Categorisation of CT Reconstruction Kernels

Using Image Features Directly Extracted from Patient Scans

by

Toke Camps

to obtain the degree of Master of Science
in Biomedical Engineering, Medical Physics
at the Delft University of Technology,
to be defended publicly on Thursday November 2, 2023 at 14:30.

Student number:	4544145	
Project duration:	Feb 1, 2023 – November 2, 2023	
Thesis committee:	M. Goorden,	TU Delft, supervisor
	V. Zuanazzi de Abreu	Aidence, supervisor
	dr. ir. D. Lathouwers	TU Delft, thesis committee

TU Delft

Abstract

Introduction CT is a versatile medical imaging method to diagnose and monitor patient diseases. However, varying patient characteristics and scan settings create challenges in maintaining consistent image quality, complicating image comparisons, especially across different sources. The reconstruction kernel in CT image reconstruction is a key parameter in the reconstruction process. It affects image characteristics, such as sharpness, contrast, and noise. There is an urgent need for a method that effectively compares and categorises reconstruction kernels from different vendors using real patient scans. Therefore, this thesis focuses on extracting features from real patient images to facilitate kernel comparisons within and across manufacturers.

Objective This research aims to create a machine learning (ML) method that categorises reconstruction kernels from various vendors into groups based on their sharpness. This categorisation will rely on image features extracted directly from real patient scans with diverse scan parameters.

Methods Two distinct methods were explored to achieve the objective, each utilising different image features and applied to a selected subset of the CT datasets from the National Lung Screening Trial (NLST) and the Lung Image Database Consortium image collection (LIDC-IDRC). The first method focused on noise features, specifically the standard deviation (SD) of the most homogeneous region of interest (ROI) to measure CT scan noise magnitude and the central frequency (CF) derived from the noise power spectrum (NPS) to represent scan noise texture. These noise features were used as input for a linear support vector machine (SVC), creating the *SVC_noise* model. Additionally, an approach that incorporated radiomic features was explored. These radiomic features were extracted from 30-pixel-sized ROIs selected from the ten most homogeneous patches. The radiomic feature sets were then used to train a random forest classifier (RFC), creating the *RFC_radiomics* model. The models were evaluated using accuracy and Receiver Operating Characteristic Area Under the Curve (ROC AUC) scores. McNemars test was employed to determine if one model significantly outperformed the other. Evaluating the categorisation results presented a significant challenge due to the lack of a ground truth. Consequently, a subset of the smoothest and sharpest kernels from each manufacturer was selected to train, validate, and test the models. Subsequently, the models were applied to the remaining kernels, and ground truth was established for each kernel by identifying the predominant class within each one.

Results Both models demonstrated strong performance when applied to 270 cases featuring 37 distinct reconstruction kernels. The *SVC_noise* model achieved an impressive ROC AUC score of 0.97 and misclassified eight of the 270 cases based on its smooth and sharp categorisation definition. The *RFC_radiomics* model achieved a slightly lower ROC AUC score of 0.96, with ten misclassifications out of the 270 cases. McNemars test indicated that the difference in performance between the two models was not statistically significant. Moreover, the ground truth approach, applied manually, resulted in only one inconsistent kernel between the two models; specifically, the determination of the ground truth of kernel $B50s$ differed.

Conclusion In summary, the *SVC_noise* and *RFC_radiomics* models displayed promising performances, with neither significantly surpassing the other. Both models exhibited the capacity to effectively identify sharpness-related patterns within the two classes while disregarding the noise caused by variations in scan parameters and patient characteristics in real patient data. This capability offers valuable insights that can bridge the divide between research and clinical applications. However, it is important to note that the findings from this research are preliminary, and caution should be exercised when applying these results to broader contexts, including newer reconstruction kernels and techniques.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

Abbreviation	Definition	Page
AI	Artificial Intelligence	2
CAD	Computer-Aided Diagnosis	2
CF	Central Frequency	16
CT	Computed Tomography	1
DL	Deep Learning	4
DLR	Deep Learning Reconstruction	9
FBP	Filtered Back-Projection	2
FPR	False Positive Rate	17
FT	Fourier Transform	8
GLCM	Gray-Level Co-occurrence Matrices	16
GLDM	Gray Level Dependence Matrix	16
GLRLM	Gray Level Run Length Matrix	16
GLSZM	Gray Level Size Zone Matrix	16
HMPV	Hierarchical Multi-Patch Voting	24
HL, LL, LH, HH	High-Low, Low-Low, Low-High, High-High	17
HU	Houndsfield Unit	19
IR	Iterative Reconstruction	2
LIDC-IDRC	Lung Image Database Consortium image collection	19
ML	Machine Learning	4
MPSLV	Multi-Patch Slice-Level Voting	24
NGTDM	Neighbouring Gray Tone Difference Matrix	16
NLST	National Lung Screening Trial	19
NPS	Noise Power Spectrum	2
RFC	Random Forest Classifier	13
RFC_radiomics	RFC model trained with radiomic features extracted from homogeneous patches	22
RFC_random	RFC model trained with radiomic features extracted from random patches	36
ROC AUC	Receiver Operating Characteristic Area Under the Curve	17
ROI	Region of Interest	15
SD	Standard Deviation	15
SLV	Slice-Level Voting	24
SVC	Support Vector Classifier	13
SVC_noise	linear SVC model that is trained using noise features	22
TPR	True Positive Rate	17

1 Introduction

In medical imaging, Computed Tomography (CT) is employed in a wide range of applications [1]. From facilitating precise patient diagnoses [2] to monitoring the progression of diseases [3, 4], CT is a cornerstone of modern medicine. However, this versatility presents a significant challenge: each specific application, in combination with a patient's characteristics, such as weight and age, necessitates a tailored approach regarding scan parameters. A tailored approach means that acquisition and reconstruction settings are adjusted to create CT images that yield sufficient image quality for gathering the necessary diagnostic information while carefully managing the radiation dose [5]. The wide diversity of scan parameters and scanners employed across different clinical contexts can lead to significant differences in the image quality [5, 6]. This diversity complicates comparing CT images, for example, when attempting to analyse data acquired from different scanners, time points, operators and institutions [7].

Most CT scanners from different vendors offer uniform settings for the majority of scan parameters. For instance, the slice thickness, which determines the thickness of each cross-sectional image, can be adjusted to a specific measurement in millimetres. However, each manufacturer offers its proprietary set of reconstruction algorithms and kernels regarding the reconstruction parameters, including its unique naming system.

Both reconstruction algorithms and kernels are essential components of the image reconstruction process, working together to transform raw X-ray data into human interpretable cross-sectional images [8]. A reconstruction algorithm is a mathematical technique to convert the raw X-ray projection data acquired during a CT scan into detailed cross-sectional images of the scanned area. This process involves complex calculations to determine the attenuation of X-rays as they pass through various tissues within the body, where a reconstruction kernel is a filter applied during the reconstruction process [9]. It influences the appearance and characteristics of the final image. Different kernels emphasise specific features or qualities in the image, such as sharpness, contrast, or noise [8, 10]. The choice of kernel depends on the diagnostic goals, the type of tissues being imaged and the user's preference.

This thesis focuses on extracting features directly from reconstructed images of actual patients, facilitating the comparison of various kernels from the same and different vendors.

1.1. Problem Statement

The lack of uniformity in the CT reconstruction process carries significant implications. In medical research, the diversity in data collection practices across hospitals compromises the statistical power of studies and undermines the reliability of their conclusions. The diversity in CT scan acquisition procedures complicates the ability to generalise findings, potentially constraining results to specific hospital environments, particular equipment manufacturers, or the unique preferences of individual healthcare practitioners.

Furthermore, selecting a reconstruction kernel substantially influences the characteristics of CT scan images, impacting critical factors like image sharpness and noise. This variability introduces uncertainty into the interpretation of CT scans, whether performed by human radiologists, algorithmic systems, or artificial intelligence (AI) models. AI models trained on specific sets of reconstruction kernels may demonstrate suboptimal performance when presented with scans reconstructed using different kernels, as the latter may alter the tissue characteristics crucial for generating accurate diagnostic outputs. Recent research confirms the lack of reproducibility in radiomic features due to variance in CT reconstruction parameters [11–14], which are pivotal in providing essential insights into tumour characteristics and are extensively utilised in medical image analysis, including cancer diagnosis, treatment prediction, and patient outcomes' assessment [15].

In addition, a study by Blazis et al. [7] revealed that a commercially available Computer-Aided Diagnosis (CAD) exhibited distinct performance disparities when analysing CT images reconstructed with iterative reconstruction (IR) compared to filtered back projection (FBP) kernels. This CAD system was developed by Aidence B.V., an Amsterdam-based company specialising in AI-powered clinical applications for lung cancer. Aidence and similar companies operating in this space would greatly benefit from developing a classification model capable of categorising CT scans based solely on objective image-based features and scan acquisition parameters rather than relying on manufacturer-specific information. Such a model has the potential to facilitate more accurate and reliable performance evaluations for their products; this, in turn, could lead to valuable insights for improving scan stratification, thereby creating more balanced datasets for enhanced AI model training. Additionally, this model makes it possible to more reliably select the most optimal CT scan series for a given diagnostic task, ultimately resulting in more accurate and clinically relevant outcomes.

Hence, a compelling need exists for a method to effectively compare and categorise reconstruction kernels across vendors based on real patient scans, including a wide range of scan parameters. This leads to the following problem statement:

Problem statement

Lack of standardisation in CT scan reconstruction kernels across vendors creates various problems, leading to data inconsistencies in research and clinical settings. To overcome this challenge, a method is needed to categorise these kernels based on scan characteristics extracted from actual patient scans, enhancing research, model performance, and healthcare education.

This research underlines the need to develop a systematic framework for categorising reconstruction kernels, mainly focusing on their sharpness and noise attributes. By doing so, it aims to enhance the consistency and reliability of CT image analysis, ultimately advancing the field of medical imaging.

1.2. Related Works

Two studies have attempted to identify the kernels across vendors that produce similar images based on characteristics extracted from a phantom [16, 17]. Solomon et al. [17] used the noise power spectrum (NPS), estimated from the uniform section of the phantom, to quantitatively compare noise texture across a wide selection of reconstruction kernels. On the other hand, Mackin et al. [16] aimed to identify reconstruction kernels that produced the most similar radiomics feature values for the materials in a specially designed radiomics phantom to enable more effective comparison of images produced using scanners from different manufacturers. While the previous research endeavours by Solomon et al. [17] and Mackin et al. [16] have made valuable contributions to the understanding of CT reconstruction kernels, it is important to recognise several shared limitations in these studies that necessitate the development of a novel approach based on scan characteristics extracted from actual patients.

Both Solomon et al. [17] and Mackin et al. [16] conducted their investigations using phantom scans, which inherently restricts their applicability to real-world clinical scenarios. Though valuable in controlled settings, phantoms cannot fully encapsulate the intricacies and variabilities in actual patient scans. For instance, they lack the physiological nuances and pathologies encountered in clinical practice, rendering the findings less directly transferable to the clinical setting.

A second shared limitation is that both studies are constrained to evaluating a single set of acquisition settings per manufacturer. This limitation does not align with the practical reality of clinical operations, where various acquisition parameters may be employed based on specific diagnostic needs. Consequently, the findings derived from these studies may not adequately represent the full spectrum of clinical scenarios, limiting their practical utility. Furthermore, while maintaining consistent acquisition settings for evaluating reconstruction kernels is methodologically systematic, the unmanageable magnitude of potential configurations, due to their exponential nature, makes the execution of an exhaustive search impractical.

Lastly, both studies primarily employ methods that compare individual kernels to one another. For instance, Mackin et al. [16] define a ’standard’ kernel for comparison. However, this approach cannot systematically compare all kernels relative to each other and does not provide a systematic method to categorise kernels into groups. Instead, it primarily reveals the extent of differentiation between kernels based on the feature values they extract, which, while informative, leaves an unaddressed need for a more comprehensive classification system for kernels.

2

Research Design

As highlighted in the introduction, the absence of standardized CT reconstruction kernels across vendors requires attention, preferably through a categorisation approach based on scan characteristics in the clinical environment. While the categorisation of reconstruction kernels is still in its early stages, the field of standardization of CT images is in active development, making it an intriguing subject for exploration within the literature review. The literature review aims to answer the following question:

Research question - Literature Review

What are the current state-of-the-art deep learning strategies for standardization across CT scans that vary due to reconstruction techniques?

This literature review offers a comprehensive overview of state-of-the-art deep learning (DL) strategies aimed at standardizing diverse CT scans with varying reconstruction techniques, detailed in Appendix A. A comparative analysis is conducted among these strategies, considering the standardization approach, DL architectures, and performance evaluation methods. Valuable insights are gained from this literature review that contribute to the development of this present thesis.

Foremost, a primary observation is the challenge of obtaining sufficiently large datasets for DL-based standardization techniques. Existing datasets often constrain themselves to specific scanner types or a restricted number of reconstruction kernels. Furthermore, many techniques involve training with pairs of images from the same source but reconstructed using different methods, raising questions about their generalizability to diverse patient scans and scanners. Consequently, the persisting issue of incomparability in reconstruction kernels across various vendors remains unresolved by these strategies.

Another noteworthy insight is using radiomic features as evaluation metrics for assessing standardization methods, as identified in several studies. This underscores the potential of radiomic features in characterizing reconstruction kernels based on actual patient scans.

Given these insights, this thesis aims to develop a machine learning (ML) method that categorises reconstruction kernels from different vendors into two groups based on their sharpness. Reconstruction kernels are commonly described by their sharpness. 'Sharp' kernels yield images with the highest achievable spatial resolution but come at the cost of increased pixel noise and the presence of streak artefacts. Conversely, 'smooth' kernels produce images with lower spatial resolution, but, on the other hand, they mitigate noise and artefacts, thereby enhancing the capability to detect low-contrast objects against the background [18]. The aimed method enables reconstruction kernels to be comparable with each other, not only within one vendor but also across vendors. Different methods and scan characterises are analysed to accomplish the main research objective following:

Main Research Objective

The development of an ML-based method for categorizing reconstruction kernels from different vendors into groups based on their sharpness. This categorisation will be accomplished by utilizing image features directly extracted from patient scans with varying scan parameters.

To accomplish this objective, the thesis first focuses on the question of whether it is possible to extract image features directly from patient scans that allow for kernel categorisation, forming research question 1:

Research Question 1

Can scan characteristics directly extracted from real patient scans facilitate the categorisation of reconstruction kernels?

To answer this question, several sub-questions have been developed that are listed below. First, image features that can effectively enable kernel categorisation are identified. From the related works, it became clear that two different sets of image features have the potential to effectively enable kernel categorisation: quantitative noise measures and radiomics features. Therefore, the thesis is divided into two distinct research components where the first delves into the potential of quantitative noise features, drawing inspiration from the study of Solomon et al. [17]. In parallel, the second component explores the practicability of evaluating radiomic features, inspired by the research conducted by Mackin et al. [16].

It is worth noting that these two studies have not been directly compared before, and their scope has been confined to phantom studies. As a result, the most suitable method for categorizing reconstruction kernels derived from actual patient scans acquired with a diverse set of scan parameters is an active research field. As a result, using both approaches, the thesis aims to answer the following sub-questions:

Research Sub-Questions 1

1.1. Which image features can effectively enable kernel categorisation?

1.2. Which machine learning (ML) models are suitable for utilizing these image features for categorisation?

1.3. What is the performance of the ML models using the identified image features in the context of kernel categorisation?

A significant challenge arises when it comes to evaluating the categorisation results. The absence of a ground truth poses a fundamental issue. The sole information accessible regarding the sharpness of the reconstruction kernels originates from the vendors' descriptions. Consequently, this thesis attempts to address Research Question 2 by seeking a methodology to assess the performance of the developed categorisation methods.

Research Question 2

How can the kernel categorisation be effectively evaluated when ground truth data is unavailable?

3
Theory

3.1. Computed Tomography

"Computed Tomography", abbreviated as CT, describes a computerized X-ray imaging technique. In this procedure, an X-ray beam is directed at a patient and rapidly rotated around the body. This rotation generates signals that are then processed by the machine's computer system to create cross-sectional images, often referred to as "slices". These slices are known as tomographic images and provide clinicians with more intricate and detailed information than traditional X-rays. Once a series of successive slices is gathered and processed by the machine's computer, they can be digitally assembled to construct a three-dimensional (3D) patient representation. This 3D image facilitates the easier identification of anatomical structures and the detection of potential tumours or abnormalities [8, 19].

CT scanners use a rotating X-ray source within a donut-shaped gantry, unlike traditional X-rays with stationary tubes. The patient lies on a moving bed as the X-ray tube orbits around them, emitting a polychromatic X-ray beam, thus consisting of a range of energies. The tube current (mA) and voltage (kV) control the X-ray beam. Tube current dictates the X-ray intensity, with higher current producing more photons, while tube voltage determines X-ray energy, influencing penetration. Digital detectors, instead of the X-ray source, capture exiting X-rays and transmit data to a computer [19, 20]. Figure 3.1 displays a schematic visualization of a CT scanner set-up.

Between the source and the detectors, collimators are present for multiple purposes. They protect the patient by confining the X-ray beam to the specific anatomical area of interest, shaping it, and reducing the impact of scattered radiation. Additionally, collimators assist in determining the slice thickness [20]. Scattered radiation arises when the X-ray beam interacts with matter and changes direction. This phenomenon can diminish image quality in medical imaging since the scattered X-rays lack valuable diagnostic information and can introduce undesired noise into the X-ray image [19].

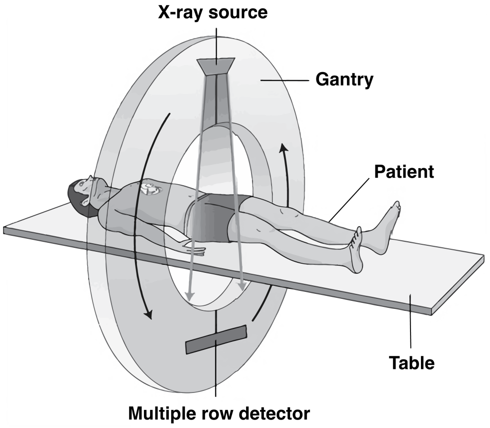


Figure 3.1: Schematic overview of a CT scanner set-up including the X-ray source, gantry, patient bed and multiple row detector.

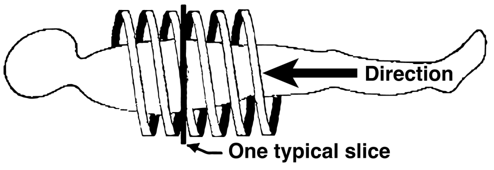


Figure 3.2: The motorized bed incrementally advances into the gantry while the data are acquired continuously during 360-degree scans, creating a helical path.

With each complete rotation of the X-ray source, the CT computer applies advanced mathematical techniques to construct a two-dimensional image slice of the patient, as elaborated in section 3.1.2. The motorized bed incrementally advances into the gantry while the data are acquired continuously during 360-degree scans until the desired number of slices has been acquired; this process is shown in Figure 3.2.

3.1.1. Image Quality

Image quality in CT can be defined by how accurately the CT image reproduces the 3D attenuation distribution of the x-ray beam through the patient, which is influenced by the following factors: noise, spatial resolution, contrast and artefacts [21]. This thesis will mainly focus on the noise and spatial resolution factors. Therefore, these two are explained in detail below.

Noise

In CT images, noise arises from random fluctuations within the image, which is associated with the number of X-rays contributing to each detector measurement. Factors affecting the number of detected X-rays and thus the noise in the image are, for example, the scan (rotation) time, electronic interference, radiation exposure and the slice thickness [21]. Noise results in unwanted variations that obscure image details and compromise overall image quality. This noise manifests as a grainy or speckled pattern within the image, making it challenging to differentiate between various tissues or structures [22].

Reducing noise is essential in CT imaging, as it directly impacts the radiologist's ability to provide accurate diagnoses. Mitigating noise may entail adjusting scan parameters, carefully increasing the radiation dose within safe limits, or applying image processing techniques to filter out noise while preserving essential image details selectively [23, 24].

Spatial Resolution

Spatial resolution refers to an imaging system's capacity to distinguish objects in the spatial dimensions of an image [21]. It quantifies the system's ability to identify two objects as they become smaller and closer together. The better the spatial resolution, the closer these objects can be without merging into one another.

It is worth noting that in CT imaging, spatial resolution and sharpness are often used interchangeably, especially when discussing reconstruction kernels. In this thesis, the term "sharpness" will specifically describe this image quality aspect. Further elaboration on reconstruction kernels and their impact on sharpness can be found in Section 3.1.4.

3.1.2. Scan Reconstruction

The X-rays transmitted through the patient interact with the body tissues they encounter, which causes an exponential reduction in distance travelled in beam intensity based on tissue density and composition. As the X-ray photons exit the patient, they are absorbed by a CT detector and converted into an electronic signal. The attenuation of the X-ray beam as it passes through a material along a line in direction s can be calculated using the line integral in Equation 3.1. Using the measured intensity I and the incoming initial intensity I_0, the linear attenuation coefficient $\mu(x,y)$ of tissue at the position(x,y) in a 2D plane can be computed.

$$I = I_0 e^{-\int \mu(x(s),y(s))ds} \tag{3.1}$$

A set of line integrals along all the ray paths in the X-ray beam creates a projection. Figure 3.3 visualized a schematic visualization of the incoming X-rays, with a line in direction s and the final projection along all the ray paths. Finally, the complete collection of line integrals that traverse the patient's body for every possible ray trajectory within the X-ray beam, encompassing all gantry angles, is called the Radon transform [8].

The basic idea behind CT scan reconstruction is to use mathematical algorithms to estimate the linear attenuation coefficient in each voxel of the image volume, which is strongly related to the tissue density. This is done by solving an inverse problem (back-projection), where the goal is to find the tissue density distribution that best explains the measured X-ray projections.

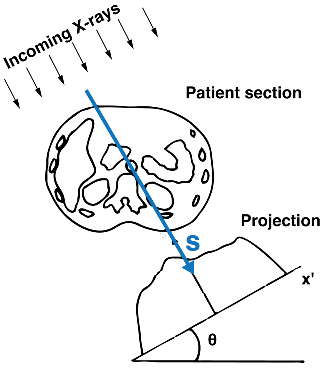


Figure 3.3: Illustration of the incoming X-rays, with a line in direction s and the final projection of the patient section along all the ray paths.

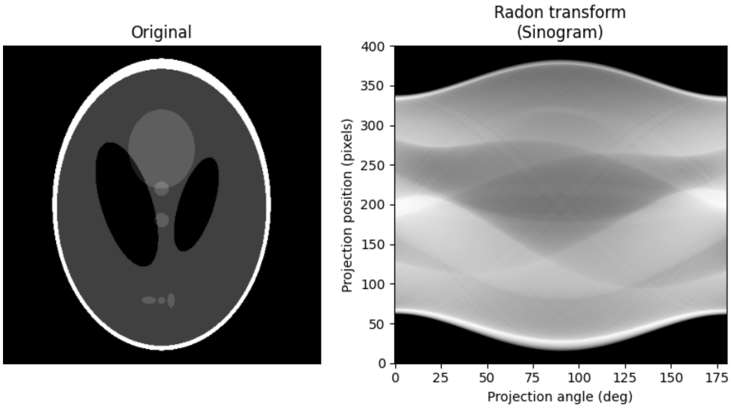


Figure 3.4: *Left*: the original image is the Shepp-Logan phantom, a standard test image. *Right*: the sinogram of the original image.

Radon Transform

The Radon transform of a CT slice is a graphical representation of the intensity losses measured by the CT scanner, also called the sinogram, where the vertical axis represents the distance various beams are from the origin, and the horizontal axis represents the angle at which the slice is measured [8, 25]. Therefore, a single point in the sinogram represents the measured change in intensity for a given distance and angle, equal to the line integral over μ in Equation 3.1. Figure 3.4 shows an example of a sinogram.

Central Slice Theorem

The central slice theorem is one of the fundamental concepts in CT image reconstruction, which enables the transformation of the complex collection of line integrals into a spatial representation of the object's internal structure [26]. This theorem, also known as the Fourier slice theorem, states that the 2D Fourier transform (FT) of an object is equivalent to the 1D FT of the object's projection passing through its centre and perpendicular to the plane of the 2D FT [27]. This means that the 1D Fourier transform of a projection is identical to a 1D profile through the origin of the 2D Fourier transform of the irradiated object (x,y). This concept is visualized in Figure 3.5.

By transforming all projections of the object into the 1D Fourier transform and interpolating them into a 2D Fourier space, the complete 2D FT of the object can be reconstructed. The original object is reconstructed from the full 2D FT using the inverse FT.

The theorem has provided a mathematical foundation for reconstructing images from X-ray projections and has facilitated the development of various CT reconstruction algorithms. The subsequent section explains how CT scans are reconstructed in practice based on the abovementioned theory.

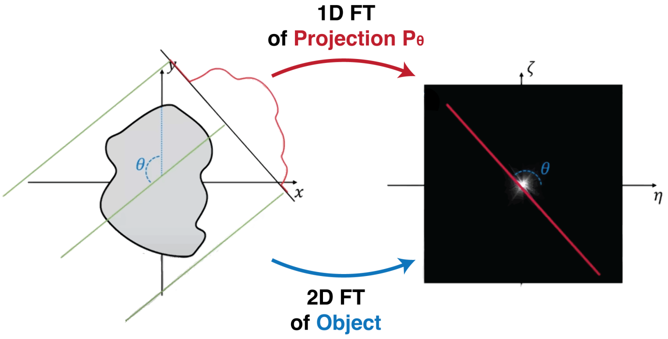


Figure 3.5: This illustration demonstrates the fundamental concept of the central slice theorem. In the left image, the process of generating the projection, denoted as P_θ, from a specific gantry angle θ is depicted in red. On the right, the object's 2D Fourier transform (FT) displayed in the left image is visualized. The red-marked slice in the right image corresponds to the 1D FT of the projection shown on the left. This 1D FT of the projection is identical to a 1D profile through the centre of the 2D FT of the object.

3.1.3. Reconstruction Techniques

Various strategies are used for CT image reconstruction, including FBP, IR, and deep learning reconstruction (DLR). For decades, FBP was the standard image reconstruction method because of its simplicity and computational efficiency [28, 29]. It was not until 2009 that the initial IR algorithms were introduced to the market, replacing the conventional FBP technique [28]. The current state-of-the-art method for CT image formation is image reconstruction based on DL, although currently, only three DLR algorithms are commercially available [30]. In the following subsections, each technique is briefly explained.

Filtered Back-Projection

Filtered back projection is a common reconstruction method that reconstructs a 3D image of an object from its projection data (sinograms) by taking the exact inverse. In short, the method evenly distributes the measured filtered signal over the projection line to compute 2D slices, which are combined to form a 3D volume representing the object [8, 31]. More specifically, for each gantry angle in a sinogram, the attenuation value is divided by the number of image pixels along the path of the projection from the X-ray source to the detector. The resulting average attenuation value is then allocated to those pixels. This process is carried out for every gantry angle. The back-projected data is then summed to form the final back-projected image; this process is displayed in Figure 3.6.

Before back-projection, the projection data is filtered to counteract blurring that occurs because of evenly spreading the attenuation value; this blurring is also visualized in Figure 3.6 [8, 32]. Applying an ideal ramp filter to modify the original projection through convolution in the spatial domain or multiplication in the Fourier domain filters out the low frequencies and passes the high frequencies, with a linear behaviour in between [9]. This enhances sharp boundaries between different anatomical structures while minimising blurring (low frequencies). However, this amplification also increases image noise, particularly at high spatial frequencies where noise is more prominent in the raw signal.

Iterative Reconstruction

The underlying principle of IR is to calculate image data that accurately corresponds to the acquired projection data using iterative algorithms [29, 32]. It formulates the problem as a constrained optimization task, seeking the unknown image data that best fits the measured projection data while satisfying the constraints. The optimization involves matching the reconstruction to the measured data and suppressing noise through regularization.

To reconstruct CT images with IR, a cycle of forward- and back-projection steps minimises a cost function, quantifying how well the reconstructed image matches the acquired projection data. The process is iterated until a predefined stopping criterion is met. Two adjustable parameters impact the IR outcome: the reconstruction kernel used in the back projection step, affecting image noise and sharpness, and the algorithm's strength, influencing noise reduction [33, 34]. Nevertheless, excessive IR strength, particularly at higher levels, can produce unsightly "blooming" artefacts that hinder the display of minute structures [35]. As a result, the impact of the IR algorithm's strength on image quality must be weighed carefully to strike a balance.

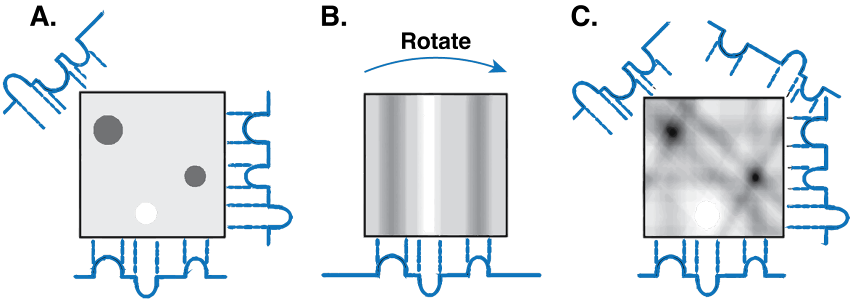


Figure 3.6: **A.** Back-projection reconstruction is applied to a simple phantom comprising three objects with different attenuation values. In addition, the projections at three different angles are shown. **B.** Attenuation values are spread out evenly along their ray path; this process is done for each angle. **C.** The final image results from the summation of the four angles of the phantom. Despite its efficiency, the back-projection method produces images that exhibit significant blurriness.

Iterative reconstruction approaches can be classified into statistical (hybrid) and model-based iterative algorithms based on how the imaging process is modelled [36, 37]. Medical manufacturers all have their own IR approaches that use different modelling and apply different cost functions.

Deep Learning Reconstruction

Deep-learning-based techniques for CT image reconstruction is an emerging approach that can potentially improve image quality further and thus reduce dose [38]. Unlike traditional CT reconstruction methods that use analytical models, DLR uses neural networks to either learn the mapping between the raw projection data (sinograms) and the corresponding high-quality CT images or to differentiate between signal and noise in low-quality images.

Training a neural network that learns the mapping demands a large dataset of CT projection data and corresponding high-quality CT images. Such a dataset originates from phantom images and patient scans conducted in a clinical setting. However, this amount of data is not always available [39]. Once this type of neural network is trained, it can take raw projection data as input and generate high-quality CT images as output [38, 39].

Conversely, a neural network that differentiates between noise and signal requires a sizable dataset containing low-quality and corresponding high-quality CT images. The low-quality CT images can be intentionally generated by introducing noise into the high-quality ones, facilitating the creation of this dataset. Once trained, this neural network can effectively enhance the signal and reduce noise in reconstructed CT images acquired with low-dose radiation [30].

3.1.4. Reconstruction Kernels

As mentioned before, in the reconstruction process of CT scans, a ramp filter is applied to eliminate blurriness caused by back-projection in both FBP and IR approaches [9]. This ramp filter can be paired with filters of varying intensities (kernels) to heighten the spatial resolution of the ultimate image, dependent on the specific application [8, 10]. This combination is called the reconstruction kernel, a parameter that can be adjusted to emphasize different tissue characteristics and influence image sharpness and noise.

Various kernels with distinct features are at clinicians' disposal in their day-to-day operations [32]. 'Smooth' kernels are designed to lower image noise and bolster the display of low-contrast details but can lead to a drop in image sharpness. Meanwhile, 'sharp' kernels aim to enhance the illustration of intricate elements in high-contrast structures, although they can increase image noise to a level that hinders the recognition and distinction of low-contrast structures [8, 10, 40]. In Figure 3.7, an example of a CT image pair is displayed, a CT scan that is reconstructed with a 'smooth' kernel, the standard kernel of GE medical systems (left scan), as well as with a 'sharp' kernel, the bone kernel of GE medical systems (right scan).

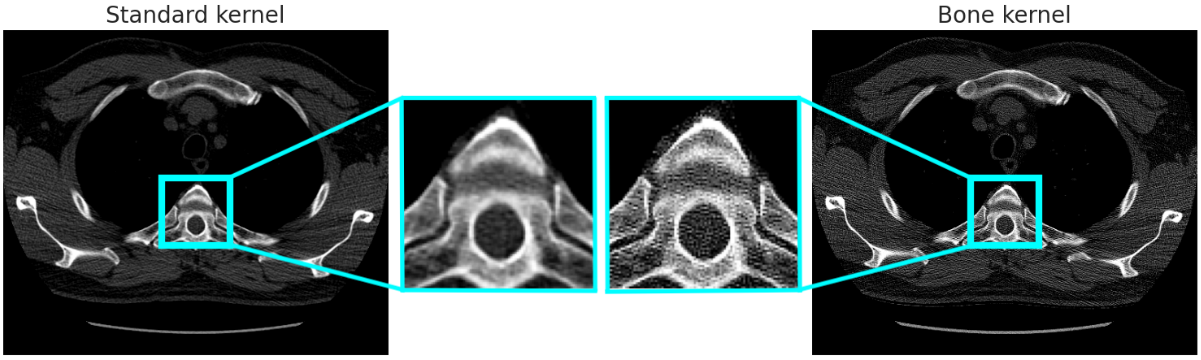


Figure 3.7: The left CT scan is reconstructed with the standard reconstruction kernel, whereas the right CT scan is reconstructed with the sharper kernel, called the bone reconstruction kernel. Both reconstruction kernels are developed at GE Medical Sytems. The standard, smoother kernel shows lower image noise and displays more low-contrast details, but, on the downside, it has a lower image sharpness. On the contrary, the scan reconstructed with the bone kernel, a sharper kernel, enables a better edge definition and shows more structural details, which are visible around the bones. However, it also shows an increased image noise.

The four largest manufacturers of CT scanners are Philips Healthcare, Siemens Healthineers, GE HealthCare and Toshiba Medical (currently known as Canon Medical Systems Corporation). The manufacturers are respectively referred to as Philips, Siemens, GE and Toshiba in this thesis. Each manufacturer has its own reconstruction techniques and reconstruction kernels they offer. In addition, each manufacturer also has its own designation system for the reconstruction kernels, some more complex than others. Explanations of these systems are hard to find, very limited and vague. Nevertheless, each designation system per manufacturer is explained in the following subsections, and the reconstruction kernels relevant to this thesis are described (if possible) based on their usage and sharpness. The complete overview of all kernels is added in the appendices.

GE Healthcare

GE Healthcare mentions nine different kernels in their manual for the CT Revolution [41]. The manufacturer provides insights into the suitable applications for each kernel and arranges the kernels in descending order, moving from higher spatial resolution to lower contrast detection capability. Table 3.1 describes and orders the kernels used in this thesis. The complete overview of all the available kernels for GE is shown in Table B.1 in Appendix B.

Table 3.1: Descriptions of the relevant GE reconstruction kernels. Extracted from the user manual of GE Healthcare Revolution CT scanner.

Kernel	Description
Soft	for tissues with similar densities, but not useful for un-enhanced scans
Std	for routine exams, e.g., chest, abdomens, and pelvis scans
Lung	for interstitial lung pathology
Bone	for high-resolution exams and sharp bone detail.

Philips Healthcare

The CT scanners of Philips Healthcare use an alphabetical system for the designation of their reconstruction kernels [42]. In general, the sharpness increases for an increasing letter in the alphabet, and different resolution strengths are available per kernel. This research only focuses on standard resolution. The relevant kernels are outlined in Table 3.2, and the full overview is given in Table B.2 in Appendix B.

Table 3.2: Descriptions of the relevant Philips reconstruction kernels. Extracted from the user manual of Philips Brilliance CT scanner.

Kernel	Description
A	Very smoothed, can be used to decrease noise significantly. Recommended for use when the patient is very large and the dose inadequate for the patient's size
B	Smoothed, but sharper and noisier than A. Recommended for CTA (for example, COW), routine abdomen, and pelvis.
C	Sharper, creates relatively low-noise images. Recommended for CTA (for example, COW), routine abdomen, and pelvis to get slightly higher sharpness than with Filter B
D	Sharp and edge-enhancing. Creates relatively high-noise images and raises the bone density

Siemens Healthineers

Siemens Healthineers offers various appropriate kernels for different applications [43]. They have three different types of kernels: "H "stands for Head, "B "stands for Body, "C "stands for Child Head and "S" stands for Special Application. The numbers define the image's sharpness: the higher the number, the sharper the image; the lower the number, the smoother the image. The last letter indicates the scanning mode: "s" stands for standard rotation time, and "f" stands for fast rotation time. In addition, the vendor offers special kernels that are indicated by a "+-sign" in the description. These kernels have an added fine-grained noise structure, which improves the low contrast detectability, usually a characteristic of a smoother kernel. The relevant kernels for this research are shown in Table 3.3, and the other kernels offered by Siemens are overviewed in Table B.3 in Appendix B.

Table 3.3: Descriptions of the relevant Siemens reconstruction kernels. Extracted from a Siemens Somatom Sensation manual.

Kernel series	Description
B20s/B20f	Smooth
B30s/B30f	Medium smooth
B31s/B31f	Medium smooth +
B35f	HeartView medium
B40s/B40f	Medium
B41s	Medium +
B45s/B45f	Medium
B50s/B50f	Medium sharp
B60s/B60f	Sharp
B70f	Very sharp
B70f	Ultra sharp

Toshiba Medical

Toshiba Medical, currently known as Canon Medical Systems Corporation, uses a naming system where each ten has a different application [44]. In addition, within a ten, a higher number indicates an increase in sharpness. Unfortunately, the manufacturer does not indicate how the tens compare to each other regarding sharpness. The kernels from FC01 and FC10 are the same reconstruction algorithm; the only difference is whether beam hardening correction processing is used.

This processing mitigates artefacts resulting from X-ray beam hardening, a phenomenon that arises as the beam passes through a patient due to its polychromatic nature. Lower-energy X-rays are more likely to be absorbed as X-rays pass through the body, leaving the higher-energy X-rays to dominate the beam. Consequently, the X-ray beam's average energy rises as it traverses the body, leading to beam hardening.

Table 3.4 describes the relevant kernels and the complete overview can be found in Table B.4 in Appendix B.

Table 3.4: Descriptions of the relevant Toshiba reconstruction kernels. Extracted from a Toshiba Aquilion16 manual.

Kernel series	Description
From FC01	For the abdomen, with beam hardening correction (BHC) processing
From FC10	For the abdomen
From FC30	For the inner ear and bones
From FC50	For the lung field
From FC82	For high resolution, for the lung field (high-resolution CT)

3.1.5. Scan Parameters

Besides the reconstruction kernels, many other scan parameters can be adjusted to control image quality. Each of these parameters has a different impact on the output of the CT scanner and influences the image quality, such as noise and sharpness [45]. Therefore, they can act as confounding factors when analysing reconstruction kernels using a dataset with a large variety of other scan parameters. The most important CT scan parameters are explained to understand each parameter's impact, including their impact on the sharpness and noise of the CT scan output.

Slice Thickness
determines the thickness of each cross-sectional image (slice) acquired during the scan, and thus the number of detected x-rays [21].

- *Sharpness Impact:* Thinner slice thickness improves spatial resolution in the axial direction and image sharpness by reducing volume averaging of adjacent structures. Conversely, thicker slices lead to reduced sharpness due to increased partial volume effects.
- *Noise impact:* Thinner slices tend to have higher noise levels because fewer X-ray photons contribute to each slice; thus, a lower signal is detected. Since fewer photons are involved, the statistical fluctuations in their detection become more pronounced, leading to increased image noise.

Tube Current (mA)
controls the amount of radiation emitted by the X-ray tube (beam intensity).

- *Sharpness Impact:* Higher tube current improves signal-to-noise ratio by increasing the number of X-ray photons detected, enhancing sharpness.
- *Noise Impact:* A lower tube current increases the image noise by decreasing the number of X-ray photons detected. The influence of this noise can be more noticeable in homogeneous materials compared to textured materials [46].

Tube Voltage (kVp)
determines the energy level of the X-rays and affects the contrast and penetration ability.

- *Sharpness Impact:* Lower kVp settings can enhance tissue contrast and sharpness by reducing beam hardening artefacts.
- *Noise Impact:* Lower kVp settings can increase image noise due to reduced penetration of X-rays through the body. Higher kVp settings reduce noise by increasing the number of detected photons.

3.2. Machine Learning

Machine learning is a subfield of Artificial Intelligence that focuses on developing algorithms and models that enable computers to learn from and make predictions or decisions based on data [47]. Machine learning systems use statistical techniques to identify patterns, relationships, and insights within datasets. These systems can then generalize from their findings to make predictions or decisions when presented with new, unseen data. Machine learning has a wide range of applications, from image and speech recognition to recommendation systems and autonomous vehicles, and it plays a crucial role in today's data-driven world.

Machine learning relies on different algorithms to solve data problems; in this research, a support vector classifier (SVC) and a random forest classifier (RFC) have been used, which are schematically displayed in Figure 3.8 and Figure 3.9, respectively. The concepts are explained in detail in the sections below.

3.2.1. Support Vector Classifier

A support vector classifier, also known as a support vector machine for binary classification, is a powerful supervised machine learning algorithm for separating data points into two distinct classes [48]. In this thesis, a linear SVC is applied to perform linear classification. Consequently, this section focuses only on the theory behind a linear SVC.

The fundamental concept behind a linear SVC is to find a decision boundary (hyperplane) that best separates data points belonging to two different classes. The key objective is to maximize the margin, which is the distance between the decision boundary and the nearest data points (support vectors) from each class. This margin maximization helps ensure that the classifier can make accurate predictions on the training data and new, unseen data. The hyperplane in a linear SVC is a flat, linear boundary that divides the feature space into two regions corresponding to the two classes. For a two-dimensional feature space, the hyperplane is a line. In higher-dimensional spaces, it becomes a hyperplane. The linear SVC identifies the optimal hyperplane based on its ability to maximize the margin [48, 49].

One important hyperparameter that can be optimised to achieve optimal results is the regularization parameter C. This parameter is the degree to which the model will accept misclassification in the dataset of each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of correctly classifying all the training points. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points [50].

3.2.2. Random Forest Classifier

The random forest classifier is a powerful ensemble learning algorithm for solving classification problems [51]. It is particularly popular due to its ability to handle complex, high-dimensional data and provide accurate predictions [52]. The algorithm is based on an ensemble of decision trees, combining multiple individual trees to make collective predictions. A decision tree is a flowchart-like structure where each internal decision node represents a feature, and each leaf node represents a class label. The tree is built by repeatedly partitioning the input data based on the selected features, aiming to create homogeneous subsets of data at each leaf node. In an RFC, multiple trees are constructed instead of using a single decision tree. Each tree is built using a random subset of the training data and a random subset of the features [51]. This randomness introduces diversity among the trees, making them less prone to overfitting and more capable of capturing different aspects of the data. Overfitting occurs when an ML model learns the training data too well, capturing noise and random fluctuations instead of the underlying patterns. This results in poor generalization to new, unseen data [53].

During the training phase of the random forest classifier, the individual trees are constructed independently. Each tree is built by repeatedly selecting a subset of data with replacement (bootstrapping) and randomly selecting a subset of features at each split. The splitting process identifies the feature and threshold that maximally separate the classes. Once all the trees are constructed, the RFC predicts the class label of a new

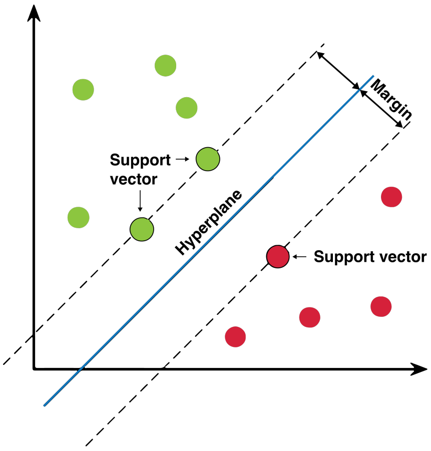


Figure 3.8: Illustration of a linear support vector classifier (SVC). This graphic shows the key elements of an SVC, including the support vectors, the margin, and the separating hyperplane, highlighting the principles of this classification algorithm.

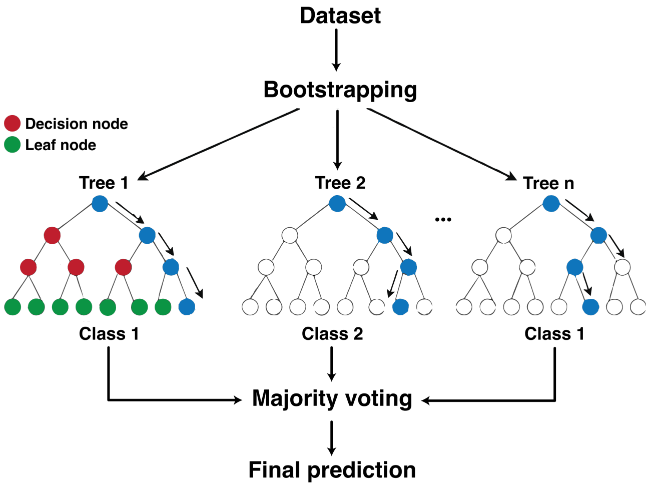


Figure 3.9: Illustration of a random forest classifier (RFC) showing the inner workings of an RFC. Using bootstrapping, the RFC generates diverse decision trees, each based on different subsets of the training data. These individual trees make their predictions for unseen data points, which are aggregated through majority voting to yield the final prediction. The diagram further illustrates decision nodes (highlighted in red), which mark points where decisions are made within the tree, and leaf nodes (highlighted in green), which represent the endpoints where final predictions are determined.

data point by aggregating the predictions of each tree [51]. This aggregation can be done by majority voting, where the class label that receives the most votes is assigned as the final prediction. Alternatively, probabilistic predictions can be obtained by taking the class probabilities from each tree and averaging them.

Unfortunately, RFC has a risk of overfitting, so hyperparameter tuning is crucial. By setting a maximum depth (max_depth), you can control the depth of tree growth. Similarly, by specifying the number of estimators ($n\_estimators$), you determine the ensemble size at each step [54]. Finding optimal values for these hyperparameters is determined empirically.

3.3. Noise Features

As mentioned in Section 3.1.4, reconstruction kernels influence the sharpness and noise of a CT image. Smoother kernels use low-pass filters to block high-frequency content, improving low-contrast resolution and reducing noise. Conversely, sharper kernels retain high-frequency details to enhance spatial resolution at the expense of increasing noise in the final CT image.

Image noise can be characterized by its magnitude and texture [55, 56]. Noise magnitude refers to the random pixel value fluctuations in a homogeneous region. In contrast, noise texture concerns relationships between neighbouring pixels that manifest as the grainy appearance in CT scans [55].

3.3.1. Noise Magnitude

In CT scans, noise magnitude refers to the extent of random fluctuations or variations in pixel values within homogeneous or uniform regions of the image. This noise is primarily caused by statistical variations in the number of detected X-ray photons [55]. In a clinical image, measurements of noise magnitude are commonly performed based on calculating the standard deviation (SD) within a region of interest (ROI) in the most uniform region. Anam et al. [56] claim that they have developed an automated method for quantifying noise in CT images, capable of distinguishing variations in noise magnitude stemming from input parameters like tube currents and image reconstruction kernels. This method estimates noise magnitude by pinpointing the minimum SD value within an SD map. The SD map is produced through a sliding window operation, where the SD value for each pixel location (x, y) is computed using the following equation:

$$\mathrm{SD}_0 = \sqrt{\frac{1}{n \times n} \sum \sum_{\mathrm{i}=1}^{\mathrm{n}\times\mathrm{n}} (I_{\mathrm{b,i}} - \bar{I}_b)^2} \tag{3.2}$$

With 'n' representing the dimensions of the sliding window $(n \times n)$, $I_{\mathrm{b,i}}$ denotes the pixel value at a certain location i within the window of the image b and $\bar{I}_b$ stands for the mean intensity value of all pixels within the window. Once the SD calculation is finalized for a single pixel, the window shifts to the next pixel to calculate the SD value similarly. This iterative procedure continues until SD values have been computed for all pixels, resulting in the generation of the SD map.

3.3.2. Noise Texture

The NPS characterises the noise texture, thus giving a better and more complete description of noise than the simple pixel's standard deviation [56, 57]. In a stationary system, the NPS gives a complete description of the noise by providing its amplitude over the entire frequency range of the image [58]. If the image noise is not stationary, the NPS is not a complete description, and the whole covariance matrix would be needed for a complete description. However, if applied with care, for example, working with small ROIs extracted from a restricted image region, the NPS can be applied to CT images [57]. To compute the NPS of a CT image, it is necessary to select homogeneous ROIs within the CT image. The 2D NPS can then be computed as:

$$NPS_{2D}(f_x, f_y) = \frac{\Delta_x \Delta_y}{L_x L_y} \frac{1}{N_{ROI}} \sum_{i=1}^{N_{ROI}} \left| FT_{2D} \left\{ ROI_i(x, y) - \overline{ROI_i} \right\} \right|^2 \tag{3.3}$$

Where Δ_x, Δ_y are the pixel sizes in the x and y dimension in millimetres, L_x, L_y are the ROI's lengths (in pixel) for both dimensions, N_{ROI} is the number of ROIs used in the average operation and $\overline{ROI_i}$ is the mean pixel value of the ith ROI.

Subsequently, the 2D NPS is averaged along a 1D radial frequency using the equation $f_r = \sqrt{f_x^2 + f_y^2}$, so the central frequency (CF) can be determined from the NPS curve using Equation 3.4. This value indicates the dominant frequency in the spectrum [55]. A smaller CF indicates that the centre of gravity is skewed toward lower frequencies. This implies a greater degree of blurred image texture or a more pronounced loss of higher-frequency components after noise reduction [55, 59].

$$CF = \frac{\int f \times NPS(f)df}{\int NPS(f)df} \tag{3.4}$$

3.4. Radiomic Features

Radiomic features are used in the field of radiomics, which involves extracting and analysing various quantitative features from medical images [60]. These features provide valuable information invisible to the human eye about the underlying tissue characteristics, such as shape, texture, intensity, and spatial relationship [61]. Some key areas where radiomic features are used include cancer diagnosis and prognosis, treatment response assessment, predictive modelling, and personalized medicine [62].

When the noise values influence noise values in each voxel in neighbouring voxels, it is called correlated noise. In the reconstruction process, the kernel contributes to creating a correlated noise texture [63]. Texture features might be sensitive to this correlated noise in a CT image because most of these features describe spatial relationships of voxel intensities within an ROI. For example, features based on grey-level co-occurrence matrices (GLCM) characterize an image's texture by counting the number of occurrences where pairs of voxels sharing identical grey levels within a specific spatial relationship occur within an ROI [63]. Therefore, radiomic features may provide useful information on the underlying noise texture related to the applied reconstruction kernel.

3.4.1. Feature Classes

Various feature types can be derived from clinical images, which can either be extracted directly from the ROIs or after applying different filters or transforms, such as the wavelet transform (explained in the next section). The features are normally categorised into the following classes:

First-order statistics features are derived from histograms and focus solely on the distribution of individual voxel values, disregarding any spatial correlations. They encompass statistical characteristics such as the mean, median, maximum, and minimum voxel intensities within an ROI, as well as measures of skewness (asymmetry), kurtosis (flatness), uniformity, and randomness (entropy) [64].

Second order statistics features consider the interrelationships between neighbouring pixels in an ROI and are the so-called texture features [65]. Using grey-level dependence matrices, the second-order statistical features can be classified into five groups [64]:

1. Gray Level Co-occurrence Matrix, GLCM features (22 features) describe combinations of grey levels of neighbouring pixels.
2. Gray Level Run Length Matrix, GLRLM features (16 features) quantify grey level runs in an image, defined as the number of pixels with the same grey level value.
3. Gray Level Size Zone Matrix, GLSZM features (16 features) quantify grey level zones in an image, defined as the number of connected pixels with the same grey level values.
4. Neighbouring Gray Tone Difference Matrix, NGTDM features (5 features) quantify the differences in grey-level intensities between each pixel or voxel and its neighbouring pixels or voxels within the ROI.
5. Gray Level Dependence Matrix, GLDM features (14 features) quantify how often pairs of pixels or voxels with specific grey levels are adjacent within the ROI.

Higher order statistics features refer to statistical measures that capture complex relationships and patterns beyond first-order and second-order statistics. These features are generated by applying filters, e.g. the wavelet transform, to the ROI before extracting features [66].

3.4.2. Wavelet Transform Filtering

The wavelet transform is a technique that decomposes an image into different frequency components by applying high-pass and low-pass filters to accentuate or suppress certain frequency bands [67]. A one-level wavelet decomposition produces four distinct filtered images denoted as LL (low-low), HL (high-low), LH (low-high), and HH (high-high), where "low" signifies low-frequency components and "high" signifies high-frequency components.

High-pass filtering in both directions (HH) captures diagonal details. High-pass filtering followed by low-pass filtering (HL) captures vertical edges. Conversely, low-pass filtering followed by high-pass filtering (LH) captures horizontal edges. Finally, low-pass filtering in both directions (LL) captures the lowest frequencies at varying scales.

The low-frequency components, representing smooth variations, serve as the foundation of an image. In contrast, the high-frequency components, responsible for capturing edges and fine details, refine the image, resulting in a more detailed representation [68].

3.5. Performance Metrics

This research uses two performance metrics to assess the developed ML models: the accuracy and the Receiver Operating Characteristic Area Under the Curve (ROC AUC). Accuracy is a common and straightforward metric used to assess the performance of classification models, particularly in ML and data analysis. It measures how often a model correctly predicts the class labels of the data points in a dataset [69] and is predicted using the following equation:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{3.5}$$

While it is easy to calculate and interpret, it has a few limitations when evaluating model performance. Firstly, for an imbalanced dataset, where one class significantly outnumbers the other, a model that predicts the majority class for all instances can still achieve a high accuracy. Furthermore, this metric does not provide insights into the types of errors the model makes. For example, it doesn't distinguish between false positives and false negatives.

The ROC AUC is also a metric used to evaluate the performance of binary classification models. It quantifies the ability of a model to distinguish between two classes across different classification thresholds. It derives its name from the receiver operating characteristic (ROC) curve, a graphical representation of a model's performance, see Figure 3.10.

The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold values. TPR, also known as Sensitivity or Recall, represents the proportion of actual positive instances correctly classified as positive by the model. FPR, on the other hand, represents the proportion of actual negative instances incorrectly classified as positive by the model. Subsequently, the ROC AUC score is calculated as the area under the ROC curve, which values from 0.5 to 1.0 [69, 70].

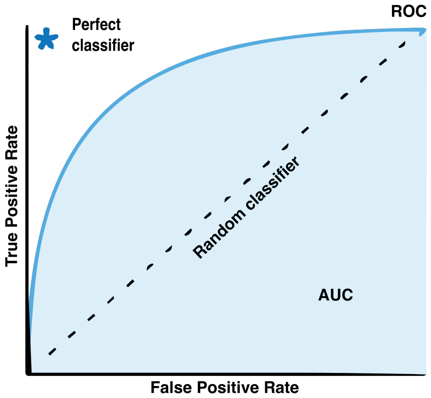
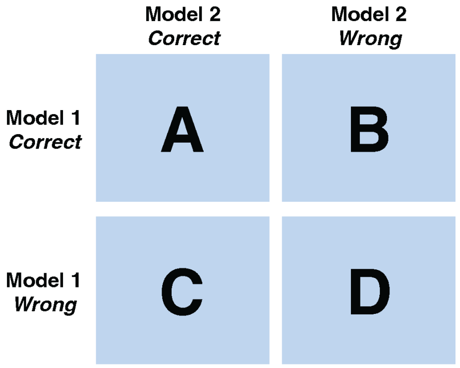


Figure 3.10: This graph showcases a Receiver Operating Characteristic (ROC) curve, a valuable tool in assessing the performance of classification models. It illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). A random classifier's curve is depicted as a baseline, while a perfect classifier is expected to create a curve that passes through the upper left corner, marked in the graph by a blue star. The Area Under the Curve (AUC) quantifies the classifier's ability to distinguish between classes, with a higher AUC indicating better performance.

A higher ROC AUC score implies that the model is more effective at distinguishing between two classes. Consequently, ROC AUC serves as a valuable metric for assessing the overall quality of a binary classification model. It is particularly useful when comparing different models or selecting the most suitable model for a specific classification task.

3.6. McNemar's Test

McNemar's test is a statistical method employed to evaluate the significance of disparities in the performance of two closely related ML models. It helps determine if there is a statistically significant difference in the classification results on the same dataset of two models. This test constructs a 2 × 2 contingency table using the classifiers' binary classification outcomes (see Figure 3.11). It calculates the McNemar statistic, which measures the contradictions in their classifications while considering interdependencies. The resulting statistic is evaluated against a significance level 0.05 to confirm whether one model significantly outperforms the other. McNemar's test is particularly valuable for assessing differences in classifier performance and is a useful tool for hypothesis testing in machine learning model evaluation [71].



	Model 2 *Correct*	Model 2 *Wrong*
Model 1 *Correct*	A	B
Model 1 *Wrong*	C	D

Figure 3.11: A 2×2 contingency table using the classifiers' binary classification outcomes.

4 Methods

This section outlines the methodology for categorising reconstruction kernels based on features extracted from real patient scans. Those features serve as input for different models that aim to distinguish between sharp and soft kernels.

4.1. Dataset

This study used thoracic CT scans from the National Lung Screening Trial (NLST) [72] and the Lung Image Database Consortium image collection (LIDC-IDRC) [73]. These data sets comprise scans obtained from multiple vendors using various reconstruction kernels and other reconstruction and acquisition parameters. In the NLST, approximately 54,000 participants were enrolled, which resulted in the acquisition of over 75,000 CT screening exams [72]. The smaller LIDC-IDRC dataset contains 1018 cases from 1010 patients [73]. The datasets represent a wide range of scanner manufacturers and models, as well as scan parameters. The datasets comprise CT scans acquired before 2010, all using the conventional FBP reconstruction technique.

In the datasets, 37 distinct reconstruction kernels from four manufacturers are available; for each reconstruction kernel, ten cases were selected if possible, otherwise the maximum number of cases available. Patient scans with less than ten slices available were excluded because these scans might not provide enough information or coverage to represent the case accurately. Similarly, kernels with less than 100 applicable slices were not included.

4.2. Selecting Extremes

Two kernel types are chosen for each manufacturer: one representing the smoothest kernel and the other representing the sharpest kernel among their products. This selection is based on the vendor descriptions described in the previous theory section (Section 3.1.4) and on the number of applicable patients and slices to ensure a sufficient data set size and scan parameter diversity.

4.3. Preprocessing

Each scan undergoes processing before computing its characteristics. The pixel values are initially rescaled to convert them into a more interpretable and consistent range, the Houndsfield unit (HU) [74]. This rescaling process utilizes the 'RescaleSlope' and 'RescaleIntercept' attributes for a linear transformation of the data using the following formula: $rescaled\_value = (pixel\_value * rescaleSlope) + rescaleIntercept$.

Following the rescaling process, the patient is segmented from the scan. This involves applying a threshold of $-200HU$ to the image, identifying the largest cluster, and then employing a morphological algorithm to fill any holes within the cluster, thus creating a final mask that covers the entire patient. By multiplying this mask with the rescaled CT scan, only the pixels corresponding to the patient are selected and utilized for further research purposes. Figure 4.1 shows the complete preprocessing process.

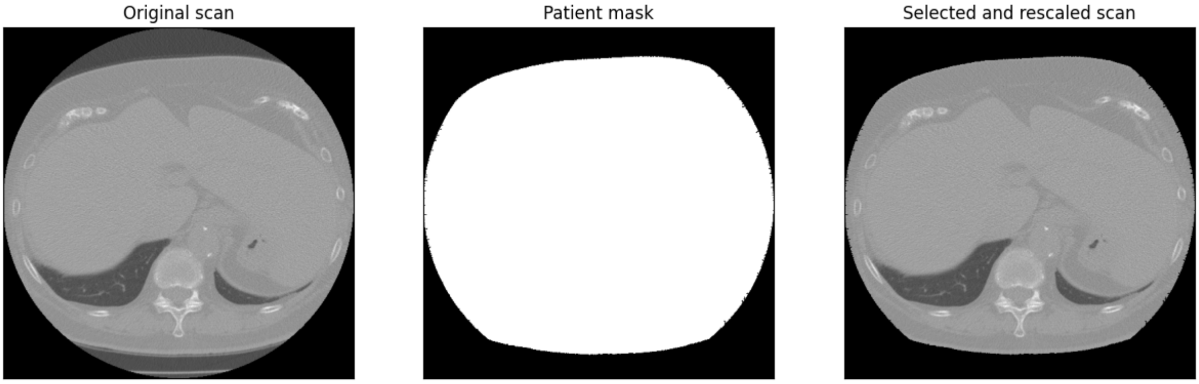


Figure 4.1: The preprocessing steps that are applied to each CT scan. The first (left) image shows the original scan retrieved from the dataset. The middle image shows the patient mask that segments only the patient from the whole scan. This mask is created by applying a threshold of $-200HU$, identifying the largest cluster, and utilizing a morphological algorithm to address gaps within the cluster. The last (right) image shows the final image, where the patient is segmented using the mask and the pixel values are rescaled to HU.

4.4. Feature Extraction

This section explains the extraction of two distinct sets of image features that have the potential to facilitate kernel categorisation: noise features and radiomic features.

4.4.1. Noise Features

Noise Magnitude

The measurement of the noise magnitude is based on the calculation of the SD following the automated noise calculation method proposed by Anam et al. [56], which was explained more in-depth in Theory Section 3.3.1. This approach comprises three steps: initially, the patient's image is segmented as explained in the preprocessing step (Section 4.3). Subsequently, an SD heatmap is generated by computing the SD value for each pixel using a sliding window of 30×30 pixels. Finally, the noise is estimated by identifying the minimum SD value from the SD heatmap. This method is applied to every slice in the data set, which results in one noise magnitude value per slice.

Noise Texture

The determination of noise texture relies on computing the CF of the NPS. The ten most homogeneous ROIs from each slice, indicated by the ten lowest SD values, are extracted with pixel dimensions $(L_x \times L_y)$ of 30×30 pixels. The ROIs are allowed to overlap with a maximum of 50%; this overlap has traditionally been recommended when calculating the NPS [75, 76].

The mean value is subtracted for each ROI to recentre the inputs around zero. The recentring simplifies the analysis and ensures that the frequency components of the image are correctly interpreted. Following that, a zero-padding of 30 pixels is applied to enhance the accuracy of the Fourier transform. This results in a new image with the original content centred in the middle and the newly added values set to zero. Zero padding increases the resolution of the NPS curve, but it also introduces a slight increase in noise [59, 77].

Subsequently, the 2D NPS is computed according to Equation (3.3) by averaging over the extracted ten ROIs. Next, the radial symmetry of the 2D NPS is utilized to generate a 1D spectrum: the data is binned according to its radial spatial frequency, followed by averaging the data in each bin. The radial bin size for all calculations was selected as 0.1 mm^{-1}. This value was established through manual adjustment to achieve the desired level of spectral smoothness. Ultimately, the 1D NPS curve is normalized by calculating the area under the curve.

The CF was computed from the normalized NPS curves, using Equation 3.4, which is equivalent to calculating the centre of gravity of the NPS [55, 78].

4.4.2. Radiomic Features

The radiomic feature extraction process uses the PyRadiomics library, an open-source package widely used for radiomic analysis [79]. This package provides a comprehensive set of features that can be extracted from the CT scans. These features encompass a range of quantitative measurements, including intensity-based, shape-based, texture-based, and wavelet-based features, calculated using PyRadiomics's built-in functions. The features are extracted using the default settings of PyRadiomics: a fixed bin width of 25HU.

The following feature classes are included: first-order statistics (18 features), GLCM (22 features), GLRLM (16 features), GLSZM (16 features), NGTDM (5 features), and GLDM (14 features). The shape features are excluded as this study is not interested in a certain shape of the tissue because this has no relation to the applied reconstruction kernel. In total, 91 features are extracted from the original image shape.

In addition, wavelet filtering is applied to decompose the original image into different frequency bands, which allows for the analysis of image details at different scales and resolutions. From the resulting four images, the same 91 features are extracted. This results in the extraction of 455 radiomic features per ROI.

ROI Selection

In radiomics, feature extraction typically involves isolating an ROI, either manually or automatically, to separate it from the background and surrounding tissues. Commonly used ROIs in radiomics include tumours, organs and lesions [62]. However, it is important to note that the primary goal of this study is not to analyse the intrinsic tissue characteristics of a specific anatomical region for medical purposes.

Instead, the focus here is on analysing the specific attributes of reconstruction kernels, which can provide valuable insights for quantifying different reconstruction kernels. Consequently, the extraction of radiomic features must be carried out on an ROI that minimizes the influence of patient-specific characteristics. Unfortunately, the most suitable ROI selection approach for this analysis remains uncertain. As an initial approach, the analysis will be performed on the ten most homogeneous patches per CT slice, extracted from the patient area measuring 30×30 pixels using the SD heatmap developed in Section 4.4.1. These patches are not allowed to overlap to ensure that each patch represents a distinct and independent region of interest.

4.4.3. Number of ROIs

It is worth emphasizing that noise characteristics can only be derived once per slice, unlike radiomic features, which are not subject to such limitations. This thesis quantifies noise magnitude by the SD value obtained from the most uniform area, representing the lowest SD value. In contrast, noise texture is established by averaging the NPS values from the ten most uniform areas. Consequently, a single collection of noise features can be generated per slice, whereas radiomic features can be calculated from various ROIs within a single slice, allowing for the extraction of multiple sets of radiomic features from the same slice.

4.5. Distribution Analysis

In total, one set of noise features and ten sets of radiomic features are computed for each slice. Each set of noise features, consisting of the SD and the CF value, is used to gain insight into the distribution of the dataset and the relationship between these two features. This distribution analysis exclusively considers the noise features due to their high interpretability, particularly compared to the extensive set of radiomic features. Moreover, prior knowledge concerning the noise features and their association with reconstruction kernels makes them more relevant to this analysis.

4.5.1. Distribution

For each patient, the median value for both features, the SD and CF, is determined and documented. The median is used because it is robust to outliers, making it more resistant to extreme values [80]. The median values are graphically analysed per kernel by displaying the box plots, which offer valuable insights into a dataset, including identifying outliers and information about the symmetry and tightness of data clustering [81].

4.5.2. Relationship

An optimal reconstruction shall maintain a high CF, indicating a small loss of higher frequency components while having a low SD, which indicates a larger amount of noise reduction. Unfortunately, in practice, most noise reductions are paired with a CF shift toward the lower frequency [55]. The relationship of the computed CF and SD values will be visualized by mapping the patients in a 2D space.

Subsequently, the Spearman's correlation coefficient (ρ) between the two features is calculated to assess the relationship between the two features. It measures the strength and direction of the monotonic relationship between two variables. Specifically, it assesses whether there is a consistent trend in how the two variables change together without assuming that the relationship is strictly linear. The values of ρ range from -1 to 1, where -1 indicates a perfect negative monotonic relationship, 1 indicates a perfect positive one, and 0 suggests no monotonic relationship. The Spearman's correlation coefficient is calculated using the following Equation:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\,(n^2-1)} \tag{4.1}$$

Where d_i is the difference between the ranks of the two variables for each data pair and n is the number of data points. Spearman's correlation coefficient is preferred here over other correlation coefficient approaches, like Pearson's correlation coefficient, since it does not necessitate normally distributed variables as input, in contrast to Pearson's coefficient [82].

4.6. Model Implementation

Two ML models are trained to categorise 'smooth' and 'sharp' kernels based on image features extracted from CT scans. The first model, the SVC_noise, is a linear SVC model trained using noise features. The second model, referred to as the $RFC\_radiomics$ model, is an RFC model trained with radiomic features.

A linear SVC is chosen for the noise features because it is assumed that the data can be well-separated by a linear decision boundary based on the expected relationship between CF and SD values. Furthermore, a linear SVC provides easily interpretable results, is computationally efficient, is less prone to overfitting compared to more complex models, and tends to have good generalization performance.

On the other hand, for the radiomic features, an RFC model is selected due to its simplicity in implementation and fast operational speed. Furthermore, it has been proven to be extremely successful in various domains and, most notably, it can handle large and complex datasets effectively [83].

4.6.1. Input data

For the models' training, validation and testing, the kernels selected as the smoothest and sharpest following Section 4.2 are used. The data points are labelled as smooth (0) and sharp (1), and the data has been split into a training and test dataset with a ratio of 80:20.

Each CT slice has one set of noise features and ten sets of radiomic features available through the feature extraction performed in Section 4.4. This tenfold difference in the number of radiomic feature sets arises because radiomic features are extracted ten times for each slice, while noise measures are extracted only once. To maintain consistency, the split was performed based on the slice numbers. This approach guarantees that different patches extracted from the same slice are included in the same subset, ensuring that the training and test datasets for the RFC model match those utilized in the SVC model.

The training dataset is used for the hyperparameter tuning and performance analysis, using ten-fold cross-validation, and the test dataset for model evaluation, see Figure 4.2. Ten-fold cross-validation means the training dataset is partitioned into ten equally sized folds. Each fold acts as a validation set once, while the remaining nine folds serve as the training set. This partitioning ensures that every dataset sample is used for training and validation purposes.

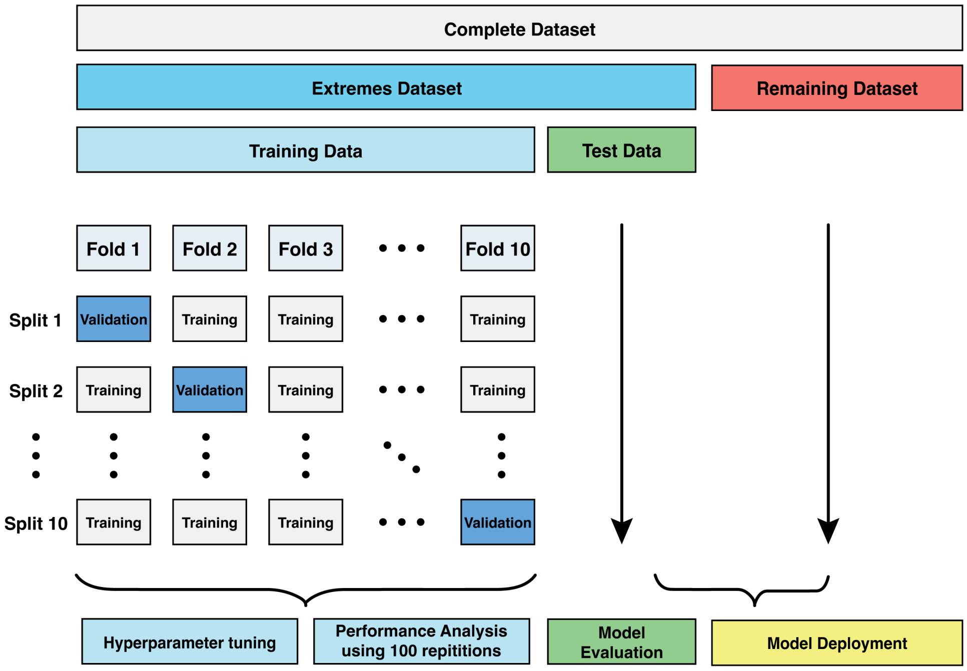


Figure 4.2: Overview of the datasets used in the methods. The extremes dataset comprises the CT scans reconstructed with the eight selected extreme kernels in Section 4.2. This dataset is split into a training dataset (80%) and a test dataset (20%). The training data is used for the hyperparameter tuning and the performance analysis, employing ten-fold cross-validation. In the ten-fold cross-validation, the training dataset is split into ten folds. In each split, nine folds are used for training, and one is used for validation. This process is repeated ten times, using a different fold as the validation group in each split. The test dataset is reserved for model evaluation. Additionally, the test dataset is combined with the remaining dataset, which contains all CT scans reconstructed with kernels not identified as "extreme", for model deployment

4.6.2. Hyperparameter Tuning

For each model, a subset of hyperparameters is tuned to tweak the model performance for optimal performance. Only the regularization parameter C has been tuned for the SVC model using the $GridSearchCV$ function in the Scikit-learn library [84]. This function performs a search over a specified parameter grid, systematically trying all hyperparameters and cross-validating the model's performance to find the best value for the hyperparameter. For the regularization parameter, the specified parameter grid is the list of values: $[0.1\ \ 1.\ \ 10.]$ and a ten-fold cross-validation is performed to evaluate the performance of each parameter.

Hyperparameter tuning for the RFC model is performed for two parameters: $n\_estimators$ and max_depth using the $RandomizedSearchCV$ function in Scikit-learn [84]. This function searches through a hyperparameter space randomly and finds the optimal combination of parameters for the models. It randomly selects a fixed number of combinations from the specified distributions for each hyperparameter. The function randomly selects a hundred combinations (specified by n_iter) from the hyperparameter space. The distribution of the number of decision trees to be created (n_estimators) is between, $[10-100]$ and of the maximum depth of each decision tree (max_depth) between $[10-120]$.

4.6.3. Performance Analysis

For the performance analysis of the models, a ten-fold cross-validation with a hundred repeats is applied using the optimal hyperparameters determined in the previous step. This method ensures robust classifier performance evaluation so the results are as generalizable as possible.

The cross-validation process proceeds iteratively, with each fold serving as the validation set exactly once. This entire cross-validation process is repeated a hundred times to enhance the reliability and stability of the evaluation. The repeated evaluation helps account for potential variability and randomness in the data, ensuring that a specific dataset partitioning does not bias the results. This approach increases the reliability of performance measures and provides a more comprehensive assessment of the model.

For each fold, the accuracy and ROC AUC of the training and validation set are computed and saved. The mean and standard deviation ROC AUC and accuracy are determined as the final performance measure for the training and validation performances.

4.6.4. Model Evaluation

The trained and validated models are evaluated using the test set. For each set of features in the test dataset, a prediction is obtained by both models by applying the *model.predict* function in the Scikit-learn library [84]. Subsequently, the accuracy and ROC AUC scores are calculated to measure the performance of both models on unseen data. This process ensures that the model's effectiveness and generalizability are assessed, allowing meaningful conclusions and insights to be drawn from the results.

4.7. Model Deployment

The models are deployed to classify smooth (0) kernels and sharp (1) kernels. The cases reconstructed with the kernels that were not identified as the extremes, together with the test dataset, are used as input; see Figure 4.2. This will result in one prediction per slice by the *SVC_noise* model and ten predictions per slice by the *RFC_radiomics* model (one per patch).

4.7.1. True Class Label

Since no true class labels (the ground truth) are available for the newly added kernels, the performance of the deployed model can not be evaluated yet. Therefore, the true class label of each kernel is determined by identifying the predominant class within each kernel. For instance, if 50.1% of the slices extracted from kernel "C" are classified as smooth, the true class label for kernel "C" is assumed to be smooth.

4.7.2. Aggregation

A final prediction for each patient is determined by aggregating the predictions for that patient. The predictions are aggregated by applying majority voting per patient; this involves classifying a data point based on the majority among a group of predictions. Majority voting is a simple and effective way to make decisions based on consensus when there are multiple sources of information, such as multiple slices from the same patient. It can help improve the robustness and accuracy of predictions. Several majority voting approaches have been applied, explained below; for each approach, the ROC AUC and accuracy scores are calculated using the previously determined true class labels and newly predicted labels per patient.

Majority Voting Approaches

The aggregation approaches for the noise and radiomic features differ based on the number of feature sets available for each slice. Noise features provide one set of features per slice extracted from the most homogeneous patch. This allows for majority voting based on the slice predictions. Conversely, radiomic features yield ten sets per slice, offering multiple options for majority voting. These aggregation approaches are described below and are visualized in Figure 4.3.

1. **Slice-Level Voting (SLV):** In this approach, one prediction per slice is aggregated to determine the final patient prediction, with predictions based on the most homogeneous patch of the slice. It applies to both the *SVC_noise* and *RFC_radiomics* models.
2. **Multi-Patch Slice-Level Voting (MPSLV):** This method aggregates ten predictions per slice to derive the final patient predictions, with predictions made on the ten patches per slice. This approach is specifically applicable to the *RFC_radiomics* model.
3. **Hierarchical Multi-Patch Voting (HMPV):** In this strategy, the initial aggregation combines the ten predictions per slice to determine slice predictions, which are further aggregated to establish the final patient predictions. This strategy applies exclusively to the *RFC_radiomics* model.

4.8. Radiomic Features Analysis

An extensive set of radiomic features is extracted from each slice's ten most homogeneous ROIs. As mentioned, the most suitable ROI selection method for extracting radiomic features remains uncertain for this

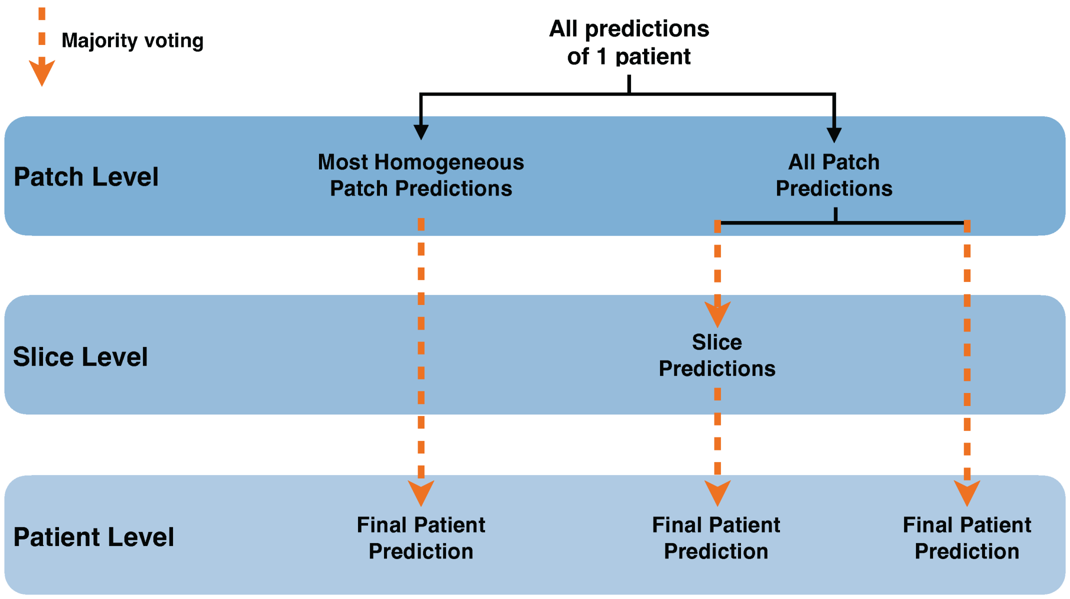


Figure 4.3: Visualization of the three distinct majority voting approaches employed as data aggregation methods to determine the final prediction per patient. The left branch displays Slice-Level Voting (SLV), the middle Hierarchical Multi-Patch Voting (HMPV) and the right Multi-Patch Slice-Level Voting (MPSLV).

application. In addition, selecting the ten most homogeneous patches of each slice is computationally heavy. Therefore, in this section, a random ROI selection approach for extracting radiomic features is explored and compared with the original ROI selection method.

Furthermore, a feature importance analysis is conducted to understand the underlying dataset relationships better and improve the model interpretability. Also, it is a first step for feature selection, which may improve the performance of the RFC and reduce the computational costs of the radiomic feature extraction and the training time.

4.8.1. ROI Selection: Random versus Most Homogeneous

A random patch selection approach is evaluated and compared to the original homogeneous patches approach. Ten random non-overlapping patches are extracted from the patient of 30×30 pixels for each slice. Subsequently, the same process has been completed: hyperparameter tuning, performance analysis, model evaluation and finally, model deployment. The final results are compared to the model performance of the RFC trained on homogeneous patches using McNemar's test (explained in theory section 3.6).

4.8.2. Feature Importance

In total, the RFC has been fed with an extensive set of 455 radiomic features; in this section, the significance of each feature is investigated. RFC models typically provide a feature importance measure known as the Mean Decrease in Impurity. This measure quantifies each feature's contribution to the model's predictive performance. However, it has a limitation; it may assign high importance to features not necessarily predictive on unseen data, particularly when the model is overfitting.

To mitigate this issue, a permutation-based feature importance approach is employed. Unlike MDI, this technique allows computing feature importance on unseen data, making it more robust. Permutation feature importance is the reduction in a model's score when a single feature's values are randomly shuffled [51]. This shuffling disrupts the relationship between the feature and the target variable. Consequently, the drop in the model's score indicates the feature's importance, revealing how much the model relies on that particular feature.

This permutation-based technique offers the advantage of being model-agnostic and applicable across various models. It can be repeatedly calculated with different feature permutations, providing a robust assessment of feature importance.

4.9. Model Comparison

In the previous steps, two models have been developed for categorising reconstruction kernels based on their sharpness, *SVC_noise* and *RFC_radiomics*. In this section, a model comparison is performed, including several aspects.

Firstly, a comparative assessment is conducted to evaluate the performance of the two models, *SVC_noise* and *RFC_radiomics*. For this comparison, for each model, the final model with the best-performing majority approach is used as input to McNemar's test (explained in section 3.6) to confirm whether one model significantly outperforms the other.

Additionally, a thorough analysis explores the differences in the categorisation of sharp and soft kernels employed by each model. This investigation aims to shed light on the implications of designating one of these categorisations as the gold standard and its impact on the performance of the other model. An examination of misclassified patients for each model is also carried out, with particular attention to identifying any overlap between the two models' misclassifications. Finally, an overall performance comparison is performed, evaluating the advantages and disadvantages of each model.

5
Results

5.1. Dataset

As a result of the predefined criteria, 14 cases were excluded from the dataset due to having fewer than ten slices. It is worth noting that 10 of these excluded cases were exclusively reconstructed with the Toshiba kernel FC11, excluding this entire kernel. Additionally, the case reconstructed with FC52 was not included due to the total number of available slices for this kernel being less than 100, leading to the removal of this entire kernel. This selection process led to a dataset of 270 cases, reconstructed using 35 distinct kernels, each characterized by a wide range of acquisition and reconstruction parameters. A concise summary of the chosen dataset is presented in Table 5.1, while a comprehensive overview can be found in Appendix C.

5.2. Selecting Extremes

The chosen smoothest and sharpest kernels per vendor, respectively, are Standard, shortened as STD & Lung (GE), A & D (Philips), B20f & B80f (Siemens), and FC01 & FC82 (Toshiba). Standard has been chosen instead of Soft due to the limited number of patients (1) and slices (116) available for the Soft kernel. This is the same reason for selecting B20f instead of B20s; only one patient is available containing 331 slices for B20s.

Table 5.1: The arrangement of kernels follows a vendor-based order derived from the sharpness descriptions given by the vendors. It ranges from smooth (upper) to sharp (lower). The kernels in bold and highlighted with a green colour represent the selected smoothest and sharpest kernels per vendor.

kernel	Tube current	Tube voltage	Slice thickness	Patient	Slices
Philips					
A	67-417	120	3.2	10	920
B	67-180	120	3.2	10	1255
C	60-120	120	1.3-3.2	10	1656
D	93-187	120	2.0-3.2	10	1770
EC	93-187	120	3.2	3	460
GE					
SOFT	60	120	2.5-10	1	116
STD	45-160	120-140	1.25-2.5	10	1706
BONE	40-90	120-140	1.25-2.5	10	987
LUNG	80-160	120	2.5	10	1671
Siemens					
B20s	225	130	1.25	1	331
B20f	90-120	120	2.0	10	1579
B30s	38-173	110-130	2.0-5.0	10	1499
B30f	75-210	120	1.0-2.0	10	1769
B31s	38-275	130	2.5-3.0	10	1271
B31f	40-500	120	2.0-3.0	10	1327
B35f	120	120	2.0	2	347
B40s	133	130	3.0	1	278
B40f	80-330	120-140	5.0	4	264
B41s	270	120	3.0	1	138
B45f	120-513	120	1.0-3.0	10	2826
B50s	80-100	120-130	2.0	4	672
B50f	90-160	120-140	2.0-5.0	10	1445
B60s	63-270	110-130	2.0-3.0	10	1492
B60f	90-160	120	2.0	10	1652
B70f	105-381	120	2.0	9	1980
B80f	150-250	120	1.0	10	3372
Toshiba					
FC01	80-260	120-135	2.0-3.0	10	1419
FC02	160	120	3.0	9	844
FC03	260	130	2.0	3	445
FC10	80-150	120	2.0	10	1790
FC50	80-160	120	2.0	9	1523
FC51	80-160	120	2.0	10	1703
FC53	80	120	1.0-2.0	3	585
FC30	160	120	2.0-3.0	10	1348
FC82	80-160	120	2.0	10	1513

5.3. Feature Extraction

5.3.1. Noise Features

The noise magnitude in terms of the lowest SD of a 30×30 pixel's ROI has been extracted from each slice. An example of this process is visualized in Figure 5.1.

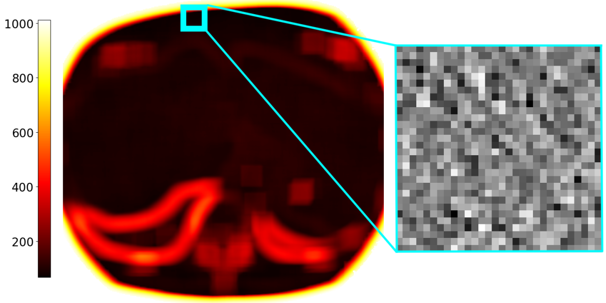


Figure 5.1: The SD heatmap of a patient CT scan is visualized. The heatmap is created using a sliding window of 30×30 pixels. The most homogeneous ROI of the image is indicated by blue and represents the ROI with the lowest SD. On the right, a zoom-in of this most homogeneous patch is added.

The ten most homogeneous patches have been identified and selected using the SD heatmap. Subsequently, the average NPS of each slice has been computed. These NPS values have been radially averaged to create a 1D spectrum and compute the CF for each slice; this process is shown in Figure 5.2.

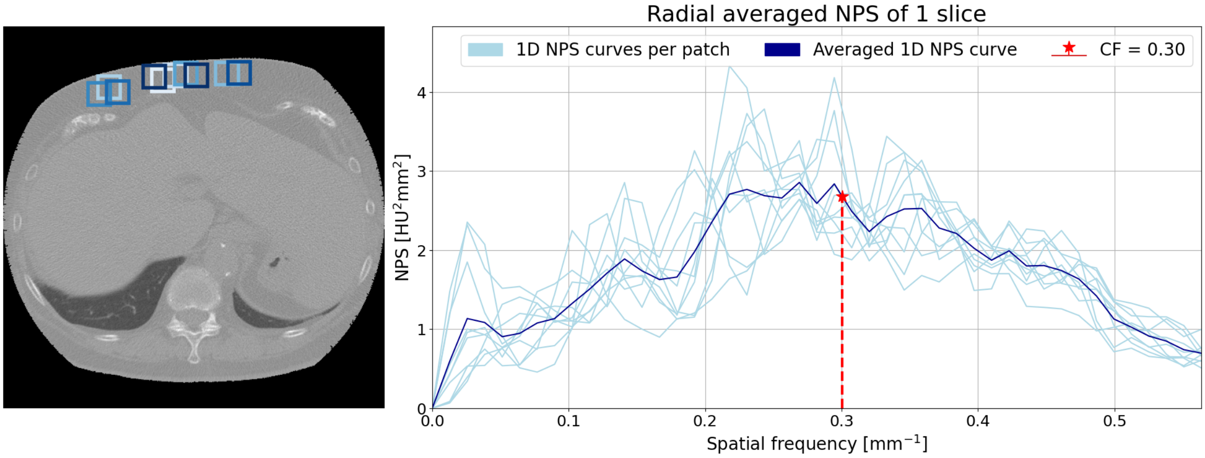


Figure 5.2: The left image displays the ten most homogeneous patches $(30 \times 30$ pixels) with a maximum overlap of 50% on top of a preprocessed CT scan. These patches have been identified using an SD heatmap. The right side of the figure shows the 1D NPS curves of each patch in light blue and the final radial averaged NPS in dark blue. The radial averaged NPS is utilized to compute the CF of the slice, indicated by the red line in the graph.

5.3.2. Radiomic Features

An example of the ROI selection process for the radiomic features is visualized in Figure 5.3. The ten patches indicated by blue are the ten most homogeneous non-overlapping patches of the slice, and 455 radiomic features are extracted from each patch.

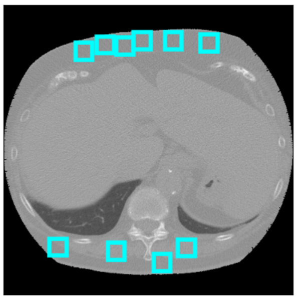


Figure 5.3: A preprocessed CT scan including the ten non-overlapping most homogeneous patches $(30 \times 30$ pixels) indicated by blue.

5.4. Distribution Analysis

5.4.1. Noise Magnitude Distribution

Subsequently, for each patient, the median of the noise magnitude has been computed; these results are added to Appendix D in Table D.1 and visualized in Figure 5.4 using box plots. The following observations are made:

- The kernels with a lower noise magnitude are distributed closer together than those with a higher noise magnitude, which means that the noise values are more similar within these kernels.
- The kernels with a higher median noise magnitude are more spread out.
- The kernels chosen as the extremes do not always show the smallest or largest values, e.g. B20f and FC82.
- The kernels of Siemens with a higher noise magnitude have multiple outliers.
- GE and Philips tend to show two distinct groups based on their noise magnitude values, whereas Toshiba and, in particular, Siemens show a more linear increase in the noise magnitude.
- Philips generally has lower noise magnitude values than the other manufacturers.

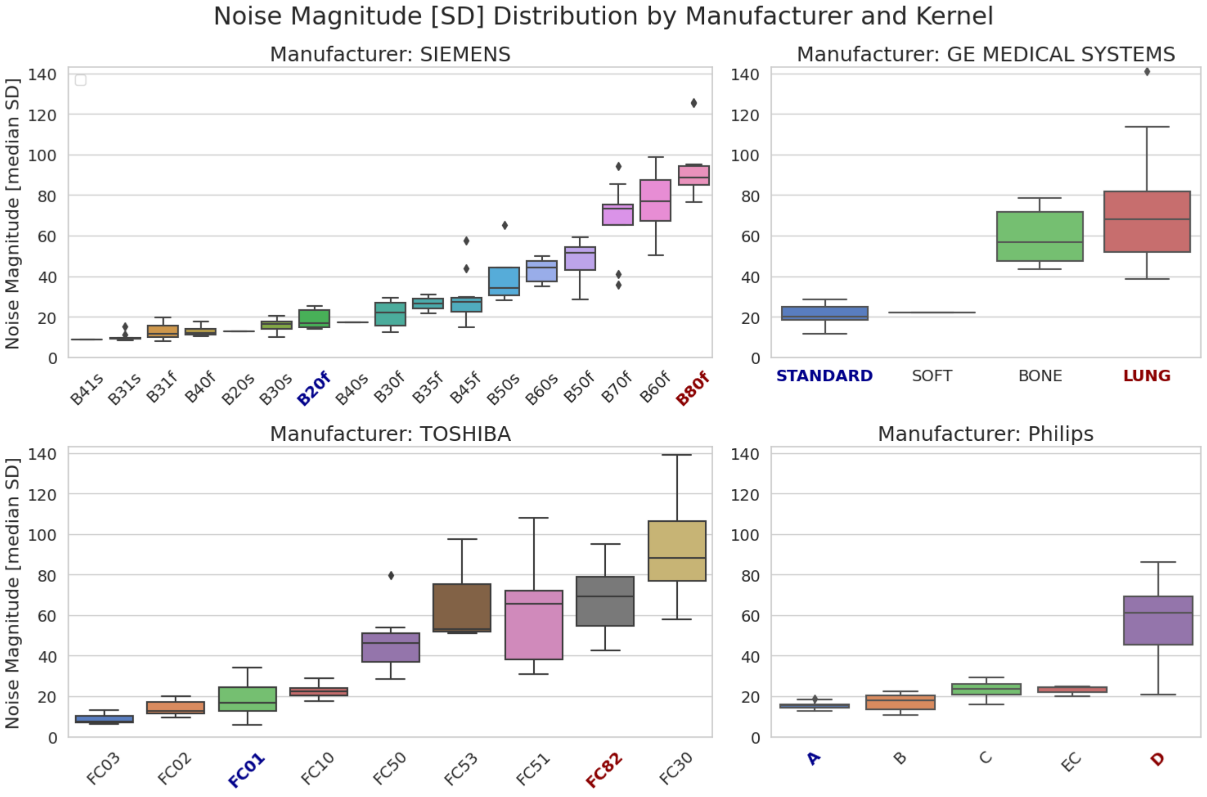


Figure 5.4: The distribution of the median noise magnitude values per kernel is displayed using box plots. The four manufacturers are visualized separately; for each graph, the kernels are ordered from left (lowest) to right (highest) based on the median SD value for that kernel. The kernels that have been identified as 'extremes' in Section 4.2 are highlighted by blue (smoothest) and red (sharpest). The black points indicate the outliers.

5.4.2. Noise Texture Distribution

For each patient, the median of the noise texture has been calculated by taking the median central frequency value of all slices of one patient. These results have been included in Appendix E and are presented in Table E.1. Additionally, the distribution per kernel has been visualized through box plots in Figure 5.5, and the following findings are noted:

- Kernels with a lower noise texture are distributed slightly closer together, less obvious than the noise magnitude.
- Kernels chosen as the extremes do not always show the smallest or largest values, e.g. FC01 and D.
- Kernels of Toshiba with a higher noise texture have several outliers.
- The results per manufacturer show less clear discernible groups compared to the distribution of the noise magnitude values.
- Philips generally has lower noise texture values than the other manufacturers.

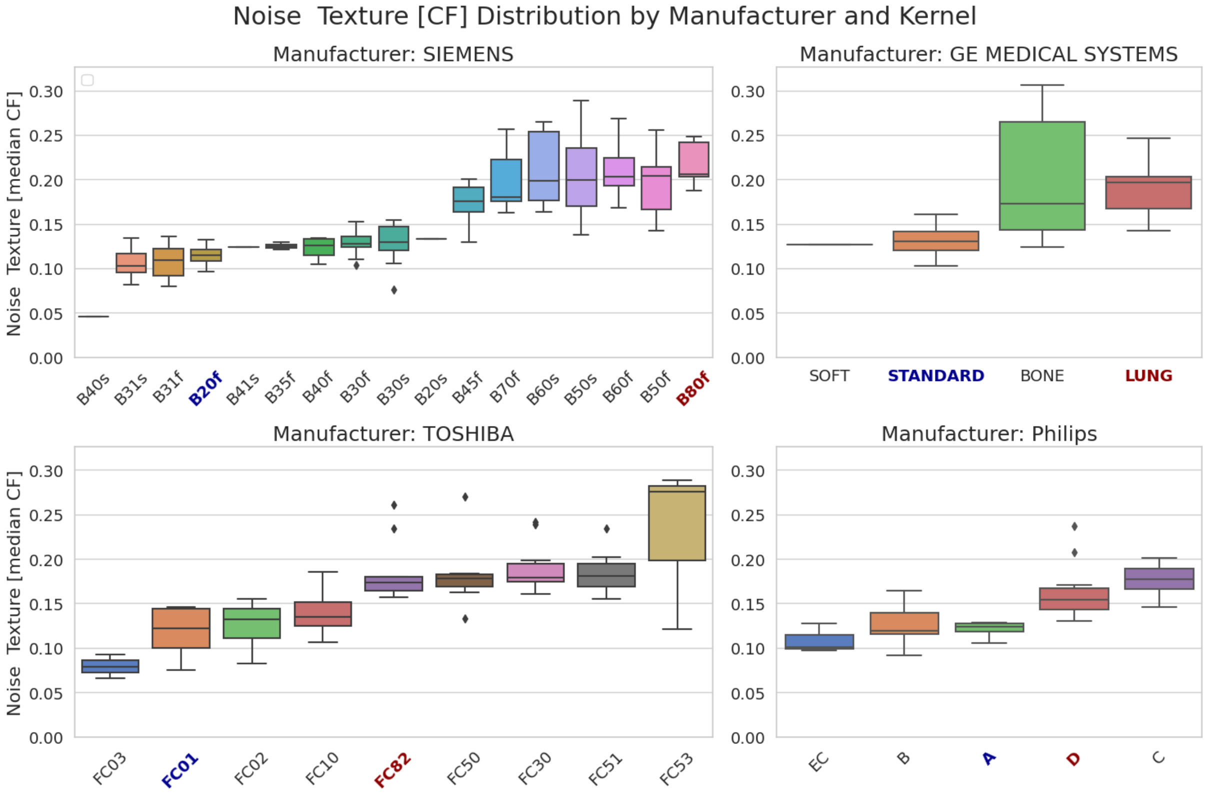


Figure 5.5: The distribution of the median noise texture values per kernel is displayed using box plots. The four manufacturers are visualized separately; for each graph, the kernels are ordered from left (lowest) to right (highest) based on the median CF value for that kernel. The kernels that have been identified as 'extremes' in Section 4.2 are highlighted by blue (smoothest) and red (sharpest). The black points indicate the outliers.

5.4.3. Relationship

The relationship between the noise magnitude and the noise texture is visualized in Figure 5.6 by plotting the median values of the central frequency and the standard deviation per kernel in a scatter plot.

The graph illustrates the anticipated relationship: when the central frequency rises, signifying a transition towards higher frequencies, it is accompanied by an increase in the standard deviation, resulting in a decrease in noise reduction. This finding is further supported by the calculated Spearman's correlation coefficient: r=0.79, p<0.001. This outcome underscores a strong, positive, statistically significant monotonic relationship between the two noise features.

Moreover, the scatter plot visually reveals that two linearly separable clusters emerge when utilizing noise magnitude and texture as features. This discovery encourages further exploring a linear classification algorithm that employs noise features as input variables.

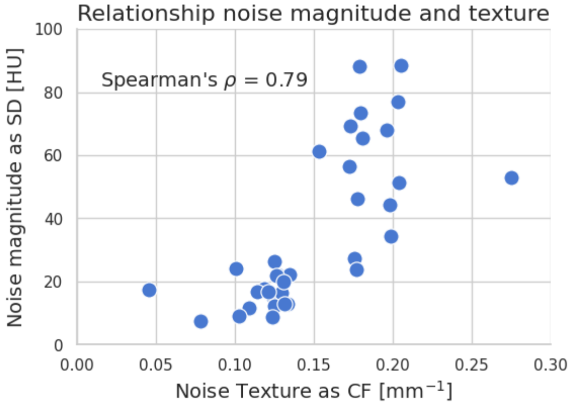


Figure 5.6: The relationship between the noise magnitude measured as SD [HU] and the noise texture measured as CF [mm^{-1}]. Each point is the median value computed over all patients from the same kernel. The scatter plot shows a strong, positive monotonic relationship, confirmed by Spearman's correlation coefficient of 0.79.

5.5. Model Implementation

5.5.1. Input data

Table 5.2 shows the number of slices per kernel and dataset. For each slice, one set of noise features and ten sets of radiomic features are available. When considering the total number of slices within the smooth and sharp categories, it becomes evident that there exists a class imbalance. Specifically, the sharp category comprises a significantly larger number of slices (8326 slices) compared to the smooth category (5624 slices) despite a similar number of patients in both groups.

Table 5.2: Specifications of the input data split in training and test dataset, per kernel, the number of slices per dataset is specified. The training set is used for the hyperparameter tuning and model performance analysis, whereas the test set is used for model evaluation.

Vendor	Smooth			Sharp		
	Kernel	Training	Test	Kernel	Training	Test
GE	STD	1384	322	LUNG	1342	329
Philips	A	747	173	D	1399	371
Siemens	B20f	1259	320	B80f	2683	689
Toshiba	FC01	1135	284	FC82	1211	302
Total		4525	1099		6635	1691

5.5.2. Hyperparameter Tuning

For the *SVC_noise* model, the regularization parameter C has been tuned to find the optimal value. This search yielded an optimal value of 0.1 for the regularization parameter, with a mean test ROC AUC score of 0.9803.

The hyperparameter search conducted for the $n\_estimators$ and max_depth parameters of the $RFC\_radiomics$ model led to the identification of the most effective parameter combination, which consisted of $n\_estimators$ = 20 and max_depth = 18. This combination achieved a mean test ROC AUC score of 0.9789, as highlighted by the red cross in Figure 5.7, which shows the performance of the hundred randomly selected combinations.

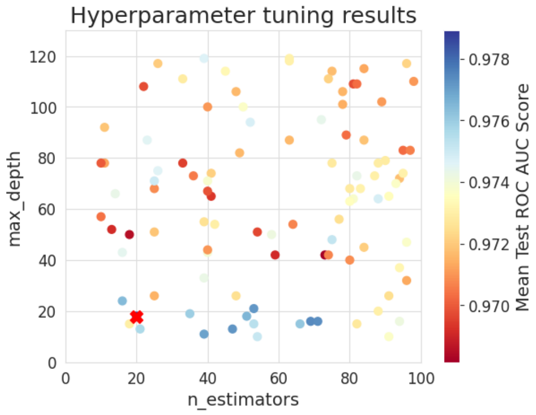


Figure 5.7: The outcomes of a hyperparameter search are presented for the 'n_estimators' and 'max_depth' parameters of the 'RFC_radiomics' model, employing a random grid search. The x-axis illustrates the 'n_estimators' parameter values, while the y-axis represents 'max_depth'. Each scatter point represents a selected hyperparameter combination, with the colour signifying the mean test ROC AUC score achieved through 10-fold cross-validation. The red cross highlights the most effective hyperparameter combination.

5.5.3. Performance Analysis

Using the optimal hyperparameters, a ten-fold cross-validation with one hundred iterations was performed to analyse the performance of both models. The results per iteration are visualized in Figure 5.8 for the SVC_noise model and in Figure 5.9 for the $RFC\_radiomics$ model. Table 5.3 shows the averaged performance scores, including their SD.

The performance of the SVC_noise model demonstrates an average accuracy score of 0.94 in both the training and validation sets. Notably, the ROC AUC metric shows a slightly elevated value of 0.98 in both sets. It is worth highlighting minimal variation across the iterations, with an SD of less than 0.01 for both metrics. On the other hand, the performance of the $RFC\_radiomics$ model surpasses this, achieving scores exceeding 0.99 for both ROC AUC and accuracy in both the training and validation sets. Furthermore, the SD between the repetitions is minimal, measuring less than 0.001.

The impressive performance scores in the training group affirm that both models can efficiently learn from the training data, effectively capturing patterns and relationships within both groups. Furthermore, the strong performance, as evidenced by the high scores on the validation data, suggests that both models can generalize effectively to unseen data.

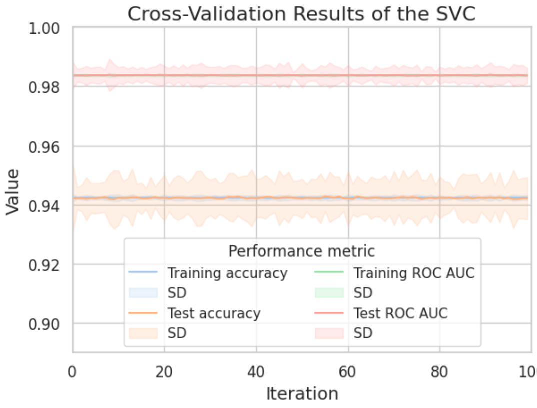


Figure 5.8: The performance results of the SVC using 10-fold cross-validation with 100 iterations, showing the training accuracy, training ROC AUC, test accuracy and test ROC AUC scores for each iteration. Additionally, the SD within each iteration is shown for each performance metric.

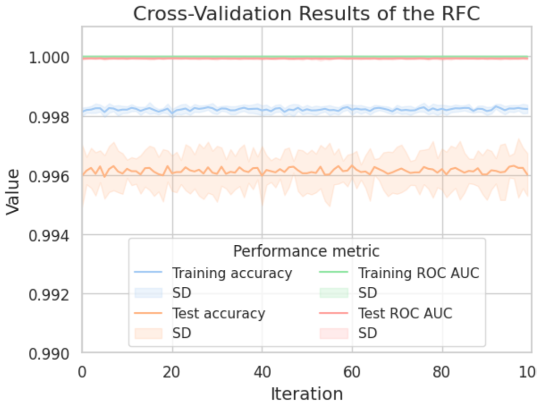


Figure 5.9: The performance results of the RFC using 10-fold cross-validation with 100 iterations, showing the training accuracy, training ROC AUC, test accuracy and test ROC AUC scores for each iteration. Additionally, the SD within each iteration is shown for each performance metric.

Additionally, the low SDs observed across cross-validation repeats indicate that, for both models, their performance remains consistent across different data folds or splits. Consequently, the specific data partitioning into training and test sets does not significantly influence the model's performance. The minimal variability between the iterations underscores the robustness and reliability of the models.

5.5.4. Model Evaluation

The trained models underwent evaluation using the test dataset, which led to the performance scores displayed in Table 5.3. The evaluation of the *SVC_noise* model resulted in a score of 0.94 for both the ROC AUC and the accuracy. Furthermore, the confusion matrix was computed to represent the classification results (Figure 5.10), illustrating the number of correct and incorrect predictions per class. In this case, it shows that 118 smooth slices (4.23%) were incorrectly classified as sharp. Conversely, only 48 slices (1.72%) were misclassified as smooth.

When applying the test dataset to the trained *RFC_radiomics* model, a score of 0.99 was achieved for both the ROC AUC and the accuracy. The confusion matrix of the classification results for the *RFC_radiomics* model is visualized in Figure 5.11. It reveals that 203 (0.73%) smooth patches were wrongly classified as sharp, while only 66 (0.24%) sharp patches were misclassified as smooth.

Notably, in both models, the misclassification rate for smooth slices/patches is higher than that of sharp slices/patches, despite the smaller number of smooth slices/patches used in the input for this model evaluation.

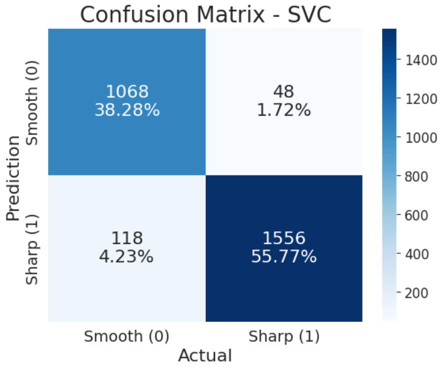


Figure 5.10: The confusion matrix representing the classification results of the *SVC_noise* model. The figure shows the number of correct and incorrect predictions per class of the test dataset.

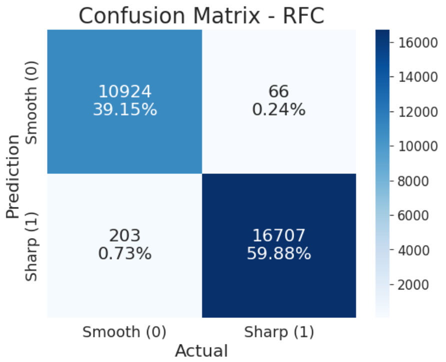


Figure 5.11: The confusion matrix representing the classification results of the *RFC_radiomics* model. The figure shows the number of correct and incorrect predictions per class of the test dataset.

5.6. Model Deployment

The trained models are deployed to classify each data point in the dataset, comprising the test dataset of "extreme" kernels together with all the other available kernels. In total, 270 patients are included with 35 different kernels. The relative frequency of the data classified as smooth and sharp according to each model is visualized using a stacked bar chart (Figure 5.12).

On the left side of this figure, the relative frequencies per kernel of smooth predictions are visualized for both models. The right side shows the relative frequency per kernel of the sharp predictions. The white line serves as the threshold boundary for determining the true label. In the case of light colours (smooth), when the relative frequency lies above this threshold, it is presumed to belong to the smooth category, while if it falls below, it is considered sharp, and vice versa for dark colours (sharp). Figure 5.12 shows that only for one kernel the true class label identified by the threshold boundary differs between the two models; this kernel is the *B50s* kernel from Siemens.

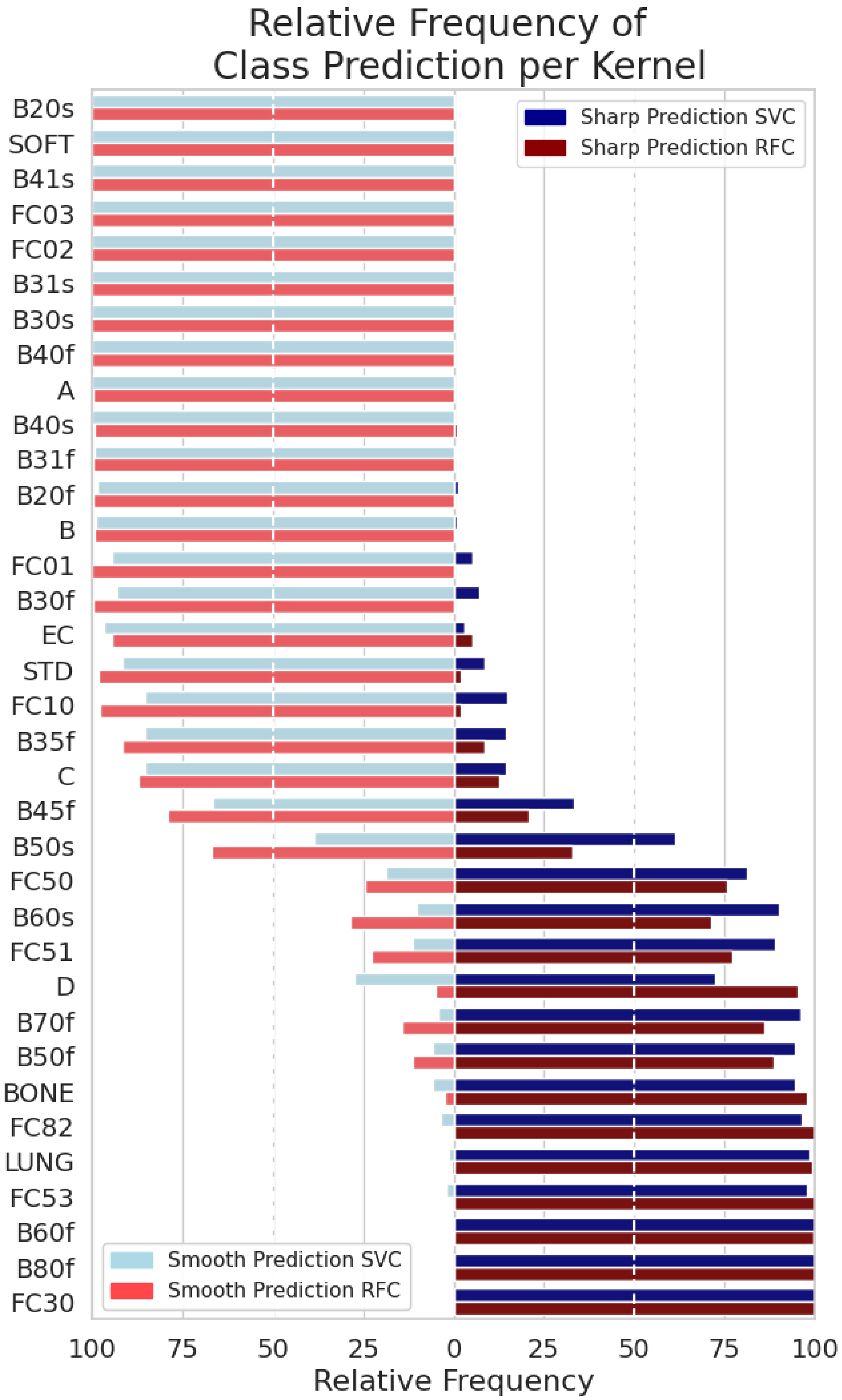


Figure 5.12: Graph of the relative frequencies of the data classified as smooth and sharp according to each model per kernel. A relative frequency of 100 per cent means that all data points of that kernel have been classified as that class. In the graph, the light and dark blue bars represent the predictions of the *SVC_noise* model, whereas the light and dark red bars indicate the predictions made by the *RFC_radiomics* model. On the left side, the relative frequencies of kernels classified as smooth are displayed; on the right side, the relative frequencies of kernels are classified as sharp. The two white dotted lines indicate the threshold boundary for determining the true labels.

5.7. Aggregation

The final model is created by aggregating the data through applying majority voting on patient level; for the *SVC_noise*, only one approach was possible. On the other hand, for the *RFC_radiomics*, three different approaches have been applied for majority voting. The results are shown in Table 5.3.

The majority voting for the *SVC_noise* model resulted in an ROC AUC and accuracy score of 0.97. The model misclassified, according to its definition of smooth and sharp, eight of the 270 patients with five distinct kernels: B45f (n=2), B50f (n=1), B50s (n=2), FC50 (n=1) and D (n=2).

The three majority voting approaches (SLV, MPSLV, HMPV) for the *RFC_radiomics* model all increased performance regarding ROC AUC and accuracy (see Table 5.3). MPSLV yielded the highest performance, yielding an ROC AUC and accuracy score of 0.96.

This approach wrongly classified ten of the 270 patients with eight distinct kernels according to its definition of smooth and sharp: B45f (n=2), B50f (n=1), B50s (n=1), B60s (n=1), B70f (n=1), FC50 (n=1), FC51 (n=2) and C (n=1).

Table 5.3: The performance results of the *SVC_noise* and *RFC_radiomics* models in terms of accuracy and ROC AUC scores. The table shows the performance analysis results, model evaluation and final prediction through aggregation. SLV, MPSLV and HMPV indicate the three different majority approaches explained in Figure 4.3.

Dataset		**Accuracy**		**ROC AUC**	
		SVC_noise	*RFC_radiomics*	*SVC_noise*	*RFC_radiomics*
5.5.3 Performance Analysis					
Training (SD)	Train	0.9423 (7.52E-4)	0.9993 (1.50E-4)	0.9837 (3.14E-4)	1.00 (0.00)
Validation (SD)	Train	0.9422 (65.6E-4)	0.9918 (8.48E-4)	0.9837 (28.2E-4)	0.9996 (1.22E-4)
5.5.4 Model Evaluation					
Testing	Test	0.9405	0.9904	0.9432	0.9910
5.7 Aggregation					
SLV	Test + remaining	0.9703	0.9407	0.9691	0.9370
MPSLV	Test + remaining	N/A	0.9630	N/A	0.9617
HMPV	Test + remaining	N/A	0.9556	N/A	0.9527

5.8. Radiomic Features Analysis

5.8.1. ROI selection: Random versus Most Homogeneous

In the previous experiments, the ROI selection was based on the most homogeneous patches; this section delves into the impact of using randomly selected patches, which are considerably more cost-effective to acquire. The RFC model trained with radiomic features extracted from randomly selected patches is called the *RFC_random* model.

The ten random patches of size 30×30 pixels were selected from each slice. Subsequently, hyperparameter tuning resulted in the best-performing combination of parameters: $n\_estimators$ = 39 and max_depth = 11 with a mean test AUC score of 0.9763.

The performance analysis resulted in the mean performance scores shown in Table 5.4 under the heading *RFC_random*. The model performs well in the training and validation sets, with minimal standard deviation among repeated runs. This indicates that the model's performance is largely unaffected by how the data is divided into training and validation subsets. The minimal variability across iterations underscores the model's robustness and reliability.

Subsequently, the trained model is applied to the test set, and the predictions are evaluated using the true labels. The performance scores are shown in terms of accuracy and ROC AUC in Table 5.4. Additionally, the model is deployed to classify each patch in the test dataset combined with the data of the unseen kernels to determine the true labels following the same procedure as previously described.

Last, majority voting has been applied using the three proposed approaches; the results are noted in Table

5.4. SLV and MPSLV achieved the same high-performance level, attaining an accuracy and ROC AUC score of 0.97. Both these approaches resulted in the misclassification of nine out of the 270 patients with seven distinct kernels: B45f (n=2), B50f (n=1), B50s (n=1), FC50 (n=1), FC51 (n=2), C (n=1) and D (n=1).

The performances of both models following data aggregation with MPSLV (best-performing approach) showed no significant difference, as the p-value of the McNemar test performed using the contingency table [[259,2], [1,8]] is 1.0. The p-value of 1.0 indicates that there is no evidence to reject the null hypothesis, which, in this context, would mean that the two models perform similarly in terms of misclassification.

Also, the categorization of the kernels is equal to that of the *RFC_radiomics* model. However, there are a few interesting differences. Firstly, in contrast to the *RFC_radiomics* model, MPSLV did not increase the model's performance compared to SLV. Furthermore, the ROI selection process of *RFC_random* is computationally less heavy. Finally, the training and validation results are slightly lower and have more variation (higher SD); this may indicate less overfitting compared to the *RFC_radiomics* model.

Table 5.4: The performance results of the *RFC_radiomics* and *RFC_random* models in terms of accuracy and ROC AUC scores. The table shows the performance analysis results, model evaluation and final prediction through aggregation. SLV, MPSLV and HMPV indicate the three different majority voting approaches explained in Figure 4.3.

	Dataset	**Accuracy**		**ROC AUC**	
		RFC_random	*RFC_radiomics*	*RFC_random*	*RFC_radiomics*
5.5.3 Performance Analysis					
Training (SD)	Train	0.9758 (2.94E-4)	0.9993 (1.50E-4)	0.9986 (0.62E-4)	1.00 (0.00)
Validation (SD)	Train	0.9684 (16.0E-4)	0.9918 (8.48E-4)	0.9961 (3.58E-4)	0.9996 (1.22E-4)
5.5.4 Model Evaluation					
Testing	Test	0.9649	0.9904	0.9687	0.9910
5.7 Aggregation					
SLV	Test + remaining	0.9667	0.9407	0.9659	0.9370
MPSLV	Test + remaining	0.9667	0.9630	0.9659	0.9617
HMPV	Test + remaining	0.9411	0.9556	0.9444	0.9527

5.8.2. Feature Importance

Permutation importance was assessed using the test set; the top 20 features are shown in Figure 5.13, revealing that the feature with the highest permutation importance yielded a modest decrease of 0.000534 in accuracy when its values were permuted. This result suggests that, on average, permuting this particular feature led to a negligible impact on model performance, implying that none of the features are individually important. However, this finding contradicts the model's high performance, indicating that there must be features of significance.

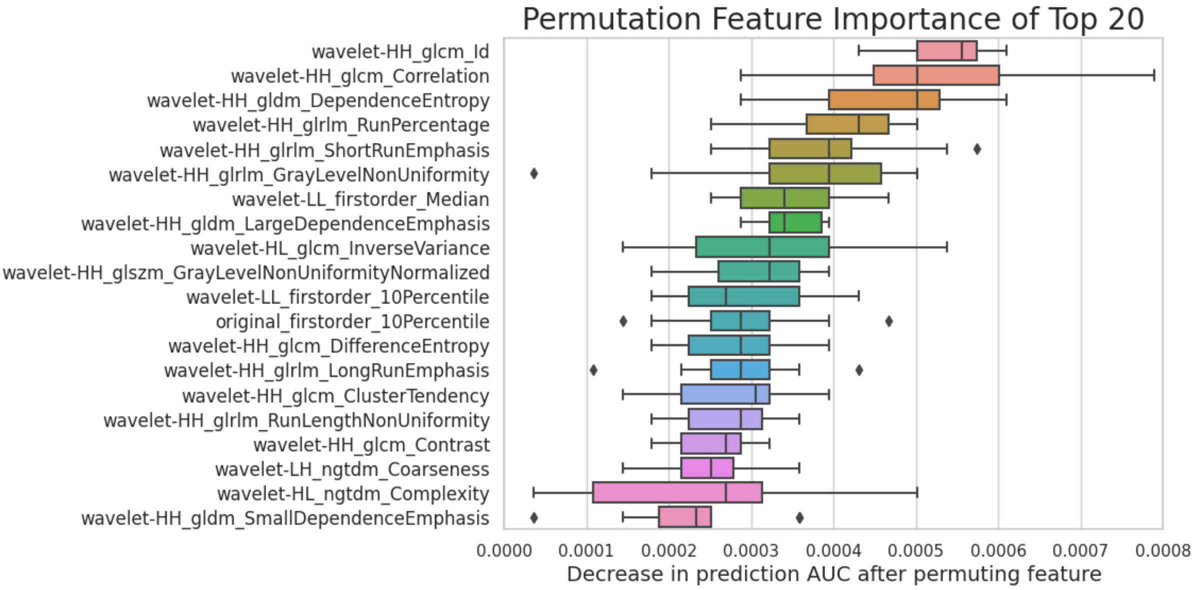
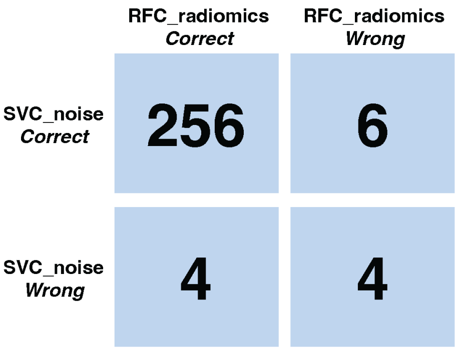


Figure 5.13: The distribution of the permutation importance values of the top 20 features, visualized using box plots. The black dots indicate the outliers. The permutation importance values are defined as the decrease in prediction ROC AUC score following permuting that specific feature. Each feature name is composed as **filteringtype_featureclass_feature**. All the features stated here are extracted from the wavelet-filtered images.

5.9. Model Comparison

5.9.1. McNemar's Test

The best-performing models for the comparison are used; for SVC_noise the model applying SLV, and for $RFC\_radiomics$ the model applying MPSLV. This created the 2x2 contingency table of [[256,6], [4,4]] shown in Figure 5.14 and resulted in a p-value of 0.75. The relatively large p-value of 0.75 suggests that the null hypothesis that none of the two models performs better than the other can not be rejected. Given these results, there is no sufficient statistical evidence to conclude that one classifier model significantly outperforms the other.



	RFC_radiomics ***Correct***	**RFC_radiomics** ***Wrong***
SVC_noise ***Correct***	**256**	**6**
SVC_noise ***Wrong***	**4**	**4**

Figure 5.14: The 2×2 contingency table using the SVC_noise and $RFC\_radiomics$ binary classification outcomes.

5.9.2. Differences in Categorization

The true class label for each unseen kernel in each model is determined by the majority class within the kernel. In some cases, this leads to inconsistent true class labels between the two models, resulting in different kernel categorizations. When comparing these true class labels (see Appendix F), we found only one kernel, $B50s$, categorized differently. The SVC_noise model assigns it the true class label "sharp", while the $RFC\_radiomics$ model labels it "smooth". Siemens describes $B50s$ as a medium sharp kernel with "s" denoting standard rotation time.

Changing the true class label for $B50s$ to "smooth" in the SVC_noise model does not affect its performance since it equally classifies the kernel as "smooth" and "sharp". However, in the $RFC\_radiomics$ model with MPSLV, altering the true class label for $B50s$ would lead to a slight performance decrease of less than 0.1 in accuracy and ROC AUC score. This is because the model predominantly classifies the kernel as "smooth" (three times) compared to "sharp" (once).

5.9.3. Misclassified Patients

Eight patients have been misclassified by the SVC_noise model and ten by the $RFC\_radiomics$ model out of 270 patients, both by their own definitions of sharp and smooth categorization. Four of those patients are misclassified by both models; their specifications, together with the median values for that kernel, are outlined in Table 5.5. The patients are reconstructed with three distinct kernels: $B45f$, $B50f$ and $FC50$, which are all on the border of smooth and sharp, as shown in Figure 5.12.

Table 5.5: The four cases misclassified by both models, including their specifications in terms of manufacturer, CT scanner model, kernel, tube current and slice thickness compared with the median tube current and slice thickness values of that specific kernel.

	Manufacturer	**Model**	**Kernel**	**Tube current** (mA)		**Slice thickness** (mm)	
				Patient	Median	Patient	Median
1.	SIEMENS	Volume Zoom	B45f	140	417	2.0	1.0
2.	SIEMENS	Sensation 16	B45f	120	417	1.0	1.0
3.	SIEMENS	Sensation 16	B50f	90	120	5.0	2.0
4.	TOSHIBA	Aquilion	FC50	140	100	2.0	2.0

By comparing these four patients with the other patients of the same kernels in Appendix C, to see if the other scan parameters potentially explain the misclassifications, some notable observations emerge:

1st Patient: *Patient reconstructed with $B45f$ misclassified as 'sharp'.*
This patient is the sole instance of a $B45f$ case acquired with a Volume Zoom scanner model. Furthermore, it has a slice thickness of 2.0 mm, unlike most $B45f$ cases, with a 1.0 mm slice thickness. This suggests the presence of less noise and reduced sharpness, creating a smoother appearance compared to cases with a 1.0 mm slice thickness. Additionally, it exhibits one of the lowest tube current values, typically resulting in increased image noise and reduced sharpness. From the investigated scan parameters, only the increased image noise caused by the low tube current may explain the misclassification of this case.

2nd Patient: *Another patient reconstructed with $B45f$ misclassified as 'sharp'*
This case is acquired with a Sensation 16 model, similar to the majority of $B45f$ cases. The only noteworthy difference in acquisition parameters compared to the other cases is using the lowest tube current value. The increased image noise resulting from this low value might have contributed to the misclassification of this patient.

3rd Patient: *Patient reconstructed with $B50f$ misclassified as 'smooth'*
This could be attributed to the exceptionally high slice thickness of 5.0 mm, whereas the others, except one, all have a slice thickness of 2.0 mm. Higher slice thickness is typically associated with reduced sharpness due to increased partial volume effects, potentially explaining the misclassification. Nevertheless, an increased slice thickness also results in lower noise levels because more X-ray photons contribute to each slice.

4th Patient: *Patient reconstructed with $FC50$ misclassified as 'smooth'*

This case does not substantially differ in scan parameters compared to the other $FC50$ cases. The only notable parameter distinction is that this patient is acquired with the lowest tube current value compared to the other patients within this kernel. This could potentially explain the misclassification since a lower tube current value increases image noise.

6 Discussion

The main objective of this thesis was to develop a method for categorising reconstruction kernels from different vendors based on their sharpness. This approach involves using scan characteristics directly extracted from actual patient scans with varying parameters. Two distinct methods were investigated to achieve this objective, and several research questions were formulated to guide the research process, all of which will be addressed below.

The first research question aims to research if scan characteristics directly extracted from actual patient scans can facilitate the categorisation of reconstruction kernels and is answered using several sub-questions, all shown below.

Research (Sub-)Question(s) 1

Can scan characteristics directly extracted from actual patient scans facilitate the categorisation of reconstruction kernels?

1.1. Which image features can effectively enable kernel categorisation?

1.2. Which ML models are suitable for utilising these image features for categorisation?

1.3. What is the comparative performance of the ML models using the identified image features in the context of kernel categorisation?

In this thesis, two distinct sets of image features were extracted from CT patient scans: noise features and radiomic features. The noise features extracted from the CT examines in the extremes dataset were utilised as input to train, validate, and evaluate the SVC_noise model. This model achieved an average ROC AUC score of 0.9837 (28.2E-4) in the validation group and a score of 0.9432 in the testing group. Similarly, the radiomic features from the same extremes dataset were used to train, validate, and evaluate the $RFC\_radiomics$ model. In the validation group, this model obtained an average ROC AUC score of 0.9996 (1.22E-4), while the testing group achieved a score of 0.9910.

These outcomes demonstrate that both sets of image features facilitate the categorisation of kernels found in the extremes dataset. They exhibit impressive ROC AUC performance in the testing groups, signifying the effective generalisation of the models to unseen data. Moreover, the SVC proves to be a suitable machine learning classification model for the noise features, while the RFC is aptly suited for the radiomic features.

Following this, the two models were deployed to categorise smooth and sharp kernels. The test dataset, combined with the remaining dataset containing kernels not classified as extremes, was used for this deployment. Given the absence of ground truth data for the remaining dataset, evaluating this deployment posed a significant challenge. To address this, we determined the true class label for each kernel by identifying the predominant class within that kernel, aiming to address the second research question:

Research Question 2

How can the kernel categorisation be effectively evaluated when ground truth data is unavailable?

The true class labels were determined for both models using the previously mentioned method. This resulted in a nearly consistent categorisation of the true labels for most kernels between the two models, except for kernel $B50s$, which had a different true class label. Subsequently, a final prediction was made for each patient by aggregating the predictions for that patient. The final predictions were compared to the established true class labels. The SVC_noise model applied aggregation approach SLV, which yielded an ROC AUC score of 0.97, per the model's definition of smooth and sharp. In total, the model misclassified eight out of the 270 patients. On the other hand, MPSLV performed best for the $RFC\_radiomics$ model, achieving a ROC AUC score of 0.96, following the model's definition of smooth and sharp. This model misclassified ten out of the 270 included patients.

In the following sections, the chapter presents research findings and critically reflects on the research process. It delves into the limitations and potential complications of the study's design and their implications for result interpretation. Furthermore, the chapter offers several suggestions for future research.

6.1. Dataset

The dataset used in this research comprises thoracic CT scans from the NLST and LIDC-IDRC open-source datasets. It encompasses a diverse range of employed scan parameters, and the exceptional performance of both models suggests their ability to discern patterns associated with both classes accurately. Moreover, the models demonstrate robustness against the noise introduced by variations in scan parameters and patient characteristics. Notably, these models consistently classify previously unseen kernels into the same categories within each model and across both models, with the sole exception being the classification of the $B50s$ kernel, which differs between the two models.

For the implementation of the models, a specific subset of the dataset was chosen (the "extremes dataset"). This subset consists of each manufacturer's smoothest and sharpest kernels available in the NLST and LIDC-IDRC datasets. However, the distribution analysis has unveiled an interesting observation: the extreme kernels selected for analysis do not consistently correspond to those with the lowest and highest values in the dataset's noise magnitude and texture distribution. Despite this inconsistency, it is essential to recognise that employing these extreme cases has been empirically validated as reasonable. The models consistently demonstrate their ability to generalise effectively when applied to unseen kernels. This indicates the robustness of the models and the practicality of using these extreme kernels for the intended categorisation tasks.

Furthermore, all scans within the dataset were reconstructed using the outdated FBP technique due to their acquisition predating 2010. Consequently, the extent to which the proposed method can be applied to more advanced and intricate reconstruction techniques, such as iterative and deep learning-based methods, is left for future work.

It is also relevant to highlight that the dataset exclusively comprises thoracic CT scans and reconstruction kernels suitable for thoracic scans. This limitation confines the generalizability of the models to other body parts and reconstruction techniques. Moreover, the dataset is imbalanced, with a predominant representation of the sharp class in contrast to the smooth class. This imbalance likely affects the heightened misclassification rate observed for smooth slices/patches. The models' limited ability to categorise the smooth class stems from insufficient training data, a direct outcome of this class imbalance.

For future research, several paths are worth exploring. Deploying the models on a larger dataset with the same kernels could help confirm their performance and categorisations. Additionally, extending the application of the models to datasets reconstructed with other unseen kernels, reconstruction techniques, and CT scans of different body parts can help ascertain the models' consistency and ability to identify precise categorisations. Training the models on different kernels identified as extremes and evaluating potential differences in categorisation distributions between the models could offer insights into the models' adaptability and performance. Lastly, addressing the issue of class imbalance by training the models on a balanced dataset may improve the misclassification rate for smooth classes and enhance overall model performance.

6.2. Noise Features

The analysis of the noise features extracted from each slice showed a notable pattern where the noise values of kernels assumed to be smoother are distributed closely together, thus having a smaller deviation. Conversely, kernels with higher sharpness displayed greater variability and occasionally featured outliers within the distribution of these measures. Moreover, the relationship between CF and SD showed a strong, positive, and statistically significant monotonic relationship supported by the Spearman's correlation coefficient yielding a value of 0.79 (p<0.001), confirming the anticipated relationship as has previously been demonstrated [55]. In addition, the relationship between the two features reveals two linearly separable clusters, which well-founded the use of the linear classifier to classify the reconstruction kernels.

In this thesis, the automated technique developed by Anam et al. [56] was employed to assess the noise magnitude in CT images. This method effectively distinguishes differences in noise magnitude resulting from reconstruction kernels when combined with noise texture values as input for a linear SVC model. While using a 30×30 pixel ROI has been demonstrated as adequate for this purpose, it is important to acknowledge that its optimality is still uncertain. Further research is required to explore and determine the most optimal ROI size in this context.

In the context of real patient scans, the applicability of NPS encounters notable challenges owing to the inherent non-stationary nature of such imaging systems, as highlighted by Dainty et al. [58]. For CT scans reconstructed using FBP, the assumption may be that CT noise exhibits local stationarity within a small ROI situated in a uniform background. This assumption underpinned the selection of a compact 30-pixel ROI extracted from the most homogenous patch recommended in the literature [77]. Nevertheless, it is essential to acknowledge that this assumption may not hold for nonlinear IR algorithms, which can manifest highly non-stationary noise patterns when anatomical structures are present [85]. Consequently, further research is needed to ascertain the suitability of the CF as an image feature for the categorisation model in the context of CT scans reconstructed using IR algorithms.

6.3. Radiomic Features

In comparing random ROI selection and utilising the most homogeneous patches, noteworthy findings emerged. The model trained on random patches demonstrated slightly superior performance, achieving an ROC AUC score of 0.97 compared to 0.96 for the homogeneous ROI selection. However, the improvement observed was not statistically significant according to the McNemar test. This suggests that radiomic features extracted from both homogeneous and random patches can effectively serve as input for an RFC when categorising reconstruction kernels. The random ROI selection might be preferred due to its computational efficiency. Additionally, the RFC_random model exhibited identical performance results for the SLV and MPSLV approaches, indicating that extracting ten patches per slice could be considered redundant, potentially enhancing computational efficiency. Further research is warranted to validate this finding and determine the optimal RFC performance configuration.

Furthermore, the analysis of permutation feature importance sheds light on the model's robustness against disturbances in individual features. This resilience is likely attributed to the presence of collinearity among the features. Perturbing a single feature had a negligible impact on the model's performance, as it could still derive similar information from correlated features. To address the challenge of multicollinearity, future research could explore techniques such as hierarchical clustering based on Spearman rank-order correlations [86]. Establishing a correlation threshold and retaining only one representative feature from each cluster could help mitigate redundancy and ensure that the selected features genuinely contribute to the model's effectiveness.

Finally, the selection of a 30-pixel ROI size for radiomic feature extraction was made arbitrarily, as there is a lack of literature-based recommendations regarding the ideal ROI size. Therefore, this aspect presents an intriguing avenue for future research exploration.

6.4. Model Development

Both the RFC model employed for radiomic features and the SVC model utilised for noise features showed exceptional consistency in their predictions when classifying reconstruction kernels within the smooth and sharp groups. This success underscores the suitability of these specific ML algorithms for this task. In addition, the models were trained on a subset of kernels known as the 'extremes', yet they have demonstrated their adaptability to other kernels, indicating their potential for generalisation to unseen data. This adaptability further underlines their utility in real-world scenarios.

Nevertheless, it is worth highlighting that this research did not encompass an exhaustive comparison of a wide array of machine learning models, which could be an intriguing avenue for future exploration. Such a comprehensive comparative analysis of diverse machine learning models would provide valuable insights into how the RFC and SVC models stack up against alternative methods. This comparative study could evaluate factors like accuracy, computational efficiency, and the robustness of different models, offering a more comprehensive understanding of their respective strengths and weaknesses.

Moreover, this thesis primarily focuses on machine learning models, but it is worth acknowledging the potential benefits of exploring DL models in this context. DL models have shown considerable promise in image classification tasks [87]. By leveraging DL models, researchers could achieve even more refined and nuanced results in categorising reconstruction kernels. This research direction opens up a wide range of possibilities, including applying convolutional layers to automatically learn relevant features from the image data, ultimately enhancing the accuracy and generalizability of the classification model.

The model performance evaluation relies mainly on the ROC AUC score. However, this score does not always provide a complete picture of a model's performance. While it effectively quantifies a model's ability to discriminate between different classes, it may not reflect the overall classification performance. In cases where the class distribution is imbalanced, which is the case in this research, the ROC AUC score can appear favourable, even if the model struggles with accurate classification within the minority class. To mitigate this limitation, incorporating other metrics such as precision, recall, and F1 score can offer a more comprehensive evaluation of model performance, mainly when class imbalances exist. These metrics can be particularly informative when one class is more prevalent than the other.

Furthermore, determining the true class labels for each kernel relies on identifying the predominant class within each kernel group. This operation inherently depends on the model's performance and may introduce a degree of bias into the categorisation results. Consequently, the ROC AUC scores used to assess the model's performance can be regarded as measures of how well the model aligns with its own predictions, rather than indicators of its overall performance in classification. This is because it cannot be guaranteed that the assumed true classes are correct. Nevertheless, it should be noted that the differences in the assigned true class labels between the two models, specifically SVC_noise and $RFC\_radiomics$, as well as the RFC_random model, are limited. The sole exception is the kernel $B50s$, which receives a different true class label based on the models. This observation suggests that, for most kernels, all three models consistently identify similar patterns associated with both classes. This alignment in the models' outputs supports the reliability of the true class labels assigned in this context.

Lastly, categorising kernels solely into two broad groups, "sharp" and "smooth", raises a valid point of debate. This binary classification may oversimplify the inherent complexity of kernel characteristics, potentially leading to a loss of important details and nuances in the data. To address this issue, the consideration of introducing an intermediate sharpness category is a concept worth investigating more thoroughly. The rationale behind this suggestion lies in the recognition that kernels can exhibit a spectrum of sharpness levels rather than fitting neatly into just two extreme categories. Introducing an intermediate category could create a more finely-grained classification system. This intermediate category would allow kernel characterisation with characteristics falling between those typically associated with sharp and smooth kernels, such as kernel $B50s$.

6.5. Conclusion

In conclusion, this thesis aimed to develop an ML-based method for categorising reconstruction kernels based on their sharpness. The *SVC_noise* and *RFC_radiomics* models demonstrated promising performances, neither outperforming the other. The ability of both models to accurately discern patterns associated with the sharpness of each class while ignoring the noise introduced by variation in scan parameters and patient characteristics in real patient data provides valuable insights, bridging the gap between research and clinical applications. The results point towards the feasibility of scan characteristics extracted from real patient scans in combination with ML models, addressing the challenge of kernel categorisation but also providing practical, versatile, and efficient tools that can benefit the broader medical imaging community. Nevertheless, the results of this research are still preliminary, and caution is warranted when extrapolating the observed results to broader contexts, such as newer reconstruction kernels and techniques.

References

[1] Shah Hussain et al. “Modern diagnostic imaging technique applications and risk factors in the medical field: A review”. In: *BioMed Research International* 2022 (2022).

[2] Xin-ying Xue et al. “Computed tomography for the diagnosis of solitary thin-walled cavity lung cancer”. In: *The Clinical Respiratory Journal* 9.4 (2015), pp. 392–398.

[3] Errol Levine et al. “Comparison of computed tomography and other imaging modalities in the evaluation of musculoskeletal tumors”. In: *Radiology* 131.2 (1979), pp. 431–437.

[4] Sumaira Naz et al. “COVID-19 and SARS-CoV-2: everything we know so far–a comprehensive review”. In: *Open Chemistry* 19.1 (2021), pp. 548–575.

[5] Minsoo Chun et al. “Fully automated image quality evaluation on patient CT: Multi-vendor and multi-reconstruction study”. In: *PloS one* 17.7 (2022), e0271724.

[6] J. Greffier et al. “CT iterative reconstruction algorithms: a task-based image quality assessment”. In: *European Radiology* 30 (1 Jan. 2020), pp. 487–500. ISSN: 14321084. DOI: 10.1007/S00330-019-06359-6/FIGURES/6.

[7] Stephan P Blazis et al. “Effect of CT reconstruction settings on the performance of a deep learning based lung nodule CAD system”. In: *European Journal of Radiology* 136 (2021), p. 109526.

[8] R. Schofield et al. “Image reconstruction: Part 1 understanding filtered back projection, noise and image acquisition”. In: *Journal of Cardiovascular Computed Tomography* 14 (3 May 2020), pp. 219–225. ISSN: 1934-5925. DOI: 10.1016/J.JCCT.2019.04.008.

[9] Gengsheng L Zeng. “Revisit of the ramp filter”. In: *2014 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*. IEEE. 2014, pp. 1–6.

[10] Lucas L. Geyer et al. “State of the Art: Iterative CT reconstruction techniques1”. In: *Radiology* 276 (2 Aug. 2015), pp. 339–357. ISSN: 15271315. DOI: 10.1148/RADIOL.2015132766/ASSET/IMAGES/LARGE/RADIOL.2015132766.FIG13.JPEG.

[11] Barbaros S Erdal et al. “Are quantitative features of lung nodules reproducible at different CT acquisition and reconstruction parameters?” In: *PLoS One* 15.10 (2020), e0240184.

[12] Nastaran Emaminejad et al. “Reproducibility of lung nodule radiomic features: multivariable and univariable investigations that account for interactions between CT acquisition and reconstruction parameters”. In: *Medical physics* 48.6 (2021), pp. 2906–2919.

[13] Sarah Denzler et al. “Impact of CT convolution kernel on robustness of radiomic features for different lung diseases and tissue types”. In: *The British journal of radiology* 94.1120 (2021), p. 20200947.

[14] Binsheng Zhao et al. “Exploring variability in CT characterization of tumors: a preliminary phantom study”. In: *Translational oncology* 7.1 (2014), pp. 88–93.

[15] Guangyao Wu et al. “The emerging role of radiomics in COPD and lung cancer”. In: *Respiration* 99.2 (2020), pp. 99–107.

[16] Dennis Mackin et al. “Matching and homogenizing convolution kernels for quantitative studies in computed tomography”. In: *Investigative radiology* 54.5 (2019), p. 288.

[17] Justin B Solomon, Olav Christianson, and Ehsan Samei. “Quantitative comparison of noise texture across CT scanners from different manufacturers”. In: *Medical physics* 39.10 (2012), pp. 6048–6055.

[18] K Eldevik, W Nordhøy, and A Skretting. “Relationship between sharpness and noise in CT images reconstructed with different kernels”. In: *Radiation protection dosimetry* 139.1-3 (2010), pp. 430–433.

[19] Euclid Seeram. *Computed Tomography-E-Book: Physical Principles, Patient Care, Clinical Applications, and Quality Control*. Elsevier Health Sciences, 2022.

[20] Jiang Hsieh. “Computed tomography: principles, design, artifacts, and recent advances”. In: (2003).

[21] Lee W Goldman. “Principles of CT: radiation dose and image quality”. In: *Journal of nuclear medicine technology* 35.4 (2007), pp. 213–225.

[22] PF Judy. “CT image quality and parameters affecting the CT image”. In: *Goldman LW, Fowlkes JB, eds* (2000).

[23] Atul Padole et al. “CT radiation dose and iterative reconstruction techniques”. In: *AJR Am J Roentgenol* 204.4 (2015), W384–W392.

[24] Choirul Anam et al. “New noise reduction method for reducing CT scan dose: Combining Wiener filtering and edge detection algorithm”. In: *AIP Conference Proceedings*. Vol. 1677. 1. AIP Publishing. 2015.

[25] Martti Kalke and Samuli Siltanen. “Sinogram interpolation method for sparse-angle tomography”. In: *Applied Mathematics* 2014 (2014).

[26] Sebastian Schafer and Jeffrey H Siewerdsen. “Technology and applications in interventional imaging: 2D X-ray radiography/fluoroscopy and 3D cone-beam CT”. In: *Handbook of Medical Image Computing and Computer Assisted Intervention*. Elsevier, 2020, pp. 625–671.

[27] Rupert CD Young and Christopher R Chatwin. “Computation of the forward and inverse Radon transform via the central slice theorem employing a nonscanning optical technique”. In: *Optical Pattern Recognition VII*. Vol. 2752. SPIE. 1996, pp. 306–316.

[28] Martin J. Willemink and Peter B. Noël. “The evolution of image reconstruction for CTfrom filtered back projection to artificial intelligence”. In: *European Radiology* 29 (5 May 2019), p. 2185. ISSN: 14321084. DOI: 10.1007/S00330-018-5810-7.

[29] Wolfram Stiller. “Basics of iterative reconstruction methods in computed tomography: A vendor-indepe-ndent overview”. In: *European Journal of Radiology* 109 (Dec. 2018), pp. 147–154. ISSN: 0720-048X. DOI: 10.1016/J.EJRAD.2018.10.025.

[30] Timothy P. Szczykutowicz et al. “A Review of Deep Learning CT Reconstruction: Concepts, Limitations, and Promise in Clinical Practice”. In: *Current Radiology Reports* 10 (9 Sept. 2022), pp. 101–115. ISSN: 21674825. DOI: 10.1007/S40134-022-00399-5/FIGURES/8.

[31] TM Peters. “Algorithms for fast back-and re-projection in computed tomography”. In: *IEEE transactions on nuclear science* 28.4 (1981), pp. 3641–3647.

[32] James Anthony Seibert. “Iterative reconstruction: how it works, how to apply it”. In: *Pediatric radiology* 44 (2014), pp. 431–439.

[33] Sonja Gordic et al. “Optimizing radiation dose by using advanced modelled iterative reconstruction in high-pitch coronary CT angiography”. In: *European radiology* 26 (2016), pp. 459–468.

[34] Andrew D Hardie et al. “What is the preferred strength setting of the sinogram-affirmed iterative reconstruction algorithm in abdominal CT imaging?” In: *Radiological physics and technology* 8 (2015), pp. 60–63.

[35] Lu Liu. “Model-based iterative reconstruction: a promising algorithm for today’s computed tomography imaging”. In: *Journal of Medical imaging and Radiation sciences* 45.2 (2014), pp. 131–136.

[36] A. Löve et al. “Six iterative reconstruction algorithms in brain CT: A phantom study on image quality at different radiation dose levels”. In: *British Journal of Radiology* 86 (1031 Nov. 2013). ISSN: 00071285. DOI: 10.1259/BJR.20130388.

[37] Marcel Beister, Daniel Kolditz, and Willi A. Kalender. “Iterative reconstruction methods in X-ray CT”. In: *Physica Medica* 28 (2 Apr. 2012), pp. 94–108. ISSN: 11201797. DOI: 10.1016/j.ejmp.2012.01.003.

[38] Ziyu Zhang and Euclid Seeram. “The use of artificial intelligence in computed tomography image reconstruction-a literature review”. In: *Journal of medical imaging and radiation sciences* 51.4 (2020), pp. 671–677.

[39] Clemens Arndt et al. “Deep learning CT image reconstruction in clinical practice”. In: *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 193. 03. Georg Thieme Verlag KG. 2021, pp. 252–261.

[40] D Mehta et al. “Iterative model reconstruction: simultaneously lowered computed tomography radiation dose and improved image quality”. In: *Med Phys Int J* 2.1 (2013), pp. 147–55.

[41] *Revolution CT User Manual*. URL: https://www.manualslib.com/manual/1304880/Ge-Revolution-Ct.html?page=294%5C#manual).

[42] *Brilliance CT Operation Manual*. URL: https://www.scribd.com/document/488973555/Operation-Manual-pdf.

[43] *Siemens SOMATOM Sensation 64 Open Application Guide*. URL: https://www.scribd.com/document/407313712/Simens-SOMATOM-Sensation-64-user-manual-pdf.

[44] *Toshiba Aquilion 16 Basic Operation Manual*. URL: https://www.scribd.com/document/355514488/2B201-313E-B-Aquilion16PC-Basic-Operation-pdf.

[45] Young Jae Kim et al. “The effect of CT scan parameters on the measurement of CT radiomic features: a lung nodule phantom study”. In: *Computational and Mathematical Methods in Medicine* 2019 (2019).

[46] Dennis Mackin et al. “Effect of tube current on computed tomography radiomic features”. In: *Scientific reports* 8.1 (2018), p. 2354.

[47] Batta Mahesh. “Machine learning algorithms-a review”. In: *International Journal of Science and Research (IJSR).[Internet]* 9.1 (2020), pp. 381–386.

[48] Marti A. Hearst et al. “Support vector machines”. In: *IEEE Intelligent Systems and their applications* 13.4 (1998), pp. 18–28.

[49] Shan Suthaharan and Shan Suthaharan. “Support vector machine”. In: *Machine learning models and algorithms for big data classification: thinking with examples for effective learning* (2016), pp. 207–235.

[50] Huan Xu, Constantine Caramanis, and Shie Mannor. “Robustness and Regularization of Support Vector Machines.” In: *Journal of machine learning research* 10.7 (2009).

[51] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[52] Vladimir Svetnik et al. “Random forest: a classification and regression tool for compound classification and QSAR modeling”. In: *Journal of chemical information and computer sciences* 43.6 (2003), pp. 1947–1958.

[53] Xue Ying. “An overview of overfitting and its solutions”. In: *Journal of physics: Conference series*. Vol. 1168. IOP Publishing. 2019, p. 022022.

[54] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. “Hyperparameters and tuning strategies for random forest”. In: *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9.3 (2019), e1301.

[55] Tinsu Pan et al. “impact on central frequency and noise magnitude ratios by advanced CT image reconstruction techniques”. In: *Medical physics* 47.2 (2020), pp. 480–487.

[56] Choirul Anam et al. “An improved method of automated noise measurement system in CT images”. In: *Journal of Biomedical Physics & Engineering* 11.2 (2021), p. 163.

[57] FR Verdun et al. “Image quality in CT: From physical measurements to model observers”. In: *Physica Medica* 31.8 (2015), pp. 823–843.

[58] John Christopher Dainty, Rodney Shaw, and LJ Cutrona. “Image Science: Principles, analysis and evaluation of photographic-type imaging processes”. In: *Physics Today* 29.1 (1976), pp. 71–72.

[59] Kirsten L Boedeker, Virgil N Cooper, and Michael F McNitt-Gray. “Application of the noise power spectrum in modern diagnostic MDCT: part I. Measurement of noise power spectra and noise equivalent quanta”. In: *Physics in medicine & biology* 52.14 (2007), p. 4027.

[60] Laura J Jensen et al. “Stability of radiomic features across different region of interest sizesA CT and MR phantom study”. In: *Tomography* 7.2 (2021), pp. 238–252.

[61] Francesca Ng et al. “Assessment of tumor heterogeneity by CT texture analysis: can the largest cross-sectional area be used as an alternative to whole tumor analysis?” In: *European journal of radiology* 82.2 (2013), pp. 342–348.

[62] Philippe Lambin et al. “Radiomics: the bridge between medical imaging and personalized medicine”. In: *Nature reviews Clinical oncology* 14.12 (2017), pp. 749–762.

[63] Muhammad Shafiq-ul-Hassan et al. “Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra”. In: *Journal of Medical Imaging* 5.1 (2018), pp. 011013–011013.

[64] Stefania Rizzo et al. “Radiomics: the facts and the challenges of image analysis”. In: *European radiology experimental* 2.1 (2018), pp. 1–8.

[65] Robert M Haralick, Karthikeyan Shanmugam, and Its’ Hak Dinstein. “Textural features for image classification”. In: *IEEE Transactions on systems, man, and cybernetics* 6 (1973), pp. 610–621.

[66] Gary Ge and Jie Zhang. “Feature selection methods and predictive models in CT lung cancer radiomics”. In: *Journal of Applied Clinical Medical Physics* 24.1 (2023), e13869.

[67] Andrew Laine and Jian Fan. “Texture classification by wavelet packet signatures”. In: *IEEE Transactions on pattern analysis and machine intelligence* 15.11 (1993), pp. 1186–1191.

[68] Hugues Benoit-Cattin. *Texture analysis for magnetic resonance imaging*. Texture Analysis Magn Resona, 2006.

[69] Mohammad Hossin and Md Nasir Sulaiman. “A review on evaluation metrics for data classification evaluations”. In: *International journal of data mining & knowledge management process* 5.2 (2015), p. 1.

[70] Alex J Bowers and Xiaoliang Zhou. “Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes”. In: *Journal of Education for Students Placed at Risk (JESPAR)* 24.1 (2019), pp. 20–46.

[71] Jan De Leeuw et al. “Comparing accuracy assessments to infer superiority of image classification methods”. In: *International Journal of Remote Sensing* 27.1 (2006), pp. 223–232.

[72] National Lung Screening Trial Research Team. “The national lung screening trial: overview and study design”. In: *Radiology* 258.1 (2011), pp. 243–253.

[73] Samuel G Armato III et al. “The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans”. In: *Medical physics* 38.2 (2011), pp. 915–931.

[74] J Ambrose. “Computerized trans-verse axial tomography”. In: *Brit. J. Radiol.* 46 (1973), p. 148.

[75] Saul N Friedman et al. “A simple approach to measure computed tomography (CT) modulation transfer function (MTF) and noise-power spectrum (NPS) using the American College of Radiology (ACR) accreditation phantom”. In: *Medical physics* 40.5 (2013), p. 051907.

[76] International Electrotechnical Commission et al. “Characteristics of digital x-ray imaging devices–Part 1: Determination of the detective quantum efficiency”. In: *Medical electrical equipment IEC* (2003), pp. 62220–1.

[77] Steven Dolly et al. “Practical considerations for noise power spectra estimation for clinical CT scanners”. In: *Journal of applied clinical medical physics* 17.3 (2016), pp. 392–407.

[78] Yoshinori Funama et al. “Noise power spectrum properties of deep learning–based reconstruction and iterative reconstruction algorithms: Phantom and clinical study”. In: *European Journal of Radiology* (2023), p. 110914.

[79] Joost JM Van Griethuysen et al. “Computational radiomics system to decode the radiographic phenotype”. In: *Cancer research* 77.21 (2017), e104–e107.

[80] Elise Whitley and Jonathan Ball. “Statistics review 1: presenting and summarising data”. In: *Critical Care* 6.1 (2001), pp. 1–6.

[81] David F Williamson, Robert A Parker, and Juliette S Kendrick. “The box plot: a simple visual method to interpret data”. In: *Annals of internal medicine* 110.11 (1989), pp. 916–921.

[82] Mavuto M Mukaka. “A guide to appropriate use of correlation coefficient in medical research”. In: *Malawi medical journal* 24.3 (2012), pp. 69–71.

[83] Jaime Lynn Speiser et al. “A comparison of random forest variable selection methods for classification prediction modeling”. In: *Expert systems with applications* 134 (2019), pp. 93–101.

[84] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[85] Justin Solomon and Ehsan Samei. “Quantum noise properties of CT images with anatomical textured backgrounds across reconstruction algorithms: FBP and SAFIRE”. In: *Medical physics* 41.9 (2014), p. 091908.

[86] Annalisa Laghi and Gabriele Soffritti. “A collinearity based hierarchical method to identify clusters of variables”. In: *New Developments in Classification and Data Analysis: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Bologna, September 22–24, 2003*. Springer. 2005, pp. 55–62.

[87] VH Arul. “Deep learning methods for data classification”. In: *Artificial Intelligence in Data Mining*. Elsevier, 2021, pp. 87–108.

A

Literature Review

Deep learning based techniques for standardization of variability across CT scans due to reconstruction: a scoping review

Toke M. Camps[a,b]

[a] *MSc Biomedical Engineering, track Medical Physics, Delft University of technology, Delft, Netherlands*
[b] *Aidence B.V., Netherlands*

Abstract— Computed tomography (CT) is vital for diagnosing, monitoring, and treating various medical conditions. However, CT scans pose challenges in large, multi-site or longitudinal studies due to variations in reconstruction techniques used by different manufacturers and hospitals. Such variations result in significant differences in image quality, hindering the comparison and analysis of CT images across different scanners and institutions. The development of a deep-learning based post-processing framework offers a new solution to address inconsistent CT images. It has the potential to standardize and normalize existing CT images while preserving most anatomical details simultaneously. The aim of this scoping review is to comprehensively overview the state-of-the-art DL strategies for standardizing CT scans that vary due to reconstruction techniques. The strategies are compared based on the type of standardization, underlying DL architecture, and performance evaluation. In total, thirteen studies were included and reviewed after a systematic literature search in PubMed and IEExplore. The techniques evaluated all atttempt to solve intra-scanner variation resulting from different reconstruction kernels, whereas only four studies aim to tackle intra- and cross-scanner variation. GAN-based model alterations demonstrate the most promising results in terms of non-paired image usage, content preservation, forward and backward mapping, and combined intra- and cross-standardization. Despite the promosing developments, further research is still required before these innovative standardization approaches can be implemented in a stable manner in clinical settings.

Keywords—Computed Tomography, standardization, deep learning, reconstruction

I. Introduction

Computed tomography (CT) is an essential imaging technique for the diagnosis, monitoring, and treatment of various medical conditions [1]. However, the use of CT scans can be challenging in large multi-site studies or longitudinal studies due to the variations in the reconstruction techniques used by different manufacturers and hospitals [2]. These variations can lead to significant differences in the image quality [3, 4], making it difficult to compare and analyse CT images across different scanners and institutions [5].

Several research studies have reported that the performance of medical image analysis techniques depends on the image variations arising from different reconstruction techniques. For instance, Blazis et al. [5] discovered that a commercially available Computer-Aided Diagnosis system based on Deep Learning (DL) performs differently on images reconstructed with iterative reconstruction (IR) compared to filtered back projection (FBP). Furthermore, recent research indicates a lack of reproducibility of radiomic features in response to the variance in CT reconstruction parameters [6–8]. Radiomic features provide valuable information about tumour characteristics, and are employed in medical image analysis, e.g. to diagnose and differentiate between different types of cancer or predict treatment response and patient outcomes [9]. These findings stress the importance of taking caution while implementing automatic image analysis in a hospital with multiple CT scanners and different reconstruction protocols.

A new solution to tackle the issue of inconsistent CT images has emerged through the development of a deep-learning based post-processing framework. This framework has the potential to standardize and normalize existing CT images, while simultaneously preserving most of the anatomic details.

1. Outline

This systematic review aims to provide a comprehensive overview of the state-of-the-art DL strategies for standardization across CT scans that vary due to reconstruction techniques. The strategies are compared with each other by comparing the type of standardization, underlying DL architecture and the performance evaluation that has been performed.

To answer this, background information is provided in section II about CT image reconstruction, clinically available reconstruction techniques and their strengths and limitations, and lastly, about DL including interesting architectures for image-to-image translation. In section III, the literature search and eligibility criteria are described. Consecutively, the results of the literature search are discussed in Section IV. In Section V, the found results are discussed and finally, in Section VI conclusions are drawn on the current status of CT image standardization based on DL models.

II. Technical Background

1. Computed tomography image reconstruction

The CT scan is performed by rotating an X-ray source and detector around a patient to acquire a series of two-dimensional (2D) X-ray projections, or 'slices' at different angles [10]. These slices represent thin cross-sections of the patient's anatomy, which can be processed to reconstruct a three-dimensional (3D) object.

The X-rays transmitted through the patient interact with the body tissues they encounter, which causes an exponential reduction with distance travelled in beam intensity based on tissue density and composition. As the X-ray photons exit the patient, they are absorbed by a CT detector and converted into an electronic signal. The attenuation of the X-ray beam as it passes through a material along the line γ can be calculated using the line integral (Eq. 1). In the equation, $\mu(x,y)$ is the linear attenuation coefficient of the tissue at position (x,y) in a 2D plan and $\gamma(t) = (x(t),y(t))$ is a parametric equation for a straight line that intersects the plane. The integral is taken over the entire length of the line γ.

$$I = e^{-\int \mu(x(s),y(s))ds} \tag{1}$$

A set of line integrals along all the ray paths in the X-ray beam form an x-ray projection (or attenuation profile) that is used to compute a 3D representation of the patient's internal anatomy. Finally, the complete collection of line integrals that traverse the patient's body for every possible ray trajectory within the X-ray beam, encompassing all gantry angles, is called the Radon transform [10].

The basic idea behind CT image reconstruction is to use mathematical algorithms to estimate the tissue density in each voxel of the image volume. This is done by solving an inverse problem (back-projection), where the goal is to find the tissue density distribution that best explains the measured X-ray projections.

1.1. Sinogram

For CT image reconstruction, a sinogram is created, which is a 2D representation of a CT scan [10, 11]. In the sinogram each column represents a single row in the raw projection data arranged in increasing angular order, thus provides a compact representation that can be easily filtered and back-projected to obtain an estimate of the 3D object. Typically, the x-axis represents the gantry angle (θ) and the y-axis represents the x-ray projection by a detector element along the detector row, thus the distance along the projection direction (l). Consequently, the sinogram comprises a series of sine functions that overlap, where the amplitude and phase rely on the voxel's location, and the gray value corresponds to the volume element's gray value, while the wavelength remains constant. An example of a sinogram is visualized in Figure 1.

1.2. Central slice theorem

The central slice theorem is one of the fundamental concepts in CT image reconstruction [12]. This theorem, also known as the Fourier slice theorem, states that the 2D Fourier transform (FT) of an object is equivalent to the 1D FT of the object's projection passing through its centre and perpendicular to the plane of the 2D FT [13]. By transforming all projections of the object into the 1D Fourier transform and interpolating them into a 2D Fourier space, the complete 2D FT of the object can be reconstructed. Using the inverse FT, the original object is reconstructed from the full 2D FT.

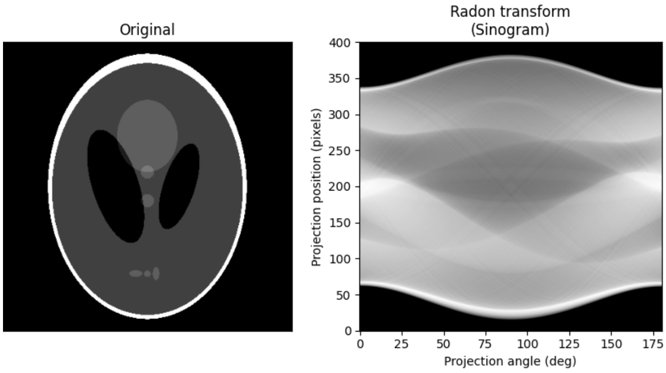


Fig. 1: *Left*: the original image, which is the Shepp-Logan phantom, a standard test image. *Right*: the sinogram of the original image.

The theorem has provided a mathematical foundation for the reconstruction of images from X-ray projections and has facilitated the development of various CT reconstruction algorithms. The subsequent section explains how CT scans are reconstructed in practice based on the aforementioned theory.

2. Reconstruction techniques

There are various strategies used for CT image reconstruction, including FBP, iterative reconstruction (IR), and deep learning reconstruction (DLR). FBP was the standard image reconstruction method for decades because of its simplicity and computational efficiency [14, 15]. In fact, it was not until 2009 that the initial IR algorithms were introduced to the market, replacing the conventional FBP technique [14]. The current state-of-the-art method for CT image formation is image reconstruction based on DL, currently only three DLR algorithms are commercially available [16]. In the following subsections, each technique is explained and the strengths and limitations are reviewed per technique. In addition, in Appendix A the commercially available algorithms are briefly explained.

2.1. Filtered back-projection

Filtered back projection is a common reconstruction method that reconstructs a 3D image of an object from its projection data (sinograms). In short, the method evenly distributes the measured filtered signal over the projection line to compute 2D slices, which are then combined to form a 3D volume that represents the object [10, 17].

More specific, for each gantry angle in a sinogram, the attenuation value is divided by the number of image pixels along the path of the projection from the X-ray source to the detector. The resulting average attenuation value is then allocated to those pixels. This process is carried out for every gantry angle. The back-projected data is then summed to form the final back-projected image.

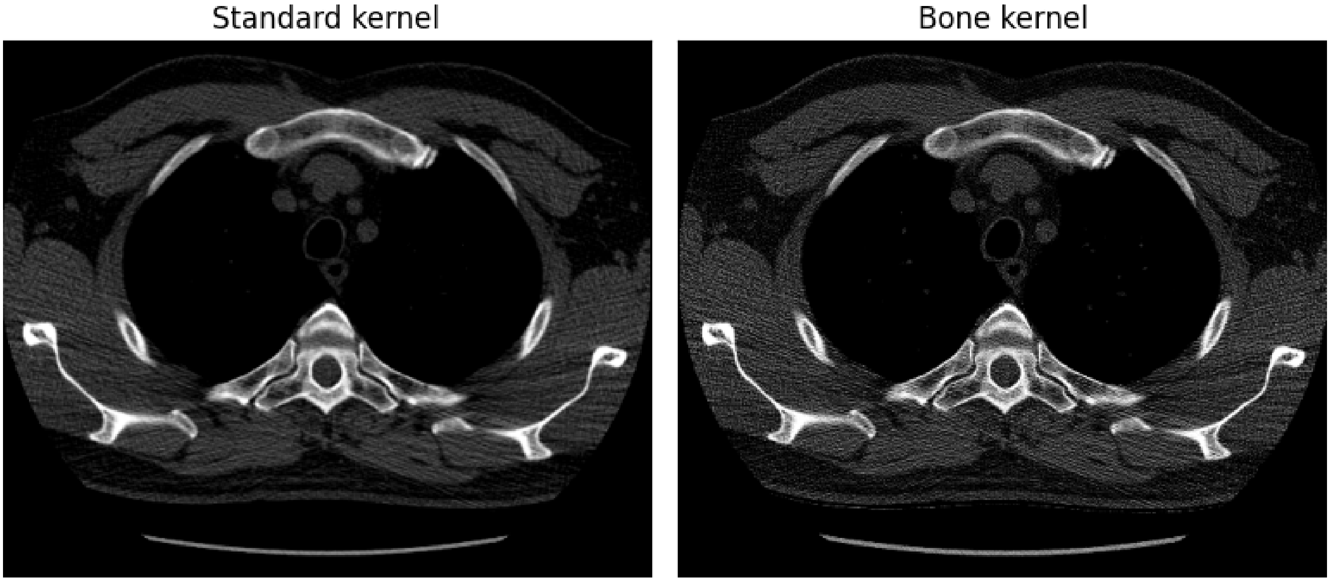


Fig. 2: The left CT scan is reconstructed with the standard reconstruction kernel, whereas the right CT scan is reconstructed with the sharper kernel, called bone reconstruction kernel. Both reconstruction kernels are developed and GE Medical Sytems. The standard, smoother kernel, shows lower image noise and displays more low-contrast details, but, on the downside it has a lower image sharpness. On the contrary, the scan reconstructed with the bone kernel, a sharper kernel, enables a better edge definition and shows more structural details, which is clearly visible around the bones, however, it also shows an increased image noise.

Prior to back-projection, the projection-data is filtered to counteract blurring that occurs because of evenly spreading the attenuation value [10, 18]. Applying a ramp filter (reconstruction kernel) to the projection, either through convolution in the spatial domain or multiplication in the Fourier domain, generates a "filtered" projection featuring negative sidelobes. This results in a spatial-frequency-amplified rendition of the initial projection, where the high-frequency augmentation matches the high-frequency damping applied during back-projection. The method heightens rapid spatial variations in the attenuation pattern, such as boundaries between anatomical structures of contrasting densities. It suppresses low spatial frequency components of the attenuation profiles, thus reduces blurring. However, the filter also enhances image noise, which exists in the raw signal primarily at high spatial frequencies.

Reconstruction kernel The ramp filter is required mathematically to eliminate blurriness from back-projection, but it can be paired with filters of varying intensities (kernels) to heighten the spatial resolution of the ultimate image, dependent on the specific application [10, 19]. For example, a 'sharper' kernel, with higher filtration, may be employed to enable a better definition of edges and a clearer delineation of structural detail [15]. However, the downside of attaining a greater spatial resolution is the increase in image noise [10, 15].

Various kernels with distinct features are at the disposal of clinicians in their day-to-day operations [18]. 'Smooth' kernels are designed to lower image noise and bolster the display of low-contrast details, but can lead to a drop in image sharpness. Meanwhile, 'sharp' kernels aim to enhance the illustration of intricate elements in high-contrast structures, although they can increase image noise to a level that hinders the recognition and distinction of low-contrast structures [19, 20]. In Figure 2 an example of a CT image pair is displayed, a CT scan that is reconstructed with a 'smooth' kernel, the standard kernel of GE medical systems (left scan), as well as with a 'sharp' kernel, the bone kernel of GE medical systems (right scan).

Strengths and limitations The FBP technique is computationally efficient, and therefore a very simple and fast method [10]. The user has control over image characteristics, i.e. by indication-specific choice of reconstruction kernel. Also, FBP reconstructs images with well-known image texture, and the conventional image quality metrics are globally valid for this method.

However, FBP makes certain assumptions that are not accurate, such as treating the focal spot on the X-ray tube as a point source with a perfectly shaped pencil beam that hits the patient's body at a single point, and measuring the intensity at the central point of each detector element [10, 14, 15, 19]. Consequently, each pixel value in the resulting image comes with an inherent uncertainty that stems from the X-ray detection process and the image reconstruction process.

While the ramp filter applied during FBP helps eliminate blurring and improve edge detection in the image, it also amplifies the image noise already present in the raw signal, particularly at high spatial frequencies. This results in FBP CT images having a characteristic speckled or mottled appearance. Low-dose CT scans are not feasible since the resultant image quality would not be sufficient for diagnostic purposes due to its noisy nature [14, 19].

2.2. Iterative reconstruction

The underlying principle of iterative image reconstruction is to calculate image data that accurately corresponds to the acquired projection data using iterative algorithms [15, 18]. The iterative algorithm can be defined as a constrained optimization problem, in which the image data is the unknown optimum solution to the problem. Mathematically, this can be expressed as a cost function that seeks to optimize two aspects of the reconstruction simultaneously: conformity of the reconstructed image data with the measured projection data (data term), and noise suppression through a regularization term that penalizes noisy solutions to the optimization problem. By incorporating constraints in the optimization problem, it is possible to account for the CT imaging process model's statistical and system optic properties. In essence, IR reconstruction requires repeatedly updating the image data to minimize the cost function, thus enhancing conformity between measured and reconstructed data while minimizing image noise.

To reconstruct CT images through IR, an optimal process consists of a cycle of forward- and back-projection steps. This involves repeatedly converting between raw projection data and image space. During the forward projection step, synthetic projection data is created and compared with measured projection data. The correction obtained from the difference between the simulated and measured projections is propagated to image space via the back projection step, typically through filtered back projection. This correction is then applied as an update to the current image data estimate. The iterative cycle of forward and back-projection steps is repeated until a predefined stopping criterion is met.

Two adjustable parameters can be identified that influence the outcome of the IR method: the reconstruction kernel and the strength of the algorithm. The reconstruction kernel is employed in the FBP step of the algorithm, and influences the image noise and image sharpness. The selection of the strength level leads to varying degrees of noise reduction [21, 22]. Nevertheless, excessive IR strength, particularly at higher levels, can produce unsightly "blooming" artefacts that tend to hinder the display of minute structures [23]. As a result, the impact of the IR algorithm's strength on image quality must be weighed carefully to strike a balance.

The major medical manufacturers have their own IR approach(es) that can be roughly categorized into statistical (hybrid) and model-based iterative algorithms, depending on the extent to which they model the imaging process [24, 25]. These categories are further explained in the next sections.

Statistical iterative reconstruction (hybrid) IR algorithms based on statistical models employ iterative data filtration, which is performed separately in projection space and/or image space [15]. However, despite this, the actual image reconstruction process often still relies on FBP. As a result, the speed of image reconstruction for this category of IR algorithms, often referred to as hybrid IR, is generally similar to that of FBP.

In projection space, statistical filtration involves iterative analysis of data variation. By using statistical models, neighbouring projection data is compared to identify overly noisy or photon-starved projections. These projections are then either replaced or modified to ensure maximum data consistency, i.e., to minimize variation. Without modification, such projections would contribute significantly to image noise and artefacts, such as streaking, while providing limited information for the reconstructed image data [26]. Unlike FBP, where all projections have equal weighting, modified projections can be assigned a lower weight to prevent potential bias, resulting in reduced contribution to the reconstructed image data compared to unaltered projections [19].

After transitioning to image space, for example, via FBP, statistical models of the noise structures characteristic of the imaged body regions are employed to iteratively filter the image data, further reducing image noise [14, 19, 26]. Edge-preserving filters are used to minimize the impact on the depiction of fine structure and low-contrast detail [26]. Although it is possible to apply iterative filtration in projection or image space alone, current state-of-the-art statistical IR algorithms typically perform iterative optimization in both spaces.

Model-based iterative reconstruction Model-based iterative reconstruction (MBIR) differs from statistical IR in that it involves simulating projection data by at least one forward projection from image space to projection space based on the current image estimate [15]. This requires a model of the CT imaging process for forward projection, as well as a model or estimate of the imaged object, also known as a prior for initializing the iterative cycle (e.g., gained by FBP reconstruction). The closer the image prior is to the imaged object, the faster the MBIR algorithm will converge. Back projection of a correction term computed by comparing synthetic and measured projection data yields updated image data, which can be used to initialize the subsequent forward projection of the iterative cycle. Since simulating synthetic raw data by forward projection in MBIR is complex and requires a large amount of computation time, iterative filtration in projection and image space, similar to the statistical filtration processes used in statistical IR, can also be used to limit the number of required iterative forward projection steps and facilitate faster convergence [26].

In contrast to statistical IR, which only models photon statistics, MBIR models the technical properties of the CT system used, such as system optics and further details of CT imaging physics, as part of the CT imaging process [15]. As a result, the modelling of the CT imaging process in MBIR is more precise and detailed than in statistical IR, making it more complex and computationally intensive.

Strengths and limitations The primary advantage of IR techniques over FBP is their ability to significantly reduce image noise and artefacts while preserving accurate attenuation values. As a result, signal-to-noise (SNR) and contrast-to-noise (CNR) ratios are increased, and the visualization of low-contrast features is enhanced. In addition, for MBIR slight improvements in spatial resolution can be observed [15, 19].

On the downside, the image reconstruction time increases with the complexity of the CT imaging process modelling. Especially, MBIR is more computationally demanding, resulting in increased reconstruction times [14, 15]. Another weakness of IR is the risk of over smoothing and of altered, unfamiliar appearing image texture, e.g., at tissue interfaces [14, 19]. Furthermore, alteration of image characteristics through the use of IR may have a potential impact on quantitative CT analyses. Consequently, applying reference standards determined from image data reconstructed with FBP without proper consideration might lead to significant disparities in outcomes [19, 27].

2.3. Deep learning reconstruction

Deep-learning-based techniques for CT image reconstruction is an emerging technique that has the potential to further improve image quality and thus reduce dose [28]. Opposed to traditional CT reconstruction methods that use analytical models, DLR uses neural networks to learn the mapping between the raw projection data (sinograms) and the corresponding high-quality CT images. Training this neural network requires a large dataset of CT projection data and corresponding high-quality CT images. Such a dataset originates from phantom images as well as patient scans conducted in a clinical setting [29]. Once the neural network is trained, it can be used to reconstruct CT images from new raw projection data [28, 29]. The network takes the raw projection data as input and produces a high-quality CT image as output.

There are several types of deep neural networks that can be used for CT image reconstruction, including convolutional neural networks (CNNs), residual neural networks, and generative adversarial networks (GANs). GAN-based methods are particularly useful for CT image reconstruction because they can generate high-quality images with fine details and textures, even when there is limited or noisy data. In the next section (section II.3), Deep learning and interesting networks are explained more in depth.

Strengths and limitations Deep learning-based CT image reconstruction has shown promising results in improving image quality and reducing radiation dose in CT scans. It allows for noise reductions relative to FBP without suffering from the unnatural appearing noise textures associated with IR solutions [16, 30]. However, it requires a large amount of training data and significant computational resources to train the neural network. Therefore, there is ongoing research in developing more efficient and effective deep learning-based CT reconstruction methods.

The reliability of the reconstructed image is another limitation of DL-based image reconstruction, as there is no guarantee that important patient structures will not be changed during reconstruction [29]. Even if a DL algorithm produces an accurate image, it could be based on wrong reasoning. For instance, a particular lesion might be eliminated or blurred out because it was not adequately represented in the training data, leading the model to mistake it for noise. On the other hand, the reconstruction technique might introduce a non-existent lesion into the reconstructed images.

Worth noting is that while commercially available DLR algorithms have primarily been trained to reduce noise, DLR has the potential to solve a variety of image reconstruction issues, including cone-beam artefacts, motion artefacts, and truncation artefacts [16].

3. Deep learning

Deep learning is a subfield of machine learning that is based on artificial neural networks [31]. Such a network consists of many individual artificial neurons, which are modelled after the structure and function of the biological neuron [31, 32] The artificial neurons modify the connections between them through the training process in the artificial neural network, similar to a biological neuron [33]. Deep learning models are made up of multiple layers of interconnected artificial neurons that process input data and learn hierarchical representations of data, which enables it to solve complex tasks [34]. This is achieved through a process known as backpropagation, which involves iteratively adjusting the weights of the neural network based on the error between the predicted and actual outputs [35].

3.1. Convolutional neural network

The convolutional neural network is the most famous and commonly employed model, it typically has an architecture that is structured as a series of stages [36]. The first few stages are composed of two types of layers: convolutional layers and pooling layers.

The convolutional layer, the main building block of a CNN, contains a set of filters, parameters of which are to be learned throughout the training. Each filter convolves with the input (the image or previous feature maps), and creates a new feature map, where each component is a neuron. The output volume of the layer is generated by stacking the feature maps of every filter along the depth dimension. Each neuron is connected to local patches in the previous feature maps through a set of weights [34, 36]. Due to the local connectivity of the convolutional layer, the network is forced to learn filters that have the maximum response to a local region of the input. The initial convolutional layers capture the low-level features (e.g., lines) of images, while the later layers extract the mid-level features (e.g., shapes and specific objects) [36, 37]. A nonlinearity (activation) function, e.g. a rectified linear unit (ReLU), is applied in between convolution layers changing the linear nature of the neural network, allowing to approximate non-linear functions [37].

The pooling layer down-samples every feature maps in the sub-sampling layers, which leads to a reduction in the representation dimensions, thereby accelerating the training process. In addition, it allows for handling of overfitting problems [37, 38].

At the end of its architecture, a CNN can include fully-connected layers that take in mid- and low-level features to create high-level abstractions, thereby representing the last-stage layers [37]. This is an optional feature of CNNs, which depends on the purpose of the model.

U-net The U-Net network is a type of CNN designed by Ronnebergert et al. [39] and allows preserving the spatial distribution of the image while abstracting image features due to its design. Originally, the network was developed for segmentation purposes, and later its applications extended to image-to-image translation [40]. The U-Net consists of an encoder, containing down-sampling layers, a decoder, containing up-sampling layers and connections from down-sampling layers to the corresponding up-sampling layers to recover lost information during down-sampling [40].

The encoder takes the input image and convolves it, generating increasingly complex feature maps as it moves deeper into the network, while the spatial scale reduces with the convolutional and pooling operations. The decoder combines the feature maps and spatial information through a sequence of deconvolution block and concatenation with high resolution features from the connection path [37, 40].

3.2. Generative adversarial network

The generative adversarial network is a class of DL models that learns the data distribution of training images and generates synthesized images under the same distribution. Two neural networks contest with each other in this model: the generator (G) and the discriminator (D) [41]. The first generates synthesized data from random noise, and the latter learns a data distribution from the training data and determines whether the synthesized data generated by G is drawn from real or fake data [40, 42]. The goal of G is to generate data that is so realistic that D cannot tell the difference between the generated and real data.

The training process of a GAN involves iteratively training the generator and the discriminator [43]. In each iteration, the generator generates new data based on the random noise input, and the discriminator evaluates both the real and generated data to determine which is real and which is fake. The discriminator's output is then used to update both the generator and discriminator's weights, so that they can better compete against each other in the next iteration.

CycleGAN CycleGAN is a type of GAN that is used to perform image-to-image translation [44]. It is designed to learn the mapping between two different image domains, without requiring paired examples of corresponding images in the two domains.

The key idea behind CycleGAN is to have two GANs, each with a generator and a discriminator, and train them in a cycle. The two domains are referred to as the source domain and the target domain. The generators in the two GANs are trained to learn the mapping between the two domains in both directions: from the source domain to the target domain, and from the target domain to the source domain. This is achieved by introducing a cycle consistency loss, which enforces that if an image is translated from the source domain to the target domain and then back to the source domain, it should be similar to the original image.

III. Methods

1. Search strategy

This literature review was carried out by following the guidelines of the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) statement [45]. A systematic literature search was conducted in the PubMed and IEEEXplore databases using the search queries stated in Appendix B. All articles published prior to the 14th of March 2023 were included.

2. Study selection

Primary studies that were available as English full text and published in 2012 or later were found eligible when it described a new DL-based method for CT scan standardization of variability caused by reconstruction techniques. Study selection was cut off before 2012, because the application of deep learning for medical imaging is relatively recent, only since 2012 deep learning started to outperform conventional methods [37]. The eligibility criteria are outlined in Table 1.

Table 1: Eligibility criteria

Category	**Criteria**
Date of publication	1 January 2012 - 14 March 2023
Language	English
Type of article	Primary study
Publications	Published as full-text article in a peer-reviewed Journal
Study objective	Description of a scan standardization method, that meets the following requirements: (1) Deep learning based (2) Target imaging modality is conventional CT (3) New technique is presented (4) The variability caused by reconstruction methods is addressed (5) Creates a standardized image

3. Results synthesis

For each study, the DL-based standardization model was extracted. More in depth, for each model, specifications of its purpose, the type of DL model, specifications of the data sets used for training and for evaluation, and performance results were extracted and summarized. To provide a structured overview, the studies are categorized and reviewed based on three different characteristics:

1. Purpose of the standardization method, defined as the standardization type. Either intra-scanner, cross-scanner or intra- & cross-scanner standardization.
2. Type of deep learning model, which is subdivided in two main DL classes: CNN or GAN.
3. The performance evaluation that the study has performed. For this characteristic, two main types of evaluation can be identified: image similarity and radiomic feature similarity.

IV. Results

1. Study selection

A total of 68 studies were identified from the electronic database search. After removing one duplicate study, the remaining 67 studies were screened on title and abstract. This resulted in the removal of 43 studies, with as predominant reason for exclusion that the study did not describe a scan standardization method (n=25). The remaining 24 studies continued for eligibility assessment based on the full text, which lead to the final selection of thirteen studies for this review. The complete overview of the search flow is shown in Figure 3.

The characteristics of the included articles are summarized in Table 2 and discussed in more details in the upcoming sections. The proposed standardization models are compared with each other based on the type of standardization (section IV.2), underlying DL architecture (section IV.3), and model performance evaluation (section IV.4).

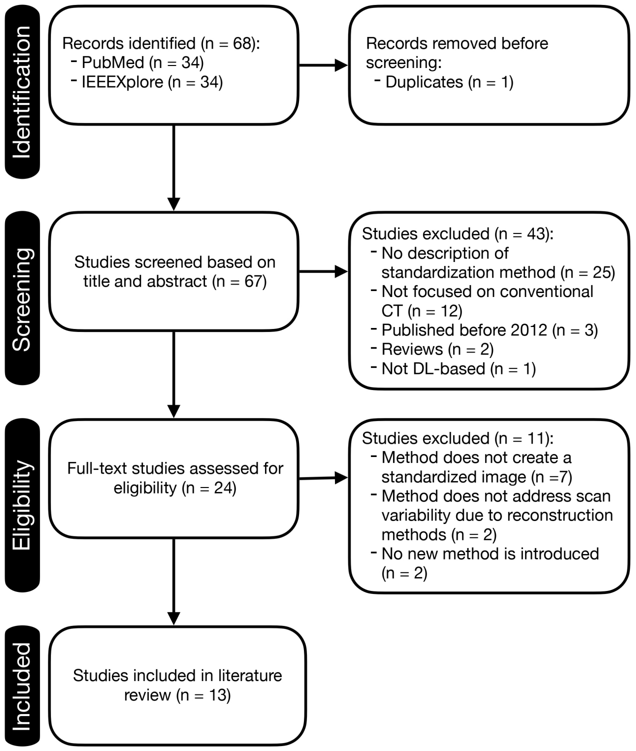

Identification
Records identified (n = 68):
- PubMed (n = 34)
- IEEEXplore (n = 34)
Records removed before screening:
- Duplicates (n = 1)
Screening
Studies screened based on title and abstract (n = 67)
Studies excluded (n = 43):
- No description of standardization method (n = 25)
- Not focused on conventional CT (n = 12)
- Published before 2012 (n = 3)
- Reviews (n = 2)
- Not DL-based (n = 1)
Eligibility
Full-text studies assessed for eligibility (n = 24)
Studies excluded (n = 11):
- Method does not create a standardized image (n =7)
- Method does not address scan variability due to reconstruction methods (n = 2)
- No new method is introduced (n = 2)
Included
Studies included in literature review (n = 13)

Fig. 3: Flow chart of the literature search in PubMed and IEEEXplore

2. Standardization type

Broadly speaking, there are two types of CT image standardization models that serve different purposes. The first type is known as intra-scanner image standardization, which generally requires paired image data. In this context, a pair of images is generated from the same object, either a patient or phantom scan, but with varying reconstruction kernels or algorithms. The image created using non-standard settings is referred to as the source image, while the image produced using standard settings is known as the target image. With access to paired image training data, a model can learn how to transform source images into target images. The second type of models is designed for cross-scanner image standardization and does not require paired image data. In this scenario, paired images are not needed; instead, images acquired using standard and non-standard protocols are stored separately. Acquiring paired training data is relatively straightforward, but is typically restricted to a single scanner.

In this review, nine articles are included that developed a model that addresses the variability in reconstruction within the scanner [40–42, 46–51]. No included studies focus on cross-scanner standardization alone, whereas four studies designed a model for the combination of intra- and cross-scanner standardization [2, 52–54]. A summary of the studies regarding their standardization type and data used for training is given in the Tables in Appendix C. In specific, the studies are divided in intra-scanner and cross- & intra scanner standardization and the specifications of the training data set and the purpose of the proposed method are stated.

2.1. Intra-scanner CT image standardization

The simplest standardization models perform conversion of CT images that are reconstructed with one kernel to images reconstructed with a different kernel. As mentioned before, image pairs are used to train these type of models, which consist of a pair of images reconstructed with different kernels, generally it contains a smooth and sharp reconstructed image.

The two CNN-based models proposed by Choe et al. [46] and Lee et al. [47] employ a similar intra-scanner standardization method. These models were trained to learn the difference between the target and source images (also known as residual images), and create the converted image by adding the residual images to the input. For each distinctive sharp-smooth reconstruction pair and conversion direction, a new model was developed, which resulted in the development of two models by Choe et al. [46] and twelve models by [47]. Both developed a model for the conversion from smooth (B30f kernel) to sharp (B50f kernel) and vice versa. The developed models of Lee et al. [47] also included kernel conversion from and to the B10f and B70f kernels.

A similar standardization technique has been introduced by Jin et al. [40], however, only for the conversion of a CT scan reconstructed with a sharp kernel to that of a standard (smoother) kernel. The model is based on a U-net architecture, and in total four models are trained using different data sets that are acquired from different vendors.

Liang et al. [42] present a simple intra-scanner standardization method that attempts to convert images reconstructed with the Bl57 kernel to the predefined standard-kernel, Bl64. In contrast to the previous two described models, this model learns the data distribution of the target data and generates synthesized images from the source images under the same distribution of the target images using adversarial learning.

The studies of Selim et al. [41] and Wu et al. [49] both present a model to convert various non-standard reconstruction settings to one predefined standard setting. Wu et al. [49] uses images reconstructed with four different kernels, varying in smoothness, and a slice thickness of 5-mm as input and converts them to an 1-mm sharp-kernel image that was defined as the standard protocol. On the other hand, the study of Selim et al. [41] defines the Bl64 kernel as standard, since according to them, it has been widely used in clinical practice. The non-standard protocols consist of two reconstruction kernels that are smoother compared to Bl64.

Using a U-net based architecture, Tanabe et al. [48] have proposed a standardization technique for sharp to smooth conversion, where the novelty of the study lies in the region-wise learning. The authors found that the accuracy of the conversion decreased for the region with CT values ranging from -200 to 200 HU. To overcome this challenge, the authors trained an additional network (the partial-model) to convert images from sharp-kernel to soft-kernel, truncating both to the range of -300 to 300 HU. They also trained a full-model with non-truncated images. The final converted image was generated by combining the outputs of the two models using a weighted sum.

Opposed to the other studies, Yang et al. [50] attempt to translate images between any pairs of kernel domains along the interpolation path. This approach enables the effective utilization of intermediate kernel images, which facilitates image conversion between sharp-to-smooth and

Table 2: Included studies overview. From each study the publishing year, standardization type (type), conversion direction: smooth-to-sharp (sm-to-sh) or sharp-to-smooth (sh-to-sm), underlying architecture, #of trained models by the authors, performance evaluation metric(s) and the use of paired-data for training has been extracted.

Article	Year	Type	Conversion direction	Architecture	#of trained models	Performance evaluation metric			Paired data (for training)
						Feature similarity	Image similarity	Alternative	
Choe et al. [46]	2019	Intra-scanner	Sm-to-sh & sh-to-sm	CNN	2	CCC, # of reproducible features	N/A	N/A	Yes
Lee et al. [47]	2019	Intra-scanner	Sm-to-sh & sh-to-sm	CNN	12	N/A	RMSE of CT value	N/A	Yes
Jin et al. [40]	2019	Intra-scanner	Sm-to-sh & sh-to-sm	CNN: U-net based	4	N/A	N/A	Variability in lung density: RA950, perc15, MLA	Yes
Yang et al. [50]	2021	Intra-scanner	Sm-to-sh & sh-to-sm	GAN: Switchable + split AdaIN	2	N/A	SSIM, PSNR	N/A	Yes
Tanabe et al. [48]	2022	Intra-scanner	Sh-to-sm	CNN: U-net based	1	N/A	Difference in CT value	N/A	Yes
Liang et al. [42]	2019	Intra-scanner	Sm-to-sh	GAN: Alternative training	1	AE, RE	N/A	N/A	Yes
Selim et al. [41]	2020	Intra-scanner	Sm-to-sh	GAN	1	AE	N/A	N/A	Yes
Wu et al. [49]	2021	Intra-scanner	Sm-to-sh	GAN: CycleGAN	1	N/A	MSE, PSNR	N/A	Yes
Lee et al. [51]	2022	Intra-scanner	N/A	GAN	1	CCC	SSIM, PSNR	Image quality: CNR	Yes
Du et al. [52]	2022	Cross- & Intra-scanner	Sm-to-sh & sh-to-sm	CNN	2	PNFD, ICC	N/A	N/A	Yes
Kim et al. [55]	2022	Cross- & Intra-scanner	Sm-to-sh & sh-to-sm	GAN: Routable decoder	1	N/A	N/A	Domain classification accuracy	No
Selim et al. [2]	2022	Cross- & Intra-scanner	Sm-to-sh	GAN: CycleGAN + domain adaptation	1	RE, # of reproducible features	N/A	N/A	Yes
Li et al. [53]	2021	Cross- & Intra-scanner	N/A	GAN: SingleGAN based	1	% of aligned features	N/A	N/A	No

Abbreviations: *CNN* = convolutional neural network, *GAN* = generative adversarial network, *AdaIN* = adaptive instance normalization, *CCC* = concordance correlation coefficient, *PNFD* = patient-normalized feature difference, *ICC* = intraclass correlation, *AE* = absolute error, *RE* = relative error, *SSIM* = structural similarity index, *PSNR* = peak signal-to-noise ratio, *RMSE* = root mean squared error, *CT* = computed tomography, *MSE* = mean squared error, *RA950* = relative lung area under 950 HU, *perc15* = lower 15th percentile threshold, *MLA* = mean lung attenuation, *CNR* = contrast-to-noise ratio.

smooth-to-sharp, as well as between intermediate-to-smooth or intermediate-to-sharp and vice versa, with the training of just one model.

All previous studies focused on the conversion between reconstruction kernels, whereas the study of Lee et al. [51] aims to convert images reconstructed with the same kernel but with different reconstruction algorithms. The model that has been presented by the authors generates synthetic CT images similar to images reconstructed with IR, in specific SAFIRE or ADMIRE, from images reconstructed using conventional FBP. This is the only study that includes a reconstruction algorithm different from FBP in their research.

2.2. Intra- & cross-scanner CT image standardization

Four studies presented a standardization method that included both intra- and cross-scanner standardization, the methods differ greatly from each other. The authors of [2] propose a standardization tool to convert any CT image acquired differently than the defined standard protocol into the standard one. For the training, they use a data set containing paired images reconstructed with Br40 and Bl64 kernels acquired with a Siemens scanner. Consequently, a second data set consisting of single images obtained with the LUNG kernel using a GE scanner has been used to evaluate the model's conversion from LUNG kernels to the standard kernel, the Bl64 kernel.

The model of Du et al. [52] also converted a broad variation of non-standard kernel images from a variety of scanners to a standard protocol. Contrary to the method of Selim et al. [2], it utilizes two different standard protocols, one where reconstruction is performed with a sharp kernel (B70f) and the other with a smooth kernel (B30f), referred to as protocol A and B, respectively. In line with defining two different standard protocols, two separate standardization models are trained, each with one of the standard protocols as target and the other as source.

Subsequently, to apply the standardization to varied kernels of different vendors, their median patient-normalized feature values were compared with the feature values calculated from protocol A and B. In cases where the non-standard protocol had a difference greater than 1 compared to protocol A or B, identified as dissimilar feature values, the non-standard protocols were converted to that specific standard protocol using a model trained specifically with that standard protocol as the target.

Kim et al. [54] propose a multi-domain translation network based on a routable GAN (RouteGAN) to effectively convert diverse CT images across several settings. The RouteGAN's encoder was designed to extract a shared feature for improved conditioning power and expressiveness. The decoder layers enable the selection of different target routes by altering the condition vector, thereby simulating the use of distinct decoders for each domain. The research covered seven distinct domains with varying radiation dose, vendor and reconstruction kernel.

A cross- and intra-scanner standardization technique that aims to normalize CT images acquired with different scanners, and different reconstruction protocols, to one standard reconstruction protocol has been proposed by Li et al. [53]. The input images are obtained with three different scanners from different vendors: GE Healthcare, Philips Healthcare & Siemens Healthcare. The input images underwent normalization based on the features of images reconstructed using a Siemens reconstruction protocol, that was acquired from the same Siemens scanner from the input data set, however with different reconstruction parameters. Additionally, the model acquired a reverse mapping from the destination images to the source images, facilitating the learning process despite non-paired images from various reconstruction protocols in a single patient that were not pixel-aligned.

3. Type of deep learning model

Two main types of deep learning models are identified: CNN and GAN models, which mainly differ in their purpose and training process. CNN is generally used for image recognition and classification, and is trained to minimize the error between the predicted output and the actual output by updating the weights of the network. On the other hand, GAN is mainly used for generating new data that resembles the training data, and during training, the generator network generates fake data, and the discriminator network tries to distinguish between the real and fake data.

It is worth mentioning that while the generator and discriminator networks in a GAN model can be based on the CNN architecture, the GAN model is fundamentally different from a standard CNN due to its distinct purpose and training processes. Therefore, GAN and CNN are considered as two different classes of DL models.

Of the included articles in this review, five articles fit the first category: CNN-models [40, 46–48, 52], however, the largest number of articles (8) present a GAN-based standardization model [2, 41, 42, 49–51, 53, 54].

3.1. CNN-based model

In 2019, the first two DL models were proposed by Choe et al. [46] and Lee et al. [47] for image conversion among different reconstruction kernels using CNN. Both studies are based on the findings of Kim et al. [55] that demonstrate that CNNs can be taught differences between high- and low-resolution images (residual images) and that CNNs can be used to accurately and rapidly convert low-resolution images to high-resolution images. The articles propose identical CNN architectures of six convolutional layers with 3 x 3 kernel size. The CNNs were trained to learn the difference between the target and input images (residual image), and to create the converted image by adding the residual images to the input.

The developed CNN model by Du et al. [52] was also inspired by the model of Kim et al. [55]. However, the authors used perceptual loss in the training phase, which optimized the model in feature space instead of image space by minimizing the mean squared error between features.

U-net-based model Jin et al. [40] presented a kernel normalization network based on the U-net architecture for image-to-image translation from a sharp kernel image (input) to a standard (smoother) kernel image (target). The input and output nodes are connected with a sum operator that allows the model to learn the residual components, which is inspired by residual learning.

The study of Tanabe et al. [48] propose a standardization method that is U-net based as well. Contrary to Jin et al. [40], they introduce region-wise learning by training an extra network with images that were truncated in pixel values. By fusing the outputs of the two models by the weighted sum, the final converted image was generated, which resulted in a better performance.

3.2. GAN-based models

Between 2016 and 2018, GAN models have shown promising performance in image-to-image mapping tasks, which motivated Liang et al. [42] to utilize them as the foundation for their standardization method for CT scans from multiple sources. Unfortunately, GAN models do not contain any constraints to control what modes of data it shall generate, therefore synthesized images are not guaranteed to be similar to the target images. The authors address this challenge by proposing a novel DL framework, GANai, that makes the GAN model applicable for medical image synthesis, where great image details have to be maintained. This novel framework has a similar architecture as conditional GAN (cGAN), but with a significantly different training process. The architecture of a cGAN learns the conditional distribution of the source image given the target image, and then performs image transference from one domain to another, which makes it suitable for medical image synthesis [56]. More importantly, the authors introduce an alternative improvement training strategy which contains two alternate training phases, and enables a series of technical improvements, including phase-specific loss functions, phase-specific training data, and the adoption of ensemble learning.

With the architecture and alternative training strategy of GANai as foundation, Selim et al. [41] introduces STAN-CT, a DICOM-to-DICOM image standardization framework. The novel framework includes two new key components: the latent loss and the feature loss. The latent based loss function enforces one-to-one mapping between the synthesized image and the standard image. Whereas, the feature loss aims to improve generator diversity.

In [51] a GAN-based model is presented where both the generator and discriminator have a CNN architecture. The architecture of G was inspired by a residual feature aggregation network that was proposed for a single image super resolution task. After the first layers of spatial average pooling, convolution and activation by Leaky ReLU, the encoded features were subjected to two sequential hierarchical feature synthesis modules. Each module contains multiple feature attention blocks that each enhances the residual feature. After going through the whole module, a high-level feature map is encoded.

CycleGAN-based models A cycle-consistent adversarial network (cycleGAN) is an unsupervised image style transfer method that can learn to translate between two different domains without the use of paired training data [44]. The model uses, besides an adversarial loss, a cycle-consistency loss to ensure that the translation has one-to-one correspondence and maintains the underlying structure of the original images. The architecture typically consists of two generators and two discriminators. Two studies focused on using cycle GANs to convert CT scans into different imaging settings or conditions [2, 50].

Unlike the conventional cycleGAN, Yang et al. [50] employ a switchable cycleGAN architecture for kernel conversion. This approach utilizes a single conditional generator with adaptive instance normalization (AdaIN) for both forward and backward kernel conversion. This singular generator can produce every conceivable interpolating path along an optimal transport path between two target domains during the inference phase. The AdaIN layer calculates the mean and variance of style features to adjust the mean and variance of content features. Furthermore, split AdaIN code generators are incorporated to effectively use intermediate domain kernel images during training, which considerably enhances the feature domain interpolation performance.

Selim et al. [2] introduces a conventional cycle GAN architecture containing two generators and two discriminators. The 'master generator' is responsible for domain A image synthesis (intra-scanner standardization), where paired training data are provided. On the other hand, the other generator, is responsible for domain B data synthesis where unpaired images are provided (cross-scanner standardization). The authors tend to improve model generalizability using a comprehensive data augmentation approach that perturbs the inputs with adversarial noise generated from a Gaussian distribution.

Alternative GAN-based models Two articles propose alternative GAN-based models [53, 54]. Li et al. [53] adopted a modified architecture from the single GAN that consists of one generator and multiple discriminators. The generator was trained to establish a one-to-one mapping between three distinct domains and a target domain. It learned to convert images from each domain to the target domain and vice versa, by cycling through this process. This allowed the generator to learn without having access to pixel-aligned images from all domains in a single patient. The generator was able to capture the unique features of images in each domain and translate them to images in another domain with minimal loss of content.

Kim et al. [54] presents a routable GAN (routeGAN) architecture to address the limited scalability to multi-domain translation of CycleGAN. The key innovation of this architecture lies in the independent functions of the encoder and decoder components of the generator in image translation. To elaborate, the encoder is responsible for extracting shared latent information from data across multiple domains, while the decoder is trained to generate specific target domain images by transforming the shared latent feature vectors using different routing codes.

4. Performance evaluation

Two clear performance evaluation categories can be distinguished from the reviewed studies: the evaluation of image similarity and the evaluation of radiomic feature similarity between the target image and the converted (standardized) image. A total of six studies report a performance evaluation of the image similarity using varying metrics [40, 47–51]. The other category is evaluated by seven studies, again using a broad variety of metrics [2, 41, 42, 46, 51–53]. One study uses alternative metrics to evaluate the performance of their proposed standardization model [55].

An overview of the studies is given in the Tables in Appendix D summarizing the data sets used for performance evaluation. The studies are again divided based on their standardization type, and the characteristics of the data sets are outlined.

4.1. Image similarity

Different image similarity metrics have been performed as a means of evaluating performance, which share the common condition of requiring a pair of images as input. Image similarity metrics basically measure the difference between two images, which is used to check if a predicted/synthesized image is similar to its target image. Additionally, the structural similarity index (SSIM), attempts to take into account the quality of the image itself as well by considering the image structure.

Four articles only measure how similar the converted image is to its target image by comparing the CT values pixel-by-pixel using the mean-squared-error (MSE) [49] and peak signal-to-noise ratio (PSNR) [49], root mean squared error (RMSE) [47], difference in CT value [48] or lung density metrics, such as the mean lung attenuation (MLA) [40]. For each of these metrics, except PSNR, applies that a smaller value indicates an increased similarity. Conversely, a higher PSNR value suggests greater image similarity.

Lee et al. [47] report that the models trained on the conversion of sharp-to-smooth reconstructed images perform much better than the models trained vice versa. The RMSE is 53.48% reduced between the synthesized sharp images and their target image averaged over all smooth-to-sharp models, compared to 78.11% for the converted smooth images, also averaged. A potential reason for the improved performance may be attributed to the process of removing noise from the image while preserving its key features, as done in sharp-to-smooth conversion [57]. This approach results in a clearer and more precise image compared to the alternative method of enhancing previously weakened high-frequency components, which can introduce unwanted details and distortions. This alternative method is employed in smooth-to-sharp conversion. In addition, the very small data set should be taken into account when interpreting these results, only CT data acquired from two patients, 631 CT images in total, were included for the evaluation.

A large increase in the image similarity measured by lung density metrics was recorded by Jin et al. [40] for the conversion of images from sharp-to-smooth. They measured the relative lung area under 950 HU (RA950), the lower 15th percentile threshold (perc15) and the MLA and calculated the pair wise mean differences between the original image pair and between the converted and target image pair. This resulted in a reduction of 99.16% for RA950, 98.43% for perc15 and 84.75% for MLA pair wise differences averaged over all four models, each trained with images from a different scanner.

The other two articles did not report the difference in image similarity before and after standardization. Tanabe et al. [48] only showed the reduction in difference in CT values before and after conversion using graphs. Whereas, the study of Wu et al. [49] solely reports the MSE and PSNR between the converted image and its ground truth image, without showing the similarity before the conversion.

The SSIM quantifies the image quality degradation caused by processing, such as data compression, or losses in data transmission. SSIM in combination with the PSNR, not only image similarity is measured, but the perceptual difference between two images is also taken into account. Lee et al. [51] and Yang et al. [50] report this combination of image similarity metrics to evaluate the performance of their proposed model. The images acquired with a low-dose that were converted by the model of Lee et al. [51] showed a significantly higher SSIM (0.759 ± 0.023 vs 0.817 ± 0.003, P< 0.001) and PSNR (26.92 ± 2.21 vs 29.32 ± 1.21, P< 0.001) compared with the original images. On the contrary, converted images acquired with an equivalent dose or high dose did not show a significant difference in SSIM (0.825 ± 0.033 vs 0.824 ± 0.002, P = 0.394), and even showed a significantly lower PSRN (32.61 ± 1.66 vs 30.76 ± 0.13, P< 0.001) [51].

Yang et al. [50] did not compare the PSNR and SSIM values of their converted images with the original image. On the other hand, it did compare their findings with various conversion methods: classical kernel conversion using smoothing and sharpening, supervised learning using the MSE loss and conventional CycleGAN. The average PSNR and SSIM values obtained using their 2-domain switchable CycleGAN method compared with the other methods are given in Table 3. Their proposed method outperformed the classical conversion and cycleGAN methods for generating sharp and smooth images, however, only the difference with the classical conversion was significant. Although higher PSNR and SSIM values were obtained from conversion using supervised learning, the visual investigation showed that blurring artefacts were present in the converted images by supervised learning, which indicates the limitations of these quantitative metrics for the evaluation of the conversion model. The study also proposes a three domain learning, which slightly enhances the performance of the model in terms of PSNR and SSIM values. This model utilizes intermediate kernel images to generate other kernel images using self-consistency loss.

4.2. Radiomic feature similarity

An alternative approach to assess the effectiveness of a standardization method is to compare the radiomic features of the resulting image with those of the target images. Radiomic features are a set of quantitative features extracted from medical images, including CT scans, using image processing techniques [58]. These features can be used to analyse the spatial and temporal characteristics of a tumour or other regions of interest in the image and have the potential to provide important diagnostic, prognostic, and predictive information in a variety of medical applications. Radiomic features can be classified into several categories, such as intensity-based features, texture-based features, shape-based features, and wavelet-based features.

The similarity between radiomic features can be measured using various metrics, such as the absolute error (AE) [41, 42], patient-normalized feature difference (PNFD) [52], and the concordance correlation coefficient (CCC) [46, 51]. Subsequently, the reproducibility of the features can be determined by defining a threshold for these metrics, as been

TABLE 3: QUANTITATIVE COMPARISON OF VARIOUS METHODS IN TWO- AND MULTI-DOMAIN LEARNING, REPORTED BY YANG ET AL. [50]. 2-DOMAIN AND 3-DOMAIN SWITCHABLE ARE THE PROPOSED METHODS BY [50]

	PSNR		**SSIM**	
	to sharp	to smooth	to sharp	to smooth
Average data set 1 & 2, two-domain				
Classical method	12.7781	12.4653	0.5916	0.6390
Supervised	**30.4943**	**22.7256**	**0.8551**	0.8212
CycleGAN	28.2048	19.6636	0.8024	0.8064
2-domain switchable	29.1500	21.7453	0.8243	**0.8551**
Average data set 3, multi-domain				
3-domain switchable	**25.7611**	**19.2320**	0.7137	**0.8524**
2-domain switchable	25.3679	18.8973	**0.7345**	0.8139
CycleGAN	25.2707	17.6606	0.7084	0.8240

done by several studies [2, 46, 53].

The proposed standardization methods of Choe et al. [46] and Lee et al. [51] both assess feature similarity by calculating the CCC between the converted and target image and compare this with the CCC before standardization. The CCC measures the degree to which pairs of observations fall on a straight line [59]. It is based on the correlation coefficient, but also takes into account the agreement in the mean and variability of the two sets of measurements. The two studies included intensity-, texture- and wavelet-based features, and found that especially, intensity-based features showed better improvement compared to the wavelet-based features before and after image standardization [46, 51]. The performance evaluation performed by Du et al. [52] using the PNFD found similar results that indicate an improved reproducibility of the intensity-based features, and a poorer improved reproducibility of the wavelet features after conversion.

Separate models for sharp-to-smooth conversion and smooth-to-sharp conversion have been trained and evaluated by both Choe et al. [46] and Du et al. [52]. The conversion to smooth by Choe et al. [46] increased the number of reproducible features, defined as features with a CCC value higher than 0.85, from 107 (15.2%) to 460 (65.5%) features. Their model for the conversion of smooth-to-sharp showed an increase of 107 (15.2%) to 388 (55.3%) reproducible features. The mean pair wise difference of the PNFD before and after conversion has been calculated by Du et al. [52] for the assessment of their models. They reported an average of 60.88% decrease of the mean pair wise difference for the conversion of sharp-to-smooth, whereas the other model even showed an average decrease of 67.72%. It should also be noted that the models of Du et al. [52] has been trained on the conversion between the B30f (smooth) and B70f (sharp) kernels of Siemens while the evaluation was assessed on the conversion of a broad range of different kernels and scanners.

The multi-domain standardization method introduced by Li et al. [53] was evaluated by performing the Wilcoxon rank-sum test on the paired features between the converted and target images. This test is used to assess whether two samples are likely to derive from the same population, where the p-value is the probability that both populations are the same is true. They define features with a p-value greater than 0.05 as aligned features and tested in total 77 radiomic features. The percentage of aligned features before standardization is 10.4% between domain A and the target (T), 18.2% between domain B and T and 50.1% between C and T, after standardization these percentages increase to 93.5% (A vs T), 89.6% (B vs T) and 77.9% (C vs T).

The models introduced and evaluated by Liang et al. [42] (GANai), Selim et al. [41] (STAN-CT) and Selim et al. [2] (UDA-CT) are compared with each other in terms of RE and number of reproducible features in the most recent published study of Selim et al. [2]. A reproducible feature is defined as a feature with an absolute relative error below 0.15, and thus a similarity of 85% or higher. UDA-CT outperforms GANai and STAN-CT for two out of three assessed tumours for intra-scanner standardization. The results of the comparison are shown in Table 8 in Appendix E. Only intra-scanner standardization has been compared between these models, as GANai and STAN-CT do not include cross-scanner standardization, opposed to UDA-CT.

4.3. Alternative evaluation

Image and radiomic feature similarity can only be assessed when the converted image can be compared with the ground truth image, therefore image pairs have to be available. For the study of Kim et al. [55] this data was not accessible, therefore they address this issue by training an extra classifier for the domain classification which was used for the quantitative evaluation. The classifier evaluates the performance by checking whether their model produces an image that is classified as a target image by reflecting the characteristics of the target domain well. This means that if their model can convert images to the target domain appropriately, the classification accuracy should be high. They report that their proposed routable GAN model shows a relatively stable accuracy in all domains. Furthermore, the total accuracy is comparable to that of a Cycle-GAN based model.

Besides the classification accuracy, Kim et al. [55] also evaluated the translation results in the frequency domain by calculating the radial average (average along the radial direction in k-spaces). This can function as performance evaluation, because in medical image translation preserving low-frequency details, such as the overall content of images, while converting high-frequency details, such as sharp edges and artefacts in low dose images, is crucial. Thus, an appropriate translation of an image to the target domain requires similar high-frequency regions of the converted and target domain image without the change of the low-frequency region. According to the findings presented in [55], their method successfully produces converted images that present comparable intensity levels to those of the target domain in the high-frequency region of k-space.

V. Discussion

CT scans are routinely employed in clinical settings, aiding in diagnostic and treatment-related decision-making processes. However, due to the use of varying reconstruction techniques across different manufacturers and hospitals, image quality can differ significantly, posing a challenge for the development of reliable and universally applicable software. In response to this issue, a range of standardization techniques based on deep learning have been recently introduced to address these limitations.

In this paper, a systematic review investigating the state-of-the-art of deep learning approaches for standardization across CT scan variations due to reconstruction techniques. A search key was used to specifically look for studies reporting a DL based scan standardization method for conventional CT scans that standardizes images in the image-domain. In total, 68 studies were identified from the electronic database search, of which thirteen were included in the review after full-text eligibility assessment. The search reveals that since 2019 different DL models have been proposed for standardization of CT scans that serve different purposes: intra-scanner standardization or intra- and cross-scanner standardization. The studies are further compared based on their underlying architecture and their performance evaluation.

The techniques evaluated in this review all attempt to solve intra-scanner variation resulting from the utilization of diverse reconstruction kernels by kernel conversion from sharp-to-smooth or the reverse. Older models, that are based on a CNN architecture, were trained separately per distinctive kernel pair and conversion direction [40, 46, 47]. The rise of GAN-based deep learning models for image-to-image mapping tasks has motivated various authors to apply this method as foundation for CT scan standardization [2, 9, 41, 42, 50, 51, 53, 54]. Multiple variations on the GAN-based architecture have been introduced, such as an alternative training strategy [42], or the addition of latent and feature loss to enforce one-to-one mapping and improve generator diversity [41]. With the use of GAN-based architectures, models have been developed that utilize unpaired images [53, 55], and the combination of cross- and intra-scanner standardization is proposed for the first time [53]. The study of Yang et al. [50] is the only one that has shown promising results regarding standardization using the intermediate kernel images to translate images between any pairs of kernel domains along the interpolation path.

Despite the recent developments in DL based CT scan standardization, challenges remain. Firstly, in the majority of the reviewed articles, the size of the data set used for training is small (Tables in Appendix C). Besides, often only one scanner is used and a limited amount of reconstruction kernels are included. On top of this, the training data sets are usually outdated, which means only old scanner models, and thus reconstruction techniques and kernels, are involved in the training process. Only one study attempted to convert FBP reconstruction to the more novel reconstruction technique, iterative reconstruction [51]. All together, this means that it is hard to predict how the reviewed methods work on images acquired from different scanners, e.g. newer models of the same vendor, and thus newer reconstruction techniques, or scanners from other manufacturers. Moreover, tackling this challenge is complicated as most of the reviewed strategies rely on paired images for training, which are very difficult to acquire over a wide range of reconstruction kernels. In addition, acquiring such pairs from different scanners is nearly impossible since the images have to be captured from a single patient.

Secondly, evaluating the potential of each method is challenging because the studies employ a broad range of evaluation metric, which makes it impossible to compare the models directly. Furthermore, most studies do not compare their model with other models within their research, further complicating the assessment of their effectiveness.

Finally, the suitability of commonly used similarity metrics for evaluating the performance of medical image translation methods is questionable. Content preservation is the most critical aspect of a standardization method as it ensures that important textures in the image, which are essential for radiologists or algorithms to inform a diagnosis, are not lost. Therefore, the effectiveness of a standardization method cannot be solely determined based on similarity metrics

1. Limitations

Several potential limitations can be identified for this review. The literature search was restricted to studies that presented a novel standardization method for CT scans to handle the variation across reconstruction parameters. DL-based models that standardize other medical images, such as MRI images, or those that standardize across acquisition parameters, such as dosage, may also serve as a valuable source of inspiration for standardizing CT images. For instance, Liu et al. [60] carried out a study to harmonize MRI images from multiple arbitrary sites using a style transferable GAN. They demonstrated that their model was effective on previously unseen images, as long as enough data from multiple sites was available for training. Furthermore, Wei et al.'s study [61] utilized a 3D GAN to standardize CT images acquired with varying slice thicknesses and dosage scenarios, which is a unique approach as it simultaneously performs denoising and super-resolution. Although the standardization of medical scans other than CT scans, as well as standardization of variations caused by factors other than reconstruction parameters, were not within the scope of this review, it would be worthwhile to analyse these areas in future studies.

Secondly, only methods that work on their own were included, which resulted in the exclusion of domain adaptation techniques. Domain adaptation is a field of computer vision, where the goal is to train a neural network on a source dataset and secure a good accuracy on the target dataset, which is significantly different from the source dataset [62]. This approach has also been proposed as solution for the heterogeneity in CT scans, e.g. by Xu et al. [63]. They integrate an unsupervised content-preserved adaptation network in a pulmonary texture classification network to alleviate the performance degradation caused when applied to data from other scanners. Utilizing domain adaptation techniques, as suggested in existing literature, could be a promising solution for addressing CT scan variations when utilizing DL-based image analysis software to analyse them.

Lastly, due to the variation between the included studies, a quantitative comparison was not feasible. The studies differ in several aspects, including performance evaluation methods, data used for model training and evaluation, and conversion direction. Another factor preventing quantitative comparison of the models, is the lack of comparison within the studies itself with other models by performing standardization on the same data set. In the future, it would be highly valuable to conduct a research comparing these standardization methods by training and evaluating on the same data set with a sufficient size.

VI. Conclusion

Several CT scan standardization techniques based on deep learning have the potential to address the difficulties in comparison and analysis of CT images across different scanners and institutions. Alterations of GAN-based models report the most promising results in terms of usage of non-paired images, content preservation, both forward and backward mapping and combined intra- and cross-standardization.

However, the strategies available in the literature also suggest that further research is needed before these innovative standardization approaches can enter the clinic in a stable manner. Large-scale research from multiple sites is necessary to quantitatively compare the standardization techniques. Furthermore, conducting extensive research on the effects of reconstruction techniques and kernels from diverse vendors and models on CT patient scans could yield valuable insights. These findings could then be utilized to develop a generalized and resilient standardization approach that can be implemented across multiple CT scanner manufacturers.

In addition, valuable insights for CT standardization of reconstruction effects may be gained from standardization techniques developed for other imaging modalities, as well as from strategies for standardizing variation caused by factors unrelated to reconstruction methods.

References

[1] Z. T. Al-Sharify, T. A. Al-Sharify, N. T. Al-Sharify *et al.*, "A critical review on medical imaging techniques (ct and pet scans) in the medical field," in *IOP Conference Series: Materials Science and Engineering*, vol. 870, no. 1. IOP Publishing, 2020, p. 012043.

[2] M. Selim, J. Zhang, B. Fei, M. Lewis, G.-Q. Zhang, and J. Chen, "Uda-ct: A general framework for ct image standardization," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1698–1701.

[3] M. Chun, J. H. Choi, S. Kim, C. Ahn, and J. H. Kim, "Fully automated image quality evaluation on patient ct: Multi-vendor and multi-reconstruction study," *PLOS ONE*, vol. 17, p. e0271724, 7 2022.

[4] J. Greffier, J. Frandon, A. Larbi, J. P. Beregi, and F. Pereira, "Ct iterative reconstruction algorithms: a task-based image quality assessment," *European Radiology*, vol. 30, pp. 487–500, 1 2020.

[5] S. P. Blazis, D. B. Dickerscheid, P. V. Linsen, and C. O. M. Jarnalo, "Effect of ct reconstruction settings on the performance of a deep learning based lung nodule cad system," *European Journal of Radiology*, vol. 136, p. 109526, 2021.

[6] B. S. Erdal, M. Demirer, K. J. Little, C. C. Amadi, G. F. Ibrahim, T. P. O'Donnell, R. Grimmer, V. Gupta, L. M. Prevedello, and R. D. White, "Are quantitative features of lung nodules reproducible at different ct acquisition and reconstruction parameters?" *PLoS One*, vol. 15, no. 10, p. e0240184, 2020.

[7] N. Emaminejad, M. W. Wahi-Anwar, G. H. J. Kim, W. Hsu, M. Brown, and M. McNitt-Gray, "Reproducibility of lung nodule radiomic features: multivariable and univariable investigations that account for interactions between ct acquisition and reconstruction parameters," *Medical physics*, vol. 48, no. 6, pp. 2906–2919, 2021.

[8] S. Denzler, D. Vuong, M. Bogowicz, M. Pavic, T. Frauenfelder, S. Thierstein, E. I. Eboulet, B. Maurer, J. Schniering, H. S. Gabryś *et al.*, "Impact of ct convolution kernel on robustness of radiomic features for different lung diseases and tissue types," *The British journal of radiology*, vol. 94, no. 1120, p. 20200947, 2021.

[9] G. Wu, A. Ibrahim, I. Halilaj, R. T. Leijenaar, W. Rogers, H. A. Gietema, L. E. Hendriks, P. Lambin, and H. C. Woodruff, "The emerging role of radiomics in copd and lung cancer," *Respiration*, vol. 99, no. 2, pp. 99–107, 2020.

[10] R. Schofield, L. King, U. Tayal, I. Castellano, J. Stirrup, F. Pontana, J. Earls, and E. Nicol, "Image reconstruction: Part 1 – understanding filtered back projection, noise and image acquisition," *Journal of Cardiovascular Computed Tomography*, vol. 14, pp. 219–225, 5 2020.

[11] M. Kalke and S. Siltanen, "Sinogram interpolation method for sparse-angle tomography," *Applied Mathematics*, vol. 2014, 2014.

[12] S. Schafer and J. H. Siewerdsen, "Technology and applications in interventional imaging: 2d x-ray radiography/fluoroscopy and 3d cone-beam ct," in *Handbook of Medical Image Computing and Computer Assisted Intervention*. Elsevier, 2020, pp. 625–671.

[13] R. C. Young and C. R. Chatwin, "Computation of the forward and inverse radon transform via the central slice theorem employing a nonscanning optical technique," in *Optical Pattern Recognition VII*, vol. 2752. SPIE, 1996, pp. 306–316.

[14] M. J. Willemink and P. B. Noël, "The evolution of image reconstruction for ct—from filtered back projection to artificial intelligence," *European Radiology*, vol. 29, p. 2185, 5 2019.

[15] W. Stiller, "Basics of iterative reconstruction methods in computed tomography: A vendor-independent overview," *European Journal of Radiology*, vol. 109, pp. 147–154, 12 2018.

[16] T. P. Szczykutowicz, G. V. Toia, A. Dhanantwari, and B. Nett, "A review of deep learning ct reconstruction: Concepts, limitations, and promise in clinical practice," *Current Radiology Reports*, vol. 10, pp. 101–115, 9 2022. [Online]. Available: https://link-springer-com.tudelft.idm.oclc.org/article/10.1007/s40134-022-00399-5

[17] T. Peters, "Algorithms for fast back-and re-projection in computed tomography," *IEEE transactions on nuclear science*, vol. 28, no. 4, pp. 3641–3647, 1981.

[18] J. A. Seibert, "Iterative reconstruction: how it works, how to apply it," *Pediatric radiology*, vol. 44, pp. 431–439, 2014.

[19] L. L. Geyer, U. J. Schoepf, F. G. Meinel, J. W. Nance, G. Bastarrika, J. A. Leipsic, N. S. Paul, M. Rengo, A. Laghi, and C. N. D. Cecco, "State of the art: Iterative ct reconstruction techniques1," *Radiology*, vol. 276, pp. 339–357, 8 2015.

[20] D. Mehta, R. Thompson, T. Morton, A. Dhanantwari, and E. Shefer, "Iterative model reconstruction: simultaneously lowered computed tomography radiation dose and improved image quality," *Med Phys Int J*, vol. 2, no. 1, pp. 147–55, 2013.

[21] S. Gordic, L. Desbiolles, M. Sedlmair, R. Manka, A. Plass, B. Schmidt, D. B. Husarik, F. Maisano, S. Wildermuth, H. Alkadhi *et al.*, "Optimizing radiation dose by using advanced modelled iterative reconstruction in high-pitch coronary ct angiography," *European radiology*, vol. 26, pp. 459–468, 2016.

[22] A. D. Hardie, R. M. Nelson, R. Egbert, W. J. Rieter, and S. V. Tipnis, "What is the preferred strength setting of the sinogram-affirmed iterative reconstruction algorithm in abdominal ct imaging?" *Radiological physics and technology*, vol. 8, pp. 60–63, 2015.

[23] L. Liu, "Model-based iterative reconstruction: a promising algorithm for today's computed tomography imaging," *Journal of Medical imaging and Radiation sciences*, vol. 45, no. 2, pp. 131–136, 2014.

[24] A. Löve, M. L. Olsson, R. Siemund, F. Stålhammar, I. M. Björkman-Burtscher, and M. Söderberg, "Six iterative reconstruction algorithms in brain ct: A phantom study on image quality at different radiation dose levels," *British Journal of Radiology*, vol. 86, 11 2013.

[25] M. Beister, D. Kolditz, and W. A. Kalender, "Iterative reconstruction methods in x-ray ct," *Physica Medica*, vol. 28, pp. 94–108, 4 2012.

[26] S. Skornitzke, "Iterative algorithms for artifact reduction in computed tomography," *Der Radiologe*, vol. 58, pp. 202–210, 2018.

[27] J. G. Fletcher, S. Leng, L. Yu, and C. H. McCollough, "Dealing with uncertainty in ct images," pp. 5–10, 2016.

[28] Z. Zhang and E. Seeram, "The use of artificial intelligence in computed tomography image reconstruction-a literature review," *Journal of medical imaging and radiation sciences*, vol. 51, no. 4, pp. 671–677, 2020.

[29] C. Arndt, F. Güttler, A. Heinrich, F. Bürckenmeyer, I. Diamantis, and U. Teichgräber, "Deep learning ct image reconstruction in clinical practice," in *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 193, no. 03. Georg Thieme Verlag KG, 2021, pp. 252–261.

[30] J. Hsieh, E. Liu, B. Nett, J. Tang, J.-B. Thibault, and S. Sahney, "A new era of image reconstruction: Truefidelity™," *White Paper (JB68676XX), GE Healthcare*, 2019.

[31] D. Jakhar and I. Kaur, "Artificial intelligence, machine learning and deep learning: definitions and differences," *Clinical and experimental dermatology*, vol. 45, no. 1, pp. 131–132, 2020.

[32] A. Krenker, J. Bešter, and A. Kos, "Introduction to the artificial neural networks," *Artificial Neural Networks: Methodological Advances and Biomedical Applications. InTech*, pp. 1–18, 2011.

[33] S.-H. Han, K. W. Kim, S. Kim, and Y. C. Youn, “Artificial neural network: understanding the basic concepts without mathematics,” *Dementia and neurocognitive disorders*, vol. 17, no. 3, pp. 83–89, 2018.

[34] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[35] R. Hecht-Nielsen, “Theory of the backpropagation neural network,” pp. 65–93, 1992.

[36] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, “Review of image classification algorithms based on convolutional neural networks,” *Remote Sensing*, vol. 13, no. 22, p. 4712, 2021.

[37] A. Anaya-Isaza, L. Mera-Jiménez, and M. Zequera-Diaz, “An overview of deep learning in medical imaging,” *Informatics in Medicine Unlocked*, vol. 26, p. 100723, 1 2021.

[38] M. Ribeiro, A. E. Lazzaretti, and H. S. Lopes, “A study of deep convolutional auto-encoders for anomaly detection in videos,” *Pattern Recognition Letters*, vol. 105, pp. 13–22, 2018.

[39] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[40] H. Jin, C. Heo, and J. H. Kim, “Deep learning-enabled accurate normalization of reconstruction kernel effects on emphysema quantification in low-dose ct,” *Physics in Medicine & Biology*, vol. 64, no. 13, p. 135010, 2019.

[41] M. Selim, J. Zhang, B. Fei, G.-Q. Zhang, and J. Chen, “Stan-ct: Standardizing ct image using generative adversarial networks,” in *AMIA Annual Symposium Proceedings*, vol. 2020. American Medical Informatics Association, 2020, p. 1100.

[42] G. Liang, S. Fouladvand, J. Zhang, M. A. Brooks, N. Jacobs, and J. Chen, “Ganai: Standardizing ct images using generative adversarial network with alternative improvement,” in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2019, pp. 1–11.

[43] F. Zhang, J. Bai, J. Zhang, Z. Xiao, and C. Pei, “An optimized training method for gan-based hyperspectral image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 10, pp. 1791–1795, 2020.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[45] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks *et al.*, “Prisma extension for scoping reviews (prisma-scr): checklist and explanation,” *Annals of internal medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[46] J. Choe, S. M. Lee, K.-H. Do, G. Lee, J.-G. Lee, S. M. Lee, and J. B. Seo, “Deep learning–based image conversion of ct reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses,” *Radiology*, vol. 292, no. 2, pp. 365–373, 2019.

[47] S. M. Lee, J.-G. Lee, G. Lee, J. Choe, K.-H. Do, N. Kim, and J. B. Seo, “Ct image conversion among different reconstruction kernels without a sinogram by using a convolutional neural network,” *Korean journal of radiology*, vol. 20, no. 2, pp. 295–303, 2019.

[48] N. Tanabe, S. Kaji, H. Shima, Y. Shiraishi, T. Maetani, T. Oguma, S. Sato, and T. Hirai, “Kernel conversion for robust quantitative measurements of archived chest computed tomography using deep learning-based image-to-image translation,” *Frontiers in Artificial Intelligence*, vol. 4, p. 209, 2022.

[49] Q. Wu, W. Huang, S. Wang, H. Yu, L. Wang, Z. Wu, Y. Zhu, Z. Liu, H. Ma, and J. Tian, “A generative adversarial network-based ct image standardization model for predicting progression-free survival of lung cancer,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3411–3414.

[50] S. Yang, E. Y. Kim, and J. C. Ye, “Continuous conversion of ct kernel using switchable cyclegan with adain,” *IEEE transactions on medical imaging*, vol. 40, no. 11, pp. 3015–3029, 2021.

[51] S. B. Lee, Y. J. Cho, Y. Hong, D. Jeong, J. Lee, S.-H. Kim, S. Lee, and Y. H. Choi, “Deep learning-based image conversion improves the reproducibility of computed tomography radiomics features: a phantom study,” *Investigative Radiology*, vol. 57, no. 5, pp. 308–317, 2022.

[52] D. Du, W. Lv, J. Lv, X. Chen, H. Wu, A. Rahmim, and L. Lu, “Deep learning–based harmonization of ct reconstruction kernels towards improved clinical task performance,” *European Radiology*, pp. 1–13, 2022.

[53] Y. Li, G. Han, X. Wu, Z. H. Li, K. Zhao, Z. Zhang, Z. Liu, and C. Liang, “Normalization of multicenter ct radiomics by a generative adversarial network method,” *Physics in Medicine & Biology*, vol. 66, no. 5, p. 055030, 2021.

[54] H. Kim, G. Oh, J. B. Seo, H. J. Hwang, S. M. Lee, J. Yun, and J. C. Ye, “Multi-domain ct translation by a routable translation network,” *Physics in Medicine & Biology*, vol. 67, no. 21, p. 215002, 2022.

[55] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[56] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.

[57] C. Pérez-Benito, S. Morillas, C. Jordán, and J. A. Conejero, “Smoothing vs. sharpening of colour images: Together or separated,” *Applied mathematics and nonlinear sciences*, vol. 2, no. 1, pp. 299–316, 2017.

[58] M. R. Tomaszewski and R. J. Gillies, “The biological meaning of radiomic features,” *Radiology*, vol. 298, no. 3, pp. 505–516, 2021.

[59] I. Lawrence and K. Lin, “A concordance correlation coefficient to evaluate reproducibility,” *Biometrics*, pp. 255–268, 1989.

[60] M. Liu, P. Maiti, S. Thomopoulos, A. Zhu, Y. Chai, H. Kim, and N. Jahanshad, “Style transfer using generative adversarial networks for multisite mri harmonization,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 313–322.

[61] L. Wei, Y. Lin, and W. Hsu, “Using a generative adversarial network for ct normalization and its impact on radiomic features,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 844–848.

[62] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, “A brief review of domain adaptation,” *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[63] R. Xu, Z. Cong, X. Ye, S. Kido, and N. Tomiyama, “Unsupervised content-preserved adaptation network for classification of pulmonary textures from different ct scanners,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1060–1064.

[64] R. C. Nelson, S. Feuerlein, and D. T. Boll, “New iterative reconstruction techniques for cardiovascular computed tomography: how do they work, and what are the advantages and disadvantages?” *Journal of cardiovascular computed tomography*, vol. 5, no. 5, pp. 286–292, 2011.

[65] K. Grant and T. Flohr, “Iterative reconstruction in image space (iris),” *White Paper-Siemens Technical Report*, 2010.

[66] K. Grant and R. Raupach, “Safire: Sinogram affirmed iterative reconstruction,” *Whitepaper, Siemens AG*, 2012.

[67] S. Baumueller, A. Winklehner, C. Karlo, R. Goetti, T. Flohr, E. W. Russi, T. Frauenfelder, and H. Alkadhi, “Low-dose ct of the lung: potential value of iterative reconstructions,” *European radiology*, vol. 22, pp. 2597–2606, 2012.

[68] J. Leipsic, T. M. Labounty, B. Heilbron, J. K. Min, G. J. Mancini, F. Y. Lin, C. Taylor, A. Dunning, and J. P. Earls, “Adaptive statistical iterative reconstruction: assessment of image noise and image quality in coronary ct angiography,” *American Journal of Roentgenology*, vol. 195, no. 3, pp. 649–654, 2010.

[69] J. Fan, M. Yue, and R. Melnyk, “Benefits of asir-v reconstruction for reducing patient radiation dose and preserving diagnostic quality in ct exams,” *White paper, GE Healthcare*, 2014.

[70] A. Scibelli, “idose4 iterative reconstruction technique. philips healthcare whitepaper,” 2011.

[71] A. Laqmani, J. Buhk, F. Henes, T. Klink, S. Sehner, H. Von Schultzendorff, D. Hammerle, H. Nagel, G. Adam, and M. Regier, “Impact of a 4th generation iterative reconstruction technique on image quality in low-dose computed tomography of the chest in immunocompromised patients,” in *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, vol. 185, no. 08, 2013, pp. 749–757.

[72] F. Tatsugami, M. Matsuki, G. Nakai, Y. Inada, S. Kanazawa, Y. Takeda, H. Morita, H. Takada, S. Yoshikawa, K. Fukumura *et al.*, “The effect of adaptive iterative dose reduction on image quality in 320-detector row ct coronary angiography,” *The British journal of radiology*, vol. 85, no. 1016, pp. e378–e382, 2012.

[73] R. Irwan, S. Nakanishi, and A. Blum, “Aidr 3d–reduces dose and simultaneously improves image quality,” *Toshiba Med Syst*, pp. 1–8, 2011.

[74] A. Gervaise, B. Osemont, S. Lecocq, A. Noel, E. Micard, J. Felblinger, and A. Blum, “Ct image quality improvement using adaptive iterative dose reduction with wide-volume acquisition on 320-detector ct,” *European radiology*, vol. 22, pp. 295–301, 2012.

[75] J. Thibault, “Veo ct model-based iterative reconstruction,” *White Papper GE Healthcare*, pp. 1–12, 2010.

[76] E. Maeda, N. Tomizawa, S. Kanno, K. Yasaka, T. Kubo, K. Ino, R. Torigoe, and K. Ohtomo, “The feasibility of forward-projected model-based iterative reconstruction solution (first) for coronary 320-row computed tomography angiography: a pilot study,” *Journal of Cardiovascular Computed Tomography*, vol. 11, no. 1, pp. 40–45, 2017.

[77] R. M. Joemai and J. Geleijns, “Forward projected model-based iterative reconstruction solution “first”,” *Tustin, CA: Toshiba Medical Systems Corporation*, vol. 7, 2017.

[78] L. R. Koetzier, D. Mastrodicasa, T. P. Szczykutowicz, N. R. van der Werf, A. S. Wang, V. Sandfort, A. J. van der Molen, D. Fleischmann, and M. J. Willemink, “Deep learning image reconstruction for ct: Technical principles and clinical prospects,” *Radiology*, p. 221257, 2023.

[79] “Ai for significantly lower dose and improved image quality. philips healthcare whitepaper,” 2021.

[80] K. Boedeker, “Aice deep learning reconstruction: bringing the power of ultra-high resolution ct to routine imaging,” *Canon Medical Systems Corporation*, 2019.

A. Commercially available reconstruction techniques

Reconstruction method	Vendor	Explanation
Hybrid iterative reconstruction		
Iterative reconstruction in image space (IRIS)	Siemens Healthineers	IRIS is based in image space, where it reconstructs an image from the raw data using three to five iterations [19, 64]. IRIS allows for dose reduction up to 60%, while maintaining spatial and low-contrast resolution and not affecting image texture according to the manufacturer [65].
Sinogram-affirmed iterative reconstruction (SAFIRE)	Siemens Healthineers	SAFIRE estimates the initial reconstruction by weighted FBP, from which new synthetic raw data is calculated using forward projection. The synthetic data is compared with the original raw data to reconstruct a correction image and update the original image. This loop is repeated a number of times and corrects imperfections and removes artefacts [66, 67]. Within each iteration, regularization is also performed to reduce image noise while maintaining image structures. SAFIRE employs a precise local image noise model that is derived by analysing the statistical significance of the raw data contributing to that pixel in the raw data sinogram [19, 67]. SAFIRE can control image impression and noise reduction through the five strength levels that are available for adaptation of the regularization term, with strength 1 being noisier and strength 5 being smoother [66, 67].
Adaptive statistical iterative reconstruction (ASIR(-V))	GE Healthcare	ASIR involves comparing a measured projection to a modeled projection, which is based on the system statistics. It uses information generated by the FBP algorithm as a building block. The difference between the two projections is used to update the original projection, and this process is repeated until the final estimated projection ultimately converges to the ideal projection. ASIR reduces image noise, however a higher percentage of ASIR can lead to a decline in image quality [19, 68]. ASIR-V is the next generation of ASIR and compared to ASIR it contains more advanced noise modeling and object modeling, and physics modeling has been added to the process [69].
iDose^4	Philips Healthcare	The iDose^4 algorithm starts by analyzing the projection data to identify and correct the noisiest raw CT data. Through an iterative diffusion process, the noisy data is penalized and edges are preserved. Following this process, uncorrelated noise that remains is propagated to the image space, which is highly localized and therefore can be effectively removed through iterations [19, 70]. The strength of the IR algorithms can be ranged from level 1 to 7, increasing levels indicate increase noise reduction [70, 71]
Adaptive iterative dose reduction (AIDR 3D)	Canon Healthcare	The original AIDR algorithm applied image noise reduction in the image domain, which required that the original high-noise images undergo several loops of iteration until the desired noise level is achieved [72]. AIDR 3D is the replacement of the original technique. It uses a 3D processing algorithm and performs IR in both the image and raw-data domain. In the raw-data domain, a statistical noise and a CT model are used together with projection noise estimation to reduce electronic noise. The reconstructions are then optimized by an iterative technique that detects and preserves sharp details and smooths the images at the same time. Finally, the output image is created by blending the initial reconstruction with the final iterative image [73, 74].
Model-based iterative reconstruction		
Advanced modeled iterative reconstruction (ADMIRE)	Siemens Healthineers	ADMIRE is an iterative algorithm for image reconstruction in CT that incorporates an adaptive regularization term to control the smoothness of the image estimate while preserving image features. It is an effective technique for producing high-quality images from noisy projection data.
Veo	GE Healthcare	Veo includes an extensive 3D model of the data acquisition process that considers various factors such as the shape of the beam as it leaves the X-ray source and to the focal spot. The interaction of the beam with the patient, and with the X-ray detector, are also taken into account [64, 75]. Due to its incorporation of system statistics and physics as well as multiple back and forward reconstructions, Veo is very time-consuming [64].
Iterative model reconstruction (IMR)	Philips Healthcare	IMR takes into account data statistics, image statistics and system models and applies forward and backward reconstruction steps. The algorithm provides the user some control over the desired image characteristics by incorporating knowledge that constraints the optimization [20].

Forward projected iterative reconstruction solution (FIRST)	Canon Medical Systems	FIRST is a full IR algorithm that jointly enhances image quality in both the sinogram and image domains. The technique integrates a forward and statistical model into the projection data fidelity term, which leads to high spatial resolution and reduced noise streaks, respectively. Moreover, an organ-specific regularization process is subjected to the images to further reduce image noise, e.g. in the bone, heart, lung and abdomen. After multiple iterations, the resulting pair of images are combined to produce the final converged solution that incorporates the benefits of increased spatial resolution and decreased image noise [76, 77].
Deep learning based reconstruction		
True Fidelity	GE Healthcare	The True Fidelity DLR utilises deep CNNs (DCNNs) that are trained with high quality images obtained through high dose FBP, so the DCNNs learn to differentiate between noise and signal [30, 78]. The goal is to produce images of comparable quality from low-dose examination. This is achieved by noise reduction while restoring preferred noise texture, that therefore improves overall image quality when compared to other reconstruction methods. The algorithm reconstructs images directly from input sinogram acquired at low radiation doses [28, 30]. The resulting images retain FBP-like noise texture, sharpness, and artefact properties because the model was trained on FBP images [16].
Precise Image	Philips Healthcare	For Precise Image a CNN was trained on lower-dose simulated sinograms (input) and matched routine-dose FBP images as the ground truth. Simulated noise was introduced prior to reconstruction. The CNN was trained on patient data from different groups and populations, with a range of scan parameters [79]
AiCE	Canon Medical Systems	AiCE is a DLR that uses DCNNs to produce high-quality MBIR images from hybrid-IR images without the longer processing time that is associated with MBIR [80]. The difference between MBIR and hybrid-IR images are given as feedback to be ’learned’ and ’updated’ by the DCNN. Before DCNN-based restoration, the input data from the scan undergoes data domain filtering and hybrid-IR reconstruction [78]. Because MBIR images are used ground truth, the technique allows for artifact reduction due to its ability to model system optics, system physics, scanner statistical properties and human anatomy [78, 80]

B. Search terms

Pubmed: ("CT" [tiab] OR "computed tomography"[tiab] OR "Tomography, X-Ray Computed" [MeSH]) AND ("generalization" [tiab] OR "harmonization" [tiab] OR "domain adaptation" [tiab] OR "standardization" [tiab] OR "homogenizing" [tiab] OR "conversion" [tiab] OR "normalization" [tiab] OR ("characterizing" [tiab] AND "matching" [tiab])) AND ("reconstruction setting*" [tiab] OR "reconstruction algorithm*" [tiab] OR "reconstruction method*" [tiab] OR "reconstruction kernel*" [tiab] OR "reconstruction filter*" [tiab] OR "convolution kernel*" [tiab] OR "convolution filter*" [tiab] OR "imaging protocol*" [tiab]) AND ("deep learning" [all fields] OR "DL" [all fields] OR "CNN" [all fields] OR "GAN" [all fields] OR "neural network*" [all fields] OR "Deep Learning" [MeSH])

IEEE Xplore: (("Abstract": CT OR "Abstract": computed tomography) AND ("Abstract": generalization OR "Abstract": harmonization OR domain adaptation OR "Abstract": homogenizing OR "Abstract": conversion OR "Abstract": normalization OR "Abstract": standardization) AND ("Abstract": reconstruction settings OR "Abstract": reconstruction kernels OR "Abstract": reconstruction algorithms OR "Abstract": reconstruction methods OR "Abstract": reconstruction kernels OR "Abstract": reconstruction filters OR "Abstract": convolution kernels OR "Abstract": convolution filters OR "Abstract": imaging protocols) AND ("All Metadata": deep learning OR "All Metadata": DL OR "All Metadata": CNN OR "All Metadata": GAN OR "All Metadata": neural network*))

C. Method and training data set overview

Table 4: Overview of data sets used for training in the included studies in this systematic review that propose a cross- and intra-scanner standardization method.

Article	Type	Conversion	Data set	Paired data	Tissue	Series (n)	Slices (n)	Data augmentation	Scanner	Scanner type	Kernel (pairs)	Source data	Target data
Du et al. [52]	Cross- & Intra-scanner	Sharp-to-smooth & Smooth-to-sharp	Private	Yes	Chest	30	N/A	No	Siemens	N/A	B30f & B70f	B30f	B70f
												B70f	B30f
Kim et al. [54]	Cross- & Intra-scanner	Sharp-to-smooth & Smooth-to-sharp	Private	No	Chest	218	18.142	Yes	Siemens	N/A	B60f	All possible combinations	
											B70f		
											B46f		
			Private	No	Chest	214	26.473	Yes	GE	N/A	BONE		
			Private	No	Chest	125	33.797	Yes	Philips	N/A	YC		
			Private	No	Chest	45	9133	Yes	Toshiba	N/A	FC85		
Selim et al. [2]	Cross- & Intra-scanner	Smooth-to-sharp	Private	Yes	Chest	N/A	9900	Yes	Siemens	Force	Br40 & Bl64	Br40	Bl64
			TCIA	No	Chest	N/A	4255	Yes	GE	Revolution EVO	LUNG	LUNG	Bl46
Li et al. [53]	Cross & Intra-scanner	N/A	Private	No	Chest	60	10.000	No	GE	N/A	N/A	GE	Siemens^{T}
			Private	No	Chest	60	10.000	No	Philips	N/A	N/A	Philips	Siemens^{T}
			Private	No	Chest	60	10.000	No	Siemens	N/A	N/A	Siemens	Siemens^{T}
			Private	No	Chest	60	10.000	No	Siemens^{T}	N/A	N/A		

Table 5: Overview of data sets used for training by the included studies in this systematic review that propose an intra-scanner standardization method.

Article	Type	Conversion	Data set	Paired data	Tissue	Series (n)	Slices (n)	Data augmentation	Scanner	Scanner type	Kernel (pairs)	Source data	Target data
Choe et al. [46]	Intra-scanner	Sharp-to-smooth & Smooth-to-sharp	Private	Yes	Chest	40	N/A	No	Siemens	Definition Edge	B30f & B50f	B30f	B50f
												B50f	B30f
Lee et al. [47]	Intra-scanner	Sharp-to-smooth & Smooth-to-sharp	Private	Yes	Chest	8	2.669	No	Siemens	Definition Edge	B10f & B30f & B50f & B70f	All possible combinations	
Yang et al. [50]	Intra-scanner	Sharp-to-smooth & Smooth-to-sharp	Private	Yes	Head	9	450	Yes	Siemens	Definition Edge	J30s & J70h	J30s	J70h
												J70h	J30s
			Private	Yes	Facial bone	10	1.355	Yes	Siemens	Definition Edge	J40s & J70h	J40s	J70h
												J70h	J40s
			Private	Yes	Facial bone	7	1.282	Yes	Siemens	Definition Edge	Hr40 & Hr49 & Hr68	All possible combinations	
Jin et al. [40]	Intra-scanner	Sharp-to-smooth	NLST	Yes	Chest	111	15.317	No	Siemens	Volume zoom	B30f & B50f	B50f	B30f
						44	6.308	No	GE	Lightspeed 16	STANDARD & BONE	BONE	STANDARD
						14	2.711	No	Philips	Mx8000	C & D	D	C
						6	991	No	Canon	Aquilion	FC10 & FC51	FC51	FC10
Tanabe et al. [48]	Intra-scanner	Sharp-to-smooth	Private	Yes	Chest	30	11.052	No	Canon	Aquilion Prime	FC13 & FC51	FC51	FC13
Liang et al. [42]	Intra-scanner	Smooth-to-sharp	Private	Yes	Chest	N/A	2448	yes	Siemens	Force	Bl57 & Bl64	Bl57	Bl64
Selim et al. [41]	Intra-scanner	Smooth-to-sharp	Private	Yes	Chest	N/A	14688	yes	Siemens	Force	Bl57 & Bl64 & Br40	Bl57	Bl64
												Br40	Bl64
Wu et al. [49]	Intra-scanner	Smooth-to-sharp	Private	Yes	Chest	174	8352	Yes	Siemens	Definition AS	5-mm B30f & 5-mm B31f & 5-mm B60f & 5-mm B80f	5-mm B30f	1-mm sharp
												5-mm B31f	1-mm sharp
												5-mm B60f	1-mm sharp
												5-mm B80f	1-mm sharp
Lee et al. [51]	Intra-scanner	N/A	Private	Yes	Phantom	60	N/A	No	Siemens	Definition Flash	FBP 30f & SAFIRE 30f	FBP 30f	SAFIRE 30f

D. Performance evaluation data set overview

Table 6: Overview of data sets used for evaluation in the included studies in this systematic review that propose a cross- and intra-scanner standardization method.

Article	Target	Data set	Paired data	Tissue	Series (n)	Slices (n)	Scanner	Scanner type	Source data
Du et al. [52]	B30f or B70f	Private	Yes	Chest	85	N/A	Siemens	Definition AS	B30f & B70f
		Private	Yes	Chest	164	N/A	Siemens	Definition AS	B30f & B70f
		Private	Yes	Phantom	22	N/A	GE	Discovery STE	Soft, Detail, Standard, Lung, Edge
							Philips	Brilliance 64	A, B, C, L, YA
							Siemens	Definition AS	I26f-2, I30f-2, I40f-2, I44f-2, I50f-2, I70f-2
							Siemens	Sensation 64	B10f, B20f, B30f, B50f, B60f, B70f
Kim et al. [54]	B60f, B70f, B46f, BONE, YC or FC85	Private	No	Chest	24	1.912	Siemens	N/A	B60f, B70f, B46f
					23	1.842	GE	N/A	BONE
					14	1.596	Philips	N/A	YC
					5	1.016	Toshiba	N/A	FC85
Selim et al. [2]	Bl64	Private	Yes	Phantom	1	N/A	Siemens	Force	Br40
			No	Phantom	1	N/A	GE	Revolution EVO	LUNG
Li et al. [53]	Siemens^{T}	Private	No	Chest	80	N/A	GE	N/A	N/A
					80	N/A	Philips	N/A	N/A
					80	N/A	Siemens	N/A	N/A
		Private	No	Chest	38	N/A	GE	N/A	N/A
					28	N/A	Philips	N/A	N/A
					32	N/A	Siemens & Siemens^{T}	N/A	N/A

TABLE 7: OVERVIEW OF DATA SETS USED FOR EVALUATION IN THE INCLUDED STUDIES IN THIS SYSTEMATIC REVIEW THAT PROPOSE AN INTRA-SCANNER STANDARDIZATION METHOD.

Article	Target	Data set	Paired data	Tissue	Series (n)	Slices (n)	Scanner	Scanner type	Source data
Choe et al. [46]	B50f or B70f	Private	Yes	Chest	104	N/A	Siemens	Definition Edge	50f & B70f
Lee et al. [47]	B10f, B30f, B50f or B70f	Private	Yes	Chest	2	2.669	Siemens	Definition edge	B10f, B30f, B50f, B70f
Yang et al. [50]	J30s or J70h	Private	Yes	Head	1	44	Siemens	Definition edge	J30s, J70h
	J40s or J70h	Private	Yes	Facial bone	1	165	Siemens	Definition edge	J40s, J70h
	Hr50, Hr49 or Hr68	Private	Yes	Facial bone	1	209	Siemens	Definition edge	Hr50, Hr49, Hr68
Jin et al. [40]	B50f	Private	Yes	Chest	110	14.732	Siemens	Volume Zoom	B50f
	BONE	Private	Yes	Chest	45	664	GE	Lightspeed 16	STANDARD
	C	Private	Yes	Chest	14	2.694	Philips	Mx8000	D
	FC10	Private	Yes	Chest	9	1.703	Canon	Aquilion	FC51
Tanabe et al. [48]	FC13	Private	Yes	Chest	30	N/A	Canon	Aquilion Prime	FC51
Liang et al. [42]	Bl64	Private	Yes	Chest	N/A	3.554	Siemens	Force	Bl57
Selim et al. [41]	Bl64	Private	Yes	Chest	N/A	7620 (patches)	Siemens	Force	Bl57, Br40
Wu et al. [49]	1-mm sharp	Private	No	Chest	108	N/A	Siemens	Definition AS	5-mm B30f, 5-mm B31f, 5-mm B60f or 5-mm B80f
Lee et al. [51]	ADMIRE 32f/36f	Private	Yes	Phantom	80	N/A	Siemens	Force	FBP 32f/36f

E. Performance evaluation results

Table 8: Quantitative comparison of intra-scanner and cross-scanner standardization employed to images from a Lungman chest phantom embedded with three tumours using the number of reproducible features and the relative absolute error. The proposed method, STAN-CT [2], includes both intra- and cross-scanner standardization in a unified manner. Only the intra-scanner standardization is compared with GANai [42] and STAN-CT [41]. For UDA-CT$_{BASIC}$, the data augmentation component of the original is removed.

	Tumor 1		Tumor 2		Tumor 3	
	# of reproducible features	RE	# of reproducible features	RE	# of reproducible features	RE
Intra-scanner standardization						
Baseline	557	1.59 ± 4.68	699	1.32 ± 7.89	651	5.12 ± 76.99
GANai	851	0.46 ± 1.29	760	0.74 ± 4.96	714	0.72 ± 4.9
STAN-CT	903	0.23 ± 0.66	896	0.45 ± 4.36	734	0.74 ± 7.96
UDA-CT$_{BASIC}$	1036	0.24 ± 0.64	902	0.18 ± 0.28	651	0.24 ± 1.61
UDA-CT	1174	0.20 ± 0.56	1162	0.08 ± 0.45	714	0.21 ± 2.54

B

Reconstruction Kernel Specifications

Table B.1: Descriptions of the available GE reconstruction kernels. The algorithms going from top to bottom increase spatial resolution and decrease low contrast detectability. Extracted from user manual of GE Healthcare Revolution CT scanner.

Kernel	**Description**
Soft	for tissues with similar densities, but not useful for un-enhanced scans
Stnd	for routine exams, e.g., chest, abdomens, and pelvis scans
Lung	for interstitial lung pathology
Chest	for mediastinum and lung detail studies.
Detail	for post myelograms, where hybrid tissue detail and bone edges are important.
Bone	for High resolution exams and sharp bone detail.
Bone Plus	for sub mm detailed head work.
Edge	for small bone work in the head, as well as high resolution scans.
Ultra	for inner ear scan.

Table B.2: Descriptions of the available Philips reconstruction kernels. Extracted from user manual of Philips Brilliance CT scanner.

Kernel	Description	Resolution availability
A	Very smoothed, can be used to significantly decrease noise. Recommended for use when the patient is very large and the dose inadequate for the patients size	Standard, High, Ultra-high
B	Smoothed, but sharper and noisier than A. Recommended for CTA (for example, COW), routine abdomen, and pelvis.	Standard, High, Ultra-high
C	Sharper, creates relatively low-noise images. Recommended for CTA (for example, COW), routine abdomen, and pelvis to get slightly higher sharpness than with Filter B	Standard, High, Ultra-high
D	Sharp and edge-enhancing. Creates relatively high-noise images and raises the bone density	Standard, High, Ultra-high
E	Sharper, delivers relatively correct CT values, even for small details.	Standard, High, Ultra-high
L	Sharper than E. Delivers relatively correct CT values even for small details.Recommended for reconstruction of low -noise lung image	Standard, High
YA	Sharper and noisier. Recommended for reconstruction of sinuses, facial bones, dental, etc.	Standard
YB	Sharper and noisier than YA, recommended for reconstruction of sinuses, facial bones, etc.	Standard
YC	Sharper and noisier than YB. Recommended for reconstruction of lungs, sinuses, facial bones, dental, and orthopaedics.	High
YD	Extremely sharp and noisy. Recommended for reconstruction of IAC (when the scan is HR rather than UHR) and sinuses. Also for reconstruction of lungs and orthopaedics	High
YE	Very sharp and noisy, recommended for extremities.	Ultra-high
YF	Extremely sharp, in fact the sharpest filter of the system. It is also the noisiest filter. Recommended for extra-sharp extremity images.	Ultra-high
UA	Designed for head scans only. Minimizes the beam-hardening artefacts and significantly improves the bone-soft tissue interface (in areas such as brain or orbits). Low noise, allows detection of small lesions with relatively low noise.	Standard, High
UB	Designed to detect small lesions with improved bone/ soft tissue interface (in areas such as brain or orbits). Low contrast, for moderate resolution	Standard, High
UC	Designed to detect small lesions with improved bone/ soft tissue interface (in areas such as brain or orbits). Increases noise in images.X	Standard, High

Table B.3: Descriptions of the available Siemens reconstruction kernels. Extracted from a Siemens Somatom Sensation manual.

Kernel series	Description
B10s/B10f	Very smooth
B20s/B20f	Smooth
B30s/B30f	Medium smooth
B31s/B31f	Medium smooth +
B35s/B35f	HeartView medium
B36f	HeartView medium
B40s/B40f	Medium
B41s/B41f	Medium +
B45s/B45f	Medium
B46f	HeartView sharp
B50s/B50f	Medium sharp
B60s/B60f	Sharp
B70s/B70f	Very sharp
B80s/B80f	Ultra sharp
H10s/H10f	Very smooth
H20s/H20f	Smooth
H30s/H30f	Medium smooth
H31s/H31f	Medium smooth +
H32s/H32f	Medium smooth FR+
H40s/H40f	Medium
H41s/H41f	Medium +
H42f	Medium FR+
H42s	Medium FR
H45s/H45f	Medium
H50s/H50f	sharp
H60s/H60f	medium
H70h	Very sharp
H80	Very sharp
C20s/C20f	Smooth
C30s/C30f	Medium smooth
C60s	Sharp

Table B.4: Descriptions of the relevant Canon/Toshiba reconstruction kernels. Extracted from a Toshiba Aquilion16 manual.

Kernel series	Description
From FC01	For the abdomen, with beam hardening correction (BHC) processing
From FC10	For the abdomen
From FC20	For the head, with beam hardening correction (BHC) processing
From FC30	For the inner ear and bones
From FC40	For the head
From FC50	For the lung field
From FC60	Xe-study
From FC70	For system maintenance
From FC80	For high resolution 1, for the inner ear and bones
From FC82	For high resolution 2, for the lung field (High resolution CT)
From FC90	For high resolution 3

C

Data Set Specifications

Table C.1: Specifications of the selected data acquired from CT scanners from GE Medical Systems manufacturer. The different models of the CT scanners are stated, and the reconstruction kernels are listed, in smooth to sharp order, according to the manufacturer. In addition, the tube peak voltage, slice thickness, tube current are given, which are acquisition parameters.

Manufacturer	Reconstruction kernel	CT scanner model	Tube peak voltage [kVp]	Slice thickness [mm]	Tube current [mA]	Patients [n]	Slices [n]
GE MEDICAL SYSTEMS	SOFT	LightSpeed16	120	2.5	60	1	116
	STANDARD	HiSpeed QX/i	120	2.5	80	1	122
		LightSpeed Plus	120	2.5	120	1	225
			120	2.5	160	1	108
		LightSpeed QX/i	120	2.5	60	1	212
			120	2.5	80	3	444
			140	2.5	140	1	133
		LightSpeed Ultra	120	1.25	120	1	291
		LightSpeed16	120	2.5	45	1	171
	BONE	HiSpeed QX/i	120	2.5	80	2	231
		LightSpeed Pro 16	120	2.5	60	1	106
		LightSpeed QX/i	120	2.5	60	1	130
			120	2.5	70	1	120
			120	2.5	90	1	136
			140	2.5	40	1	139
		LightSpeed Ultra	120	1.25	60	1	32
		LightSpeed16	120	2.5	60	1	100
			120	2.5	80	1	132
	LUNG	LightSpeed Plus	120	2.5	80	2	508
			120	2.5	100	1	246
			120	2.5	160	5	630
		LightSpeed16	120	2.5	80	1	123
			120	2.5	140	1	164
					Total	31	4619

Table C.2: Specifications of the selected data acquired from CT scanners from Philips manufacturer. The different models of the CT scanners are stated, and the reconstruction kernels are listed, in smooth to sharp order, according to the manufacturer. In addition, the tube peak voltage, slice thickness, tube current are given, which are acquisition parameters.

Manufacturer	Reconstruction kernel	CT scanner model	Tube peak voltage [kVp]	Slice thickness [mm]	Tube current [mA]	Patients [n]	Slices [n]
Philips	A	Mx8000	120	3.2	67	2	179
			120	3.2	8.3	1	92
			120	3.2	100	5	486
			120	3.2	133	1	75
			120	3.2	417	1	88
	B	Mx8000	120	3.2	67	3	253
			120	3.2	92	1	66
			120	3.2	93	2	329
			120	3.2	100	1	99
			120	3.2	150	2	325
		Mx8000 IDT 16	120	2.0	180	1	183
	C	Mx8000	120	3.2	60	1	149
			120	3.2	67	1	94
			120	3.2	93	5	797
			120	1.3	120	2	473
			120	3.2	140	1	143
	D	Mx8000	120	3.2	93	6	993
			120	3.2	100	1	160
			120	3.2	150	1	161
			120	3.2	187	1	156
		Brilliance 16P	120	2.0	240	1	301
	EC	Mx8000	120	3.2	93	1	175
			120	3.2	140	1	130
			120	3.2	187	1	155
					Total	43	6062

Table C.3: Specifications of the selected data acquired from CT scanners from Siemens manufacturer. The different models of the CT scanners are stated, and the reconstruction kernels are listed, in smooth to sharp order, according to the manufacturer. In addition, the tube peak voltage, slice thickness, tube current are given, which are acquisition parameters.

Manufacturer	Reconstruction kernel	CT scanner model	Tube peak voltage [kVp]	Slice thickness [mm]	Tube current [mA]	Patients [n]	Slices [n]
SIEMENS	B20f	Sensation 16	120	2.0	90	3	477
			120	2.0	120	2	298
		Volume Zoom	120	2.0	120	5	804
	B20s	Emotion 6	130	1.25	225	1	332
	B30f	Sensation 16	120	1.0	150	1	358
			120	2.0	75	1	183
			120	2.0	80	1	169
			120	2.0	90	1	151
		Volume Zoom	120	2.0	120	2	330
			120	2.0	150	2	247
			120	2.0	188	1	172
			120	2.0	210	1	159
	B30s	Emotion 6	130	2.5	38	1	10
			130	2.5	51	1	158
			130	4.0	173	1	320
			130	5.0	66	1	66
		Emotion 16	110	2.0	63	1	176
			130	2.0	100	4	619
		Volume Zoom	120	2.0	80	1	152
	B31f	Sensation 16	120	3.0	40	1	117
			120	3.0	212	1	122
			120	3.0	256	1	102
			120	3.0	292	1	154
			120	3.0	392	1	126
			120	3.0	494	1	119

Continued on next page

Table C.3: Specifications of the selected data acquired from CT scanners from Siemens manufacturer. The different models of the CT scanners are stated, and the reconstruction kernels are listed, in smooth to sharp order, according to the manufacturer. In addition, the tube peak voltage, slice thickness, tube current are given, which are acquisition parameters. (Continued)

Manufacturer	Reconstruction kernel	CT scanner model	Tube peak voltage [kVp]	Slice thickness [mm]	Tube current [mA]	Patients [n]	Slices [n]
			120	3.0	500	1	113
		Volume Zoom	120	2.0	120	3	482
	B31s	Emotion 6	130	2.5	38	1	155
			130	3.0	40	1	124
			130	3.0	64	1	112
			130	2.5	75	1	141
			130	3.0	226	1	151
			130	3.0	256	1	120
			130	3.0	275	4	476
	B35f	Volume Zoom	120	2.0	120	2	347
	B40f	Sensation 4	120	5.0	80	1	78
		Volume Zoom	120	5.0	210	1	65
			120	5.0	295	1	63
			140	5.0	330	1	58
	B40s	Emotion Duo	130	3.0	133	1	279
	B41s	Sensation 16	120	3.0	270	1	138
	B45f	Definition	120	1.0	332	1	251
		Sensation 16	120	1.0	120	1	401
			120	1.0	345	1	331
			120	1.0	412	1	299
			120	1.0	421	1	328
			120	1.0	432	1	349
			120	1.0	440	1	339
		Sensation 64	120	1.0	513	1	266
			120	3.0	435	1	104

Continued on next page

Table C.3: Specifications of the selected data acquired from CT scanners from Siemens manufacturer. The different models of the CT scanners are stated, and the reconstruction kernels are listed, in smooth to sharp order, according to the manufacturer. In addition, the tube peak voltage, slice thickness, tube current are given, which are acquisition parameters. (Continued)

Manufacturer	Reconstruction kernel	CT scanner model	Tube peak voltage [kVp]	Slice thickness [mm]	Tube current [mA]	Patients [n]	Slices [n]
	B45f	Volume Zoom	120	2.0	140	1	166
	B50f	Sensation 16	120	2.0	90	2	325
			120	5.0	90	1	68
		Sensation 16	120	5.0	108	1	58
		Volume Zoom	120	2.0	120	4	623
			120	2.0	150	1	168
			140	2.0	160	1	203
	B50s	Emotion 16	130	2.0	100	3	520
		Volume Zoom	120	2.0	80	1	152
	B60f	Sensation 16	120	2.0	90	6	1029
		Volume Zoom	120	2.0	120	1	141
			140	2.0	140	2	321
			120	2.0	160	1	161
	B60s	Emotion 16	110	2.0	63	2	302
			130	2,0	90	1	151
			130	2.0	100	6	947
		Sensation 16	120	3.0	270	1	92
	B70f	Sensation 16	120	2.0	105	4	803
			120	2.0	208	1	346
			120	2.0	360	1	344
			120	2.0	381	1	325
		Volume Zoom	120	2.0	150	3	510
	B80f	Sensation 16	120	1.0	150	7	2423
			120	1.0	250	3	949
					Total	123	22618

Table C.4: Specifications of the selected data acquired from CT scanners from Toshiba Healthcare manufacturer. The different models of the CT scanners are stated, and the reconstruction kernels are listed, in smooth to sharp order, according to the manufacturer. In addition, the tube peak voltage, slice thickness, tube current are given, which are acquisition parameters.

Manufacturer	Reconstruction kernel	CT scanner model	Tube peak voltage [kVp]	Slice thickness [mm]	Tube current [mA]	Patients [n]	Slices [n]
TOSHIBA	FC01	Aquilion	120	2.0	80	1	140
			120	2.0	160	5	797
			135	3.0	260	4	486
	FC02	Aquilion	120	3.0	160	9	844
	FC03	Aquilion	135	2.0	260	3	458
	FC10	Aquilion	120	2.0	80	3	519
			120	2.0	100	1	159
			120	2.0	120	5	962
			120	2.0	150	1	150
	FC30	Aquilion	120	2.0	160	6	920
			120	3.0	169	4	428
	FC50	Aquilion	120	2.0	80	4	630
			120	2.0	100	2	369
			120	2.0	120	1	193
			120	2.0	140	1	161
			120	2.0	160	1	170
	FC51	Aquilion	120	2.0	80	5	832
			120	2.0	120	2	334
			120	2.0	150	2	359
			120	2.0	160	1	180
	FC53	Aquilion	120	1.0	80	1	263
			120	2.0	80	2	322
	FC82	Aquilion	120	2.0	80	7	1039
		Aquilion	120	2.0	100	1	164
			120	2.0	160	2	310
					Total	87	11214

D

Noise magnitude results

Table D.1: Median noise magnitude measured as standard deviation value per patient (1-10), order per kernel from low to high, left to right. The median standard deviation per kernel has also been computed together with the median absolute deviation (MAD)

Kernel	1	2	3	4	5	6	7	8	9	10	Median	MAD
A	12.80	13.73	14.40	14.46	15.19	15.64	15.73	16.05	18.27	18.71	15.41	0.98
B	10.86	12.59	13.40	13.47	16.32	19.43	19.99	20.38	21.39	22.56	17.88	3.96
C	15.87	17.25	19.91	22.75	23.28	24.22	25.60	26.16	26.43	29.10	23.75	2.55
D	20.75	28.33	45.42	45.63	58.23	64.42	68.86	69.13	83.77	86.11	61.33	15.80
EC	19.85	24.26	24.93	n/a	n/a	n/a	n/a	n/a	n/a	n/a	24.26	0.67
SOFT	21.94	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	21.94	n/a
STD	11.57	14.86	18.21	18.63	18.76	21.11	21.60	26.00	26.90	28.64	19.93	3.40
BONE	43.54	44.22	45.34	54.04	55.28	57.85	64.55	73.91	73.95	78.56	56.57	11.78
LUNG	38.78	38.78	51.52	53.04	65.77	70.15	80.72	82.36	113.52	140.97	67.96	15.68
B20s	12.91	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	12.91	n/a
B20f	13.96	14.22	14.37	15.39	15.77	17.59	18.20	24.88	24.90	25.42	16.68	2.38
B30s	9.88	11.20	14.00	14.31	16.36	16.75	17.64	17.75	19.95	20.45	16.55	2.40
B30f	12.37	14.94	15.01	18.10	19.20	25.14	25.60	27.36	28.89	29.17	22.17	5.95
B31s	8.45	8.83	8.96	9.06	9.09	9.26	9.42	9.46	11.33	15.11	9.17	0.26
B31f	8.08	9.58	9.66	10.53	10.61	12.72	14.02	16.36	18.06	19.64	11.66	2.22
B35f	21.66	31.12	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	26.39	4.73
B40f	10.40	11.46	12.90	17.78	n/a	n/a	n/a	n/a	n/a	n/a	12.18	1.25
B40s	17.38	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	17.38	n/a
B41s	8.89	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	8.89	n/a
B45f	14.89	20.30	21.65	25.30	27.39	27.42	27.88	29.89	44.04	57.63	27.41	4.12
B50f	28.42	41.67	41.83	47.48	50.92	51.75	53.83	54.51	54.94	59.39	51.34	3.73
B50s	27.94	31.30	37.19	65.18	n/a	n/a	n/a	n/a	n/a	n/a	34.24	4.63
B60f	50.18	55.47	66.12	70.48	73.60	80.50	80.52	89.61	96.56	98.70	77.05	11.74
B60s	34.93	35.13	35.66	42.95	44.33	44.36	46.66	48.02	49.44	49.74	44.34	4.39
B70f	35.95	41.11	65.33	71.55	73.39	73.88	75.41	85.46	94.17	n/a	73.39	8.06
B80f	76.63	79.45	84.22	87.15	87.86	89.20	92.88	94.99	125.34	125.90	88.53	5.41
FC01	6.01	7.27	12.65	12.98	14.61	18.60	20.71	25.92	27.38	34.13	16.61	6.71
FC02	9.41	10.00	11.38	11.89	12.78	14.63	17.02	17.36	19.91	n/a	12.78	2.78
FC03	6.47	7.53	13.34	n/a	n/a	n/a	n/a	n/a	n/a	n/a	7.53	1.07

Continued on next page

Table D.1: Median noise magnitude measured as standard deviation value per patient (1-10), order per kernel from low to high, left to right. The median standard deviation per kernel has also been computed together with the median absolute deviation (MAD) (Continued)

Kernel	1	2	3	4	5	6	7	8	9	10	Median	MAD
FC10	17.50	18.67	20.26	21.50	22.21	22.22	23.59	24.37	28.77	29.00	22.22	2.05
FC50	28.61	36.73	37.06	46.15	46.28	48.82	50.89	53.71	79.70	n/a	46.28	7.44
FC51	30.85	31.83	38.21	38.78	64.82	66.17	69.70	72.99	94.50	107.87	65.50	27.00
FC53	51.07	52.91	97.44	n/a	n/a	n/a	n/a	n/a	n/a	n/a	52.91	1.83
FC30	57.73	72.70	74.98	82.82	86.23	90.02	90.58	111.40	118.97	139.18	88.13	14.29
FC82	42.58	48.79	53.14	59.15	66.69	72.08	72.84	80.69	90.26	94.84	69.39	13.78

E

Noise texture results

Table E.1: Median noise texture measured as central frequency extracted from the 1D noise power spectrum per patient (1-10), order per kernel from low to high, left to right. The median standard deviation per kernel has also been computed together with the median absolute deviation (MAD)

Kernel	1	2	3	4	5	6	7	8	9	10	Median	MAD
A	0.1053	0.1167	0.1175	0.1190	0.1241	0.1241	0.1242	0.1282	0.1285	0.1287	0.1241	4.525E-3
B	0.0918	0.1155	0.1156	0.1164	0.1190	0.1191	0.1273	0.1440	0.1484	0.1646	0.1190	5.888E-3
C	0.1461	0.1616	0.1656	0.1676	0.1745	0.1799	0.1807	0.1923	0.1980	0.2014	0.1772	13.32E-3
D	0.1303	0.1375	0.1410	0.1510	0.1516	0.1559	0.1587	0.1702	0.2073	0.2366	0.1538	14.53E-3
EC	0.0973	0.1011	0.1278	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.1011	3.788E-3
SOFT	0.1266	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.1266	n/a
STD	0.1023	0.1184	0.1191	0.1219	0.1257	0.1359	0.1395	0.1418	0.1448	0.1603	19.93	11.35E-3
BONE	0.1237	0.1404	0.1422	0.1445	0.1614	0.1840	0.2282	0.2762	0.2894	0.3059	0.1727	40.62E-3
LUNG	0.1424	0.1466	0.1643	0.1739	0.1961	0.1972	0.2014	0.2040	0.2065	0.2463	67.96	16.33E-3
B20s	0.1334	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.1334	n/a
B20f	0.0966	0.1046	0.1060	0.1132	0.1137	0.1154	0.1195	0.1219	0.1262	0.1318	0.1145	7.927E-3
B30s	0.0756	0.1054	0.1195	0.1230	0.1244	0.1351	0.1435	0.1478	0.1520	0.1546	0.1334	15.91E-3
B30f	0.1041	0.1097	0.1235	0.1244	0.1270	0.1274	0.1305	0.1382	0.1397	0.1522	0.1272	7.339E-3
B31s	0.0813	0.0898	0.0932	0.1026	0.1030	0.1032	0.1115	0.1188	0.1233	0.1297	0.1031	11.56E-3
B31f	0.0796	0.0853	0.0906	0.0959	0.1060	0.1129	0.1177	0.1233	0.1282	0.1095	0.1095	16.29E-3
B35f	0.1210	0.1290	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.1250	3.984E-3
B40f	0.1044	0.1180	0.1330	0.1341	n/a	n/a	n/a	n/a	n/a	n/a	0.1255	8.003E-3
B40s	0.0460	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.0460	n/a
B41s	0.1238	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.1238	n/a
B45f	0.1294	0.1528	0.1625	0.1669	0.1690	0.1824	0.1862	0.1926	0.1955	0.2001	0.1757	15.07E-3
B50f	0.1427	0.1624	0.1639	0.1732	0.1970	0.2117	0.2133	0.2139	0.2353	0.2551	0.2043	40.01E-3
B50s	0.1376	0.1810	0.2176	0.2881	n/a	n/a	n/a	n/a	n/a	n/a	0.1993	31.02E-3
B60f	0.1680	0.1733	0.1920	0.1970	0.2028	0.2036	0.2110	0.2290	0.2478	0.2679	0.2032	18.53E-3
B60s	0.1631	0.1722	0.1738	0.1821	0.1965	0.1998	0.2533	0.2537	0.2563	0.2650	0.1982	30.52E-3
B70f	0.1630	0.1687	0.1754	0.1780	0.1797	0.1811	0.2221	0.2415	0.2567	n/a	0.1797	11.06E-3
B80f	0.1870	0.1873	0.2028	0.2042	0.2055	0.2060	0.2269	0.2465	0.2481	0.2484	0.2057	18.55E-3
FC01	0.0748	0.0894	0.0987	0.1016	0.1106	0.1327	0.1422	0.1452	0.1456	0.1460	0.1217	23.24E-3
FC02	0.0828	0.1102	0.1108	0.1268	0.1320	0.1322	0.1440	0.1542	0.1551	n/a	0.1320	21.16E-3

Continued on next page

Table E.1: Median noise texture measured as central frequency extracted from the 1D noise power spectrum per patient (1-10), order per kernel from low to high, left to right. The median standard deviation per kernel has also been computed together with the median absolute deviation (MAD) (Continued)

Kernel	1	2	3	4	5	6	7	8	9	10	Median	MAD
FC03	0.0656	0.0787	0.0926	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.0787	13.07E-3
FC10	0.1061	0.1216	0.1232	0.1303	0.1341	0.1357	0.1493	0.1525	0.1674	0.1857	0.1349	13.84E-3
FC50	0.1334	0.1628	0.1689	0.1750	0.1780	0.1802	0.1830	0.1833	0.27	n/a	0.1779	5.325E-3
FC51	0.1549	0.1645	0.1691	0.1696	0.1780	0.1838	0.1914	0.1951	0.2021	0.2342	0.1809	13.00E-3
FC53	0.1215	0.2756	0.2884	n/a	n/a	n/a	n/a	n/a	n/a	n/a	0.2756	12.76E-3
FC30	0.1609	0.1663	0.1740	0.1755	0.1765	0.1816	0.1833	0.1985	0.2388	0.2418	0.1791	8.935E-3
FC82	0.1568	0.1573	0.1637	0.1675	0.1698	0.1771	0.1793	0.1797	0.2345	0.2608	0.1734	8.008E-3

F

Comparison Model Categorization

Table F.1: Categorisations per kernel. The table shows the correct percentage of classifications per kernel for the *SVC_noise* and *RFC_radiomics* models. Additionally, it shows the manually determined true class label per kernel.

		SVC_noise		*RFC_radiomics*	
Kernel	**N**	**True class**	**Correct [%]**	**True class**	**Correct [%]**
A	10	smooth	100	smooth	100
B	10	smooth	100	smooth	100
C	10	smooth	100	smooth	90
D	10	sharp	80	sharp	100
EC	3	smooth	100	smooth	100
SOFT	1	smooth	100	smooth	100
STD	10	smooth	100	smooth	100
BONE	10	sharp	100	sharp	100
LUNG	10	sharp	100	sharp	100
B20s	1	smooth	100	smooth	100
B20f	10	smooth	100	smooth	100
B30s	10	smooth	100	smooth	100
B30f	10	smooth	100	smooth	100
B31f	10	smooth	100	smooth	100
B31s	10	smooth	100	smooth	100
B35f	2	smooth	100	smooth	100
B40f	4	smooth	100	smooth	100
B40s	1	smooth	100	smooth	100
B41s	1	smooth	100	smooth	100
B45f	10	smooth	80	smooth	80
B50f	10	sharp	90	sharp	90
B50s	4	sharp	50	smooth	75
B60f	10	sharp	100	sharp	100
B60s	10	sharp	100	sharp	90
B70f	9	sharp	100	sharp	89
B80f	10	sharp	100	sharp	100
FC01	10	smooth	100	smooth	100

Continued on next page

Table F.1: Categorisations per kernel. The table shows the correct percentage of classifications per kernel for the *SVC_noise* and *RFC_radiomics* models. Additionally, it shows the manually determined true class label per kernel. (Continued)

		SVC_noise		*RFC_radiomics*	
Kernel	**N**	**True class**	**Correct [%]**	**True class**	**Correct [%]**
FC02	9	smooth	100	smooth	100
FC03	3	smooth	100	smooth	100
FC10	10	smooth	100	smooth	100
FC50	10	sharp	89	sharp	89
FC51	10	sharp	100	sharp	80
FC53	3	sharp	100	sharp	100
FC30	10	sharp	100	sharp	100
FC82	10	sharp	90	sharp	100