FABRICATION OF ANGIOGRAPHY QUALITY CONTROL PHANTOM FOR IMAGE QUALITY EVALUATION USING MACHINE LEARNING

MUHAMMAD HAZIQ BIN ABD AZIZ

SCHOOL OF HEALTH SCIENCES

UNIVERSITI SAINS MALAYSIA

2025

FABRICATION OF ANGIOGRAPHY QUALITY CONTROL PHANTOM FOR IMAGE QUALITY EVALUATION USING MACHINE LEARNING

by

MUHAMMAD HAZIQ BIN ABD AZIZ

Dissertation submitted in partial fulfilment of the requirement for the degree of Bachelor of Medical Radiation

JUNE 2025

CERTIFICATE

This is to certify that the dissertation entitled "**FABRICATION OF ANGIOGRAPHY QUALITY CONTROL PHANTOM FOR IMAGE QUALITY EVALUATION USING MACHINE LEARNING**" is the bona fide record of research work done by MUHAMMAD HAZIQ BIN ABD AZIZ during the period from October 2024 to June 2025 under my supervision. I have read this dissertation and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation to be submitted in partial fulfilment for the degree of Bachelor of Health Science (Honours) (Medical Radiation).

Main Supervisor:

……..

Siti Aishah Abd Aziz
Senior Lecturer
School of Health Sciences
Universiti Sains Malaysia
Health Campus
16150 Kubang Kerian
Kelantan, Malaysia

Field Supervisor:

……..

En.Nik Kamarullah Ya Ali
Medical Physicist
Specialist Hospital
Universiti Sains Malaysia
16150 Kubang Kerian
Kelantan, Malaysia

Date: June 2025

DECLARATION

I, MUHAMMAD HAZIQ BIN ABD AZIZ hereby declare that the dissertation entitled "**FABRICATION OF ANGIOGRAPHY QUALITY CONTROL PHANTOM FOR IMAGE QUALITY EVALUATION USING MACHINE LEARNING**" is the result of my own investigations, except where otherwise stated and duly acknowledged. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at Universiti Sains Malaysia or other institutions. I grant Universiti Sains Malaysia the right to use the dissertation for teaching, research and promotional purposes.

…….

MUHAMMAD HAZIQ BIN ABD AZIZ

Date: June 2025

ACKNOWLEDGEMENT

In the name of Allah S.W.T., the Most Gracious and the Most Merciful, I begin with profound gratitude. This challenging yet rewarding journey has shaped me in ways I never imagined, and I will forever cherish the memories and the beautiful souls who walked alongside me. To my beloved parents, **Abd Aziz bin Kassim** and **Hamidah binti Mohd Jadi**, you have been my pillars of strength and my guiding light. Your endless sacrifices, unwavering faith in me, and the values you instilled have made this dream possible. No words can capture my gratitude for your love, financial support, and the countless nights you worried about my progress. May Allah S.W.T. bless you abundantly for everything you have done.

My deepest gratitude to my supervisor, **Madam Siti Aishah Abd Aziz**, whose wisdom, patience, and dedication have been extraordinary. Over two semesters, you have been more than an academic guidance. Your constructive criticism, feedback, passion for excellence, and sincere care for your students' success will forever be etched in my heart. Warmest thanks to my field supervisor, **En. Nik Kamarullah Ya Ali**, whose practical guidance and field expertise were invaluable during the data collection phase. His support ensured the smooth execution of fieldwork activities.

My sincere appreciation to **Dr. Muhammad Akmal Remli** for his exceptional computational guidance throughout this research. His expertise and insights were instrumental in navigating the technical challenges. I am equally grateful to his dedicated PhD students, **Ainin Sofia Jusoh** and **Meor Muhammad Muaz** from the Institute for Artificial Intelligence and Big Data (AiBIG), whose assistance and knowledge sharing made complex computational tasks manageable.

To my cherished friends, **Alya**, **Qila**, **Shameera, Zahra, and Shahril,** you brought joy and light to my life during these challenging times. Our friendship has been a precious gift, offering comfort during stressful moments and celebrating with me during joyful ones. Your emotional support and the laughter we shared touched my soul deeply and reminded me that I was never alone in this journey. This whole-hearted achievement belongs to all of us. Each person mentioned has contributed not just to the completion of this research, but to the person I have become. I pray that Allah S.W.T. blesses each of you with happiness and success in all your future endeavours. May He reward your kindness in this life and the hereafter.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Abbreviation	Definition
2D	Two-Dimensional
3D	Three-Dimensional
3DRA	Three-Dimensional Rotational Angiography
ACR	American College of Radiology
AHE	Adaptive Histogram Equalisation
AI	Artificial Intelligence
ASIC	Application-Specific Integrated Circuit
AUC	Area Under the Curve
CATIA	Computer Aided Three-Dimensional Interactive Application
CIRS	Computerized Imaging Reference Systems
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CT	Computed Tomography
CTA	Computed Tomography Angiography
CUDA	Compute Unified Device Architecture
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DLR	Deep Learning Reconstruction
DSA	Digital Subtraction Angiography
DT	Decision Tree
FBP	Filtered Back Projection

Abbreviation	Definition
FPGA	Field-Programmable Gate Array
FPR	False Positive Rate
GPU	Graphics Processing Unit
HDL	Hardware Description Language
HPUSM	Hospital Pakar Universiti Sains Malaysia
IR	Iterative Reconstruction
KNN	k-Nearest Neighbors
LAO	Left Anterior Oblique
LBP	Local Binary Pattern
MBIR	Model-Based Iterative Reconstruction
ML	Machine Learning
MQSA	Mammography Quality Standards Act
MRI	Magnetic Resonance Imaging
NPS	Noise Power Spectrum
OBJ	Object File Format
PACS	Picture Archiving and Communication System
PA	Posterior-Anterior
PCA	Principal Component Analysis
PLA	Polylactic Acid
PMMA	Polymethyl Methacrylate
PVA	Polyvinyl Alcohol
QA	Quality Assurance
RF	Random Forest
RGB	Red, Green, Blue

Abbreviation	Definition
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SID	Source-to-Image Distance
SMO	Sequential Minimal Optimisation
SMOTE	Synthetic Minority Over-sampling Technique
STL	Standard Tessellation Language
SVM	Support Vector Machine
TPR	True Positive Rate
UNet	U-shaped Convolutional Neural Network
USA	United States of America

FABRIKASI FANTOM KAWALAN KUALITI ANGIOGRAFI UNTUK PENILAIAN KUALITI IMEJ MENGGUNAKAN PEMBELAJARAN MESIN

ABSTRAK

Kawalan kualiti (QC) angiografi terjejas oleh penilaian subjektif dan kekurangan fantom khusus. Kajian ini membangunkan fantom angiografi dalaman yang berpatutan dan menilai kualiti imej menggunakan pembelajaran mesin (ML). **Tujuan:** 1) Mereka bentuk dan cipta fantom dalaman untuk kontras tinggi dan resolusi ruang; 2) Menilai prestasi dan pengesahan model ML; 3) Menyesahkan ML terbaik untuk penilaian kualiti imej fantom. **Kaedah:** Fantom dalaman dicetak 3D (LW-PLA-HT) dengan manik tungsten karbida (kontras tinggi) dan pasangan garisan Huttner 18 (resolusi ruang). 14 imej angiografi diperoleh dari HPUSM dan dianalisis dalam MATLAB R2024a. Analisis imej melibatkan pra-pemprosesan, segmentasi, pengekstrakan ciri, dan augmentasi. Pengelas SVM, KNN, dan RF dinilai menggunakan ketepatan, kepersisan, kepekaan, skor-F1, dan AUC, dengan validasi silang 10-lipatan dan pembahagian 80/20. **Dapatan Kajian:** Penilaian manusia menunjukkan variasi. Antara SVM, KNN, dan RF, Random Forest (RF) menunjukkan prestasi keseluruhan terbaik. Untuk klasifikasi kontras tinggi, RF mencapai ketepatan 100% (skor F1 1.0000), diikuti KNN (76.11% ketepatan, skor F1 0.7503), dan SVM (61.95% ketepatan, skor F1 0.6095). Klasifikasi resolusi ruang lebih mencabar; RF mendahului (90.32% ketepatan, skor F1 0.9050), diikuti KNN (64.52% ketepatan, skor F1 0.6650), dan SVM (32.26% ketepatan, skor F1 0.3180). **Kesimpulan:** Random Forest menunjukkan prestasi terbaik dalam penyelidikan ini, yang menyerlahkan daya maju penghasilan fantom angiografi yang kos-efektif dan penggunaan ML untuk penilaian kualiti imej.

FABRICATION OF ANGIOGRAPHY QUALITY CONTROL PHANTOM FOR IMAGE QUALITY EVALUATION USING MACHINE LEARNING

ABSTRACT

Angiography's QC suffers from subjective evaluations and a lack of specialised phantoms. This study addresses this by developing an affordable, in-house angiography phantom and evaluating the image quality using a machine learning (ML) approach. **Purpose:** 1) Design and fabricate an in-house phantom for high contrast and spatial resolution; 2) Assess ML model performance and validation; 3) Validate the best ML for evaluation of phantom image quality. **Method:** An in-house phantom was 3D-printed using LW-PLA-HT, incorporating tungsten carbide beads for high contrast and a Huttner Type 18-line pair for spatial resolution. 14 angiographic images were acquired from HPUSM and analysed in MATLAB R2024a. Image analysis involved pre-processing, segmentation, feature extraction and augmentation were applied. Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF) classifiers were evaluated using accuracy, precision, recall, F1-score, and AUC, with 10-fold cross-validation and an 80/20 training/testing. **Results:** Human evaluations showed variability. Among SVM, KNN, and RF, Random Forest demonstrated the best overall performance. For high-contrast image classification, RF achieved exceptional results (100% accuracy, 1.0000 F1 score), followed by KNN (76.11% accuracy, 0.7503 F1 score), and SVM (61.95% accuracy, 0.6095 F1 score). Spatial resolution classification was more challenging, with RF again leading (90.32% accuracy, 0.9050 F1 score), followed by KNN (64.52% accuracy, 0.6650 F1 score), and SVM (32.26% accuracy, 0.3180 F1 score). **Conclusion:** Random Forest demonstrated the best performance in this research, which highlights the viability of fabricating a cost-effective angiography phantom and utilising ML for image quality assessment.

CHAPTER 1

INTRODUCTION

1.1. Background of Study

Angiography is used for the surgical or endovascular treatment of intracranial aneurysms (Benomar et al., 2021). It can visualise the anatomy and vascular structures system by detecting contrast medium injected into a blood vessel. This contrast highlights the inner vessel walls and flows through the lumen, which is then captured in a series of X-ray images. Originally developed as a diagnostic tool, angiography has transformed significantly over the years, evolving into a key foundation for interventional therapies. Initially a static two-dimensional (2D) method of recording vascular structures on screen film, angiography has advanced to real-time 2D visualisation on monitors and even three-dimensional (3D) reconstructions using computed tomography (CT) and magnetic resonance imaging (MRI). However, conventional angiography remained the gold standard, although it is invasive for diagnosing many intravascular conditions. Technological advancements have broadened the scope of angiography to include non-invasive methods, such as computed tomography angiography (CTA) and magnetic resonance angiography (MRA) (Omeh & Shlofmitz, 2024).

Image quality of angiography is influenced by several factors, including temporal resolution, spatial resolution, contrast resolution and radiation dose. These factors may lead to inconsistent image quality, potentially affecting diagnostic accuracy (Ghekiere et al., 2017). To address this, medical imaging facilities often use “phantoms” as a

physical model that simulates human anatomy and tissue characteristics (Christie et al., 2023).

Phantoms serve three primary purposes in medical imaging and treatment. First and foremost, the purpose is quality assurance and calibration, which may be quantitative or qualitative, focusing on controlling and evaluating imaging or treatment systems. This includes phantom used to assess image quality metrics like spatial resolution and dosimetry, crucial for quality control in radiotherapy. Beyond quality assurance, phantoms support research and development by facilitating new instrument designs, procedural setups, or intervention methods, allowing for simulations of treatments or surgeries without the use of human or animal subjects. Finally, phantoms are essential for education and training, providing hands-on practice for procedures under image guidance, such as angiography, with the aid of personnel and radiologists (Wegner et al., 2023).

Despite their benefits, commercially available phantoms can be expensive, limiting access for many hospitals and research institutions, especially in resource-constrained settings (Groenewald & Groenewald, 2016). Additionally, these commercially manufactured phantoms may not accurately replicate specific clinical scenarios or unique patient anatomies. Consequently, there is a strong motivation to develop an in-house angiography phantom that is affordable, customisable, and can provide simulations of vascular structures (Soloukey et al., 2024).

Over the years, image quality assessment has relied heavily on subjective and non-repeatability evaluation by personnel and radiologists, who visually inspect the images for clarity, contrast, and resolution. While these expert evaluations are the gold standard, they can introduce variability, as different evaluators may interpret image quality

differently (Ho et al., 2022). Machine learning offers a great solution by providing a standardised, objective method to assess image quality. Hence, training a machine learning model on testing images labelled for quality can automate the evaluation process, making it more consistent and sensitive to subtle quality changes or issues that human evaluators might overlook (Shurrab & Duwairi, 2022).

1.2. Problem Statement

Angiography plays a critical role in diagnosing and treating vascular diseases, requiring high-quality imaging to accurately visualise blood vessels and vascular structures (Lubis et al., 2021). However, current quality control (QC) practices for angiography are inadequate due to the lack of specialised phantoms tailored to this imaging modality. Existing QC techniques primarily rely on fluoroscopy-based phantoms, which do not accurately reflect clinical conditions encountered during angiographic procedures, leading to suboptimal image quality assessments and challenges in optimising radiation dose (Pancholy et al., 2022). Furthermore, image quality evaluation in clinical settings remains largely subjective, relying on radiologist interpretation, which introduces variability and potential bias that can affect diagnostic accuracy and consistency (Ho et al., 2022; Oh et al., 2022). Despite progress in phantom development for other imaging modalities such as mammography, angiography still lacks dedicated, modality-specific QC tools that reflect its unique technical and clinical demands.

1.3. Objective

1.3.1. General Objective

To design, fabricate and evaluate an in-house angiography phantom combined with a machine learning-based image quality assessment technique to enhance the reliability and accuracy of angiographic imaging for diagnostic purposes.

1.3.2. Specific Objectives

1. To design and fabricate an in-house phantom for high contrast and spatial resolution image quality assessment.
2. To assess machine learning model performance and validation in evaluating the image quality of the fabricated angiographic phantom.
3. To validate the best machine learning for the evaluation of phantom image quality in classifying high contrast and spatial resolution of the fabricated angiographic phantom.

1.4. Hypothesis

1.4.1. Null Hypothesis (H_o)

The fabricated in-house angiography phantom combined with a machine learning-based image quality assessment does not significantly improve the accuracy or consistency of angiographic image quality evaluation compared with each model's algorithms.

1.4.2. Alternative Hypothesis (H_A)

The fabricated in-house angiography phantom combined with a machine learning-based image quality assessment significantly improves the accuracy and consistency of angiographic image quality evaluation compared with each model's algorithms.

1.5. Significant of Study

This study addresses critical limitations in current angiography quality control (QC) practices, which rely on non-specialised phantoms and subjective evaluations that introduce variability and reduce diagnostic accuracy (Pancholy et al., 2022; Ho et al., 2022; Oh et al., 2022). By fabricating a dedicated angiography phantom using affordable materials like PLA (Groenewald, 2017; Li, 2020) and incorporating machine learning for image quality assessment, the study aims to provide a cost-effective method in achieving the objectives in QC solution.

Similar success in other modalities, such as mammography, demonstrates the potential of 3D-printed phantoms to improve standardisation and meet clinical standards (Celina et al., 2023). Machine learning offers accurate, reproducible evaluations, reducing reliance on subjective interpretation and improving workflow efficiency (Ho et al., 2022; Oh et al., 2022). The proposed approach is expected to enhance diagnostic consistency, optimise radiation dose management, and support more reliable QC processes in angiography.

1.6. Conceptual Framework

The conceptual framework of this study, illustrated in Figure 1.1, centered on the image quality evaluation of angiography imaging using a machine learning-based approach, supported by an in-house fabricated phantom. These custom-designed phantom images will undergo a structured pipeline comprising image pre-processing, region-based detection, and feature extraction stages. During pre-processing, operations such as cropping, contrast enhancement, and noise reduction will be applied. Feature extraction will include both geometric and region-based features, such as dot count, area, perimeter, and eccentricity. Extracted features will be converted into numerical representations (pixel values or structured feature vectors), which are then fed into machine learning classifiers, supporting automated and objective quality control in angiographic imaging.

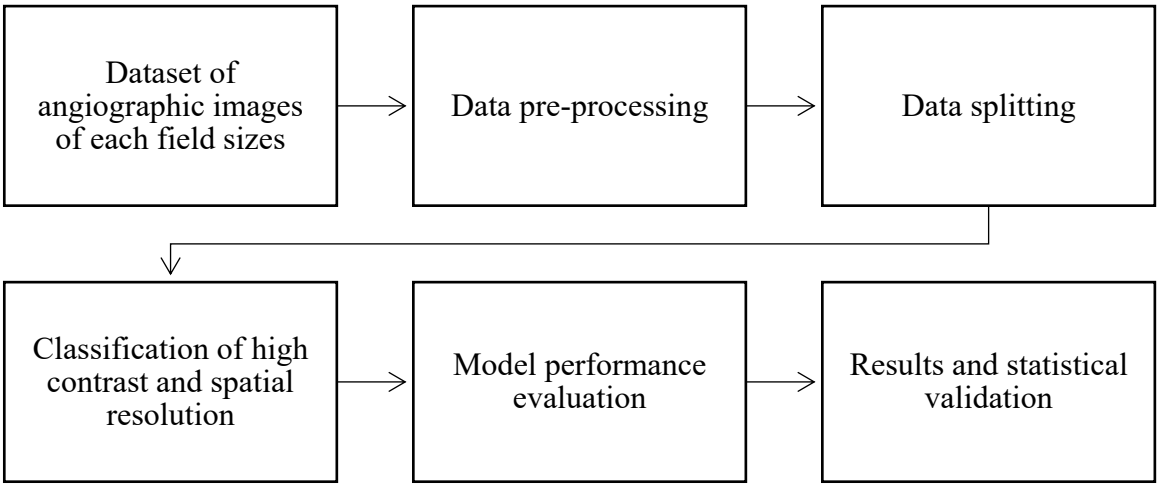

Figure 1.1: Conceptual framework

CHAPTER 2

LITERATURE REVIEW

2.1. Image Quality Control in Angiography

Image quality control in angiography ensures accurate and reliable diagnostic imaging results. Various imaging modalities, such as computed tomography angiography (CTA) and three-dimensional rotational angiography (3DRA), require quality control measures to maintain image quality and diagnostic accuracy. The authors stated that fluoroscopy and digital subtraction angiography (DSA) were the benchmarks in intervention radiology for diagnosis and treatment. However, computed tomography angiography (CTA) is widely used for three-dimensional vascular imaging and plays a significant part in the delineation of vascular diseases (Lubis et al., 2021).

This highlights the importance of utilising phantoms in evaluating image quality in specific medical imaging techniques. Phantom studies are significant in evaluating image quality in medical physics, particularly in techniques such as 3DRA and CTA. By assessing the effects of various parameters on image quality and implementing corrections for artefacts, researchers aim to enhance the accuracy and reliability of diagnostic imaging modalities (Svenson & Irvine, 2024).

2.2. Limitations of Existing Quality Control Phantoms

Due to the limited ability to archive image loops on photographic film and the initial digital platform, fluoroscopy has been used to guide real-time image setup and equipment navigator, while cineangiography has captured video angiograms. The image resolution for both modalities is essential for assessing the specifics of the vasculature and treatments (Pancholy et al., 2022).

A study conducted by Pancholy et al. (2022) involved a phantom-based total of 40 experiments, half of which were done using fluoroscopy and the other half using cine angiography. They used a comparative evaluation of high-contrast and low-contrast resolution of images. 5 acquisitions were carried out, for posterior-anterior (PA) or left anterior oblique (LAO). The study found out there is no difference in high or low contrast resolution between PA and LAO projections using fluoroscopy. Cineangiography showed that PA projections had higher contrast resolution than LAO projections, but there was no significant difference in low contrast resolution. There was no significant difference in high-contrast or low-contrast resolution between low and high table positions with fluoroscopy or cineangiography (Pancholy et al., 2022).

In short, the use of fluoroscopy-based QC phantoms in angiography may not accurately reflect the clinical scenario encountered during angiographic procedures, leading to potential drawbacks in image quality assessment and radiation dose optimisation. To achieve optimal results and enhance the overall performance of the imaging equipment, specialised phantoms designed for angiography must be used when performing quality control tests.

2.3. Potential for an In-House Angiography Phantom

The Mammography Quality Standards Act (MQSA) has approved image quality testing using the American College of Radiology (ACR) accreditation phantom, CIRS Model 015, which is made of Polymethyl Methacrylate (PMMA) material (Celina et al., 2023).

The study was conducted by Celina et al. (2023) to fabricate a standard ACR phantom with various 3D printer filaments to imitate the fibres, specks and masses. In the study, they used Polylactic Acid (PLA) and Polyvinyl Alcohol (PVA) to fabricate

the in-house phantoms. CATIA v5R2016/2/2010 was the software that the authors used to model the breast ACR phantom fabrication using a 3D printer. The 3D designs were saved in Standard Tessellation Language (STL) format with a 95% fill density setting for the printing process. Fibres were nylon fibre, and specks were aluminium carbonate (Al_2CO_3). As a result, all fabricated phantoms met the ACR standard. PLA and PVA 3D printer filaments showed excellent texture at relatively affordable phantom prices (Celina et al., 2023).

Another study, Groenewald (2017) and Li (2020) suggested using PLA as the phantom housing material because the PLA attenuation coefficient is much closer to human soft tissue. In addition to Groenewald (2017), uses small metallic beads made from tungsten carbide with diameters (0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 1.0) mm were used to evaluate spatial uniformity, noise, and sensitometry. Therefore, fabricating a phantom specific for angiography using listed and tested materials is valuable for image quality assessments.

2.4. Challenges in Subjective Image Quality Assessment

Subjective evaluation by radiologists and medical physicists plays a crucial role in the diagnostic accuracy and efficiency of quality control (QC) (Ho et al., 2022) (Oh et al., 2022). The research that has been done by (Ho et al., 2022) used two experienced radiologists who rely on their 13 and 15 years of experience and perception to assess image quality, which directly impacts the overall diagnostic accuracy of mammography images. This reliance on subjective assessment can influence the success rate of outcomes in image quality evaluation. While subjective evaluation is essential in assessing image quality, the integration of machine learning (ML) and deep learning (DL) has been shown to have a profound impact on diagnostic processes and clinical workflows (Ho et al., 2022) (Oh et al., 2022).

To sum up, subjective evaluation by radiologists is essential for assessing image quality and diagnostic accuracy in radiology. However, the integration of ML and DL and objective evaluation methods can enhance efficiency and accuracy in image quality control results. Hence, continuous learning algorithms implementation in angiography principles can further improve angiography image quality by providing radiologists or medical physicist with feedback loops for ongoing improvement.

2.5. Emerging Solutions in Machine Learning for Image Quality Evaluation

The use of machine learning (ML) in medical imaging has shown promising results in improving image quality and diagnostic accuracy across various modalities. Studies from Ho et al. (2022) and Oh et al. (2022) have demonstrated the potential of deep learning reconstruction algorithms to enhance image quality, particularly in mammogram images.

The results of both studies of the mammography phantom with artificial intelligence (AI) showed high accuracy, reasonable object scoring, and less time-consuming. Ho et al. (2022) stated that the obtained accuracies for fibers, specks, and masses were 90.2%, 98.2%, and 88.9%, respectively.

Hence, the studies demonstrated values that are almost perfectly in agreement between manual evaluation and predicted labels. Thus, implementing machine learning to evaluate image quality in angiography is beneficial to increase accuracy.

2.6. Image Pre-processing Techniques

Image pre-processing techniques are important in visualising structures on the phantom images related to the quality of the image obtained, which helps with diagnosis and treatment planning in a clinical setup. Several research works have focused on

optimising algorithms and methods to improve the precision and effectiveness in phantom images.

Ho et al. (2022) had obtained ACR phantom images from Digital Imaging and Communications in Medicine (DICOM) format files were processed using custom MATLAB scripts. These scripts automatically cropped the images and divided them into a 4 × 4 grid of sub-images. The authors converted the raw images to a grayscale level beforehand. Grayscale images contain only intensity values, reducing the complexity of computations compared to coloured images, which have multiple channels (RGB). In medical imaging, grayscale enhances the visibility of structures, making it easier to detect abnormalities. Many computer vision models, including convolutional neural networks (CNNs), perform better with grayscale images when colour is not a significant factor (Sundell et al., 2022).

Doria et al. (2021) and Sato et al. (2023) explored deep learning-based denoising methods in CT imaging using phantom models, focusing on image pre-processing impacts. The first study from Sato et al. (2023) evaluated a commercial DL-based image processing software (DLIP, FCT PixelShine), applied as a post-processing tool to filtered back projection (FBP) images. The authors demonstrated effective noise suppression while preserving spatial resolution comparable to model-based iterative reconstruction (MBIR) and deep learning reconstruction (DLR), with slight noise texture smoothing indicated by a shift in noise power spectrum (NPS) peak frequency. The second study from Doria et al. (2021) investigated convolutional neural network (CNN)-based denoising using two architectures, an encoder-decoder and a UNet model that had been trained on a large dataset of phantom CT images. The UNet outperformed the encoder-decoder in noise reduction and spatial resolution preservation. However, radiomic analysis revealed that UNet-based denoising could unintentionally alter

texture features, raising concerns about its impact on quantitative imaging analysis. Both studies highlighted the effectiveness of DL-based denoising while emphasizing the need for careful evaluation of potential alterations in image characteristics, especially when quantitative metrics are used for diagnosis or treatment planning.

Sundell et al. (2022) conducted an automated image pre-processing pipeline for phantom-based image analysis. Convolutional Neural Network (CNN) input normalisation was achieved by isolating the phantom's wax area using intensity-based segmentation and correcting its orientation based on target position. The phantom label and the corner area were replaced with noise to prevent model bias. The cleaned phantom image was divided first into sub-images corresponding to specific targets (fibres, masses, or specks). Background intensity non-uniformities were corrected using 2D polynomial subtraction. Finally, all sub-images were resized to 128×128 pixels, normalised to a range of 0–1 intensity scale, and stored as 64-bit floating-point images.

2.7. Feature Extraction and Data Labelling

Feature extraction and data labelling are fundamental components in automated phantom image evaluation pipelines, particularly when integrated with machine learning or deep learning algorithms. These steps facilitates the translation of complex image information into quantifiable data for quality control and diagnostic accuracy (Torfeh et al., 2023).

Ho et al. (2022) implemented a unique feature extraction technique in a phantom study, where a total of 159 features were derived from each pattern image using in-house MATLAB algorithms. The extracted features encompassed multiple categories: position, global, local, edge, and texture information. Position features represented the specific location of the pattern within the phantom, encoded from 1 to 16. Global

features included statistical descriptors such as the mean and standard deviation of gray levels, matrix size, and overall image gradients. Local features were derived from the signal region of interest (ROI) and background, incorporating intensity metrics, edge characteristics, contrast values, and contrast-to-noise ratios. Notably, texture features and gradients were also computed, enhancing the sensitivity of pattern recognition. The signal ROIs were automatically detected, enabling consistency across image sets and reducing human bias (K. Bharodiya, 2022).

In another study focusing on mammography phantom analysis by Oh et al. (2023) highlighted a structured data labelling approach to train a deep learning model for phantom quality assessment. The labelling process was conducted in two stages. First, the phantom area containing the 16 standard test objects (fibers, specks, and masses) was isolated from the raw mammogram using rectangular bounding boxes, effectively removing irrelevant background regions. Second, a detailed scoring system based on the American College of Radiology (ACR) digital mammography quality control guidelines was employed. Each object was scored as 1 (fully visible), 0.5 (partially visible), or 0 (not visible), depending on visibility criteria such as complete structure recognition or partial presence. This process resulted in the labelling of 2,208 phantom images, with classification outcomes of 1,878 images as “pass” and 330 as “fail.” The labelled dataset was a reference for developing supervised learning models capable of scoring objective image quality.

Together, these studies emphasise the critical role of structured feature extraction and accurate data labelling in phantom-based quality control systems. Feature engineering tailored to phantom structure and consistent labelling criteria grounded in clinical guidelines significantly improves the reliability and reproducibility of automated evaluations.

2.8. Image Augmentation

Torfeh et al. (2023) and Doria et al. (2021) employed data augmentation as a critical step to enhance the effectiveness and generalisability of deep learning models trained on phantom images for quality assurance and image denoising, respectively. Torfeh et al. (2023), incorporate augmentation techniques to the ACR MR phantom images to address the limited availability of training data and to simulate a broader range of possible acquisition variations. Their augmentation included spatial transformations such as rotations and translations, aiming to improve the neural network's ability to generalise across diverse image orientations and slight positioning inconsistencies typical in real-world scans.

In contrast, Doria et al. (2021) designed their augmentation process specifically for CT phantom slices, where each 2D slice from a 3D volume was augmented in 90° rotations and flipping. This was performed across 24 axial slices reconstructed along the phantom’s depth, introducing spatial variability in the insert locations to prevent the network from overfitting to fixed object positions. This augmentation was carried out separately for FBP and IR reconstruction methods, resulting in two distinct datasets used for training and testing validation. Both studies underscore that augmentation not only combats overfitting but also simulates real-world variability, ultimately improving the reliability of deep learning models in automated quality assessment and denoising of medical images.

2.9. Machine learning (ML) algorithms

Phantom and machine learning evaluation in the clinical setting requires reliable image quality and automated evaluation methods. Machine learning algorithms have emerged as effective tools for improving the consistency and accuracy of quality

assurance (QA). A recent study applied deep learning models to assess MRI images acquired from ACR phantoms, which are standard test objects used to evaluate the MRI machine performance. By processing phantom images with convolutional neural networks (CNN), the system was able to automatically detect and evaluate key parameters such as spatial resolution and geometric distortion. This approach enables consistent and objective monitoring of MRI system performance, reducing reliance on manual inspection and supporting more efficient imaging workflows (Torfeh et al., 2023; Ramos et al., 2022).

Support Vector Machines (SVM) have attracted significant attention due to their adaptability and effectiveness in classification tasks. In the context of image quality evaluation, SVM has been successfully applied to phantom images in mammography, offering a robust method for automating QA processes. A study by Ho et al. (2022) implemented a one-versus-one SVM classifier trained on segmented pattern images from ACR mammographic phantoms, using dimensionality reduction through Principal Component Analysis and Sequential Minimal Optimisation for efficient training. The model achieved high classification accuracies of 90.2% for fibers, 98.2% for specks, and 88.9% for masses. The authors demonstrated its reliability in identifying visibility levels of phantom features. These findings confirm the practical applicability of SVM in supporting consistent and objective quality assessments in mammographic imaging.

K-Nearest Neighbors (KNN) classifiers have been employed for their simplicity and effectiveness in pattern recognition tasks, including phantom imaging analysis. In the study by Chow et al. (2020), the authors implemented a KNN classifier within a broader machine learning framework to support the correction of deterministic geometric errors in single-plane and dual-plane X-ray fluoroscopy using phantom images. The classifier was used to map error distributions by referencing labelled

instances based on spatial proximity in the feature space, which was derived from phantom-based fluoroscopic projections. Performance of the KNN classifier showed that the model contributed to robust and accurate self-supervised learning of projection errors, precision of 3D reconstruction in fluoroscopy was improved. While KNN is generally sensitive to noise and data scaling, in this controlled phantom-based setting, it proved to be a reliable tool for error detection and correction when supported by well-prepared training data and consistent acquisition conditions.

Random Forest (RF) classifiers are known for their ability to handle high-dimensional data, making them well-suited for radiomics or phantom image applications. A recent study by Hertel et al. (2023) experimented with the stability of radiomics features extracted from phantom scans acquired using a photon-counting detector CT system and evaluated using RF classifier. The classifier was applied in the context of test-retest analysis, aiming to assess which radiomics features remained consistent across repeated scans under varying acquisition and reconstruction parameters. The RF model utilised an ensemble of decision trees to rank feature importance and identify those with high reproducibility. The RF classifier effectively distinguished between stable and unstable features, providing valuable insight into which radiomics metrics are reliable for quantitative imaging studies. Given its ability to manage feature variability and deliver consistent results, the RF model proved to be a reliable approach for analysing the true extracted feature in phantom-based radiomics research.

2.10. ROC curve and AUC analysis

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of classification models across different thresholds. Originally developed during World War II for radar signal detection, ROC

analysis has since been adapted in various fields, including medical imaging, machine learning, and diagnostic testing. In a classification context, the ROC curve plots the True Positive Rate (TPR or sensitivity) against the False Positive Rate (FPR or 1-specificity) at different discrimination thresholds. This visualisation allows researchers and clinicians to assess how well a model distinguishes between two classes, which typically have positive and negative outcomes (Nahm, 2022).

To construct a ROC curve, multiple threshold values are applied to a classifier's prediction probabilities, producing a series of (TPR, FPR) pairs. These points are then plotted in a 2D space, with the x-axis representing FPR and the y-axis representing TPR. A model that performs perfectly would achieve a point in the top-left corner of the ROC space (0, 1), indicating 100% sensitivity and 100% specificity. As the decision threshold varies, the ROC curve is formed, tracing the trade-off between sensitivity and specificity. The curve essentially visualises how many correct positive classifications the model makes at the cost of incorrect positive classifications (false positives) (Gupta, 2024).

The ROC curve provides crucial insights into a model's behaviour. A steep curve that hugs the top-left corner reflects a high-performing model, as it indicates a high TPR with a low FPR at most thresholds. On the other hand, a curve that lies close to the diagonal line from (0, 0) to (1, 1) represents a model with no discrimination capacity, performing no better than random guessing. Models that produce curves below the diagonal may be misclassifying classes or might be reversed in their predictions. This characteristic is useful for model validation and comparison, especially in binary classification settings where both false positives and false negatives carry different clinical or operational consequences (Chan et al., 2022).

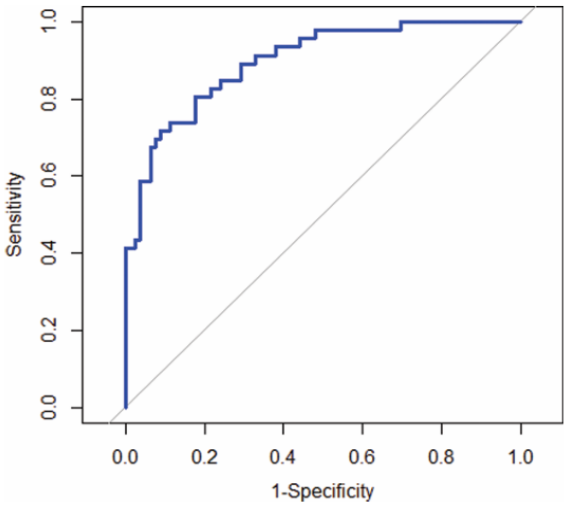

Figure 1.1: A hypothetical ROC curve demonstrating the trade-off between sensitivity and specificity (S. Yang & Berdine, 2017)

One of the key summary metrics derived from the ROC curve is the Area Under the Curve (AUC) or also known as the c-statistics. The AUC quantifies the overall ability of the model to discriminate between positive and negative classes. AUC values range between 0 and 1, where a score of 1.0 indicates perfect classification, and a score of 0.5 implies performance no better than chance. In practice, an AUC between 0.7–0.8 is considered acceptable, 0.8–0.9 is excellent, and above 0.9 is outstanding. AUC is particularly advantageous because it provides a single scalar value that summarises the entire ROC curve, making it easier to compare model performance across different datasets or algorithms (Nahm, 2022).

Table 2.1: Rule of thumb interpreting AUC (S. Yang & Berdine, 2017)

AUC = 0.5	No discrimination, e.g., randomly flip a coin
$0.6 \geq AUC > 0.5$	Poor discrimination
$0.7 \geq AUC > 0.6$	Acceptable discrimination

$0.8 \geq AUC > 0.7$	Excellent discrimination
$AUC > 0.9$	Outstanding discrimination

Additionally, AUC helps in selecting optimal decision thresholds. While the ROC curve itself allows visual exploration of sensitivity and specificity trade-offs, the AUC score can be used in algorithmic tuning or threshold adjustment by identifying the point on the curve that offers the best balance between sensitivity and specificity for a given application. In image quality control tasks, such as those performed in computed tomography (CT), deep learning classifiers are evaluated using AUC to determine how reliably they can identify images that meet or fail diagnostic standards (Gupta, 2024).

In more advanced applications, such as predictive analytics for audit selection (Chan et al., 2022), ROC analysis supports performance evaluation in contexts where skewed class distributions are common. The flexibility of ROC curves in being threshold-independent allows them to assess classifiers on unbalanced datasets, as the curve does not rely on specific class proportions. This property is useful in domains where positive cases such as tax fraud or rare diseases, are rare but critical to identify correctly, and precision-recall curves may complement ROC curves in those settings.

In summary, the ROC curve is a vital diagnostic tool for evaluating binary classification models, providing a clear visualisation of the trade-offs between true positive and false positive rates. The AUC offers a concise metric for comparing model discriminative ability across datasets and threshold settings. As shown in diverse applications from the medical field, such as CT image quality control to financial auditing, the ROC and AUC metrics play a role in verifying and refining classification systems in both clinical and operational domains.

2.11. Performance evaluation metrics

In the medical imaging field, particularly in automating diagnostic quality control (QC) processes such as in computed tomography (CT), evaluation metrics are indispensable for quantifying a model's predictive performance. Gupta (2024) comprehensively applied a range of performance metrics including accuracy, recall, precision, F1 score, specificity, and cross-entropy loss to assess classification models that distinguish between images from calibrated and miscalibrated scanners. These metrics serve as diagnostic tools in determining where models succeed or fall short and enabling continuous and reliable results.

Accuracy represents the ratio of correct predictions to the total predictions made and is commonly used due to its intuitive interpretation. However, its utility diminishes in imbalanced datasets, where a high accuracy might mask poor detection of the minority class (Gupta, 2024 & Litjens et al., 2017). Therefore, accuracy must be interpreted alongside other, more sensitive measures, especially in medical contexts where class imbalance is common.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (1)$$

Recall also known as sensitivity, defined as the proportion of true positives correctly identified out of all actual positives, is crucial in medical diagnostics where failing to detect a condition could have serious consequences. For instance, a low recall in detecting poor-quality CT images may result in undetected miscalibrations, posing clinical risks (Gupta, 2024). Esteva et al. (2019) similarly emphasised the importance of high recall in clinical-grade AI systems to ensure that all relevant pathology or image quality failures are detected.

$$\text{Recall(R)} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

On the other hand, precision measures how many of the positively predicted instances are correct. In QC applications, a high precision implies the model does not raise unnecessary alarms by misclassifying good-quality images as defective. The F1 score, as the harmonic mean of precision and recall, balances these two aspects, offering a single composite metric that is especially useful in scenarios where neither false positives nor false negatives can be tolerated (Gupta, 2024).

$$\text{Precision(P)} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{F1 score} = 2 \times \frac{\text{P}\times\text{R}}{\text{P+R}} \quad (4)$$

Specificity also known as the true negative rate, complements recall by evaluating the proportion of correctly identified negative images that are truly miscalibrated and correctly classified as such. It is particularly important when false positives (incorrectly marking good images as bad) lead to costly re-acquisition or unnecessary concerns. Litjens et al. (2017) pointed out that in high-stakes environments such as radiology, both sensitivity and specificity must be optimised to reduce misdiagnoses and improve decision confidence.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (5)$$

Loss of function, specifically cross-entropy loss used by Gupta (2024), provides a continuous measure of how far the model's predicted probabilities deviate from the true labels. This function is particularly suited to binary classification tasks and is widely used in neural networks and logistic regression models. A decreasing loss over training epochs indicates that the model is improving in aligning its predictions with the

ground truth. Binary Cross-Entropy (BCE) measures how well the model's predicted probabilities match the actual binary labels. A smaller loss indicates better prediction performance. Loss is low when predicted probabilities are close to the true labels (as predicting 0.95 when the label = 1). Loss is high when predicted probabilities are far from the true labels (as predicting 0.1 when the label = 1) (Terven et al., 2023).

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (6)$$

where each components represents,

Table 1.2: Explanation of each mathematical components Gupta, (2024) and Spindelböck et. al, (2021)

Components	Explanation
$\mathbf{n}$	The total number of samples in the dataset (or mini batch)
$\mathbf{y_i}$	The true label for the i-th sample
$\boldsymbol{\hat{y}_i}$	The predicted probability (by the model) that the i-th sample belongs to class 1. This value is in the range (0,1) typically produced by a sigmoid activation function.
$\mathbf{log}(\boldsymbol{\hat{y}_i})$	The natural logarithm of the predicted probability for the positive class (when y_i=1)
$\mathbf{log}(\mathbf{1} - \boldsymbol{\hat{y}_i})$	The natural logarithm of the predicted probability for the negative class (when y_i=0)

$\boldsymbol{y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)}$	This term selects the correct log-probability based on the actual label • If y_i=1: the second term becomes 0, and the loss becomes $-\log(\hat{y}_i)$ • If y_i=0: the second term becomes 1, and the loss becomes $-\log(1 - \hat{y}_i)$
$\sum_{i=1}^{n}$	Adds the loss from all samples in the dataset.

Furthermore, comprehensive evaluations using these metrics are considered best practice in the development of trustworthy medical AI tools (Esteva et al., 2019). Together, these metrics provide a multidimensional assessment of model performance. When interpreted collectively, they offer valuable insight into the trade-offs between different types of errors and help developers make informed decisions in model selection and tuning for clinical applications.

2.12. Ensemble learning

Ensemble learning is a sophisticated machine learning technique that enhances predictive performance by combining the outputs from multiple models. These individual models, referred to as base learners, each contribute to the final output, resulting in more reliable and accurate predictions. The core idea behind ensemble learning is that the strengths of diverse models can be leveraged to compensate for individual weaknesses, thereby improving the generalisation ability of the system, especially when dealing with noisy or complex datasets.

Several ensemble techniques have been developed, each designed to harness different model advantages. Bagging (Bootstrap Aggregating) involves training multiple models on different random subsets of the training data and averaging their outputs to reduce variance. Boosting, on the other hand, builds models sequentially, with each new model attempting to correct the errors made by its predecessor. This approach effectively reduces both bias and variance. Stacking trains a meta-model that learns to combine the outputs of several base models, often yielding better results by capturing non-linear relationships among predictions (Ferrouhi & Bouabdallaoui, 2024).

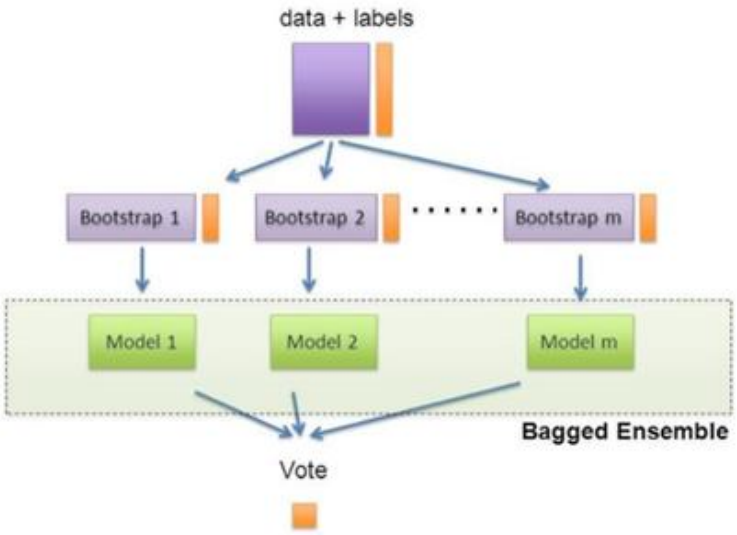

Figure 2.2: Bagging structure (Akbulut et al.,2022)

Another simpler and widely used technique is the voting ensemble. In this approach, predictions from multiple base classifiers are combined through a majority voting rule for classification tasks (hard voting) or by averaging predicted probabilities (soft voting) (Akbulut et al.,2022). Voting ensembles are particularly effective when the base classifiers are diverse in nature such as mixing Decision Trees (DT), Support Vector Machines (SVM), Random Forest (RF) and k-Nearest Neighbors (KNN)