **generative adversarial network (GAN)** by Goodfellow et al. [52] has advanced the field of image denoising. A GAN consists of two competing neural networks: a generator and a discriminator (see **Figure 3**). The generator aims to create realistic images from input data, while the discriminator behaves as a binary classifier that distinguishes real images from generated fake images. During training, the generator optimizes its weights to generate realistic images that fool the discriminator, while the discriminator learns to detect them. The generator's weights are updated through backpropagation based on the discriminator's classification loss. Ideally, upon convergence, the generator produces reconstructions of such high quality that the discriminator can no longer tell the difference, thereby reducing its classification accuracy. In the context of CT denoising, the generator learns to transform noisy LDCT inputs into denoised images that are statistically indistinguishable from real SDCT scans. In the domain of supervised learning, Pix2Pix [53], originally designed for natural image translation, emerged as a foundation framework. It employs a conditional GAN that learns the mapping from source to target using aligned image pairs, allowing the generation of realistic images indistinguishable by humans. In medical imaging, Wolterink et al. [54] trained a CNN with an adversarial loss to accurately capture the statistical properties of SDCTs. Their work demonstrated that adversarial network allows the generator to better learn the noise distribution in SDCT.

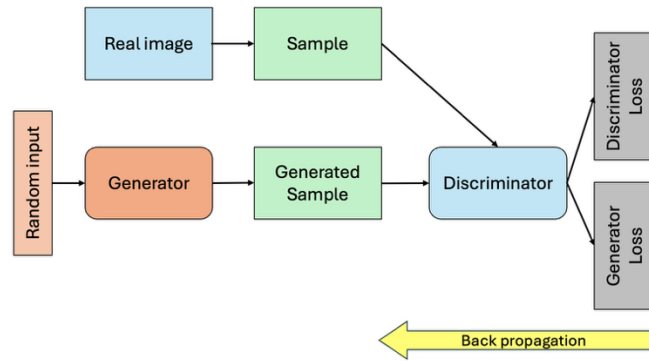


Figure 3 – GAN architecture consists of a generator that aims to create realistic images from the input data and a discriminator that attempts to distinguish real images from generated fake images. Adapted from [55]

While the aforementioned studies relied on custom low-dose simulation, the majority of supervised CT denoising utilizes the *2016 NIH AAPM-Mayo Clinic LDCT Grand Challenge* dataset. As one of the few publicly available datasets for this task, it has become the benchmark for evaluating denoising and iterative reconstruction algorithms. The original challenge training dataset consists of abdominal CT images from 10 patients, paired as full-dose (or standard dose) and quarter-dose (or low dose) filtered back projection images. Following the challenge, an extended version was released in 2020 in the Cancer Image Archive (TCIA) named *Low-dose CT and Projection data* [56]. This extended version includes 299 cases, including 99 head, 100 chest and 100 abdominal CT scans. All CT scans were acquired at standard dose level, and for each clinical case, a corresponding simulated lower dose level projection dataset was provided. Examples of studies that used this dataset include: [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69]. Since authors of each method use their own training, validation, and test splits, performance variability across the literature is significant, with PSNR values ranging from 22.31 to 47.90 dB, RMSE values ranging from 0.02 to 15.04, and SSIM values ranging from 0.69 to 0.98 [70]. These wide ranges reflect differences in image quality and fidelity and suggest discrepancies in the accuracy of reconstructed images. While SSIM values also span a broad range (0.69 to 0.98), this metric highlights the varying degrees of success models have in preserving anatomical structures.

Despite being the standard reference dataset used in denoising research, it is important to note that even these benchmarks are generated using validated noise insertion methods rather than clinically acquired paired data [12], [17]. Therefore, as paired clinical data is not publicly

available, the field of denoising relies on synthetic approaches. While these approaches could be effective in benchmarks, there are two drawbacks: (1) the noise simulation is often treated as a secondary step, potentially compromising the fidelity of the synthetic images and the real-world generalizability of models trained on them; and (2) the cycles of synthetic degradation and subsequent denoising are computationally expensive and time-consuming during both development and clinical implementation. Therefore, alternative training approaches without paired datasets should be explored.

To address the scarcity of paired data, **unsupervised learning** has emerged as an important learning strategy. Unlike supervised methods, unsupervised approaches focus on learning the statistical distribution of a set of clean images (domain A) and noisy images (domain B) without requiring a 1:1 correspondence. This allows models to be trained using SDCT scans from one group of patients and LDCT scans from a different group, expanding the availability of usable training data.

The most prevalent unsupervised framework for unpaired translation is **Cycle-consistent GAN (CycleGAN)**, proposed by Zhu et al. [71]. Distinct from the GAN previously described, CycleGAN is composed of two generators (one mapping Domain A to B, and another B to A) and two discriminators (one to distinguish generated A from real A and another to distinguish generated B from real B). Additionally, to ensure structural integrity without paired data, CycleGAN introduces a cycle-consistency loss, which requires that the generated image when translated back to the source domain should match the original input. For instance, Kang et al. [61] have used CycleGAN to ensure that an image translated from LDCT to SDCT and back again remains identical to the original input. This constraint is critical for medical imaging as it prevents the network from altering underlying anatomy. In their work they compared the SNR between the original noisy and denoised images and showed that the denoised images had higher SNR ($p < 0.005$). Recent literature has focused on improving the efficiency and structural fidelity of the GAN models, including the work of Gu and Ye [72] who proposed a tunable CycleGAN using a single generator with adaptive instance normalization (AdaIN). Instead of training two separate generators, the same baseline generator converting a LDCT image to a SDCT image can be switched to a generator converting SDCT to LDCT by simply changing the AdaIN code. Consequently, this architecture minimizes memory requirements and stabilizes training on small datasets (PSNR 30.8730 dB, SSIM = 0.6605).

Beyond adversarial networks, some researchers developed unsupervised denoise models based on statistical estimators. For instance, Kim et al. [73] leveraged the physical properties of the CT acquisition pipeline and introduced a framework that selects the objective function based on the data domain, i.e., using the Poisson unbiased risk estimator (PURE) as loss function to train deep neural networks on noisy raw measurements, and using weighted Stein’s unbiased risk estimator (WSURE) as objective functions for sinograms (log-transformed raw data). Their physics-driven approach yielded PSNR of 41.7 dB and SSIM of 0.978 using PURE, and PSNR of 37.4 dB and SSIM of 0.927 using WSURE, on simulated low-dose images derived from the *2016 LDCT Grand Challenge*.

Distinct from pure statistical estimators, **self-supervised learning** methods enable models to learn denoising directly from LDCT data itself. This strategy leverages the statistical properties of the noise, which is random and uncorrelated, and the characteristics of the anatomy, which is redundant and structured.

The conceptual foundation of many SSL approaches in natural image restoration is Noise2Noise, introduced by Lehtinen et al. [74], which demonstrated that a network can be trained using pairs of noisy observations of the same scene. Theoretically, if the noise has a zero-mean distribution, the L2 loss function will cause the network to converge towards the mean value of the observations, effectively recovering the clean signal. While Noise2Noise doesn’t require clean targets, it still requires paired noisy observations of the same structure, which can be challenging in clinical environments. To address this challenge, Noise2Void [75] and Noise2Self [76] were proposed to train only by using single noisy images. These models predict a center pixel based on its local neighbors, achieving promising results using single noisy images assuming independent noise among neighbor pixels. However, because CT noise is often correlated or structured due to the reconstruction process, these methods cannot suppress correlated or structured noise. To address this, Niu et al. [77] proposed Noise2Sim, a non-local similarity-based approach. By identifying similar patches across the image as internal pairs, Noise2Sim behaves like a supervised method under mild conditions and suppresses correlated noise more effectively than local blind-spot methods. In the *2016 LDCT Grand Challenge*, the method achieved a PSNR of 28.38 ± 1.51 and a SSIM of 90.45 ± 2.76 .

Another approach for single-image training is iterative data refinement (IDR), proposed by Zhang et al. [78], which only requires a single noisy image and a noise model to train a

denoising neural network. It works by iteratively refining the training set with generated noisy data. However, the framework is computationally expensive as it requires multiple rounds of training on a fixed and known model, limiting its effectiveness against the complex, non-stationary noise found in clinical CT images.

Recent research has integrated classical image processing constraints, such as edge-preserving filters and histogram equalization, with DL to improve parameter efficiency and structural fidelity. Wagner et al.[79] introduced a differentiable bilateral filter layer into the gradient flow, achieving a PSNR of 43.07 dB. Similarly, Choi et al.[80] proposed multi-view projection synthesis for sparse-view CBCT. By restoring sinograms in the projection domain before reconstruction, this dual-domain approach reached a PSNR of 37.63 dB, demonstrating the value of addressing noise before it is transformed into image-space artifacts.

As the field moves toward more complex architectures, attention networks have become popular for LDCT restoration. Li et al. [81] introduced self-attention modules to leverage structural relationships both within and between CT slices. This volumetric context allows the model to preserve features regardless of distance, achieving a PSNR of 27.73 dB and SSIM of 0.8876 on the AAPM-Mayo dataset. Similarly, Jing et al. [82] integrated attention blocks into the decoding phase of a self-supervised framework, further demonstrating that attention networks are flexible for real-world clinical scenarios, as they bypass the logistical burden of collecting paired datasets entirely.

Despite the acceptable visual improvements achieved in CT restoration using current attention-based DL methods, Kulathilake et al. [35] note that attention-based models often yield sub-optimal quantitative results as measured by PSNR and SSIM. The main reason is the lack of importance given to the structural feature preservation and the reliance on pixel-wise loss functions like MSE during training, which can lead to smoothing or structure deformation. Hence, balancing high PSNR and SSIM metrics with the preservation of diagnostic anatomical integrity remains a primary challenge for future unsupervised research.

From denoising to image degradation

Over the past few decades, the problem of image denoising in computer vision has been extensively studied. However, as Zainulina et al. [83] concluded in their work, artificially

adding noise to images can bias the denoising model's predictions, since accurate noise simulation requires a comprehensive understanding of the actual CT system and may not be feasible at times. In fact, noise in LDCT measurement data is influenced by many factors, such as quantum noise, the logarithmic transformation of measurements, and pre-reconstruction corrections for system calibration, making modeling the noise in the reconstructed images particularly challenging. Therefore, as suggested in [84], there is a need for more sophisticated noise simulation approaches to bridge the gap between simulated and real-world data in CT image denoising applications.

Building on the simulation techniques discussed in the supervised strategy section, it is important to highlight that accessible and reproducible CT noise simulation methods are scarce in the literature. In this context, two CT degradation methods are particularly relevant: (1) a sinogram-domain simulation proposed by Zeng et al. [51] due to its feasibility as it is applied directly on SDCT images, and (2) a validated physics-based noise modeling technique from Yu et al.[85], serving as a reference as it was used to generate the paired benchmarks for the *2016 LDCT Grand Challenge*.

Briefly, for the Zeng et al.'s approach, a standard-dose sinogram data is first obtained by applying the Radon transform to the original SDCT images; subsequently, simulated Poisson and Gaussian noise are added to mimic the effects of electronic and quantum noise, generating low-dose sinogram data. Equation 1 describes the calculation for the low-dose sinogram data ($p_{ld,sim}$):

$$p_{ld,sim} = \log \frac{I_{ld,sim}^0}{\text{Poisson}(I_{ld,sim}^0(s) * \exp(-p_{sd})) + \text{Gaussian}(m_e, \sigma_e^2)} \quad (1)$$

where $I_{ld,sim}^0$ is the simulated low-dose scan incident flux, p_{sd} is the standard-dose sinogram data obtained from the original SDCT images using Radon transformation, m_e and σ_e^2 are the mean and variance of the electronic noise, respectively. Specifically, $I_{ld,sim}^0$ is computed by multiplying the incident flux by the ratio between the high and simulated low-dose incident flux, k . This ratio is computed based on the proposed equation:

$$k(mAs) = a * (mAs) + b \quad (2)$$

where $a=0.0095$ and $b=0.0309$.

On the other hand, the simulation technique used to generate the *2016 LDCT Grand Challenge* is the validated physics-based noise modeling technique from Yu et al. [85]. This technique incorporates the effect of bowtie filters, automatic exposure control and electronic noise. The noise insertion method requires scaling the incident photon number and adding normally distributed stochastic noise, as described in their equation:

$$\widetilde{P}_B \approx P_A + \sqrt{\frac{1-a}{a} \cdot \frac{\exp(P_A)}{N_{0A}} \cdot \left(1 + \frac{1+a}{a} \cdot \frac{N_e \cdot \exp(P_A)}{N_{0A}}\right)} \cdot x \quad (3)$$

where $\widetilde{P}_B$ is the simulated low dose data, P_A is data at a higher tube current setting, a ($0 < a < 1$) is a dose scaling factor, N_{0A} is the incident photon flux corresponding to P_A , x is a normally distributed stochastic process with zero mean and a unit variance, N_e represents the electronic noise. However, directly reproducing this model requires access to raw projection data, incident flux calibration and detailed scanner-specific variables, which are often inaccessible.

Metrics to evaluate denoising and degradation

The performance of denoising and degradation models is evaluated using different metrics depending on the availability of paired data. For instance, when LDCT and SDCT are available from the same patient, the most used metrics in the literature are **PSNR and SSIM** [86].

PSNR measures the ratio of the maximum possible power of the signal to the power of the corrupting noise. Given a reference image x and the generated image y , both of size $M \times N$, and with maximum intensity value MAX , the PSNR is defined as:

$$PSNR(x, y) = 10 \log_{10} \left(\frac{MAX^2}{MSE(x, y)} \right) \quad (4)$$

where MSE is defined as:

$$MSE(x, y) = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (x_{ij} - y_{ij})^2 \quad (5)$$

A higher PSNR indicates higher image quality and superior noise suppression, since the PSNR value approaches infinity as the MSE approaches zero.

To measure the similarity between two images, the **SSIM** metric is computed. It is considered to be correlated with the quality perception of the human visual system. SSIM considers three components: luminance (l), contrast (c) and structure (s) and is defined as:

$$SSIM(x, y) = l(x, y)c(x, y)s(x, y) \quad (6)$$

where the components are defined as:

$$\begin{cases} l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \\ c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \\ s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \end{cases} \quad (7)$$

Here, μ is the mean pixel value, σ^2 the variance and C are positive constants used to avoid a null denominator. SSIM index is a value between 0 and 1, where higher values indicate better quality and 1 represents perfect structural similarity.

In the context of degradation and noise simulation, the **multi-scale SSIM (MS-SSIM)** [87] is preferred over SSIM. Since CT noise and artifacts manifest at different spatial frequencies, a single-scale metric such as SSIM fails to capture the noise accurately. MS-SSIM calculates structural similarity at multiple downsampled resolutions, providing a robust evaluation of the simulated noise. It is computed by iteratively applying a low-pass filter and downsampling the filtered image by a factor of 2 for M scales. The original image is indexed as Scale 1 and the highest scale as Scale M , which is obtained after $M-1$ iterations. At the j -th scale, the contrast and the structure are calculated and denoted as $c_j(x, y)$ and $s_j(x, y)$, respectively. Instead, the luminance is computed only at Scale M and is denoted as $l_M(x, y)$. The overall SSIM evaluation is obtained by combining the measurements at different scales using:

$$MS - SSIM(x, y) = [l_M(x, y)]^{(\alpha_M)} \cdot \prod_{j=1}^M [c_j(x, y)]^{(\beta_j)} [s_j(x, y)]^{(\gamma_j)} \quad (8)$$

where the exponents $\alpha_M, \beta_j, \gamma_j$ are weights used to adjust the relative importance of each component. Similar to SSIM, MS-SSIM ranges from 0 to 1, with higher values indicating greater fidelity of the simulated noise to the reference image.

Additionally, the **noise power spectrum (NPS)** [88] is often used to evaluate noise simulations. Unlike PSNR and SSIM, the NPS analyzes the frequency content of the noise. Hence, NPS characterizes the noise texture, giving a better and more complete description of noise across the frequency domain. Considering regions of interests (ROIs) in 2D, the 2D NPS is computed as:

$$NPS_{2D}(f_x, f_y) = \frac{\Delta_x \Delta_y}{L_x L_y} \frac{1}{N_{ROI}} \sum_{i=1}^{N_{ROI}} |FT_{2D}\{ROI_i(x, y) - \overline{ROI_i}\}|^2 \quad (9)$$

where f_x, f_y represent the spatial frequencies, Δ_x, Δ_y are the pixel sizes in x and y dimension, L_x, L_y are the ROI's lengths in pixel, N_{ROI} is the number of ROIs used in the operation and $\overline{ROI_i}$ is the mean pixel value of the i -th ROI. The resulting spectral values are typically radially averaged to produce a 1D curve, allowing for a visual comparison of noise texture between simulated and real images.

In unsupervised scenarios, such as CycleGAN, when paired ground truth is unavailable, direct pixel-to-pixel comparison is not possible. Hence, distributional metrics such as the **Fréchet Inception Distance (FID)** [89] and the **Kernel Inception Distance (KID)** are computed to evaluate the quality of the generated images.

FID measures the distance between the feature distribution of the real and generated images. To compute this metric, a pre-trained InceptionV3 network is used to extract high-dimensional feature vectors from both image sets. These feature distributions are assumed to follow a multivariate Gaussian distribution, defined by a mean (μ) and a covariance matrix (Σ). The FID is calculated as the squared Fréchet distance using the following equation:

$$FID(x, y) = \|\mu_x - \mu_y\|_2^2 + Tr\left(\Sigma_x + \Sigma_y - 2(\Sigma_x \Sigma_y)^{1/2}\right) \quad (10)$$

Where μ_x, Σ_x represent the mean and covariance of the real data features; μ_y, Σ_y represent the mean and covariance of the generated data features. $\|\cdot\|_2^2$ denotes the squared Euclidean distance, Tr denotes the trace of the matrix (sum of diagonal elements). A lower FID score

indicates that the generated images are statistically more similar to the target distribution, translating into a higher perceptual realism.

KID is an alternative metric to FID as it measures the squared maximum mean discrepancy (MMD) between feature vectors. Unlike FID, which calculates a distance between fitted Gaussian distributions, KID calculates the MMD directly on the features using a polynomial kernel $k(x, y) = \left(\frac{1}{d}x^T y + 1\right)^3$, where d is the dimensionality of the feature vector. Therefore, KID, an unbiased estimator, is defined as:

$$KID(X, Y) = \underbrace{\frac{1}{m(m-1)} \sum_{i \neq j}^m k(x_i, x_j)}_{\text{Sum of real data}} + \underbrace{\frac{1}{n(n-1)} \sum_{i \neq j}^n k(y_i, y_j)}_{\text{Sum of generated data}} - \underbrace{\frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(x_i, y_j)}_{\text{Cross-term}} \quad (11)$$

Where $X = \{x_1, \dots, x_m\}$ and $Y = \{y_1, \dots, y_n\}$ are the sets of feature vectors obtained from the Inception network for the real images and generated images, respectively; m and n are the number of samples in the real and generated batches, respectively. The lower the KID value, the closer the generated distribution is to the real data distribution.

1.2.2. Lung nodule detection

Lung diseases include a wide range of pathologies affecting the thoracic cavities. Depending on the anatomical region affected (see **Figure 4**), these conditions can be categorized into: **airways diseases affecting the trachea**, such as asthma, chronic obstructive pulmonary disease (COPD), bronchitis and cystic fibrosis; **the alveoli**, such as pneumonia, tuberculosis, pulmonary edema; the area between the air sacs known as **the interstitium**, such as interstitial lung disease; and **the chest wall** conditions like pleural effusion and pneumothorax [90].

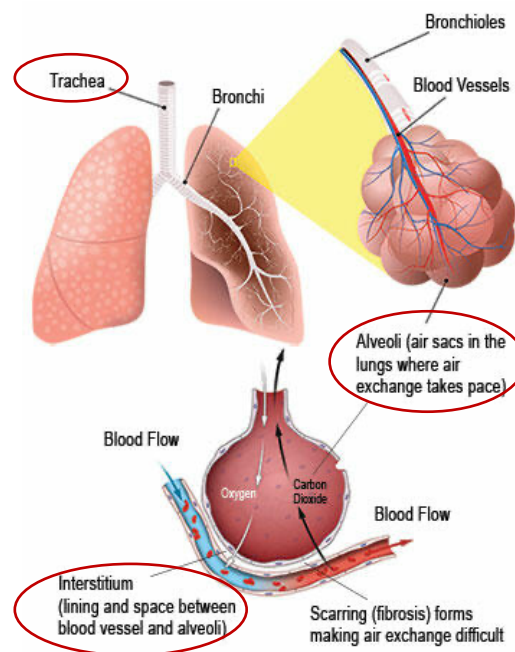


Figure 4 – Anatomical structures of the lung, including the trachea, bronchi, alveoli, and interstitium, highlighting the distinct regions that could be affected by different respiratory pathologies. Image from <https://www.froedtert.com/interstitial-lung-disease> (accessed on 22/01/2026).

Among these lung diseases, **the development of pulmonary nodules**, typically appearing in the form of irregular circular or oval growths inside the lung parenchyma, is the earliest manifestation of lung cancer. Consequently, detecting all pulmonary nodules is the critical first step in the lung cancer diagnostic workflow. Despite its importance, detection is a difficult task due to the complex, irregular and cluttered structure of the lungs (see **Figure 5**), from which nodules must be distinguished from a dense network of normal anatomical structures such as alveoli, blood vessels and bronchioles. As highlighted by [91], this complexity presents significant challenges for human readers. Radiologists often fail to identify all nodules and

inter-observer variability regarding the classification of suspicious findings is considerable [92], [93]. These limitations are enhanced under time pressure and the unknown number of nodules. **This subjectivity, combined with the risk of misdetection, highlights the necessity for objective and automated diagnostic tools.**

In clinical practice, nodules are neither uniform nor regular. Their morphology, size and shape for example, dictate their malignancy risk and subsequent management. Radiologists primarily classify nodules into two categories: **solid and subsolid** (see Figure 5). Solid nodules are dense and can obscure the underlying lung parenchyma. Subsolid nodules are divided into (a) pure ground-glass nodules which appear to be opaque and do not obscure the underlying vessels; and (b) part-solid nodules which contain both ground-glass and solid components.

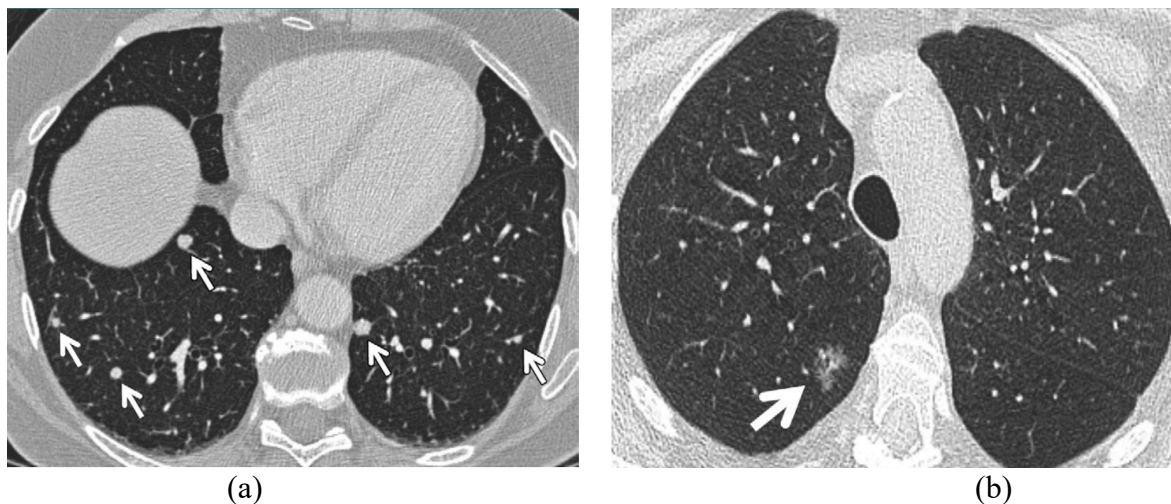


Figure 5 – Example of CT images showing (a) multiple solid nodules (arrow); (b) a part solid nodule (arrow), ground glass. Images from [94].

While subsolid nodules often grow more slowly than solid ones, they frequently correspond to pre-invasive adenocarcinoma (early stage of cancer where abnormal cells are confined to where they started) or early invasive adenocarcinomas (the cancer cells have broken through the original tissue and begun to grow into the surrounding tissues) [95]. From a diagnostic perspective, these stages are distinguished by their density on CT scans: pre-invasive lesions often appear as ground-glass opacities, while invasive growth is usually associated with a solid internal component. Subsolid nodules require careful long-term monitoring, as they often have a better overall prognosis when found early [96]. For instance, pre-invasive adenocarcinoma is associated with a nearly 100% 5-year survival rate after complete surgical resection [97]. Therefore, identifying these nodules is critical for the patient's survival.

The primary objective of lung cancer screening is to maximize the detection of early-stage cancers while minimizing false positives. However, this balance is difficult to maintain due to the complex anatomy of the lungs and the inherent difficulty in detecting small nodules at early stages. Retrospective analysis reveals that detection errors occur even for nodules that are clearly visible. The European Society of Radiology and the European Respiratory Society have stated that most screening studies require double reading by two radiologists to minimize errors [96]. However, this approach is both time-consuming and expensive. Furthermore, there is a shortage of radiologists trained specifically for lung cancer screening to ensure quality assurance of reading and avoid missed diagnosis or overdiagnosis which can result in overtreatment of insignificant [96]. As noted by Cui et al. [98], radiologists screening LDCT images face challenges of mechanical repetition and the easy omission of small nodules. These limitations have led to the integration of CAD and AI as second readers, as modern deep learning models address these issues by providing consistent identification criteria and high identification sensitivity.

Overview of lung nodule detection methods

The standard architecture for nodule detection consists of a two-stage pipeline: **(1) candidate detection**, in which all potential suspicious regions are identified and **(2) false positive reduction (FPR)**, where a classifier analyzes the candidates from the first step to distinguish true nodules from anatomical noise. In the candidate detection phase, architectures such as U-Net, YOLO and ResNet are often utilized, while 3D CNNs are commonly used in the false positive reduction stage. For instance, Ji et al.[99] proposed an improved YOLO-v5s based model incorporating attention modules and multi-scale contextual information. This model achieved mean average precision (mAP) and F1 scores of 0.720 and 0.740 on the LUNA16 dataset, and 0.794 and 0.766 on the X-Nodule dataset. Li et al.[100] introduced a candidate nodule detection using Res2Net with global channel spatial attention mechanisms and a false positive reduction network, achieving a CPM of 0.929 on LUNA16. Ma et al.[101] proposed an end-to-end network embedding a transformer module within a 3D CNN framework, combining an attention block and a multi-scale skip pathway, with a dual-head detector to optimize sensitivity and specificity, achieving a competition performance metric (CPM) of 0.9073 on the LUNA16 dataset. Zhou et al.[102] proposed a cascaded multi-stage framework combining a 2D YOLOv5 model for initial nodule localization, a candidate selection algorithm for redundancy reduction, a 3D multi-size fusion model for false positive reduction, and a 3D

U-Net for segmentation, achieving a CPM of 0.895 on LUNA16. Beyond pulmonary nodule detection, similar pipelines have been applied to non-small cell lung cancer (NSCLC) tumor detection and segmentation. Primakov et al.[103] developed and validated a 2D U-Net pipeline across 1,328 CT scans from eight institutions, achieving a sensitivity of 0.97 and AUC of 0.98. Dlamini et al.[104] employed YOLOv4 combined with a region-based active contour model for detection and 3D volumetric reconstruction of NSCLC tumors, achieving a sensitivity of 97.02% and F1 score of 0.968. Recent research has focused on bridging the gap between experimental CAD and real-world screening scenarios. One example is the work of Cui et al. [98] which trained a ResNet using chest LDCT screening cases from three hospitals to identify candidate nodules, followed by a neural network classifier to reduce false positives. Their model obtained FROC scores (95% CI) of 0.75 [0.73, 0.78] on the validation set and 0.80 [0.78, 0.82] on the LUNA16 dataset. Aresta et al.[105] also employed the YOLOv3 architecture for nodule detection, but their goal was to understand the role of CAD and eye-tracking systems in providing a second, independent, and unsupervised opinion during lung cancer screening. In their study, the AI achieved a detection sensitivity of 0.69 compared to the specialists' 0.67. Besides the work of Aresta et al.[105], other comparative studies in screening settings further demonstrate that DL systems can exceed human expert sensitivity. For instance, Cui et al. [106] evaluated a deep learning-based CAD within a LDCT screening program. Their architecture leveraged a maximum intensity projection technique, employing an ensemble of U-Nets for candidate detection followed by VGG-net classifiers for false positive reduction. Results indicate that the CAD system achieved 90.1% sensitivity with 1.0 false positives per scan, significantly outperforming a double-reading protocol by radiologists, which achieved 76.0% sensitivity. This performance gap is particularly evident in small nodules (4–6 mm) and Lung-RADS 2 cases (solid nodules < 6mm) [107], suggesting that AI's ability to process subtle textural patterns exceeds manual inspection. Noteworthy, the analysis of Cui et al. [106] revealed that the CAD system identified 63 nodules missed by double reading, while radiologists identified 27 nodules missed by AI. Their findings implied that AI functions best as a complementary tool. This observation is further supported by Aresta et al. [105] who demonstrated that combining radiologist annotations with AI predictions improves global detection sensitivity. **Table 3** summarizes the architectures, datasets and key performance metrics of the reviewed lung nodule detection methods.

Table 3 - Summary of deep learning-based lung nodule detection methods.

Study	Model	Pipeline	Dataset	Metrics
Ji et al. [99]	YOLOv5s + attention	Single-stage	LUNA16 X-Nodule	mAP 0.720, F1 0.740
Li et al. [100]	Res2Net + attention + FPRNet	Two-stage	LUNA16	CPM 0.929
Ma et al. [101]	3D CNN + transformer	Single-stage	LUNA16	CPM 0.907
Cui et al. [98]	ResNet + NN classifier	Two-stage	Hospital LDCT + LUNA16	FROC 0.80 (LUNA16)
Aresta et al. [105]	YOLOv3	Single-stage	Screening cohort	Sensitivity 0.69
Cui et al. [106]	U-Net ensemble + VGG-net	Two-stage	LDCT screening program	Sensitivity 0.901
Primakov et al. [103]	2D U-Net	Single-stage	Multi-institutional (1328 scans)	Sensitivity 0.97, AUC 0.98
Zhou et al. [102]	YOLOv5 + 3D fusion + 3D U-Net	Multi-stage (cascaded)	LUNA16	CPM 0.895
Dlamini et al. [104]	YOLOv4 + active contour	Single-stage	NSCLC (TCIA)	Sensitivity 97.02%

However, despite these advancements, translating algorithmic success into a clinical screening situation remains difficult. Schreuder et al. [91] highlight that while algorithms achieved sensitivity scores of up to **97.2% on benchmarks like LUNA16, performance decreased in a real-world screening scenario**. This decrease is often due to variations in CT vendors, dosage (SDCT vs. LDCT), and the exclusion of nodules smaller than 3 mm in common research datasets. This generalization gap reveals several bottlenecks in current deep learning methodologies. First, the majority of high-performing systems are trained on curated datasets such as LIDC-IDRI or LUNA16, which may not represent the diversity of real-world clinical data [98]. Second, traditional CNNs are often limited by their local receptive fields, which may struggle to model the broader anatomical context required to distinguish a nodule from a complex vascular bifurcation across multiple slices. As highlighted by Thakur et al. [108], the success of traditional supervised DL models is highly dependent on the availability of a large, high-quality labeled dataset. This condition is rarely met in medical image analysis, as obtaining labeled data is a laborious task that requires clinical expertise.

To address this data insufficiency, the field has increasingly adopted **self-supervised learning (SSL)** and semi-supervised approaches to exploit the available unlabeled medical data [109], [110], [111], [112]. Unlike the generative applications discussed previously for image restoration, SSL in this context is employed for representation learning by automatically generating pseudo-labels from the data itself. This approach allows models to learn meaningful features and representations without requiring extensive, costly human annotation. SSL is typically trained in two distinct phases: **pretraining and fine-tuning**. In the pretraining phase, the neural network is trained on a pretext task using unlabeled data, for example, predicting image rotations or reconstructing masked parts of the image. The goal is to force the model to learn the fundamental underlying structure of the data (e.g., the anatomical structure of the lungs) without needing manual annotations. Once pretrained, the model is refined on a specific downstream task (such as nodule detection), using a smaller set of annotated data.

Earlier breakthroughs in medical SSL focused on CNN-based architectures. Most notably, Model Genesis [113] provided a foundational framework for pre-training CNNs by teaching them to restore original CT images from various distorted versions. The model is trained to recover original 3D sub-volumes from inputs that have undergone specific transformations, including **non-linear transformations** to learn organ appearance and intensity distributions; **local pixel shuffling** to discern textures and boundaries; and **inner/outer cutouts** to understand global geometry and local continuity (e.g., rotating, shuffling, or masking patches). By solving these reconstruction challenges, Model Genesis demonstrated that models could learn generic medical representations without manual labels.

Following these foundational works, the field has increasingly shifted towards **contrastive learning and masked image modeling**. **Contrastive learning** is a technique that learns image-level representations by training the model to recognize similar and dissimilar versions of the same data across different augmentations. The objective is to group similar or positive pairs (augmentations of the same image) in the feature space while pushing different or negative pairs apart. One example of a contrastive learning technique is the momentum contrast (MoCo) [112]. This framework introduced a dynamic queue and a momentum-updated encoder to build a large, consistent dictionary of negative samples independent of batch size. This approach proved that unsupervised pretraining outperformed supervised pretraining in downstream tasks such as object detection and segmentation.

In the medical domain, Chaitanya et al. [109], [110] extended these principles by proposing strategies to learn both global and local features. By leveraging structural similarity across volumetric images and using pseudo-labels for unlabeled data, they demonstrated that models could achieve benchmark performance using as few as two labeled 3D volumes. Specifically addressing nodule detection, Niu and Wang [114] applied region-based contrastive learning to a 3D vision transformer. By dividing CT volumes into non-overlapping cubes and learning embedding features via self-attention, their model achieved a sensitivity of 0.901 at 1/8 false positives per scan in the LUNA16 dataset (specifically nodules $> 6\text{mm}$), demonstrating the power of pre-training for small-dataset scenarios.

Parallel to contrastive approaches, **masked image modeling (MIM)** has emerged as an effective SLL technique. In this strategy, the model is trained to reconstruct hidden or masked portions of an input image. By predicting missing pixels or voxels, the network is forced to learn contextual representations that are more robust than supervised labels. For instance, He et al. [111] used this technique and introduced **masked autoencoder (MAE)**, which utilizes an asymmetric encoder-decoder architecture. The encoder processes only the visible subset of patches (typically 25% of the image), while a lightweight decoder reconstructs the original image from the latent representation and the mask tokens. This high masking ratio (75%) creates a challenging pretext task that forces the model to learn a deep contextual understanding of image structure, resulting in highly scalable vision learners. Building on this, Xie et al. [115] introduced SimMIM, a simplified framework demonstrating that complex designs like special tokenization are unnecessary. They showed that simple random masking combined with a lightweight linear prediction head is sufficient to learn state-of-the-art representations.

The shift toward SSL is occurring in parallel with an architectural transition from CNNs toward **Vision Transformers (ViTs)**. Originally designed for Natural Language Processing (NLP) to model relationships between words in a sentence, Transformers were adapted for computer vision by treating an image as a sequence of patches, analogous to tokens in a sentence. Unlike CNNs, which rely on localized filters, ViTs utilize a self-attention mechanism. This allows the model to calculate the relationship between a specific patch (e.g., a potential nodule) and its surrounding anatomical environment across the entire volume. Recent studies have begun to explore ViTs for nodule detection. Mkindu et al. [116] proposed 3D-NodViT, a model based on a shifted window vision transformer integrated with a 3D region proposal network. Employing Bayesian optimization to tune hyperparameters such as the number of transformer

encoders and channels, their approach demonstrated that ViTs could outperform deep CNN-based architectures with fewer resources, achieving a CPM of 0.909 on the LUNA16 dataset. However, the effectiveness of such models remains constrained by data limitations. Mkindu et al.'s reliance on the relatively small LUNA16 dataset highlights a critical bottleneck. While ViTs demonstrate potential, their efficacy is constrained without the large-scale datasets that are required to learn robust features from scratch. Nonetheless, this characteristic makes ViTs suited for SSL. By leveraging large-scale pre-training, such as MAE, on unlabeled data, Transformers can learn rich, high-dimensional representations necessary for 3D medical reasoning before being applied to a specific downstream task. Feichtenhofer et al. [117] extended the MAE framework to video with Video-MAE. A key finding of their work is that volumetric data is highly redundant, requiring an optimal masking ratio as high as 90% (compared to 75% for 2D images) to create a sufficiently challenging pretext task. By using spacetime-agnostic random masking, Video-MAE enables vanilla Vision Transformers to learn powerful representations with minimal domain-specific inductive bias. Although originally designed for natural video, this framework is directly applicable to volumetric medical imaging, as video is a temporal sequence of 2D frames, a CT scan is a spatial sequence of 2D slices (see **Figure 6**).

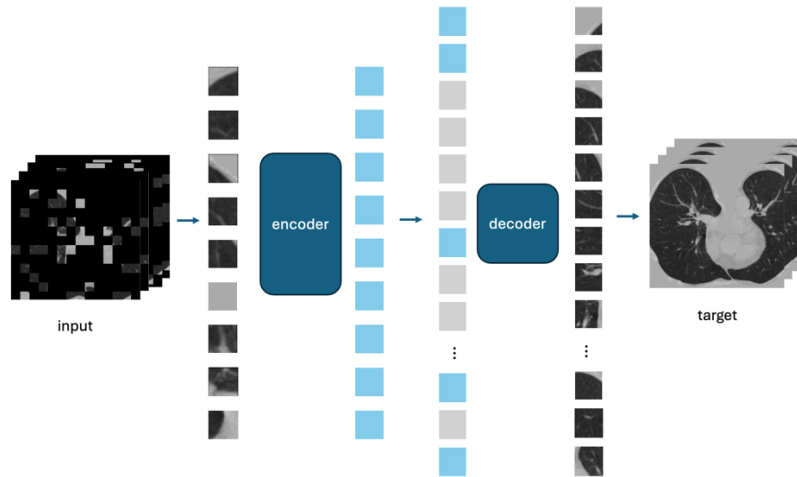


Figure 6 – Adaptation of the Video-MAE framework for volumetric medical imaging. The model processes a 3D input volume (e.g., a lung CT scan) by applying a high masking ratio (90%), leaving only a small portion of visible patches. These patches are processed by an encoder to generate latent representations, which a lightweight decoder then uses to reconstruct the original anatomical volume. Adapted from [25], [117].

Recent research has shown that this pretraining strategy significantly improves performance on downstream tasks like classification [118] and segmentation [119]. For instance, Buess et al.[118] introduced Video-CT MAE for vertebral fracture diagnosis. They demonstrated that by initializing models with weights pre-trained on natural videos and performing self-supervised domain adaptation on unlabeled CT data, they could close the domain gap between natural videos and medical images. Their approach improved the downstream task performance by 8.1 in F1 score and training speedup, proving that Transformers can excel in low-data medical regimes when supported by robust SSL.

Metrics to evaluate lung nodule detection

To quantitatively evaluate the performance of the nodule detection system, metrics such as the **free-response receiver operating characteristic (FROC)**, **competition performance metric (CPM)**, **average precision (AP)** and **mean average precision (mAP)** are commonly used.

The **FROC** [120] is the primary metric for evaluating nodule detection techniques. Unlike a standard ROC curve, which plots the true positive rate against the false positive rate, the FROC curve plots **sensitivity** in the y-axis against the average **false positives per scan (FPs/scan)** in the x-axis. Sensitivity measures the ability of the model to identify true positives (TP) considering the false negatives (FN), i.e., $sensitivity/recall = \frac{TP}{TP+FN}$. False positives per scan, as the name indicates, is the total number of false positives found, divided by the number of scans.

An ideal FROC curve represents a system that detects and localizes all nodules without making any mistakes, this appears as a curve that rises steeply near the origin and rapidly plateaus at a sensitivity of 1.0. In practice, FROC is reported as a value at a specific false positive per scan threshold, e.g., a sensitivity of 0.8 at 0.1 FPs/scan.

AP and **mAP** are standard metrics used in computer vision detection tasks [121]. These metrics evaluate both **localization** and **classification** simultaneously by considering the spatial overlap between the predicted bounding box and the ground truth using Intersection over Union (IoU) and the accuracy of the predicted class. IoU measures the ratio of the intersection area and the union area between the predicted bounding box and the ground truth:

$$IoU = \frac{\text{area of intersection}}{\text{area of union}} \quad (12)$$

To calculate AP for a specific object category (C_m) at a specific IoU threshold (t), a prediction is considered a TP if its $\text{IoU} \geq t$, otherwise it is a false positive (FP). For each category C_m , precision is calculated at each instance (n) where a true positive is identified. Then, these precision values are averaged over the total number of positive instances (i) within that specific class. Mathematically, it is defined as:

$$AP_{C_m} = \frac{1}{i} \sum_{n=1}^i \text{Precision}_{C_{mn}} \quad (13)$$

where precision is computed for the n -th instance of class C_m as:

$$\text{Precision}_{C_{mn}} = \frac{TP_{C_{mn}}}{TP_{C_{mn}} + FP_{C_{mn}}} \quad (14)$$

The higher the AP, the higher the model precision. In medical imaging, an AP at IoU of 0.1 is accepted.

The mAP provides a global view of the detection performance by averaging AP scores calculated at different IoU thresholds, e.g., mean average precision calculated over IoU thresholds from 0.1 to 0.5 with increments of 0.05 or across different classes (if detecting multiple nodule types). Formally, considering one class of objects, mAP is defined by:

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_{IoU=t_i} \quad (15)$$

where N is the number of IoU thresholds and $AP_{IoU=t_i}$ is the AP at the i -th threshold.

1.2.3. Lung nodule classification

Once a candidate nodule is detected, the subsequent task is classification. During the nodule classification phase, the radiologist should differentiate between benign, malignant and indeterminate findings. In screening scenarios, accurate classification is important to prevent unnecessary invasive procedures and patient anxiety while ensuring early malignancies are not missed.

The differentiation of nodules relies mainly on morphological and volumetric criteria defined by international guidelines depending on the detection context. For instance, for incidental findings, i.e., nodules found on CT scans performed for reasons other than cancer screening, the Fleischner Society [94] guidelines are considered the global standard, focusing on reducing unnecessary follow-up for low-risk incidental nodules. In the specific context of lung cancer screening, the Lung-RADS (Lung Imaging Reporting and Data System) [107] from the American College of Radiology is the mandatory standard in US-based screening. It consists of a structured reporting system with discrete assessment categories (0-4) to standardize management. Based on nodule characteristics such as size, density and growth, it has specific follow-up intervals. For example, solid nodules $< 6\text{mm}$ or no solid nodules $< 30\text{ mm}$ are considered Lung-RADS 1 and 2 and therefore benign nodules and should continue annual screening, while suspicious findings in Lung-RADS 4B or X should undergo immediate diagnostic workup.

In contrast to the categorical approach proposed by Lung-RADS, the Pan-Canadian (PanCan) [122] model and British Thoracic Society (BTS) [123] guidelines utilize probabilistic risk calculators (the Brock model). This guideline estimates a continuous percentage risk of malignancy and advocates for volumetric analysis. Other guidelines include the European position statement [124] based on the NELSON trial's findings and the National Comprehensive Cancer Network [125] composed of guidelines for oncologists to use after screening detection. While these guidelines differ in their specific algorithms, there is a unified consensus shifting toward volumetry-based management [96]. The European Society of Radiology and European Respiratory Society [96] emphasizes that measuring nodule volume and volume doubling time is superior to manual diameter measurement, which is prone to high inter-observer variability. This clinical pivot toward quantitative volumetric analysis highlights the necessity for 3D automated systems.

Deep learning-based techniques have emerged as an efficient tool for lung nodule classification tasks, demonstrating strong performance across diverse architectures and input strategies. These approaches can be differentiated based on their input strategy, such as end-to-end models that process CT volumes directly, patch-based methods centered on the nodule and segmented-based approaches. Among end-to-end approaches, Ardila et al.[126] proposed a 3D CNN processing full LDCT volumes, including both current and prior scans to predict lung cancer risk, achieving an AUC of 0.944 on 6716 NLST cases. Shuvo et al.[127] introduced a three-stage hybrid pipeline integrating 3D Res-U-Net for lung segmentation, YOLOv5 for nodule detection and a Vision Transformer for binary classification, achieving 96.29% accuracy on LUNA16. Patch-based approaches use cropped nodule-centered regions without segmentation mask as input image for model training. Halder et al.[128] introduced a two-pathway architecture combining adaptive morphology operations and Gabor filtering to simultaneously capture textural and morphological nodule characteristics, reporting an AUC of 0.9936 on LIDC-IDRI dataset. Wang et al.[129] designed MResNet, which combines residual learning with a pyramid pooling module to fuse multi-scale coarse and fine-grained features from nodule-centered patches, obtaining an AUC of 0.928. Lv et al.[130] addressed data scarcity using a lightweight multi-scale interleaved fusion network on 50x50 nodule image patches, obtaining an accuracy of 94.82% on LUNA16. Annotation-dependent approaches require nodule segmentation to train the DL classifier. Astaraki et al.[131] proposed a dual-pathway model integrating intra-nodule characteristics with contextual features extracted from 1,297 manually segmented nodules, with a random forest classifier on top of the network, achieving an AUROC of 0.937. Chang et al.[132] developed a multi-view residual selective kernel network fed with axial, coronal and sagittal planes and added handcrafted texture features, achieving an AUC of 0.9711. Sun et al. [133] proposed Nodule-CLIP, which performs U-Net-based 3D nodule segmentation before aligning the image, class, and attribute features through contrastive learning, achieving 90.6% classification accuracy on LIDC-IDRI. While these methods demonstrate remarkable classification performance, their reliance on learned latent representations limits the direct extraction of clinically meaningful features, motivating the exploration of quantitative imaging approaches such as radiomics.

Radiomics aims to extract quantitative and reproducible information from diagnostic images such as CT, MRI, and PET, including complex patterns that are relevant but difficult to recognize or quantify by the human eye. As defined in the literature [134] [135], radiomics

refers to the high-throughput extraction of large amounts of quantitative features from medical images, effectively converting medical images into mineable, high-dimensional data. Unlike traditional radiological interpretation, which relies on subjective visual assessment of semantic features, radiomics mathematically quantifies the tumor. Combining radiomic features with machine learning algorithms enables the development of models to predict malignancy, patient survival, and other outcomes[136]. According to Pastorino et al.[8] and van der Aalst et al.[10], individual risk stratification using **radiomic and AI analysis of baseline LDCT findings represents a promising development for future screening programs, improving the efficacy of limited health care resources and reducing costs.**

The typical radiomics pipeline consists in five main steps as shown in **Figure 7: (1) image acquisition; (2) segmentation**, where the region of interest (ROI) is defined; **(3) features extraction**, where the mathematical descriptors from the segmented ROI (2D or 3D) is calculated using software like *Pyradiomics*; **(4) feature analysis**, in this step the redundant and correlated features are removed, the stability and discriminative analysis are also computed in this step to ensure feature robustness and relevancy; finally, **(5) model development**, in this step machine learning models such as decision tree, support vector machine, random forest, logistic regression etc. are trained and validated. Additionally, according to recommendations by the ESR [137], explainability analysis should be included as an important step in the radiomics pipeline to ensure transparency and interpretability, which are essential for facilitating the transition to clinical practice.

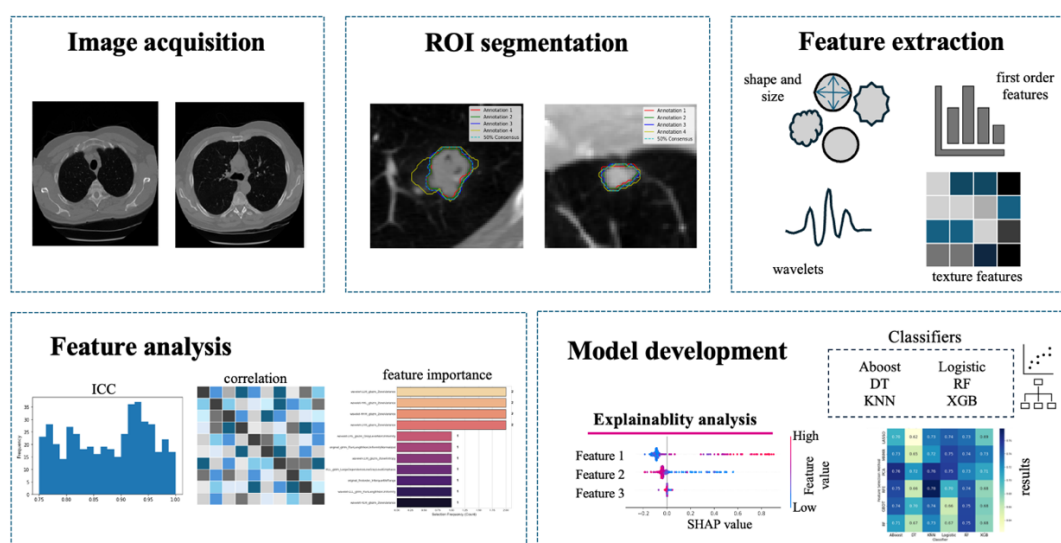


Figure 7 – Radiomics pipeline consisting of five main steps: image acquisition, ROI segmentation, feature extraction, feature analysis and model development.

Radiomic features

The quantitative features extracted in a radiomics pipeline are categorized into the following classes: **shape and size features, first-order (statistical) features, and second-order (texture) features**. The shape and size features describe the geometric properties of the ROI such as volume, area, sphericity and compactness. First-order features consist of global gray-level histogram and describe the distribution of voxel intensities within the ROI, including mean, max, min, variance and percentiles, skewness and kurtosis, entropy and uniformity or energy. Texture features quantify the spatial interrelationships between voxels and include gray-level cooccurrence matrix (GLCM), gray-level run-length matrix (GLRLM), gray-level size zone matrix (GLSZM), gray-level distance zone matrix (GLDZM), neighborhood gray-tone difference matrix (NGTDM), and neighborhood gray-level dependence matrix (NGLDM).

Beyond these original features, higher-order features can be extracted by applying filters to the images. A prominent example is wavelet filtering, which decomposes the original image into different frequency components. This process yields eight separate decompositions for each level by applying either a High (H) or Low (L) pass filter across the three spatial dimensions, resulting in combinations such as LLL, LLH, LHL, LHH, etc. Extracting first-order and texture features from these wavelet-transformed images allows the model to capture complex multiscale information that may be hidden in the original image.

The primary advantage of radiomics is its ability to provide a non-invasive virtual biopsy. As highlighted by the *PyRadiomics* study [135], solid tumors are spatially and temporally heterogeneous, however, a physical tissue biopsy can only capture a small sample of the tumor, which can potentially miss aggressive subclones. Conversely, radiomics captures the comprehensive phenotypic characteristics of the entire 3D volume. By correlating these textural patterns with malignancy, radiomics bridges the gap between medical imaging and personalized medicine.

This capability is particularly valuable for managing indeterminate pulmonary nodules discovered incidentally or during CT screening, since traditional radiologist assessment relies on subjective visual indicators, such as size, borders and shape, which often leads to diagnostic uncertainty [138]. While invasive image-guided biopsies remain the reference standard for determining cancer subtype and confirming malignancy, radiomics could offer a non-invasive

alternative by extracting reproducible, quantitative data, thereby guiding management and reducing the need for invasive intervention.

Furthermore, radiomics, usually paired with traditional machine learning algorithms, has the advantage of being interpretable and explainable compared to DL approaches. This distinction is particularly important because, despite the high theoretical performance of DL models, their translation into clinical practice is compromised by the black box phenomenon. For clinicians, a diagnosis score without interpretation is often insufficient to justify aggressive treatments such as resection or radiotherapy. Consequently, the transparency provided by radiomic features remains a significant factor in building clinical trust and facilitating the adoption of AI in screening workflows.

Radiomics applications in lung cancer

Recent literature has demonstrated the versatility of radiomics in complex lung diagnostic tasks such as histological subtyping, staging and malignancy classification. Song et al.[139] introduced latent representation learning as a novel feature engineering method, projecting original shape, intensity, and texture features into a latent space through a unique hybrid loss function. When tested on a multi-center dataset for non-small cell lung cancer (NSCLC) subtype classification from 8 international databases, this approach significantly improved classification performance compared to traditional methods (all $p < 0.001$). Dunn et al.[140] demonstrated the integration of DL segmentation with radiomic analysis for automated lung cancer subtype classification. Using the incremental multiple resolution residual network for automatic segmentation, they achieved an accuracy of 92.7% and an AUC of 97% when classifying three histological subtypes: adenocarcinoma, small cell carcinoma, and squamous cell carcinoma. Their analysis revealed that higher-order texture features were the most predictive for subtype classification. D'Arnese et al.[141] developed an automated pipeline using PET/CT images to characterize NSCLC histological subtypes with a mean accuracy of 94.2%, enabling earlier diagnosis without immediate invasive intervention. Beyond primary tumor characterization, radiomics has shown promise in predicting lymph node metastasis in NSCLC, which is crucial for patient management and prognosis. Li et al.[142] conducted a comprehensive meta-analysis revealing that CT-based radiomics achieved a sensitivity of 0.84 and specificity of 0.82 with an AUC of 0.90, while PET/CT radiomics demonstrated comparable performance with a sensitivity of 0.82, specificity of 0.86, and an AUC of 0.90.

These findings highlight radiomics as a non-invasive tool for predicting metastatic spread and the potential in supporting surgical planning and personalized treatment decisions.

While subtyping and staging are critical for confirmed cancer patients, the primary challenge in screening scenarios is distinguishing malignant nodules from the vast number of benign false positives detected on LDCT. Although extensive research in the field of lung cancer screening using radiomics has been conducted, some studies report **low sensitivity or specificity**. For instance, Alahmari et al. [143] used 18 Rider features and trained them with a Random Forest model, achieving a sensitivity of 0.49 and a specificity of 0.93. Rundo et al. [144] used 32 radiomic features trained with logistic regression, achieving a sensitivity of 0.67 and a specificity of 0.86. Conversely, Hunter et al. [145] achieved a higher sensitivity of 0.86 using 5 radiomic features selected by LASSO and the K-means clustering algorithm, but at the cost of a low specificity of 0.56. Causey et al. [146] achieved notable results (sensitivity of 0.88 and specificity of 0.88) using the LIDC-IDRI dataset with features obtained from convolutional neural networks. However, when applying only radiomic features, the sensitivity dropped to 0.70. Moreover, Causey et al. [146] removed nodules with an average score for the likelihood of malignancy between 3 and 4 from the LIDC-IDRI dataset, which are the most challenging nodules to classify.

While some models in the literature achieved good performance, they often used **smaller sample sizes**, for example, Choi et al. [147] (72 nodules), Selvam et al. [138] (88 nodules), Garau et al. [148] (113 nodules), and Liu et al. [149] (141 nodules), Mao et al. [150] (294 nodules). Larger datasets contribute to more stable, unbiased ML models, thereby enhancing reliability and generalizability. Peikert et al. [151] achieved good performance metrics (sensitivity = 0.90, specificity = 0.86) using 8 radiomic features trained with LASSO model with a relatively large dataset, but it is worth noting that their study included larger nodules (≥ 7 mm) than those analyzed in our study (≥ 3 mm). To date, radiomic analyses focusing on small lung nodules have not been extensively explored in the literature. **Table 4** summarizes recent studies that have applied radiomics to predict nodule malignancy in lung cancer screening.

Besides the aforementioned limitations, some other challenges highlighted in the literature [152] include **reproducibility issues**, i.e., a significant portion of published studies as not reproducible because many studies rely on private dataset that are not publicly available, making independent validation impossible; **lack of technical transparency**, development environment, specific library and programming languages are often omitted as very few research

groups release their source code. Making code accessible would benefit the entire scientific community, fostering knowledge sharing and allowing researchers to build upon existing state-of-the-art baselines rather than starting from scratch.

Table 4 - Summary of recent studies applying radiomics for nodule malignancy prediction in lung cancer screening.

Study	Dataset (nodule size)	LDCT?	Model description	Sen.	Sp.	Acc.	AUC	Limitations
Choi et al. 2018 [147]	72 patients LIDC-IDRI (≥ 3 mm)	yes	2 radiomic features, LASSO feature selection and SVM classification	87.2%	81.2%	84.6%	0.89	small dataset
Peikert et al. 2018 [151]	726 nodules from NLST (≥ 7 mm)	yes	8 radiomic features selected and trained with LASSO model	90.4%	85.5%	-	0.94	selected data to balance the dataset
Alahmari et al. 2018 [143]	467 cases from NLST (≥ 4 mm)	yes	18 Rider features, ReliefF feature selection, Random Forest	49.0%	93.0%	-	0.82	low sensitivity
Causey et al. 2018 [146]	664 nodules from LIDC-IDRI (≥ 3 mm)	no	a logistic regression model trained only on a size metric (square root of the largest cross-sectional area)	69.7%	95.5%	82.6%	0.94	low sensitivity
Mao et al. 2019 [150]	294 nodules (6-15 mm)	yes	11 radiomic features, LASSO feature selection, linear regression model	81.0%	92.2%	89.8%	0.90	small sample size
Garau et al. 2020 [148]	113 nodules COSMOS (all sizes)	yes	5 radiomic features, SVM-LASSO test set: 72 nodules from LIDC-IDRI.	90.0%	FPR = 35.5%	79.1%	0.86	small sample size
Liu et al. 2021 [149]	141 nodules (not specified)	yes	3 radiomic features, radiomic monogram	86.4%	85.7%	86.0%	0.98	small sample size, potential selection bias
Rundo et al. [144]	703 nodules from the bioMILD trial (all sizes)	yes	32 radiomic features, Borderline-SMOTE over-sampling method, elastic net regularization for logistic regression, tested with a blind test	67.0%	86.2%	76.6%	0.82	low sensitivity
Selvam et al. [138]	88 nodules (not specified)	no	10 radiomic features selected by random forest classifier, XGBoost algorithm	89%	85%	89%	0.89	small sample size,
Hunter et al. [145]	990 nodules (5-15 mm)	mix	5 radiomic features selected by LASSO, K-means clustering	86%	56%	73%	0.78	low specificity

LIDC-IDRI - the Lung Image Database Consortium and Image Database Resource Initiative; NLST - National Lung Screening Trial; Sen. - sensitivity; Sp. - specificity; Acc. - accuracy; AUC - area under the ROC curve; FPR - false positive rate.

Metrics to evaluate lung nodule classification

To evaluate the performance of machine learning models in lung nodule classification, specifically distinguishing between benign and malignant nodules, metrics such as **sensitivity, specificity, accuracy, balanced accuracy, area under the ROC curve (AUC) and false positive rate** are often reported. These metrics are computed from the confusion matrix, which summarizes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

AUC is based on the receiver operating characteristic (ROC) curve, which plots sensitivity on the y-axis against false positive rate (FPR) on the x-axis at various classification thresholds. The AUC provides a single value ranging from 0.5 to 1, where 1 indicates a perfect model that correctly distinguishes all samples, whereas 0.5 indicates a random guess.

1.3. Thesis motivation and objectives

Increasing survival rates through lung cancer screening have been demonstrated in several large-scale clinical trials. However, the implementation of these screening programs still faces technical and clinical challenges. While potentially cost-effective in the long term, screening programs impose an economic burden because they require manual analysis of large volumes of CT scans. Furthermore, the high rate of false positives remains a critical barrier as it leads to overdiagnosis, triggering unnecessary invasive biopsies and causing significant psychological distress for patients.

CAD and AI offer promising solutions to streamline this process. However, their effectiveness is often constrained by imaging modality. Although LDCT reduces patient radiation exposure, it inevitably compromises image quality, resulting in lower signal-to-noise ratios, increased noise, and artifacts that can obscure diagnostic details. Additionally, the development of robust AI tools to assist radiologists in screening scenarios faces a bottleneck due to the scarcity of annotated LDCT datasets. Even though many models with good performance have been developed and extensive research has been done in the field of lung cancer, most of the studies focus on the usual SDCT scan rather than LDCT, which is used in screening protocols. This creates a domain gap that limits their reliability in real-world screening scenarios. **This PhD project aims to bridge these gaps by developing AI-based tools specifically tailored for lung cancer screening. To this end, the research focuses on overcoming data scarcity and image degradation issues to build robust algorithms for both nodule detection and interpretable classification.**

Objective 1

A substantial portion of current literature focuses on image restoration (denoising) to improve LDCT quality. However, these conventional methods either require raw projection data, which is scanner-specific and often inaccessible, or paired SDCT-LDCT image datasets for training, which is problematic as it is unethical to expose patients to repeated CT scans. To overcome these limitations, some researchers create synthetic LDCT images by degrading existing SDCT data to train denoising [45], [46], [49]. Others rely on the dataset from *the 2016 LDCT Grand Challenge*, yet even this standard benchmark was generated using a validated noise-insertion

method rather than raw clinical paired [12], [17]. Therefore, as paired clinical data is not publicly available, the field of denoising relies on synthetic approaches.

This PhD project shifts the focus from denoising LDCT to generating diagnostically representative LDCT images from existing SDCT data through a degradation step. By using these generated images to train and validate robust radiomic classifiers, this strategy directly addresses the domain gap. This approach bypasses the high computational cost of image restoration and ensures that downstream models are trained on data that accurately reflects the noise profile of clinical screening scans.

Objective 1: To evaluate and compare multiple degradation techniques on CT images. This involves: (1) providing a systematic comparison of different physical and synthetic degradation methods; and (2) validating the feasibility of using degraded SDCT images for developing downstream classification models.

Objective 2

Automated diagnostic tools powered by artificial intelligence are critical for alleviating radiologists' workload in detecting nodules on screening scans. However, the reliability of these models depends on the quality and quantity of the dataset, which remains costly and time-consuming to collect and curate. State-of-the-art transformer-based methods have shown superior performance, but their effectiveness is limited by data availability, requiring robust pretraining strategies [99], [114], [116]. To address data insufficiency, various self-supervised and semi-supervised approaches [109], [110], [112] have been proposed to leverage large amounts of unlabeled data. Among them, MIM [115], [117], [153] has proven effective by training models to reconstruct masked regions of inputs, forcing them to learn rich, contextual visual representations, leading to improved performance for downstream tasks in 3D medical reasoning, such as classification [118], and segmentation [119].

Consequently, this thesis proposes to advance lung nodule detection by leveraging self-supervised pretraining of transformer backbones to enhance the state-of-the-art nnDetection framework. [154]. By treating 3D CT volumes as sequential spatiotemporal data (analogous to video), vast amounts of unlabeled data can be exploited to pre-train the model. This forces the network to learn rich, contextual anatomical representations by reconstructing masked regions, providing a robust initialization for detecting nodules in 3D volumes without relying on large, labeled datasets.

Objective 2: To leverage state-of-the-art self-supervised training methods on large domain-specific lung screening CT datasets. This involves adapting MIM-based pre-training, specifically Video-MAE, to 3D transformer backbones and fine-tuning them on annotated datasets to achieve state-of-the-art lung nodule detection performance.

Objective 3

Once a nodule is detected, the subsequent challenge is distinguishing benign from malignant. However, LDCT screening is associated with a high false-positive rate: up to 96.4% of positive screening results were false positives [4]. While deep learning classifiers exist, they often lack transparency, which limits clinical trust. Radiomics offers an interpretable alternative, extracting quantitative biomarkers (shape, texture) that align with biological characteristics.

However, developing a robust radiomics classifier for screening is hindered by dataset limitations: the NLST (screening) dataset lacks segmentation, whereas the LIDC-IDRI (segmented) dataset is mostly SDCT. To address this, we leverage the mixed nature of the LIDC-IDRI dataset, which contains both SDCT and LDCT images. One of the degradation strategies studied in **Objective 1** is applied to the SDCT images, while the subset of real LDCT images in the LIDC-IDRI dataset is reserved for testing. Training the radiomics model on simulated LDCT images generated from SDCT scans enables the development of a classifier robust to the noise characteristics of screening scans, while retaining the interpretability necessary for clinical decision support. This approach not only addresses the scarcity of annotated LDCT data but also offers a novel solution to explore the potential inclusion and utilization of SDCT images to develop screening-specific AI tools.

Objective 3: To develop an algorithm for lung nodule classification using radiomics, addressing the lower signal-to-noise ratio resulting from LDCT acquisitions. The goal is to build an interpretable model trained on synthetic LDCT data that can accurately distinguish likely benign from likely malignant nodules in real-world screening scenarios.

1.4. Thesis Outline

The present thesis is organized into five chapters, structured to guide the reader from the clinical context and theoretical background to the state of the art and methodological innovations. Chapter 1 consists of a cover essay, which describes the clinical problem addressed, covers the theoretical background and related work, and outlines the specific objectives of this thesis. Chapters 2 through 4 detail the methodological contributions. Chapter 5 concludes with a summary of the main contributions, a discussion of limitations and a reflection on future research developments. A brief overview of the remaining chapters is provided below:

- **Chapter 2** provides a systematic comparison of three CT image degradation methods for generating synthetic LDCT images and evaluates the efficacy of each method in training robust nodule classification models. The study included approaches ranging from sinogram-domain statistical simulation (method 1), to the replication of a validated physics-based method using Pix2Pix (method 2), and unpaired image generation using CycleGAN (method 3). To compare these simulation methods, an evaluation pipeline based on a radiomics-based nodule classification task is proposed. This pipeline uses radiomic features extracted from degraded SDCT images generated by applying the three methods to train the classifier, and reserves LDCT for validation and testing. For each degradation method, the pipeline proceeds with feature extraction from nodule annotations, feature selection and the training and testing of machine learning models. Finally, non-parametric statistical tests were used to compare the performance of the degradation methods.
- **Chapter 3** proposes a lung nodule detection framework that leverages state-of-the-art self-supervised training methods on a large domain-specific lung screening CT dataset. This chapter details the adaptation of masked image modeling to 3D transformer-based backbones, specifically Video-Vit and SwinUNETR, for lung nodule detection using the nnDetection framework for the finetuning lung nodule detection task. The methodology involves large-scale pre-training using the unlabeled NLST dataset to learn robust spatial representations, followed by fine-tuning on the annotated LUNA16 dataset. Furthermore, the work assesses the framework's effectiveness across architectural variations, including the inclusion of skip connections and varying input

resolutions ($16 \times 224 \times 224$ and $64 \times 128 \times 128$). This study advances the use of pretrained transformers for medical detection tasks.

Chapter 4 presents the development of an explainable radiomics-based pipeline for classifying lung nodules, using degraded CT images generated by one of the methods investigated in Chapter 2. This chapter includes research presented at a conference and an extended journal paper. While both studies share the same concept, their methodologies differ in dataset management and feature selection strategies. Although both studies used the LIDC-IDRI dataset, in the conference paper a repeated random sampling strategy was employed, selecting 400 subjects across ten iterations. Within each iteration, data were split into training and test sets at an 8:2 ratio, resulting in a mix of simulated (degraded SDCT) and real LDCT images in both sets. In the journal paper, the full dataset was utilized and to ensure more rigorous validation, degraded SDCT was used for model training, while real LDCT images were reserved for model validation and testing. This strategy effectively addresses the limited availability of annotated LDCT images for training while ensuring the model's validity through testing on real LDCT images. In terms of feature selection strategies, it was limited to the gradient-boosted decision tree method in the conference paper, whereas six different feature selection techniques (1. principal component analysis, 2. recursive feature elimination, 3. feature selection using random forest, 4. least absolute shrinkage and selection operator, 5. feature selection using gradient-boosted decision tree and 6. minimum redundancy maximum relevance) were assessed in the journal paper. Furthermore, feature selection was performed on three different feature subsets independently: (a) only shape and size features, (b) all features but SS features, and (c) all features. These three sets were chosen to assess the relevance of nodule morphological characteristics on the prediction. Both studies addressed class imbalance through a ROI perturbation technique for oversampling combined with random undersampling strategy. Subsequently, six machine learning models were trained and validated for both studies: random forest, AdaBoost, XGBoost, decision tree, K-nearest neighbors, and logistic regression. In both works, the optimal developed model was evaluated on the test set and SHAP (SHapley Additive exPlanations) was used to explain the predictions.

- **Chapter 5** summarizes the main findings, discusses the limitations of the current work and outlines potential future developments.

2 A comparative analysis of CT degradation for robust LDCT nodule classification using radiomics

A Comparative Analysis of CT Degradation for LDCT Nodule Classification using Radiomics

Jiaying Liu, Anna Corti, Valentina D.A. Corino, Luca Mainardi

Abstract—Low-dose computed tomography (LDCT) is the standard modality for lung cancer screening, known for its low radiation dose but high noise levels. While existing literature focuses on denoising LDCT images, comparative research on simulating LDCT characteristics to directly use these images for model development is lacking. This study shifts the focus from denoising images to degrading available standard-dose CT (SDCT) data, generating synthetic images for data augmentation to train classifiers for screening-detected nodules. We compare three degradation methods: (1) a sinogram domain statistical noise insertion; (2) replicate a validated physics-based simulation using Pix2Pix; and (3) unpaired CycleGAN. The generated images were utilized to simulate LDCT screening scenario replacing 695 SDCT cases from the LIDC-IDRI dataset, from which radiomic features were extracted to train machine learning models for lung nodule classification. Regarding image quality, CycleGAN achieved the best Fréchet inception distance (0.1734) and kernel inception distance (0.0813 ± 0.1002) scores, indicating distributional alignment with the target low-dose domain. In the nodule classification task, results confirmed the necessity of domain adaptation since a baseline model trained on non-degraded SDCT data failed to generalize to the real LDCT set (AUC 0.789) with a low sensitivity (0.571). Degraded images generated using CycleGAN approach led to the most balanced performance on the classification task using Adam Booster classifier, achieving an AUC of 0.861, sensitivity of 0.743 and specificity of 0.858 in the independent test. Our findings confirm that generating synthetic LDCT data from standard-dose scans is a viable strategy for training robust nodule classifiers for screening detected nodules.

Index Terms—low-dose CT, image degradation, lung nodule classification, radiomics, generative adversarial networks, synthetic data augmentation.

I. INTRODUCTION

LUNG cancer remains the leading cause of cancer-related mortality worldwide [1]. Early detection through lung cancer screening has been shown to significantly reduce mortality as evidenced in several clinical trials [2], [3]. Through these trials, low-dose computed tomography (LDCT) has emerged as the preferred imaging modality for this screening, as it provides the necessary diagnostic capability

while minimizing the inherent risk of cumulative radiation exposure associated with standard-dose computed tomography (SDCT) scans (also referred to as routine-dose and full-dose). Nonetheless, lower radiation dose results in a lower signal-to-noise ratio and in images with lower contrast, blurring and increased noise and artifacts, all of which may affect clinicians' diagnostic results. Consequently, a substantial portion of the current literature focuses on LDCT image restoration methods to improve LDCT image quality [4]. Conventional restoration methods can be categorized into three groups: sinogram domain filtering, iterative reconstruction and image domain restoration [5]. The sinogram domain filtering directly works on the raw projection data before back-projection, allowing for accurate computation of noise statistics. Iterative reconstruction methods alternate between the sinogram and image domains, which results in high computational expense. Image domain restoration is considered a post-processing method directly applied to reconstructed images. These conventional methods either requires access to raw projection data, which is scanner-specific and inaccessible, or paired SDCT-LDCT image datasets for training, which is problematic as it is unethical to expose patients to repeated computed tomography (CT) scans. To overcome these limitations, some researchers create synthetic LDCT images by degrading existing SDCT data to train denoising [6], [7], [8]. Others rely on the dataset from the 2016 LDCT Grand challenge, yet even this standard benchmark was generated using a validated noise insertion method rather than raw clinical paired [9], [10]. Therefore, as paired clinical data is not publicly available, the field of denoising relies on synthetic approaches. While these approaches could be effective in benchmarks, there are two drawbacks: (1) the noise simulation is often treated as a secondary step, potentially compromising the fidelity of the synthetic images and the real-world generalizability of models trained on them; and (2) the cycles of synthetic degradation and subsequent denoising are computationally expensive and time-consuming during both development and clinical implementation.

This study shifts the focus from denoising LDCT to generating diagnostically representative LDCT images from existing SDCT data, by a degradation step. Instead of denoising the images, this study aims to (1) compare degradation methods

A. Corti is funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvanCed Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

J. Liu, A. Corti, V.D.A. Corino and L. Mainardi are with the Department of Electronics, Informatics and Bioengineering, Politecnico di Milano, Milano, Italy

V.D.A. Corino is also with Cardiotech Lab, Centro Cardiologico Monzino IRCCS, Milan, Italy

Corresponding author: J. Liu (jiaying.liu@polimi.it)

for synthetic image generation and (2) use these generated images to train and validate a robust classifier using radiomics.

Radiomics is an approach that extracts high-throughput quantitative features from a region of interest, allowing for the acquisition of more accurate information that may be hidden from the human eye for analysis and predictive modeling in precision medicine. This approach faces a practical limitation: the lack of annotated LDCT datasets. To overcome this, we leverage the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset, which contains mainly SDCT images but also LDCT, and propose a framework to generate simulated LDCT scans to bridge the domain gap for model training, reserving the real LDCT scans for model validation and testing.

This paper aims to: (1) provide a systematic comparison and evaluation of different LDCT degradation methods; and (2) introduce an approach that bypasses denoising and instead validates the feasibility of using SDCT images for developing downstream classification models.

II. MATERIALS AND METHODS

A. Datasets

This study uses two datasets: the LIDC-IDRI [11] and the Low-Dose CT Image and Projection Data [12], which expands upon the subset originally released for the 2016 LDCT Grand Challenge.

A.1. LIDC-IDRI dataset

The LIDC-IDRI dataset consists of 1018 thoracic CT scans acquired from 1010 participants, with nodule annotations performed by one to four experienced radiologists. Consistent with the literature [13], [14], we defined LDCT as acquisitions with a tube current ≤ 80 mAs, equivalent to 40 mAs assuming a typical gantry rotation time of 0.5 seconds. Fig.1 shows examples of SDCT (a) and LDCT (b) subjects from the LIDC-IDRI dataset and examples of the corresponding nodule annotations.

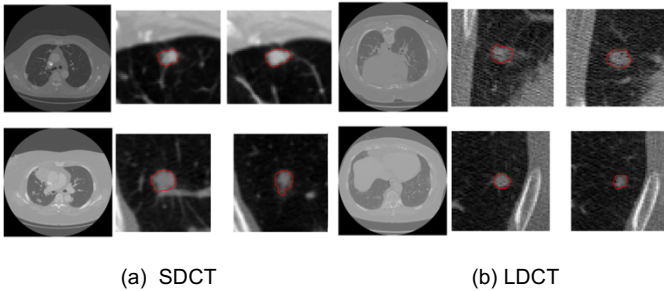


Fig. 1. Examples of SDCT (a) and LDCT (b) subjects from the LIDC-IDRI dataset and examples of the corresponding nodule annotations.

All images were preprocessed to reduce the image-related sources of variability. The preprocessing steps included Hounsfield Unit (HU) conversion, intensity range clipping to a window of $[-1200, 600]$ HU, denoising using a 3D Gaussian filter with a $3 \times 3 \times 3$ voxel kernel and $\sigma = 0.5$, and image normalization to the $[0, 1]$ range.

After excluding the duplicated participants, a total of 2625 nodules, each assigned as ≥ 3 mm (lesions with the greatest in-

plane dimension in the range 3–30 mm) marked by at least one radiologist, were included in this study. Table I describes the LIDC-IDRI dataset.

TABLE I

SUMMARY OF THE LIDC-IDRI DATASET USED IN THIS STUDY.

LIDC-IDRI dataset	SDCT	LDCT
Nr. participants	695	315
Nr. nodules	1950	675
Non-malignant (average likelihood < 4)	1725	606
Malignant (average likelihood > 4)	225	69

A.2. Low-dose CT image and projection dataset

The Low-Dose CT Image and Projection dataset [12] is composed of full-dose and quarter-dose projection data and images from 299 patients, including CT scans of the head, chest, and abdomen. For our study, we focused on a subset of 50 paired chest CT scans, which included both SDCT and simulated LDCT images. Fig. 2 shows an example of paired SDCT and simulated LDCT slices from the dataset. Reduced dose projection data were simulated for each scan using a validated noise-insertion method that considers the effects of the bowtie filter, automatic tube current modulation and electronic noise [15].

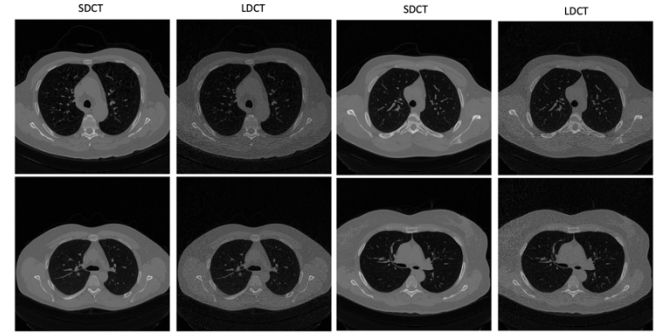


Fig. 2. Example of paired SDCT and simulated LDCT slices from the Low-dose CT image and projection dataset.

B. Methods

LDCT images are degraded by quantum noise and various artifacts during the image acquisition, which lead to a non-uniform distribution of noise over the image space as well as the presence of blurring and streaking artifacts (see Fig.1 and Fig.2). While typical LDCT simulation algorithms inject Poisson noise into SDCT sinograms to mimic the dominant quantum noise, this approach oversimplifies the complex, stochastic nature of real LDCT noise. Therefore, to more accurately simulate LDCT images using SDCT data, we compared three degradation techniques: 1) sinogram domain noise injection [16], 2) a validated physics-based degradation model using Pix2Pix [15], [17], and 3) cycle-consistent generative adversarial network (CycleGAN), a popular unpaired image-to-image translation method [18]. A detailed description of these methods and the implemented training strategies is provided in the following sections.

To compare the LDCT simulation methods, we propose a pipeline (Fig.3) that includes a radiomics-based nodule

classification task. First, the LIDC-IDRI dataset was split into SDCT, which is reserved for training, and LDCT, which is used for 50% validation and 50% testing. Then, the SDCT training set is degraded following the three selected degradation methods. For each degradation method, the pipeline proceeds to extract radiomic features from nodule annotations, perform feature selection, train and test machine learning models.

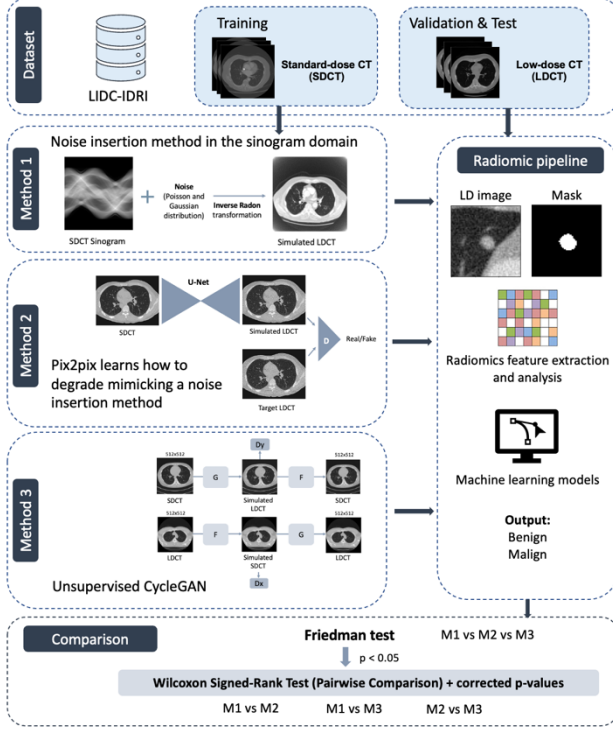


Fig. 3. Overview of the pipeline.

The pipeline was developed using Python 3.9. The code can be found at <https://github.com/jiayingliu423/LDCT-degradation-and-nodule-classification>.

B.1. Method 1 – Simple sinogram-domain simulation

The first degradation method (**method 1**) is a simple low-dose CT simulation strategy in the sinogram domain using the raw data from a high-dose scan as proposed by [16] and utilized in the work of [19]. This method simulates LDCT by adding Poisson and Gaussian noise to mimic the effect of electronic and quantum noise and has the advantage of being computationally lightweight while requiring minimal scanner-specific parameters. Briefly, standard dose sinogram data were obtained by using Radon transformation; subsequently, simulated Poisson and Gaussian noise were added so that the low dose sinogram data were generated as shown in Fig. 4. Equation 1 describes the calculation for the low dose sinogram data ($p_{ld,sim}$):

$$p_{ld,sim} = \log \frac{I_{ld,sim}^0}{\text{Poisson}(I_{ld,sim}^0 \cdot \exp(-p_{sd})) + \text{Gaussian}(m_e, \sigma_e^2)} \quad (1)$$

where $I_{ld,sim}^0$ is the simulated low-dose scan incident flux, p_{sd} is the standard-dose sinogram data obtained from the original

SDCT images using Radon transformation, m_e and σ_e^2 are the mean and variance of the electronic noise, respectively. Finally, the simulated LDCT images were reconstructed from the low-dose sinogram data using the inverse Radon transform.

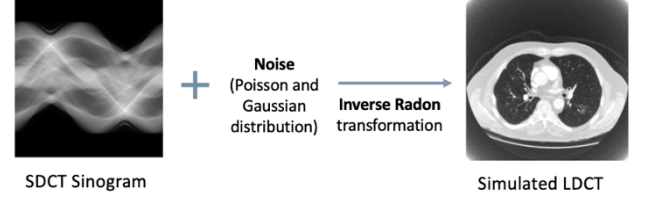


Fig. 4. Pipeline to simulate low-dose computed tomography (LDCT). The Radon transformation is applied to the standard-dose computed tomography (SDCT) image, then electronic and quantum noise effects are included by noise injection. The final step involves inverse Radon transformation to reconstruct the LDCT image. Adapted from [19].

B.2. Method 2 – Physics-based degradation model replication using Pix2Pix

The second degradation method (**method 2**) studied in this paper is inspired by the LDCT grand challenge, which uses a validated physics-based noise modeling technique from Yu et al. [15] to generate the low-dose images present in the challenge dataset. Yu et al.'s technique incorporates the effect of bowtie filters, automatic exposure control and electronic noise. The noise insertion method requires scaling the incident photon number and adding normally distributed stochastic noise, as described in their equation:

$$\tilde{P}_B \approx P_A + \sqrt{\frac{1-a}{a} \cdot \frac{\exp(P_A)}{N_{0A}} \cdot (1 + \frac{1+a}{a} \cdot \frac{N_e \cdot \exp(P_A)}{N_{0A}})} x \quad (2)$$

where $\tilde{P}_B$ is the simulated low dose data, P_A is data at a higher tube current setting, a ($0 < a < 1$) is a dose scaling factor, N_{0A} is the incident photon flux corresponding to P_A , x is a normally distributed stochastic process with zero mean and a unit variance, N_e represents the electronic noise.

Directly reproducing this model requires access to raw projection data, incident flux calibration and detailed scanner-specific variables, which are often inaccessible. We, therefore, leveraged the publicly available LDCT image and projection data dataset (challenge dataset) [12], which was generated using the [15] model and we estimate the scanner specific SDCT-to-LDCT transformation using the Pix2Pix [17].

Pix2pix, a conditional generative adversarial network (GAN), was selected for its ability to utilize paired training data to enforce strict anatomical consistency while synthesizing noise textures. The framework, as shown in Fig. 5, consists of 1) a generator (G), composed by U-Net architecture with skip connections, and 2) a discriminator (D) PatchGAN that classified local $N \times N$ image patches as real or fake. In this way, the generator with the help of the discriminator which focuses on local patches, is forced to learn to simulate low dose images noise characteristics. The original objective function minimizes a combination of (a) the conditional adversarial loss (L_{CGAN}), which encourages the generator to produce realistic low-dose images that are indistinguishable from the low-dose domain; and (b) the $L1$ reconstruction loss (L_{L1}) which enforces pixel-wise similarity between the generated LD image output and the paired ground truth image. We included the structural similarity

index (SSIM) loss term in addition to the standard adversarial and $L1$ loss to further enhance the perceptual quality and structural preservation of the CT images. The loss function is defined as:

$$L(G, D) = L_{GAN}(G, D) + \lambda_{L1}L_{L1}(G) + \lambda_{SSIM}L_{SSIM}(G) \quad (3)$$

where λ_{L1} and λ_{SSIM} are the weights for the correspondent losses. In our experiment, we set both $\lambda_{L1} = 50$ and $\lambda_{SSIM} = 50$.

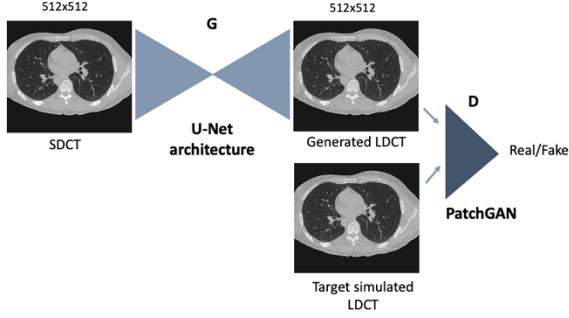


Fig. 5. Architecture of Pix2Pix, composed by a generator (G) and a discriminator (D), used to replicate the degradation method from [15], leveraging paired images provided by the low-dose grand challenge dataset. SDCT: standard-dose CT, LDCT: low-dose CT.

Training strategy

The 50 paired chest CT scans, with consistent 512x512 pixels per slice, from the low-dose grand challenge dataset was partitioned into an 80-20 split for training and validation, respectively. The model was trained for 200 epochs (100 epochs at a constant learning followed by 100 epochs of linear decay) with a batch size of 32. Evaluation was performed using multi-scale SSIM (MS-SSIM) and mean absolute error (MAE) to assess structural accuracy.

B.3. Method 3 – Unpaired image generation using CycleGAN

For the third degradation method (**method 3**), we utilize CycleGAN[18], a popular GAN designed for unpaired image-to-image translation, capable of learning the characteristic noise distribution and artifacts of the LDCT domain directly from unpaired SDCT and LDCT images. CycleGAN, as shown in Fig. 6, is composed of two mapping functions (generators): $G: X \rightarrow Y$ and $F: Y \rightarrow X$; and two associated adversarial discriminators: D_x and D_y . In our study, generator G translates images from the SDCT (domain X) to the LDCT (domain Y), and F performs the reverse translation from LDCT to SDCT. The discriminator D_y aims to distinguish between real LDCT images and generated LDCT images ($G(x)$). In this way, D_y encourages generator G to produce realistic LDCT images that are indistinguishable from the true LDCT images. Similarly, D_x should differentiate between real SDCT images and generated SDCT images $F(y)$, guiding generator F to create realist SDCT images.

The overall objective function minimizes a combination of a) the adversarial loss ($\mathcal{L}_{GAN}$), which encourages the generated

images to match the distribution of images in the target domain (low-dose); and b) cycle consistency loss ($\mathcal{L}_{cyc}$) which enforces the intuition that an image translated from X to Y and back to X (and vice versa) should be close to the original image, i.e., $F(G(x)) \approx x$ and $G(F(y)) \approx y$. The complete loss function is defined as:

$$\mathcal{L}(G, F, D_x, D_y) = \mathcal{L}_{GAN}(G, D_y, X, Y) + \mathcal{L}_{GAN}(F, D_x, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \quad (4)$$

where λ controls the relative importance of the two losses.

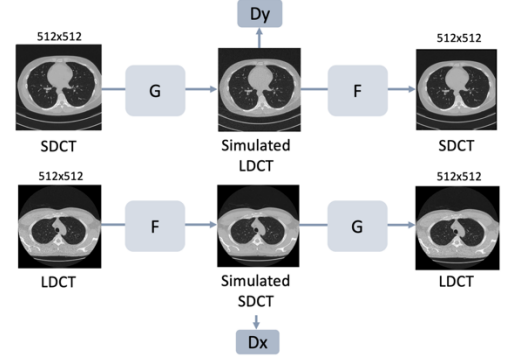


Fig. 6. Schematic representation of the CycleGAN model. Generator G translates images from the SDCT to the LDCT, and F performs the reverse translation from LDCT to SDCT. Discriminator D_y aims to distinguish between real LDCT images and generated LDCT images ($G(x)$) and D_x should differentiate between real SDCT images and generated SDCT images $F(y)$.

Training strategy

The SDCT images were split into 80-20 for training and testing (see Table II), resulting in 556 subjects and 139,301 2D images in the training set and 139 subjects and 35,280 2D images for testing. All the LDCT images, consisting of 315 subjects and 68,302 2D images, were used for training. The input and output image sizes were 512x512. For every 5, 10, 15, 20 and 25 epochs, we used the reserved SDCT test set to evaluate the CycleGAN model using Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) metrics to quantify distributional similarity to the LDCT target domain. The FID score measures the similarity between the feature distributions of the real LDCT images and the generated LDCT images. The lower the FID score, the closer the generated images are to the real data distribution. The KID score uses the maximum mean discrepancy (MMD) to measure the distance between feature distributions. Similar to FID, the lower the KID score, the better the model's performance in generating LDCT images.

TABLE II

NUMBER OF SUBJECTS AND IMAGES FROM THE LIDC DATASET USED IN THE TRAINING AND TESTING CYCLEGAN.

LIDC-IDRI dataset for CycleGAN training	Subjects	Images
Training X – SDCT (80%)	556	139301
Training Y – LDCT	315	68302
Test X – SDCT (20%)	139	35280

B.4. Radiomic pipeline

The three degradation methods were employed as image generator for training a pulmonary nodule classification, following the methodology proposed in our previous work [19]. This classification task ensures a comparative assessment of each simulation method's ability to preserve diagnostically relevant information and enhance classification performance. The pipeline is summarized in Fig.7 and consists of (1) radiomic feature extraction, feature selection, (2) balancing the training set, (3) model training, (4) model validation and (5) evaluation including explainability analysis. Radiomic features were extracted using PyRadiomics[20], following the methodology of our prior work [19] with settings adapted for normalized CT volumes. Extraction was performed in 3D on isotropically resampled volumes ($1 \times 1 \times 1$ mm, sitkBSpline interpolation). Intensity values were Z-score normalized prior to extraction, and gray-level discretization used a fixed bin count of 32. Features were computed on both the original image and wavelet-decomposed images, spanning seven feature classes: shape, first-order, GLCM, GLRLM, GLSZM, NGTDM, and GLDM, yielding 851 features in total. PyRadiomics implements features in accordance with the Image Biomarker Standardisation Initiative (IBSI) guidelines, ensuring reproducibility across implementations. The YAML configuration file used for extraction is provided as supplementary material. Feature selection step included stability and discriminative analysis, the removal of redundant features and the feature selection using absolute shrinkage and selection operator (LASSO), minimum redundancy maximum relevance (MRMR), principal component analysis (PCA), recursive feature elimination (RFE), feature selection using gradient-boosted decision tree (GBDT) and random forest (RF). To address the imbalanced training dataset (1725 non-malignant nodules and 225 malignant nodules), a data augmentation technique introduced in [21] was first applied to the minority class (225 malignant nodules). This augmentation technique consisted in the application of ROI perturbations, including 15% dilation and erosion, as well as random contour adjustments, resulting in a total of 900 malignant nodules. Subsequently, random undersampling was applied to the majority class (1725 benign nodules) to achieve 900 benign nodules. Model training included the training of six classification models: AdaBoost, Decision Tree (DT), k-Nearest Neighbors (KNN), Logistic Regression, Random Forest (RF), and eXtreme Gradient Boosting (XGB).

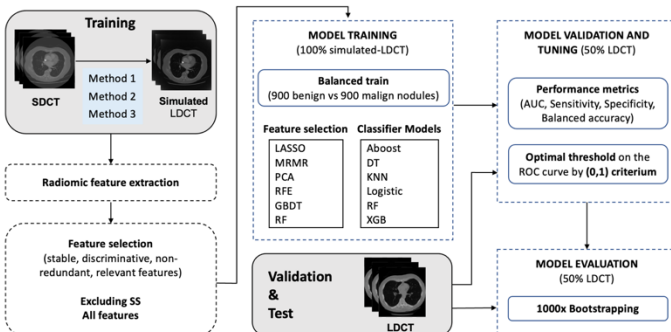


Fig. 7. Pipeline for pulmonary nodule classification using radiomics.

During the validation phase, 50% of the LDCT cases was used to optimize model parameters, including: (1) selecting the optimal decision threshold for the receiver operating characteristic (ROC) curve determined by the minimum Euclidean distance to point (0,1); (2) identifying the optimal feature subset size from 5, 10, 15, 20 features (except for LASSO) based on the area under the ROC curve (AUC); and (3) determining the best feature selection and classifier combination based on a metric obtained by averaging the AUC, balanced accuracy, sensitivity and specificity. The final selected model and feature set for each degradation method were tested on the reserved test set (the remaining 50% of the real LDCT cases). Finally, during testing, 1000 bootstrapping iterations with 10% replacement were computed to obtain the average and 95% confidence interval (CI) for each performance metric.

To ensure experimental validity, we emphasize the strict separation between the image synthesis and classification pipelines. CycleGAN was trained in the SDCT to LDCT direction using real SDCT images as source and real LDCT images as unpaired target-domain references. The resulting synthetic LDCT images, generated from SDCT subjects, constitute the classifier training set. The 315 real LDCT subjects are used exclusively as the classifier validation and test set and are never or included in classifier training. This design ensures that no test subject contributes, directly or indirectly, to classifier training, and that the evaluation reflects genuine generalization to unseen real LDCT data.

B.5. Statistical analysis for method comparison

Non-parametric statistical tests were used to compare the performance of the degradation methods. The primary analysis was conducted using the Friedman non-parametric test to determine if there was a statistically significant difference in the performance among the three degradation methods. If the p -value from the Friedman test was less than 0.05, it indicates that the difference is significant and a pairwise comparison is computed using the Wilcoxon signed-rank test with corrected p -values using the Bonferroni correction.

III. RESULTS

This section presents the results of this study, organized in the following way: (1) evaluation of the degradation methods based on visual and quantitative metrics; (2) performance of the radiomics-based pulmonary nodule classification model when trained on the simulated LDCT data; and (3) statistical comparison of the three degradation approaches.

A. Degradation method evaluation

In this section, we report the intrinsic performance metrics used during model development, as well as the final cross-method qualitative and quantitative comparison.

A.1. Intrinsic results during model development

Due to the distinct architectural nature of each degradation method (deterministic, paired and unpaired), different validation metrics were required for optimization prior to the final comparative analysis.

For method 1, as it was a deterministic noise insertion method rather than a learnable parameter-based model, no optimization metrics were applicable.

For method 2, Pix2Pix paired translation model, performance was evaluated using pixel-wise reconstruction metrics. After 200 epochs, the model achieved a MAE of 0.0961 and a MS-SIM of 0.7192 on the validation.

For method 3, the CycleGAN unpaired translation model was evaluated by translating the reserved SDCT test set to the LDCT domain and comparing it against the real LDCT reference distribution. The training performance over epochs is presented in Table III. The model consistently improved until epoch 20 achieving the lowest FID and KID scores and was thus selected to degrade the SDCT images from the LIDC-IDRI dataset.

TABLE III

RESULTS OBTAINED FROM CYCLEGAN AFTER TRAINING FOR 5, 10, 15, 20 EPOCHS AND 21 EPOCHS.

Testing epoch	FID	KID (mean; std)
Epoch 5	1.0458	2.9562; 1.7077
Epoch 10	0.3505	0.5087; 0.5918
Epoch 15	0.1645	0.4115; 0.7622
Epoch 20	0.0925	0.1304; 0.5235
Epoch 21	0.1891	0.2578; 0.4185

A.2. Cross-method qualitative and quantitative comparison

The visual characteristics of the simulated LDCT patches in comparison to the original SDCT patch are presented in Fig.8. All three methods introduce visible non-uniform noise compared to the original SD patch. Method 1 presents a noticeable fine granular noise, similar to salt-and-pepper noise, while method 2 presents more pronounced, coarse-grained noise with distinct streak artifacts visible in high-density regions, such as chest wall. Method 3 shows blurriness, a characteristic found in LDCT images.

To provide a quantitative comparison, distributional similarity metrics FID and KID scores were computed using 128x128 center-cropped CT slice patches from degraded images, referencing the real LDCT slices as the target ground truth distribution (see Table IV).

The baseline SDCT images obtained an FID of 1.1494 and mean KID of 1.3695. Notably, method 1 and method 2 obtained a higher FID and KID scores compared to the SDCT baseline (method 1 FID: 4.0717; method 2 FID: 4.6729), indicating that while visually these methods introduced noise artifacts, the statistical characteristics of this synthetic noise deviated from the noise of real LDCTs, causing a divergence in the deep feature space. In contrast, method 3 outperformed both the baseline and the paired degradation methods, achieving the

lowest scores (FID: 0.1734; KID: 0.0813). This means method 3 effectively translated SDCT images into a distribution that statistically aligns with the real LDCT target domain. This superior performance is consistent with the validation strategy used for the CycleGAN framework, where FID and KID metrics were utilized as the selection criteria to identify the optimal training checkpoint that maximized distributional alignment.

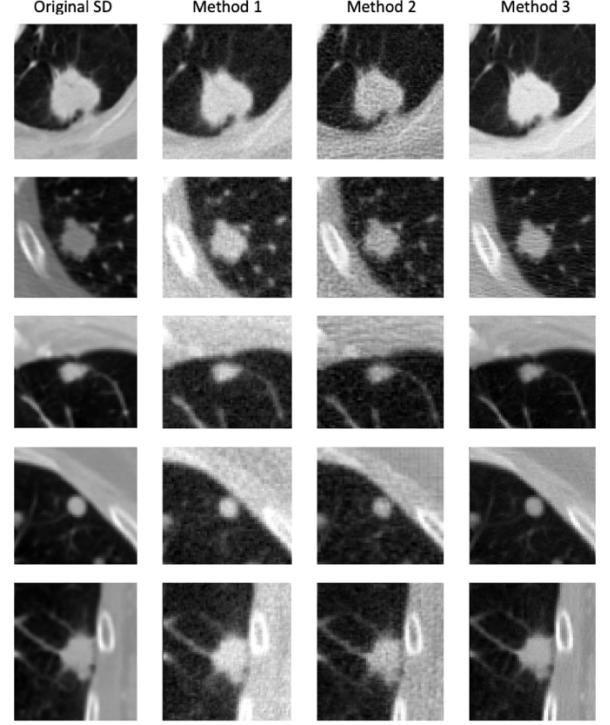


Fig. 8. Visual analysis of nodule patches from degradation methods. The figure compares patches from the original SDCT dataset with the corresponding simulated images generated by method 1 (simple sinogram-domain simulation), method 2 (physics-based degradation model replication using Pix2Pix), and method 3 (CycleGAN).

TABLE IV

FID AND KID SCORES COMPUTED ON IMAGE PATCHES. LOWER SCORES REPRESENT BETTER DISTRIBUTIONAL ALIGNMENT WITH THE REAL LDCT DOMAIN.

	FID	KID (mean $\pm$ std)
SDCT	1.1494	1.3695 $\pm$ 0.5763
Method 1	4.0717	3.0200 $\pm$ 0.5110
Method 2	4.6729	8.4469 $\pm$ 1.1596
Method 3	0.1734	0.0813 $\pm$ 0.1002

B. Radiomics-based pulmonary nodule classification

Prior to model training, radiomic feature analysis including stability, discriminative power, and redundancy removal, was performed on the features extracted from the nodules from each simulated dataset to obtain a smaller set of robust radiomic features [19].

Fig. 9 illustrated the outcome of the feature analysis process across the three degradation methods and the SDCT baseline. Despite the varying noise characteristics introduced by each simulation technique, as described before and shown in Fig. 8, the number of stable, discriminative and non-redundant features identified was consistent among the degradation methods: 99 from methods 1 and 2, and 103 from method 3. In comparison, the SDCT baseline identified 116 features. Table V provides a categorical feature count of the final stable, discriminative and non-redundant feature sets, indicating wavelet-derived features as the most selected across all methods. The intersection ($\cap$) columns highlight the cross-method overlap: 60 features were found in common considering the three degradation methods ($m1 \cap m2 \cap m3$), while 52 of these features were also present in the SDCT baseline ($SDCT \cap m1 \cap m2 \cap m3$), representing the features robust to both degradation and domain shifts. See supplementary materials for detailed feature list.

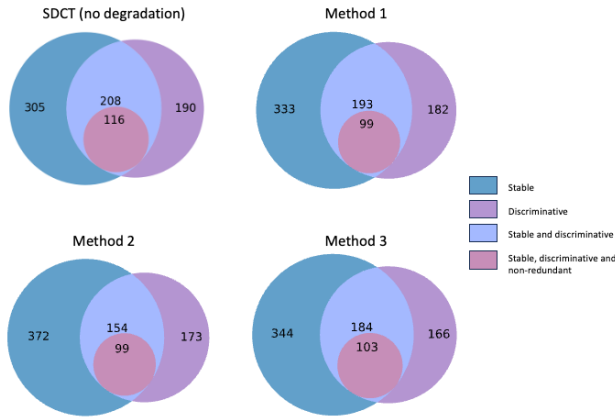


Fig. 9. Venn diagram representing the number of features considered stable, discriminant and non-redundant for each method.

TABLE V

FEATURE COUNTS BY CATEGORY AND CROSS-METHOD INTERSECTION ANALYSIS. THE "M1 $\cap$ M2 $\cap$ M3" COLUMN REPRESENTS FEATURES PRESENT IN ALL THREE DEGRADATION METHODS, WHILE THE FINAL COLUMN INCLUDES THE SDCT BASELINE IN THE INTERSECTION.

	m1	m2	m3	m1 $\cap$ m2 $\cap$ m3	SDCT	SDCT $\cap$ m1 $\cap$ m2 $\cap$ m3
Original Shape	2	2	2	2	2	2
Original First Order	7	7	7	5	5	4
Original Textural	9	9	9	5	7	3
Wavelet	81	81	85	48	102	43
Total	99	99	103	60	116	52

m1: method 1; m2: method 2; m3: method 3

The optimal feature set subset size (5, 10, 15, 20) and the best pair of feature selector and classifier combination are reported in the supplementary material. The validation and test results for each degradation scenario are summarized in Table VI. Performance metrics for the test set are reported as average and 95% CI from 1000 bootstrapped samples.

The overall classification performance on the real LDCT test set highlights the importance of degrading the SDCT images for domain-specific data augmentation. The model trained on baseline SDCT (without degradation) performed well during validation (AUC 0.863) but decreased on the testing phase (AUC 0.789). Specifically, the sensitivity decreased to 0.571 which indicates that the model failed to classify the nodules from the LDCT test set. A detailed statistical comparison of these results is presented in the following section.

On the other hand, all three degradation methods effectively simulated LDCT features to train a robust nodule classification model, achieving a test AUC above 0.84 for all the methods and settings. From validation to testing, the AUC metric remained robust across all method. Method 1 maintained an identical AUC (0.844) on both validation and test, while method 2 and method 3 improved slightly by 1.6% and 0.7%, respectively. Conversely, balanced accuracy showed general decrease across all methods, with method 1 obtaining a decrease of 4.6%, method 2 of 3.3%, and method 3 showing the most stability with only a reduction of only 1.4%. In terms of sensitivity, method 1 and 2 showed a notable decrease of 10.9% and 8.1%, respectively, in the test set. This decrease was compensated by high specificity, increasing 1.7% for method 1 and 1.3% for method 2. Method 3 demonstrated the most balanced performance on the independent test set, maintaining the highest sensitivity (0.743) and balanced accuracy (0.800), with increases in AUC (0.7%) and specificity (5.3%) from validation to test. This suggests that the unpaired CycleGAN simulation (method 3) may preserve subtle nodule features more effectively than the paired approaches, reducing the rate of false negatives.

C. Statistical test for degradation methods comparison

The performance of the three degradation methods was formally compared using the 1000 bootstrapped test set AUC, balanced accuracy, sensitivity and specificity values.

For every metric, the Friedman test consistently yielded a p -value <0.05 (see supplementary materials) when comparing the three degradation methods, indicating that at least one method's performance distribution was statistically significantly different from the others. Subsequent Bonferroni-corrected pairwise Wilcoxon signed-rank tests confirmed these differences, reporting p -value <0.05 for all metrics (see supplementary materials).

TABLE VI
NODULE CLASSIFICATION VALIDATION AND TEST RESULTS.

	AUC	Balanced accuracy	Sensitivity	Specificity
SDCT				
validation	0.863	0.814	0.824	0.805
test	0.789 [0.750; 0.827]	0.744 [0.707; 0.781]	0.571 [0.500; 0.643]	0.917 [0.904; 0.930]
Method 1				
validation	0.844	0.813	0.765	0.861
test	0.844 [0.804; 0.886]	0.767 [0.731; 0.803]	0.656 [0.586; 0.727]	0.878 [0.862; 0.893]
Method 2				
validation	0.843	0.821	0.765	0.878
test	0.859 [0.824; 0.900]	0.788 [0.752; 0.822]	0.684 [0.618; 0.757]	0.891 [0.877; 0.905]
Method 3				
validation	0.854	0.814	0.824	0.805
test	0.861 [0.826; 0.900]	0.800 [0.768; 0.836]	0.743 [0.684; 0.811]	0.858 [0.842; 0.875]

Fig. 10 represents bootstrapped distribution violin plots, highlighting a distinct domain shift between the SDCT baseline and the degradation methods. The SDCT baseline showed lower distribution across most metrics compared to the degradation approaches. More specifically, the SDCT baseline presented a drop in sensitivity distribution (0.571 [0.500; 0.643]), indicating a failure to classify nodules in the noisy LDCT domain.

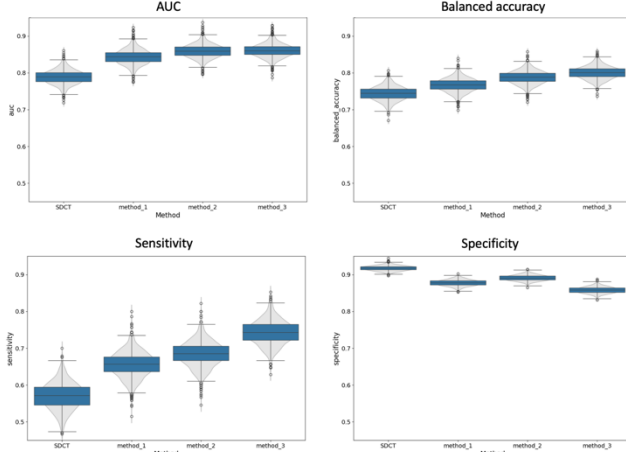


Fig. 10. Bootstrapped distribution violin plots of the nodule classification model on the independent real LDCT test set. The boxplots illustrate the variability across 1000 bootstrap iterations for AUC, balanced accuracy, sensitivity and specificity.

Focusing on comparing the degradation methods, AUC distributions for methods 2 and 3 are visually similar distributions with highly overlapping confidence intervals (method 2: 0.859 [0.824; 0.900] and method 3: 0.861 [0.826; 0.900]), both slightly superior to method 1 (0.844 [0.804; 0.886]). In terms of balanced accuracy, a sequential increase is observed from method 1 to method 3, with method 3 achieving the highest results: method 3 (0.800 [0.768; 0.836]) > method 2 (0.788 [0.752; 0.822]) > method 1 (0.767 [0.731; 0.803]). Sensitivity and specificity reveal a distinct performance trade-off as well as a difference in variability. A stepwise increase is also observed in sensitivity moving from method 1 to method 3, where method 3 achieves the highest median sensitivity of 0.743 [0.684; 0.811]. However, compared to the other metrics, sensitivity exhibits the widest confidence intervals, indicating greater variability in the model's ability to detect positive cases.

Conversely, the specificity plot shows the inverse trend characterized by significantly narrower confidence intervals, reflecting higher stability. Method 2 dominates with the highest specificity (0.891 [0.877; 0.905]), while Method 3 shows a lower, albeit still robust, specificity distribution (0.858 [0.842; 0.875]).

Despite the partial overlap in the 95% confidence intervals, the statistical analysis confirms that the underlying performance distributions are distinct. The consistent shift in the distributions confirms that method 3 effectively prioritizes the detection of true positives (sensitivity) compared to methods 1 and 2.

IV. DISCUSSION

This study provided a systematic comparison of different LDCT degradation methods for synthetic image generation and evaluated the efficacy in improving nodule classification models using radiomic features extracted from those synthetic images. We included approaches ranging from sinogram-domain statistical simulation (method 1), to the replication of a validated physics-based method using Pix2Pix (method 2), to unpaired image generation using CycleGAN (method 3). In this way, we bypassed the conventional denoising step and instead validated the feasibility of using LDCT images to develop downstream classification models.

The main contributions of this paper are:

- We provide the first systematic comparison and evaluation of different low-dose CT degradation methods.
- We introduce an approach that bypasses denoising and instead validates the feasibility of using SDCT images for developing downstream classification models.
- By providing a reliable method for simulating LDCT scans, our findings will not only aid in the development of more accurate denoising algorithms but also accelerate the inclusion of SDCT images into the development of nodule detection and classification models, ultimately improving the efficacy of lung cancer screening.

The need for realistic degradation was confirmed by the poor generalization of the model trained on SDCT data, i.e., images without any degradation. While the SDCT baseline model performed well during validation, it failed to generalize to the test set, yielding a collapsed sensitivity of 0.571. In contrast, the results demonstrated that synthetic data generated by all three

methods can successfully train a classifier to perform well on real LDCT data (AUC > 0.84). However, the methods showed distinct qualitative visual features and distinct quantitative performance in the nodule classification task. Method 1 produced visible but more subtle granular noise compared to method 2, which replicated pronounced streak artifacts and high-frequency noise that appeared visually consistent with the photon starvation. However, despite visible noise in methods 2, it yielded relatively lower sensitivity. Conversely, method 3 produced perceptually blurrier images with subtle noise but obtained the best FID (0.0316) and KID (0.0102) scores computed on nodule. This indicates that while method 3 may lack high-frequency sharpness, it achieved superior distributional alignment with the target domain. In the nodule classification task, method 3 yielded the most balanced classification performance using with the highest sensitivity, achieving a test set AUC of 0.861, balanced accuracy of 0.800, sensitivity of 0.743, specificity of 0.858.

In terms of radiomic feature analysis, despite the visual differences in the images generated by different methods, the final features identified as stable, discriminant and non-redundant were similar, with 60 features found in common. The dominance of wavelet-derived features (80% of this common set), specifically those with low-pass components (92 low-pass directional filter components compared to 52 high-pass components), indicates that the underlying structural characteristics of the nodules are preserved regardless of the degradation technique, while high-frequency noise variations are less relevant for discriminating features. When considering the feature subsets selected to train the classifiers, only two features were commonly selected across all methods, *original shape major axis length* and *original shape sphericity*. This highlights the importance of these SS features and is coherent with the criteria of the Lung Imaging Reporting and Data System (Lung-RADS) [22], which primarily quantifies tumor risk based on nodule diameter and morphological regularity. These findings are consistent with our previous work [19]. Further details regarding the selected features can be found in the supplementary materials.

This study is not exempt from limitations. First, the simulation methods relied on image-domain or simplified statistic assumptions in the sinogram domain, future work should consider scanner specific parameters or leverage deep learning in the projection domain to better include non-uniform realistic LDCT noise. Second, regarding the statistical analysis, although the Friedman and pairwise Wilcoxon tests indicated significant differences, the overlap in the 95% confidence intervals suggest that these differences might be partially attributed to the bootstrap sample size which makes the statistical tests more susceptible to small differences. Finally, while the classification results are promising, the sensitivity (0.743) of the models could be improved, despite being comparable to previous findings [19]. However, the high specificity (0.858) is clinically significant, as a major limitation of LDCT screening programs is the high rate of false positives leading to unnecessary interventions. Future work should validate the proposed pipeline on independent annotated LDCT datasets such as LUNA25 to assess generalizability of the model.

V. CONCLUSION

Our findings confirm that generating synthetic LDCT data from SDCT is a viable strategy for development of nodule classification models for screening-detected nodules. This work supports the use of unpaired generative adversarial networks as a scalable, cost-effective solution to the data scarcity in LDCT imaging.

REFERENCES

- [1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.
- [2] H. J. de Koning *et al.*, "Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial," *New England Journal of Medicine*, vol. 382, no. 6, pp. 503–513, Feb. 2020, doi: 10.1056/NEJMOA1911793;WGROU:STRING:MMS.
- [3] The National Lung Screening Trial Research Team, "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, Aug. 2011, doi: 10.1056/NEJMOA1102873;PAGEGROUP:STRING:PUBLICATION.
- [4] W. Kim, S. Y. Jeon, G. Byun, H. Yoo, and J. H. Choi, "A systematic review of deep learning-based denoising for low-dose computed tomography from a perceptual quality perspective," *Biomedical Engineering Letters* 2024 14:6, vol. 14, no. 6, pp. 1153–1173, Aug. 2024, doi: 10.1007/S13534-024-00419-7.
- [5] K. A. S. H. Kulathilake, N. A. Abdullah, A. Q. M. Sabri, and K. W. Lai, "A review on Deep Learning approaches for low-dose Computed Tomography restoration," *Complex & intelligent systems*, vol. 9, no. 3, pp. 2713–2745, Jun. 2023, doi: 10.1007/S40747-021-00405-X.
- [6] T. Zhao, M. McNitt-Gray, and D. Ruan, "A convolutional neural network for ultra-low-dose CT denoising and emphysema screening," *Med. Phys.*, vol. 46, no. 9, pp. 3941–3950, Sep. 2019, doi: 10.1002/MP.13666;JOURNAL:JOURNAL:24734209;ISSUE:ISSUE:DOI.
- [7] M. Gholizadeh-Ansari, J. Alirezaie, and P. Babyn, "Deep Learning for Low-Dose CT Denoising Using Perceptual Loss and Edge Detection Layer," *Journal of Digital Imaging* 2019 33:2, vol. 33, no. 2, pp. 504–515, Sep. 2019, doi: 10.1007/S10278-019-00274-4.
- [8] H. Chen *et al.*, "Low-dose CT denoising with convolutional neural network," *Proceedings - International Symposium on Biomedical Imaging*, pp. 143–146, Jun. 2017, doi: 10.1109/ISBI.2017.7950488.
- [9] E. Eulig, B. Ommer, and M. Kachelrieß, "Benchmarking deep learning-based low-dose CT image denoising algorithms," *Med. Phys.*, vol. 51, no. 12, pp. 8776–8788, Dec. 2024, doi: 10.1002/MP.17379;WGROU:STRING:PUBLICATION.
- [10] C. H. McCollough *et al.*, "Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge," *Med. Phys.*, vol. 44, no. 10, pp. e339–e352, Oct. 2017, doi: 10.1002/MP.12345.
- [11] S. G. Armato *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Med. Phys.*, vol. 38, no. 2, 2011, doi: 10.1118/1.3528204.
- [12] T. R. Moen *et al.*, "Low-dose CT image and projection dataset," *Med. Phys.*, vol. 48, no. 2, pp. 902–911, Feb. 2021, doi: 10.1002/MP.14594.
- [13] T. Snawsill *et al.*, "Low-dose computed tomography for lung cancer screening in high-risk populations: a systematic review and economic evaluation," *Health Technol. Assess.*, vol. 22, no. 69, pp. 1–276, Nov. 2018, doi: 10.3310/HTA22690.
- [14] C. Rampinelli, D. Origi, and M. Bellomi, "Low-dose CT: technique, reading methods and image interpretation," *Cancer Imaging*, vol. 12, no. 3, p. 548, 2013, doi: 10.1102/1470-7330.2012.0049.
- [15] 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," *J. Comput. Assist. Tomogr.*, vol. 36, no. 4, pp. 477–487, 2012, doi: 10.1097/RCT.0B013E318258E891.

-
- [16] D. Zeng *et al.*, “A simple low-dose X-Ray CT simulation from high-dose scan,” *IEEE Trans. Nucl. Sci.*, vol. 62, no. 5, pp. 2226–2233, Oct. 2015, doi: 10.1109/TNS.2015.2467219.
 - [17] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, doi: 10.1109/CVPR.2017.632.
 - [18] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 2242–2251, Mar. 2017, doi: 10.1109/ICCV.2017.244.
 - [19] J. Liu, A. Corti, V. D. A. Corino, and L. Mainardi, “Lung nodule classification using radiomics model trained on degraded SDCT images,” *Comput. Methods Programs Biomed.*, vol. 257, p. 108474, Dec. 2024, doi: 10.1016/J.CMPB.2024.108474.
 - [20] J. J. M. van Griethuysen *et al.*, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.
 - [21] 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 Informatics in Medicine* 2024 37:5, vol. 37, no. 5, pp. 2401–2414, May 2024, doi: 10.1007/S10278-024-01013-0.
 - [22] J. Christensen *et al.*, “ACR Lung-RADS v2022: Assessment Categories and Management Recommendations,” *Chest*, vol. 165, no. 3, pp. 738–753, Mar. 2024, doi: 10.1016/j.chest.2023.10.028.

3 Leveraging self-supervised pretraining using transformers for enhanced lung nodule detection in CT scans



Leveraging Self-supervised Pretraining Using Transformers for Enhanced Lung Nodule Detection in CT Scans

Jiaying Liu^{1(✉)}, Qi Ma^{2,3}, Anna Corti¹, Valentina D. A. Corino^{1,4},
Luca Mainardi¹, and Ender Konukoglu²

¹ B3Lab, DEIB, Politecnico di Milano, Milan, Italy
jiaying.liu@polimi.it

² Computer Vision Lab, ETH Zürich, Zürich, Switzerland
qi.ma@vision.ee.ethz.ch

³ INSAIT, Sofia University “St. Kliment Ohridski”, Sofia, Bulgaria

⁴ Cardiotech Lab, Centro Cardiologico Monzino IRCCS, Milan, Italy

Abstract. Lung nodule detection is critical for early diagnosis of lung cancer, but remains challenging due to the nodules resemblance to normal tissues. Recent transformer-based approaches have made significant progress; however, their large number of parameters necessitates extensive annotated datasets to achieve robust and reliable results. To address this, we leverage state-of-the-art self-supervised training methods, specifically Masked Image Modeling, on a large domain-specific dataset of lung screening CTs, followed by finetuning on the annotated LUNA16 dataset. Our method achieves an AP of 82.63% and an mAP of 81.23%, outperforming the baseline nnDetection. The experiments demonstrate the effectiveness of pretraining, yielding an increase of 24.0% in performance on the Video-ViT backbone and 4.1% on the Swin Transformer. Additionally, we examine the effect of RGB video pretraining and architectural variations during both pretraining and fine-tuning stages. This work highlights the potential of self-supervised learning in improving efficiency and accuracy in lung cancer screening. Code: github.com/vit-swin-lung-nodule-detection.

Keywords: lung nodule detection · self-supervise pretraining · transformers

1 Introduction

Lung cancer was the most diagnosed cancer in 2022 and the leading cause of cancer deaths [29]. Early detection through low-dose computed tomography (CT) screening can reduce mortality [26], but the high volume of scans burdens radiologists. Identifying lung nodules is challenging due to their small size and similarity to normal lung tissue and vessels, leading to high false positives [22].

J. Liu and Q. Ma—These authors contributed equally to this work and share first authorship.

Automated diagnostic tools powered by artificial intelligence have improved the accuracy and efficiency of lung cancer screening [2]. However, the reliability of these models depends on the quality and quantity of the dataset, which remains costly and time-consuming to collect and curate. State-of-the-art transformer-based methods have shown superior performance, but their effectiveness is still influenced by data limitations, requiring robust pretraining strategies [12, 19, 20]. To address data insufficiency, various self-supervised and semi-supervised approaches [5, 6, 8, 9, 21] have been proposed to leverage large amounts of unlabeled data. Among them, Masked Image Modeling (MIM) [7, 10, 28] has proven effective by training models to reconstruct masked regions of inputs, forcing them to learn rich, contextual visual representations, leading to improved performance for downstream tasks in 3D Medical reasoning such as classification [4], and segmentation [18].

In this work, we incorporate self-supervised pretraining into transformer backbones for lung nodule detection building on the nnDetection framework. While most closely related studies [4, 18] employ domain-adapted MIM pretraining, such as Video Vision Transformer (Video-ViT) on K700 [13], for tasks like vertebral fracture detection or dental segmentation, our goal is to assess it for 3D detection. CNN-based approaches such as Model Genesis were not considered, as our focus is on evaluating transformer-specific architectures. Despite DINOv2’s promising results in 2D vision tasks, it involves significantly higher computational for 3D medical detection. Therefore, we extend prior work by demonstrating the utility of large-scale masked pretraining for 3D lung nodule detection and assess its effectiveness under architectural variations, specifically models with or without skip connections, and across input resolutions of 16224224 and 64128128 volumes. The main contribution of this paper are:

- We are among the first to adapt MIM-based pretraining to a 3D nodule detection task and demonstrate its effectiveness on transformer-based backbones such as Video-ViT and SwinUNETR.
- The proposed framework leverages the spatial reasoning capabilities of the MIM mechanism, leading to impressive finetuning results on the LUNA16 dataset, outperforming the Retina U-net baseline.
- Through extensive experiments, we analyze the impact of different architectural designs and input patch sizes on model’s performance. Our method advances the use of pretrained transformers for medical detection tasks.

2 Methods

2.1 Dataset and Preprocessing

Table 1 lists the datasets used in our study. For pretraining, we used the unlabeled dataset from National Lung Screening Trial (NLST) [25], containing low-dose CT scans from 26,254 participants screened over 3 years [1]. We included 24,820 low-dose CT scans, selecting only the latest scan per participant to avoid redundancy, as only 1.1% were diagnosed with cancer. Corrupted scans were

Table 1. Datasets used for pretaining and finetuning.

Task	Dataset	Size	Scans	Modality	Slice thickness	Annotations
Pretraining	NLST [25]	3.78 TB	24,820	low-dose CT	≥ 1.25 mm	-
Finetuning	LUNA16 [22]	112 GB	888	CT	≥ 2.5 mm	1186 nodules

excluded. For finetuning, we used the LUNA16 dataset which is composed by 888 CT scans containing 1186 nodules annotated by multiple radiologists [22].

The preprocessing pipeline followed the work of [16] and was applied to both datasets. It involved steps such as converting raw CT data to Hounsfield Units, segmenting the lungs, padding, bone removal, intensity normalization, resampling, and background cropping to ensure standardized intensity, consistent resolution, and higher processing speed.

2.2 Self-supervise Pretraining

Masked Image Modeling. As shown in Fig. 1, given a 3D CT scan $X \in \mathbb{R}^{H \times W \times D}$ with height H , width W , and depth D , we randomly crop a volume $X_c = \text{Crop}(X)$, where $X_c \in \mathbb{R}^{H_c \times W_c \times D_c}$. Here, H_c, W_c, D_c are the cropped dimensions. The cropped CT scan undergoes a patchify operation $\text{patchify}(\cdot)$, followed by an embedding layer $P(\cdot)$, which maps the patches into a latent space: $E = P(\text{patchify}(X_c)) \in \mathbb{R}^{F \times \hat{H} \times \hat{W} \times \hat{D}}$, where F is the feature dimension and $\{\hat{H}, \hat{W}, \hat{D}\}$ are number of patches cross dimension. Then, we apply a random masking with a ratio $r \in [0, 1]$ to create masked tokens $E_m \in \mathbb{R}^{F \times r \hat{H} \hat{W} \hat{D}}$ and visible tokens $E_v \in \mathbb{R}^{F \times (1-r) \hat{H} \hat{W} \hat{D}}$. The corresponding masked input patches are X_m . The spatial relationships are preserved while r fraction of tokens are removed. Next, the tokenizer processes embeddings with positional information: $T = \text{Tokenizer}(E) = \text{Flatten}(E) + \text{SpatialPE}(\hat{H}, \hat{W}) + \text{TemporalPE}(\hat{D})$. Here, SpatialPE encodes spatial coordinates within the transverse plane (e.g., leftright and anteriorposterior axes), while TemporalPE captures positional information along the superiorinferior axis, now treated as the temporal dimension. The visible tokens, $T_v \in \mathbb{R}^{d_z}$, are encoded to a latent representation: $z = f_\theta(T_v) \in \mathbb{R}^{d_z}$, where f_θ is the encoder network and d_z is the latent dimension. Afterwards, the latent code is concatenated with learnable token, $T_l \in \mathbb{R}^{d_z}$, to obtain mask tokens ($\hat{T}_m$) for prediction: $\hat{T}_m = g_\phi(z \oplus T_l) \in \mathbb{R}^{r \hat{H} \hat{W} \hat{D}}$ where $\oplus$ denotes concatenation and g_ϕ is the decoder network. Then, a reconstruction head (Φ) maps the predicted tokens back to the input space: $\hat{X}_m = \Phi(\hat{T}_m) \in \mathbb{R}^{F \times r \hat{H} \hat{W} \hat{D}}$. Finally, we minimize the reconstruction loss ($\mathcal{L}_2$) between the predicted and original masked input: $\mathcal{L}_2 = \min_{\theta, \phi} \mathbb{E} \left[\|X_m - \hat{X}_m\|_2^2 + \lambda \mathcal{L}_{\text{reg}} \right]$, where $\mathcal{L}_{\text{reg}}$ is an optional regularization term and λ is a weighting coefficient.

Large scale Video Dataset Pretraining. Building on previous works [4, 14, 18], we leverage the transformer model Video-ViT Large [7], pretrained

on the large-scale video dataset Kinetics-700 [13]. This model demonstrates strong performance when adapted to medical domain tasks [4, 14, 18]. Notably, the model is pretrained on inputs of shape 16224224 during the video pretraining stage.

Large scale Lung Dataset Pretraining. While the spatial reasoning learned from the MIM method on natural videos benefits medical image applications, a domain gap remains. CT scans, for example, exhibit distinct statistical properties, such as differing means and variances, requiring the embedding layer and reconstruction head to adapt accordingly.

To address this, we further pretrain the model in an MIM manner on the target domain. Our method differs from [4, 18] in the following aspects: **(1)** We redesign the kernel sizes in the embedding layer $P(\cdot)$ (originally $2 \times 16 \times 16$) to accommodate different input shapes. This adjustment allows training on larger input sizes. **(2)** we use a larger dataset for pretraining; **(3)** we investigate multiple transformer backbones with different architecture designs. Our findings highlight the potential of self-supervised learning to improve lung nodule detection by leveraging video-based and domain-specific pretraining. We conduct ablation studies with two input sizes: $16 \times 224 \times 224$ and $64 \times 128 \times 128$.

We further investigate pretraining mechanisms using SwinUNETR [24], a 3D transformer-based model originally designed for medical image analysis. To adapt it for self-supervised learning, we follow the original implementation which integrates two auxiliary losses: contrastive loss and rotation prediction loss. These losses enhance the models ability to learn robust representations transferable to downstream tasks such as nodule detection.

2.3 Finetuning

After pretraining, we evaluate the pretrained encoder on detection tasks using Retina U-net decoder [11, 17] which is also exploited in nnDetection [3], a state-of-the-art medical object detection framework that automates hyperparameter tuning. In addition to the box regression loss and classification loss, we incorporate a multi-task segmentation loss by leveraging an additional decoder head. The segmentation map is generated by box center and a given diameter, which has been shown to be beneficial in the work of [3].

Building on insights from [15], which explored how decoder architectures impact vision transformer (ViT) encoder performance in 2D tasks, we design two decoder configurations (see Fig. 2) tailored for 3D medical imaging: 1) plain backbone with skip connection: the intermediate features in layers [5, 11, 17, 23] are connected to the hierarchical decoder; 2) plain backbone without skip connection: only the last layer of encoder is used and connected to hierarchical decoder. To connect the encoder and decoder, we use 3D transposed convolutions to upsample spatiotemporal features, aligning them with the Retina U-Net decoder. When the spatial dimensions are already aligned, we apply two 3D convolutional layers. If not specified, we follow work [15] and use 1) as the default.

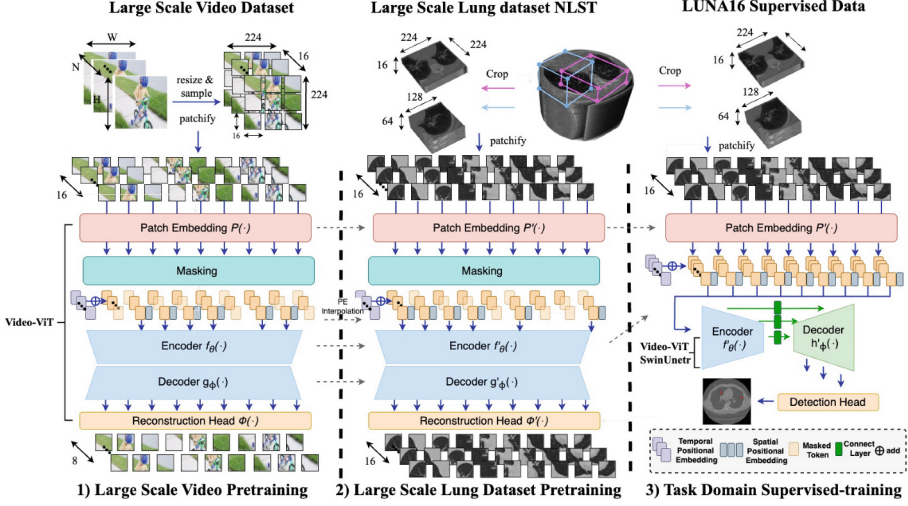


Fig. 1. Illustration of the Video-MAE pretraining mechanism. A high proportion of 3D patches are randomly masked, and the model is trained to reconstruct them. Figure adapted from [7].

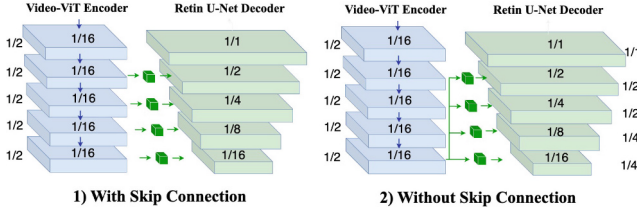


Fig. 2. Illustration of Skip and Non-Skip connections used to link the transformer encoder to the detector head.

3 Experiments

3.1 Pretraining

Video-Vit model was initialized with weights pretrained on the Kinetics-700. We further trained the model on the NLST dataset for 150 epochs (approx. 2h/epoch) using NVIDIA RTX A6000 GPU. The training hyparameters were set as follows: learning rate of $1e-5$, weight decay of $5e-4$ and warmup period of 10 epochs, using the AdamW optimizer and a cosine learning rate scheduler, random mask ratio of $r = 0.9$. $\mathcal{L}_2$ reconstruction loss was considered to minimizing the squared difference between reconstructed and original images. For the 16224224 input size, a 3D convolution (21616 kernel) output 8 temporal tokens and 1414 spatial tokens (81414 tokens per volume). For 64128128, the embedding layer was reinitialized with a 81616 kernel, producing 8 temporal tokens and 88 spatial tokens (888 tokens per volume). The spatial positional embedding is interpolated

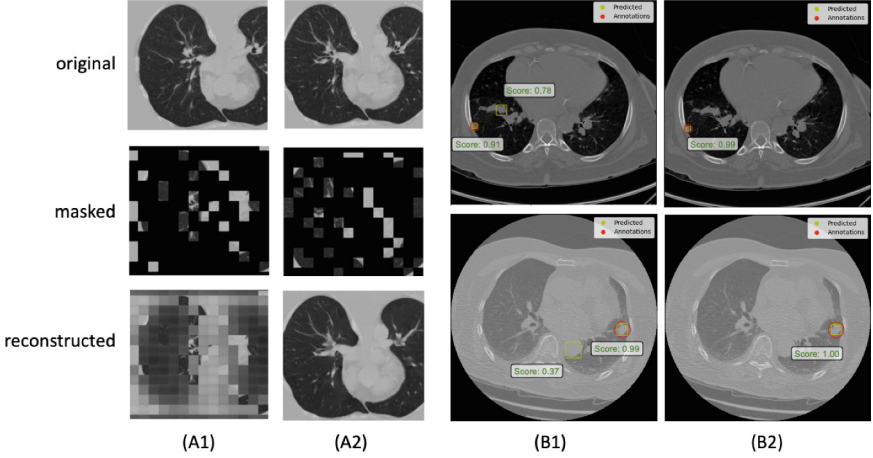


Fig. 3. Examples of Video-Vit reconstructions during pretraining (A1) from scratch ($\mathcal{L}_2 = 4.13e - 2$) and (A2) using video weights ($\mathcal{L}_2 = 6.04e - 4$). Nodules detected during finetuning using (B1) scratch model and (B2) lung pretrained weights.

from 14×14 to 8×8 using bilinear interpolation. A batch size of 64 per GPU was used for the former input and 16 for the later case.

SwinUNETR was initialized with the official MONAI weights pretrained on multiple medical imaging datasets, including lung CTs. Thus, we only pretrained with NLST for 20 epochs (approx. 2h/epoch) using NVIDIA RTX A6000. We use crop size of $64 \times 128 \times 128$ and after a 6-layer encoder, it outputs a shape of $2 \times 4 \times 4$ with 768 channels. The training hyperparameters were set as follows: learning rate of $1e - 6$, weight decay of 0.1 and momentum of 0.9, using warmup cosine learning rate scheduler with 500 warmup steps, and gradient clipping ($\text{max_norm} = 1.0$). A batch size of 2 per GPU was used.

3.2 Finetuning

We report finetuning results on LUNA16, using subsets 1 to 9 for training and subset 10 for testing. The finetuning was conducted for 60 epochs, with 5000 mini-batches per epoch. For optimization, we used AdamW (learning rate: $1e - 4$) for transformer-based methods and SGD (learning rate: $1e - 2$) for the Retina U-Net baseline. A cosine annealing learning rate scheduler with a 10-epoch warm-up was applied. Further implementation details can be found in our [code](#).

Model performance was evaluated using the following metrics: **(1)** average precision (AP) at IoU 0.1; **(2)** mean average precision (mAP) calculated over IoU thresholds from 0.1 to 0.5 with increments of 0.05; and **(3)** free-response operating characteristics (FROC) at IoU 0.1, which measures detection sensitivity across varying false positive rates.

4 Results and Discussion

4.1 Model Performance Across Different Configurations

We present qualitative and quantitative pretraining results in Fig. 3. The $\mathcal{L}_2$ reconstruction loss, relative to the maximum input intensity, is 4% in (A1) training from scratch, i.e., without using any pretrain weights and 0.01% in (A2) with large-scale video pretrained weights. The pretrained model significantly enhances reconstruction quality, accurately reconstructing lung structures, including fine details like blood vessels. This is possible using only 10% of visible tokens since different patches are masked in successive slices. However, training from scratch exhibits severe patch artifacts. In Fig. 3, we also compare nodule detection results from Video-ViT trained from scratch (B1) which detects more false positives, and with pretrained weights (B2). The lung pretraining weights initialized with video pretrain weights (V+L pretrain) model (B2) successfully detects small nodules without suffering from similar anatomical structure.

Table 2. Comparison of lung nodule detection performance across different pretraining strategies and input configurations.

Model	mAP _{IoU0.10-0.50}	AP _{IoU0.10}	FROC _{IoU0.10}
Input shape - 16224224			
Retina U-net	75.231	76.723	80.867
V-Vit (scratch)	65.523	70.881	74.617
V-Vit (V+L pretrain)	81.234	82.629	<u>84.184</u>
Input shape - 64128128			
Retina U-net	<u>78.924</u>	81.135	85.331
V-Vit (scratch)	67.303	70.570	72.959
V-Vit (V+L pretrain)	72.550	75.258	77.806
Swin (scratch)	75.596	78.320	81.505
Swin (M+L pretrain)	78.723	<u>81.327</u>	77.806
V-Vit: Video-Vit; Swin: SwinUNETR; V: video; L: lung; M: MONAI.			

Table 2 presents the detection performance including the Retina U-net baseline, our model from scratch, and using pretrained weights. Among the configurations, Video-Vit (V+L pretrain) achieved the best results with a mAP of 81.234% and an AP of 82.629% with $16 \times 224 \times 224$ input, aligning with the resolution used in video pretraining. In contrast, the $64 \times 128 \times 128$ configuration requires partial reinitialization of embedding layer, reducing the benefit of pretraining. On the other hand, while we pretrained both transformers the same way, Video-Vit (V+L pretraining) outperformed SwinUNETR primarily for two reasons: 1) the medical image dataset used to pretrain SwinUNETR (5050 CT volumes) is significantly smaller compared to the video dataset (500,000 video clips); 2) SwinUNETR model has less trainable parameters (26.7 million) than Video-Vit (338

million). Our results are comparable to recent approaches in the literature. For instance, Wu et al. [27] proposed YOLOv7-based framework using 2D images, achieving a mAP of 80.94% and 95.26% with data augmentation. Similarly, Ji et al. [12] modified YOLOv5 and achieved a mAP of 72.0%. Su et al. [23] applied Faster R-CNN to the LIDC-IDRI dataset and achieved an AP of 82.4% for the original model and 91.2% for the optimized version. Some of these methods rely on 2D CNN architectures, while our 3D transformer-based approach achieves a mAP of 81.234% and AP of 82.629%, highlighting its competitiveness. Future work could explore pretraining a Video-Vit with $64 \times 128 \times 128$ input size using the video dataset and complete 10-fold cross-validation to enable more direct comparisons with studies that report results using the official challenge metrics.

4.2 Model Performance Freezing the Encoder

Table 3 presents the performance of transformer encoders under a frozen-encoder setting, using linear probing, where only the last layer is retrained to evaluate feature representation effectiveness with minimal finetuning.

Table 3. Evaluation of detection performance with a frozen encoder using different pretraining weights and architectural configuration (with or without skip connection).

Model	Skip	mAP _{IoU0.10–0.50}	AP _{IoU0.10}	FROC _{IoU0.10}
Input shape - 16224224				
V-Vit (V pretrain)	no	73.425	76.649	78.700
V-Vit (V+L pretrain)	no	74.139	78.050	82.015
V-Vit (V+L pretrain)	yes	75.772	78.558	81.633
Input shape - 64128128				
V-Vit (V pretrain)	no	64.308	67.659	71.173
V-Vit (V+L pretrain)	no	64.503	67.860	71.301
V-Vit (V+L pretrain)	yes	64.100	67.399	71.301
Swin (M pretrain)	yes	69.573	72.848	75.510
Swin (M+L pretrain)	yes	68.830	71.755	74.745
V-Vit: Video-Vit; Swin: SwinUNETR; V: video; L: lung; M: MONAI.				

Results show when the encoder is frozen, domain-specific pretraining is beneficial, especially for models (Video-Vit) not already exposed to lung CT data. However, for SwinUNETR initialized with MONAI weights, further in-domain pretraining does not yield improvements, which is not a surprising finding since MONAI weights were determined with medical imaging data including Lung CT. Additionally, we evaluated the models with and without skip connections between the transformer encoder and detector head. Results in Table 3 indicate that skip connections improve performance for smaller temporal input sizes

$(16 \times 224 \times 224)$, but become less beneficial as the temporal input size increases $(64 \times 128 \times 128)$.

5 Conclusion

In this work, we explore the effectiveness of self-supervised pretraining using Video-ViT and SwinUNETR in lung nodule detection. Results show that pretraining in video and lung yields an increase of 24.0% in performance on Video-ViT backbone and 4.1% on SwinUNETR. Video-ViT backbone achieved an AP of 82.629% and a mAP of 81.234%, surpassing the baseline and confirming the effectiveness of leveraging external large-scale and domain-specific data. To conclude, this work highlights the potential of self-supervised learning to improve lung nodule detection, contributing toward the development of more efficient and accurate lung cancer screening tools.

Acknowledgments. The authors acknowledge the support of Align Technology for their funding and resources that contributed to this research. AC is funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvanCed Technologies for Human-centrEd Medicine (ANTHEM)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aberle, D.R., et al.: Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**(5), 395–409 (2011)
2. Adams, S.J., Stone, E., Baldwin, D.R., Vliegenthart, R., Lee, P., Fintelmann, F.J.: Lung cancer screening. *The Lancet* **401**(10374), 390–408 (2023). [https://doi.org/10.1016/S0140-6736\(22\)01694-4](https://doi.org/10.1016/S0140-6736(22)01694-4)
3. Baumgartner, M., Jäger, P.F., Isensee, F., Maier-Hein, K.H.: nnDetection: a Self-configuring Method for Medical Object Detection, pp. 530–539. Springer (2021). https://doi.org/10.1007/978-3-030-87240-3_51
4. Buess, L., et al.: Video-ct mae: self-supervised video-ct domain adaptation for vertebral fracture diagnosis. In: *Medical Imaging with Deep Learning* (2024)
5. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations (2020). <https://arxiv.org/abs/2006.10511>
6. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. *Med. Image Anal.* **87**, 102792 (2023)
7. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. *Adv. Neural. Inf. Process. Syst.* **35**, 35946–35958 (2022)

8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009 (2022)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning (2020). <https://arxiv.org/abs/1911.05722>
10. Hondru, V., Croitoru, F.A., Minaee, S., Ionescu, R.T., Sebe, N.: Masked image modeling: A survey (2025). <https://arxiv.org/abs/2408.06687>
11. Jaeger, P.F., et al.: Retina u-net: embarrassingly simple exploitation of segmentation supervision for medical object detection. In: *Machine Learning for Health Workshop*, pp. 171–183. PMLR (2020)
12. Ji, Z., Wu, Y., Zeng, X., An, Y., Zhao, L., Wang, Z., Ganchev, I.: Lung nodule detection in medical images based on improved yolov5s. *IEEE Access* (2023)
13. Kay, W., et al.: The kinetics human action video dataset (2017). <https://arxiv.org/abs/1705.06950>
14. Ke, A., Huang, S.C., O’Connell, C.P., Klimont, M., Yeung, S., Rajpurkar, P.: Video pretraining advances 3d deep learning on chest ct tasks (2023). <https://arxiv.org/abs/2304.00546>
15. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: *European Conference on Computer Vision*, pp. 280–296. Springer (2022)
16. Liao, F., Liang, M., Li, Z., Hu, X., Song, S.: Evaluate the malignancy of pulmonary nodules using the 3d deep leaky noisy-or network. *IEEE Trans. Neural Networks Learning Syst.* **30**(11), 3484–3495 (2019)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988 (2017)
18. Ma, Q., et al.: A large-scale dataset of gaussian splats and their self-supervised pretraining. In: *3DV 2025* (2024)
19. Mkindu, H., Wu, L., Zhao, Y.: Lung nodule detection in chest ct images based on vision transformer network with bayesian optimization. *Biomed. Signal Process. Control* **85**, 104866 (2023)
20. Niu, C., Wang, G.: Unsupervised contrastive learning based transformer for lung nodule detection. *Phys. Med. Biol.* **67**(20), 204001 (2022)
21. Oquab, M., et al.: Dinov2: learning robust visual features without supervision (2024). <https://arxiv.org/abs/2304.07193>
22. Setio, A.A.A., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Med. Image Anal.* **42**, 1–13 (2017)
23. Su, Y., Li, D., Chen, X.: Lung nodule detection based on faster r-cnn framework. *Comput. Methods Programs Biomed.* **200**, 105866 (2021)
24. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740 (2022)
25. Team, N.L.S.T.R.: Data from the national lung screening trial (nlst) (version 3) (201). <https://doi.org/10.7937/TCIA.HMQ8-J677>
26. Thai, A.A., Solomon, B.J., Sequist, L.V., Gainor, J.F., Heist, R.S.: Lung cancer. *The Lancet* **398**(10299), 535–554 (2021). [https://doi.org/10.1016/S0140-6736\(21\)00312-3](https://doi.org/10.1016/S0140-6736(21)00312-3)
27. Wu, X., Zhang, H., Sun, J., Wang, S., Zhang, Y.: Yolo-msrf for lung nodule detection. *Biomed. Signal Process. Control* **94**, 106318 (2024)

28. Xie, Z., et al.: Simmim: a simple framework for masked image modeling (2022). <https://arxiv.org/abs/2111.09886>
29. Zhou, J., Xu, Y., Liu, J., Feng, L., Yu, J., Chen, D.: Global burden of lung cancer in 2022 and projections to 2050: Incidence and mortality estimates from globocan. *Cancer Epidemiol.* **93**, 102693 (2024)

4 Lung nodule classification using radiomics model trained on degraded SDCT images



Lung nodule classification using radiomics model trained on degraded SDCT images

Jiaying Liu^{a,*}, Anna Corti^a, Valentina D.A. Corino^{a,b}, Luca Mainardi^a

^a Department of Electronics, Information and Bioengineering, Politecnico di Milano, Politecnico di Milano, Via Giuseppe Ponzio, 34, 20133 Milan, Italy

^b Carditech Lab, Centro Cardiologico Monzino IRCCS, Milan, Italy

ARTICLE INFO

Keywords:

Standard-dose CT degradation
Low-dose CT
Lung nodules
Machine learning
Radiomics

ABSTRACT

Background and objective: Low-dose computed tomography (LDCT) screening has shown promise in reducing lung cancer mortality; however, it suffers from high false positive rates and a scarcity of available annotated datasets. To overcome these challenges, we propose a novel approach using synthetic LDCT images generated from standard-dose CT (SDCT) scans from the LIDC-IDRI dataset. Our objective is to develop and validate an interpretable radiomics-based model for distinguishing likely benign from likely malignant pulmonary nodules.

Methods: From a total of 1010 CT images (695 SDCTs and 315 LDCTs), we degraded SDCTs in the sinogram domain and obtained 1950 nodules as the training set. The 675 nodules from the LDCTs were stratified into 50%-50% partitions for validation and testing. Radiomic features were extracted from nodules, and three feature sets were assessed using: a) only shape and size (SS) features, b) all features but SS features, and c) all features. A systematic pipeline was developed to optimize the feature set and evaluate multiple machine learning models. Models were trained using degraded SDCT, validated and tested on the LDCT nodules.

Results: Training a logistic regression model using three SS features yielded the most promising results, achieving on the test set mean balanced accuracy, sensitivity, specificity, and AUC-ROC scores of 0.81, 0.76, 0.85, and 0.87, respectively.

Conclusions: Our study demonstrates the feasibility and effectiveness of using synthetic LDCT images for developing a relatively accurate radiomics-based model in lung nodule classification. This approach addresses challenges associated with LDCT screening, offering potential implications for improving lung cancer detection and reducing false positives.

1. Introduction

Lung cancer is the deadliest malignancy in the world, being responsible for 18% of total cancer deaths in 2020 [1]. Being often diagnosed at advanced stages, it is associated with a low 5-year survival rate (being only 13% in Europe) [2]. Several trials have recently proved the efficacy of low-dose computed tomography (LDCT) screening in reducing lung cancer mortality [3–5]. However, LDCT screening is associated with high false positive rates: up to 96.4% of the positive screening results were false positive [3]. With the implementation of screening programs in the US and ongoing efforts in Europe to assess the benefits and cost savings of lung cancer screening through trials, the workload on radiologists increases significantly. To alleviate the burden on radiologists and overcome the high number of false positives, extensive research has been conducted to offer objective and efficient

diagnostic tools [6].

Radiomics analysis is based on the premise that image-captured characteristics of tumors, such as shape and texture, are potential tumor biomarkers that can be used to stratify patients, predict patient clinical outcomes or further characterize tumor phenotypes. In this context, radiomic features can enable the development of computer-aided diagnostic tools supporting clinical decisions [7,8]. Moreover, radiomics has the advantage of being more explainable than deep learning features, which are intrinsically more abstract and difficult to interpret [8,9].

Several studies applied radiomics with machine learning for nodule malignancy prediction in lung cancer screening, as detailed in Table 1. Other studies using deep learning approaches, such as Artificial Neural Networks and Convolutional Neural Networks [10], achieved good performances but they were out the scope of this paper as we intend to

* Corresponding author.

E-mail address: jiaying.liu@polimi.it (J. Liu).

<https://doi.org/10.1016/j.cmpb.2024.108474>

Received 16 May 2024; Received in revised form 11 October 2024; Accepted 22 October 2024

Available online 23 October 2024

0169-2607/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

develop a more explainable model. Although the extensive research in the field of lung cancer screening, some studies report low sensitivity rates (e.g., 56.0% in [11], 70.0% in [9], 10.2% in [12]), others considered small datasets (e.g., 72 nodules in [13], 113 nodules in [14]) or deliberately selected balanced datasets to address class imbalance [9, 15]. All studies but one used LDCT images in their research.

Publicly available datasets, as the National Lung Screening Trial (NLST) and the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI), have been extensively used for the development of nodule detection and classification but present distinct limitations [3,18]. NLST is a randomized controlled trial designed to determine if lung cancer screening using LDCT or chest radiography reduces mortality. Despite its size, containing around 25,000 participants, it lacks tumor segmentation [3]. Nevertheless, segmentation is a critical step in a radiomic approach, as features are extracted from the segmented volumes. On the other hand, LIDC-IDRI consists of diagnostic and lung cancer screening CT scans with marked-up annotated lesions from experienced radiologists. However, the number of LDCT images is relatively limited (315 out of the 1010 CTs, considering images acquired with tube current times seconds < 80 mAs).

In order to extract radiomic features from the nodules and build a radiomic-based classifier, the nodules require to be segmented. Thus, we intend to develop a classifier capable of distinguishing likely benign nodules from likely malignant ones during the screening stage. In this scenario, without the annotations from the screening NLST dataset, we adopted an innovative strategy leveraging the annotations from the non-screening LIDC-IDRI dataset by degrading standard-dose CT (SDCT) scans. The aim of this study is to develop an interpretable radiomics-based model using degraded SDCT images for the classification of likely benign pulmonary nodules from likely malignant ones and validate the classifier by testing on real LDCT images. Being of its first, this approach not only addresses the scarcity of annotated LDCT data but also offers a novel solution to explore the potential inclusion and utilization of SDCT images in the model development for detecting screening pulmonary nodules, especially in situations where LDCT screening is not widely available or incorporated into protocols.

2. Methods

2.1. SDCT degradation and image preprocessing

To develop a LDCT-based lung classifier utilizing both SDCT and LDCT images from the LIDC-IDRI dataset, we degraded the SDCT images. This degradation step was essential to ensure compatibility between the different image types and facilitate the development of a classifier capable of effectively analyzing both SDCT and LDCT modalities.

Additionally, all images (both SDCT and LDCT) were preprocessed to reduce the image-related sources of variability. The preprocessing steps were thresholding, to ensure all pixel values were positive with negative pixel values set to zero, and denoising, using a 3D Gaussian filter with a $3 \times 3 \times 3$ voxel kernel and $\sigma = 0.5$. Following the thresholding step, SDCT images were degraded to obtain simulated LDCT. All steps were performed using MATLAB version 2022b.

SDCT degradation (Fig. 1) was performed using a low-dose simulation technique in the sinogram domain, as proposed by Zeng et al. [19]. Briefly, standard-dose sinogram data were obtained by using Radon transformation (360 projection angles with 1-degree increments); subsequently, simulated Poisson and Gaussian noise was added to mimic the effect of electronic and quantum noise, generating low-dose sinogram data. Eq. (1) describes the calculation for the low-dose sinogram data ($p_{ld, sim}$):

$$p_{ld, sim} = \log \frac{I_{ld, sim}^0}{\text{Poisson}(I_{ld, sim}^0(s) * \exp(-p_{sd})) + \text{Gaussian}(m_e, \sigma_e^2)} \quad (1)$$

where $I_{ld, sim}^0$ is the simulated low-dose scan incident flux, p_{sd} is the standard-dose sinogram data obtained from the original SDCT images using Radon transformation, and m_e and σ_e^2 are the mean and variance of the electronic noise, respectively. Specifically, $I_{ld, sim}^0$ is computed by multiplying the incident flux with the ratio between the high and simulated low-dose incident flux, k . This ratio is computed based on the proposed equation Eq. (2):

Table 1
Summary of recent studies applying radiomics for nodule malignancy prediction in lung cancer screening.

Study	Dataset (nodule size)	LDCT?	Model description	Sen.	Spc.	Acc.	AUC	Limitations
Choi et al. 2018 [13]	72 patients LIDC-IDRI (≥ 3 mm)	yes	2 radiomic features, LASSO features selection and SVM classification	87.2%	81.2%	84.6%	0.89	small dataset
Peikert et al. 2018 [15]	726 nodules from NLST (≥ 7 mm)	yes	8 radiomic features selected and trained with LASSO model	90.4%	85.5%	-	0.94	selected data to balance the dataset
Alahmari et al. 2018 [11]	467 cases from NLST (≥ 4 mm)	yes	18 Rider features, ReliefF feature selection, Random Forest	49.0%	93.0%	-	0.82	low sensitivity
Causes et al. 2018 [9]	664 nodules from LIDC-IDRI (≥ 3 mm)	no	a logistic regression model trained only on a size metric (square root of largest cross-sectional area)	69.7%	95.5%	82.6%	0.94	low sensitivity
Mao et al. 2019 [16]	294 nodules (6–15 mm)	yes	11 radiomic features, LASSO feature selection, linear regression model	81.0%	92.2%	89.8%	0.90	small sample size
Garau et al. 2020 [14]	113 nodules COSMOS (all sizes)	yes	5 radiomic features, SVM-LASSO test set: 72 nodules from LIDC-IDRI.	90.0%	FPR = 35.5%	79.1%	0.86	small sample size
Liu et al. 2021 [17]	141 nodules (not specified)	yes	3 radiomic features, radiomic monogram	86.4%	85.7%	86.0%	0.98	small sample size, potential selection bias
Rundo et al. 2021 [12]	703 nodules from bioMILD trial (all sizes)	yes	32 radiomic features, Borderline-SMOTE over-sampling method, elastic net regularization for logistic regression, tested with a blind test	67.0%	86.2%	76.6%	0.82	low sensitivity

LIDC-IDRI - the Lung Image Database Consortium and Image Database Resource Initiative; NLST - National Lung Screening Trial; Sen. – sensitivity; Spc. – specificity; Acc. – accuracy; AUC – area under the ROC curve; FPR – false positive rate.

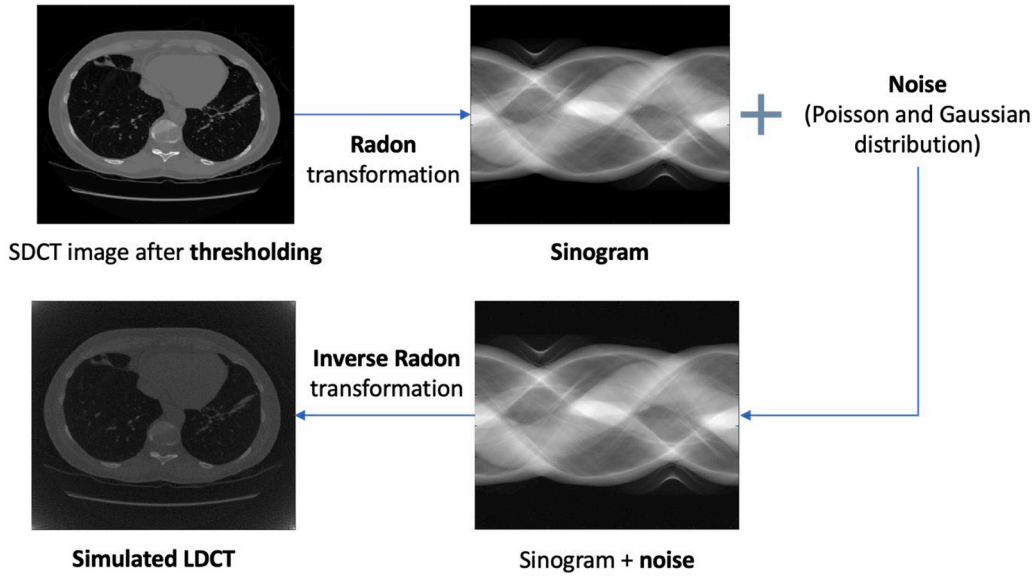


Fig. 1. Pipeline to simulate low-dose computed tomography (LDCT). The Radon transformation is applied to the standard-dose computed tomography (SDCT) image, then electronic and quantum noise effects are included by noise injection. The final step involves inverse Radon transformation to reconstruct the LDCT image.

$$k \text{ (mAs)} = a * (\text{mAs}) + b \quad (2)$$

where $a = 0.0095$ and $b = 0.0309$. In the present study, three different low-dose scenarios were simulated using different k-values corresponding to 40, 50, and 75 mAs to model variability in dose levels. We assigned $k1(40\text{mAs})$ to half of the data points, $k2(50\text{mAs})$ to a quarter, and $k3(75\text{mAs})$ to the remaining quarter. This distribution was based on the proportions found in a subset of LDCT images from the LIDC-IDRI dataset. Electronic noise was modeled using Gaussian noise with a standard deviation $\sigma = \sqrt{11}$, consistent with the scanner noise characteristics outlined in Zeng et al [19]. and Ma et al [20]. Finally, the simulated LDCT images were reconstructed from the low-dose sinogram data using the inverse Radon transform. For parameters settings see more details in [19].

2.2. Dataset description

The dataset used in this study derives from LIDC-IDRI, available in [21]. It contains 1018 thoracic CT scans acquired from 1010 participants, with nodule annotations performed by one to four experienced radiologists. After excluding the duplicated participants, a total of 2625 nodules, each assigned as ≥ 3 mm (lesions with greatest in-plane dimension in the range 3–30 mm) marked by at least one radiologist, was included in this study.

The dataset includes some participants with SDCT and others with LDCT images. Tube current ranged from 40 to 650 mAs, while tube voltage ranged from 120 to 140 kV. Even if there is no common consensus, previous studies considered as LDCT those images acquired with tube current-time product ranging from 20 to 100 mAs and voltages spanning from 80 to 140 kV [22,23]. For our analysis, LDCT images were defined as those acquired with a tube current ≤ 80 mAs, similar to the threshold used in the NLST [22].

Each nodule was evaluated by the radiologists for the likelihood of malignancy on a scale of 1 to 5: 1 – highly unlikely, 2 – moderately unlikely, 3 – indeterminate, 4 – moderately suspicious, 5 – highly suspicious.

Given that multiple radiologists might provide scores for a single nodule, a consensus was reached by averaging the scores to derive a single score for each nodule. In our study, nodules with an average score ≥ 4 were considered malignant. The summary of nodule characteristics in each subset is presented in Table 2.

Table 2

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

	SDCT	LDCT
Nr. participants	695	315
Nr. nodules	1950	675
Averaged likelihood of malignancy		
[1;2]	280	86
[2;3]	766	222
[3;4]	679	298
[4;5]	225	69
Non-malignant (average likelihood <4)	1725	606
Malignant (average likelihood ≥ 4)	225	69

SDCT: standard-dose CT; LDCT: low-dose CT.

2.3. Pipeline description

Fig. 2 details the workflow of the study, consisting of radiomic feature extraction, feature selection, model training, validation and tuning, and model evaluation. In this study, nodules in SDCT subset were used as the training set, while nodules in the LDCT subset were stratified into 50–50% partitions, designated as the validation and test set, respectively.

2.3.1. Radiomic feature extraction

Radiomic features were extracted from the region of interest (ROI), the segmented nodules, using Pyradiomics [24]. When the nodule had more than one annotation, the consensus mask was used to extract the radiomic features. A total of 851 features were extracted from each nodule, including 14 shape features, 18 first-order features, 75 texture features (24 gray-level co-occurrence matrix (GLCM) features, 14 gray-level dependence matrix (GLDM) features, 16 gray-level run length matrix (GLRLM) features, 16 gray-level size zone matrix (GLSZM), and 5 neighboring gray tone difference matrix (NGTDM) features), and 744 wavelet-based features. For more details on radiomic features, see Pyradiomics documentation [24].

2.3.2. Feature selection

Feature selection step was performed on three different feature subsets independently: a) only shape and size (SS) features, b) all features but SS features, and c) all features. These three sets were chosen to assess the relevance of nodule morphological characteristics on the

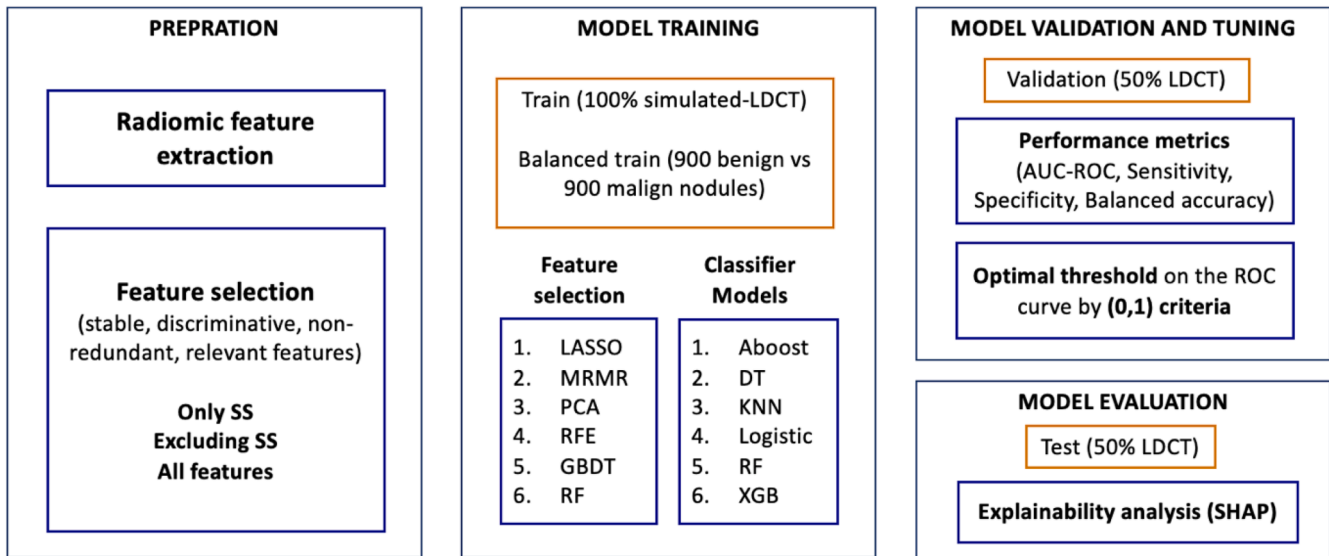


Fig. 2. Pipeline consisting of radiomic feature extraction, feature selection (using a) only shape and size (SS) features, b) all features but SS features, and c) all features), model training and validation, and explainability analysis.

prediction.

All feature selection steps were performed on the training set and then applied to the test set. Additionally, the selection of stable, discriminant and non-redundant features were computed before balancing the training set (see Section 2.3.3), while the selection of more relevant features was performed after obtaining a balanced dataset.

2.3.2.1. Selection of stable and discriminant features. Firstly, a stability analysis was performed to identify the features that exhibit consistent behavior under small transformations. This analysis was conducted by considering nodules that had at least four annotations (705 nodules): features were computed for each ROI and the inter-class correlation coefficient (ICC) was obtained. Features were deemed stable if ICC was > 0.75 , as defined by Koo et al. [25]. Second, a discriminative analysis was conducted to discard features that have similar values under large transformations and, therefore, can be considered non-informative. The analysis consisted in translating the original ROIs 150% of the size of the box surrounding the ROI (Fig. 3) and computing the ICC: features were considered discriminative (i.e., informative) when ICC was < 0.5 .

2.3.2.2. Selection of non-redundant features. To address redundancy, a pairwise correlation analysis was performed: when a pair of features had

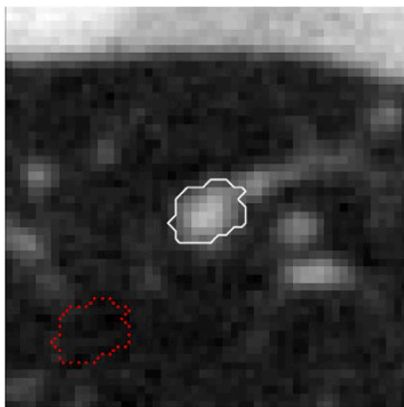


Fig. 3. Example of a large translation for discriminative analysis. The white contour represents the consensus masks, while the red dots indicate the translation masks.

an absolute Spearman's correlation coefficient above 0.85, only one of the two was kept. In particular, the one with a lower mean correlation (computed through Spearman's correlation coefficient) with all the other $n-2$ features was selected. Afterwards, the non-redundant features underwent normalization using z-score standardization.

2.3.2.3. Selection of more relevant features. Six different feature selection techniques were assessed after balancing the training set (see Section 2.3.3): principal component analysis (PCA), recursive feature elimination (RFE), feature selection using random forest (RF), least absolute shrinkage and selection operator (LASSO), feature selection using gradient-boosted decision tree (GBDT) and minimum redundancy maximum relevance (MRMR). Specifically, we explored varying numbers of selected features (5, 10, 15, and 20) using RFE, RF, GBDT, and MRMR. For PCA, dimensionality reduction was examined with 5, 10, 15, and 20 components, retaining 56%, 67%, 73%, and 78% of variance, respectively, for the feature subset including all the features and 56%, 68%, 74% and 79% of variance, respectively, for the feature subset without SS features. For LASSO, features were selected based on non-zero coefficients with a penalty value of 0.1.

2.3.3. Balancing the training set

To address the imbalanced training dataset, we first applied a data augmentation technique introduced in [26] consisting in the application of ROI perturbations on the 225 malignant nodules of the training set: 15% dilation and erosion, as well as random contour adjustments. Fig. 4 illustrates the augmentation process, including erosion (a), dilation (b), and contour randomization (c). As a result, the size of the minority class in the training set quadruplicated (from 225 to 900). Afterwards, a random undersampling approach was adopted to reduce the non-malignant nodules (majority class) from 1725 to 900, to match the number of nodules in the minority class post-augmentation. Random undersampling was opted in our methodology since it has shown to perform well in the literature according to [27].

2.3.4. Model development

For each subset of selected features, six machine learning models were trained and validated: random forest (RF), AdaBoost (ABoost), XGBoost (XGB), decision tree (DT), K-nearest neighbors (KNN), and logistic regression (LR). These algorithms were chosen based on the studies mentioned in Table 1 and their extensive utilization in the field of medical image analysis and radiomics in oncology, aligning with

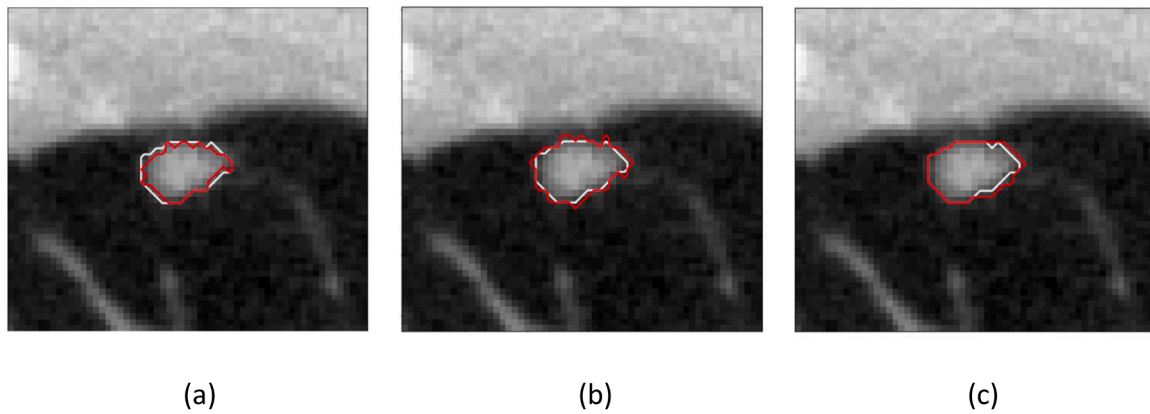


Fig. 4. Perturbation technique applied to the minority class for data augmentation: (a) erosion, (b) dilation, (c) random contour. The white contour represents the original consensus mask, while the red represents the perturbation.

Parmar et al.'s review [28]. For hyperparameter tuning, grid search was used. Details of the parameter grid for each model can be found in the Supplementary Material. During the validation phase, the optimal threshold on the ROC curve was determined by the minimum Euclidean distance to the point (0,1). Bootstrapping was performed 1000 times to obtain the mean value and the 95% confidence interval (CI) for each metric. Model development was implemented using the Scikit-Learn Python library [29].

Finally, the developed model was evaluated on the test set and SHAP analysis (SHapley Additive exPlanations) [30] was used to explain the predictions.

3. Results

The analysis focused on the three different feature sets. To determine the optimal number of features for each feature selection method combined with the different classifiers, we selected the configuration that yielded the highest area under the ROC curve (AUC-ROC) value on the validation set (see Supplementary Material). Additionally, sensitivity, specificity, balanced accuracy and geometric mean were computed to evaluate models' performances on the validation and test set after applying the optimal threshold on the ROC curve determined during the validation phase.

3.1. Model using only SS features

When considering only the SS features, three features were identified as stable, discriminant and non-redundant: *minor_axis_length*, *sphericity* and *elongation*. Consequently, further feature selection was not required. In this scenario (Fig. 5), RF and LR achieved similar scores on the validation set, with RF obtaining balanced accuracy, sensitivity, specificity and AUC-ROC scores of 0.80, 0.76, 0.83 and 0.88, respectively, and LR yielding 0.80, 0.76, 0.84 and 0.87 for the same metrics, respectively.

Considering the simplicity of LR compared to RF, LR was evaluated on the test set and achieved balanced accuracy, sensitivity, specificity and AUC-ROC scores of 0.82, 0.74, 0.89 and 0.82, respectively. SHAP analysis (Fig. 6) on LR revealed that *minor_axis_length*, corresponding to the second-largest axis length of the ROI-enclosing ellipsoid, was the most important feature, with lower values (indicating smaller nodules) associated to reduced predicted malignancy risk. The second most significant feature was *sphericity* which measures the roundness of the shape of the tumor region relative to a sphere. Higher *sphericity* values (indicating rounder nodules) were associated with reduced predicted malignancy risk. Additionally, *elongation* measures the relationship between the two largest principal components in the ROI shape. Higher *elongation* values suggested a more circle-like nodule shape, correlating with a lower likelihood of malignancy.

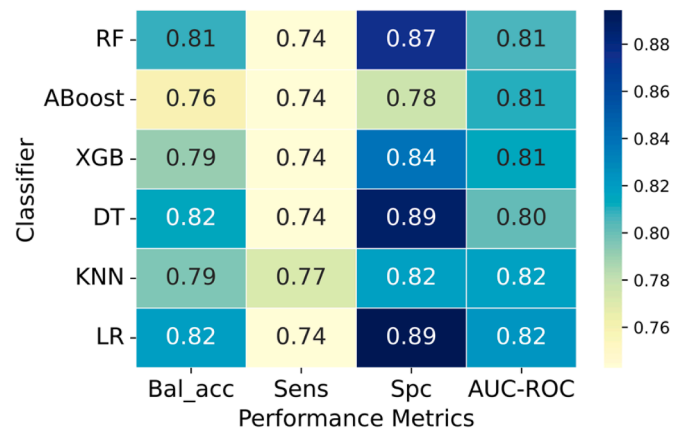


Fig. 5. - Heatmap showing performance metrics (balanced accuracy (bal_acc), sensitivity (sens), specificity (spc), AUC-ROC) on the validation set for different classifiers (random forest (RF), AdaBoost (ABoost), XGBoost (XGB), decision tree (DT), k-nearest neighbors (KNN), logistic regression (LR)) using only SS features.

3.2. Model without SS features

When using all features but SS features, 111 features resulted stable, discriminative, and non-redundant. Subsequently, the selection of more relevant features methods was performed, leading to a varying number of features utilized for model training. Fig. 7 depicts heatmaps of sensitivity and balanced accuracy on the validation set for each feature selection and classifier combination considering the optimal number of selected features. The best performance was obtained by LR trained with the 20 features selected by RF, being associated with a balanced accuracy of 0.79, a sensitivity of 0.82, specificity of 0.76, and AUC-ROC of 0.83. Evaluating the same model on the test set reported balanced accuracy, sensitivity, specificity and AUC-ROC scores of 0.60, 0.43, 0.77 and 0.53, respectively. Given these results on the test set, which were insufficient for meaningful interpretation, SHAP analysis was not performed.

3.3. Model using all features

Considering the set including all the 851 radiomic features, 113 were considered stable, discriminative, and non-redundant, thus undergoing further selection of more relevant features. Fig. 8 depicts heatmaps of sensitivity and balanced accuracy on the validation set. The best performance on the validation set was obtained with the combination of MRMR feature selection (5 features) and ABoost classifier, leading to a

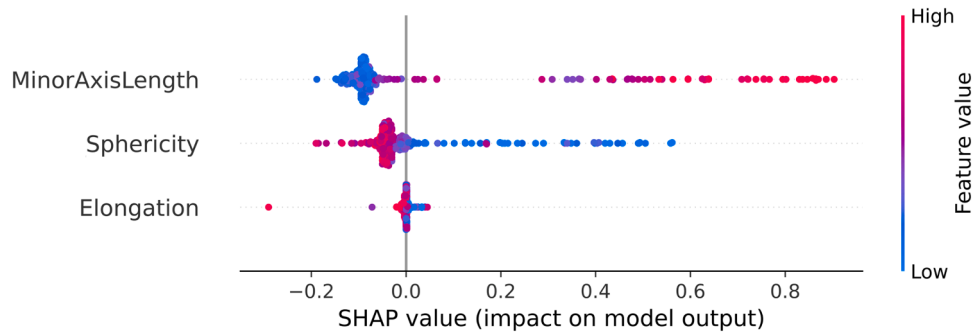


Fig. 6. Beeswarm plot of the test set generated from SHAP analysis of the developed model using logistic regression.

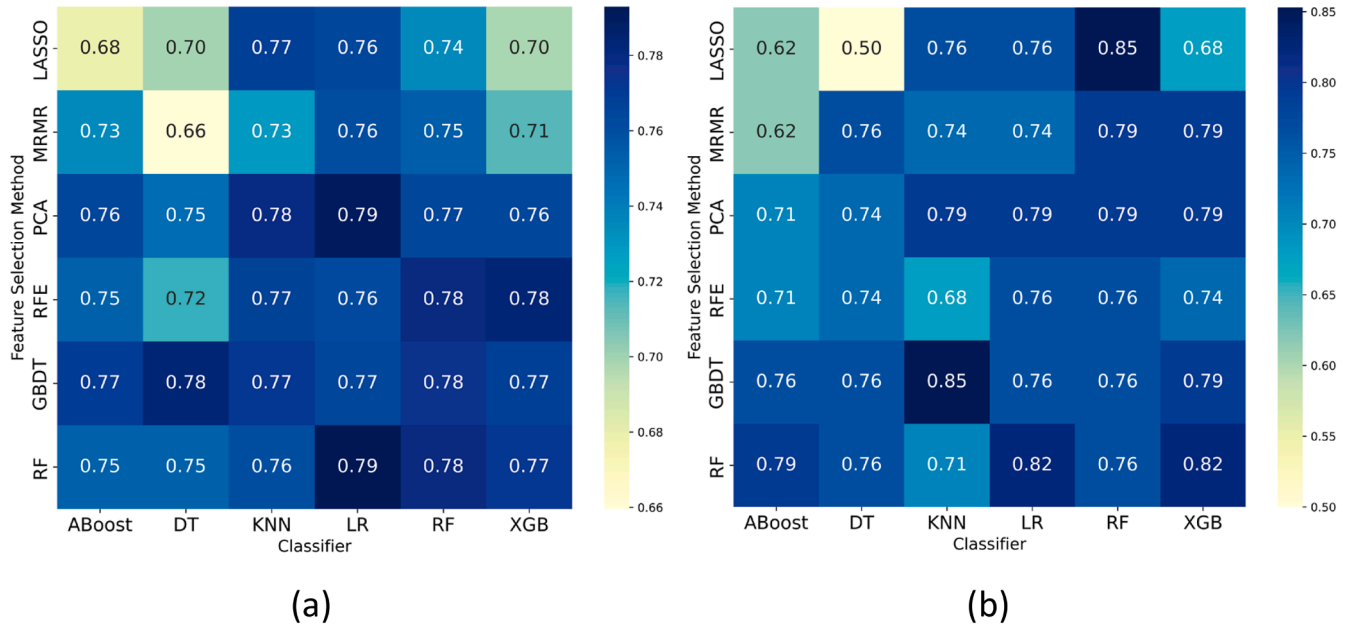


Fig. 7. Heatmaps of the (a) balanced accuracy and (b) sensitivity obtained during the validation phase considering different combinations of feature selection and classifiers for models using all but SS features.

balanced accuracy of 0.85, sensitivity of 0.85, specificity of 0.85 and AUC-ROC of 0.87. Although the combination of features selected by MRMR with the LR classifier and RF with the XGB model demonstrated competitive balanced accuracy (0.84 for MRMR with LR and 0.83 for RF with XGB) and sensitivity (0.88 for MRMR with LR and 0.91 for RF with XGB) values, the achieved specificity was relatively lower (0.79 for MRMR with LR and 0.75 for RF with XGB).

Considering the combination MRMR with ABOost, the results obtained on the test set were 0.78 for balanced accuracy, 0.74 for sensitivity, 0.83 for specificity and 0.82 for AUC-ROC. SHAP analysis on this model identified *sphericity*, *wavelet_LLH_glcM_MCC* (maximal correlation coefficient (MCC)), and *major_axis_length* as the three most impactful features (Fig. 9). *Sphericity* displayed consistent behavior with the model using only SS features, i.e., higher values of *sphericity* (indicating rounder nodules) were found to be associated to lower likelihood of malignancy. *Wavelet_LLH_glcM_MCC* feature, derived from wavelet transformations involving low-pass (L) and high-pass (H) filters, reflects the complexity of the texture within a ROI region. Lower *wavelet_LLH_glcM_MCC* values, indicating more intricate or complex textures, are associated with a higher probability of malignancy. *Major_axis_length*, which represents the largest axis length of the ROI-enclosing ellipsoid, behaves analogously to the *minor_axis_length* feature explained in models using only SS features, where higher values indicate a greater likelihood of malignancy.

3.4. Comparative analysis

Table 3 summarizes the best results obtained using different feature sets during validation and testing, presented as mean values and their 95% CI. When using three SS features, the model achieved consistent results on both validation and test sets, suggesting robustness with a minimal feature set. When considering all but SS features, performance declined notably on the test set compared to validation. When using all features, the classifier achieved the highest performance during validation with a slightly decrease on the test set. These results suggest that using a small subset of SS features yielded robust performance across validation and testing phases, while excluding these features led to a significant decrease (p -value < 0.05 for most of the metrics) in model's performance on the test set. The superior performance of the only SS feature set can be attributed to the radiologists' reliance on these features for nodule scoring in this dataset. For more details on p -values, see Supplementary Material.

Additionally, as the original labels ranged from 1 to 5, with nodules having an average score ≥ 4 considered malignant, we sought to determine how classified and misclassified instances in the test set were spread across malignancy levels. This analysis was performed for the model using only SS features. The histogram in Fig. 10 reveals that a significant portion of misclassified nodules belonged to the category with an average score between 3 and 4, representing the most uncertain

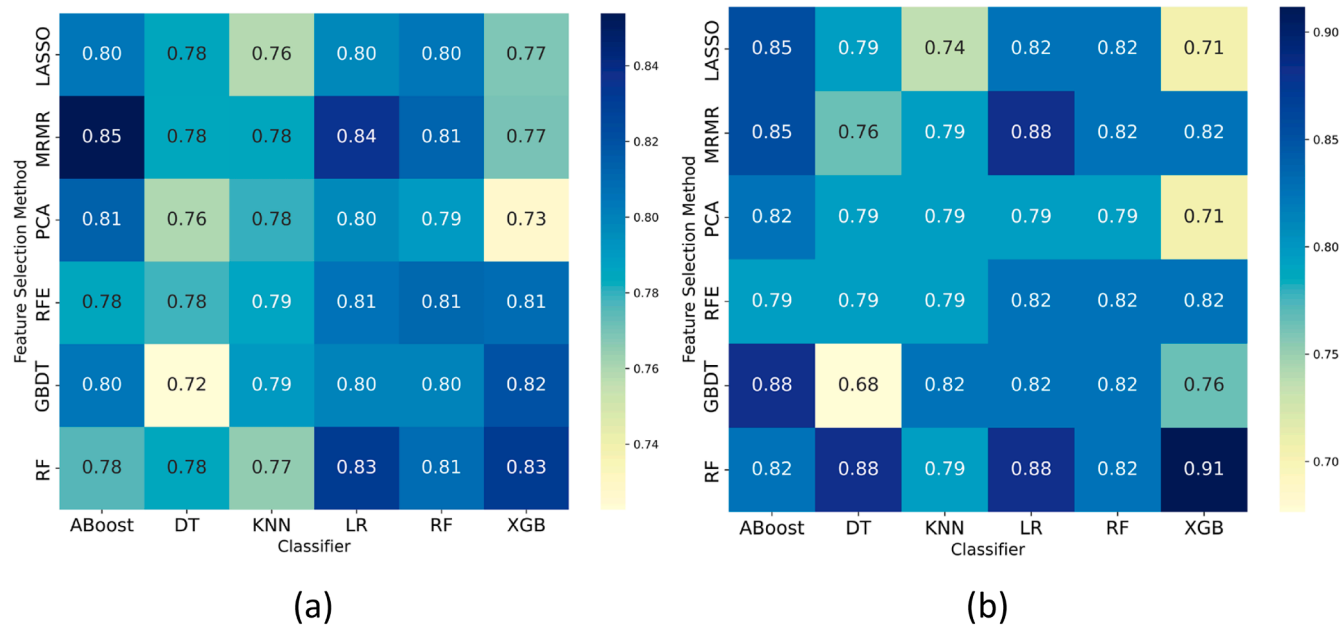


Fig. 8. Heatmaps of the (a) balanced accuracy and (b) sensitivity obtained during validation phase considering different combinations of feature selection and classifiers for models using all features.

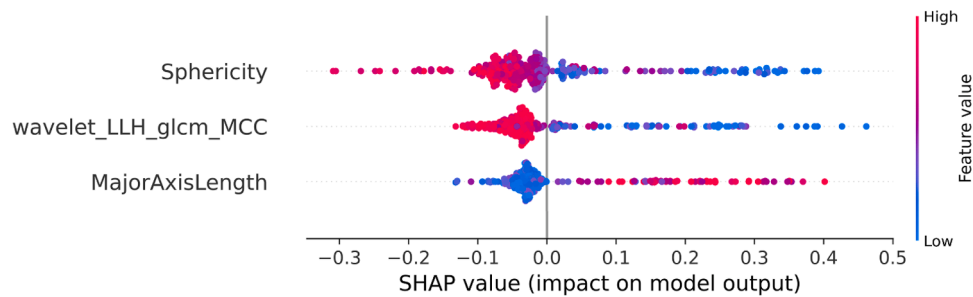


Fig. 9. Beeswarm plot of the predictions made for the instances present in the test set generated from SHAP analysis. It shows the first 3 features selected by MRMR feature selection when used to train an ABoost model.

Table 3
Summary of the best results obtained during validation and the correspondent testing phase for different feature sets using: a) only shape and size (SS) features, b) all features but SS features, and c) including all features. The results, presented as mean (CI), were calculated using 1000 bootstrapping iterations.

Feature set	Bal_acc.	Sen.	Spc.	AUC-ROC	G-mean	#Feat
Only SS features						
validation	0.807 (0.807–0.807)	0.796 (0.793–0.798)	0.819 (0.817- 0.821)	0.874 (0.874–0.875)	0.806 (0.806–0.807)	3
test	0.805 (0.804–0.80)	0.759 (0.757–0.760)	0.850 (0.847 - 0.853)	0.874 (0.874–0.874)	0.802 (0.802–0.803)	
All but SS features						
validation	0.784 (0.783–0.785)	0.796 (0.794–0.799)	0.772 (0.770–0.774)	0.822 (0.821–0.823)	0.783 (0.782–0.784)	20
test	0.664 (0.659–0.669)	0.544 (0.535–0.553)	0.783 (0.780–0.786)	0.822 (0.821–0.823)	0.647 (0.642–0.653)	
All features						
validation	0.826 (0.825 - 0.827)	0.843 (0.841 - 0.845)	0.810 (0.808 - 0.811)	0.865 (0.864 - 0.866)	0.826 (0.825 - 0.826)	5
test	0.769 (0.768 - 0.770)	0.740 (0.738 - 0.742)	0.798 (0.796 - 0.800)	0.865 (0.864 - 0.866)	0.768 (0.767 - 0.769)	

Sen. – sensitivity; Spc. – specificity; AUC-ROC – area under the ROC curve; Bal_acc. – balanced accuracy; G-mean – geometric mean; # Feat. – number of features.

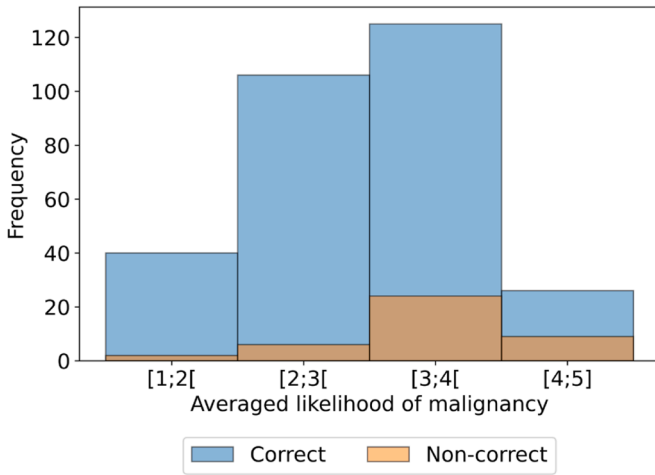


Fig. 10. Histogram representing the original label distribution (average likelihood of malignancy) of the correctly and incorrectly classified instances in the test set.

cases.

4. Discussion and conclusion

In the present study, we developed an explainable radiomics-based model to classify lung nodules using CT images. In particular, we demonstrated the effectiveness of training a classifier from synthetic LDCT, obtained by degrading SDCT images, and testing it on real LDCT images. This strategy enabled us to address the limited availability of annotated LDCT images for training while ensuring the validity of our results through testing on real LDCT images.

Our findings revealed that extracting only three SS features from the lung nodules and using a simple LR classifier, yielded the most consistent results, achieving mean (CI) for balanced accuracy, sensitivity, specificity, AUC-ROC, and geometric-mean scores of 0.805 (0.804–0.805), 0.759 (0.757–0.760), 0.850 (0.847–0.853), 0.874 (0.874–0.874), 0.802 (0.802–0.803), respectively, on the test set. These results were comparable to or exceeded the performance of several studies in the literature, such as Alahmari et al. [11], who used 18 Rider

features trained on Random Forest model, obtained sensitivity of 0.49 and specificity of 0.93; and Rundo et al. [12], who used 32 radiomic features trained on logistic regression, obtained a sensitivity of 0.67 and specificity of 0.86. Causey et al. [9] achieved notable results (sensitivity of 0.88 and specificity of 0.88) using the LIDC-IDRI dataset with features obtained from convolutional neural networks. However, when applying only radiomic features the sensitivity dropped to 0.70. Moreover, Causey et al. [9] neglected nodules with an average score for the likelihood of malignancy between 3 and 4 (included in this study), which are the most challenging nodules to classify.

While some models in the literature achieved better performance, they often used smaller sample sizes, for example, Choi et al. [13] considered 72 nodules and developed a support vector machine (SVM) classifier using two radiomic features, Mao et al. [16] extracted 11 radiomic features from each of the 294 nodules and used linear regression model, Garau et al. [14] considered 113 nodules and extracted five radiomic features they trained on SVM classifier, and Liu et al. [17] used 141 nodules to develop a radiomic monogram using three features. We recognize that larger datasets contribute to more stable and non-biased ML models, enhancing reliability and generalizability. Notably, our dataset comprised a larger number of nodules (2625), surpassing by sample sizes all the studies listed in Table 1 (see Fig. 11). Peikert et al. [15] achieved better performance metrics (sensitivity = 0.90, specificity = 0.86) using 8 radiomic features trained with LASSO model with a relatively large dataset, but it is worth noting that their study included larger nodules (≥ 7 mm) than those analyzed in our study (≥ 3 mm). To date, radiomic analyses focusing on small lung nodules have not been extensively explored in the literature. We specifically included nodules with a size of ≥ 3 mm, considered optimal for screening and important for achieving earlier diagnosis [31]. In addition, smaller nodules present greater difficulty in classification and are associated with higher rates of false positives [31], emphasizing the need for robust classification models in this context.

Including SHAP analysis in the pipeline allowed us to identify the most impactful features contributing to the classification of the lung nodules. Specifically, *axis_length* and *sphericity* were selected as the top 3 most impactful features both when only SS feature were included in model training and also when all the features were considered. This result is coherent and highlights the importance of SS feature in distinguishing likely benign lung nodules from likely malign ones. Moreover, identifying *minor axis length* as the most important feature in the

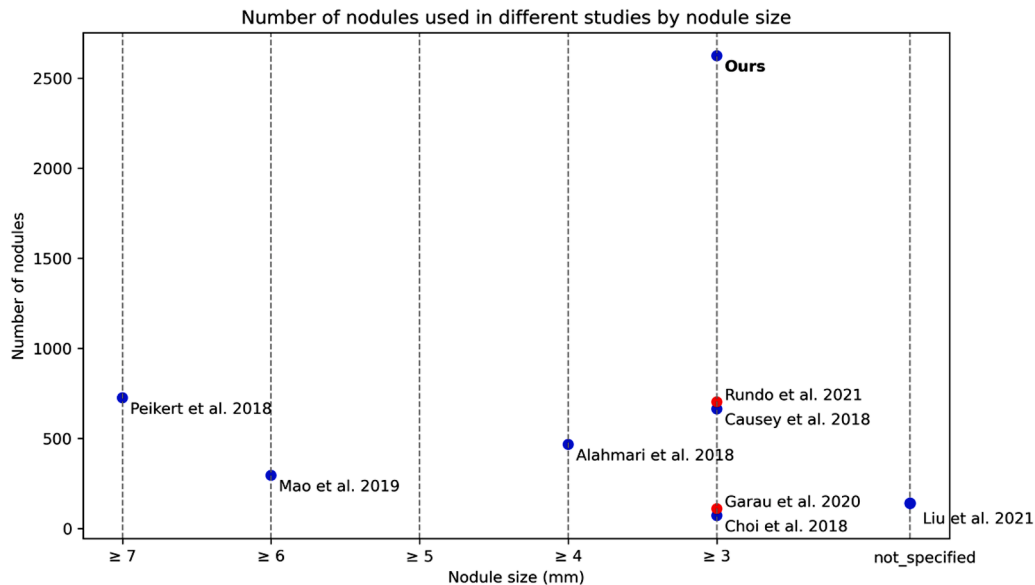


Fig. 11. Distribution of nodule sizes ($\geq x$ mm) and the corresponding number of nodules used across different studies. Each point represents a study, with different color markers used to differentiate studies with similar numbers of nodules analyzed.

best model aligns with one of the criteria used in the Lung Imaging Reporting and Data System (Lung-RADS) developed by the American College of Radiology to standardize the screening of lung cancer on CT images [32].

One limitation of our study was the nature of the dataset, which was not tailored for nodule classification or screening but for developing segmentation methods. Consequently, the nodule labels were based on radiologists' assessments of malignancy likelihood. In this perspective, our study successfully emulated the radiologists' approach to nodule evaluation, proposing a tool to support radiologists' screening activity. In clinical practice, further improvements are needed before it can be considered ready for clinical use. The current study represents a preliminary study aimed at assessing the feasibility of training models using simulated LDCT images for lung nodule classification, overcoming challenges associated with limited annotated LDCT datasets. Additionally, the degradation approach was not further validated, due to the absence of publicly available paired LDCT-SDCT datasets, thus relying on prior validation from Zeng et al. [19]. Although the model's performance on real LDCT data supports the goodness of the degradation method, future studies will include comparison and validation on multiple LDCT simulation techniques.

Finally, although the results of our analysis can be considered robust, being LIDC-IDRI based on a large and multicentric dataset, further validation on external datasets could strengthen our findings.

While an extensive comparison has been made between the proposed model and existing radiomic models for lung nodule classification using screening LDCT images, deep learning approaches were considered out of the scope of this paper. As future work, training deep learning models on degraded SDCT would allow for a more direct comparison with other advanced state-of-the-art methods.

In conclusion, utilizing degraded SDCT images offers a promising avenue for overcoming challenges associated with limited annotated LDCT images. By using both SDCT and synthetic LDCT images for model training, this approach expands the potential for the development of better and more generalizable models, ultimately enhancing patient care by lowering the false positive rate and improving diagnostic accuracy. Additionally, our study underscores the potential of shape and size features to enhance lung nodule classification accuracy.

Data availability

The dataset can be found in: <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254>.

The best trained model can be found in GitHub: https://github.com/jiayingliu423/LDCT_radiomics.

Funding

AC is funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvanCed Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

CRedit authorship contribution statement

Jiaying Liu: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Anna Corti:** Writing – review & editing, Visualization, Investigation, Formal analysis. **Valentina D.A. Corino:** Writing – review & editing, Visualization, Supervision, Investigation, Formal analysis, Conceptualization. **Luca Mainardi:** Writing – review & editing, Visualization, Supervision,

Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.cmpb.2024.108474](https://doi.org/10.1016/j.cmpb.2024.108474).

References

- [1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, *CA Cancer J. Clin.* 71 (2021) 209–249, <https://doi.org/10.3322/caac.21660>.
- [2] R. De Angelis, M. Sant, M.P. Coleman, S. Francisci, P. Baili, D. Pierannunzio, A. Trama, O. Visser, H. Brenner, E. Ardanaz, M. Bielska-Lasota, G. Engholm, A. Nennecke, S. Siesling, F. Berrino, R. Capocaccia, H. Amash, C. Amati, F. Di Salvo, S. Bonfarnuzzo, L. Botta, R. Foschi, G. Gatta, C. Margutti, P. Minicozzi, M. Caldora, E. Carrani, S. Mallone, P. Roazzi, S. Rossi, M. Santaquilani, A. Tavilla, M. Hackl, N. Zielonke, W. Oberaigner, K. Henau, E. Van Eycken, N. Dimitrova, Z. Valerianova, M. Sekerija, L. Dušek, M. Zvolský, H. Storm, M. Magi, T. Areleid, N. Malila, K. Seppä, M. Velten, A. Belot, V. Bouvier, X. Troussard, G. Launoy, A. V. Guizard, A.M. Bouvier, J. Faivre, P. Arveux, M. Maynadie, A.S. Woronoff, M. Robaszewicz, I. Baldi, A. Monnereau, B. Tretarre, N. Bossard, M. Colonna, F. Molinié, S. Bara, C. Schvartz, J. Clavel, E. Desandes, B. Lacour, B. Lapôte-Ledoux, P. Grosclaude, M. Grosclaude, R. Stabenow, A. Eberle, S. Luttmann, J. Engel, G. Schubert-Fritschle, P. Kaatsch, B. Holleczeck, A. Katalinic, M. Garami, Z. Jakab, J. Jónasson, L. Tryggvadóttir, H. Comber, A. Bulatko, G. Mazzoleni, C. Buzzoni, A. Giacomini, P. Mancuso, A. Mazzei, A. Suter Sardo, A. Caldarella, E. Crocetti, G. Manneschi, L. Dal Maso, D. Serraino, A. Zucchetto, S. Busco, M. Natali, F. Pannozzo, V. Gennaro, P. Ricci, C. Pascucci, M. Auteliano, G. Randi, M. Federico, M. Fusco, M.F. Vitale, M. Usala, R. Cusimano, B. Ravazzolo, F. Vitale, R. Staiti, A. Traina, M. Zarcone, M. Michiara, M. Maule, F. Merletti, R. Tumino, E. Di Felice, P. Giorgi Rossi, M. Vicentini, F. Falcini, L. Bonelli, M. Vercelli, A. Iannelli, R. Cesaraccio, O. Sechi, S. Piffer, M.L. Contrino, A. Madeddu, F. Tisanu, A.C. Fanetti, S. Maspero, S. Rosso, R. Zanetti, P. Candela, T. Scuderi, F. Bianconi, F. La Rosa, F. Stracci, P. Contiero, G. Tagliabue, P. Dei Tos, S. Guzzinati, S. Pildava, G. Smalley, D. Agius, N. Calleja, R. Micallef, T.B. Johannesen, S. Gózd, R. Mężyk, J. Rachtan, M. Bębenek, J. Błaszczyk, K. Kępska, L. Antunes, M.J. Bento, C. Castro, A. Mayer-Da-Silva, A. Miranda, D. Coza, C. Safaei Diba, M. Primić-Zakeli, E. Almar, A. Mateos, C. Ramirez, J. Bidaurrezaga, N. Larrañaga, A. Torrella-Ramos, M. Vicente Raneda, J.M. Díaz García, R. Jimenez-Chillaro, A.I. Marcos-Navarro, A. Izquierdo Font, R. Marcos-Gragera, C. Martinez, M.J. Sanchez, C. Navarro, M. D. Chirlaque, C. Moreno-Iribas, S. Felipe García, E. Pardo Romaguera, R. Peris-Bonet, M. Carulla, J. Galceran, S. Khan, M. Lambe, M. Mousavi, C. Bouchardy, M. Usel, H. Frick, C. Herrmann, M. Lorez, S.M. Ess, A. Bordoni, A. Spitale, I. Konzelmann, K. Aben, P. Huijgens, V. Lemmens, R. Verhoeven, R. Otter, C. Allemani, B. Rachet, N. Eassey, G. Lawrence, T. Moran, J. Rashbass, M. Roche, J. Verne, J. Wilkinson, C. Stiller, L. Anderson, F. Bannion, A. Gavin, D. Brewster, D. W. Huws, C. White, Cancer survival in Europe 1999–2007 by country and age: results of EURO CARE-5—A population-based study, *Lancet Oncol.* 15 (2014) 23–34, [https://doi.org/10.1016/S1470-2045\(13\)70546-1](https://doi.org/10.1016/S1470-2045(13)70546-1).
- [3] National Lung Screening Trial Research Team, Reduced lung-cancer mortality with low-dose computed tomographic screening, *New Engl. J. Med.* 365 (2011) 395–409, <https://doi.org/10.1056/NEJMoa1102873>.
- [4] U. Pastorino, M. Silva, S. Sestini, F. Sabia, M. Boeri, A. Cantarutti, N. Sverzellati, G. Sozzi, G. Corrao, A. Marchianò, Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: new confirmation of lung cancer screening efficacy, *Ann. Oncol.* 30 (2019) 1162–1169, <https://doi.org/10.1093/annonc/mdz169>.
- [5] H.J. de Koning, C.M. van der Aalst, P.A. de Jong, E.T. Scholten, K. Nackaerts, M. A. Heuvelmans, J.-W.J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg, S. van 't Westeinde, M. Prokop, W.P. Mali, F.A.A. Mohamed Hoessein, P.M.A. van Ooijen, J.G.J.V. Aerts, M.A. den Bakker, E. Thunnissen, J. Verschakelen, R. Vliegthart, J. E. Walter, K. ten Haaf, H.J.M. Groen, M. Oudkerk, Reduced lung-cancer mortality with volume CT screening in a randomized trial, *New Engl. J. Med.* 382 (2020) 503–513, <https://doi.org/10.1056/NEJMoa191793>.
- [6] A. Khawaja, B.J. Bartholmai, S. Rajagopalan, R.A. Karwowski, C. Varghese, F. Maldonado, T. Peikert, Do we need to see to believe?—Radiomics for lung nodule classification and lung cancer risk stratification, *J. Thorac. Dis.* 12 (2020) 3303, <https://doi.org/10.21037/JTD.2020.03.105>.
- [7] J.D. Shur, S.J. Doran, S. Kumar, D. Ap Dafydd, K. Downey, J.P.B. O'connor, N. Papanikolaou, C. Messiou, D.M. Koh, M.R. Orton, Radiomics in oncology: a practical guide, *Radiographics*. 41 (2021) 1717–1732, <https://doi.org/10.1148/rg.2021210037>.

- [8] T. Peikert, B.J. Bartholmai, F. Maldonado, Radiomics-based Management of Indeterminate Lung Nodules? Are We There Yet? *Am. J. Respir. Crit. Care Med.* 202 (2020) 167–169, <https://doi.org/10.1164/rccm.202004-1279ED>.
- [9] J.L. Causey, J. Zhang, S. Ma, B. Jiang, J.A. Qualls, D.G. Politte, F. Prior, S. Zhang, X. Huang, Highly accurate model for prediction of lung nodule malignancy with CT scans, *Sci. Rep.* 8 (2018), <https://doi.org/10.1038/S41598-018-27569-W>.
- [10] W. Quanyang, H. Yao, W. Sicong, Q. Linlin, Z. Zewei, H. Donghui, L. Hongjia, Z. Shijun, Artificial intelligence in lung cancer screening: detection, classification, prediction, and prognosis, *Cancer Med.* 13 (2024), <https://doi.org/10.1002/CAM4.7140>.
- [11] S.S. Alahmari, D. Cherezov, D.B. Goldgof, L.O. Hall, R.J. Gillies, M.B. Schabath, Delta radiomics improves pulmonary nodule malignancy prediction in lung cancer screening, *IEEE Access.* 6 (2018) 77796–77806, <https://doi.org/10.1109/ACCESS.2018.2884126>.
- [12] L. Rundo, R.E. Ledda, C. Di Noia, E. Sala, G. Mauri, G. Milanese, N. Sverzellati, G. Apolone, M.C. Gilardi, M.C. Messa, I. Castiglioni, U. Pastorino, A low-dose CT-based radiomic model to improve characterization and screening recall intervals of indeterminate prevalent pulmonary nodules, *Diagnostics* 11 (2021) 1610, <https://doi.org/10.3390/diagnostics11091610>.
- [13] W. Choi, J.H. Oh, S. Riyahi, C.J. Liu, F. Jiang, W. Chen, C. White, A. Rimner, J. G. Mechalakos, J.O. Deasy, W. Lu, Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer, *Med. Phys.* 45 (2018) 1537–1549, <https://doi.org/10.1002/mp.12820>.
- [14] N. Garau, C. Paganelli, P. Summers, W. Choi, S. Alam, W. Lu, C. Fanciullo, M. Bellomi, G. Baroni, C. Rampinelli, External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis, *Med. Phys.* 47 (2020) 4125–4136, <https://doi.org/10.1002/mp.14308>.
- [15] T. Peikert, F. Duan, S. Rajagopalan, R.A. Karwoski, R. Clay, R.A. Robb, Z. Qin, J. R. Sicks, B.J. Bartholmai, F. Maldonado, Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial, *PLoS One* 13 (2018), <https://doi.org/10.1371/JOURNAL.PONE.0196910>.
- [16] L. Mao, H. Chen, M. Liang, K. Li, J. Gao, P. Qin, X. Ding, X. Li, X. Liu, Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose CT screening, *Quant. Imaging Med. Surg.* 9 (2019) 263–272, <https://doi.org/10.21037/qims.2019.02.02>.
- [17] J. Liu, H. Xu, H. Qing, Y. Li, X. Yang, C. He, J. Ren, P. Zhou, Comparison of radiomic models based on low-dose and standard-dose CT for prediction of adenocarcinomas and benign lesions in solid pulmonary nodules, *Front. Oncol.* 10 (2021) 1, <https://doi.org/10.3389/fonc.2020.634298>.
- [18] S.G. Armato, 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, E.A. Kazerooni, H. MacMahon, E.J.R. Van Beek, D. Yankelevitz, A.M. Biancardi, P.H. Bland, M.S. Brown, R. M. Engelmann, G.E. Laderach, D. Max, R.C. Pais, D.P.Y. Qing, R.Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G.W. Gladish, C.M. Jude, R. F. Munden, I. Petkovska, L.E. Quint, L.H. Schwartz, B. Sundaram, L.E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Vande Castele, S. Gupte, M. Sallam, M.D. Heath, M.H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B.Y. Croft, L.P. Clarke, The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a Completed Reference Database of Lung Nodules on CT Scans, *Med. Phys.* 38 (2011) 915, <https://doi.org/10.1118/1.3528204>.
- [19] D. Zeng, J. Huang, Z. Bian, S. Niu, H. Zhang, Q. Feng, Z. Liang, J. Ma, A Simple Low-dose X-ray CT Simulation from High-dose Scan, *IEEE Trans. Nucl. Sci.* 62 (2015) 2226, <https://doi.org/10.1109/TNS.2015.2467219>.
- [20] J. Ma, Z. Liang, Y. Fan, Y. Liu, J. Huang, W. Chen, H. Lu, Variance analysis of x-ray CT sinograms in the presence of electronic noise background, *Med. Phys.* 39 (2012) 4051–4065, <https://doi.org/10.1118/1.4722751>.
- [21] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository, *J. Digit. Imaging* 26 (2013) 1045–1057, <https://doi.org/10.1007/S10278-013-9622-7>.
- [22] C. Rampinelli, D. Origi, M. Bellomi, REVIEW Low-dose CT: technique, reading methods and image interpretation, *Cancer Imaging* 12 (2012) 548–556, <https://doi.org/10.1102/1470-7330.2012.0049>.
- [23] T. Snowsill, H. Yang, E. Griffin, L. Long, J. Varley-Campbell, H. Coelho, S. Robinson, C. Hyde, Low-dose computed tomography for lung cancer screening in high-risk populations: a systematic review and economic evaluation, *Health Technol. Assess. (Rockv)* 22 (2018) 1–276, <https://doi.org/10.3310/HTA22690>.
- [24] J.J.M. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H. Beets-Tan, J.C. Fillion-Robin, S. Pieper, H.J.W.L. Aerts, Computational radiomics system to decode the radiographic phenotype, *Cancer Res.* 77 (2017) e104, <https://doi.org/10.1158/0008-5472.CAN-17-0339>.
- [25] T.K. Koo, M.Y. Li, Cracking the code: providing insight into the fundamentals of research and evidence-based practice a guideline of selecting and reporting intraclass correlation coefficients for reliability research, (2016). <https://doi.org/10.1016/j.jcm.2016.02.012>.
- [26] F. Lo Iacono, R. Maragna, G. Pontone, V.D.A. Corino, A novel data augmentation method for radiomics analysis using image perturbations, *J. Imaging Inf. Med.* 2024 (2024) 1–14, <https://doi.org/10.1007/S10278-024-01013-0>.
- [27] K.M. Hasib, M.S. Iqbal, F.M. Shah, J. Al Mahmud, M.H. Popel, M.I.H. Showrov, S. Ahmed, O. Rahman, A survey of methods for managing the classification and solution of data imbalance problem, *J. Comput. Sci.* 16 (2020) 1546–1557, <https://doi.org/10.3844/JCSSP.2020.1546.1557>.
- [28] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, H.J.W.L. Aerts, Machine learning methods for quantitative radiomic biomarkers, *Sci. Rep.* 5 (1) (2015) 1–11, <https://doi.org/10.1038/srep13087>, 5 (2015).
- [29] F. Pedregosa, FABIANPEDREGOSA, V. Michel, O. Grisel, OLIVIERGRISEL, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Courapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot, Scikit-learn: machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot, J. Mach. Learn. Res. 12 (2011) 2825–2830. <http://scikit-learn.sourceforge.net>. accessed March 9, 2024.
- [30] S.M. Lundberg, P.G. Allen, S.-I. Lee, A Unified approach to interpreting model predictions, in: Conference on Neural Information Processing Systems, 2017. <https://github.com/slundberg/shap>. accessed March 7, 2024.
- [31] M. Silva, G. Milanese, R.E. Ledda, U. Pastorino, N. Sverzellati, Screen-detected solid nodules: from detection of nodule to structured reporting, *Transl. Lung Cancer Res.* 10 (2021) 2335, <https://doi.org/10.21037/TLCR-20-296>.
- [32] American College of Radiology Committee on Lung-RADS, Lung-RADS Assessment Categories 2022, (2022). <https://www.acr.org/-/media/ACR/Files/RADS/Lung-RADS/Lung-RADS-2022.pdf> (accessed May 8, 2024).

Low-Dose CT-Based Radiomic Model to Identify Malignant Lung Nodules

Jiaying Liu, Anna Corti, Valentina D.A. Corino, Luca Mainardi
Department of Electronics, Information and Bioengineering
Politecnico di milano, Milan, Italy
{jiaying.liu, anna.corti, valentina.corino, luca.mainardi}@polimi.it

Abstract— Lung cancer screening using low-dose computed tomography (LDCT) faces challenges in nodule classification and radiologist burden. This study introduces a novel approach leveraging degraded standard-dose CT (SDCT) images and develops an explainable radiomics-based model to improve lung nodule classification. Using data from LIDC-IDRI public dataset, our pipeline extracts and preprocesses 851 radiomic features per nodule, ensuring stability, discriminant characteristics, and non-redundancy. To address dataset imbalance, we applied data augmentation and random undersampling techniques. Varying the number of features selected by gradient-boosted decision tree across multiple classifiers, we identified an optimal combination of 15 features with random forest classifier, achieving balanced accuracy, sensitivity, specificity and AUC-ROC scores of 0.82 ± 0.04 , 0.81 ± 0.08 , 0.83 ± 0.05 , 0.88 ± 0.05 , respectively. Explainability analysis using SHAP highlights features influencing classification, including *major axis length*, *sphericity*, and *wavelet_LHH_ngtldm_Busyness*. In conclusion, our study demonstrates the potential of integrating degraded SDCT and radiomics to improve lung nodule classification.

Keywords— radiomics, low-dose computed tomography (LDCT), lung cancer, nodule classification.

I. INTRODUCTION

Lung cancer is the deadliest malignancy in the world, being responsible for 18% of total cancer deaths in 2020 [1]. While low-dose computed tomography (LDCT) screening has shown promise in reducing mortality rates compared to X-rays [2], the prevalence of false positive rates, reaching 96.4% of the positive screening, poses a significant challenge [2]. As screening programs are being implemented in the US and ongoing trials in Europe are assessing the benefits and cost-effectiveness of lung cancer screening, the workload on radiologists increases significantly. To overcome these challenges, extensive research has focused on developing objective and efficient diagnostic tools, with radiomics and deep learning emerging as promising methods [3].

Radiomics is based on the premise that image-captured characteristics of tumors, such as shape and texture, are potential tumor biomarkers that can be used to stratify patients, predict patient clinical outcomes or further characterize tumor phenotypes.

The development of radiomics-based models requires annotated datasets, which are currently lacking in the context of lung screening LDCT images. For instance, the extensive National Lung Screening Trial (NLST) contains LDCTs from 53,000 participants but lacks segmented nodules. On the other hand, the Lung Imaging Database Consortium and Image Database Resource Initiative (LIDC-IDRI) [4] consists of diagnostic and lung cancer screening CT scans with marked-up annotated lesions by experienced radiologists. However, the number of LDCT images is relatively limited (315 out of the 1010 CTs, considering images acquired with tube current < 80 mAs). To address this

challenge, we leverage the annotations from the LIDC-IDRI dataset by degrading standard-dose CT (SDCT) scans. This degradation step was essential to ensure compatibility between the different image types and facilitate the development of a classifier capable of effectively analyzing both SDCT and LDCT modalities.

Focusing on radiomics with machine learning studies applied to pulmonary nodule prediction, many have been shown to be promising in previous studies[5]. However, some considered small datasets ([6], [7]) and others achieved low sensitivity ([8], [9], [10]). For instance, Choi et al. [6] used two radiomics features and trained a support vector machine, obtaining scores of 0.89 and 0.87 for AUC and sensitivity. However, [6] only included 72 nodules in their study. On the other hand, Alahmari et al. [9] considered 467 patients in their study but obtained a low sensitivity of 0.56. Causey et al. [8] trained a logistic regression model using only one feature (square root of largest cross-sectional area) and obtained AUC score of 0.94, but also a relatively low sensitivity of 0.70. Likewise, Rundo et al. [10] obtained a sensitivity of 0.67 considering 32 radiomic features and logistic regression model. Other studies deliberately selected balanced datasets to address class imbalance [8], [11]. None of the aforementioned studies used degraded SDCT images to simulate LDCT.

The aim of this study is to build a novel explainable radiomic model to distinguish between likely benign and likely malign lung nodules, as annotated by radiological assessment, addressing the limitations associated with different dosage CT images, imbalanced dataset and false positive rates.

II. METHODS

The pipeline (Fig. 1) consists of data partition, image preprocessing, feature extraction and preprocessing, feature selection, model training and testing, iterated ten times with different seeds for increased robustness. In the end, the best model is chosen to perform explainability analysis.

A. Data description and partition

We randomly sampled ten times 400 participants from the publicly available LIDC-IDRI [4], which includes both SDCT and LDCT images. From the images, lesions ≥ 3 mm were segmented and evaluated by one to four radiologists in terms of likelihood of malignancy on a scale of 1 (highly unlikely) to 5 (highly suspicious). The consensus segmentation was considered and multiple scores for a single nodule were averaged to derive a consensus score. In this study, nodules with an average score ≥ 4 were considered likely malignant. Table I summarizes the mean and standard deviation of the number of nodules present in each randomly selected group of 400 participants across all ten seed selections. We split the 400 participants into train and test sets with an 8:2 ratio.

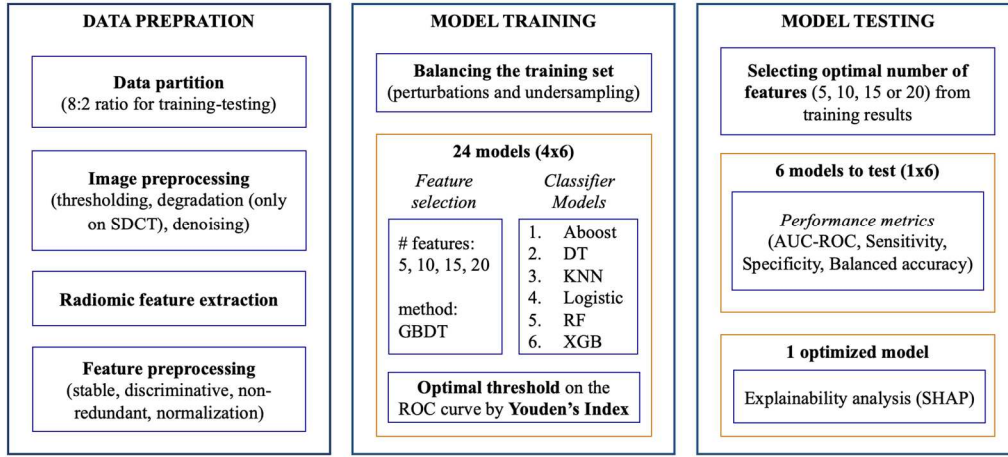


Fig. 1. Overview of the pipeline used in this study, consisted in data preparation, model training and testing. Data preparation includes partitioning, image preprocessing, radiomic feature extraction, and feature preprocessing. During model training, the training set is balanced, and the performance of six classifiers is evaluated using varying numbers of features selected with gradient boosted decision trees. The optimal ROC curve threshold is determined using Youden's index. Subsequently, models are tested with the selected optimal number of features, and the optimized model is selected based on performance metrics, followed by explainability analysis.

TABLE I. SUMMARY OF THE NUMBER OF NODULES USED IN THIS STUDY

Nodules	mean	std*
Total	1187	32
Train set		
Likely benign	853	29
Likely malign	102	7
Test set		
Likely benign	205	15
Likely malign	27	6

*Standard deviation

B. Image preprocessing

Image preprocessing involved thresholding, degradation (for SDCT only), and denoising, implemented in MATLAB version 2022b:

1. Thresholding ensured all pixels were positive by setting negative values to zero.
2. For SDCT images (CT acquired at tube current ≥ 80 mA), degradation based on Zeng et al.'s LDCT simulation technique [12] was applied. This method utilizes raw data from high-dose scans, obtains the raw sinogram data and injects Gaussian and Poisson noise into the sinogram domain to model electronic and quantum noise. The simulated LDCT images are reconstructed from the sinogram data. See more details in [12].
3. Denoising was performed using a 3D Gaussian filter with a $3 \times 3 \times 3$ voxel kernel and $\sigma = 0.5$.

C. Feature extraction and preprocessing

Radiomic features were extracted from the region of interest (ROI), i.e., the segmented nodules, using Pyradiomics [13], resulting in 851 features per nodule (14 shape features, 18 first-order features, 75 texture features and 744 wavelet-based features). For more details on radiomic features, see Pyradiomics documentation [13].

Feature preprocessing was performed on the training set and applied to the test set: it comprised steps to assure feature stability, discriminant characteristics, and non-redundancy. Stability analysis assessed feature consistency under small transformations. This analysis was conducted on nodules with

four radiologists' annotations using intra-class correlation coefficient (ICC). Features were deemed stable if the $ICC > 0.75$, as defined in [14]. Discriminative analysis identified informative features by comparing original ROIs with translated ROIs with no overlap with the nodule. In this analysis, features with $ICC < 0.5$ were considered discriminative. Finally, highly correlated features were removed by using the absolute Spearman's correlation coefficient > 0.85 .

D. Balancing the dataset

To address the imbalanced dataset, we first applied a data augmentation technique introduced in [15] consisting of the application of ROI perturbations on the malignant nodules of the training set (Fig. 2): 15% erosion (a) and dilation (b), as well as random contour adjustments (c). Afterwards, a random undersampling approach was adopted to reduce the non-malignant nodules to match the number of nodules in the minority class post-augmentation. Considering the ten seeds, the mean and standard deviation for nodule counts is 407 (± 28) per class after balancing the dataset.

E. Feature selection, model training and testing

Feature selection was performed on the training sets after balancing the classes using gradient-boosted decision tree (GBDT) feature selection method. The number of features selected was varied (5, 10, 15 or 20 features) centered on the sample-to-feature ratio guideline [16] suggesting "10-15 samples per feature" when the class is balanced or, when

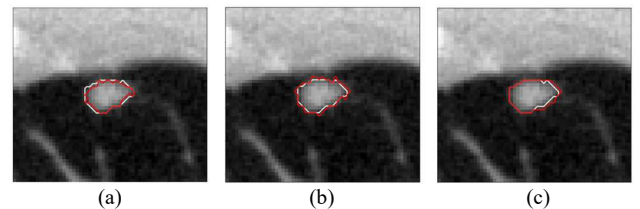


Fig. 2. Perturbation technique applied to the minority class for data augmentation: (a) erosion, (b) dilation, (c) random contour. The white contour represents the original consensus mask, while the red represents the perturbation.

imbalanced, apply this rule to the minority class. In our case, having an average of 129 nodules in the minority class, we could select 9-13 features. Guided by this rule, we explored an interval (5-20) centered on this range (9-13). Additionally, based on our previous work [17], which extensively explored feature selection methods, supports the decision to start with 5 features and incrementally increase by 5, up to 20. Afterwards, six machine learning classifiers were trained: random forest (RF), AdaBoost (ABOOST), XGBoost (XGB), decision tree (DT), K-nearest neighbors (KNN), and logistic regression (LR). The optimal cut-off on the ROC curve was determined by Youden's index during the training phase and applied to the testing.

This comprehensive analysis, resulting in a total of 24 models (4 number of features, 6 classifiers), aimed to identify the most effective number of features. To determine the optimal number of features (5, 10, 15 or 20), we compared the mean balanced accuracy of each number of features across all classifiers. After the identification of the optimal number of features, we obtain 6 models (1 number of features, 6 classifiers) which are evaluated on the test set using four performance metrics: balanced accuracy, sensitivity, specificity, area under the ROC curve (AUC-ROC).

Based on these performance metrics, we selected the best-performing configuration and used SHapley Additive exPlanations (SHAP) [18] to explain the predictions.

III. RESULTS

A systematic analysis was conducted to assess the impact of the number of features and machine learning algorithms by computing the mean and standard deviation across multiple seeds.

A. Optimal number of features

Fig. 3 represents the mean balanced accuracy achieved by each classifier in combination with a specific number of features selected by GBDT during the training phase. The graph shows consistent patterns of increase and decrease in mean balanced accuracy across different classifiers as the number of selected features varies. Moreover, each classifier was characterized by a general trend of increasing mean balanced accuracy as the number of selected features increased from 5 to 20 across most classifiers. Specifically,

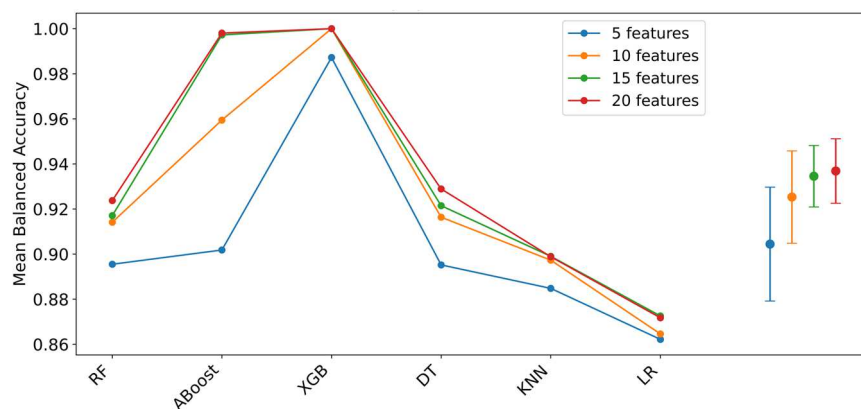


Fig. 3. Mean balanced accuracy based on the 10 seeds achieved by each classifier in combination with a specific number of features (5, 10, 15 and 20) selected by GBDT during the training phase.

XGB consistently achieved high mean balanced accuracy scores across all feature counts. By examining the overall mean and standard deviation calculated for each number of features, it can be observed that the results for 15 features and 20 features exhibit similar mean and standard deviation (t-test: $t = 0.78$, $p > 0.05$). In a parsimony way, we decided to select 15 features as the optimal number. Table II shows the number of features selected from each category, the frequency of the features selected and the average frequency considering 10 iterations by category. For instance, there were three features (major axis length, elongation and sphericity) from the original shape category that were selected 22 times globally (i.e., major axis length was selected 10 times, elongation 7 times and sphericity 5 times). Therefore, the average frequency for a feature from this category was 7.33 (22/3) out of the 10 iterations. Comparatively, the original shape category was the most predominant category of features selected and original first-order category the least selected.

B. Performance of classifiers in the test set

Fig. 4 shows the results on the test set of the classifiers trained with 15 features. The values reported were computed after utilizing the optimal cut-off determined by Youden's index during the training phase.

In terms of balanced accuracy and AUC-ROC, RF, KNN and LR demonstrated comparable performance, overperforming the other classifiers. When examining sensitivity and specificity, a trade-off is particularly visible in ABOOST and XGB classifiers, where high specificity values are associated with low sensitivity. In contrast, RF classifier maintained high scores for both sensitivity and specificity compared to the other classifiers.

TABLE II. FREQUENCY AND AVERAGE SELECTION OF FEATURES BY CATEGORY OVER 10 ITERATIONS.

Feature category		Nr. features	Total times selected	Average
Original	Shape	3	22	7.33
	First-order	2	2	1
	Textural	6	10	1.67
Wavelet	First-order	23	43	1.87
	Textural	43	73	1.70

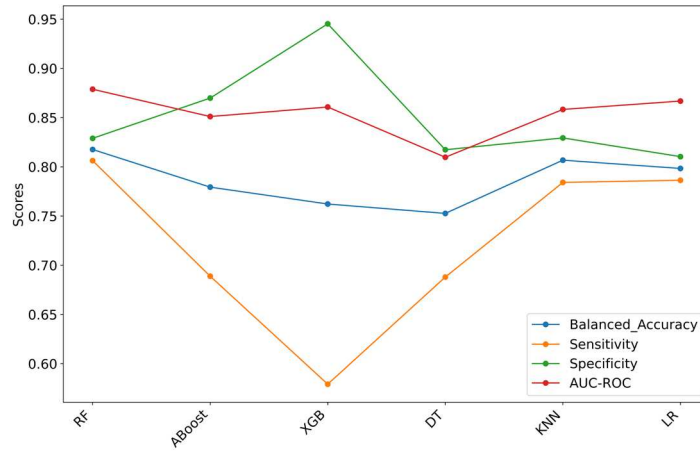


Fig. 4. Results achieved by different classifiers using 15 features selected by gradient boost decision tree on the test set.

Considering all performance metrics, RF emerged as the best performing classifier, achieving for balanced accuracy, sensitivity, specificity and AUC-ROC scores of 0.82 ± 0.04 , 0.81 ± 0.08 , 0.83 ± 0.05 , 0.88 ± 0.05 , respectively.

C. Explainability analysis

Considering RF model for explainability analysis, the top three most important features identified from the SHAP analysis (Fig. 5) were: *MajorAxisLength*, *sphericity* and *wavelet_LHH_ngtdm_Busyness*. The feature *MajorAxisLength*, representing the largest axis length of the ROI-enclosing ellipsoid, showed a positive correlation with malignancy likelihood, indicating that medium to higher values are associated with a greater likelihood of malignancy. Conversely, *sphericity*, which measures the roundness of the shape of the tumor region relative to a sphere, demonstrated an inverse relationship with malignancy likelihood. Specifically, lower values of sphericity (more irregular the nodule) were found to be associated with a higher likelihood of malignancy. The *wavelet_LHH_ngtdm_Busyness* feature is a textural feature computed from neighboring gray tone difference matrix (NGTDM) obtained after applying low-pass (L) and high-pass (H) filters. It measures the intensity change of a pixel considering its neighbors. A low busyness value indicates a gradual change of intensity between pixels and their neighborhood, which was overall associated with a lower likelihood of malignancy.

D. Comparison with original malignancy scores

As the original malignancy scores given by the radiologists ranged from 1 to 5, with nodules having an average score ≥ 4 considered malignant, we sought to determine how classified and misclassified instances in the

test set were spread across malignancy levels. This analysis was performed for the RF model using the 15 features selected by GBDT. The histogram in Fig. 6 reveals that a significant portion of misclassified nodules belonged to the category with an average score between 3 and 4, representing the most uncertain cases. In contrast, none of the nodules from class [1;2[(average score less than 2) were misclassified and there were very few misclassifications from the class [4;5] (average score of 4 or higher).

IV. DISCUSSION

In the present study, a novel explainable radiomic-based pipeline was developed using degraded SDCT and LDCT images to classify lung nodules based on the likelihood of malignancy. A key innovation stands in degrading SDCT images to leverage the annotations present in the LIDC-IDRI dataset, aiming to obtain a screening scenario where only LDCTs are present while addressing the limitations associated with different dosage CT images in the same dataset. Furthermore, we also addressed the challenge of imbalanced datasets by implementing a recent data augmentation technique involving creating perturbations in nodules present in the minority class and random undersampling of the majority class. To ensure the robustness of our findings, we conducted the pipeline ten times with varying seeds, enabling the generation of multiple data partitions. This approach enhances the reliability of our testing results by mitigating the influence of seed selection bias.

The final optimized model, consisting of 15 features selected by GBDT and trained with RF classifier, achieved the highest performance scores with balanced accuracy of

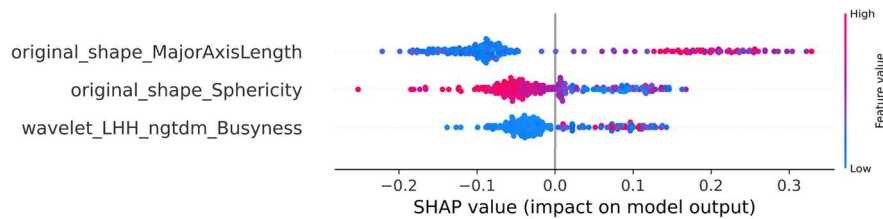


Fig. 5. SHAP beeswarm plot of the test set (seed 8) generated from the optimized model, 15 features selected by GBDT with RF classifier.

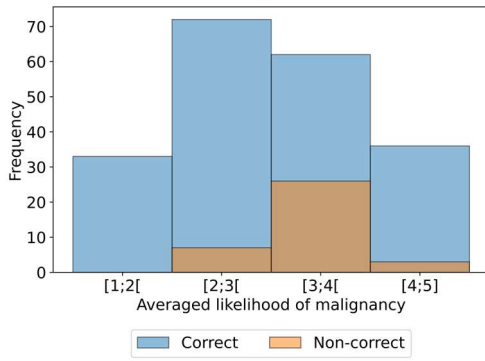


Fig. 6. Histogram representing the original label distribution (average likelihood of malignancy) of the correctly and incorrectly classified instances in the test set (seed 8).

0.82 ± 0.04 , sensitivity of 0.81 ± 0.08 , specificity of 0.83 ± 0.05 and auc-roc of 0.88 ± 0.05 . These results are comparable to or exceeded the performance of several studies in the literature (Table III), such as Alahmari et al. [9] (sensitivity of 0.49 and specificity of 0.93) and Rundo et al. [10] (sensitivity of 67.0% and specificity of 86.2%). Causey et al. [8] achieved notable results (sensitivity of 0.88 and specificity of 0.88) when using deep learning convolutional neural network features on the LIDC-IDRI dataset. However, when applying only radiomic features the sensitivity dropped to 0.70. Moreover, Causey et al. [8] also omitted nodules with an average score for the likelihood of malignancy between 3-4, which are the most challenging nodules to classify.

We recognize that larger datasets contribute to more stable and accurate models, enhancing reliability and generalizability. Our dataset comprised a larger number of nodules (400), surpassing the sample sizes of studies that achieved better performance with smaller datasets. For example, Choi et al. [6] considered 72 subjects, Mao et al. [19] 295 nodules, Garau et al. [7] 110 subjects, and Liu et al. [20] 141 nodules.

Although Peikert et al. [11] achieved better performance metrics (sensitivity = 0.90, specificity = 0.86) with a relatively large dataset, it is worth noting that their study included nodules larger than those analyzed in our study. Their dataset comprised nodules with a size threshold of ≥ 7 mm, whereas our study focused on nodules with a size threshold of ≥ 3 mm, considered optimal for screening [21]. In addition, research suggests that the lower the size of the nodule, the higher the rate of false positives [21], emphasizing the need for robust classification models in this context. Furthermore, Peikert et al. [11] selected data to balance the dataset, potentially introducing bias.

TABLE III. COMPARISON WITH RESULTS FOUND IN THE LITERATURE.

Author	N	Nodule size (mm)	Sen. %	Spec. %	AUC
Choi et al. [6]	72	≥ 3	87.2	81.2	0.89
Peikert et al. [11]	726	≥ 7	90.4	85.5	0.94
Alahmarie et al. [9]	467	≥ 4	49.0	93.0	0.82
Causey et al. [8]	664	≥ 3	69.7	95.5	0.94
Mao et al. [19]	294	≥ 6	81.0	92.2	0.90
Garau et al. [7]	110	-	90.0 ^a , 35.5 ^b	-	0.86
Liu et al. [20]	141	-	86.4	85.7	0.89
Rundo et al. [10]	11 012	-	67.0	86.2	0.82
Proposed	400	≥ 3	81.0	83.0	0.88

N: sample size; ^a true positive rate; ^b false positive rate.

Additionally, we employed the SHAP method to interpret the model's classification and identify key features. This interpretability analysis enabled us to mimic radiologists' scoring processes, providing valuable insights into the decision-making process. Specifically, identifying *MajorAxisLength* and *sphericity* as the most important features aligns with criteria used in the Lung Imaging Reporting and Data System (Lung-RADS) developed by the American College of Radiology to standardize the screening of lung cancer on CT images [22]. Moreover, the inclusion of textural wavelet features (*wavelet_LHH_ngtdm_Busyness*) among the top three important features suggests that the wavelet transform provided additional information not visible to the naked eye.

The present study is not exempt from limitations. For instance, the dataset under consideration was primarily curated to develop segmentation methods rather than for classification purposes or screening. In addition, nodule evaluation relied on radiologist assessments rather than histology-confirmed malignancy, leading to subjectivity in nodule evaluation.

However, our study successfully emulated the radiologists' approach to nodule evaluation, demonstrating the effectiveness of combining visually perceptible shape and size features with wavelet features, which offer insights beyond human perception.

V. CONCLUSION

In conclusion, the approach proposed herein demonstrated the potentiality of utilizing degrading SDCT images as a promising method for overcoming challenges associated with limited annotated LDCT images, as well as radiomics in the discrimination between the likely malign nodules and less likely ones.

REFERENCES

- [1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.
- [2] National Lung Screening Trial Research Team, "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, Aug. 2011, doi: 10.1056/NEJMoa1102873.
- [3] A. Khawaja *et al.*, "Do we need to see to believe?—radiomics for lung nodule classification and lung cancer risk stratification," *J Thorac Dis*, vol. 12, no. 6, p. 3303, Jun. 2020, doi: 10.21037/JTD.2020.03.105.
- [4] S. G. Armato *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," *Med Phys*, vol. 38, no. 2, p. 915, 2011, doi: 10.1118/1.3528204.
- [5] F. Zhu, C. Yang, J. Zou, W. Ma, Y. Wei, and Z. Zhao, "The classification of benign and malignant lung nodules based on CT radiomics: a systematic review, quality score assessment, and meta-analysis," <https://doi.org/10.1177/02841851231205737>, vol.

- 64, no. 12, pp. 3074–3084, Oct. 2023, doi: 10.1177/02841851231205737.
- [6] W. Choi *et al.*, “Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer,” *Med Phys*, vol. 45, no. 4, pp. 1537–1549, Apr. 2018, doi: 10.1002/mp.12820.
- [7] N. Garau *et al.*, “External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis,” *Med Phys*, vol. 47, no. 9, pp. 4125–4136, Sep. 2020, doi: 10.1002/mp.14308.
- [8] J. L. Causey *et al.*, “Highly accurate model for prediction of lung nodule malignancy with CT scans OPEN,” doi: 10.1038/s41598-018-27569-w.
- [9] S. S. Alahmari, D. Cherezov, D. B. Goldgof, L. O. Hall, R. J. Gillies, and M. B. Schabath, “Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening,” *IEEE Access*, vol. 6, pp. 77796–77806, 2018, doi: 10.1109/ACCESS.2018.2884126.
- [10] L. Rundo *et al.*, “A low-dose CT-based radiomic model to improve characterization and screening recall intervals of indeterminate prevalent pulmonary nodules,” *Diagnostics*, vol. 11, no. 9, p. 1610, Sep. 2021, doi: 10.3390/diagnostics11091610.
- [11] T. Peikert *et al.*, “Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial,” *PLoS One*, vol. 13, no. 5, May 2018, doi: 10.1371/JOURNAL.PONE.0196910.
- [12] D. Zeng *et al.*, “A Simple Low-dose X-ray CT Simulation from High-dose Scan,” *IEEE Trans Nucl Sci*, vol. 62, no. 5, p. 2226, Oct. 2015, doi: 10.1109/TNS.2015.2467219.
- [13] J. J. M. Van Griethuysen *et al.*, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Res*, vol. 77, no. 21, p. e104, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.
- [14] T. K. Koo and M. Y. Li, “Cracking the Code: Providing Insight Into the Fundamentals of Research and Evidence-Based Practice A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research,” 2016, doi: 10.1016/j.jcm.2016.02.012.
- [15] 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 Informatics in Medicine* 2024, pp. 1–14, May 2024, doi: 10.1007/S10278-024-01013-0.
- [16] J. D. Shur *et al.*, “Radiomics in oncology: A practical guide,” *Radiographics*, vol. 41, no. 6, pp. 1717–1732, Oct. 2021, doi: 10.1148/rg.2021210037.
- [17] J. Liu *et al.*, “Developing a robust two-step machine learning multiclassification pipeline to predict primary site in head and neck carcinoma from lymph nodes,” *Heliyon*, vol. 10, no. 2, Jan. 2024, doi: 10.1016/J.HELIYON.2024.E24377.
- [18] S. M. Lundberg, P. G. Allen, and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Conference on Neural Information Processing Systems*, 2017, Accessed: Mar. 07, 2024. [Online]. Available: <https://github.com/slundberg/shap>
- [19] L. Mao *et al.*, “Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose CT screening,” *Quant Imaging Med Surg*, vol. 9, no. 2, pp. 263–272, 2019, doi: 10.21037/qims.2019.02.02.
- [20] J. Liu *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,” *Front Oncol*, vol. 10, p. 1, Feb. 2021, doi: 10.3389/fonc.2020.634298.
- [21] M. Silva, G. Milanese, R. E. Ledda, U. Pastorino, and N. Sverzellati, “Screen-detected solid nodules: from detection of nodule to structured reporting,” *Transl Lung Cancer Res*, vol. 10, no. 5, p. 2335, May 2021, doi: 10.21037/TLCR-20-296.
- [22] American College of Radiology Committee on Lung-RADS, “Lung-RADS Assessment Categories 2022.” Accessed: May 08, 2024. [Online]. Available: <https://www.acr.org/-/media/ACR/Files/RADS/Lung-RADS/Lung-RADS-2022.pdf>

5 Discussion and Conclusions

5.1. Main findings

The present PhD aimed to advance the field of lung cancer screening by addressing three critical pillars of the screening workflow: (1) mitigating low-dose image degradation; (2) enhancing automated nodule detection; (3) improving radiomics-based malignancy classification.

The first limitation identified in the literature was the lack of reproducible CT degradation methods to simulate low-dose characteristics present in LDCT images used for lung cancer screening and the lack of annotated LDCT images to support the development of robust AI models to alleviate radiologist workload. Current literature focuses on image denoising to improve LDCT quality. However, these methods require raw projection data or paired SDCT-LDCT datasets which are non-existent. Most studies rely on paired data from *the 2016 LDCT Grand Challenge*, generated using a noise-insertion method. Consequently, the first study of this PhD shifts the focus from denoising LDCT to degrading SDCT images to generate diagnostically representative LDCT images. Accordingly, the first study's (**Publication I**) aim was to evaluate and compare multiple degradation techniques on CT images. To achieve this objective, a systematic comparison of different physical and synthetic degradation methods was conducted, including sinogram-domain statistical simulation (method 1), the replication of a validated physics-based method using Pix2Pix (method 2), and unpaired image generation using CycleGAN (method 3). This comparison was supported by the downstream task of lung nodule classification using radiomic features extracted from images degraded by each method. In this way, the lack of annotated LDCT images was overcome by applying degradation to annotated SDCT datasets.

The three degradation approaches showed distinct qualitative visual features and distinct quantitative performance in the nodule classification task. Method 1 produced visible but more subtle granular noise, whereas method 2 replicated pronounced streak artifacts and high-frequency noise that appeared visually consistent with photon starvation. However, despite visible noise in method 2, it yielded relatively lower sensitivity in the classification task. Conversely, method 3 produced perceptually blurrier images with subtle noise but obtained the

best FID (0.0316) and KID (0.0102) scores computed on the nodule patch. These results indicate that while method 3 may lack high-frequency sharpness, it achieved superior distributional alignment with the target low-dose domain. In the nodule classification task, method 3 yielded the most balanced classification performance with the highest sensitivity, achieving an AUC of 0.861, balanced accuracy of 0.800, sensitivity of 0.743 and specificity of 0.858 on the test set (composed of real LDCT only). The need for realistic degradation was confirmed by the poor generalization of the model trained on SDCT data, i.e., images without degradation. While the SDCT baseline model performed well during validation, it failed to generalize to the test set, yielding a collapsed sensitivity of 0.571. In contrast, the results demonstrated that synthetic data generated by all three methods can successfully train a classifier to perform well on real LDCT data ($\text{AUC} > 0.84$).

Therefore, the main contributions of the first study were:

- provided the first systematic comparison and evaluation of different low-dose CT degradation methods.
- introduce an approach that bypasses denoising and instead validates the feasibility of using SDCT images for developing downstream classification models.
- by providing a reliable method for simulating LDCT scans, these findings will not only aid in the development of more accurate denoising algorithms but also accelerate the inclusion of SDCT images into the development of nodule detection and classification models, ultimately improving the efficacy of lung cancer screening.

The second paper (Publication II) addresses the critical clinical task of lung nodule detection, where improving accuracy and reducing false positives has a significant real-world impact. While state-of-the-art transformer-based methods have demonstrated superior performance on many computer vision tasks, their effectiveness in the medical domain is limited by the lack of large-scale, curated datasets. To address this, the second study proposes and evaluates a two-stage pretraining strategy: initial pretraining on a large-scale natural video dataset (Kinetics-700), followed by further domain-specific pretraining on the large unlabeled lung CT dataset (NLST). By treating 3D CT volumes as sequential spatiotemporal data (analogous to video), vast amounts of unlabeled data can be exploited to pretrain the model. This forces the network to learn rich, contextual anatomical representations by reconstructing masked regions, providing a robust initialization for detecting nodules in 3D volumes without relying on large labeled datasets. To bridge the domain gap between natural videos and medical

volumes, this study systematically investigated the influence of input resolutions and embedding layer configurations on the Video-ViT backbone. Two volumetric configurations were evaluated: (1) $16 \times 224 \times 224$ input size aligned with the original pretraining resolution, using a 3D convolution with a $2 \times 16 \times 16$ kernel, resulting in $8 \times 14 \times 14$ tokens per volume; and (2) a second configuration to accommodate larger volumetric depth with an input size of $64 \times 128 \times 128$, using an $8 \times 16 \times 16$ kernel, producing $8 \times 8 \times 8$ tokens per volume. The study further investigated SwinUNETR pretraining mechanisms initialized with MOANI weights and using a crop size of $64 \times 128 \times 128$.

The quality of the learned representations was first assessed via the pretext task of image reconstruction. The model was trained to minimize the L2 reconstruction loss relative to the maximum input intensity. The loss decreased drastically from 4% in the model trained from scratch (random initialization) to 0.01% in the model utilizing large-scale video pretrained weights. The pretrained model significantly enhances reconstruction quality, accurately reconstructing lung structures, including fine details like blood vessels using only 10% of visible tokens. Conversely, training from scratch resulted in severe patch artifacts, demonstrating a lack of contextual understanding.

The superior representations learned during pretraining translated directly to the downstream task of lung nodule detection on the LUNA16 benchmark. Among the configurations, Video-ViT initiated with video weights and further pretrained using NLST achieved the best results on the downstream detection task with a mAP of 81.234% and an AP of 82.629% using the $16 \times 224 \times 224$ input configuration. In contrast, the $64 \times 128 \times 128$ configuration achieved lower results, mAP of 72.550% and AP of 75.258%, as the Video-ViT model required a partial reinitialization of the embedding layer, reducing the benefit of pre-training. On the other hand, while both transformers benefited from pretraining, Video-ViT outperformed SwinUNETR (mAP 78.723% and AP 81.327%) primarily for two reasons: 1) the medical image dataset used to pretrain SwinUNETR (5050 CT volumes) is significantly smaller compared to the video dataset (500,000 video clips); 2) the SwinUNETR model has fewer trainable parameters (26.7 million) than Video-ViT (338 million). The obtained results are comparable to recent approaches in the literature. For instance, Wu et al.[155] proposed YOLOv7-based framework using 2D images, achieving a mAP of 80.94% and 95.26% with data augmentation. Similarly, Ji et al.[99] modified YOLOv5 and achieved a mAP of 72.0%. Su et al. [156] applied Faster R-CNN to the LIDC-IDRI dataset and achieved an AP of 82.4% for the original model and 91.2% for the

optimized version. Some of these methods rely on 2D CNN architectures, while our 3D transformer-based approach achieves a mAP of 81.234% and AP of 82.629%, highlighting its competitiveness.

The main contributions of the second study can be summarized as follows:

- This work represents one of the first adaptations of MIM pretraining to the 3D nodule detection task, demonstrating its effectiveness on transformer-based backbones such as Video-ViT and SwinUNETR.
- The proposed framework leverages the spatial reasoning capabilities of the MIM mechanism, leading to impressive finetuning results on the LUNA16 dataset and outperforming the Retina U-net baseline.
- Through extensive experiments, the impact of different architectural designs and input patch sizes on the model's performance was analyzed. These findings advance the use of pretrained transformers for medical detection tasks.

Regarding the lung nodule classification task, the limitations identified in the literature included radiomic studies with low sensitivity or specificity [143], small sample sizes [147], [148], [150] and the absence of small lung nodules [145], [151]. The use of small datasets limits the reliability and generalizability of the models. Furthermore, some studies excluded small nodules which are critical for early diagnosis and more challenging to classify. To overcome these limitations, **chapter 4** of this thesis proposed two studies (**Publication III and IV**) with a larger sample size to develop an explainable radiomics-based model trained on larger datasets, including smaller nodules (≥ 3 mm) and achieved competitive results. Both studies used degraded SDCT images to leverage annotations present in the LIDC-IDRI dataset, aiming to simulate a screening scenario in which only LDCTs are present while addressing the limitations associated with the lack of annotated datasets. The conference paper included 400 subjects and used a mixed degraded SDCT and real LDCT in the model training and testing, with an 8:2 partition. The pipeline was conducted ten times with varying seeds, enabling the generation of multiple data partitions to ensure robustness and enhance the reliability of our testing results by mitigating the influence of seed selection bias. The final optimized model, consisting of 15 features selected by GBDT and trained with a RF classifier, achieved the highest performance scores with balanced accuracy of 0.82 ± 0.04 , sensitivity of 0.81 ± 0.08 , specificity of 0.83 ± 0.05 and AUC-ROC of 0.88 ± 0.05 . Additionally, SHAP method identified

major_axis_length and *sphericity* as the most important features. Moreover, the inclusion of textural wavelet features (*wavelet_LHH_ngtdm_Busyness*) among the top three important features suggests that the wavelet transform provided additional information not visible to the naked eye.

The extended journal paper included a bigger sample size of 2625 nodules. This study demonstrated the effectiveness of training a classifier exclusively from synthetic LDCT, generated by degrading SDCT images, and testing it exclusively on real LDCT images. This study further performed feature selection independently on three different subsets: a) only shape and size (SS) features, b) all features but SS features, and c) all features. These three sets were chosen to assess the relevance of nodule morphological characteristics on the prediction. Results revealed that extracting only three SS features from the lung nodules and using a simple LR classifier, yielded the most consistent results, achieving mean (confidence intervals) for balanced accuracy, sensitivity, specificity, AUC-ROC, and geometric-mean scores of 0.805 (0.804–0.805), 0.759 (0.757–0.760), 0.850 (0.847–0.853), 0.874 (0.874–0.874), 0.802 (0.802–0.803), respectively, on the test set. These results were comparable to or exceeded the performance of several studies in the literature, such as [143] and [144]. This extended study also included SHAP analysis in the pipeline which identified *axis_length* and *sphericity* as the top 3 most impactful features, both when only SS features were included in model training and also when all features were considered. This result is coherent with the previous study and highlights the importance of the SS feature in distinguishing likely benign from likely malignant lung nodules, aligning with one of the criteria used in the Lung-RADS to standardize the screening of lung cancer on CT images.

The main contributions of these last two studies were:

- Both studies used degraded SDCT images for classification model development, enabling to address the limited availability of annotated LDCT images for training while ensuring the validity of our results through testing on real LDCT images.
- Both studies contributed to improving nodule classification using radiomics to build an explainable machine learning model, advancing in this way the lung nodule classification.

5.2. Work limitations and future developments

Despite the contributions presented in this thesis, this work is not exempt from limitations.

First, the first study compared a broad range of degradation techniques; however, the methods relied on image domain or simplified statistical assumptions in the sinogram domain. These approximations may not fully capture the complex, non-uniform noise characteristics inherent to specific scanner properties, although quantitative metrics such as FID and KID demonstrated strong statistical similarity between the simulated and the real LDCT. Additionally, regarding the statistical analysis, although the Friedman and pairwise Wilcoxon tests indicated significant differences, the overlap in the 95% confidence intervals suggests that these differences might be partially attributed to the bootstrap sample size which makes the statistical tests more susceptible to small differences. Hence, **future work** should consider integrating scanner-specific parameters or leveraging deep learning in the projection domain to better account for non-uniform, realistic LDCT noise, leading to superior results with significant statistical differences and non-overlapping confidence intervals.

Second, the results and experiments on the second study are limited to a single fold of the LUNA16 dataset, whereas the standard LUNA16 benchmarking score requires 10-fold cross-validation. Furthermore, although the study validated the adaptation of well-established self-supervised learning frameworks (Video-MAE and SwinUNETR) to a new and important application domain, the field of computer vision evolves rapidly. Therefore, **future work** should complete 10-fold cross-validation to enable more direct comparisons with studies that report results using the official challenge metrics, explore newer self-supervised frameworks and extend the model’s capabilities to include nodule segmentation useful for radiomics analysis in nodule classification.

Third, the dataset used in the last two nodule classification studies was not originally tailored for nodule classification or screening, but for developing segmentation methods. Consequently, the nodule labels were based on radiologists’ assessments of malignancy likelihood. In this perspective, our study successfully emulated the radiologists’ approach to nodule evaluation, proposing a tool to support radiologists’ screening activity. **Future work** should use lung screening datasets with biopsy-confirmed ground truth to bridge the gap between this methodology and diagnostic clinical practice. Furthermore, although radiomics have the

advantage of being explainable, interpretable and computationally more efficient, DL models have demonstrated great potential in the literature. Therefore, **future research** should focus on training DL models directly on the degraded SDCT images generated in this thesis and develop hybrid models that complement radiomics with DL features, potentially improving classification performance while maintaining interpretability. Additionally, due to the order in which the studies were conducted, i.e., the studies in Chapter 4 were completed prior to the degradation method comparison in Chapter 2, the superior degradation technique identified in Chapter 2 could not be applied in Chapter 4. However, the classification results of the best degradation method can be found in chapter 2 for the sake of comparison.

Finally, although the results of our analysis can be considered robust given the large and multicentric nature of the LIDC-IDRI dataset, the lack of validation in external, independent cohorts remains an obstacle to clinical translation. **Future research** should focus on large-scale validation through prospective, multi-center clinical trials. To support this transition to real-time clinical contexts, it would be worthwhile to design a diagnostic system with continuous learning capabilities, i.e., implementing a cloud-based computing framework. This would guarantee machine-independent reproducibility and allow the system to adapt dynamically to new data, ensuring the tool remains effective across different medical centers [152], [157]

5.3. Conclusion

This PhD addresses multiple stages of the lung cancer screening process, including image quality, nodule detection and malignancy classification, by proposing innovative solutions to overcome the limitations of dataset and annotations scarcity.

The findings of this research confirm that:

1. Generating synthetic LDCT data from SDCT offers a promising data augmentation strategy to address challenges associated with limited annotated LDCT images. This work supports the use of unpaired generative adversarial networks as a scalable, cost-effective solution to the data scarcity in LDCT imaging.
2. By treating 3D medical scans as video sequences, it is possible to adapt self-supervised learning techniques from the natural domain to the medical domain. This work highlights the potential of leveraging vast amounts of unlabeled data to improve lung nodule detection, contributing toward the development of more efficient and accurate lung cancer screening tools.
3. Utilizing degraded SDCT images is a viable strategy for developing robust nodule classification models for screening-detected nodules. This approach expands the potential for the development of more generalizable models, ultimately enhancing model's ability to differentiate early-stage malignancy.

In conclusion, this thesis advances existing lung cancer screening methodologies, bridging the gap between data scarcity and clinical utility through the integration of synthetic image generation, self-supervised learning, and explainable AI.

Bibliography

- [1] D. M. Parkin, F. I. Bray, and S. S. Devesa, “Cancer burden in the year 2000. The global picture”, doi: 10.1016/S0959-8049(02)00739-6.
- [2] F. Bray Bsc *et al.*, “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” 2024, doi: 10.3322/caac.21834.
- [3] C. Allemani *et al.*, “Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries,” *The Lancet*, vol. 391, no. 10125, pp. 1023–1075, Mar. 2018, doi: 10.1016/S0140-6736(17)33326-3.
- [4] The National Lung Screening Trial Research Team, “Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening,” *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, Aug. 2011, doi: 10.1056/NEJMOA1102873;PAGEGROUP:STRING:PUBLICATION.
- [5] H. J. de Koning *et al.*, “Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial,” *New England Journal of Medicine*, vol. 382, no. 6, pp. 503–513, Feb. 2020, doi: 10.1056/NEJMOA1911793;WGROU:STRING:MMS.
- [6] J. H. Pedersen *et al.*, “The danish randomized lung cancer ct screening trial- overall design and results of the prevalence round,” *Journal of Thoracic Oncology*, vol. 4, no. 5, pp. 608–614, 2009, doi: 10.1097/JTO.0b013e3181a0d98f.
- [7] N. Becker *et al.*, “Lung cancer mortality reduction by LDCT screening-Results from the randomized German LUSI trial,” *Int. J. Cancer*, vol. 146, no. 6, pp. 1503–1513, Mar. 2020, doi: 10.1002/IJC.32486.
- [8] U. Pastorino *et al.*, “Ten-year results of the Multicentric Italian Lung Detection trial demonstrate the safety and efficacy of biennial lung cancer screening,” *Eur. J. Cancer*, vol. 118, pp. 142–148, Sep. 2019, doi: 10.1016/j.ejca.2019.06.009.
- [9] E. Paci *et al.*, “Mortality, survival and incidence rates in the ITALUNG randomised lung cancer screening trial,” *Thorax*, vol. 72, no. 9, pp. 825–831, Sep. 2017, doi: 10.1136/THORAXJNL-2016-209825.
- [10] C. M. van der Aalst, K. ten Haaf, and H. J. de Koning, “Lung cancer screening: latest developments and unanswered questions,” *Lancet Respir. Med.*, vol. 4, no. 9, pp. 749–761, Sep. 2016, doi: 10.1016/S2213-2600(16)30200-4.
- [11] D. C. L. Lam *et al.*, “Lung Cancer Screening in Asia: An Expert Consensus Report,” *Journal of Thoracic Oncology*, vol. 18, no. 10, pp. 1303–1322, Oct. 2023, doi: 10.1016/j.jtho.2023.06.014.
- [12] C. H. McCollough *et al.*, “Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge,” *Med. Phys.*, vol. 44, no. 10, pp. e339–e352, Oct. 2017, doi: 10.1002/MP.12345.

- [13] D. Peeters, B. Obreja, N. Antonissen, and C. Jacobs, “Benchmarking of Artificial Intelligence and Radiologists for Lung Cancer Screening in CT: The LUNA25 Challenge,” Mar. 2025, doi: 10.5281/ZENODO.15094631.
- [14] M. Gartenschläger *et al.*, “Pulmonary nodules: detection with low-dose vs conventional-dose spiral CT,” *Eur. Radiol.*, vol. 8, no. 4, pp. 609–614, 1998, doi: 10.1007/S003300050445.
- [15] K. N. Prasad, W. C. Cole, and G. M. Haase, “Radiation protection in humans: extending the concept of as low as reasonably achievable (ALARA) from dose to biological damage,” *Br. J. Radiol.*, vol. 77, no. 914, pp. 97–99, Feb. 2004, doi: 10.1259/BJR/88081058.
- [16] C. Rampinelli, D. Origgi, and M. Bellomi, “Low-dose CT: technique, reading methods and image interpretation,” *Cancer Imaging*, vol. 12, no. 3, p. 548, 2013, doi: 10.1102/1470-7330.2012.0049.
- [17] E. Eulig, B. Ommer, and M. Kachelrieß, “Benchmarking deep learning-based low-dose CT image denoising algorithms,” *Med. Phys.*, vol. 51, no. 12, pp. 8776–8788, Dec. 2024, doi: 10.1002/MP.17379;WGROU:STRING:PUBLICATION.
- [18] N. Karabulut, M. Törü, V. Gelebek, M. Gülsün, and O. M. Ariyürek, “Comparison of low-dose and standard-dose helical CT in the evaluation of pulmonary nodules,” *Eur. Radiol.*, vol. 12, no. 11, pp. 2764–2769, Nov. 2002, doi: 10.1007/S00330-002-1368-4/METRICS.
- [19] J. R. Mayo, T. E. Hartman, Kyung Soo Lee, S. L. Primack, S. Vedal, and N. L. Muller, “CT of the chest: minimal tube current required for good image quality with the least radiation dose,” *AJR Am. J. Roentgenol.*, vol. 164, no. 3, pp. 603–607, 1995, doi: 10.2214/AJR.164.3.7863879.
- [20] T. Kubo *et al.*, “Low dose chest CT protocol (50 mAs) as a routine protocol for comprehensive assessment of intrathoracic abnormality,” *Eur. J. Radiol. Open*, vol. 3, pp. 86–94, Jan. 2016, doi: 10.1016/J.EJRO.2016.04.001.
- [21] A. S. Bhalla, A. Das, P. Naranje, A. Irodi, V. Raj, and A. Goyal, “Imaging protocols for CT chest: A recommendation,” *Indian J. Radiol. Imaging*, vol. 29, no. 3, p. 236, Jul. 2019, doi: 10.4103/IJRI.IJRI_34_19.
- [22] C. A. Gatsonis *et al.*, “The national lung screening trial: Overview and study design,” *Radiology*, vol. 258, no. 1, pp. 243–253, Jan. 2011, doi: 10.1148/RADIOL.10091808/-/DC1.
- [23] T. Snowsill *et al.*, “Low-dose computed tomography for lung cancer screening in high-risk populations: a systematic review and economic evaluation,” *Health Technol. Assess.*, vol. 22, no. 69, pp. 1–276, Nov. 2018, doi: 10.3310/HTA22690.
- [24] A. L. Pegna *et al.*, “Four-Year Results of Low-Dose CT Screening and Nodule Management in the ITALUNG Trial,” *Journal of Thoracic Oncology*, vol. 8, no. 7, pp. 866–875, Jul. 2013, doi: 10.1097/JTO.0B013E31828F68D6.
- [25] J. Liu, Q. Ma, A. Corti, V. D. A. Corino, L. Mainardi, and E. Konukoglu, “Leveraging Self-supervised Pretraining Using Transformers for Enhanced Lung Nodule Detection in CT Scans,” *Lecture Notes in*

- Computer Science*, vol. 16241 LNCS, pp. 106–116, 2026, doi: 10.1007/978-3-032-09513-8_11/TABLES/3.
- [26] K. A. S. H. Kulathilake, N. A. Abdullah, A. Q. M. Sabri, and K. W. Lai, “A review on Deep Learning approaches for low-dose Computed Tomography restoration,” *Complex & intelligent systems*, vol. 9, no. 3, pp. 2713–2745, Jun. 2023, doi: 10.1007/S40747-021-00405-X.
 - [27] D. Karimi, P. Deman, R. Ward, and N. Ford, “A sinogram denoising algorithm for low-dose computed tomography,” *BMC Medical Imaging 2016 16:1*, vol. 16, no. 1, pp. 11–, Jan. 2016, doi: 10.1186/S12880-016-0112-5.
 - [28] Y. Wang, W. Li, S. Fu, and C. Zhang, “Adaptive filtering with self-similarity for low-dose CT imaging,” *Optik (Stuttgart)*, vol. 126, no. 24, pp. 4949–4953, Dec. 2015, doi: 10.1016/J.IJLEO.2015.09.128.
 - [29] A. Manduca *et al.*, “Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT,” *Med. Phys.*, vol. 36, no. 11, pp. 4911–4919, Nov. 2009, doi: 10.1118/1.3232004;JOURNAL:JOURNAL:24734209;REQUESTEDJOURNAL:JOURNAL:24734209;WGROU:STRING:PUBLICATION.
 - [30] S. Tang and X. Tang, “Statistical CT noise reduction with multiscale decomposition and penalized weighted least squares in the projection domain,” *Med. Phys.*, vol. 39, no. 9, pp. 5498–5512, Sep. 2012, doi: 10.1118/1.4745564;PAGE:STRING:ARTICLE/CHAPTER.
 - [31] Y. Chen *et al.*, “Thoracic low-dose CT image processing using an artifact suppressed large-scale nonlocal means,” *Phys. Med. Biol.*, vol. 57, no. 9, p. 2667, Apr. 2012, doi: 10.1088/0031-9155/57/9/2667.
 - [32] Y. Zhang, Y. Wang, W. Zhang, F. Lin, Y. Pu, and J. Zhou, “Statistical iterative reconstruction using adaptive fractional order regularization,” *Biomed. Opt. Express*, vol. 7, no. 3, p. 1015, Mar. 2016, doi: 10.1364/BOE.7.001015.
 - [33] Q. Xu, H. Y. Yu, X. Q. Mou, L. Zhang, J. Hsieh, and G. Wang, “Low-dose X-ray CT reconstruction via dictionary learning,” *IEEE Trans. Med. Imaging*, vol. 31, no. 9, pp. 1682–1697, 2012, doi: 10.1109/TMI.2012.2195669.
 - [34] J. F. Cai, X. Jia, H. Gao, S. B. Jiang, Z. Shen, and H. Zhao, “Cine cone beam CT reconstruction using low-rank matrix factorization: Algorithm and a proof-of-principle study,” *IEEE Trans. Med. Imaging*, vol. 33, no. 8, pp. 1581–1591, 2014, doi: 10.1109/TMI.2014.2319055.
 - [35] K. A. S. H. Kulathilake, N. A. Abdullah, A. Q. M. Sabri, and K. W. Lai, “A review on Deep Learning approaches for low-dose Computed Tomography restoration,” *Complex & intelligent systems*, vol. 9, no. 3, pp. 2713–2745, Jun. 2023, doi: 10.1007/S40747-021-00405-X.
 - [36] H. Kong, W. Gao, X. Du, and Y. Di, “An improved non-local means algorithm for CT image denoising,” *Multimedia Systems 2024 30:2*, vol. 30, no. 2, pp. 79–, Mar. 2024, doi: 10.1007/S00530-024-01283-2.

- [37] Y. Yuan, Y. Zhang, and H. Yu, "Adaptive Nonlocal Means Method for Denoising Basis Material Images from Dual-Energy Computed Tomography," *J. Comput. Assist. Tomogr.*, vol. 42, no. 6, pp. 972–981, Nov. 2018, doi: 10.1097/RCT.0000000000000805.
- [38] Y. Zhang, M. Salehjahromi, and H. Yu, "Tensor decomposition and non-local means based spectral CT image denoising," *J. Xray Sci. Technol.*, vol. 27, no. 3, pp. 397–416, May 2019, doi: 10.3233/XST-180413;WGROU:STRING:PUBLICATION.
- [39] L. Jia *et al.*, "Denoising for low-dose CT image by discriminative weighted nuclear norm minimization," *IEEE Access*, vol. 6, pp. 46179–46193, Aug. 2018, doi: 10.1109/ACCESS.2018.2862403.
- [40] S. Hashemi, N. S. Paul, S. Beheshti, and R. S. C. Cobbold, "Adaptively Tuned Iterative Low Dose CT Image Denoising," *Comput. Math. Methods Med.*, vol. 2015, no. 1, p. 638568, Jan. 2015, doi: 10.1155/2015/638568.
- [41] R. J. Geraldo, L. M. V. Cura, P. E. Cruvinel, and N. D. A. Mascarenhas, "Low Dose CT Filtering in the Image Domain Using MAP Algorithms," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 1, no. 1, pp. 56–67, Jan. 2017, doi: 10.1109/TNS.2016.2635131.
- [42] W. Kim, S. Y. Jeon, G. Byun, H. Yoo, and J. H. Choi, "A systematic review of deep learning-based denoising for low-dose computed tomography from a perceptual quality perspective," *Biomedical Engineering Letters* 2024 14:6, vol. 14, no. 6, pp. 1153–1173, Aug. 2024, doi: 10.1007/S13534-024-00419-7.
- [43] Q. Ding, Y. Long, X. Zhang, and J. A. Fessler, "Statistical Image Reconstruction Using Mixed Poisson-Gaussian Noise Model for X-Ray CT," Jan. 2018, Accessed: Dec. 23, 2025. [Online]. Available: <https://arxiv.org/pdf/1801.09533>
- [44] L. Huang, H. Jiang, S. Li, Z. Bai, and J. Zhang, "Two stage residual CNN for texture denoising and structure enhancement on low dose CT image," *Comput. Methods Programs Biomed.*, vol. 184, p. 105115, Feb. 2020, doi: 10.1016/J.CMPB.2019.105115.
- [45] H. Chen *et al.*, "Low-dose CT denoising with convolutional neural network," *Proceedings - International Symposium on Biomedical Imaging*, pp. 143–146, Jun. 2017, doi: 10.1109/ISBI.2017.7950488.
- [46] T. Zhao, M. McNitt-Gray, and D. Ruan, "A convolutional neural network for ultra-low-dose CT denoising and emphysema screening," *Med. Phys.*, vol. 46, no. 9, pp. 3941–3950, Sep. 2019, doi: 10.1002/MP.13666;JOURNAL:JOURNAL:24734209;ISSUE:ISSUE:DOI.
- [47] S. Žabić, Q. Wang, T. Morton, and K. M. Brown, "A low dose simulation tool for CT systems with energy integrating detectors," *Med. Phys.*, vol. 40, no. 3, 2013, doi: 10.1118/1.4789628.
- [48] S. Young, H. J. G. Kim, M. M. Ko, W. W. Ko, C. Flores, and M. F. McNitt-Gray, "Variability in CT lung-nodule volumetry: Effects of dose reduction and reconstruction methods," *Med. Phys.*, vol. 42, no. 5, May 2015, doi: 10.1118/1.4918919.

- [49] M. Gholizadeh-Ansari, J. Alirezaie, and P. Babyn, “Deep Learning for Low-Dose CT Denoising Using Perceptual Loss and Edge Detection Layer,” *Journal of Digital Imaging* 2019 33:2, vol. 33, no. 2, pp. 504–515, Sep. 2019, doi: 10.1007/S10278-019-00274-4.
- [50] A. Zhong, B. Li, N. Luo, Y. Xu, L. Zhou, and X. Zhen, “Image Restoration for Low-Dose CT via Transfer Learning and Residual Network,” *IEEE Access*, vol. 8, pp. 112078–112091, 2020, doi: 10.1109/ACCESS.2020.3002534.
- [51] D. Zeng *et al.*, “A simple low-dose X-Ray CT simulation from high-dose scan,” *IEEE Trans. Nucl. Sci.*, vol. 62, no. 5, pp. 2226–2233, Oct. 2015, doi: 10.1109/TNS.2015.2467219.
- [52] I. J. Goodfellow *et al.*, “Generative Adversarial Nets,” Jun. 2014, Accessed: Jan. 20, 2026. [Online]. Available: <http://www.github.com/goodfeli/adversarial>
- [53] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. doi: 10.1109/CVPR.2017.632.
- [54] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, “Generative adversarial networks for noise reduction in low-dose CT,” *IEEE Trans. Med. Imaging*, vol. 36, no. 12, pp. 2536–2545, Dec. 2017, doi: 10.1109/TMI.2017.2708987.
- [55] Google Developers, “Overview of GAN Structure,” Generative Adversarial Networks. Accessed: Jan. 20, 2026. [Online]. Available: https://developers.google.com/machine-learning/gan/gan_structure
- [56] T. R. Moen *et al.*, “Low-dose CT image and projection dataset,” *Med. Phys.*, vol. 48, no. 2, pp. 902–911, Feb. 2021, doi: 10.1002/MP.14594.
- [57] H. Chen *et al.*, “Low-Dose CT with a residual encoder-decoder convolutional neural network,” *IEEE Trans. Med. Imaging*, vol. 36, no. 12, pp. 2524–2535, Dec. 2017, doi: 10.1109/TMI.2017.2715284.
- [58] W. Du, H. Chen, P. Liao, H. Yang, G. Wang, and Y. Zhang, “Visual Attention Network for Low-Dose CT,” *IEEE Signal Process. Lett.*, vol. 26, no. 8, pp. 1152–1156, Aug. 2019, doi: 10.1109/LSP.2019.2922851.
- [59] F. Fan *et al.*, “Quadratic Autoencoder (Q-AE) for Low-Dose CT Denoising,” *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 2035–2050, Jun. 2020, doi: 10.1109/TMI.2019.2963248.
- [60] E. Kang, W. Chang, J. Yoo, and J. C. Ye, “Deep Convolutional Framelet Denosing for Low-Dose CT via Wavelet Residual Network,” *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1358–1369, Jun. 2018, doi: 10.1109/TMI.2018.2823756.
- [61] E. Kang, J. Min, and J. C. Ye, “A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction,” *Med. Phys.*, vol. 44, no. 10, pp. e360–e375, Oct. 2017, doi: 10.1002/MP.12344.

- [62] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "SACNN: Self-Attention Convolutional Neural Network for Low-Dose CT Denoising with Self-Supervised Perceptual Loss Network," *IEEE Trans. Med. Imaging*, vol. 39, no. 7, pp. 2289–2301, Jul. 2020, doi: 10.1109/TMI.2020.2968472.
- [63] H. Shan *et al.*, "3-D Convolutional Encoder-Decoder Network for Low-Dose CT via Transfer Learning From a 2-D Trained Network," *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1522–1534, Jun. 2018, doi: 10.1109/TMI.2018.2832217.
- [64] L. Yang, H. Shanguan, X. Zhang, A. Wang, and Z. Han, "High-Frequency Sensitive Generative Adversarial Network for Low-Dose CT Image Denoising," *IEEE Access*, vol. 8, pp. 930–943, 2020, doi: 10.1109/ACCESS.2019.2961983.
- [65] Q. Yang *et al.*, "Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss," *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018, doi: 10.1109/TMI.2018.2827462.
- [66] W. Yang *et al.*, "Improving Low-Dose CT Image Using Residual Convolutional Network," *IEEE Access*, vol. 5, pp. 24698–24705, Oct. 2017, doi: 10.1109/ACCESS.2017.2766438.
- [67] X. Yin *et al.*, "Domain Progressive 3D Residual Convolution Network to Improve Low-Dose CT Imaging," *IEEE Trans. Med. Imaging*, vol. 38, no. 12, pp. 2903–2913, Dec. 2019, doi: 10.1109/TMI.2019.2917258.
- [68] X. Yin *et al.*, "Domain Progressive 3D Residual Convolution Network to Improve Low-Dose CT Imaging," *IEEE Trans. Med. Imaging*, vol. 38, no. 12, pp. 2903–2913, Dec. 2019, doi: 10.1109/TMI.2019.2917258.
- [69] C. You *et al.*, "Structurally-sensitive multi-scale deep neural network for low-dose CT denoising," *IEEE Access*, vol. 6, pp. 41839–41855, Jul. 2018, doi: 10.1109/ACCESS.2018.2858196.
- [70] M. Zubair, H. B. Md Rais, F. Ullah, Q. Al-Tashi, M. Faheem, and A. Ahmad Khan, "Enabling Predication of the Deep Learning Algorithms for Low-Dose CT Scan Image Denoising Models: A Systematic Literature Review," *IEEE Access*, vol. 12, pp. 79025–79050, 2024, doi: 10.1109/ACCESS.2024.3407774.
- [71] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 2242–2251, Mar. 2017, doi: 10.1109/ICCV.2017.244.
- [72] J. Gu and J. C. Ye, "AdaIN-Switchable CycleGAN for Efficient Unsupervised Low-Dose CT Denoising," *IEEE Trans. Comput. Imaging*, vol. 7, pp. 73–85, Aug. 2020, doi: 10.1109/TCI.2021.3050266.
- [73] K. Kim, S. Soltanayev, and S. Y. Chun, "Unsupervised training of denoisers for low-dose ct reconstruction without full-dose ground truth," *IEEE Journal on Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1112–1125, Oct. 2020, doi: 10.1109/JSTSP.2020.3007326.

- [74] J. Lehtinen *et al.*, “Noise2Noise: Learning Image Restoration without Clean Data,” *35th International Conference on Machine Learning, ICML 2018*, vol. 7, pp. 4620–4631, Mar. 2018, Accessed: Dec. 28, 2025. [Online]. Available: <https://arxiv.org/pdf/1803.04189>
- [75] A. Krull, T. O. Buchholz, and F. Jug, “Noise2void-Learning denoising from single noisy images,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 2124–2132, Jun. 2019, doi: 10.1109/CVPR.2019.00223.
- [76] J. Batson and L. Royer, “Noise2Self: Blind Denoising by Self-Supervision,” May 24, 2019, *PMLR*. Accessed: Jan. 21, 2026. [Online]. Available: <https://proceedings.mlr.press/v97/batson19a.html>
- [77] C. Niu *et al.*, “Suppression of Correlated Noise with Similarity-based Unsupervised Deep Learning,” Nov. 2020, Accessed: Dec. 29, 2025. [Online]. Available: <https://arxiv.org/pdf/2011.03384>
- [78] Y. Zhang, D. Li, K. L. Law, X. Wang, H. Qin, and H. Li, “IDR: Self-Supervised Image Denoising via Iterative Data Refinement”, Accessed: Jan. 21, 2026. [Online]. Available: <https://github.com/zhangyi-3/IDR>
- [79] F. Wagner *et al.*, “Ultralow-parameter denoising: Trainable bilateral filter layers in computed tomography,” *Med. Phys.*, vol. 49, no. 8, pp. 5107–5120, Aug. 2022, doi: 10.1002/MP.15718.
- [80] K. Choi, “A Comparative Study between Image- and Projection-Domain Self-Supervised Learning for Ultra Low-Dose CBCT,” *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2022-July, pp. 2076–2079, 2022, doi: 10.1109/EMBC48229.2022.9871947.
- [81] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, “SACNN: Self-Attention Convolutional Neural Network for Low-Dose CT Denoising With Self-Supervised Perceptual Loss Network,” *IEEE Trans. Med. Imaging*, vol. 39, no. 7, pp. 2289–2301, Jul. 2020, doi: 10.1109/TMI.2020.2968472.
- [82] J. Jing *et al.*, “Training low dose CT denoising network without high quality reference data,” *Phys. Med. Biol.*, vol. 67, no. 8, Apr. 2022, doi: 10.1088/1361-6560/AC5F70.
- [83] E. Zainulina, A. Chernyavskiy, and D. V. Dylov, “Self-supervised Physics-based Denoising for Computed Tomography,” Nov. 2022, Accessed: Jan. 21, 2026. [Online]. Available: <https://arxiv.org/pdf/2211.00745>
- [84] M. B. Kiss, A. Biguri, C. B. Schönlieb, K. J. Batenburg, and F. Lucka, “Learned denoising with simulated and experimental low-dose CT data,” *Scientific Reports* 2025 16:1, vol. 16, no. 1, pp. 835–, Dec. 2025, doi: 10.1038/s41598-025-30457-9.
- [85] 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,” *J. Comput. Assist. Tomogr.*, vol. 36, no. 4, pp. 477–487, 2012, doi: 10.1097/RCT.0B013E318258E891,.
- [86] A. Horé and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” *Proceedings - International Conference on Pattern Recognition*, pp. 2366–2369, 2010, doi: 10.1109/ICPR.2010.579.

- [87] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1398–1402, 2003, doi: 10.1109/ACSSC.2003.1292216.
- [88] F. R. Verdun *et al.*, "Image quality in CT: From physical measurements to model observers," *Phys. Med.*, vol. 31, no. 8, pp. 823–843, Dec. 2015, doi: 10.1016/J.EJMP.2015.08.007.
- [89] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [90] S. Dodia, B. Annappa, and P. A. Mahesh, "Recent advancements in deep learning based lung cancer detection: A systematic review," *Eng. Appl. Artif. Intell.*, vol. 116, p. 105490, Nov. 2022, doi: 10.1016/J.ENGAPPAL.2022.105490.
- [91] A. Schreuder, E. T. Scholten, B. van Ginneken, and C. Jacobs, "Artificial intelligence for detection and characterization of pulmonary nodules in lung cancer CT screening: ready for practice?," *Transl. Lung Cancer Res.*, vol. 10, no. 5, pp. 2378–2388, May 2021, doi: 10.21037/TLCR-2020-LCS-06.
- [92] A. A. A. Setio *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, Dec. 2017, doi: 10.1016/J.MEDIA.2017.06.015.
- [93] P. F. Pinsky, D. S. Gierada, P. H. Nath, E. Kazerooni, and J. Amorosa, "National Lung Screening Trial: Variability in Nodule Detection Rates in Chest CT Studies," *Radiology*, vol. 268, no. 3, p. 865, Sep. 2013, doi: 10.1148/RADIOL.13121530.
- [94] H. MacMahon *et al.*, "Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017," *Radiology*, vol. 284, no. 1, pp. 228–243, Jul. 2017, doi: 10.1148/RADIOL.2017161659.
- [95] N. Gardiner, S. Jogai, and A. Wallis, "The revised lung adenocarcinoma classification—an imaging guide," *J. Thorac. Dis.*, vol. 6, no. Suppl 5, p. S537, Oct. 2014, doi: 10.3978/J.ISSN.2072-1439.2014.04.05.
- [96] H. U. Kauczor *et al.*, "ESR/ERS statement paper on lung cancer screening," *Eur. Radiol.*, vol. 30, no. 6, pp. 3277–3294, Jun. 2020, doi: 10.1007/S00330-020-06727-7.
- [97] H. Chen *et al.*, "Genomic and immune profiling of pre-invasive lung adenocarcinoma," *Nature Communications 2019 10:1*, vol. 10, no. 1, pp. 5472–, Nov. 2019, doi: 10.1038/s41467-019-13460-3.
- [98] S. Cui *et al.*, "Development and clinical application of deep learning model for lung nodules screening on CT images," *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 13657–, Aug. 2020, doi: 10.1038/s41598-020-70629-3.
- [99] Z. Ji *et al.*, "Lung Nodule Detection in Medical Images Based on Improved YOLOv5s," *IEEE Access*, vol. 11, pp. 76371–76387, 2023, doi: 10.1109/ACCESS.2023.3296530.

- [100] Y. Li, L. Hui, X. Wang, L. Zou, and S. Chua, "Lung nodule detection using a multi-scale convolutional neural network and global channel spatial attention mechanisms," *Scientific Reports* 2025 15:1, vol. 15, no. 1, pp. 12313-, Apr. 2025, doi: 10.1038/s41598-025-97187-w.
- [101] L. Ma, G. Li, X. Feng, Q. Fan, and L. Liu, "TiCNet: Transformer in Convolutional Neural Network for Pulmonary Nodule Detection on CT Images," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, p. 196, Feb. 2024, doi: 10.1007/S10278-023-00904-Y.
- [102] Z. Zhou, F. Gou, Y. Tan, and J. Wu, "A Cascaded Multi-Stage Framework for Automatic Detection and Segmentation of Pulmonary Nodules in Developing Countries," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5619–5630, Nov. 2022, doi: 10.1109/JBHI.2022.3198509.
- [103] S. P. Primakov *et al.*, "Automated detection and segmentation of non-small cell lung cancer computed tomography images," *Nature Communications* 2022 13:1, vol. 13, no. 1, pp. 3423-, Jun. 2022, doi: 10.1038/s41467-022-30841-3.
- [104] S. Dlamini, Y. H. Chen, and C. F. Jeffrey Kuo, "Complete fully automatic detection, segmentation and 3D reconstruction of tumor volume for non-small cell lung cancer using YOLOv4 and region-based active contour model," *Expert Syst. Appl.*, vol. 212, p. 118661, Feb. 2023, doi: 10.1016/J.ESWA.2022.118661.
- [105] G. Aresta *et al.*, "Automatic Lung Nodule Detection Combined With Gaze Information Improves Radiologists' Screening Performance," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2894–2901, Oct. 2020, doi: 10.1109/JBHI.2020.2976150.
- [106] X. Cui *et al.*, "Performance of a deep learning-based lung nodule detection system as an alternative reader in a Chinese lung cancer screening program," *Eur. J. Radiol.*, vol. 146, Jan. 2022, doi: 10.1016/j.ejrad.2021.110068.
- [107] J. Christensen *et al.*, "ACR Lung-RADS v2022: Assessment Categories and Management Recommendations.," *Chest*, vol. 165, no. 3, pp. 738–753, Mar. 2024, doi: 10.1016/j.chest.2023.10.028.
- [108] S. K. Thakur, D. P. Singh, and J. Choudhary, "Lung cancer identification: a review on detection and classification," *Cancer Metastasis Rev.*, vol. 39, no. 3, pp. 989–998, Sep. 2020, doi: 10.1007/S10555-020-09901-X.
- [109] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, Jun. 2020, Accessed: Jan. 14, 2026. [Online]. Available: <https://arxiv.org/pdf/2006.10511>
- [110] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Med. Image Anal.*, vol. 87, p. 102792, Jul. 2023, doi: 10.1016/J.MEDIA.2023.102792.
- [111] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 15979–15988, Nov. 2021, doi: 10.1109/CVPR52688.2022.01553.

- [112] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, Nov. 2019, doi: 10.1109/CVPR42600.2020.00975.
- [113] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models Genesis," *Med. Image Anal.*, vol. 67, p. 101840, Jan. 2021, doi: 10.1016/J.MEDIA.2020.101840.
- [114] C. Niu and G. Wang, "Unsupervised contrastive learning based transformer for lung nodule detection," *Phys. Med. Biol.*, vol. 67, no. 20, Oct. 2022, doi: 10.1088/1361-6560/AC92BA.
- [115] Z. Xie *et al.*, "SimMIM: A Simple Framework for Masked Image Modeling," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 9643–9653, Nov. 2021, doi: 10.1109/CVPR52688.2022.00943.
- [116] H. Mkindu, L. Wu, and Y. Zhao, "Lung nodule detection in chest CT images based on vision transformer network with Bayesian optimization," *Biomed. Signal Process. Control*, vol. 85, p. 104866, Aug. 2023, doi: 10.1016/J.BSPC.2023.104866.
- [117] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked Autoencoders As Spatiotemporal Learners," *Adv. Neural Inf. Process. Syst.*, vol. 35, May 2022, Accessed: Jan. 14, 2026. [Online]. Available: <https://arxiv.org/pdf/2205.09113>
- [118] L. Buess *et al.*, "Video-CT MAE: Self-supervised Video-CT Domain Adaptation for Vertebral Fracture Diagnosis," Dec. 23, 2024, *PMLR*. Accessed: Jun. 05, 2025. [Online]. Available: <https://proceedings.mlr.press/v250/buess24a.html>
- [119] Q. Ma *et al.*, "ShapeSplat: A Large-scale Dataset of Gaussian Splats and Their Self-Supervised Pretraining," Aug. 2024, Accessed: Jan. 14, 2026. [Online]. Available: <https://arxiv.org/pdf/2408.10906>
- [120] D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," *Med. Phys.*, vol. 16, no. 4, pp. 561–568, 1989, doi: 10.1118/1.596358.
- [121] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digit. Signal Process.*, vol. 132, p. 103812, Jan. 2023, doi: 10.1016/J.DSP.2022.103812.
- [122] A. McWilliams *et al.*, "Probability of Cancer in Pulmonary Nodules Detected on First Screening CT," *New England Journal of Medicine*, vol. 369, no. 10, pp. 910–919, Sep. 2013, doi: 10.1056/NEJMOA1214726;PAGEGROUP:STRING:PUBLICATION.
- [123] M. E. J. Callister *et al.*, "British Thoracic Society guidelines for the investigation and management of pulmonary nodules," *Thorax*, vol. 70 Suppl 2, pp. ii1–ii54, Aug. 2015, doi: 10.1136/THORAXJNL-2015-207168.
- [124] M. Oudkerk *et al.*, "European position statement on lung cancer screening," *Lancet Oncol.*, vol. 18, no. 12, pp. e754–e766, Dec. 2017, doi: 10.1016/S1470-2045(17)30861-6.
- [125] D. E. Wood *et al.*, "NCCN Guidelines® Insights: Lung Cancer Screening, Version 1.2025," *J. Natl. Compr. Canc. Netw.*, vol. 23, no. 1, Jan. 2025, doi: 10.6004/JNCCN.2025.0002.

- [126] D. Ardila *et al.*, “End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography,” *Nature Medicine* 2019 25:6, vol. 25, no. 6, pp. 954–961, May 2019, doi: 10.1038/s41591-019-0447-x.
- [127] S. B. Shuvo and T. B. Mamun, “AutoLungDx: A hybrid deep learning approach for early lung cancer diagnosis using 3D Res-U-Net, YOLOv5, and vision transformers,” *Inform. Med. Unlocked*, vol. 61, p. 101739, Mar. 2026, doi: 10.1016/J.IMU.2026.101739.
- [128] A. Halder, S. Chatterjee, and D. Dey, “Adaptive morphology aided 2-pathway convolutional neural network for lung nodule classification,” *Biomed. Signal Process. Control*, vol. 72, p. 103347, Feb. 2022, doi: 10.1016/J.BSPC.2021.103347.
- [129] H. Wang, H. Zhu, L. Ding, and K. Yang, “A diagnostic classification of lung nodules using multiple-scale residual network,” *Scientific Reports* 2023 13:1, vol. 13, no. 1, pp. 11322–, Jul. 2023, doi: 10.1038/s41598-023-38350-z.
- [130] E. Lv, X. Kang, P. Wen, J. Tian, and M. Zhang, “A novel benign and malignant classification model for lung nodules based on multi-scale interleaved fusion integrated network,” *Scientific Reports* 2024 14:1, vol. 14, no. 1, pp. 27506–, Nov. 2024, doi: 10.1038/s41598-024-79058-y.
- [131] M. Astaraki, Y. Zakko, I. Toma Dasu, Ö. Smedby, and C. Wang, “Benign-malignant pulmonary nodule classification in low-dose CT with convolutional features,” *Physica Medica*, vol. 83, pp. 146–153, Mar. 2021, doi: 10.1016/j.ejmp.2021.03.013.
- [132] H. H. Chang, C. Z. Wu, and A. H. Gallogly, “Pulmonary Nodule Classification Using a Multiview Residual Selective Kernel Network,” *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, p. 347, Feb. 2024, doi: 10.1007/S10278-023-00928-4.
- [133] L. Sun, M. Zhang, Y. Lu, W. Zhu, Y. Yi, and F. Yan, “Nodule-CLIP: Lung nodule classification based on multi-modal contrastive learning,” *Comput. Biol. Med.*, vol. 175, Jun. 2024, doi: 10.1016/j.combiomed.2024.108505.
- [134] M. E. Mayerhoefer *et al.*, “Introduction to Radiomics,” *J. Nucl. Med.*, vol. 61, no. 4, pp. 488–495, Apr. 2020, doi: 10.2967/JNUMED.118.222893.
- [135] J. J. M. van Griethuysen *et al.*, “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0339.
- [136] J. D. Shur *et al.*, “Radiomics in Oncology: A Practical Guide,” <https://doi.org/10.1148/rg.2021210037>, vol. 41, no. 6, pp. 1717–1732, Oct. 2021, doi: 10.1148/RG.2021210037.
- [137] J. Santinha *et al.*, “ESR Essentials: radiomics—practice recommendations by the European Society of Medical Imaging Informatics,” *European Radiology* 2025 35:3, vol. 35, no. 3, pp. 1122–1132, Oct. 2024, doi: 10.1007/s00330-024-11093-9.

- [138] M. Selvam, A. Chandrasekharan, A. Sadanandan, V. K. Anand, A. Murali, and G. Krishnamurthi, "Radiomics as a non-invasive adjunct to Chest CT in distinguishing benign and malignant lung nodules," *Scientific Reports* 2023 13:1, vol. 13, no. 1, pp. 19062–, Nov. 2023, doi: 10.1038/s41598-023-46391-7.
- [139] F. Song *et al.*, "A Novel Feature Engineering Method Based on Latent Representation Learning for Radiomics: Application in NSCLC Subtype Classification," *IEEE J. Biomed. Health Inform.*, vol. PP, no. 1, pp. 31–41, Jan. 2023, doi: 10.1109/JBHI.2023.3290006.
- [140] B. Dunn, M. Pierobon, and Q. Wei, "Automated Classification of Lung Cancer Subtypes Using Deep Learning and CT-Scan Based Radiomic Analysis," *Bioengineering (Basel)*, vol. 10, no. 6, Jun. 2023, doi: 10.3390/BIOENGINEERING10060690.
- [141] E. D'Arnese, G. W. Di Donato, E. Del Sozzo, M. Sollini, D. Sciuto, and M. D. Santambrogio, "On the Automation of Radiomics-Based Identification and Characterization of NSCLC," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 6, pp. 2670–2679, Jun. 2022, doi: 10.1109/JBHI.2022.3156984.
- [142] Y. Li, J. Deng, X. Ma, W. Li, and Z. Wang, "Diagnostic accuracy of CT and PET/CT radiomics in predicting lymph node metastasis in non-small cell lung cancer," *Eur. Radiol.*, vol. 35, no. 4, pp. 1966–1979, Apr. 2025, doi: 10.1007/S00330-024-11036-4.
- [143] S. S. Alahmari, D. Cherezov, D. B. Goldgof, L. O. Hall, R. J. Gillies, and M. B. Schabath, "Delta Radiomics Improves Pulmonary Nodule Malignancy Prediction in Lung Cancer Screening," *IEEE Access*, vol. 6, pp. 77796–77806, 2018, doi: 10.1109/ACCESS.2018.2884126.
- [144] L. Rundo *et al.*, "A low-dose CT-based radiomic model to improve characterization and screening recall intervals of indeterminate prevalent pulmonary nodules," *Diagnostics*, vol. 11, no. 9, p. 1610, Sep. 2021, doi: 10.3390/diagnostics11091610.
- [145] B. Hunter *et al.*, "Radiomics-based decision support tool assists radiologists in small lung nodule classification and improves lung cancer early diagnosis," *Br. J. Cancer*, vol. 129, no. 12, pp. 1949–1955, Dec. 2023, doi: 10.1038/S41416-023-02480-Y.
- [146] J. L. Causey *et al.*, "Highly accurate model for prediction of lung nodule malignancy with CT scans," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/S41598-018-27569-W.
- [147] W. Choi *et al.*, "Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer," *Med. Phys.*, vol. 45, no. 4, pp. 1537–1549, Apr. 2018, doi: 10.1002/MP.12820.
- [148] N. Garau *et al.*, "External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis," *Med. Phys.*, vol. 47, no. 9, pp. 4125–4136, Sep. 2020, doi: 10.1002/MP.14308.
- [149] J. Liu *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," *Front. Oncol.*, vol. 10, Feb. 2021, doi: 10.3389/FONC.2020.634298.

- [150] L. Mao *et al.*, “Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose CT screening,” *Quant. Imaging Med. Surg.*, vol. 9, no. 2, pp. 263–272, 2019, doi: 10.21037/QIMS.2019.02.02.
- [151] T. Peikert *et al.*, “Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial,” *PLoS One*, vol. 13, no. 5, May 2018, doi: 10.1371/journal.pone.0196910.
- [152] S. Tomassini, N. Falcionelli, P. Sernani, L. Burattini, and A. F. Dragoni, “Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey,” *Comput. Biol. Med.*, vol. 146, p. 105691, Jul. 2022, doi: 10.1016/J.COMPBIOMED.2022.105691.
- [153] V. Hondru, F. A. Croitoru, S. Minaee, R. T. Ionescu, and N. Sebe, “Masked Image Modeling: A Survey,” *Int. J. Comput. Vis.*, vol. 133, no. 10, pp. 7154–7200, Aug. 2024, doi: 10.1007/s11263-025-02524-1.
- [154] M. Baumgartner, P. F. Jaeger, F. Isensee, and K. H. Maier-Hein, “nnDetection: A Self-configuring Method for Medical Object Detection,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12905 LNCS, pp. 530–539, Jun. 2021, doi: 10.1007/978-3-030-87240-3_51.
- [155] X. Wu, H. Zhang, J. Sun, S. Wang, and Y. Zhang, “YOLO-MSRF for lung nodule detection,” *Biomed. Signal Process. Control*, vol. 94, p. 106318, Aug. 2024, doi: 10.1016/J.BSPC.2024.106318.
- [156] Y. Su, D. Li, and X. Chen, “Lung Nodule Detection based on Faster R-CNN Framework,” *Comput. Methods Programs Biomed.*, vol. 200, Mar. 2021, doi: 10.1016/j.cmpb.2020.105866.
- [157] A. Halder, S. Chatterjee, and D. Dey, “Adaptive morphology aided 2-pathway convolutional neural network for lung nodule classification,” *Biomed. Signal Process. Control*, vol. 72, p. 103347, Feb. 2022, doi: 10.1016/J.BSPC.2021.103347.

List of Figures

Figure 1 - Examples of SDCT (a) and LDCT (b) subjects from the LIDC-IDRI dataset and examples of the corresponding nodule annotations. LDCT images are characterized by increased quantum noise, streak artifacts, reduced contrast, and blurring. Adapted from [25].	10
Figure 2 – Diagram of CT imaging. Adapted from [26].	10
Figure 3 – GAN architecture consists of a generator that aims to create realistic images from the input data and a discriminator that attempts to distinguish real images from generated fake images. Adapted from [55].	14
Figure 4 – Anatomical structures of the lung, including the trachea, bronchi, alveoli, and interstitium, highlighting the distinct regions that could be affected by different respiratory pathologies. Image from https://www.froedtert.com/interstitial-lung-disease (accessed on 22/01/2026).	23
Figure 5 – Example of CT images showing (a) multiple solid nodules (arrow); (b) a part solid nodule (arrow), ground glass. Images from [94].	24
Figure 6 – Adaptation of the Video-MAE framework for volumetric medical imaging. The model processes a 3D input volume (e.g., a lung CT scan) by applying a high masking ratio (90%), leaving only a small portion of visible patches. These patches are processed by an encoder to generate latent representations, which a lightweight decoder then uses to reconstruct the original anatomical volume. Adapted from [25], [112].	30
Figure 7 – Radiomics pipeline consisting of five main steps: image acquisition, ROI segmentation, feature extraction, feature analysis and model development.	35

List of Tables

Table 1 – Comparison of SDCT and LDCT scanning protocols from selected studies in the literature.	8
Table 2 – LDCT protocols employed in lung cancer screening trials.	9
Table 3 - Summary of deep learning-based lung nodule detection methods.....	27
Table 4 - Summary of recent studies applying radiomics for nodule malignancy prediction in lung cancer screening.	39