DEEP LEARNING-BASED PREDICTION OF DAILY DELIVERED DOSE FOR ADVANCED LUNG CANCER RADIOTHERAPY

DEEP LEARNING-BASED PREDICTION OF DAILY DELIVERED DOSE FOR ADVANCED LUNG CANCER RADIOTHERAPY

By

SHIRLIE SUET-YEE CHAN, B.Sc.

A Thesis Submitted to the Radiation Sciences Graduate Program and School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

TITLE:	Deep Learning-Based Prediction of Daily Delivered Dose for Advanced Lung Cancer Radiotherapy
AUTHOR:	Shirlie Suet-Yee Chan B.Sc. (Honours Life Physics, University of Waterloo)
SUPERVISOR:	Dr. Marcin Wierzbicki
COMMITTEE:	Dr. Priscilla Dreyer & Dr. Robert Hunter
NO. OF PAGES:	xx, 97

Lay Abstract

Daily changes in the shape and position of the tumour and nearby organs can make the original radiotherapy treatment plan less accurate, which might reduce treatment effectiveness and/or increase side effects. However, creating new plans every day is time-consuming and unfeasible. There lacks a method to quickly verify whether daily changes in the lung are significant enough to render the planned dose ineffective or harmful. This thesis proposes an AI-based tool to predict if the planned dose still applies for the current day anatomy or if a new plan needs to be generated. Before each treatment, an image of the lung is taken. Using this, the AI predicts where the dose will be distributed inside the body, helping therapists confirm the tumour is treated while healthy organs are spared. The results of this study show that the AI model was successful in predicting the dose for two different treatment techniques.

Abstract

Interfractional anatomical changes reduce the accuracy of the planned radiotherapy treatment. However, developing new plans for every treatment fraction is impractical. A 3D U-Net model was adapted and optimised to predict daily dose distributions in advanced non-small-cell lung cancer (NSCLC), enabling evaluation of dose deviations due to interfractional changes and supporting timely re-planning decisions.

The U-Net was trained using planning CT, daily cone-beam CT (CBCT) images, and associated dose distributions, from 24 patients with stage III NSCLC who received 63 Gy in 30 fractions by IMRT (n=13) or VMAT (n=11). Rigid registration was performed to align CBCT images with the corresponding planning CT images, from which dose was re-calculated in the treatment planning system. These CBCT-based dose distributions were assumed to represent the daily delivered (true) dose. Model performance was assessed using a leave-one-out cross-validation (LOOCV) scheme applied separately to three cohorts: IMRT-only (n=13), VMAT-only (n=11), and combined (n=24) patients. Prediction accuracy was quantified using gamma (3%/3mm) analysis, with a 20% maximum dose threshold.

The IMRT- and VMAT-trained models achieved mean gamma pass rates of $98.1 \pm 1.4\%$ and $90.1 \pm 2.6\%$ at 3%/3mm, respectively. The combined-trained models, validated on IMRT and VMAT datasets, attained mean pass rates of $97.3 \pm 1.3\%$ and $93.6 \pm 2.4\%$ at 3%/3mm, respectively. Gamma failures appeared to correlate with heterogeneous tissue regions, steep dose gradients, and anatomical changes, particularly within the planning target volume (PTV). The higher prediction accuracy of IMRT may be attributable to its fixed beam angles, compared to the continuously rotating gantry in VMAT delivery.

The strong agreement of model predictions with true dose distributions demonstrate the potential of U-Net-based dose prediction as a patient-specific, dosimetric verification tool in advanced NSCLC radiotherapy. Further work is needed to test robustness and generalisability of the U-Net model in predicting absolute doses using larger and more diverse datasets.

Keywords: deep learning; 3D U-Net; dose prediction; non-small-cell lung cancer (NSCLC); radiotherapy; treatment planning

This work is dedicated to my 爹地 *&* 媽咪*,*

for their sacrifices, selfless love, unwavering support, and endless encouragement.

Acknowledgments

First and foremost, I want to express my sincerest gratitude to my supervisor, Dr. Marcin Wierzbicki, for his guidance and support during the last two years. Marcin, your insight and patience were instrumental in shaping, not only this thesis, but also how I think and approach problems as a researcher. I am sincerely thankful for the opportunity, your hands-on assistance with trouble-shooting code, and your willingness to bounce ideas with me whenever I knock on your door. I would also like to extend appreciation to my committee members, Dr. Priscilla Dreyer and Dr. Rob Hunter, for their constructive feedback and for the knowledge I gained from the courses they taught.

A heartfelt thank you to both Dr. Ernest Osei and Dr. Brenda Lee, whose passion for teaching and dedication to student success deeply inspired me throughout my undergrad at the University of Waterloo. Their enthusiasm for the field of medical physics helped cultivate my own interest and commitment to pursuing research in this area.

As the only student in my research "group", I am immensely grateful to my office mates–Chloe, Jigar, Curtis, Helen, and Michelle–for everything from sweet treat runs to late-night headaches trying to complete assignments, and every moment in between that made grad studies both bearable and memorable. Thank you, Chloe and Helen, for taking the time to explain concepts and solve problems together. To Jigar and Curtis, going to COMP together was an unforgettable experience and I will miss sitting right in the middle of your bickering (and also instigating arguments).

Finally, my deepest gratitude goes to Mom, Dad, Sam, Sabelle, and Arthur, for being my support system and a constant source of encouragement. I would not be where I am without all of you believing in me and being there every step of the way, through all of the tears and breakdowns. Your love and faith have been, and continue to be, my strength, for which I am eternally grateful as I face new challenges and strive to make a difference.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Γ	Gamma index
3DCRT	Three-dimensional conformal radiotherapy
AAA	Analytical anisotropic algorithm
AdaGrad	Adaptive gradient (algorithm); others include AdaDelta and Adam
Adam	Adaptive moment estimation
AP	Anterior-posterior (direction, dose profile)
ART	Adaptive radiotherapy
AXB	Acuros External Beam
BN	Batch normalisation
CBCT	Cone beam computed tomography
CCC	Collapse cone convolution
cGy	Centigray
CNN	Convolutional neural network

CT	Computed tomography
CTV	Clinical target volume
CV	Cross-validation
D95	Minimum dose that covers 95% of the target volume (usually PTV)
DD	Dose difference
DL	Deep learning
Dmax	Maximum dose received by target volume
Dmean	Mean dose received by organ
Dmin	Minimum dose received by target volume
DOF	Degrees of freedom
DSC	Dice similarity coefficient
DTA	Distance-to-agreement
DVH	Dose-volume histogram
EBRT	External beam radiotherapy
EPID	Electronic portal imaging device
FC	Fully connected
FFF	Flattening filter-free
GTV	Gross tumour volume
Gy	Gray
HD	Hausdorff distance
HI	Homogeneity index
HU	Hounsfield unit
IGRT	Image-guided radiotherapy
IMRT	Intensity-modulated radiotherapy

IOE	Intelligent optimisation engine
ISO	Isocentre
ITV	Internal target volume
JD	Jacobian determinant
kV-CBCT	Kilovoltage cone beam computed tomography
kV-iCBCT	Kilovoltage iterative cone beam computed tomography
Linac	Linear accelerator
LBTE	Linear Boltzmann transport equation
LOOCV	Leave-one-out-cross-validation
LR	Left-right (direction, dose profile)
MAE	Mean absolute error
MC	Monte Carlo
MLC	Multileaf collimator
MSD	Mean surface distance
MSE	Mean squared error
MU	Monitor units
NSCLC	Non-small cell lung cancer
OAR	Organ(s)-at-risk
OBI	On-board imaging
PB	Pencil beam
PFS	Progression free survival
PTV	Planning target volume
PRV	Planning organ-at-risk volume
PSQA	Patient-specific quality assurance

QA	Quality assurance
ReLU	Rectified linear unit
ROI	Regions-of-interest
Rx	Prescription (dose)
TERMA	Total energy released per unit mass
SABR	Stereotactic ablative body radiotherapy
SC	Superposition/convolution
SCLC	Small cell lung cancer
SD	Standard deviation
SGD	Stochastic gradient descent
SIB	Simultaneously integrated boost
TPS	Treatment planning system
TV	Target volume
VMAT	Volumetric arc radiotherapy

Declaration of Academic Achievement

I declare that all the work presented in this thesis, titled "Deep Learning-Based Prediction of Daily Delivered Dose for Advanced Lung Cancer Radiotherapy", is my own work and does not involve any plagiarism or academic dishonesty.

I certify that I have read this thesis and that, in my opinion, it fulfills the requirements in both scope and quality for the degree of Master of Science.

Chapter 1 Introduction

1.1 External Beam Radiotherapy for Advanced Stage Non-Small Cell Lung Cancer

Lung and bronchus cancer remains a leading cause of cancer-related mortality worldwide, claiming more lives than breast, colorectal, and prostate cancers combined [1]. Despite the significant 3.8% decline per year in mortality rates in both sexes since 2015, an estimated 32,100 cases and 20,700 deaths are still expected in Canada in 2024 [2,3]. Lung cancer can be classified as small cell (SCLC) or non-small cell lung cancer (NSCLC) according to its origin and histology. NSCLC accounts for 80-85% of newly diagnosed lung cancer cases, the development of which is divided into four stages [4]. Advanced NSCLC refers to stage III (locally advanced) or stage IV (metastatic) cases with disease spread beyond the primary lung site, into the mediastinum, surrounding lymph nodes, and/or to distant organs. The high mortality rate associated with lung cancer is largely attributable to the absence of symptoms in the early stages, resulting in more than 70% of patients being diagnosed at an advanced stage, most of which are not amenable to potentially curative surgery [4].

Treatment options for lung cancer depend on the type, tumour stage, and medical condition of the patient, and may be in the form of chemotherapy, radiotherapy, targeted therapy, immunotherapy, or, more frequently, a combined modality approach [4]. About one-third of patients present with locally advanced NSCLC, and population studies suggest that 39-52% of these patients may be treated with external beam radiotherapy as a monotherapy [5]. The conventional regimen of thoracic radiotherapy alone, that involves delivering 2 Gy per fraction for a total dose of 60 Gy, is associated with a 2-year progression free survival (PFS) as low as 20-30% and an even lower 5-year survival of 5-7% [6,7]. Technological advancements in planning and delivery of external beam radiotherapy (EBRT) by a linear accelerator (linac), namely the use of imaging before and during treatment, has enabled precise and conformal dose delivery to stationary targets, thus maximizing tumour cell killing and healthy tissue sparing [8,9]. However, key techniques that aid in optimising the therapeutic ratio in lung cancer treatment, including three-dimensional conformal radiotherapy (3DCRT), intensity-modulated radiotherapy (IMRT), volumetric arc radiotherapy (VMAT), image-guided radiotherapy (IGRT), and adaptive radiotherapy (ART), continue to be confounded by interfractional anatomical changes that lead to dose deviations and compromised treatment accuracy [9].

1.1.1 Radiotherapy Treatment Planning and Delivery

Treatment planning in radiotherapy is referred to as the process of optimising treatment parameters and determining the most effective way to irradiate and manage the disease. Treatment planning is a multi-step process individualised for each patient, beginning with treatment simulation, which involves patient positioning and acquisition of a CT simulation scan [10]. This is followed by contouring, where the target volume(s) and adjacent critical organs-at-risk (OARs) are delineated. Subsequent steps include the selection of appropriate beam arrangement, computation of the doses to be delivered, evaluation of the resulting dose distributions, and finally the transfer of treatment planning information to the delivery system [10].

Treatment simulation is mainly performed on a computed tomography (CT) scanner via a computer software that generates a three-dimensional representation of the patient in treatment position. At the initial stage of simulation, the patient is immobilised in the treatment position on the CT table using immobilisation devices that are included in the image study and can be indexed on the linac treatment table to ensure position reproducibility [10]. Additional reference skin marks (or tattoos) also assist in guiding patient alignment and positioning during each treatment fraction, as they coincide with lasers in the treatment room. CT images offer superior spatial resolution and accurate mapping of tissue electron density information, derived from CT Hounsfield Units (HUs), which are essential for localisation and delineation of target(s) and surrounding OARs, as well as for precise dose calculations [11,12].

Target delineation on the CT scan consists of identifying and contouring the gross tumour volume (GTV). This is expanded to account for subclinical disease spread, creating the clinical target volume (CTV). Since the extent and location of the target can change as a function of time due to motion of internal organs, cardiac activity, patient breathing, and other physiological factors, an internal target volume (ITV) is required to encompass the CTV, which includes an internal margin. For lung targets, the ITV is defined using 4D CT images of the target at different phases of the respiratory cycle, such that the contour covers the full range of motion of the target during the entire cycle. The planning target volume (PTV) is the ITV plus a setup margin to account for geometrical uncertainties in patient setup, and variations in its size, shape, and position relative to the planned treatment beam(s), respectively [10]. It follows that adequate dose to the PTV at each fraction presumes adequate treatment of the entire disease-bearing volume. Next, all relevant OARs are segmented (or contoured) on the same CT scan to prevent dosing tissues beyond established tolerance limits. In parallel to the definition of the PTV, it may be necessary to define a planning OAR volume (PRV), which includes the OAR and a margin of position and motion uncertainty [10].

The contoured CT dataset is represented within the computerised treatment planning system (TPS) to simulate treatment without physically re-positioning the patient. Treatment planning involves the

selection of appropriate beam parameters (energy, arrangement, angles, shapes), the computation of dose distributions based on contoured target(s) and OARs, and iterative adjustment to maximise target coverage while minimizing dose to nearby normal tissues [10]. To achieve highly conformal dose distributions, advanced radiotherapy delivery techniques such as IMRT and VMAT have been developed. IMRT is an advanced form of 3D conformal radiotherapy (3DCRT) in which multileaf collimator (MLC) leaves dynamically change to modulate the radiation beam, generating multiple small beamlets that each deliver a portion of the total dose, and fixed fields with variable intensities to achieve conformal dose distributions [13]. In 3D-CRT, 3-5 beams of uniform intensities are shaped by MLC leaves to match the 3D outline of the target [13–15]. In contrast, the dynamic movement of the MLC leaves in IMRT improves conformity in volumes with complex concave shapes and facilitate dose escalation [13,16]. A standard IMRT plan for locally advanced lung cancer consists of fixed gantry angle beams, each subdivided into numerous smaller beams, which prolongs treatment delivery time and compromises patient comfort, affecting position reproducibility and introducing intrafractional motion [14,15]. On the other hand, VMAT delivers radiation in a single, continuous 360° arc around the patient, providing a significant reduction in treatment time [17]. VMAT also confers superior dose conformation and homogeneity within the target volume, as well as better sparing of OARs, further reducing side effects and enhancing treatment effectiveness [13,17]. In essence, the ability of VMAT to simultaneously and continuously change gantry rotation speed, field shape, and dose rate, results in distinct dose distribution characteristics. Regardless of the technique, the treatment plan must be optimised prior to delivery.

Plan optimisation is laborious and often necessitates multiple iterations of beam adjustments through collaboration with dosimetrists, physicists, and oncologists. This optimisation utilises inverse treatment planning, wherein the desired dose distributions or clinical objectives are first set and then used to guide beam configuration and intensity choices. This approach differs from “forward planning”, as used in 3D-CRT, where beam parameters are established first, followed by dose calculation. To elaborate, the

optimisation algorithm in inverse planning is generally considered to consist of (1) an objective function that describes the clinical criteria and assigns a numerical score to each plan, and (2) a method to reach the optimum intensity distribution by minimising the objective function. The objective function is designed to quantify the agreement between a plan and clinical goals, with lower values corresponding to better adherence to the prescribed criteria and constraints [18]. For dose-based algorithms, a quadratic objective function is typically used for mathematical convenience, and measures the difference between the planned dose distribution and clinical goals (e.g., target prescription dose, dose homogeneity, critical organ maximum dose, critical organ dose-volume constraints) by summing the squares of these differences. Stochastic and deterministic algorithms can be used, and differ in how they search for optimal combinations of treatment parameters (e.g., beam angles, intensities, shape). Stochastic algorithms, such as simulated annealing, incorporates randomness in the search process by introducing a small positive or negative change of different magnitude with each iteration, which enables local minima escape and increases the chance of finding a global optimum [10,18]. Deterministic algorithms, such as a simple gradient implementation, follow a predefined path and always proceed to the nearest minimum, with the solution updating with each iteration based on the gradient of the dose difference [18]. Deterministic methods work best for convex or well-behaved optimisation problems, where the objective function is differentiable, continuous, and has a single global optimum (versus multiple local optima) [10,19]. They may also be combined with stochastic methods for complex, multi-objective optimisation scenarios. The effectiveness of the algorithm to find the global optimum directly influences the quality of the final treatment plan. At each iteration, after adjustments to beamlet weights or MLC positions, dose calculation is performed using pencil beam (PB) algorithms to assess agreement of current solution to the plan objectives [10,20]. In short, inverse planning is an iterative process, where parameters are adjusted, dose is recalculated, and the plan is against clinical goals repeatedly, until an optimal or acceptable solution is determined. Once the optimal fluence or segment weights are

established, the plan may be recalculated using Acuros XB, which is better equipped at accounting for tissue heterogeneities and complex geometries for final dose computation.

Each optimised plan then undergoes verification, which involves an assessment of the plan quality, independent dose calculations, and evaluation of treatment delivery by the machine, to ensure that the planned dose can be delivered accurately to meet clinical goals. A physicist confirms the correct monitor units (MUs) and model parameters are transferred from the TPS to the treatment machine. In addition to patient setup verification, electronic portal imaging devices (EPID) are also used for patient-specific quality assurance (PSQA) to verify deliverability and safety of the planned beam fluence and dose distribution [21,22].

The current treatment planning workflow is limited by the lack of prior knowledge regarding what is achievable or viable, leading to an inefficient and time-consuming trial-and-error process. Another challenge is that the anatomy captured in the planning CT scan may not be representative of treatment delivery. In other words, treatment delivery according to this plan is susceptible to errors arising from interfractional anatomical changes, which may culminate in deviations between the planned and delivered dose distributions.

1.1.2 Image-Guided Radiotherapy

To ensure precise verification of treatment delivery and mitigate inaccuracies related to positioning and patient motion, image-guided radiotherapy (IGRT) has been used to visualise changes in size and location of the target and adjust for positional changes through the treatment process [23–26]. Among the various technologies available, kilovoltage cone beam computed tomography (kV-CBCT) is the most commonly used IGRT technique for patient setup verification and tumour position confirmation immediately prior to each treatment fraction. The Varian TrueBeam CBCT (Varian Medical Systems, Palo Alto, CA) on-board imaging (OBI) implementation involves a cone-shaped, kV x-ray beam and a flat-panel detector at 90° to

the beam, to generate high-contrast, three-dimensional volumetric image in a single gantry rotation (~130 seconds for full standard, 360° lung CBCT) [8,26]. These systems are also integrated with robotic couches for 6-degrees-of-freedom (DOF) adjustments. In comparison, conventional CT for simulation make use of a fan-shaped beam, wide multi-detector arrays, and helical scanning, with acquisition times of ~8 seconds (64-slice helical CT) [27]. The resulting daily CBCT image is automatically or manually registered with the plan CT using rigid image registration to match bony anatomy or soft tissue landmarks. Any discrepancies between the current and planned position are translated into couch shifts to achieve alignment.

A key limitation to kV-CBCT is its inability to directly predict the effect of significant anatomical changes due to weight loss, atelectasis, pleural effusion, tumour progression or shrinkage, differential shifts of targets, etc., on the delivered dose. In clinical practice, the TPS operates independently from the linac, meaning that re-computation of the delivered dose to reflect anatomical changes is not feasible on the required time scale. Consequently, decision-making at the treatment console relies solely on geometric alignment, without consideration of whether dosimetric goals are still being met. This disconnect between image guidance and dosimetric evaluation prevents radiation therapists from making real-time decisions that maintain delivery consistency with the planned dose. Some other limitations of kV-CBCT include its susceptibility to image artefacts caused by increased scatter, including cupping, streaking, and ring artefacts [28,29]. Compared to the small, multi-row detector arrays used in CT, a large 2D flat panel detector is used in CBCT, which collects more scattered radiation. Without proper correction methods, these artefacts can distort the true attenuation values and broaden the HU peaks representing tissue density, potentially degrading HU accuracy and compromising CBCT-based dose calculations and re-planning efforts [30]. Despite these drawbacks, kV-CBCT is useful for the accurate, three-dimensional localisation of the target and OARs and daily monitoring of anatomical changes, ultimately improving treatment outcomes and facilitating adaptive radiotherapy (ART) strategies.

1.1.3 Adaptive Radiotherapy

The availability of daily in-room imaging has enabled the advancement of ART, such that PTVs and overall treatment plans can be continuously updated and refined to maximise the therapeutic ratio according to anatomical and functional information acquired over the course of treatment. This may be in the form of modifying initial treatment plan parameters (e.g., field margins, number of fractions) to account for the new representation of the patient or altering the aggressiveness of clinical goals based on the disease progression. Simply, ART is a solution to handling daily anatomical changes and individualising plans to conformally and accurately irradiate targets [31]. The frequency with which ART is used depends on the rate of anatomical changes, as cases with fast progression and random variations may require more frequent adaptation, either offline or online [32,33]. In one example of offline ART, a new CT simulation image is acquired during the treatment series and used to guide re-planning in between fractions, with the updated plan applied in subsequent treatments. This approach takes several hours or days as it uses the same software and workflows used in full radiotherapy planning. In another instance, offline CBCT-based re-planning makes use of the daily CBCT acquired by CBCT systems with iterative reconstruction, as well as improved calibration and correction methods, which have demonstrated discrepancies less than 2% between dose calculation using planning CT vs. CBCT with the Acuros XB algorithm [34]. Re-optimisation of the treatment plan using the daily CBCT scan requires manual intervention, repeat simulation and additional planning time between fractions outside the standard workflow [32]. This type of re-optimisation can take several hours, requiring manual intervention and additional planning time (re-contouring, plan review, and QA) between fractions outside the standard workflow.

In online ART, image acquisition and in-session modifications are completed in the same fraction (with the patient in treatment position), meaning that matching the CBCT to the plan CT for table correction is no longer needed [32]. For example, the daily CBCT may be imported directly in the TPS, where the reference plan CT contours are propagated to the CBCT and the treatment plan is re-optimised

based on the daily anatomy. The 2020 Varian Ethos (Siemens Healthineers Inc., Erlangen, Germany) is a commercially available, online ART system that features a software/linac combination comprising AI-based target and OAR segmentation, a machine learning-enhanced TPS, an intelligent optimisation engine (IOE), and improved kV-CBCT image quality to produce online adaptive plans in 15 to 25 minutes [35,36].

The Ethos is built upon the Halcyon, a ring-gantry linac featuring a 6 MV flattening filter-free (FFF) beam operating at 800 MU/min dose rate, 4 RPM gantry rotation, dual-layer MLC system, and ultra-fast iterative kV-CBCT (kV-iCBCT) reconstruction algorithm. Halcyon was designed to streamline IGRT with simplified workflows, high treatment throughput, and integrated kV-CBCT, forming the basis for the online adaptive functionally of Ethos [37]. The iCBCT reconstruction method offers a higher contrast-to-noise ratio and HU accuracy than conventional CBCT in support of direct CBCT dose calculations. With the addition of a 4 RPM gantry rotation, CBCT scans can be acquired in a single, 17-second breath-hold to further minimise motion artefacts. Image acquisition by the Ethos 2.0 with HyperSight CBCT also uses iCBCT but is significantly faster, providing 6-second CBCT scans with larger field-of-views (up to 70 cm) and fewer artefacts [35,38]. Notably, HyperSight incorporates advanced model-based scatter and metal artefact reduction algorithms, enhancing image quality and achieving HU accuracy close to that of diagnostic CT. These advancements in image quality play a critical role in accurate AI-driven auto-segmentation, as clarity and soft-tissue contrast directly impact auto-contouring performance and precision of ART.

The adaptive treatment delivery process of Ethos consists of three steps: influencer review, target review, and plan selection. Within the TPS, disease sites are organised into modules, or “intents”, each of which is associated with selected, auto-contoured structures and organs situated near or within the target area, known as “influencer structures” [35]. Based on these influencers, a structure-guided deformable image registration (DIR) is performed, generating a newly deformed CT that retains the HUs from the simulation CT but conforms to the CBCT anatomy. Target volumes from the original, clinician-approved plan are subsequently propagated onto this synthetic CT, where dose distributions for both targets and OARs are

recalculated using the updated contours. The reference plan re-computed on the daily anatomy, known as the "scheduled plan", is compared to the "adaptive plan", which is a plan fully re-optimised on the daily anatomy and designed to meet clinical goals. The adaptive plan is one where the beam parameters were re-optimised to better achieve current objectives associated with anatomical changes [35]. The radiation oncologist then selects the most suitable plan and transfers it to the delivery system, allowing patient-specific dose delivery in real-time. If the adaptive plan is chosen, a "clinical approval" from the oncologist and a "technical approval" from the medical physicist must be given [39]. Overall, online ART expedites real-time adjustments to interfractional deformations that a standard CBCT-guided treatment (using only couch corrections) cannot accommodate. For lung cancer specifically, online ART offers dosimetric benefits in target coverage, reduced lung dose, and OAR sparing[40,41].

Still, online ART fails to address and adapt to instantaneous intrafractional changes, such as continual bladder filling or movement of abdominal gas pockets [32]. In locally advanced NSCLC, similar intrafractional changes, including respiratory motion, transient airway obstruction, and organ motion, also lead to inaccuracies in contours, distortions in dose calculations, and deviations in delivered dose. Ideally, real-time ART would provide a plan that considers the extent of both interfractional and intrafractional changes, though limited by current technologies. To add, compared to their offline counterpart, CBCT-based online ART solutions such as the Ethos are not widely available. It is not possible to implement the Ethos workflow on an existing non-Halcyon linac and manual implementation in current treatment paradigms is complicated by personnel, resource, and time constraints, leading to concerns about potential oversights. Thus, there is a demand for a solution that offers a balance between conventional offline and fully adaptive treatment workflows to rapidly and accurately assess the effect of interfractional changes on the delivered dose.

1.2 Applications of Deep Learning in Radiotherapy

Deep learning (DL) is a subset of machine learning that relies on multilayered artificial neural networks, with emerging applications in medical imaging data analysis [42]. For instance, DL methods have been widely applied for image segmentation tasks, in which a medical image is set as the input layer and a classifier is obtained as output. In these tasks, each pixel (or voxel) in the input is assigned a label, creating a map of classifications that identifies which pixels (or voxels) belong to specific anatomical structures, such as a tumour or an OAR [43].

Among various DL architectures, convolutional neural networks (CNNs) are particularly well-suited for pattern recognition within grid-like images. CNNs are composed of convolutional, pooling, and fully connected (FC) layers [44]. Convolutional layers apply a convolution operation, wherein small matrices (or tensors) called kernels (or filters) pass over the input. At each pixel (or voxel), the weights of the kernel are multiplied elementwise with the corresponding values in some region of the input matrix function, and the resulting products are summed to generate a single value in the output feature map [44]. This process detects spatial patterns, such as edges and shapes, with each feature map representing a different aspect of the image. These feature maps are then downsampled by pooling layers, and mapped by a subset of FC layers to the final outputs of the network, such as the probabilities for each class in classification tasks [44]. In radiology, CNNs have been successfully employed for tumour detection, organ segmentation, image reconstruction, treatment response, outcome prediction, and automated annotation, leveraging large datasets to learn clinically relevant imaging structures [45,46].

A CNN is trained using supervised learning with paired inputs and known (or annotated) target outputs [47]. During training, the network iteratively updates its parameters through backpropagation to minimise the loss function, which quantifies the difference between the predictions on the input data and the ground truth (i.e., labels). Briefly, backpropagation calculates the gradients of the loss function with

respect to each network parameter using the chain rule. By propagating the error through the network layers, the effect of each parameter on the overall error can be determined [44,47]. Based on the gradients, an optimisation algorithm will subsequently update the parameters accordingly in a direction that reduces the loss, thereby improving the predictions over multiple training iterations (termed epochs).

More recently, CNNs have also been used to predict dose in radiotherapy treatment planning [48,49]. Dose prediction is a form of regression task, where the output is a continuous dose distribution rather than a binary or categorical mask as in segmentation tasks. The implementation of DL dose prediction in treatment planning can enable rapid plan evaluation, facilitate adaptive workflows, inform decision-making, and reduce reliance on manual trial-and-error approaches and inter-planner variations for different radiotherapy techniques.

1.2.1 U-Net Architecture

U-Net is an end-to-end CNN architecture initially proposed by Ronneberger et al. [50] for medical image segmentation applications. The novelty of U-Net and its superior segmentation capability lies in its U-shaped structure and use of skip connections, whereby convolutional layers are linked with corresponding upsampling layers, facilitating simultaneous feature extraction and pixel classification [44,50]. The symmetric architecture of U-Net is composed of an encoder and a decoder, respectively known as contracting (or analysis) and expansive (or synthesis) paths (Fig. 1) [50,51]. The encoder identifies and extracts relevant features of the input image and the decoder upsamples the learned features to construct a segmentation map, enabling each pixel in an image to be classified into distinct categories for object delineation. Simply, the encoder learns the "what", and the decoder learns the "where" in the image.

The three-dimensional (3D) U-Net is an augmentation of the basic two-dimensional (2D) U-Net encoder-decoder framework, in which 2D operations are substituted with equivalent 3D operations, namely 3D convolutions, 3D max pooling, and 3D up-convolutions (or transposed convolutions), that results in 3D

segmented images. Due to the abundance of repeating structures and shapes in 3D images, this type of network can be trained quickly even with scarcely annotated examples. Each depth (or level) of the encoder consists of two successive 3×3×3 convolutions, followed by batch normalisation, rectified linear unit (ReLU) activation, and a 2×2×2 max pooling layer [44]. The U-Net depicted in Fig. 1 has a depth of four, with three encoder blocks and three decoder blocks [51].

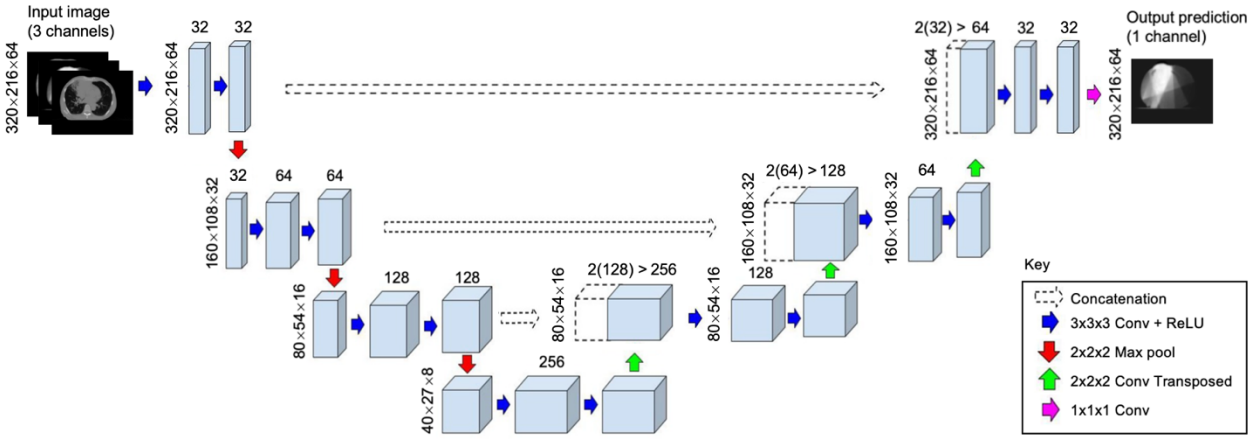

Fig. 1 Three-dimensional (3D) U-Net architecture. The network consists of a downward encoder that learns features through successive convolutional and pooling layers, and an upward decoder that reconstructs the feature maps through upsampling and transposed convolutional layers. Skip connections between corresponding encoder and decoder layers preserve spatial detail. Numerical values at the top of rectangular prisms represent the channel dimension (i.e., number of feature maps). Adapted from Aboussaleh et al. [51]

In the convolutional layer, convolutional kernels (or sliding filters) pass over the input image, such that their weights are multiplied elementwise with the corresponding values in some region of the input matrix (or tensor), and the resulting products are summed. It should be noted that the kernel itself also has a fourth channel dimension, which is equivalent to the number of channels in the input (i.e., C_{in}×3×3×3). This process is repeated as the kernel slides over the entire input, generating a new output matrix called a feature map. Each value in the feature map represents the result of applying a kernel to a specific region of the input (i.e., matrix multiplication), with each kernel capturing a different detail. The number of kernels used in a layer determines the number of feature maps produced by that layer, and is doubled at each

successive depth to capture increasingly complex representations. For example, the output feature map after the first convolution by 32 kernels would have 32 feature maps and hence a shape of 32×320×216×64.

To continue, batch normalisation (BN) normalises these outputs to have zero mean and unit variance per channel, followed by scaling and shifting with learnable parameters. This normalisation mitigates internal covariate shift, where the distribution of input values keep changing as parameters are updated, slowing training convergence [44]. By standardising the values before activation functions and subsequent layers, training stability and speed can be improved.

After each convolutional layer, the ReLU function performs an elementwise operation on the feature maps, returning the input value if it is positive and zero otherwise [44]. This prevents the problem of vanishing gradients, which occurs as gradient signals backpropagated from output to input layers become exponentially small (or "vanish"), meaning that parameters receive very small updates [47]. This can slow down or stop the training process altogether, impeding on the ability of the network to effectively learn in deep layers.

Next, a stride of two is assigned to the max pooling operation to downsample the feature map. The max pooling function splits the positive-only feature map into non-overlapping 2×2×2 regions and selects the maximal element from each [44]. This identifies the most heavily weighted features and halves the size of the input. In the bottom-most layer, the encoder is linked to the decoder by a bottleneck consisting of two convolutional layers, each preceding a ReLU, facilitating input data compression. Max pooling does not have a channel dimension and operates independently on each feature map.

As mentioned earlier, the innovation of U-Net lies in its use of skip connections [50]. In the decoder, the feature map is upsampled to the original size of the input image, wherein each block of the expansive path includes a 2×2×2 transposed convolution layer and a concatenation with the feature map from the respective layer in the encoder. Note again that the kernel has a channel dimension equal to the number of channels in the input feature map (i.e., kernel has shape C_{in}×2×2×2). Before a transposed convolution

operation, two 3×3×3 convolutions are performed, each followed by a ReLU [50]. The transposed convolution, with a kernel size of 2×2×2 and stride of two in each dimension, doubles the spatial dimensions of the input feature map. To put it another way, each voxel in the input contributes to a 2×2×2 block of voxels, rather than a single voxel, in the output. A stride of two ensures that these blocks are spaced to give an output that is twice as large. A simpler, 2D representation of transposed convolution with 2×2 kernel is shown in Fig. 2.

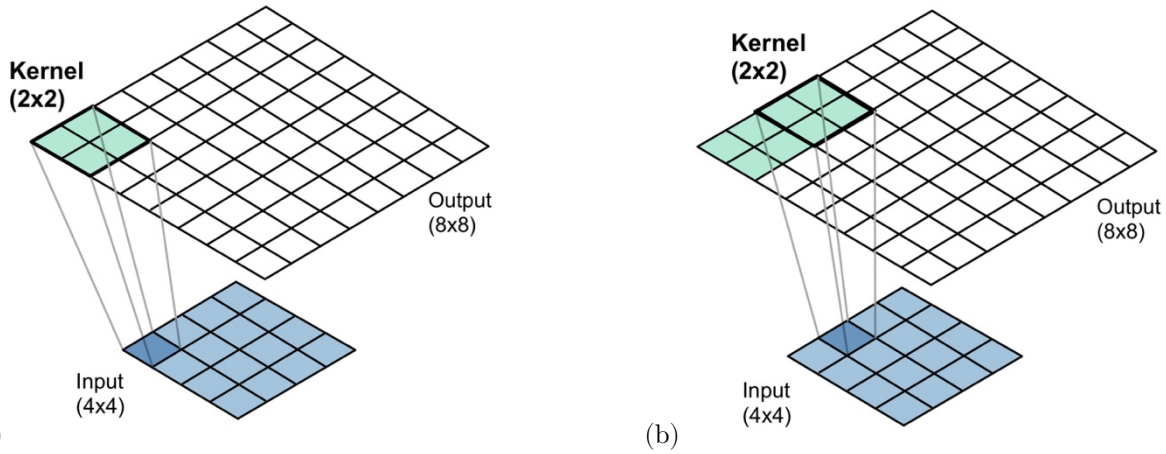

Fig. 2 Two-dimensional (2D) transposed convolution operation. Transposed convolution with a 2×2 kernel (in green), and stride of two, applied to a 4×4 input (in blue, bottom) gives an 8×8 output (top). (a) Kernel is applied to the first input element at position (0,0); and then to (b) an adjacent input element, two pixels over (due to stride of two). Essentially, each input pixel contributes to four output pixels.

Again, the transposed convolution kernel has a channel dimension that is equivalent to the number of kernels applied, such that setting the number of kernels to half of the number of feature maps at the bottleneck reduces the number of feature maps by half prior to concatenation to maintain symmetry with the corresponding encoder layer.

The up- and down-sampling of feature maps raises concerns about the potential loss of information. To ensure the preservation of details, skip connections are implemented by concatenating the cropped feature maps from the encoder with the corresponding feature maps in the decoder (i.e., at the same level) along

the channel dimension before upsampling [50]. This concatenation essentially combines feature maps from earlier layers to later ones of the network into a single tensor, forming larger feature maps that leverage both abstracted features and spatial precision, and improving object localisation by the network. Notably, cropping of encoder features is necessary given the loss of border pixels with every convolution.

Finally, the output of the last decoder block is passed through a 1×1×1 convolution to map each feature vector at each voxel to the desired number of output classes [50]. This 1×1×1 convolution reduces the number of channels in the feature maps to the desired number of output channels, all while preserving the spatial dimensions of the input image. In Fig. 1, this last convolution would employ kernels with shape 32×1×1×1, such that a weighted sum across all 32 input channels at each voxel is computed, resulting in a single value per voxel. This way, the output consists of a single channel with the same spatial dimensions as the initial input. Overall, the U-Net architecture, with its encoder-decoder structure and skip connections, provides a robust framework for learning multi-scale details and abstract patterns, making it particularly well-suited for dose prediction in radiotherapy.

1.2.2 U-Net for Dose Prediction

Multiple studies have investigated U-Net (and other variants) for dose prediction across different anatomical sites and treatment modalities. Table 1 summarises recent studies, highlighting treatment site, input data, network architecture, dataset and training, and key findings [52–57].

All six studies adopted the Adam optimiser with an initial learning rate of 1×10^{-3}, in conjunction with a mean squared error (MSE) loss function for training. Adam is a common choice in DL given its rapid convergence (i.e., minimisation of the loss function) and good generalisation performance [52–57]. In terms of data preprocessing, Liu et al. [55] and Wang et al. [56] performed normalisation of dose maps to standardise the range of dose values across all training samples, thereby ensuring effective and consistent model training. Liu et al. [55] further normalised CT images within the range of [-1000, 3000] and also

resampled CT and dose images to a uniform pixel spacing of $2.5 \times 2.5 \times 3$ mm, making it easier for the model to generate predictions. Regarding network architecture, Nguyen et al. [52] employed a 2D U-Net, which was fed 2D slices as inputs. However, a 3D U-Net may be more adept in modelling the spatial continuity or gradients of dose distributions between anatomical slices, along the superior-inferior (z) axis of the patient. Hence, a 3D approach processes volumetric data directly, providing a more complete representation of dose-volume relationship. Wang et al. [56] explored the impact of network depth on 3D U-Net-based dose prediction for cervical cancer radiotherapy and determined that a model with a depth of 5 achieved the smallest dose deviations and the highest homogeneity index (HI), outperforming depths 3 and 4 ($p < 0.05$) in D98 and D99 (i.e., dose received by 98% and 99% of the target volume, respectively). Interestingly, despite having more parameters, the model with a depth of 5 showed no evidence of overfitting, while the shallower models with depths 3 and 4 did overfit. The term overfitting is used to describe the performance of the model on unseen validation data. Models that overfit essentially memorise the training inputs and are not able to generalise well to new data. Though deeper networks may be prone to overfitting, they can also learn more complex relevant features of dose distirbutions. This underscores the importance of evaluating different depths, as the optimal depth may vary depending on the task, data available, and training strategy, with trade-offs between accuracy, robustness, and resource constraints.

Moreover, to address the complexity of multidimensional data, variants of U-Net that incorporate residual or dense dilated networks have been implemented in the literature. Liu et al. [55] proposed a 3D residual U-Net with GAN-based loss that involves an adversarial training process between a generator and a discriminator network, allowing more realistic dose distirbutions to be generated, which enhanced the clinical plausibility of the predicted distribution. Cao et al. [57] examined the use of 3D dense dilated U-Net to predict 3D dose distributions of VMAT, including SIB techniques, and identify suboptimal plans to guide re-optimisation.

Table 1 Summary of recent studies using U-Net-based dose prediction

Study	Treatment Site / Technique	Input Data	Dataset & Training	Network Architecture	Purpose	Key Findings
Nguyen et al. [52]	Prostate IMRT	Six contours (PTV, bladder, body, rectum, femoral heads)	88 patients (single slice with PTV), 10-fold CV	2D U-Net (d=7) + CNN layers	Simple dose prediction model to accelerate plan re-optimisation	Achieved DSC=0.91 between predicted and true isodose volumes, and average mean and max dose differences within 5.1% of Rx dose
Willems et al. [53]	Prostate VMAT	CT (+GTV mask), contours, ISO	79 patients (77 Gy/35 fx; n=36 with SIB), 5-fold CV	3D U-Net (d=4)	Predict dose from contours vs. CT-only to reduce plan time & variability	Models with CT+ISO+contours outperformed CT-only model in terms of Dmax, D50, D98, and achieved low error rate of 2.5±1.2%
Ma et al. [54]	Prostate IMRT	Four OARs (PTV, rectum, bladder, body); 10 optimal DVHs	97 patients (77 train, 20 test)	3D U-Net (d=4)	Dose prediction based on DVH & anatomy to address different trade-offs	Multiple Pareto optimal plan DVHs were used for training to simulate trade-offs and physician preferences, achieving dose differences $\leq 1.7 \pm 0.8\%$ of Rx dose for OARs and PTV
Liu et al. [55]	Cervix IMRT	CT, eight contours (PTV, body, bladder, rectum, bowel, spinal cord, femoral heads)	130 cases (100 train, 10 validation, 20 test)	3D residual U-Net with GAN-based loss	Predict dose for OARs, target as plan evaluation tool to improve treatment outcomes	Achieved DSC=0.87 and max PTV dose difference of 2.968 ± 2.840 Gy, indicating strong overlap between predicted and true dose regions
Wang et al. [56]	Cervix VMAT	Binary OAR mask with 6 channels (PTV, body, bladder, rectum, SI, colon)	92 patients (50.4 Gy/28 fx; 72 train, 10 validation, 10 test)	3D U-Net (d=3,4,5)	Test impact of network depth on accuracy and robustness of U-Net dose prediction	Depth 5 model achieved smallest dose differences and highest homogeneity index (HI) between actual and predicted doses, outperforming depths 3 and 4 ($p<0.05$) in D98 and D99
Cao et al. [57]	Lung VMAT (with SIB)	CT, target and OAR contours, prescribed dose	93 patients (75 train, 18 test) (35-72 Gy)	3D dense dilated U-Net	Predict dose to identify suboptimal plan quality, guide re-optimisation	Achieved a mean voxel-wise dose difference of -0.49 ± 0.54 Gy between actual and predicted distributions, and plans guided by prediction improved OAR sparing without compromising target dose coverage

Note: PTV=planning target volume; GTV=gross tumour volume; OAR=organ-at-risk; DVH= dose-volume histogram; ISO=isocentre; SIB=simultaneously integrated boost; DSC=dice similarity coefficient; CV=cross-validation; SI=small intestine; Rx=prescription

The added dense connections enable layers of the network to share information, and the dilated convolutions expand the receptive field using spaced-out kernel elements to capture wider context without losing resolution or increasing network parameter count. However, dilated convolutions can introduce gridding artefacts, as a consequence of sparse sampling, and cause uneven coverage of spatial features. Although these model modifications were able to extract more complex features and improve prediction accuracy, they generally involved higher computational costs, more challenging optimisation, and longer training times compared to the base 3D U-Net structure. These methodological choices guided the design of the present study, aiming to optimise training stability and prediction accuracy, with trade-offs in computational demand and duration of training.

The most common input data type for dose prediction was planning CT images, with or without contours. These studies typically reduced dose to simple one-dimensional metrics for specific anatomical structures from DVHs, such as mean dose and volume receiving some dose threshold, which are then related to outcome by statistical models [52–57]. To add, model robustness usually followed a cross-validation strategy, and its generalisability greatly depends on the choice of validation set. The studies reviewed here used small retrospective and single institution datasets, highlighting data scarcity as one of the biggest issues in development of deep learning models with medical data. In parallel, these papers focused on retrospective plan evaluation or guiding re-optimisation following treatment delivery, rather than integrating dose prediction into prospective planning workflows. In spite of its routine use in IGRT, CBCT is also absent as an input in the proposed models. By integrating CBCT into the prediction pipeline, the actual dose delivered can be verified in real time, reflecting both daily anatomical changes and dose deviations.

1.3 Motivation and Significance of Research

The aim of this thesis is to investigate the potential of a 3D U-Net model for daily 3D dose distribution prediction in patients with stage III NSCLC treated by IMRT and VMAT. The model utilises three routinely acquired clinical inputs, namely the planning (reference) CT image, the associated planned dose distribution, and daily kV-CBCT images, to generate a dose distribution prediction reflective of the daily patient anatomy. This DL approach is proposed as a rapid method for assessing and quantifying the dosimetric impact of interfractional anatomical changes. As re-planning for every fraction can be time-consuming and online ART strategies (such as the Varian Ethos) are resource-intensive, a dose prediction model that may be implemented on any linac with CBCT represents a practical alternative for estimating cumulative dose delivered and identifying cases with clinically relevant dose deviations that necessitate re-planning.

Chapter 2 Materials and Methods

2.1 Computational Environment

The computing workstation used in this study was equipped with an Intel® Core™ i9-14900K CPU (US Intel Corporation, Santa Clara, CA, USA), 64 GB DDR5 RAM, a 4 TB NVMe SSD, and an NVIDIA GeForce RTX 4090 GPU with 24 GB GDDR6X memory (NVIDIA Inc, Santa Clara, CA, USA). The programming language used was Python v3.12 (Python Software Foundation, Wilmington, DE, USA), and the deep learning framework employed is PyTorch 2.3.1 with CUDA v12.2 (Meta AI, Menlo Park, CA, USA). All training and evaluations were conducted on an Ubuntu v24.04 operating system.

2.2 Dataset

The dataset consists of 24 randomly selected patients with stage III NSCLC who received lung RT with daily CBCT image guidance. These patients (n=24) were treated with radical intent to a prescribed dose of

63 Gy in 30 fractions. Among them, 13 patients were treated by IMRT and 11 by VMAT. Data for each patient was anonymised and comprised the plan CT, planned dose distribution, daily CBCT, and the daily dose distributions. For each fraction, the CBCT was first aligned to the planning CT before the initial planned beam parameters were applied to recalculate the dose on the CBCT and generate the daily dose distributions. These computed CBCT-based dose distributions on a 2.5mm voxel grid were considered the ground truth (or true dose), in this study, representative of the daily delivered dose.

For IMRT treatments planned in the Pinnacle TPS v16.2.1 (Philips Medical Systems, Milpitas, CA), rigid registration was performed using an in-house software tool, aligning CBCT images with the corresponding planning CT images based on couch corrections recorded at the treatment unit. After registration, the corrected CBCT geometry was exported, on which dose was recalculated via the collapsed cone convolution (CCC) algorithm available in Pinnacle.

For VMAT plans, rigid registration was already applied on the unit to the CBCT images during six-degrees-of-freedom (6DOF) couch corrections. Prior to dose calculation on the CBCT in the Eclipse TPS v18.1 (Varian Medical Systems, Palo Alto, CA), a support structure model of the Varian Exact IGRT treatment couch (Varian Medical Systems, Palo Alto, CA) was inserted and manually positioned in the TPS to account for beam attenuation by the couch [58]. Subsequently, dose was re-calculated using Acuros External Beam (AXB) v16.1.2 algorithm in Eclipse, mapping the original beam geometry, MLC settings, and MUs on the imported CBCT.

It should be noted that CBCT scans with severe artefacts, poor image quality (caused by patient motion), or partial exclusion of the target volume (due to the limited field of view), were omitted from the dataset to facilitate consistent and accurate dose prediction. Approximately 5% of the total 720 fractions across the 24 patient datasets were excluded from further analysis.

2.3 Data Preprocessing

Several preprocessing steps were applied to prepare image data for analysis. The end goal of preprocessing was to construct, for each treatment fraction, a multi-channel input for the U-Net model comprising the plan CT, plan dose grid, and CBCT, while the ground truth output was the resampled daily CBCT-based dose distribution. This way, the U-Net can learn the spatial and contextual relationships between the planned treatment intent and daily anatomy, relate dose distribution to anatomy, and ultimately predict the daily delivered dose distribution. To achieve this, all images needed to be anatomically aligned and consistent in terms of spatial dimensions across fractions and patients.

Raw image data, including .img files from Pinnacle and .dcm (DICOM) files from Eclipse, were loaded and converted into NumPy arrays for ease of manipulations. Specifically, each .img file was read in binary mode and a custom function was used to extract known voxel dimensions from the corresponding header file to reshape the flat binary buffer into a 3D array. On the other hand, each DICOM series was parsed into a 3D volume using the ImageSeriesReader from SimpleITK module, which reads the spatial metadata to stack the 2D slices in anatomical order [59]. The voxel data from the resulting 3D volume was then converted into arrays using the GetArrayFromImage function from the same module. To aid model convergence, reduce complexity, and create comparable and interpretable results, CT images were then normalised to the maximum image intensity value and dose distributions were normalised to the maximum dose value for each fraction. Traditionally, CT images use a 12-bit depth for storage, allowing for $2^{12} = 4096$ discrete intensities or shades of gray (ranging from 0 to 4095) [60]. These raw integer values refer only to the stored digital values from the CT scanner and are not equivalent to HU by default. The Pinnacle TPS stores these values as "CT numbers" ranging from 0–4095, while the Eclipse TPS uses the HU scale directly, ranging from -1024–3071, for its CT-to-density conversion. Briefly, Pinnacle expects and uses the full digital range (0–4095), while Eclipse uses the HU values as exported from the CT scanner. Correspondingly, the

voxel intensities of both CT and CBCT images were divided by 4095, mapping their values from [0, 4095] to [0, 1] in subsequent pre-processing steps. If conversion to HU is required, the DICOM header contains metadata with rescale parameters (slope, intercept) that can be applied to the raw values. Further, instead of using absolute physical dose in units of cGy, dose normalisation helps improve training stability and generalisation such that the model learns to predict relative dose patterns.

In previous work, the exported and readily accessible CT and CBCT images of patients treated by IMRT (n=13) were already registered. In contrast, for patients treated by VMAT (n=11), rotational (pitch, roll, yaw) and translational (x, y, z) transformations had to be applied to the daily CBCT images to achieve alignment with the corresponding planning CT images. More clearly, these matrix transformations were executed in the TPS for visualisation; however, the aligned images could not be exported directly. As such, the transformations were re-applied to the CBCT NumPy arrays. These transformation parameters are collectively described by a 6DOF transformation matrix, consisting of a 3×3 matrix describing rotations and a 3×1 vector representing translation. These matrices were calculated by the IGRT system, which performs automated 6DOF patient positioning corrections based on bony anatomy and the carina. This is followed by manual adjustment to improve target matching and registration of the daily CBCT to the planning CT. In short, these matrices reflect both automatic and manual adjustments.

Next, the CT and CBCT images were trimmed by converting the couch top position from CT metadata into an array index and setting the intensity of voxels below that index to zero, effectively removing the couch from the image. This step is necessary given that the couch does not represent patient anatomy or contribute to meaningful learnable features for dose prediction. It is also a potential source of high-density artefacts, discontinuities, and non-anatomical biases that may interfere with image alignment and model generalisability. All images were cropped to extract the body region and focus on relevant structures. The cropping bounds in the x (left-right), y (anterior-posterior), and z (superior-inferior) directions were defined for the IMRT patients by evaluating the mean voxel intensity across slices orthogonal

to the desired cropping axis. For instance, bounds in the x-direction were identified by iterating from lateral edges toward the center of the image and selecting the first slice where the mean intensity exceeded a predefined threshold. A small padding margin, proportional to the voxel size with a value of zero, was added to the boundary to avoid exclusion of anatomical structures. This threshold-based cropping process is repeated in the y-direction, scanning from the anterior to the posterior surfaces of the patient. In the z-direction, cropping bounds were determined by centering on the center of the target for IMRT and expanding symmetrically above and below, for a total of 104 slices. So far, these steps ensure consistent and patient-specific image cropping while excluding non-anatomical areas, such as air, background, and treatment couch. The bounds and image size of (432×288×104) established for IMRT patients were then applied to VMAT patients, without patient-specific target localisation prior to cropping. To ensure that anatomical structures and target volumes in VMAT-treated patient images were not erroneously clipped by these IMRT-derived cropping bounds, all slices in the cropped volumes were visually inspected (using matplotlib python library) against the original images. The resulting size-consistent volumes retained information relevant to dose prediction (i.e., the entire thoracic region with sufficient margins) while reducing computational load in model training.

Furthermore, the input and output of the U-Net must be represented in the same matrix, with a 1:1-pixel spatial correspondence. However, the original plan CT scans had an in-plane pixel size of 1.17 mm and a slice thickness of 3 mm, while the dose grids consisted of isotropic voxels measuring 2.5 mm. To address this, a trilinear interpolation function was written and used to obtain the dose at each CT image voxel location, producing dose distributions that match the dimensions and spatial resolution of the CT. Trilinear interpolation estimates $D(x, y, z)$, the dose value at point (x, y, z) in the CT, as the weighted average of the intensities of its nearest eight neighbours

$$D(x, y, z) = \sum_{i=0}^{1} \sum_{j=0}^{1} \sum_{k=0}^{1} w_{ijk} D_{ijk} \tag{1}$$

where D_{ijk} is the dose value at the corner voxel indexed by $(x_{ijk}, y_{ijk}, z_{ijk})$ and $w_{ijk}$is the corresponding interpolation weight, illustrated in Fig. 3. The weights are determined by the distances along each of the three axes, such that closer neighbours have a larger influence on the interpolated value. For instance, the weight assigned to point D_a relative to its position from the point of interest $D(x, y, z)$ is $w_a = (x_1 - x)(y_1 - y)(z_1 - z)$. Both the planned and CBCT dose distributions were resampled in this manner and cropped using the same boundaries as the plan CT image.

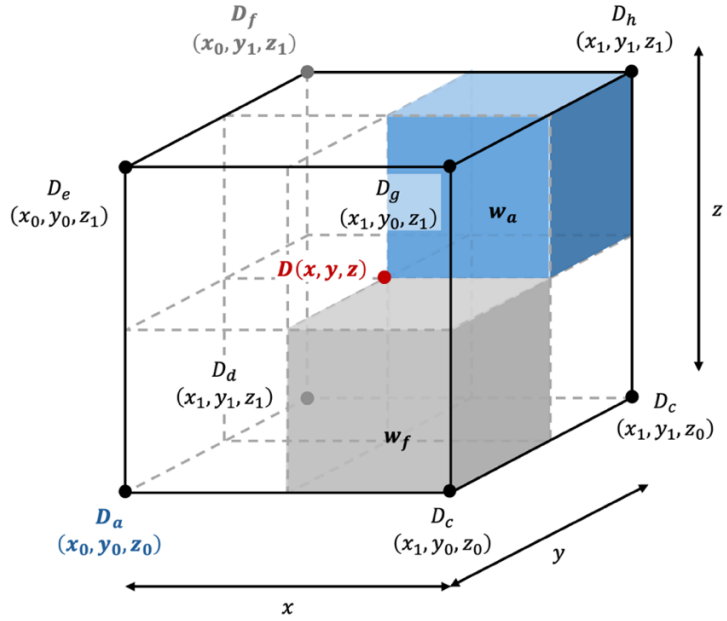

Fig. 3 Trilinear interpolation in 3D grid. Dose at an arbitrary point $D(x, y, z)$ is estimated as a weighted dose average of the eight surrounding voxels, with weights w dependent on how far the point of interest is from a particular neighbour.

Lastly, the plan CT, planned dose distribution, and CBCT images, were stacked along the channel dimension, creating a 3-channel 3D array wherein each channel corresponds to a separate image. This is analogous to an RGB image, where the red, blue, and green channels represent different image data. This

stacked representation allows the model to simultaneously leverage information from three different sources for improved dose predictions. There were two additional restrictions to the final image dimensions. First, the dimensions were required to be integer values and divisible by 2^n, where n is the depth of the network (i.e., number of encoder-decoder layers) due to the repeated downsampling (and corresponding upsampling) operations. Recall that downsampling (max pooling) is performed with each depth, reducing the spatial dimensions by a factor of 2. To ensure that the feature maps in the bottommost layer of the network have integer spatial dimensions, the starting input image dimensions must therefore by divisible by 2^n. Second, it became apparent early in the work that processing many (432×288×104) images with 3 channels each required significantly more GPU memory than was available. To bypass this constraint, the stacked images and CBCT dose distributions were rescaled to 74% in the x and 75% in the y dimensions, and excess slices in the z-dimension (either empty or not relevant to the lungs) were removed to form a final size of (320×216×64). This final dimension was chosen to accommodate both the 2^n divisibility requirement and memory constraint.

2.4 U-Net Model Implementation

The 3D U-Net architecture code was adapted from GitHub and implemented as a class in Visual Studio Code (Microsoft Corporation, Redmond, WA, USA) [61]. The input starts with dimensions (3×320×216×64), with the first value representing the channel dimension (i.e., in_channels = 3) along which the plan CT, plan dose, and CBCT images are stacked, and the last three being the image dimensions (x width, y height, z depth). As the input progresses through the network, a single channel CBCT dose distribution prediction, with dimensions (1×320×216×64), will be generated (i.e., out_channels = 1). Specifically, at each depth along the encoder path, features of the input are extracted by convolutional layers employing a (3×3×3)-sized convolutional kernel, and its dimensions are halved by a max pooling operation.

Recall that the kernel itself also has a fourth channel dimension, which is equivalent to the number of channels in the input. This way, the convolution sums over the channel dimension, collapsing it to produce a single scalar per spatial location per kernel. As an example, if 32 kernels, each with shape (3×3×3×3), are applied to the input with shape (3×320×216×64), the output will have shape (3×320×216×64). Each kernel sums the three input channels and produces one output channel, and 32 kernels generate 32 such output channels. The number of kernels applied at each convolutional layer was manually set such that the number of feature maps (containing the activation or output values) is doubled with each subsequent convolutional layer. Doubling the number of kernels, and subsequently the number of feature maps (output channels), is a common design to increase representational capacity with the reduction in spatial resolution. Each convolutional layer is followed by a normalisation and ReLU activation.

As the feature maps move up the decoder path, their dimensions are doubled through 2×2×2 transposed convolutional layers. At each level of the decoder, the upsampled feature maps are concatenated with the corresponding encoder feature maps at the same depth through skip connections [50]. This is followed by two convolutional layers involving (3×3×3)-sized kernels and ReLU activations. The number of kernels and feature maps decreases symmetrically with those of the encoder. Finally, the output of the last decoder block is convolved with a (1×1×1)-sized kernel to reduce the feature maps to a single output channel, producing the final CBCT dose prediction. This final output effectively assigns a predicted dose value to each voxel, reflecting the influence of anatomical variation (between plan and daily anatomy) and planned dose intent as described by the inputs. No activation function is applied at the output layer, as dose is a continuous, non-categorical quantity.

All convolutional kernel weights and biases were initialised using PyTorch default settings, which apply Kaiming initialisation for neural networks with ReLU activations. In a neural network, weights are numerical values that determine the influence of an input on the final output. Kaiming initialisation ensures a balanced starting point by sampling each weight from a normal distribution with a mean of zero and a

variance scaled according to the number of input units to the layer [62]. By accounting for the number of connections fed into each layer, this scaling helps the network maintain an appropriate amount of variability and support stable convergence. Without proper weight initialisation, the network may display instability during training, with the signal vanishing or growing uncontrollably, especially given the inclusion of ReLU activations which zero out any negative values. All bias values were initialised to zero, a standard approach to prevent skewing of the output in early stages of training. Biases are adjustable values added at each layer to allow the network to shift the output up or down.

2.5 Training Workflow

A leave-one-out-cross-validation (LOOCV) approach was employed to maximise usage of the relatively small dataset size. With this method, nearly the entire dataset is used for each fold, providing an unbiased and robust estimate of model generalisability. To be specific, multiple instances of the model are created, each trained on all but one patient dataset, with the remaining set used for validation. For each training iteration, the stacked input of each patient in the training set was fed into the network, and a predicted dose distribution was then generated for each fraction and compared to the corresponding true dose distribution using a voxel-wise loss function. Learnable parameters, specifically the weights and biases of the convolutional kernels, change in small increments based on the gradients of the loss function (i.e., partial derivatives of the loss with respect to each parameter). These gradients are calculated through backpropagation and guide the optimiser in adjusting the parameters to minimise the difference between the prediction and target distributions and reach the global minimum. The training workflow is described in Fig. 4 below.

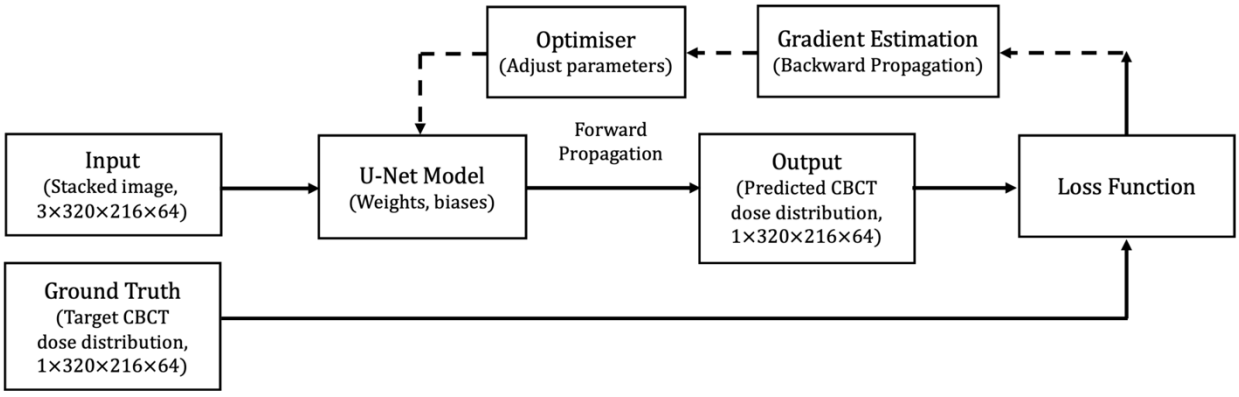

Fig. 4 Outline of training workflow with backpropagation. U-Net parameters (weights and biases) start with random values. Inputs (training samples) are fed into the network and the error between the generated output (prediction) and true dose distributions is calculated by the loss function. Based on the gradient of the loss function (gradient estimation by backpropagation), the optimiser updates the network parameters, such that the loss is minimised.

2.6 Hyperparameter Optimisation Strategy

The performance of the U-Net model is dependent on several interdependent architecture and training hyperparameters, including the depth of the network, the number of convolutional kernels per depth, loss function, number of epochs, batch size, and optimiser. The impact of these hyperparameters on model performance and prediction accuracy were assessed and subsequently optimised using different subsets of IMRT datasets prior to training. The selected values, informed by both empirical results and previous work, were then fixed throughout the training and evaluation processes.

2.6.1 Network Depth and Number of Convolutional Kernels

The depth of the network, d, is the number of downsampling/upsampling levels in the encoder-decoder structure, which defines the extent of feature extraction [62]. As depth increases, the total number of convolutional layers also increases. Thus, a greater depth value indicates a deeper network capable of

learning more abstract features at different resolutions. Still, research comparing U-Nets of varying depths has shown that, although deeper models generally perform better in complex prediction (regression) or segmentation tasks, performance gains do not scale indefinitely and there are practical limitations [56,64,65]. The optimal depth depends on the dataset, application, and available computational resources. Deeper networks not only require much more GPU memory and longer training times, but the added layers may also cause performance to plateau, leading to marginal increase or even deterioration in prediction accuracy due to overfitting or training instability [64,66].

To add, the depth determines the number of convolutional operations, with the number of convolutional kernels per layer doubling at each depth. As such, deeper U-Net have more layers and a greater number of convolutional kernels, increasing the dimensionality, or number of independent features (analogous to degrees of freedom), represented at a particular depth. Simply, the number of convolutional kernels, c, refers to the number of feature maps per network depth (or layer) [44]. For example, a convolutional layer with 32 kernels applies 32 different filters to the input, producing 32 output feature maps. A greater number of kernels enables enhanced feature extraction, though at the cost of increased computational load. In short, feature extraction at a particular image resolution depends on the depth, while the effectiveness with which the features are learned or represented depends on the number of kernels at that depth. Models of depths 3, 4, and 5, with initial kernel counts of 16, 32 and 64, were compared to assess the impact of network depth and number of convolutional kernels on model performance and prediction accuracy. Table 2 summarises the depth and kernel combinations tested.

Previous work focusing on a 2D U-Net for dose distribution prediction utilised 150 epochs and mean squared error (MSE) loss function for training on 12 patient datasets [67]. Building on this, the optimisation of network depth and number of convolutions per depth in this work was conducted with 150 epochs and MSE loss, as well as a batch size of 1 due to the small test size of the training dataset, composed two IMRT patients (5 fractions each). Each model configuration was then validated with 5 fractions of an another

randomly selected IMRT patient. This setup was designed to establish an appropriate degree of model complexity to learn the input-to-output mapping and produce accurate predictions (as measured by gamma pass rate, see Section 2.7), independent of the number of times the training dataset was passed through the network (i.e., number of epochs) and without the risk of overfitting or unstable optimisation.

Table 2 Depth and kernel configurations

Depth	Initial Kernel Count	Kernel Configuration (Per Layer)
3	16	16, 32, 64
	32	32, 64, 128
	64	64, 128, 256
4	16	16, 32, 64, 128
	32	32, 64, 128, 256
	64	64, 128, 256, 512
5	16	16, 32, 64, 128, 256
	32	32, 64, 128, 256, 512
	64	64, 128, 256, 512, 1024

2.6.2 Loss Function and Number of Epochs

It is important to note that MSE was initially used to reduce the search space and identify an effective model architecture. However, the choice of loss function can impact training dynamics, model behaviour, and prediction accuracy. The impact of loss function type on model performance was assessed by comparing prediction results obtained using mean squared error (MSE), mean absolute error (MAE), and a combined MSE and MAE function. Mean square error, also known as L2 loss, is calculated as the averaged square difference between the predicted and target image values, expressed as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \tag{2}$$

where n is the total number of voxels and y_i and x_i are the predicted and target dose value at voxel i, respectively [68]. Mean absolute error, also known as L1 loss, is calculated as the average absolute difference between the predicted and target dose values, given by

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{3}$$

with the same terms as MSE [68]. Compared with MAE, MSE offers faster convergence and consistency, and penalises large errors more heavily. By squaring the difference between predicted and target dose values, the model is encouraged to minimise larger deviations. In MAE, different errors are equally weighted and penalised linearly, meaning that outliers do not have a substantial impact on the loss. A combined MSE and MAE loss function leverages the advantages of each, defined as

$$\text{CombinedLoss} = (\alpha)\text{MSE} + (1-\alpha)\text{MAE} \tag{4}$$

where α is the weighting factor balancing the MSE and MAE components, with α=0.5 indicating that equal weight is given to both [68]. Intuitively, dose prediction requires large dose errors to be minimised, suggesting that MSE may be more suitable for this application. However, dose distributions can have noise or uncertainties, particularly near tissue boundaries, to which overfitting can be prevented by choosing MAE. With these considerations, a combined loss function that emphasises MSE (70%) with some (30%) contribution from MAE (α=0.7) would help minimise large dose errors and reduce sensitivity to noise. A combined function with equal contributions from MSE and MAE (α=0.5) was also tested.

Generally, the loss function can impact the nature of the solutions to which the model converges and how quickly it does so. As such, the loss function that (1) reached stability and convergence the fastest, trained and validated on all fractions of a single IMRT patient, and (2) achieved the highest gamma pass rate in a single instance of LOOCV on IMRT datasets, within 250 epochs would represent the "optimal" choice. To be explicit, faster convergence means the model reaches low training loss in fewer epochs (or less time), signifying increased training efficiency, while high pass rate ensures that efficiency correlates to clinical accuracy [62].

During training, multiple forward and backward passes are performed for parameter optimisation. As described earlier, the number of epochs is the number of times the entire training dataset is passed

forward and backward through the network during training [69]. The number of epochs was tested in increments of 25 (from 50 to 250) concurrently with the comparison of loss function by performing a single instance of LOOCV with IMRT datasets. The optimal number of epochs was defined as the point at which under- and over-fitting were balanced, indicated by a plateau in gamma pass rate.

2.6.3 Non-Optimised Parameters

Batch size and optimiser choice were not optimised in this study and set to commonly accepted default values and those established in previous work.

The batch size determines the number of training samples that are simultaneously processed during gradient estimation, before model weights are updated in an epoch [69]. For clarity, a training sample consists of the stacked 3-channel input (planning CT, planned dose distribution, and daily CBCT) and the target output (CBCT-based dose distribution) for a single fraction. For each batch, a forward pass computes predictions and loss, and a backward pass calculates the gradients of the loss with respect to model parameters, and the optimiser (discussed later) updates those parameters accordingly. Essentially, the batch size dictates the frequency with which weights are updated per epoch, with smaller batch sizes leading to more frequent updates and larger batch sizes resulting in fewer updates per epoch. A small batch size of 1 would allow the model to process each sample at a time and immediately update the weights, yielding better results at the cost of slower convergence, longer training times, and noisier gradient updates [69]. Conversely, a large batch size of n can leverage the entire dataset before changing model weights, attaining faster convergence and more stable gradient estimates, but requires excessive GPU memory and also risks the network falling into the local minima [54]. In this study, a batch size of 5 samples (fractions) was selected based on previous work, balancing computational efficiency and memory constraints with training stability [67].

The model was trained using backpropagation, a method that calculates the gradients of the loss function with respect to each parameter (i.e., weights and biases), working backwards from the output to the input. Specifically, for each voxel in the output feature map, an intermediate gradient (known as a delta term) is computed using the chain rule, which accounts for how changes in the weighted input of that voxel propagate through subsequent layers and contribute to the overall loss. Backpropagation utilises these intermediate gradients (deltas) to calculate the gradients of weights and biases and determine how changes to these parameters affect the weighted input of each voxel. More clearly, the gradient of a kernel weight is the delta of each voxel multiplied by the corresponding input value to which the weight was applied, summed over all output voxels (i.e., weight gradients = delta $\times$ input). The gradient of a bias is simply the sum of the deltas across all output voxels in the feature map (i.e., bias gradients = sum of deltas). A gradient descent optimisation algorithm then uses these weight and bias gradients to iteratively update the weights and biases (in the direction opposite to the gradient, down towards the global minimum) before the next batch within the epoch, minimizing the loss function and improving prediction accuracy [70]. A previous student investigated several different implementations of gradient descent, including the Stochastic Gradient Descent (SGD) optimiser, as well as adaptive gradient algorithms, such as AdaGrad, AdaDelta, and Adam. They concluded that the Adam (adaptive moment estimation) optimiser converged the fastest and to the lowest loss value, and was thus chosen for this study [71]. It is often described as a first-order, gradient-based iterative optimisation algorithm that combines the strengths of both momentum and adaptive learning strategies [72]. Adam computes adaptive learning rates for each parameter by tracking exponentially moving averages of both the gradients (first moment) and the squared gradients (second moment). In other words, it checks how steep the slope is (first moment) and how fast the slope is changing (second moment) by comparing network predictions with the target. Given that both moment estimates are initialised to zero and hence biased towards zero in early training steps, an additional bias correction is applied to normalise the estimates to improve stability.

Adam is governed by the learning rate (α), decay rates (β_1 and β_2), and constant epsilon (ϵ) added to the denominator to avoid division by zero (for numerical stability) [72]. The learning rate scales the gradient and controls the global step size (or maximum possible step size per update). While Adam adapts the learning rate, a default initial rate of $\alpha = 1 \times 10^{-3}$ has demonstrated good performance [72]. The decay rates of the first and second moment estimates, β_1 and β_2, control the direction and scaling of learning for each parameter individually, respectively. A high β_1 (close to 1) means that past gradients are more heavily weighted, with a greater influence on the optimiser compared to new changes, whereas a low β_1 renders the optimiser more sensitive to recent gradients. Considering that β_2 tracks the variance of gradients over time (for each parameter), a high β_2 means that more weight is given to past square gradients, resulting in a slow change of the estimate over time. It follows that a low β_2 would make the adaptive learning rate more reactive but possibly less stable. The default rates, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, have been proven to provide sufficient balance between stability and adaptability [72].

To discourage overfitting, a weight decay $= 1 \times 10^{-5}$ was also implemented as a parameter shrinkage step that is separate from the gradient update, leading to smaller weights, better regularisation, and more consistent training behaviour without slowing down the learning process [73]. Hence, the parameters are reduced by a factor proportional to 1×10^{-5} at each optimisation step to improve training stability. The hyperparameters introduced in this section continue to be adjusted by Adam throughout training, across multiple batches and epochs, to minimise loss.

2.7 Evaluation of Model Performance and Prediction Accuracy

The U-Net model was first evaluated on a dataset of 13 IMRT patients, wherein each model was trained with 12 of the 13 patient datasets. To test model robustness across radiotherapy techniques and anatomy, the best and worst IMRT models were validated using a single patient treated by VMAT. Second, the model

was evaluated on a dataset of 11 VMAT patients using the same LOOCV approach (training each model instance on 10 patients and validating on the remaining one). Finally, generalisability of the model across radiotherapy techniques was analysed by repeating LOOCV on a combined dataset of 13 IMRT and 11 VMAT patients, producing a total of 24 instances of the model.

For both hyperparameter optimisation and performance evaluation, the accuracy of the prediction was quantified by gamma analysis and depicted by gamma index maps. Gamma analysis is a voxel-based evaluation that considers both the distance-to-agreement (DTA) and dose difference (DD) between the reference (i.e., target) and predicted dose distributions, making it a comprehensive tool for dose quality assurance in radiotherapy.

For each voxel in the reference dose distribution, the gamma index (Γ) is calculated for all possible corresponding voxels in the predicted dose distribution within defined tolerance criteria and the smallest gamma value is sought, according to Eq. (5) and Fig. 5:

$$\Gamma(\vec{r}_m, \vec{r}_c) = \min \sqrt{\left(\frac{|\vec{r}_m - \vec{r}_c|}{\Delta d}\right)^2 + \left(\frac{|D_m(\vec{r}_m) - D_c(\vec{r}_c)|}{\Delta D}\right)^2} \tag{5}$$

where $\vec{r}_m$ and $\vec{r}_c$ are the spatial location of r_m and r_c voxels in the measured (predicted) and calculated (reference) dose distributions, respectively; $D_m(\vec{r}_m)$ and $D_c(\vec{r}_c)$ are the dose values at voxels r_m and r_c, respectively; Δd is the distance-to-agreement criterion and ΔD is the dose difference criterion.

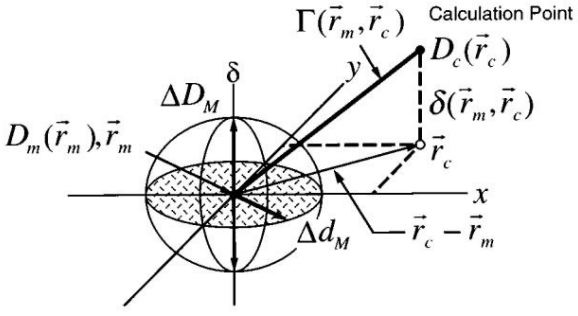

Fig. 5 Geometric representation of gamma index calculation. Gamma index combines DD and DTA criteria, first published by Low et al. [74]. The index is calculated per voxel in the reference dose distribution and compared to the predicted dose distribution.

The "min" operation selects the smallest gamma value over all comparison voxels in the reference dose grid, which represents the best match between the reference and predicted distributions at that particular voxel [74,75].

A gamma index less than or equal to 1 ($\Gamma \leq 1$) indicates that the prediction is within the acceptable range of accuracy defined by some criteria. The calculated gamma indices are then mapped to a colour scale, with red indicating regions in which the model failed to accurately predict the dose ($\Gamma > 1$) and blue highlighting regions in which the model passed ($\Gamma \leq 1$). These gamma maps serve as a visual representation of the agreement between the reference and predicted dose distirbutions. Gamma analysis was performed for each fraction of the patient that was excluded from training, the gamma pass rates of which were averaged, and tabulated with the standard deviation, median, as well as the minimum and maximum pass rates attained.

As recommended by AAPM Task Group 119, gamma analysis at a tolerance of 3% and 3mm of the prescribed dose is commonly performed in clinical practice as a baseline for IMRT QA, striking a balance between stringent requirements and practical acceptance for quality assurance [76]. This criterion specifies that, for a voxel in the predicted dose distribution to pass, there must exist a voxel in the reference distribution such that the weighted combination of dose difference and spatial distance satisfies the gamma condition. For instance, if the closest reference voxel lies 3mm away, then the allowable dose difference must be 0% to satisfy $\Gamma \leq 1$. If instead the dose difference reaches the 3% tolerance, then the voxel in the predicted distribution must coincide exactly with the reference voxel (i.e., spatial difference is 0mm). A mean gamma pass rate of at least 95% is widely accepted as the threshold for clinical applicability, in which 95% of the voxels in the predicted dose distribution lies within 3%/3mm of the dose in the reference distribution [76].

A 2%/2mm or a 5%/5mm criteria may be justified in certain scenarios, e.g., when higher precision is required for critical structures or when a more lenient threshold is appropriate due to increased anatomical variation or uncertainty. By comparing mean gamma pass rates at 2%/2mm or 5%/5mm tolerances with a

baseline 3%/3mm criterion, the sensitivity and robustness of dose prediction accuracy was assessed. This comparison across increasingly lenient criteria provides insights into the extent and clinical relevance of the discrepancies between the predicted and true dose distributions. Further, stratified gamma analyses by dose ranges enabled meaningful evaluation of model accuracy in different clinically relevant dose regions. Voxels were grouped based on the reference dose into four ranges: >90%, 70–90%, 50–70%, and 20–50% of the true maximum prescribed dose. Stratified analysis was employed in place of anatomic- or structure-specific evaluation, as contours were not used in this study. This choice reflects the goal of predicting dose distribution with a simple model trained solely on routinely available data, without segmentation.

Although AAPM Task Group 218 recommends a 3%/2mm criterion for IMRT and VMAT patient-specific QA, this study primarily uses a 3%/3mm criterion for baseline evaluation of model performance in dose prediction [76,77]. This choice aligns with other dose prediction studies and also reduces penalties for non-model-related uncertainties, particularly in CBCT-based dose calculation [78,79]. Additionally, any voxel receiving less than 20% of the maximum true dose was excluded from analysis to improve computational efficiency. It should be noted that both AAPM Task Group 119 and 218 recommend a 10% dose threshold [76,77]. Yet, a 20% dose threshold was applied in this study to ensure that low-dose regions, encompassing large volumes with minimal dose gradients, do not disproportionately influence the overall gamma pass rate, allowing the evaluation to focus on clinically relevant, high-dose areas that have a larger impact on tumour control and patient outcomes.

Chapter 3 Results and Discussion

3.1 Hyperparameter Optimisation Outcomes

Optimization of network depth, number of convolutional kernels per depth, and the number of epochs was performed based on the accuracy of the predicted dose distribution, as quantified by the gamma pass rate. By systematically varying the values of these hyperparameters and selecting those that yielded the highest gamma pass rate at a 3%/3mm tolerance, an optimal U-Net configuration, in terms of complexity, efficiency, and accuracy, was achieved. The choice of loss function was determined by the rate of convergence to the lowest loss value, indicative of training efficiency. Parameters that were not optimised in this study, namely batch size (=5) and optimiser choice (Adam), were set to commonly accepted default values and those established in previous work for all optimisation experiments and evaluation of optimised model performance on the complete dataset [67,71].

3.1.1 Network Depth and Number of Convolutional Kernels

To determine the optimal U-Net architecture, models with depths 3, 4, and 5 were tested, each with three different initial kernel counts: 16, 32, and 64. The number of convolutional kernels is doubled with each subsequent depth in the encoder path. As an example, for a U-Net model of depth 4 starting with 32 kernels, the successive levels have 32, 64, 128, then 256 kernels.

Each model configuration was initially trained on 10 fractions from two randomly selected IMRT patients (5 fractions from each) and then validated on 5 fractions from a different patient using gamma analysis at 3%/3mm. Based on previous work, a total of 150 epochs and mean squared error (MSE) loss function were used for training [67]. These parameters were later validated in Section 3.1.2. Along with gamma pass rate, supporting metrics including mean loss at the 150^{th} epoch, total training time across 150 epochs, and model size, are shown in Table 3. Notably, the mean loss at the 150^{th} refers to the average training loss value across all batches at the final epoch.

Table 3 Comparing performance of U-Net model configurations of varying depths and initial kernel counts over 150 epochs based on mean ± SD gamma pass rate, mean loss per epoch, training time for 150 epochs, and model size

Depth	Initial Kernel Count	Gamma Pass Rate (%)		Mean Loss at 150^{th} Epoch ($\times 10^{-4}$)	Training Time (min)	Model Size (GB)
		Mean	±SD			
3	16	90.4	4.0	8.93	2.48	6.85
	32	98.9	0.8	4.85	4.15	13.60
	64	98.9	0.6	7.12	7.05	27.12
4	16	96.5	2.2	7.19	3.45	7.07
	32	98.6	1.0	5.88	5.40	14.06
	64	91.5	0.5	5.40	8.78	28.07
5	16	94.0	2.1	9.80	4.13	7.14
	32	92.7	0.7	7.83	6.85	14.24
	64	96.2	1.0	1.13	9.60	28.57

The mean losses and gamma pass rates were evaluated in tandem and compared across configurations, with the former reflecting how effectively the model has learned the mapping from input to

output, and the latter serving as a clinically relevant metric for prediction accuracy. Training time provides a quantitative measure of computational efficiency of the model (i.e., total time for 150 epochs) and was considered for practicality, while model size describes memory needed for network parameters (weights and biases) and was assessed to ensure the model remains deployable on available hardware without excessive resource usage. Model size is distinct from checkpoint file size, which saves a training run and includes model parameters, optimiser state (momentum and variance estimate for each parameter), training metadata (epoch number, loss), as well as framework overhead. The checkpoint file is used to resume training and is usually 2-3 times larger than model size. Here, model size is used for simple comparison. The depth and number of kernels were considered optimised for the model that achieved the lowest loss and highest gamma pass rate, indicative of effective learning and generalisable prediction accuracy.

From the results in Table 3, the depth 3, 32-kernel model demonstrated the most favourable performance, with a mean gamma pass rate of $98.9 \pm 0.8\%$, and low mean loss per epoch of 4.85×10^{-5}. Although the depth 3, 64-kernel configuration achieved the same mean pass rate, its higher training time and larger model size means that the additional number of convolutions increased computational burden without improvement in accuracy. In comparison, depth 4 models showed slightly lower performance overall, with the best configuration (starting with 32 kernels) reaching a gamma pass rate of $98.6 \pm 1.0\%$ and slightly higher mean loss of 5.88×10^{-5}. The depth 4, 64-kernel model exhibited an even lower gamma pass rate of $91.5 \pm 0.5\%$, suggesting possible overfitting or training instability despite its higher capacity.

Model performance further declined for models of depth 5, with the 64-kernel configuration attaining the highest mean loss of 1.13×10^{-4}. In fact, several shallower and less complex models were able to reach similar or even higher pass rates than that of this configuration ($96.2 \pm 1.0\%$) with shorter training times. These findings imply that increasing the initial number of kernels beyond 32 at any given depth leads to comparable or worse performance in spite of longer training times. As such, the depth 3, 32-kernel model appears to offer the most favourable trade-off between accuracy and computational efficiency.

However, this preliminary test was conducted on a limited subset of patients (trained on 10, validated on 5 fractions) and the models may not generalise in the same way under a LOOCV framework, where interpatient variability can influence outcomes. Accordingly, a single instance of LOOCV was performed (training on 12, validating on 1 excluded patient datasets) for three model configurations: depth 3 with 32 initial kernels, and depth 4 with both 16 and 32 initial kernels (Table 4). Notably, LOOCV was not performed with the shallowest depth 3, 16-kernel model or any of the depth 5 models due to underperformance on a small subset and consistently poor results, respectively. To add, models starting with 64 kernels, regardless of depth, were not considered further due to excessive memory demands relative to their marginal or absent performance improvements.

Table 4 Effect of depth and initial kernel count on prediction accuracy across 150 epochs, quantified by gamma pass rate at 3%/3mm for a single instance of LOOCV

Depth	Initial Kernel Count	Gamma Pass Rate (%)				Mean Loss at 150th Epoch ($\times 10^{-4}$)	Training Time (min)
		Mean	±SD	Max.	Min.		
3	32	99.0	0.5	99.7	97.4	7.62	114.43
4	16	99.5	0.3	99.9	98.8	6.32	143.12
	32	99.6	0.2	99.9	98.8	4.47	176.12

Among all three tested configurations, the depth 3, 32-kernel model showed the lowest mean gamma pass rate ($99.0 \pm 0.5\%$), while depth 4, 16- and 32-kernel models attained comparably high pass rates of $99.5 \pm 0.3\%$ and $99.6 \pm 0.2\%$. Yet, the 32-kernel configuration achieved a perhaps lower mean loss per epoch (4.47×10^{-5} vs. 6.32×10^{-5}) and standard deviation (0.2% vs. 0.3%) than the 16-kernel model at depth 4 and may indicate enhanced training effectiveness and prediction consistency. Also, this configuration may offer more stable performance or generalisability over a larger dataset (i.e., several instances of LOOCV), potentially justifying the longer training duration (176 minutes vs. 143 minutes). Therefore, the U-Net model with a depth of 4 and initial kernel count of 32 was selected for subsequent optimisation experiments and training of the full dataset.

3.1.2 Loss Function and Number of Epochs

For regression tasks, including dose prediction, mean squared error (MSE) and mean absolute error (MAE) (voxel-wise) loss functions are typically used to quantify the difference between the prediction and true continuous values, with higher loss values indicating a poorer fit of the model to the data. The choice of loss function, either (1) MSE alone, (2) a combination of 70% MSE and 30% MAE ($\propto$=0.7), (3) a combination with equal weighting ($\propto$=0.5), or (4) MAE alone, was selected based on the convergence rate and prediction accuracy of U-Net model (depth 4, initial kernel count of 32) trained for 250 epochs.

To assess convergence, the model was trained using data from all fractions of a single randomly selected IMRT patient with each of the four loss functions. The mean loss per epoch computed by these loss functions were recorded and used to construct loss curves (Table 5, Fig. 6). A comparison of these loss curves showed that the MSE and MSE-weighted combined ($\propto$=0.7) configurations reached convergence more quickly within the first 75 epochs compared to the MAE and equally-weighted combined ($\propto$=0.5) configurations.

Table 5 Comparison of loss function type on convergence efficiency across 250 epochs

Loss Function	MSE*	Combined ($\propto$=0.7)	Combined ($\propto$=0.5)	MAE
No. of Epochs	Mean Loss ($\times 10^{-4}$)			
50	1.16	11.5	16.3	72.3
75	0.98	5.62	10.1	34.1
100	0.95	5.28	8.78	18.8
125	0.95	5.41	7.91	16.0
150	0.95	5.38	7.86	15.7
175	0.96	5.37	7.83	15.7
200	0.95	5.31	7.80	15.5
225	0.95	5.33	7.79	15.5
250	0.95	5.31	7.80	15.5

*Note: By scaling dose values to range [0, 1], squaring small errors reduces the magnitude of MSE values compared to MAE values, which reflect absolute differences.

Since MSE amplifies large errors with its squared term, the loss function is highly sensitive to outliers and may subsequently become disproportionately biased [68]. Although MAE penalises all errors proportionately, the resulting gradient remains constant, leading to slower convergence and increased sensitivity to local noise [68]. In Fig. 6, MAE exhibits high initial loss, frequent fluctuations throughout training, and noisier (or unstable) loss curve. In fact, the loss only starts to converge after 125 epochs, suggesting that MAE alone may not be well suited for efficient or stable training in this context. The combined loss function, with a more heavily weighted MSE component ($\propto$=0.7), attained a lower final loss value (after 250 epochs) and reached convergence approximately 50 epochs earlier than its equally weighted ($\propto$=0.5) counterpart.

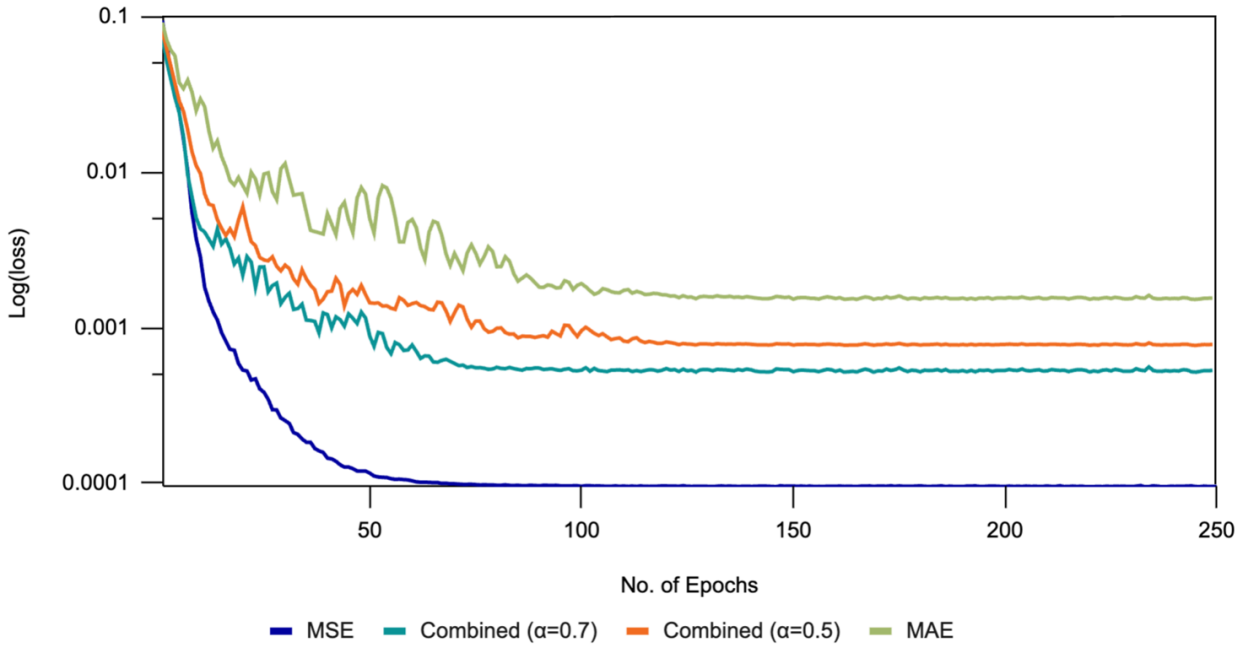

Fig. 6 Effect of loss function choice on training convergence. Comparison of four different loss functions for model training over 250 epochs using all fractions of a single IMRT patient.

To confirm that the supposed convergence efficiency of MSE translates into clinically meaningful prediction accuracy, a single instance of LOOCV with the IMRT dataset was performed in increments of 25, from 50 to 250 epochs (Table 6). In other words, the model was trained with 12 patient datasets using

the four different loss function and for varying number of epochs and validated with the remaining patient dataset (all fractions) at 3%/3mm tolerance. Notice that this test also served to establish an appropriate training duration (i.e., the number of epochs with which over- and under-fitting can be mitigated) for a larger dataset than that used to assess convergence. While all four loss functions achieved similarly high pass rates (for this particular instance of LOOCV), the MSE function underperformed in terms of gamma pass rate across all epochs compared to MAE alone, which reached 99.4% within the first 50 epochs with minimal variance through to epoch 250. The combined loss with $\propto$=0.5 yielded similarly consistent results across the 250 epochs. Still, the $\propto$=0.7 combined function managed to converge faster while maintaining consistently high performance, balancing robustness to dose outliers of MAE and stable gradient behaviour of MSE. This 70% MSE and 30% MAE ratio ensures that the model does not overfit to low-dose noise by disproportionately prioritise minimising small errors and instead, splits its attention between both small and large errors in high-dose, clinically significant regions, enabling more generalisable and reliable predictions.

Table 6 Comparison of loss function type on prediction accuracy across 250 epochs, quantified by gamma pass rate at 3%/3mm for a single instance of LOOCV

Loss Function	MSE	Combined ($\propto$= 0.7)	Combined ($\propto$= 0.5)	MAE
No. of Epochs	Mean ($\pm$SD) Gamma Pass Rate (%)			
50	97.2 ± 0.4	98.7 ± 0.2	98.6 ± 0.6	99.4 ± 0.4
75	97.5 ± 0.5	99.0 ± 0.8	99.4 ± 0.7	99.6 ± 0.3
100	98.4 ± 0.3	99.1 ± 0.7	99.4 ± 0.4	99.3 ± 0.5
125	97.9 ± 0.4	99.6 ± 0.4	99.4 ± 0.5	99.6 ± 0.3
150	98.7 ± 0.3	99.8 ± 0.2	99.6 ± 0.3	99.8 ± 0.2
175	98.1 ± 0.5	99.6 ± 0.4	99.5 ± 0.5	99.7 ± 0.3
200	96.2 ± 0.8	99.6 ± 0.4	99.6 ± 0.3	99.7 ± 0.2
225	98.8 ± 0.4	99.7 ± 0.3	99.4 ± 0.4	99.7 ± 0.2
250	98.7 ± 0.4	99.7 ± 0.2	99.7 ± 0.3	99.7 ± 0.2

To continue, the number of epochs describes the number of times the entire training dataset is passed through the network (forward and back) [69]. With more training iterations, the model is expected

to continually update its weights and minimise errors, such that more accurate output predictions are produced. The combined loss function with $\propto$=0.7 achieved peak prediction accuracy at epoch 150, with marginal changes in performance after this point. From a practical perspective, 150 epochs is sufficient for model convergence and represents a reliable and efficient training duration.

3.2 Performance Outcomes and Dose Prediction Accuracy

Following hyperparameter optimisation, dose prediction was performed using a U-Net model of depth 4 and initial kernel count of 32, trained for 150 epochs with a batch size of 5, a combined loss function weighted toward MSE ($\propto$=0.7), and the Adam optimiser. The model was evaluated in three stages using leave-one-out-cross-validation (LOOCV) across:

1) IMRT patients (n=13),
2) VMAT patients (n=11), and
3) a combined IMRT and VMAT dataset (n=24).

The accuracy of the model was quantified by gamma analysis with a baseline 3%/3mm criterion and a 20% dose threshold for the entire 3D volume, capturing global performance. For each model, the mean, standard deviation (SD), median, maximum, and minimum gamma pass rate values, across all treatment fractions for the validation patient, are tabulated along with an "overall" column that summarises these statistics for across all validation patients. Gamma analysis with either 2%/2mm or 5%/5mm tolerances were also used to estimate the sensitivity of gamma pass rates and assess the robustness of the predictions. Additionally, gamma analyses stratified by dose regions of >90%, 70–90%, 50–70%, and 20–50%, were performed to identify whether the IMRT-, VMAT-, and combined-trained models were more likely to fail in specific dose ranges.

3.2.1 Evaluation of U-Net Model for IMRT Dose Distribution Prediction

Across the 13 IMRT patients evaluated using LOOCV, the U-Net model demonstrated consistently high gamma pass rates, with an overall mean pass rate of 98.1% at 3%/3mm (Table 7). The median pass rate of 99.0% and relatively low standard deviation of 1.4% means that predictions were both accurate and stable between patients. It should be noted that the model number denotes the patient excluded from training and used for validation in each LOOCV iteration. The number is arbitrary and intended for ease of reference.

Subsequent application of a stricter 2%/2mm criterion led to a drop in the overall gamma pass rate, from 98.1% to 94.7%, as expected given the increased sensitivity to spatial and dose discrepancies. Nonetheless, most models achieved gamma pass rates above the desired 95% threshold for clinical applicability and continued to perform robustly under a 2%/2mm tolerance, particularly models 2, 3, 5, 6, 7, 9, 10, 11, and 12 (Table 8).

Table 7 Gamma pass rates for LOOCV at 3%/3mm across 13 IMRT patients

	Gamma Pass Rate (%)						
Model #	1	2	3	4	5	6	7
Mean	97.8	99.8	98.9	94.5	98.5	99.1	98.9
Std. Dev.	0.7	0.2	0.3	5.5	0.7	0.3	0.8
Median	97.8	99.8	99.0	97.3	98.5	99.3	99.0
Max.	99.2	99.9	99.3	99.8	99.7	99.5	99.6
Min.	96.7	99.6	97.9	83.0	97.1	98.6	96.0
Model #	8	9	10	11	12	13	Overall
Mean	98.7	98.8	98.8	99.9	99.6	92.2	98.1
Std. Dev.	1.7	0.9	0.6	0.2	0.3	2.1	1.4
Median	99.3	99.3	98.8	100.0	99.6	92.1	99.0
Max.	99.9	99.9	99.7	100.0	99.9	98.2	100.00
Min.	93.0	97.2	96.9	99.2	99.1	88.0	83.0

At both 3%/3mm and 2%/2mm, models 2 and 11 achieved the highest gamma pass rate, while models 4 and 13 showed the lowest mean pass rates and largest standard deviation. This suggests that specific anatomical and/or dosimetric features of the validation patient may have differed more noticeably from

those in the training dataset and greater variability between fractions exists for these patients. Moreover, the low minimum pass rates for models 4 and 13 were not outliers, with several fractions attaining similar pass rates. Visual inspection of CT/CBCT and dose grid images of these cases also revealed no abnormalities, artefacts, or preprocessing mismatches, meaning that there may be more subtle features with which the model struggles to generalise.

Table 8 Gamma pass rates for LOOCV at 2%/2mm across 13 IMRT patients

	Gamma Pass Rate (%)						
Model #	1	2	3	4	5	6	7
Mean	93.4	97.5	97.5	87.5	95.8	96.4	96.3
Std. Dev.	1.8	0.8	1.0	9.8	0.9	1.5	2.2
Median	93.1	97.5	97.7	91.1	96.0	96.5	96.5
Max.	97.4	99.1	98.6	98.6	97.9	98.3	98.5
Min.	90.4	96.1	93.4	69.7	94.0	93.3	87.5
Model #	8	9	10	11	12	13	Overall
Mean	93.4	95.3	96.9	99.4	98.0	84.1	94.7
Std. Dev.	4.9	2.7	1.6	0.5	1.1	3.1	2.4
Median	95.0	95.6	96.7	99.6	98.1	84.8	96.5
Max.	98.4	99.1	99.2	99.9	99.5	92.2	99.9
Min.	80.5	90.6	92.3	97.9	95.0	76.1	69.7

To qualitatively assess prediction accuracy, gamma index maps of an axial slice (at approximately the centre of the target) were generated for representative models 11, 4, and 13 that achieved the highest, second lowest, and lowest mean pass rate (Fig. 7). For illustrative purposes, fraction 10 was randomly selected for images and predictions (note that the standard deviation in pass rates between fractions were relatively low). These maps show the spatial distribution of passing and failing regions at both 3%/3mm and 2%/2mm criteria. The drop in gamma pass rates under the more stringent 2%/2mm tolerance is clearly depicted by larger red regions, attributable to a greater number of voxels in the prediction failing to meet the criterion.

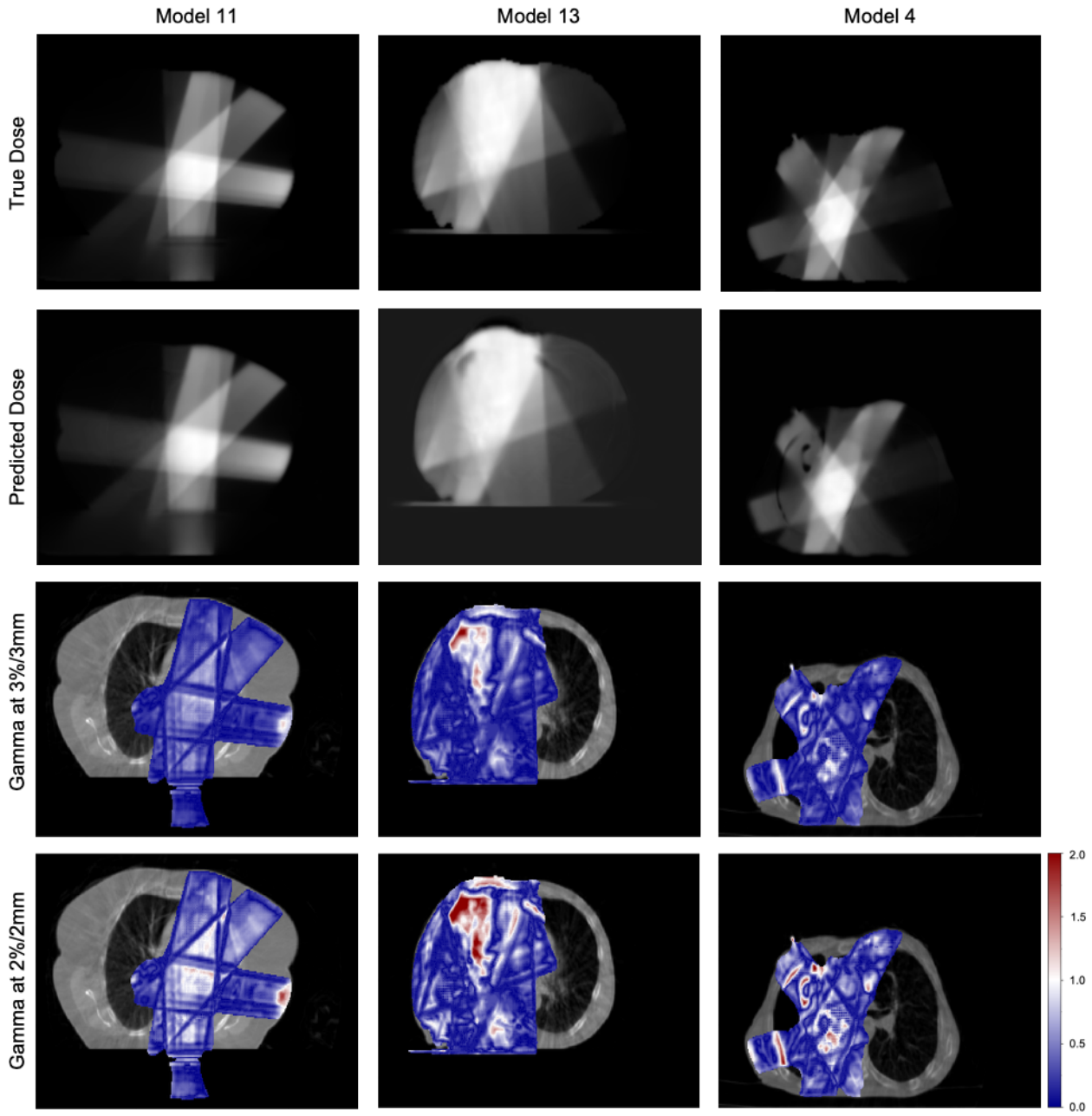

Fig. 7 Comparison of prediction accuracy and gamma index maps from IMRT-trained models with varying performance: models 11 (highest pass rate), 13 (lowest pass rate), and 4 (second lowest pass rate). Gamma maps are overlaid on top of the associated daily CBCT, with red regions representing areas where $\Gamma > 1$ and the predicted dose exceeds the specified tolerance.

Voxels along beam edges generally exhibited higher agreement, as evidenced in Fig. 7 by deeper blue regions. These edges are geometric field boundaries where dose drops from the treatment field to essentially zero outside the field, typically in areas <10–20% of the prescription dose. Although beam edges fall into transition zones between beam "on" and "off" (i.e., high-gradient regions), their behaviour and patterns within the training data may be more consistent between patients and easier to predict. The pass rates for these regions may also be higher due to the spatial tolerance of gamma analysis. In a steep gradient, dose changes rapidly over small distances, such that small spatial shifts in the predicted dose distribution can still meet the criteria if a nearby voxel, within the DTA radius, satisfies the dose tolerance. Hence, the overall gamma pass rate can arise from the proximity of nearby voxels meeting the dose criteria nearby, rather than complete dose agreement at the individual voxel [75,80].

In addition, the dose gradient at beam edges stems from the finite size of the source, collimation, MLC design, and scatter, creating similar penumbra shape and dose fall-off pattern for comparable beam energies and geometries [81]. Conversely, plan complexity and patient-specific anatomy in high-dose regions, where many beamlets overlap for conformal dose delivery, make predictions more sensitive to beam modulation variations and tissue heterogeneity, which are less systematic than at the beam edges [14]. More simply, dose gradients at beam edges are driven by physical beam constraints, whereas beam modulation near the target rely on patient-specific factors, compromising prediction accuracy.

Furthermore, it seems that gamma failures within the treated volume were not entirely random and appear to correlate with heterogeneous regions where there are variations in tissue density (e.g., tissue-to-air boundary). For instance, a closer inspection of plan CT and CBCT (Fig. 8) for the validation patients of model 13 and 4 reveals that the strong chest wall features within the predicted dose distributions (columns 2 and 3 in Fig. 8, respectively) are a consequence of anatomical changes between planning and treatment. In the CBCT acquired at fraction 10, the target near the heart has noticeably shrunk and shows more well-defined boundaries compared to the plan CT of the validation patient for model 13. The change in size and

appearance of the target following treatment response likely reduced prediction accuracy (mean±SD=92.2 ± 2.1%), as other patients in the training dataset either experienced less pronounced interfractional changes or exhibited different changes. As another example, the validation patient for model 4 had experienced more severe lung collapse between planning and delivery (indicated by green arrows of different sizes), such that the chest wall boundary from the plan CT no longer matches the CBCT geometry at fraction 10. This patient also seems to have metal objects, perhaps wires connected to a pacemaker (visible as bright white spots on CT, circle in green on dose prediction), which were not present for others in the training dataset. The slight differences in their locations may have been difficult to convert into a dose distribution for the model, resulting in poor prediction accuracy with large interfractional variability (mean±SD =94.5 ± 5.5%).

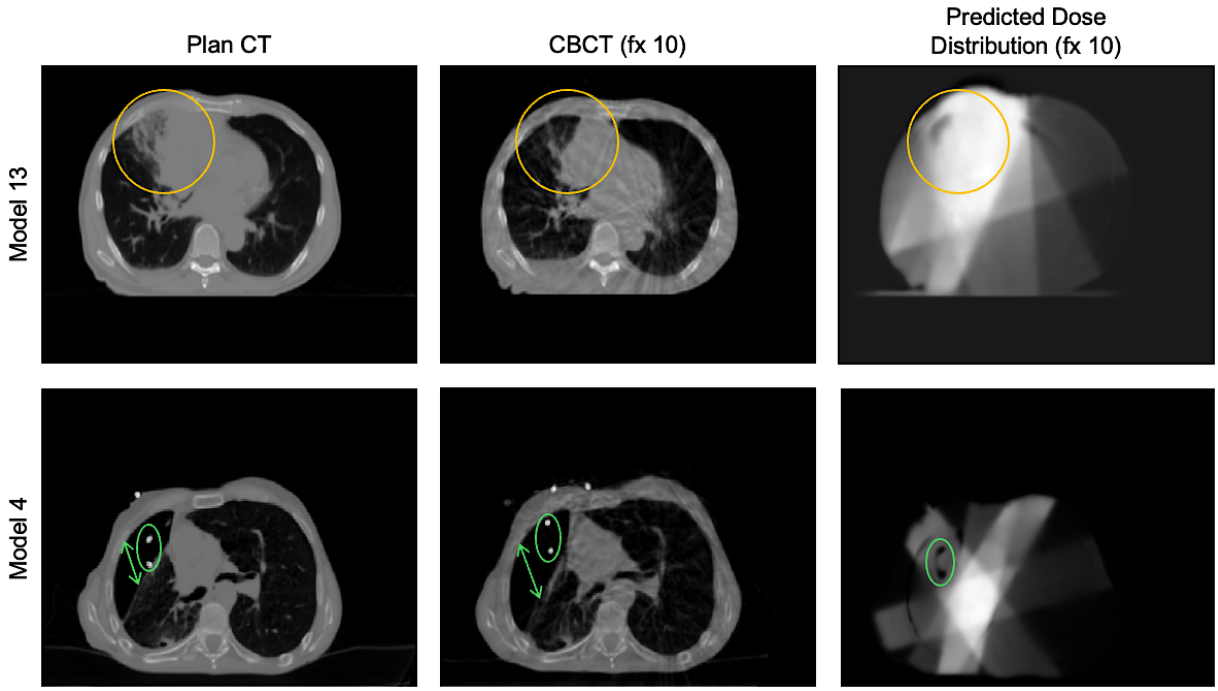

Fig. 8 Gamma failures tend to be a result of anatomical changes between plan (plan CT) and delivery (CBCT) and differences between training dataset and validation patient. Validation patients for model 13 and 4 experienced pronounced anatomical changes and had metal objects, respectively, which resulted in poorer prediction accuracy.

Due to the lack of structural and contour information, it was difficult to ascertain any correlation between prediction failures and critical structures or target volumes. Instead, a stratified gamma analysis, as opposed to global (Table 7), can provide dose region-specific insight by enabling focused evaluation of gamma pass rates within clinically relevant high-, mid-, and low-dose areas within the treatment volume. More explicitly, the dose regions correspond to voxels receiving >90%, 70–90%, 50–70%, and 20–50% of the maximum true dose in the ground truth dose distribution. The stratified results in Table 9 were obtained from gamma evaluation at 3%/3mm of each IMRT validation patient, the average of which is depicted by a heatmap (Fig. 9).

Table 9 Gamma pass rates (3%/3mm) stratified by dose region for IMRT validation patients

	Mean Gamma Pass Rate (%)						
Model #	1	2	3	4	5	6	7
Dose Region							
>90%	96.2 ± 2.0	99.1 ± 1.0	98.5 ± 1.1	94.4 ± 5.0	99.6 ± 0.5	94.7 ± 3.0	100.0 ± 0.1
70–90%	98.2 ± 1.3	98.1 ± 2.3	99.8 ± 0.3	89.1 ± 9.6	99.5 ± 0.4	98.7 ± 0.8	97.8 ± 1.0
50–70%	99.8 ± 0.2	99.8 ± 0.3	99.9 ± 0.1	92.9 ± 9.1	99.6 ± 0.3	99.8 ± 0.2	99.8 ± 0.2
20–50%	98.3 ± 0.2	99.9 ± 0.1	98.8 ± 0.3	96.0 ± 3.6	98.0 ± 1.0	99.3 ± 0.4	99.7 ± 0.3
Model #	8	9	10	11	12	13	Overall
Dose Region							
>90%	96.4 ± 4.0	98.2 ± 2.2	99.4 ± 0.7	99.3 ± 0.8	98.4 ± 1.3	90.2 ± 6.2	97.3 ± 1.9
70–90%	96.6 ± 4.7	96.5 ± 2.3	93.0 ± 4.7	99.7 ± 0.9	100.0 ± 0.0	87.7 ± 8.9	96.5 ± 3.2
50–70%	97.3 ± 3.6	99.5 ± 1.2	99.2 ± 0.8	99.9 ± 0.4	99.9 ± 0.1	95.5 ± 4.7	98.7 ± 2.7
20–50%	98.7 ± 1.2	99.5 ± 0.5	99.4 ± 0.3	99.9 ± 0.1	99.9 ± 0.1	98.7 ± 1.2	98.9 ± 0.9

From Table 9, the overall mean gamma pass rate in the high-dose region was $97.3 \pm 1.9\%$, with model 7 achieving $100.0 \pm 0.1\%$ and model 5 at $99.6 \pm 0.2\%$. Across all 13 models, model 13 achieved the lowest pass rate of $90.2 \pm 6.2\%$ in this >90% region (Fig. 9, row 1). To continue, gamma pass rates were generally highest in the 50–70% (Fig. 9, row 3) and 20–50% (Fig. 9, row 4) regions, with overall mean rates of $98.7 \pm 2.7\%$ and $98.9 \pm 0.9\%$, respectively. All models exceeded 95% accuracy in the 20–50% range, including model 13, which reached a pass rate of $98.7 \pm 1.2\%$ despite its poor global and high-dose area performances. This likely reflects the larger and more consistent representation of low-dose regions in the

training data, allowing the model to generalise better in these areas compared those near the target. In contrast, high-dose regions are often more patient-specific, involving complex beam arrangements and heterogeneous anatomy, the variability of which hinders generalisation efforts.

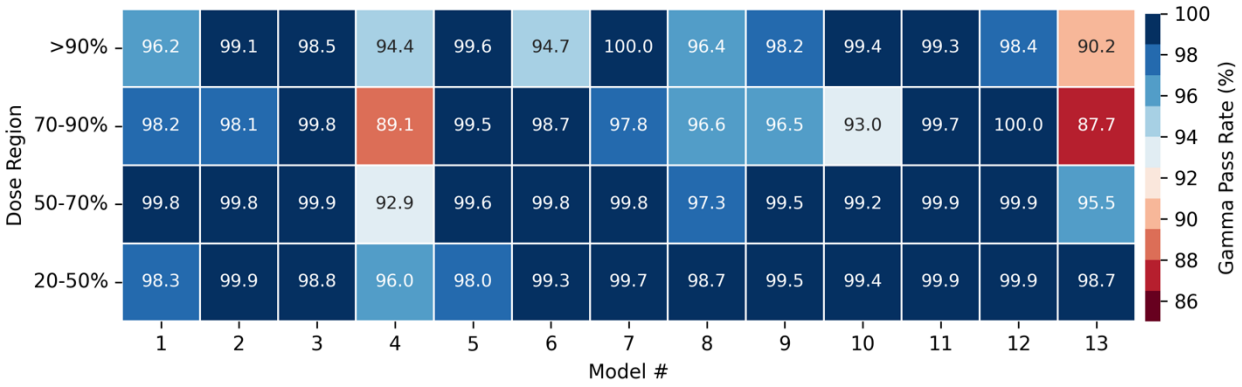

Fig. 9 Gamma pass rates (3%/3mm) for 13 IMRT models stratified by dose region. The heatmap displays gamma pass rates averaged over all fractions of each validation patient across four dose regions.

Models 4 and 13, which yielded the lowest global mean pass rates at 3%/3mm, evidently struggled in the 70–90% dose region, attaining low mean pass rates with large standard deviations of 89.1 ± 9.6 and 87.7 ± 8.9, respectively. In fact, the 70–90% dose region had the lowest overall mean pass rate ($96.5 \pm 3.2\%$) among the four stratified regions (Fig. 9, row 2), which may be attributable to the presence of steeper dose gradients, where dose falls off rapidly to spare nearby healthy tissues while maintaining adequate tumour coverage. This effect is depicted by left-right (LR) and anterior-posterior (AP) dose profiles, near the edge of the PTV (i.e., brightest region on dose image), of the worst-performing fractions from validation patients of models 4 (Fig. 10) and 13 (Fig. 11), with mean pass rates of $89.1 \pm 9.6\%$ and $87.7 \pm 8.9\%$ within the 70–90% region, respectively. Fig. 10 confirms that the most pronounced mismatch between the predicted distribution by model 4 and true distribution occurred in the steep gradient, 70–90% region, as shown by both the rapid curve divergence in the LR dose profile and the concentrated gamma failures in the corresponding axial slice. The AP dose profile at this particular vertical cross-section reveals considerable

underdosing below 30% and overdosing between 70–90% dose ranges. Recall that the dose threshold for gamma analysis was set to 20%, preventing disproportional influence of low-dose voxels occupying a larger volume. Hence, both the global and stratified gamma results overlook discrepancies in regions falling below the 20% threshold. The predicted dose distribution appears to falsely suggest increased toxicity risk to healthy tissues and concern of inadequate control of microscopic disease, thereby limiting model reliability in such regions. To elaborate, the overestimation of dose in 70–90% region by the model means that it was unable to accurately reproduce steep dose gradients (at least in the anterior aspect for this cross-section).

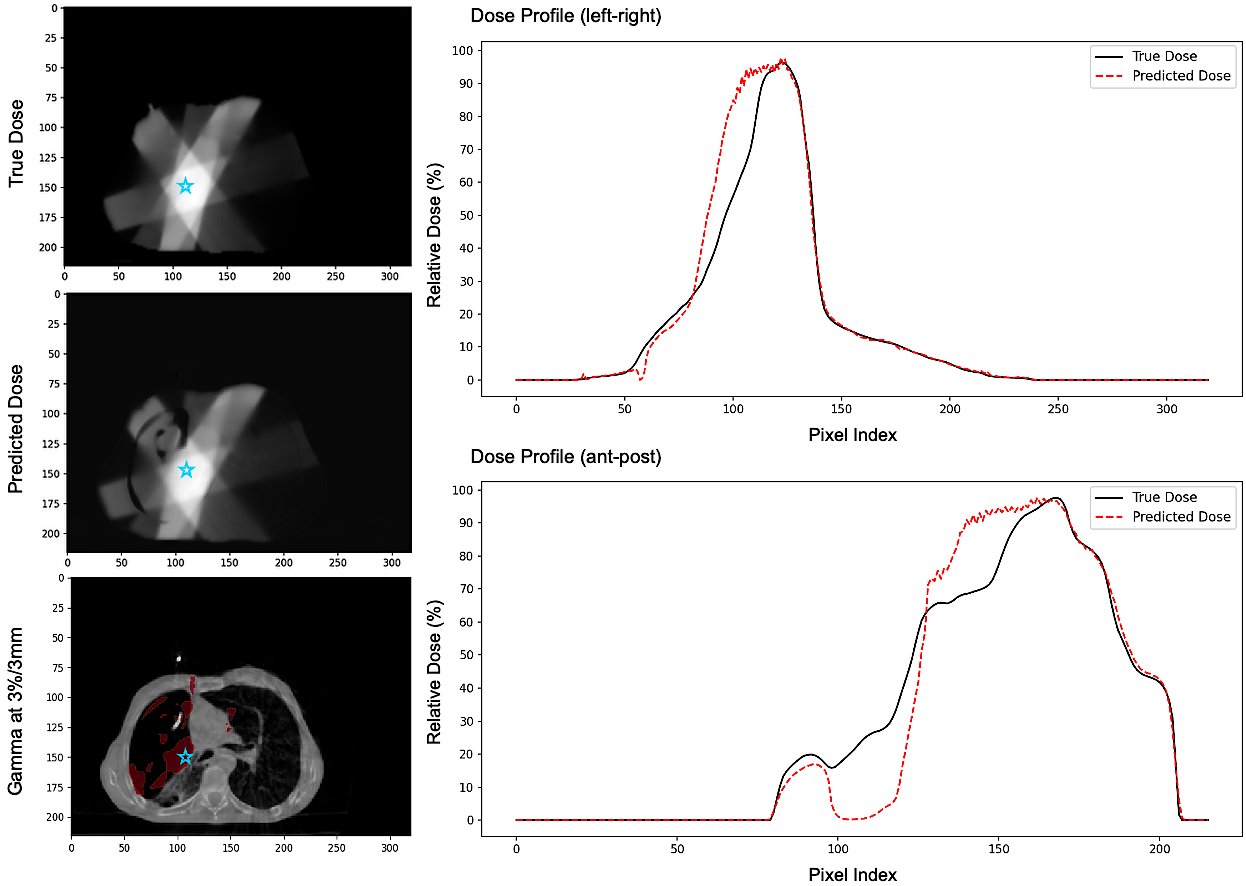

Fig. 10 Dose prediction accuracy of IMRT model 4. Axial slices of true dose predicted dose, and corresponding global 3%/3mm gamma failure map within PTV, as indicated by the blue star, are shown along with LR and AP dose profiles (True dose: —, Predicted dose: ---). Note: Each pixel represents 1.17 mm.

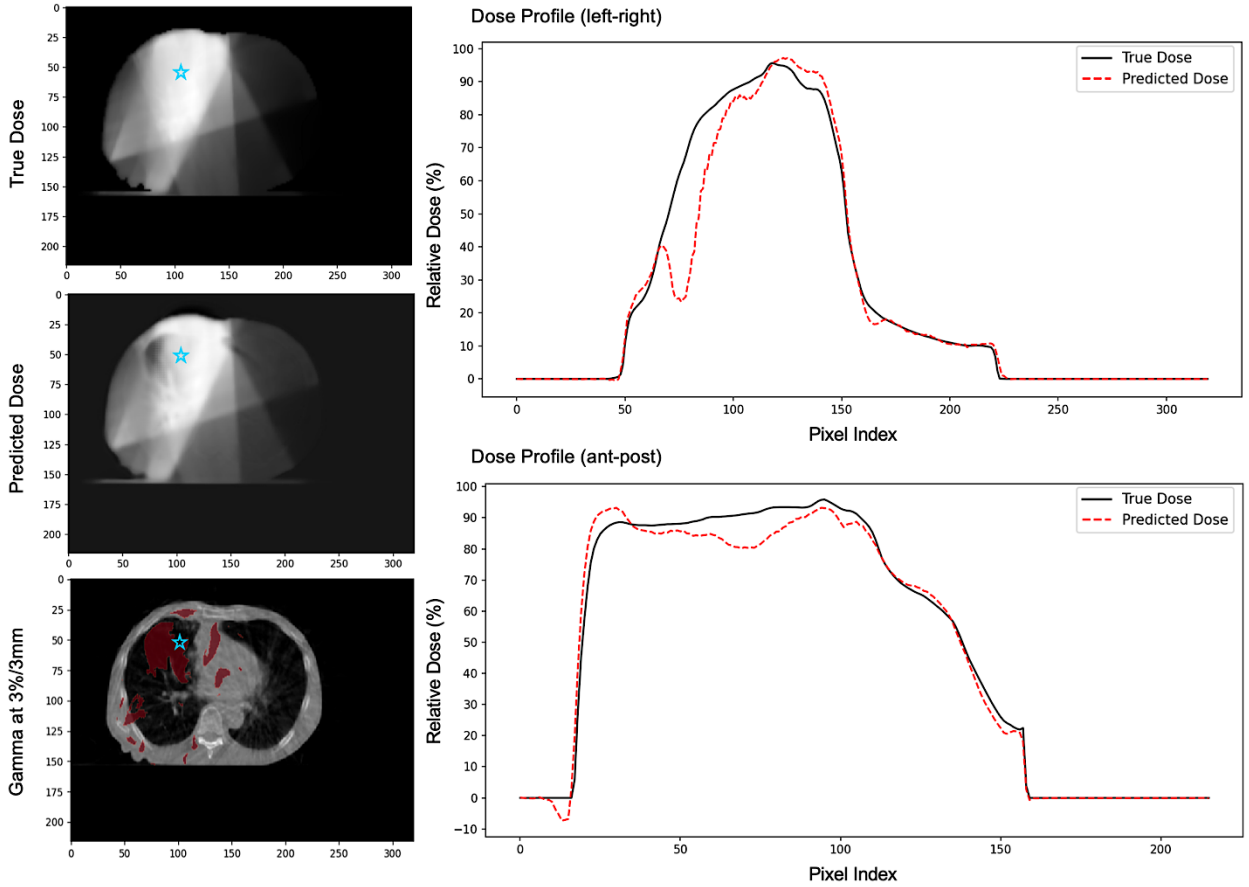

Fig. 11 Dose prediction accuracy of IMRT model 13. Axial slices of true dose predicted dose, and corresponding global 3%/3mm gamma failure map within PTV, as indicated by the blue star, are shown along with LR and AP dose profiles (True dose: —, Predicted dose: ---). Note: Each pixel represents 1.17 mm.

Though steep dose gradients can yield higher gamma pass rates due to the spatial tolerance of the gamma metric, they are also sensitive to positioning errors, meaning predictions that lay beyond the DTA search radius will cause a sharp decline in gamma pass rates [74]. This intermediate dose area typically coincides with the edges of high-dose PTV or areas in which the PTV overlaps with lower-dose normal tissues or OARs, wherein small spatial mismatches between the predicted and true distributions can result in relatively large gamma failures. Given the dose heterogeneity in these boundary regions and interplay between dose conformality and sparing objectives, accurate prediction becomes increasingly challenging.

This is reinforced by the misalignment of true and predicted dose curves, wherein model 13 underestimated the dose (Fig. 11).

Overall, predicted distributions produced by these models tend to have better agreement with the true dose in low 20–50% and 50–70% dose regions, and displayed minor declines in agreement beyond the intermediate 70–90% region. This suggests that errors in predictions are not uniformly distributed but rather concentrated in clinically, high-dose areas, and may correspond with edges of the target volume, proximity to critical structures, steeper dose gradients, and/or areas of increased dose inhomogeneity and tissue heterogeneity.

While IMRT and VMAT are both widely used in clinical practice for NSCLC radiotherapy, VMAT has gained traction in recent years for its superior target conformity, shorter treatment times, and improved delivery efficiency. Yet, many centres continue to use IMRT for specific patient cases and based on resource availability. Also, VMAT plans can take longer to optimise and result in variations in plan quality due to time constraints [82]. To support clinical decision-making across various institutions, a unified and flexible dose prediction model that performs reliably for both IMRT and VMAT cases would be beneficial. To that end, models 11 and 13 (with the highest and lowest mean pass rates, respectively) were validated using all fractions from a single VMAT patient at 3%/3mm (Table 10).

Table 10 Validation of IMRT-trained models (11 and 13) on all fractions of a single VMAT patient

	Gamma Pass Rate (%)	
Model #	11	13
Mean	91.1	90.6
Std. Dev.	8.9	8.5
Median	92.6	91.8
Max.	99.8	99.0
Min.	66.1	66.8

Interestingly, both models yielded comparable mean gamma pass rates of $91.1 \pm 8.9\%$ and $90.6 \pm 8.5\%$, respectively. However, the high standard deviations (8.9% and 8.5% vs. overall SD=1.4% for IMRT models)

and wide pass rate range (66.1% to 99.8%), indicate considerable variability between fractions. This finding suggests that, while IMRT-trained models can produce accurate predictions for some VMAT fractions, their performance can drop markedly for others, with certain fraction falling well below clinically acceptable pass rates. It also implies that models trained on IMRT may generalise well to VMAT, though the high and consistent predictive accuracy for IMRT-trained models does not always translate to VMAT. Hence, models trained exclusively on VMAT data may be needed to learn technique-specific, dose delivery features, such as dynamic beam modulation and more heterogeneous dose distributions, and to achieve improved predictions.

To sum, IMRT-trained models produced consistently accurate dose predictions, with a global mean gamma pass rate of $98.1 \pm 1.4\%$ at 3%/3mm, which is above the accepted clinical threshold of 95% [76]. Stratified gamma analysis confirmed strong performance of all 13 IMRT models across all dose regions, though the 70–90% dose region exhibited the greatest variability, signifying reduce capability of models in predicting areas with sharp dose transitions. Validation of representative IMRT-trained models on a single, random VMAT case demonstrated inadequate degree of generalisability between both techniques, suggesting the need for modality-specific models to maintain prediction accuracy.

3.2.2 Evaluation of U-Net Model for VMAT Dose Distribution Prediction

Across the 11 VMAT patients validated by LOOCV, the overall average percentage of voxels passing the 3%/3mm tolerance criterion was lower than that observed for the IMRT cohort (mean=90.1% vs. 98.1%). A Mann-Whitney U test revealed a statistically significant difference in the central tendency of mean pass rates between the IMRT and VMAT groups ($U = 1, n_1 = 13, n_2 = 11,\ p < 0.05$). The Hodges-Lehmann estimate of the median difference was 8.7 percentage points, indicating that pass rates achieved by IMRT models were about 8.7% higher than those of VMAT models. However, the 95% confidence interval of this effect size estimate (-0.6 to 16.4) includes zero, reflecting uncertainty in the true magnitude of the difference,

which likely stems from the variability and limited dataset size. Still, a Brown-Forsythe test showed no statistically significant difference in the variability of mean pass rates between IMRT and VMAT models ($p = 0.13$). Hence, the observed difference in central tendency is more likely due to modality-related performance differences rather than an artefact of uneven spread in patient data. An additional visual inspection of CT/CBCT and dose grid images verified no preprocessing mismatches or artefacts, meaning the minimum pass rates for each model were not outliers. These results suggest that IMRT-trained models tended to outperform VMAT models in dose prediction accuracy, though further data may be required to precisely quantify the extent of the difference. Notably, the model architecture and training hyperparameters for these VMAT-specific models were initially optimised with IMRT data. Consequently, IMRT-optimised configurations may not yet be adequate in accommodating the complexities associated with VMAT, and dedicated hyperparameter tuning may be of benefit.

Table 11 Gamma pass rates for LOOCV at 3%/3mm across 11 VMAT patients

	Gamma Pass Rate (%)					
Model #	1	2	3	4	5	6
Mean	91.0	95.5	82.6	90.9	87.5	88.6
Std. Dev.	8.9	2.4	6.9	7.5	2.6	4.9
Median	92.8	96.0	84.4	93.6	87.1	89.9
Max.	99.0	98.9	93.3	99.3	93.3	94.1
Min.	64.3	89.0	68.9	73.0	81.9	69.5
Model #	7	8	9	10	11	Overall
Mean	94.6	89.1	88.6	89.0	93.8	90.1
Std. Dev.	5.2	4.7	10.2	8.9	4.1	2.6
Median	96.3	90.1	92.1	91.9	94.7	92.1
Max.	98.1	93.6	99.6	99.7	98.3	99.7
Min.	76.4	72.8	63.4	70.6	82.1	63.4

The accuracy of predictions generated by VMAT models fluctuated more between treatment fractions within the same validation patient than IMRT models and may be another cause of the relatively poorer performance, secondary to the lack of VMAT-specific optimisation. The gamma maps in Fig. 12 depict this interfractional variability.

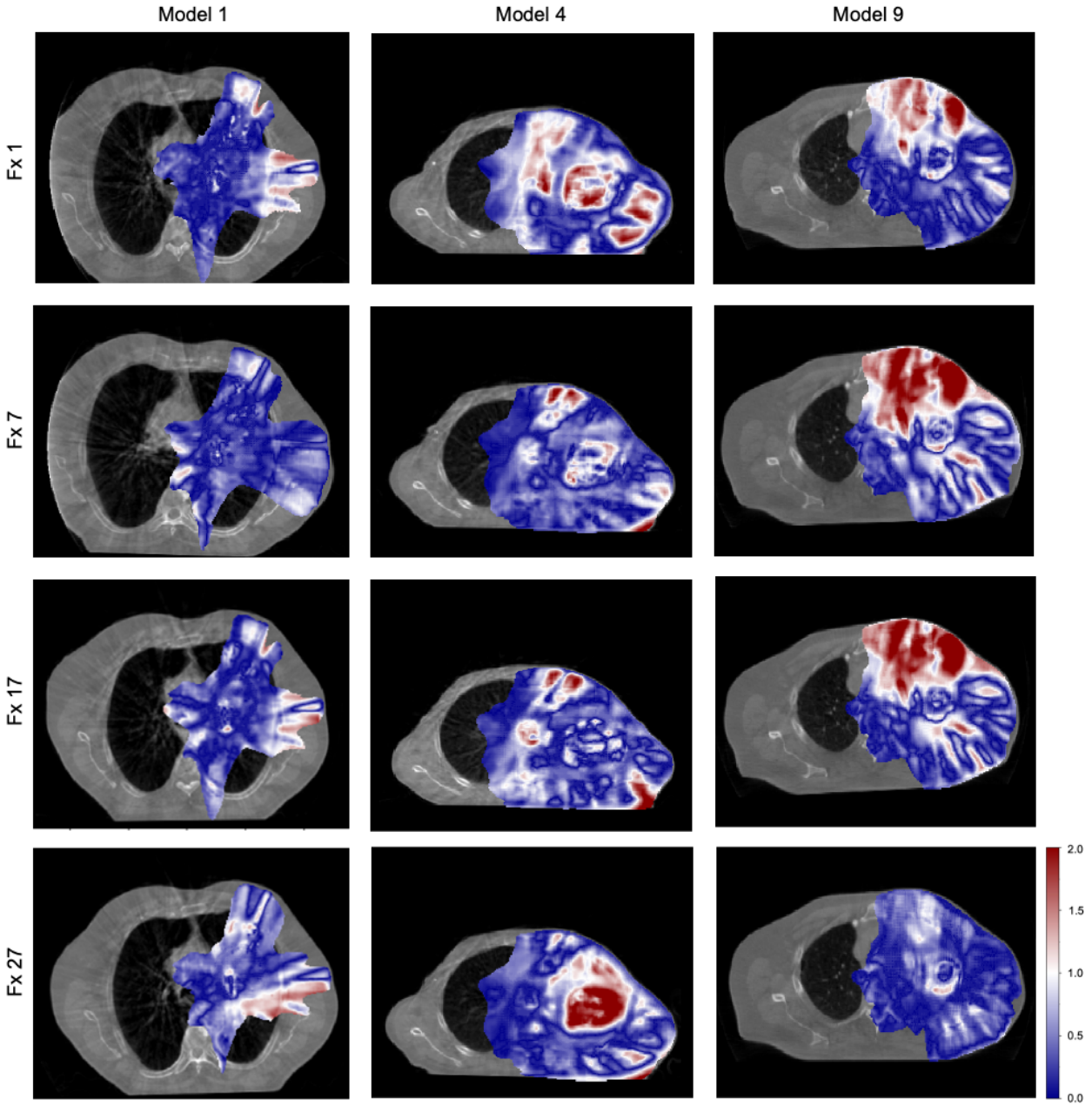

Fig. 12 Selected gamma index maps (3%/3mm) showing interfractional variability in VMAT dose distribution prediction accuracy. Columns correspond to a different patient validated by models 1, 4, and 9 (SD=8.9%, 7.5%, and 10.2%, respectively). Rows show an axial slice (near centre of target) for treatment fractions 1, 7, 17, and 27.

In particular, VMAT models 1, 4, and 9 yielded respective standard deviations of 8.9%, 7.5%, and 10.2%, each exceeding the highest standard deviation among IMRT models, which was 5.5% in IMRT model 4. Within each column, the extent and location of failures (in red) across different fractions (of the same slice) appear to vary randomly, with no discernible temporal pattern, meaning that prediction errors may be driven by subtle, patient-specific changes rather than systematic changes over time. The variability in performance may be linked to the inherent complexity of dose distributions of VMAT delivery, wherein dose is delivered in a continuous arc with dynamic modulation of beam intensity and shape compared to the more discrete and static beams of IMRT [17]. To elaborate, the continuous change in gantry rotation spreads low doses over a larger volume, creating more intricate dose maps with fine variations and subtle gradations, with which the network may struggle to reproduce. In contrast, IMRT uses fewer fixed beam angles that result in more localised and less complex dose distributions. In other words, there are more variables over which the model must generalise with VMAT, leading to poorer prediction capability and lower pass rates. Although dose predictions did not meet the 95% pass rate threshold at 3%/3mm and thus not clinically acceptable, these VMAT-trained models hold promise for daily dose prediction with the overall pass rate increasing from $90.1 \pm 2.6\%$ (at 3%/3mm) to $97.8 \pm 1.4\%$ at 5%/5mm (Table 12).

Table 12 Gamma pass rates for LOOCV at 5%/5mm across 11 VMAT patients

	Gamma Pass Rate (%)					
Model #	1	2	3	4	5	6
Mean	98.9	99.4	94.2	98.0	96.1	98.4
Std. Dev.	1.9	0.5	5.2	2.9	0.9	2.1
Median	99.7	99.6	95.8	99.3	96.0	99.2
Max.	100.0	100.0	98.5	100.0	98.8	99.8
Min.	92.5	98.6	73.0	86.5	94.4	89.4
Model #	7	8	9	10	11	Overall
Mean	99.3	96.4	97.6	98.0	99.6	97.8
Std. Dev.	1.2	1.5	4.2	3.0	0.6	1.4
Median	99.7	96.6	99.8	99.7	99.8	99.6
Max.	99.9	98.4	100.0	100.0	99.9	100.0
Min.	95.1	91.0	83.6	89.5	97.6	73.0

By loosening the tolerance, the 5%/5mm criterion quantifies the magnitude of the deviation. Gamma maps (3%/3mm and 5%/5mm) of an axial slice near target centre were generated for models 2, 8, and 3 (Fig. 13).

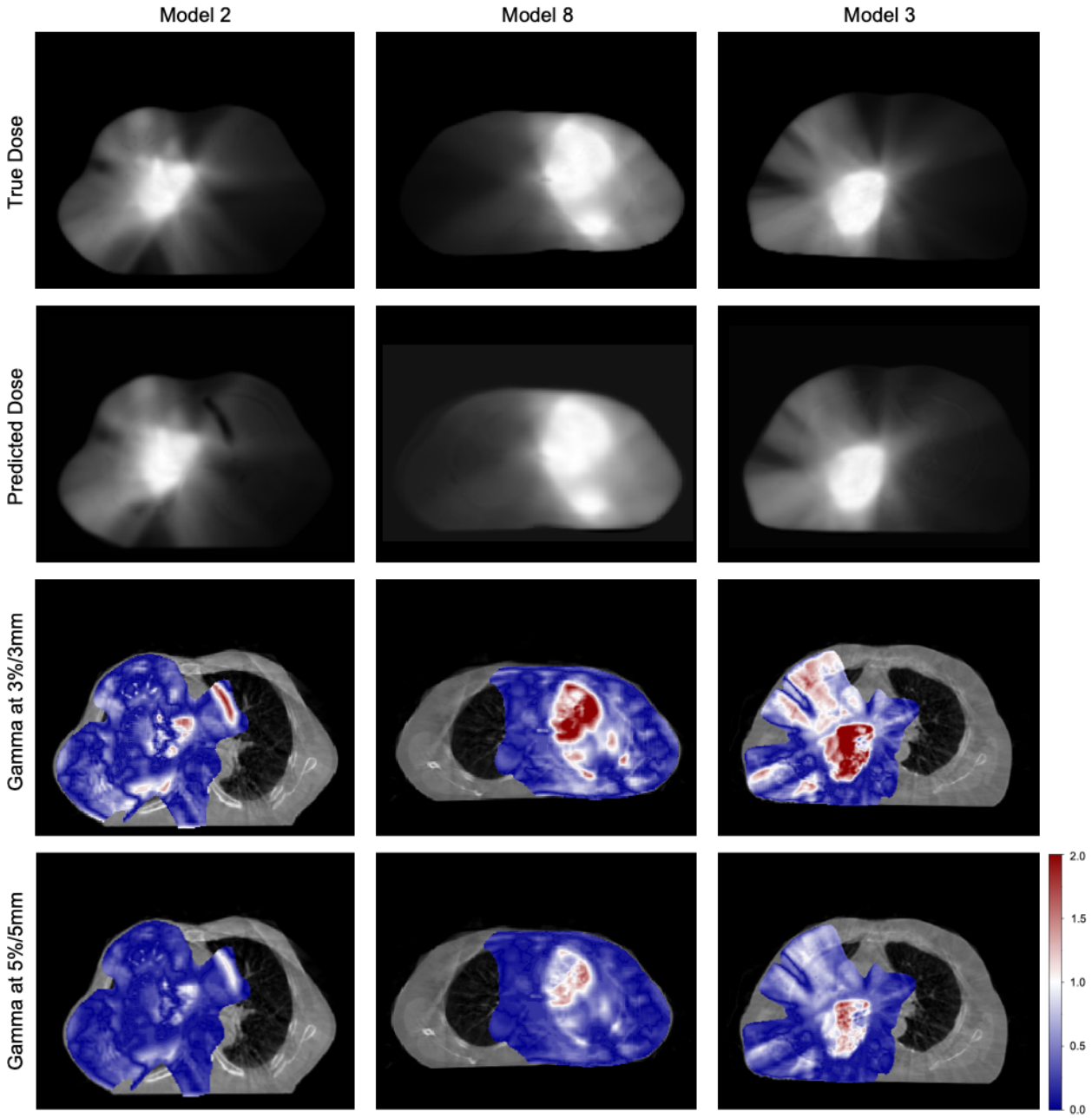

Fig. 13 Gamma index maps depicting dose prediction accuracy for VMAT-trained models: models 2 (highest pass rate), 8 (median pass rate), and 3 (lowest pass rate) for validation, respectively. Dose distributions and gamma (3%/3mm and 5%/5mm) maps correspond to fraction 23.

The observed improvement in mean pass rate implies that the degree of discrepancy between the true and predicted dose distributions were moderate, with predictions often just exceeding the stricter 3%/3mm tolerance. The gamma maps in both Fig. 12 and Fig. 13 show distinct differences in performance, as well as in the extent and spatial distribution of gamma failures across various models.

Given the continuously rotating delivery of VMAT, the high agreement along fixed beam edges observed in IMRT gamma maps (Fig. 7) is inherently absent. Instead, it appears that gamma failures correlated with clinically high-dose regions (within the PTV) or in areas where there were anatomical differences between planning (plan CT) and delivery (CBCT). For example, the validation patients for models 4 and 9 show varying degrees of gamma failures across fractions (refer to Fig. 12), though their location remained roughly the same. Based on the plan CTs, the GTVs (outlined in red) of both patients appear to be attached superiorly and laterally to the chest wall (Fig. 14, 15).

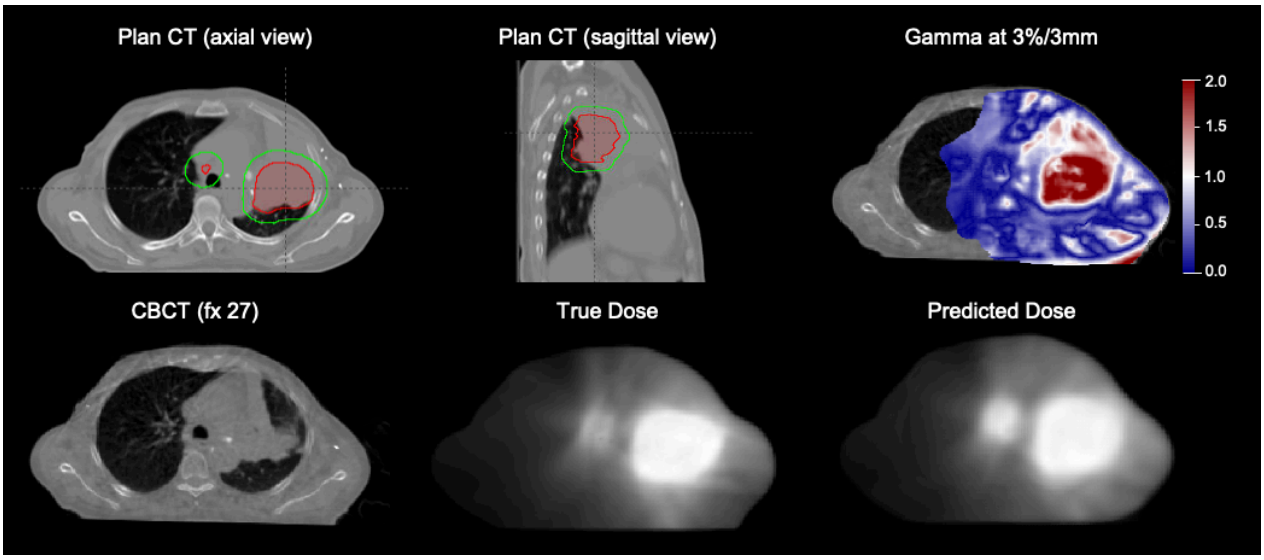

Fig. 14 Gamma failures of VMAT model 4. Validation patient presenting with GTV (outlined in red, encompassed by PTV in green) attached superiorly and laterally to the chest wall led to more failures where beam enters and/or exits the body. Agreement of true and predicted dose distributions at fraction 27 are shown by the gamma map.

The proximity of the PTV to the ribs, ipsilateral lung, and heart, as well as the thin tissue margins involved, requires more complex modulation to achieve target coverage while sparing nearby OARs [83]. This increased modulation complexity, in conjunction with PTV changes, may be the reason for gamma failures in the anterior aspect of the patient, and regions where the beam enters/exits.

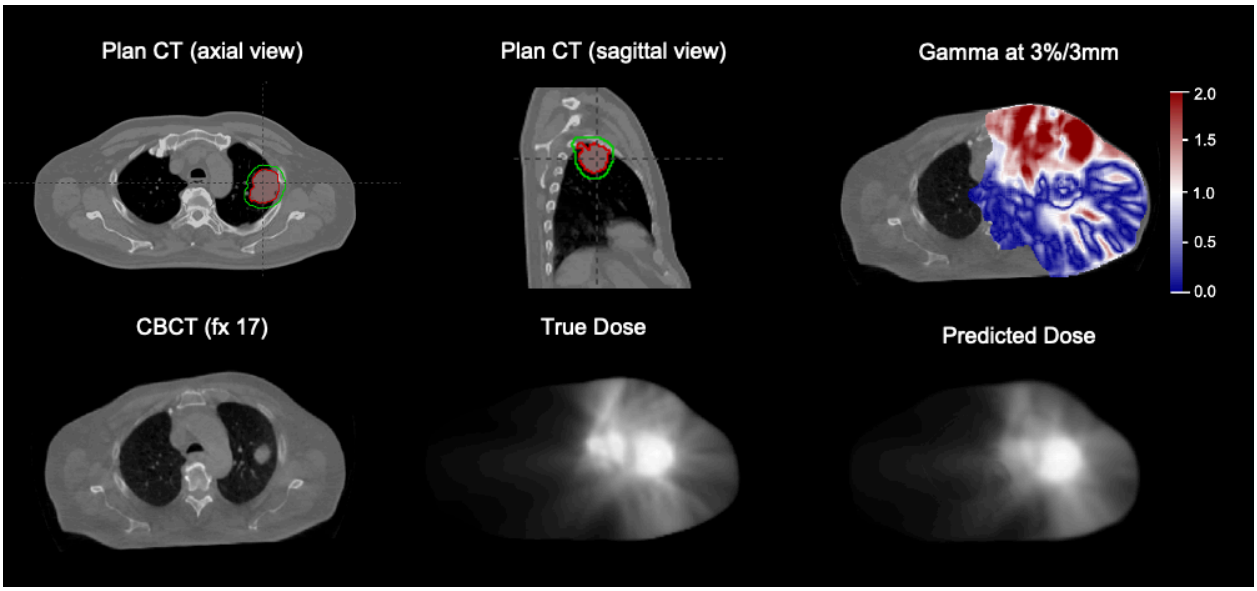

Fig. 15 Gamma failures of VMAT model 9. Validation patient presenting with GTV (outlined in red, encompassed by PTV in green) attached superiorly to chest wall led to more failures where beam enters the body anteriorly. Agreement of true and predicted dose distributions at fraction 17 are shown by the gamma map.

Gamma failures also tended to reflect changes near or within the PTV, where steep dose gradients exist for conformal delivery. In the validation case for model 2 (column 2 in Fig. 16), the patient initially presented with a posterior pleural effusion in the right lung on the plan CT (indicated by the pink arrow), which is a common symptom of lung cancer [84,85]. In the CBCT acquired at fraction 23, this effusion appears to have been resolved or treated. While this resolution is positive for the patient, it introduced a challenge for the model, which seems to predict dose assuming the effusion is still present.

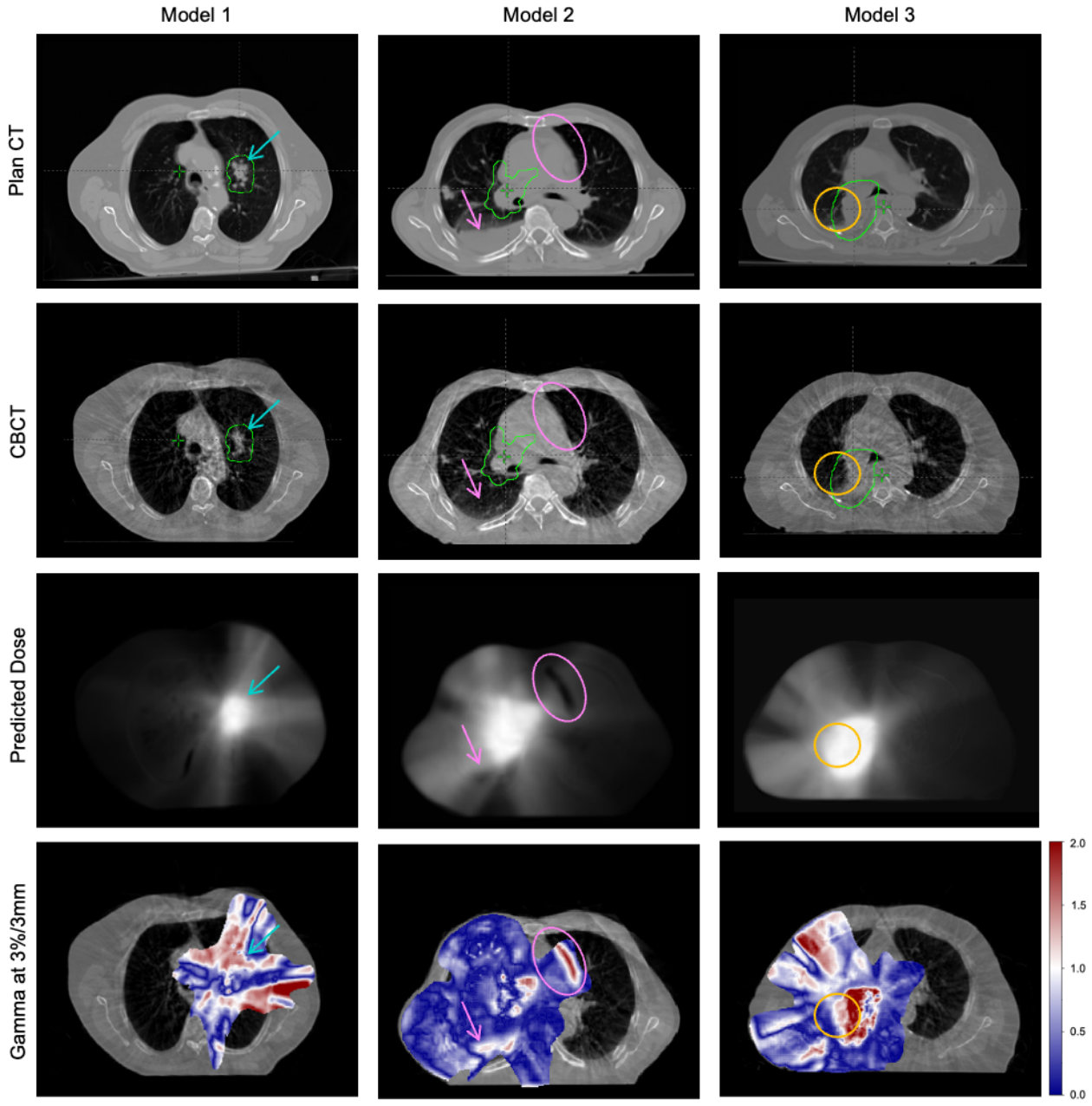

Fig. 16 Trends in gamma failures of VMAT-trained models 1 (second largest SD), 2 (highest mean pass rate), and 3 (lowest mean pass rate). Gamma failures appear to correlate with the PTV (outlined in green in plan CT and CBCT), as well as anatomical changes between planning and delivery. Validation patients of models 1 (cyan arrow) and 3 (orange circle) both show evidence of target shrinkage, whereas model 2 is challenged by presence of a pleural effusion (pink arrow) and heart displacement (pink oval).

In more detail, the apparent replacement of pleural fluid (~1.0 g/cm^3) by air-filled lung tissue (~0.3 g/cm^3) caused a substantial change in tissue density, such that the planned beam attenuation was much lower during delivery, and the actual delivered dose is higher than predicted [86,87]. No other patient in the training set had pleural effusion. Minor heart displacement is also observed in the same CBCT, potentially due to respiratory and/or cardiac motion [88]. The underprediction of dose by the model is depicted as black regions (indicated by pink arrow and circle) and reflected by gamma failures in the posterior region. Notice that the resolution of pleural effusion and heart displacement occurred away from the PTV, resulting in localised gamma failures confined only to those affected regions, without significant degradation of the overall prediction accuracy.

Conversely, the validation patients for models 1 and 3 (columns 1 and 3 in Fig. 16, respectively) experienced target (or GTV) shrinkage within the PTV, leading to widespread gamma failures that extended beyond the PTV itself. This may be associated with the heightened sensitivity of VMAT delivery to PTV changes. Given that VMAT delivers dose through a continuously modulated arc to create sharp gradients and tight conformity around the PTV, changes in the GTV relative to the plan CT, even when contained within the PTV margin, can cause the precisely optimised beam paths to be misaligned with the actual high-dose region and target volume, rendering model predictions inaccurate [89]. To be specific, the model assumes the same beam geometry and MLC modulation as for the planning CT, such that mismatches between the fixed planned beam arrangement and the altered anatomy in the CBCT complicate accurate dose prediction. In cases of shrinkage or shifts within the PTV, the model currently fails to consider that the planned beam arrangement and precise dose shaping optimised on the planning CT may no longer be adequate in target coverage or OAR sparing. Consequently, the predicted dose distribution deviates from the true (delivered) distribution, with larger relative dose differences detected by gamma analysis and errors propagating to adjacent critical structures due to the modulation complexity required for conformity. These findings suggest that anatomical changes near or within the PTV culminates in more pronounced gamma

failures compared to changes beyond the PTV, which occur in lower-dose, shallower gradient regions and produce subtler changes in dose distribution with lower incidence of gamma failures.

Although the model was trained on cases with changes within the PTV over the course of treatment, the limited number of such cases may not provide sufficient representation for effective generalisation to VMAT beam configurations, which are less standardised than IMRT. Importantly, the severity of under- or over-prediction of dose, quantified by gamma pass rate, did not appear to be directly proportional or correlated to either the extent of anatomical change or the initial target size/shape. Additional data would be needed to ascertain clear trends and to clarify the influence of other factors, such as target location, beam arrangement, proximity to OARs, etc., on predictions, and to guide improvements in prediction accuracy.

Subsequent stratified gamma analysis at 3%/3mm confirmed the reduced model accuracy in predicting dose within steep gradient regions surrounding the target, with the lowest and most variable pass rates (between validation patients) observed in the >90% dose region (mean±SD=$70.8 \pm 6.2\%$), corresponding to target-adjacent areas most affected by PTV changes (Table 13).

Table 13 Gamma pass rates (3%/3mm) stratified by dose region for VMAT validation patients

	Mean±SD Gamma Pass Rate (%)					
Model #	1	2	3	4	5	6
Dose Region						
>90%	85.2 ± 7.9	88.1 ± 8.7	19.1 ± 8.9	85.5 ± 9.8	43.7 ± 9.3	81.2 ± 11.8
70–90%	93.7 ± 9.7	96.6 ± 4.4	80.7 ± 7.2	84.6 ± 12.8	82.1 ± 4.0	88.5 ± 7.0
50–70%	94.1 ± 12.3	99.0 ± 0.8	94.9 ± 3.8	92.6 ± 8.3	94.8 ± 1.6	97.6 ± 2.5
20–50%	90.8 ± 8.8	95.4 ± 2.6	86.6 ± 7.9	92.2 ± 7.6	90.6 ± 2.7	88.1 ± 5.4
Model #	7	8	9	10	11	Overall
Dose Region						
>90%	82.7 ± 8.1	38.7 ± 5.4	86.7 ± 5.6	81.7 ± 27.9	86.5 ± 6.0	70.8 ± 6.2
70–90%	91.1 ± 11.5	88.7 ± 5.3	94.5 ± 7.1	82.5 ± 16.5	97.4 ± 1.6	89.1 ± 4.4
50–70%	96.3 ± 5.1	94.9 ± 5.0	89.4 ± 9.7	90.4 ± 8.3	95.6 ± 2.7	94.5 ± 3.7
20–50%	96.4 ± 4.3	96.4 ± 5.4	88.0 ± 11.3	89.9 ± 7.8	93.7 ± 4.9	91.7 ± 2.7

Even the highest pass rate in the >90% region, achieved by VMAT model 2, was only $88.1 \pm 8.7\%$. For the same dose region, model 3 exhibited the lowest pass rate, with an average $19.1 \pm 8.9\%$ across all validation

fractions. The deviation of predicted from true dose curves within the PTV for models 2 and 3 are depicted by LR and AP dose profiles in Fig. 17 and 18, respectively. To elaborate, the dose profiles in Fig. 17 shows slight dose underestimation in the >90% dose region, with the predicted dose curve situated below the true dose. Still, the model performed well in dose regions below >90%, with visible overlap between the predicted and true curves. This shows that dose underestimation occurred within the PTV, whereas good agreement was observed in the surrounding steep dose gradient regions given the close spatial proximity of voxels capable of satisfying the dose tolerance within the DTA radius.

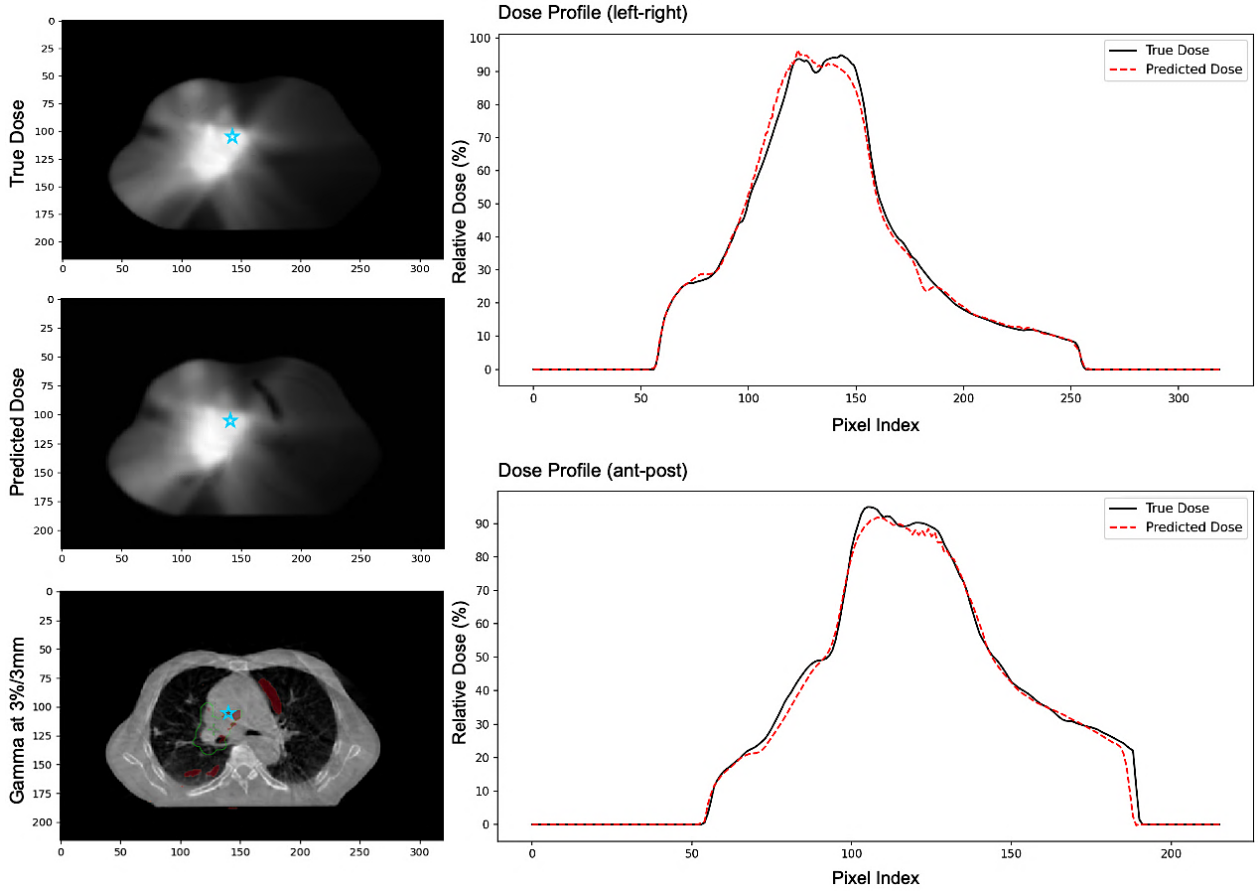

Fig. 17 Dose prediction accuracy of VMAT model 2. Axial slices of true dose distribution, predicted dose distribution, and corresponding global 3%/3mm gamma failure map within the PTV (outlined in green on underlying CBCT), for cross-section indicated by the blue star, are shown along with LR and AP dose profiles (True dose: —, Predicted dose: ---). Note: Each pixel represents 1.17 mm.

Note again that the profiles are sampling specific 1D cross-sections through the 2D axial slices at the reference point marked by the star. Similarly, model 3 underestimated the dose in the high >90% region, coinciding again with the PTV, though the extent of deviation between the predicted and true dose curves is markedly greater than that of model 2, such that very few voxels met the dose criterion within the DTA radius (Fig. 18).

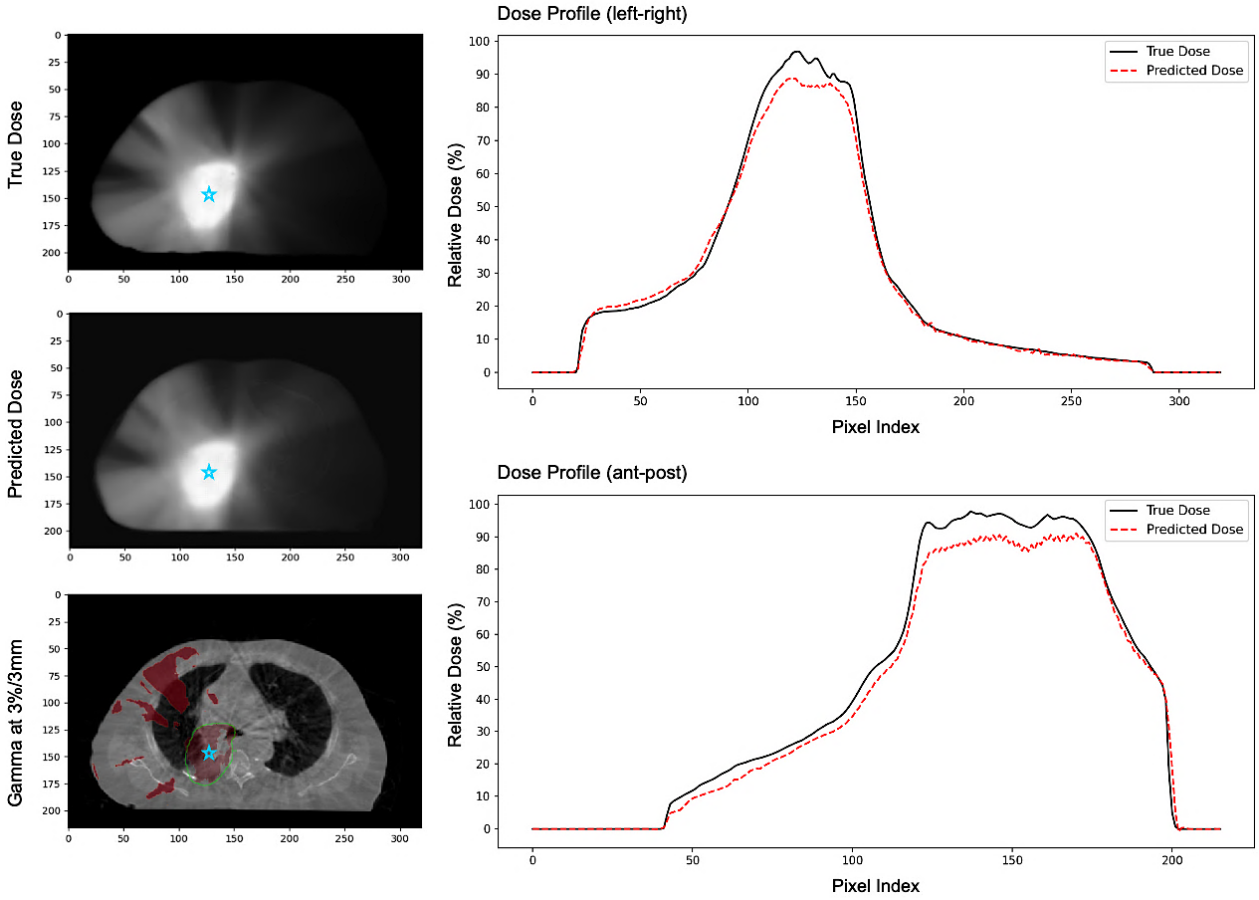

Fig. 18 Dose prediction accuracy of VMAT model 3. Axial slices of true dose distribution, predicted dose distribution, and corresponding global 3%/3mm gamma failure map within the PTV (outlined in green on underlying CBCT), for cross-section indicated by the blue star, are shown along with LR and AP dose profiles (True dose: —, Predicted dose: ---). Note: Each pixel represents 1.17 mm.

The relatively poorer agreement is maintained throughout all dose regions below 90%, consistent with the corresponding gamma failure map. These results show the possible dependence of prediction accuracy on patient-specific features. In fact, overall mean pass rates in the 20–50%, 50–70%, and 70–90% dose regions for these VMAT-trained models were $91.7 \pm 2.7\%$, $94.5 \pm 3.7\%$, and $89.1 \pm 4.4\%$, respectively, indicating comparable performance across these regions. Evidently, these models underperformed in all four dose regions and displayed greater variability in performance compared to the IMRT-trained models (overall pass rates summarised in Table 14). The heatmap in Fig. 19 illustrates the predominant occurrence of gamma failures in the >90% region for VMAT models. In contrast, IMRT models exhibited better and less variable performance in the >90% region, with most failures concentrated instead within the 70–90% region.

Table 14 Comparison of mean gamma pass rates (3%/3mm) stratified by dose region for IMRT vs. VMAT

	Mean±SD Gamma Pass Rate (%)	
Model Type	IMRT	VMAT
Dose Region		
>90%	97.3 ± 1.9	70.8 ± 6.2
70–90%	96.5 ± 3.2	89.1 ± 4.4
50–70%	98.7 ± 2.7	94.5 ± 3.7
20–50%	98.9 ± 0.9	91.7 ± 2.7

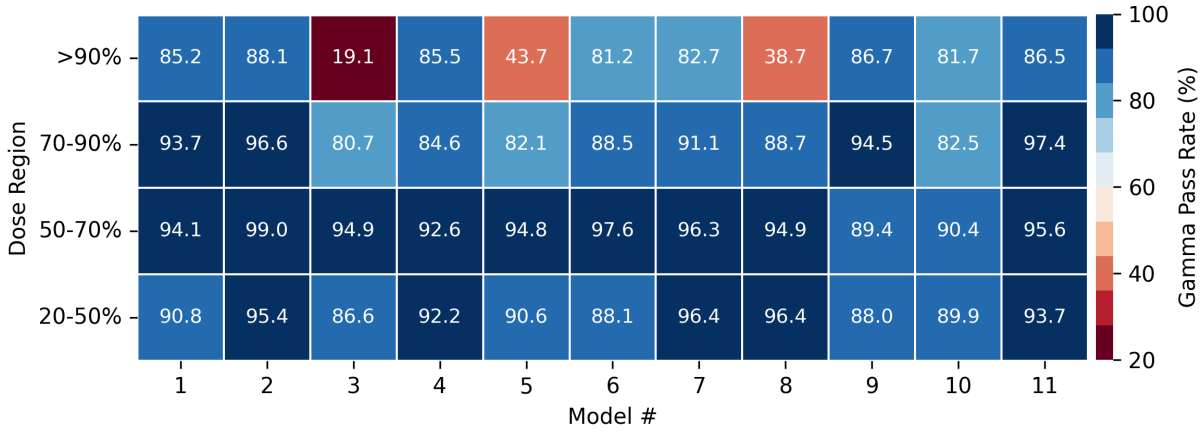

Fig. 19 Gamma pass rates (3%/3mm) for 11 VMAT models stratified by dose region. The heatmap shows gamma pass rates averaged over all fractions of each validation patient across four dose regions. Pass rates are colour-coded from dark red (<20% minimum) to dark blue (100% maximum).

As mentioned earlier, the reduced accuracy and increased variability in performance of VMAT models may be attributed to the highly modulated dose distribution near or within the PTV region as radiation is delivered by a continuously rotating gantry and simultaneous dynamic modulation of the dose rate and MLC positions [17]. On the other hand, the relatively more static and uniform delivery of IMRT, involving a fixed set of gantry angles, generates simpler dose distributions with similar spatial patterns (i.e., "star-shaped") that may be easier for the model to learn and predict. By extension, the increase in degrees of freedom of VMAT, including various beamlet angles and modulation strategies, signifies a wider range of dose distributions (or solutions) that the models must learn to predict. With the limited dataset of 11 patients, the highly modulated and conformal nature of VMAT not only introduces more uncertainty but also reduces the accuracy in generalisability to individual patient geometries, particularly in high-dose, PTV-adjacent regions.

To recap, VMAT-trained models demonstrate potential in dose prediction for VMAT plans, with an overall mean gamma pass rate of $97.8 \pm 1.4\%$ at 5%/5mm, indicating that errors at 3%/3mm tolerance were near acceptable limits rather than a result of large discrepancies. Due to the high modulation and tight conformity of VMAT delivery, changes in or near the PTV (especially those attached to the chest wall) lead to more pronounced gamma failures. It is important to note that the PTV contour was not explicitly included in the model input or training data and was solely used for visualisation. While the IMRT dataset was acquired independently by another student and the PTV could not be displayed in the TPS, similar PTV-related factors may explain the poor performance of IMRT model 13, which also seems to involve a GTV (i.e., brightest spot in the dose image, Fig. 10) attached to the chest wall. By incorporating additional training data and repeating hyperparameter optimisation with VMAT data, these models can be improved, and it may be possible to reach the desired 95% pass rate threshold despite the complex interplay of patient-specific anatomical changes and VMAT-specific delivery dynamics.

3.2.3 Performance of U-Net Trained on Combined IMRT and VMAT Datasets

Although IMRT models demonstrated high dose prediction accuracy (mean±SD=$98.1 \pm 1.4\%$), earlier results revealed limitations with models trained on a single delivery technique, including the inadequate generalisation of IMRT-trained models to VMAT cases, and the inability of VMAT-trained models to attain a 95% pass rate at 3%/3mm. Thus, the integration of both IMRT and VMAT into training may enable the model to capture a wider spectrum of anatomical and delivery scenarios seen in clinical practice. This approach was assessed using LOOCV at 3%/3mm on a combined dataset of 24 patients, with combined-trained models 1–13 each excluding one IMRT patient (Table 15a) and models 14–24 each excluding one VMAT patient (Table 15b). Separating validation this way focuses the analysis on the effect of training on a more diverse dataset.

The models trained on a combined dataset achieved a slightly lower gamma pass rate at 3%/3mm tolerance compared to the IMRT-trained models (mean=97.3% vs. 98.1%) when validated on the IMRT dataset (n=13). A Wilcoxon signed-rank test showed no statistically significant difference in median difference of gamma pass rates ($W = 18, n = 13, p > 0.05$) between these two model types, revealing that training on a more diverse dataset did not adversely impact generalisation to IMRT plans.

Table 15a Gamma pass rates (3%/3mm) for IMRT validation using model trained on combined dataset

	Gamma Pass Rate (%)						
Model #	1	2	3	4	5	6	7
Mean	97.9	92.7	98.7	94.3	95.7	97.6	98.7
Std. Dev.	0.8	1.0	0.2	5.3	1.2	1.8	1.1
Median	97.7	92.7	98.7	96.8	96.0	97.9	99.0
Max.	99.4	94.5	98.8	99.5	97.5	99.6	99.6
Min.	96.2	90.3	98.3	84.2	91.8	94.0	94.0
Model #	8	9	10	11	12	13	Overall
Mean	98.9	97.7	98.5	98.7	98.1	97.3	97.3
Std. Dev.	1.1	2.7	0.5	0.5	0.9	1.4	1.3
Median	99.4	98.5	98.5	98.7	98.1	97.3	98.1
Max.	99.9	99.4	99.3	99.4	99.6	99.7	99.9
Min.	95.1	87.5	97.3	97.1	95.9	93.9	84.2

Table 15b Gamma pass rates (3%/3mm) for VMAT validation using model trained on combined dataset

	Gamma Pass Rate (%)					
Model #	14	15	16	17	18	19
Mean	91.5	96.8	95.5	92.6	93.5	86.3
Std. Dev.	8.3	2.9	4.2	7.2	2.6	5.2
Median	92.6	97.4	96.5	95.4	93.8	87.1
Max.	98.9	99.8	99.9	99.6	98.0	93.7
Min.	66.6	86.0	86.5	73.6	85.5	66.9
Model #	20	21	22	23	24	Overall
Mean	95.8	96.7	92.1	90.3	98.0	93.6
Std. Dev.	6.3	4.9	9.6	7.0	2.3	2.4
Median	98.0	98.3	95.1	93.4	99.0	95.4
Max.	99.8	99.5	99.9	99.6	99.9	99.9
Min.	72.2	74.5	68.8	75.5	91.9	66.6

Another Wilcoxon signed-rank test revealed that the combined-trained models outperformed the VMAT-trained models on VMAT validation cases (mean=93.6% vs. 90.1%), with the median difference reaching statistical significance ($W = 6, n = 11, p < 0.05$). Table 16 provides a summary of the overall gamma pass rate at 3%/3mm by model type (i.e., which treatment technique the model was trained on, either IMRT-specific, VMAT-specific, or combined trained models). Fig. 20 displays the distribution of pass rates for each model type on (a) IMRT and (b) VMAT validation patients.

Table 16 Summary of overall gamma pass rates for LOOCV at 3%/3mm by model type

	Gamma Pass Rate (%)			
Model Type	IMRT-trained	VMAT-trained	Combined-trained	
Validation Set	IMRT	VMAT	IMRT	VMAT
Mean	98.1	90.1	97.3	93.6
Std. Dev.	1.4	2.6	1.3	2.4
Median	99.0	92.1	98.1	95.4
Max.	100.00	99.7	99.9	99.9
Min.	83.0	63.4	84.2	66.6

By training on a combined dataset, performance was significantly enhanced for VMAT, though the clinical 95% pass rate was still not met. These findings underscore again the inherent differences in delivery techniques and dose modulation between IMRT and VMAT, and the need for larger and more variable datasets to enhance dose prediction accuracy regardless of technique.

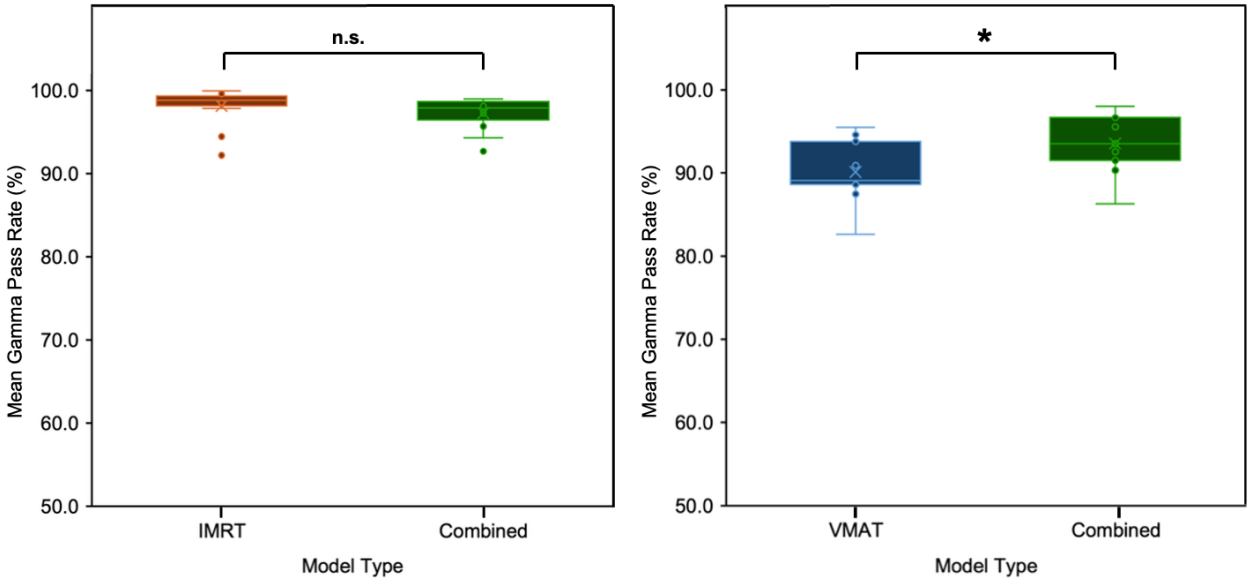

Fig. 20 Box and whisker plots of mean gamma pass rates (3%/3mm) across model types. Comparisons between (a) IMRT- (n=13, Table 7) and combined-trained (n=13, Table 15a) models validated on IMRT dataset; and (b) VMAT- (n=11, Table 11) and combined-trained (n=11, Table 15b) models validated on VMAT dataset. Note: n.s. = not significant ($p > 0.05$), * = $p < 0.05$ (i.e., median difference in gamma pass rates is significantly different), "x" = average; line inside the box = median; lower and upper whiskers = data range (min and max, respectively); dots outside whiskers = outliers.

Interestingly, model 2 for both the IMRT-trained (Section 3.2.1) and combined-trained were validated on the same IMRT patient, yet the mean pass rates were 99.8% vs. 92.7%, respectively. The reduced performance of the combined-model in this case was primarily due to greater gamma failures within the high-dose, PTV region, while the dose fall-off region at the PTV boundary showed relatively few failures. This observation suggests that incorporating both IMRT and VMAT data introduces heterogeneity that

can affect the accuracy of dose prediction in IMRT cases. In Section 3.2.2, it was determined that the U-Net model performed significantly poorer for the VMAT cohort compared to the IMRT cohort, such that a combined-trained model may prioritise minimising errors in VMAT cases, which produce larger loss gradients during the training process. More simply, the complexity and variability of VMAT plans adds noise that complicates prediction of the simpler and standardised dose patterns of IMRT, causing the combined-trained model to underperform compared to the IMRT-only model (Fig. 21a).

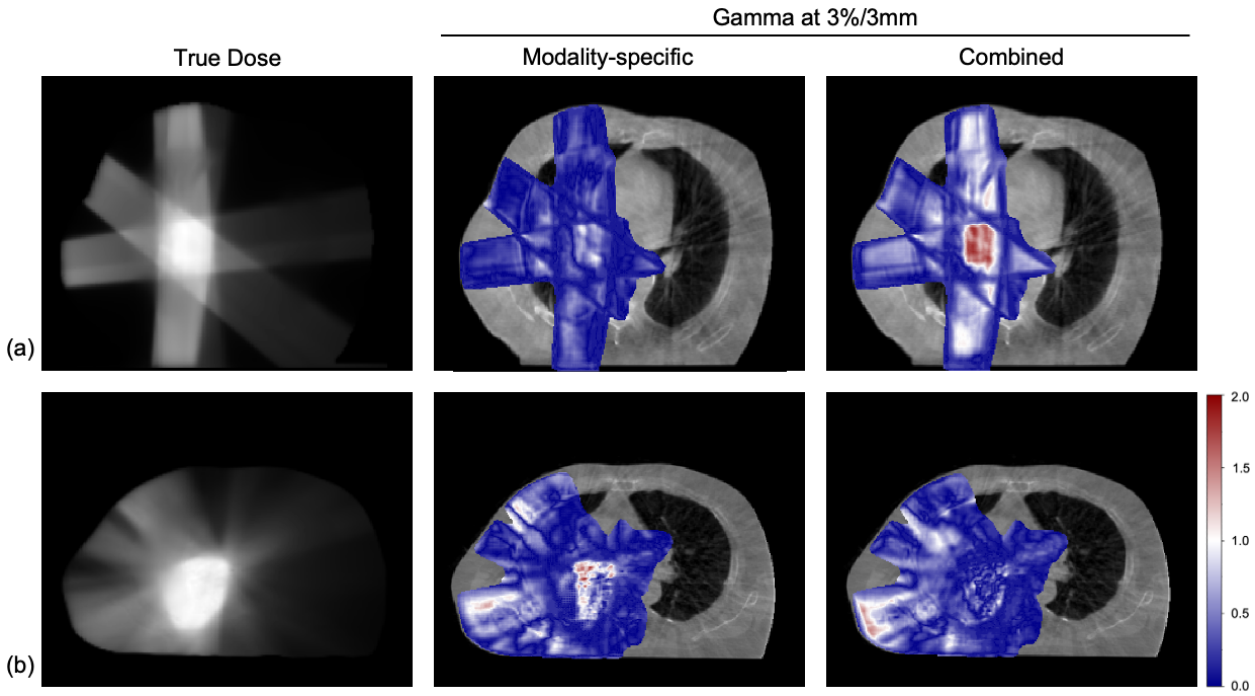

Fig. 21 Effect of model type (modality-specific vs. combined data) on dose prediction accuracy when validated on the same patient. Gamma maps at 3%/3mm tolerance are overlaid on top of axial CBCT image to illustrate regions in which the predicted dose distribution failed to reach $\Gamma \leq 1$. (a) IMRT-trained model 2 vs. combined-trained model 2, validated on the same IMRT patient (mean pass rate=99.8% vs. 92.7%). (b) VMAT-trained model 3 vs. combined-trained model 16, validated on the same VMAT patient (mean pass rate=82.6% vs. 95.5%). Here, mean pass rate refers to the mean over all fractions of the validation patient.

Conversely, validation of VMAT-trained model 3 and combined-trained model 16 on the same VMAT patient revealed more accurate prediction of dose within the PTV and mean pass rate increasing from 82.6%

to 95.5% (Fig. 21b). This may indicate that generalisation over VMAT cases actually benefits from including IMRT data, the consistency of which helps the model learn stable dose-to-anatomy mapping. The differences in dose prediction accuracy using modality-specific vs. combined-trained models are illustrated by gamma maps in Fig. 21.

In short, while combined training did slightly reduce the accuracy of IMRT predictions, the 95% clinical threshold was still reached and significant improvement in the generalisation of VMAT cases was observed, supporting its potential for clinical implementation across radiotherapy techniques and institutions for advanced NSCLC.

3.3 Advantages of Dose Prediction Over CBCT-based Dose Recalculation

With the availability of onboard kV-CBCT, it is possible to perform CBCT-based dose calculation to assess and quantify the dosimetric impact of anatomical changes during treatment [90]. CBCT-based dose calculation involves the application of the initial treatment plan beam parameters and dose calculation algorithm (e.g., CCC or AXB) to the anatomical and density information obtained from the daily CBCT image. Given proper calibration and correction methods, such as density override and curve matching-based techniques, dose calculation accuracy using CBCT images has been shown to be comparable to that of simulation CT [30,91–93]. Specifically, Tarigan et al. [34] and Hu et al. [94] have reported discrepancies within 2% between CBCT- and CT-based dose calculation. Chen et al. [90] showed that CBCT-based dose distributions achieved mean gamma pass rates of $99.0 \pm 1.9\%$, $97.6 \pm 4.4\%$, and $95.3 \pm 6.0\%$ at 3%/3mm (with respect to the planning CT) in high-, medium-, and low-dose regions, respectively. They further concluded that CBCT-based dosimetry can successfully guide re-planning decisions over the course of treatment. Still, U-Net-based dose prediction provides workflow advantages over CBCT-based dose recalculation, particularly in terms of efficiency and feasibility of clinical implementation.

To elaborate, generating ground truth dose distributions with AXB in the Eclipse TPS typically required ~30 seconds per fraction, whereas U-Net dose prediction took ~20 milliseconds (Table 17). In addition to its >1,000-fold increase in computational efficiency, the U-Net model can be easily deployed on a standard GPU-equipped workstation and integrated into a treatment unit. In contrast, CBCT-based dose recalculation using AXB (and other dose computation algorithms) requires access to the TPS, the integration of which on the treatment unit continues to be stalled despite the availability of suitable hardware. A notable exception is the Varian Ethos platform, which offers on-board re-planning but within its proprietary all-in-one system that cannot be extended as a solution for non-Ethos linacs. Ultimately, U-Net dose prediction is an alternative to CBCT-based recalculation that can offer near instantaneous dose deviation assessment at the treatment console to facilitate real-time decisions regarding plan adaptation.

Table 17 Comparison of computation time for AXB dose recalculation vs. U-Net dose prediction

	Time (s)*										
Trial #	1	2	3	4	5	6	7	8	9	10	Mean
Method											
AXB dose recalculation	22.27	20.34	36.62	31.43	39.34	39.75	21.22	30.62	20.02	30.14	29.18
U-Net dose prediction	0.023	0.019	0.023	0.020	0.024	0.022	0.020	0.019	0.018	0.021	0.021

*Note: The time tabulated reflects time required to generate a prediction. It does not include the time of gamma analysis.

3.4 Limitations and Considerations

The following limitations and considerations may restrict the interpretation and generalisability of the results, as well as the reliability and clinical applicability of the model as a clinical guidance tool in its current state.

First, to bypass GPU memory constraints (24 GB VRAM), the CT and dose grid images were rescaled to 74% and 75% of their original size, in the x and y dimensions, respectively. While necessary for

model training with 3D volumetric data and batch size of 5, this reduction likely smoothed dose gradients, as interpolation slightly averages values near sharp gradients, and may have artificially inflated gamma pass rates by obscuring voxel-level discrepancies.

Next, the small size of training and validation datasets limits model generalisability to new IMRT and VMAT cases. Here, a LOOCV scheme was used to maximise the use of the limited dataset and provide a less biased estimate of performance on new data compared to a single train/test split. Deep learning models require large and diverse datasets to generalise well and prevent overfitting. Yet, the dataset used was drawn from a single institution (Juravinski Cancer Centre) and may not reflect CT scanner, physician preferences, or planning approaches that would be expected in wider clinical use. Also, the model was trained using coplanar IMRT cases and may not perform as well for non-coplanar plans. Though non-coplanar beam arrangements are less common than coplanar delivery for NSCLC, they have been employed for some middle and lower lobe tumours to avoid irradiation of OARs, particularly the heart [95].

Another aspect to consider is that the dose distributions were normalised to the maximum dose value for each fraction, which standardises the range of dose values across all training samples and facilitates more effective and consistent model training, especially for training on the combined dataset. However, relative dose prediction limits clinical utility as dose values (cGy) are needed to evaluate compliance with prescribed dose constraints and clinical goals, and conduct dose-volume histogram (DVH) analyses and biological outcome modelling, pointing to the necessity of absolute dose prediction [10,96]. As an aside, DVH analyses also require accurate contours that define the volumes for which the dose distributions are analysed [96,97]. In this study, all patients shared the same prescription dose (63 Gy in 30 fractions), meaning that there is already a degree of standardisation and the relative dose is equal to the absolute dose scaled by a constant. This scaling does not change the spatial dose distribution, voxel-to-voxel ratios, or the patterns being learned by the model. Thus, relative and absolute dose prediction are equally difficult and expected

to yield comparable gamma pass rates. Building off of the work of a previous student, this study continued to focus on relative dose prediction.

Additionally, it should be considered that the U-Net model was initially optimised on IMRT data. Even when trained on VMAT cases, performance did not reach the 95% threshold at 3%/3mm tolerance, indicating technique-specific dose patterns. By tuning hyperparameters and choosing loss functions based on IMRT characteristics, the model may have become biased towards IMRT patterns, leading to suboptimal learning of VMAT features.

While 150 epochs (i.e., iterations over the entire training dataset) yielded high prediction accuracy for the IMRT cohort, with an overall mean pass rate of 98.1% at 3%/3mm, a single LOOCV fold took approximately 2.5 hours. Training one VMAT model was slightly faster, as the VMAT dataset contained a total of 303 fractions, compared to 369 fractions for IMRT. Note that a portion of the total number of fractions (less than 30) is excluded from each training fold for validation. Evidently, training time depends primarily on dataset size and longer training times may limit scalability with additional patient cases.

Further, the ground truth dose distribution was assumed to be the dose recalculated on the CBCT. CBCT images suffer from poor image quality and inaccurate Hounsfield units (HU) or CT numbers since a large volume is imaged by the cone-shaped x-ray beam in a single rotation and captured by a large 2D flat panel detector, whereas a conventional diagnostic CT uses fan-beam geometry and multi-row detector arrays [28]. As a result, CBCT images are more susceptible to scatter and artefacts that degrade HU uniformity and accuracy. The non-uniform HU distribution means that the same tissue or material may can display a wider range of HU values, complicating tissue differentiation and electron density mapping, which limits its application for dose calculation. Still, the Varian OBI system offers small differences in HU values between planning CT and CBCT as it is equipped with a HU calibration procedure using the Catphan phantom [91]. Moreover, with proper calibration and correction methods (e.g., density override or HU-to-electron density curve matching), CBCT-based dose calculations can achieve comparable accuracy to that of planning CT

[30,91–93]. Tarigan et al. [34] showed that the discrepancy between CBCT- and CT-based dose calculation was less than 2% in lung cases, with gamma pass rate of 87-94% for IMRT at 2%/2mm. Hu et al. [94] reported average gamma pass rates of 94.0% at 2%/2mm for thoracic patients on CBCT from Varian Halcyon/Ethos system, as well as an uncertainty within ±2% for the novel AXB iCBCT algorithm. Altogether, these findings suggest that, while HU inaccuracy may contribute to some uncertainty in gamma pass rates, it is unlikely to have caused gross errors in dose calculation. Therefore, the CBCT-based dose distributions used as ground truth in this study can be considered sufficiently accurate.

Lastly, the inclusion of dose distributions computed using both the CCC algorithm in Pinnacle TPS and AXB in Eclipse TPS may have contributed to the observed performance disparity between combined-trained and modality-specific models. The dose calculation approaches of CCC and AXB are fundamentally different. CCC is a superposition-convolution method that approximates dose deposition by convolving the total energy released per unit mass (TERMA) with anisotropic kernels (pre-calculated in water and scaled for tissue density), assuming straight-line energy transport along cone axes, which unfortunately limits accurate modelling of subtle lateral electron scatter contributes, especially in regions of tissue heterogeneity and large density differences [86]. In contrast, AXB solves the linear Boltzmann transport equation (LBTE) and computes the absorbed dose to medium for each voxel in the dose grid, allowing delivered dose distributions to be modelled with higher accuracy than CCC, even in cases of inhomogeneity [97,98]. In other words, AXB simulates actual scatter and electron transport, which is important across interfaces between different tissues where backscatter occurs [99,100]. While dose calculations by AXB have proven to be closely aligned with those of Monte Carlo simulations and direct physical dose measurements, CCC tends to overestimate dose in low-density tissues, including the lung. Inoue et al. [98] reported that the Dmax, Dmin, and D95 values calculated by CCC were 1.17±0.85, 1.95±1.36, and 1.85±0.95 Gy, respectively, which were significantly higher than those of AXB ($p < 0.05$). Their results found significant overestimation and more homogeneous distribution of dose to the PTV by CCC in stereotactic ablative body radiotherapy

(SABR) lung cases, whereas AXB provided a more conformal plan [98]. This overestimation stems from the inaccurate density scaling of water-derived kernels. Nonetheless, Han et al. [100] reported acceptable differences between AXB and CCC, with mean pass rates greater than 90% at 2%/2mm, except in the lung region for 18 MV, 10×10 cm^2 fields. Considering that AXB and CCC are from different TPS, there might exist differences in Pinnacle vs. Eclipse beam models. For these reasons, training a model on combined IMRT and VMAT data, where ground truth dose distributions are derived from two distinct algorithms (with inherently different dose patterns), may have hindered effective generalisation across both modalities. While the difference can be small, it remains clinically relevant, particularly in the lung region, and should be considered in the interpretation of results.

Chapter 4 Conclusion

4.1 Summary of Key Findings

The goal of this work was to adapt, optimise, and evaluate a 3D U-Net dose prediction model that may be implemented on any linac with CBCT, allowing radiation therapists to identify clinically relevant, daily dose deviations in patients with stage III NSCLC treated by IMRT and VMAT. This approach represents a means of quantifying the effect of interfractional anatomical changes on the planned dose distribution, intermediate to fully adaptive and conventional offline workflows, to support timely re-planning decisions. It may also serve to reduce total re-planning time by providing the physician with the predicted dose, which subsequently enables targeted guidance to dosimetrists for faster plan adaptation.

The optimised U-Net of depth 4 and initial kernel count of 32 was trained for 150 epochs with a batch size of 5, combined loss function weighted towards MSE ($\propto$=0.7), and the Adam optimiser, and validated by LOOCV across: (1) an IMRT-only cohort (n=13), (2) a VMAT-only cohort (n=11), and (3) a combined IMRT and VMAT cohort (n=24).

The IMRT-trained models achieved a mean gamma pass rate of $98.1 \pm 1.4\%$ at 3%/3mm tolerance, exceeding the accepted clinical threshold of 95%. Gamma pass rates in the predicted distirbutions generated by these models appeared to correlate positively with beam edges and negatively with regions of large anatomical changes, tissue heterogeneity, and steep dose gradients (i.e., 70–90% dose region). The VMAT-trained models attained a significantly lower mean pass rate of $90.1 \pm 2.6\%$ at 3%/3mm ($p < 0.05$), which may be attributable to the sensitivity of VMAT delivery to anatomical changes within or at the boundaries of the PTV compared to changes beyond the PTV. Unlike the fixed beam angles of IMRT, VMAT involves a continuously rotating gantry to achieve superior modulation and target conformity, generating dose distributions with more varied spatial patterns, across which the models may have struggled to generalise.

By incorporating both IMRT and VMAT data into training, the combined models achieved mean pass rates of $97.3 \pm 1.3\%$ and $93.6 \pm 2.4\%$ at 3%/3mm, when validated on the IMRT and VMAT datasets, respectively. While the addition of VMAT plans may have complicated the prediction of simpler and more standard IMRT dose distributions, the accuracy of VMAT predictions was significantly improved ($p < 0.05$). Overall, the results in this thesis demonstrate the potential of U-Net-based dose prediction as a rapid and cost-effective approach for patient-specific dosimetric verification in advanced NSCLC radiotherapy.

4.2 Future Work

Towards clinical implementation, future work should focus on expanding the dataset to represent a variety of approved clinical cases, as well as those with different anatomical changes, dose patterns, and beam orientations, to improve model robustness and generalisability. For instance, the dataset should reflect different types of IMRT and VMAT lung plans, namely for peripheral and central lung cancer (both PTV encompassing and not-encompassing the mediastinal lymphatic drainage region) [100]. It may also be

beneficial to ensure consistent use of dose calculation algorithms when generating dose distributions, which may help reduce variability in the training inputs [86].

As another next step, incorporating PTV and OAR contours as additional inputs is needed, alongside dose prediction of absolute dose values, for DVH analysis [52–57]. This would allow quantification of the agreement between predicted and true dose distributions within critical structures using clinically relevant metrics (e.g., Dmax, Dmin, D95) and provide a more direct comparison with treatment prescriptions and dose limits. Further, including PTV and OAR contours as inputs would enable the application of a targeted loss weighting scheme that assigns higher loss weights to the PTV and OARs, such that the model prioritises prediction error reduction within clinically important regions during training [57].

It may also be worthwhile to re-optimise the model (e.g., depth of the network) using a subset of combined IMRT and VMAT data, rather than just IMRT, to accommodate the complexity of VMAT [56]. Other U-Net variants or network architectures, such as GAN or V-Net, may be more suitable and should be investigated to address dose prediction for various NSCLC plans [55,101,102]. In terms of clinical integration, the model should be embedded within a user-friendly interface with access to existing TPS and delivery systems.

Collectively, these avenues of future work aim to enhance the accuracy and clinical applicability of a U-Net-based dose prediction model to facilitate daily dosimetric verification and guide clinical re-planning decisions in advanced NSCLC cases.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, May 2024.

[2] Public Health Agency of Canada, "Release notice: Canadian cancer statistics 2023," Health Promotion and Chronic Disease Prevention in Canada, vol. 44, no. 1, p. 1, Jan. 2024. [Online]. Available: https://www.canada.ca/en/public-health/services/reports-publications/health-promotion-chronic-disease-prevention-canada-research-policy-practice/vol-44-no-1-2024/canadian-cancer-statistics-2023.html. [Accessed: Jun. 9, 2025].

[3] D. R. Brenner, J. Gillis, A. A. Demers, L. F. Ellison, J.-M. Billette, S. X. Zhang, et al., "Projected estimates of cancer in Canada in 2024," *CMAJ*, vol. 196, no. 18, pp. E615–E623, 2024.

[4] Q. Guo, L. Liu, Z. Chen, Y. Fan, Y. Zhou, Z. Yuan, and W. Zhang, "Current treatments for non-small cell lung cancer," *Frontiers in Oncology*, vol. 12, p. 945102, 2022.

[5] M. Gouran-Savadkoohi, A. Mesci, G. R. Pond, A. Swaminath, K. Quan, J. Wright, and T. Tsakiridis, "Contemporary real-world radiotherapy outcomes of unresected locally advanced non-small cell lung cancer," *Journal of Thoracic Disease*, vol. 15, no. 2, p. 423, 2023.

[6] D. S. Møller, C. M. Lutz, A. A. Khalil, M. Alber, M. I. Holt, M. Kandi, et al., "Survival benefits for non-small cell lung cancer patients treated with adaptive radiotherapy," *Radiotherapy and Oncology*, vol. 168, pp. 234–240, 2022.

[7] W. Curran, C. Scott, and C. Langer, "Long-term benefit is observed in a phase III comparison of sequential vs concurrent chemoradiation for patients with unresectable stage III NSCLC: RTOG 9410," presented at the American Society of Clinical Oncology Annu. Meeting, 2003.

[8] International Atomic Energy Agency, *Introduction of image guided radiotherapy into clinical practice*, IAEA Human Health Reports No. 16, Vienna, Austria: IAEA, 2019.

[9] P. Iliopoulos, F. Simopoulou, V. Simopoulos, G. Kyrgias, and K. Theodorou, "Review on cone beam computed tomography (CBCT) dose in patients undergoing image guided radiotherapy (IGRT)," in *Advances in Dosimetry and New Trends in Radiopharmaceuticals*, IntechOpen, 2023.

[10] F. M. Khan, P. W. Sperduto, and J. P. Gibbons, *Khan's Treatment Planning in Radiation Oncology*, 5th ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2021.

[11] E. Lin and A. Alessio, "What are the basic concepts of temporal, contrast, and spatial resolution in cardiac CT?," *Journal of Cardiovascular Computed Tomography*, vol. 3, no. 6, pp. 403–408, 2009.

[12] A. T. Davis, A. L. Palmer, and A. Nisbet, "Can CT scan protocols used for radiotherapy treatment planning be adjusted to optimize image quality and patient dose? A systematic review," *The British Journal of Radiology*, vol. 90, no. 1076, p. 20160406, 2017.

[13] F. M. Khan and J. P. Gibbons, *Khan's The Physics of Radiation Therapy*, 5th ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2014.

[14] A. Taylor and M. E. B. Powell, "Intensity-modulated radiotherapy—what is it?," *Cancer Imaging*, vol. 4, no. 2, p. 68, 2004.

[15] J. Boyle, B. Ackerson, L. Gu, and C. R. Kelsey, "Dosimetric advantages of intensity modulated radiation therapy in locally advanced lung cancer," *Advances in Radiation Oncology*, vol. 2, no. 1, pp. 6–11, 2017.

[16] B. Cho, "Intensity-modulated radiation therapy: a review with a physics perspective," *Radiation Oncology Journal*, vol. 36, no. 1, p. 1, 2018.

[17] K. Otto, "Volumetric modulated arc therapy: IMRT in a single gantry arc," *Med. Phys.*, vol. 35, no. 1, pp. 310–317, 2008.

[18] C. S. Chui and S. V. Spirou, "Inverse planning algorithms for external beam radiation therapy," *Medical Dosimetry*, vol. 26, no. 2, pp. 189–197, 2001.

[19] Y. Chen, D. Michalski, C. Houser, and J. M. Galvin, "A deterministic iterative least-squares algorithm for beam weight optimization in conformal radiotherapy," *Physics in Medicine & Biology*, vol. 47, no. 10, p. 1647, 2002.

[20] J. V. Siebers, M. Lauterbach, S. Tong, Q. Wu, and R. Mohan, "Reducing dose calculation time for accurate iterative IMRT planning," *Medical Physics*, vol. 29, no. 2, pp. 231–237, 2002.

[21] L. Brualla, M. Rodriguez, and A. M. Lallena, "Monte Carlo systems used for treatment planning and dose verification," *Strahlentherapie und Onkologie*, vol. 193, no. 4, pp. 243–259, 2017.

[22] F. De Martino, S. Clemente, C. Graeff, G. Palma, and L. Cella, "Dose calculation algorithms for external radiation therapy: an overview for practitioners," *Applied Sciences*, vol. 11, no. 15, p. 6806, 2021.

[23] X. C. Ren, Y. E. Liu, J. Li, and Q. Lin, "Progress in image-guided radiotherapy for the treatment of non-small cell lung cancer," *World Journal of Radiology*, vol. 11, no. 3, p. 46, 2019.

[24] J. De Los Santos, R. Popple, N. Agazaryan, J. E. Bayouth, J. P. Bissonnette, M. K. Bucci, et al., "Image guided radiation therapy (IGRT) technologies for radiation therapy localization and delivery," *International Journal of Radiation Oncology * Biology * Physics*, vol. 87, no. 1, pp. 33–45, 2013.

[25] J. Y. Chang, L. Dong, H. Liu, G. Starkschall, P. Balter, R. Mohan, et al., "Image–guided radiation therapy for non–small cell lung cancer," *Journal of Thoracic Oncology*, vol. 3, no. 2, pp. 177–186, 2008.

[26] J. Westberg, H. R. Jensen, A. Bertelsen, and C. Brink, "Reduction of Cone-Beam CT scan time without compromising the accuracy of the image registration in IGRT," *Acta Oncol.*, vol. 49, no. 2, pp. 225–229, 2010.

[27] L. J. Kroft, J. J. Roelofs, and J. Geleijns, "Scan time and patient dose for thoracic imaging in neonates and small children using axial volumetric 320-detector row CT compared to helical 64-, 32-, and 16-detector row CT acquisitions," *Pediatric Radiology*, vol. 40, no. 3, pp. 294–300, 2010.

[28] J. Pagare, B. Landage, and N. Pagare, "Cone beam computed tomography artefacts—A review," *International Journal of Medical Science and Current Research (IJMSCR)*, vol. 4, no. 2, p. 506, 2021.

[29] R. Schulze, U. Heil, D. Groß, D. D. Bruellmann, E. Dranischnikow, U. Schwanecke, and E. Schoemer, "Artefacts in CBCT: a review," *Dentomaxillofacial Radiology*, vol. 40, no. 5, pp. 265–273, 2011.

[30] I. Fotina, J. Hopfgartner, M. Stock, T. Steininger, C. Lütgendorf-Caucig, and D. Georg, "Feasibility of CBCT-based dose calculation: comparative analysis of HU adjustment techniques," *Radiotherapy and Oncology*, vol. 104, no. 2, pp. 249–256, 2012.

[31] J. J. Sonke, M. Aznar, and C. Rasch, "Adaptive radiotherapy for anatomical changes," *Seminars in Radiation Oncology*, vol. 29, no. 3, pp. 245–257, Jul. 2019.

[32] L. Lu, Z. Zhang, and P. Qi, "A review of online adaptive radiation therapy," *Applied Radiation Oncology*, vol. 4, 2024, doi: 10.37549/ARO-D-24-00037.

[33] S. Lim-Reinders, B. M. Keller, S. Al-Ward, A. Sahgal, and A. Kim, "Online adaptive radiation therapy," *International Journal of Radiation Oncology * Biology * Physics*, vol. 99, no. 4, pp. 994–1003, 2017.

[34] S. T. Tarigan, A. Nainggolan, M. Fadli, S. Liura, and S. A. Pawiro, "Evaluation of adaptive planning of lung cases based on cone beam CT images," *Journal of Physics: Conference Series*, vol. 1505, no. 1, p. 012020, Mar. 2020.

[35] D. N. Stanley, J. Harms, J. A. Pogue, J. G. Belliveau, S. R. Marcrom, A. M. McDonald, et al., "A roadmap for implementation of kV-CBCT online adaptive radiation therapy and initial first year experiences," *Journal of Applied Clinical Medical Physics*, vol. 24, no. 7, p. e13961, 2023.

[36] Y. Archambault, C. Boylan, D. Bullock, T. Morgas, J. Peltola, E. Ruokokoski, et al., "Making online adaptive radiotherapy possible using artificial intelligence and machine learning for efficient daily re-planning," *Med Phys Int J*, vol. 8, no. 2, 2020.

[37] P. K. Pathak, S. K. Vashisht, S. Baby, P. K. Jithin, Y. Jain, R. Mahawar, and V. G. G. K. Sharan, "Commissioning and quality assurance of HalcyonTM 2.0 linear accelerator," *Reports of Practical Oncology and Radiotherapy*, vol. 26, no. 3, pp. 433–444, 2021.

[38] H. Liu, D. Schaal, H. Curry, R. Clark, A. Magliari, P. Kupelian, et al., "Review of cone beam computed tomography based online adaptive radiotherapy: current trend and future direction," *Radiation Oncology*, vol. 18, no. 1, p. 144, 2023.

[39] S. A. Yoganathan, A. Khemissi, S. Paloor, R. Hammoud, and N. Al-Hammadi, "An end-to-end quality assurance procedure for Ethos online adaptive radiotherapy," *Journal of Medical Physics*, vol. 50, no. 1, pp. 140–147, 2025.

[40] D. S. Møller, M. I. Holt, M. Alber, M. Tvilum, A. A. Khalil, M. M. Knap, and L. Hoffmann, "Adaptive radiotherapy for advanced lung cancer ensures target coverage and decreases lung dose," *Radiotherapy and Oncology*, vol. 121, no. 1, pp. 32–38, 2016.

[41] J. Duan, J. Harms, D. H. Boggs, A. J. Kole, R. A. Popple, D. N. Stanley, et al., "Assessing dosimetric benefits of cone beam computed tomography-guided online adaptive radiation treatment frequencies for lung cancer," *Advances in Radiation Oncology*, vol. 10, no. 4, p. 101740, 2025.

[42] H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," in *Deep Learning in Medical Image Analysis: Challenges and Applications*, 1st ed., 2020, pp. 3–21.

[43] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, no. 10, p. 1034, 2024.

[44] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[45] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers*, vol. 15, no. 14, p. 3608, 2023.

[46] R. Guzmán Gómez, G. Lopez Lopez, V. M. Alvarado, F. Lopez Lopez, E. Esqueda Cisneros, and H. López Moreno, "Deep learning approaches for automated prediction of treatment response in non-small-cell lung cancer patients based on CT and PET imaging," *Tomography*, vol. 11, no. 7, p. 78, 2025.

[47] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, p. 53, 2021.

[48] D. Huang, H. Bai, L. Wang, Y. Hou, L. Li, Y. Xia, et al., "The application and development of deep learning in radiotherapy: a systematic review," *Technology in Cancer Research & Treatment*, vol. 20, 2021, Art. no. 15330338211016386.

[49] C. Wang, X. Zhu, J. C. Hong, and D. Zheng, "Artificial intelligence in radiotherapy treatment planning: present and future," *Technology in Cancer Research & Treatment*, vol. 18, 2019, Art. no. 1533033819873922.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, Cham, 2015, pp. 234–241.

[51] I. Aboussaleh, J. Riffi, K. el Fazazy, A. M. Mahraz, and H. Tairi, "3DUV-NetR+: A 3D hybrid semantic architecture using transformers for brain tumor segmentation with MultiModal MR images," *Results in Engineering*, vol. 21, p. 101892, 2024.

[52] D. Nguyen, T. Long, X. Jia, W. Lu, X. Gu, Z. Iqbal, and S. Jiang, "A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning," *Scientific Reports*, vol. 9, no. 1, p. 1076, 2019.

[53] S. Willems, W. Crijns, E. Sterpin, K. Haustermans, and F. Maes, "Feasibility of CT-only 3D dose prediction for VMAT prostate plans using deep learning," in *Proc. First Int. Workshop Artificial Intelligence Radiother. (AIRT),* Shenzhen, China, Oct. 2019, pp. 10–17.

[54] J. Ma, T. Bai, D. Nguyen, M. Folkerts, X. Jia, W. Lu, L. Zhou, and S. Jiang, "Individualized 3D dose distribution prediction using deep learning," in *Proc. First Int. Workshop Artificial Intelligence Radiother. (AIRT)*, Shenzhen, China, Oct. 2019, pp. 110–118.

[55] J. Liu, X. Zhang, X. Cheng, and L. Sun, "A deep learning-based dose prediction method for evaluation of radiotherapy treatment planning," *J. Radiat. Res. Appl. Sci.*, vol. 17, no. 1, p. 100757, 2024.

[56] M. Wang, Y. Pan, X. Zhang, and R. Yang, "Exploring the impact of network depth on 3D U-Net-based dose prediction for cervical cancer radiotherapy," *Frontiers in Oncology*, vol. 14, p. 1433225, 2024.

[57] W. Cao, M. Gronberg, S. Bilton, H. Baroudi, S. Gay, C. Peeler, et al., "Dose prediction via deep learning to enhance treatment planning of lung radiotherapy including simultaneous integrated boost techniques," *Med. Phys.*, vol. 52, no. 5, pp. 3336–3347, 2025.

[58] H. Li, A. K. Lee, J. L. Johnson, R. X. Zhu, and R. J. Kudchadker, "Characterization of dose impact on IMRT and VMAT from couch attenuation for two Varian couches," *J. Appl. Clin. Med. Phys.,* vol. 12, no. 3, pp. 23–31, 2011, doi: 10.1120/jacmp.v12i3.3471.

[59] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of SimpleITK," *Front. Neuroinform.*, vol. 7, p. 45, 2013.

[60] T. Kimpe and T. Tuytschaever, "Increasing the number of gray shades in medical display systems—how much is enough?," *J. Digit. Imaging*, vol. 20, no. 4, pp. 422–432, 2007, doi: 10.1007/s10278-006-1052-3.

[61] A. Aghdam, "unet3d.py," GitHub, [Online]. Available: https://github.com/amir-aghdam/3D-UNet/blob/main/unet3d.py. [Accessed: Dec. 1, 2024].

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1026–1034. doi: 10.48550/arXiv.1502.01852.

[63] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Annals of Data Science*, vol. 9, no. 2, pp. 187–212, 2022, doi: 10.1007/s40745-020-00253-5.

[64] Z. Huang, Z. Wang, Z. Yang, and L. Gu, "Adwu-net: Adaptive depth and width u-net for medical image segmentation by differentiable neural architecture search," in *Proc. Int. Conf. Med. Imaging Deep Learn. (MIDL)*, Dec. 2022, pp. 576–589.

[65] J. Kugelman, J. Allman, S. A. Read, S. J. Vincent, J. Tong, M. Kalloniatis, F. K. Chen, M. J. Collins, and D. Alonso-Caneiro, "A comparison of deep learning U-Net architectures for posterior segment OCT retinal layer segmentation," *Scientific Reports*, vol. 12, no. 1, p. 14888, 2022.

[66] M. Khouy, Y. Jabrane, M. Ameur, and A. Hajjam El Hassani, "Medical image segmentation using automatic optimized U-Net architecture based on genetic algorithm," *Journal of Personalized Medicine*, vol. 13, no. 9, p. 1298, 2023.

[67] M. Chenier and M. Wierzbicki, "Towards AI-based dose prediction of daily delivered dose during lung cancer radiotherapy," *Radiotherapy and Oncology*, vol. 186, Suppl. 1, p. S170, 2023. doi: 10.1016/S0167-8140(23)89262-3.

[68] H. H. Rashidi, S. Albahra, S. Robertson, N. K. Tran, and B. Hu, "Common statistical concepts in the supervised Machine Learning arena," *Frontiers in Oncology*, vol. 13, p. 1130229, 2023. doi: 10.3389/fonc.2023.1130229.

[69] M. A. K. Raiaan, S. Sakib, N. M. Fahad, A. Al Mamun, M. A. Rahman, S. Shatabda, and M. S. H. Mukta, "A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks," *Decision Analytics Journal*, p. 100470, 2024. doi: 10.1016/j.dajour.2024.100470.

[70] W. K. Hong, *Artificial Intelligence-Based Design of Reinforced Concrete Structures: Artificial Neural Networks for Engineering Applications*, Woodhead Publishing Series in Civil and Structural Engineering. Elsevier, 2023.

[71] R. Singh, "Dose prediction for radiotherapy of advanced stage lung cancer," M.S. thesis, Dept. of Health and Radiation Physics, McMaster Univ., Hamilton, ON, Canada, 2020. [Online]. Available: http://hdl.handle.net/11375/26017

[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. doi: 10.48550/arXiv.1412.6980.

[73] X. Wang and L. Aitchison, "How to set AdamW's weight decay as you scale model and dataset size," arXiv preprint arXiv:2405.13698, 2024. doi: 10.48550/arXiv.2405.13698.

[74] D. A. Low, W. B. Harms, S. Mutic, and J. A. Purdy, "A technique for the quantitative evaluation of dose distributions," *Medical Physics*, vol. 25, no. 5, pp. 656–661, 1998. doi: 10.1118/1.598248.

[75] S. Diamantopoulos, K. Platoni, G. Patatoukas, P. Karaiskos, V. Kouloulias, and E. Efstathopoulos, "Treatment plan verification: a review on the comparison of dose distributions," *Physica Medica*, vol. 67, pp. 107–115, 2019. doi: 10.1016/j.ejmp.2019.10.029.

[76] G. A. Ezzell, J. W. Burmeister, N. Dogan, T. J. LoSasso, J. G. Mechalakos, D. Mihailidis, et al., "IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM Task Group 119," *Medical Physics*, vol. 36, no. 11, pp. 5359–5373, 2009. doi: 10.1118/1.3238104.

[77] M. Miften, A. Olch, D. Mihailidis, J. Moran, T. Pawlicki, A. Molineu, et al., "Tolerance limits and methodologies for IMRT measurement-based verification QA: recommendations of AAPM Task Group No. 218," *Medical Physics*, vol. 45, no. 4, pp. e53–e83, 2018. doi: 10.1002/mp.12810.

[78] Y. Wang, Z. Piao, H. Gu, M. Chen, D. Zhang, and J. Zhu, "Deep learning-based prediction of radiation therapy dose distributions in nasopharyngeal carcinomas: a preliminary study incorporating multiple features including images, structures, and dosimetry," *Technology in Cancer Research & Treatment*, vol. 23, 2024, Art. no. 15330338241256594. doi: 10.1177/15330338241256594.

[79] R. N. Kandalan, D. Nguyen, N. H. Rezaeian, A. M. Barragán-Montero, S. Breedveld, K. Namuduri, *et al.,* "Dose prediction with deep learning for prostate cancer radiation therapy: model adaptation to different treatment planning practices," *Radiotherapy and Oncology*, vol. 153, pp. 228–235, 2020. doi: 10.1016/j.radonc.2020.10.027.

[80] J. M. Steers and B. A. Fraass, "The effect of dose gradients on gamma comparison insensitivity in patient specific QA comparisons," *Med. Phys.*, vol. 50, no. 6, pp. 3671–3686, 2023.

[81] E. B. Podgorsak, "External photon beams: Physical aspects," in *Radiation Oncology Physics: A Handbook for Teachers and Students*, Vienna, Austria: IAEA, 2005, p. 169.

[82] J. Y. Chang, "Intensity-modulated radiotherapy, not 3 dimensional conformal, is the preferred technique for treating locally advanced lung cancer," *Semin. Radiat. Oncol.*, vol. 25, no. 2, pp. 110–116, 2015.

[83] H. Xu and G. Hatcher, "Treatment planning study of volumetric modulated arc therapy and three dimensional field-in-field techniques for left chest-wall cancers with regional lymph nodes," *Rep. Pract. Oncol. Radiother.*, vol. 21, no. 6, pp. 517–524, 2016.

[84] B. Jany and T. Welte, "Pleural effusion in adults—etiology, diagnosis, and treatment," *Deutsches Ärztebl. Int.*, vol. 116, no. 21, p. 377, 2019.

[85] V. S. Karkhanis and J. M. Joshi, "Pleural effusion: diagnosis, treatment, and management," *Open Access Emerg. Med.*, 2012. doi: 10.2147/OAEM.S29942.

[86] Q. Liu, J. Liang, C. W. Stanhope, and D. Yan, "The effect of density variation on photon dose calculation and its impact on intensity modulated radiotherapy and stereotactic body radiotherapy," *Med. Phys.*, vol. 43, no. 10, pp. 5717–5729, 2016.

[87] C. Altunbas, B. Kavanagh, W. Dzingle, K. Stuhr, L. Gaspar, and M. Miften, "Dosimetric errors during treatment of centrally located lung tumors with stereotactic body radiation therapy: Monte

Carlo evaluation of tissue inhomogeneity corrections," *Med. Dosimetry*, vol. 38, no. 4, pp. 436–441, 2013.

[88] M. L. Schmidt, L. Hoffmann, M. M. Knap, T. R. Rasmussen, B. H. Folkersen, J. Toftegaard, et al., "Cardiac and respiration induced motion of mediastinal lymph node targets in lung cancer patients throughout the radiotherapy treatment course," *Radiother. Oncol.*, vol. 121, no. 1, pp. 52–58, 2016.

[89] S. Shiraishi and K. L. Moore, "Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy," *Med. Phys.*, vol. 43, no. 1, pp. 378–387, 2016.

[90] S. Chen, Q. Le, Y. Mutaf, W. Lu, E. M. Nichols, B. Y. Yi, et al., "Feasibility of CBCT-based dose with a patient-specific stepwise HU-to-density curve to determine time of replanning," *J. Appl. Clin. Med. Phys.*, vol. 18, no. 5, pp. 64–69, 2017.

[91] M. De Smet, D. Schuring, S. Nijsten, and F. Verhaegen, "Accuracy of dose calculations on kV cone beam CT images of lung cancer patients," *Med. Phys.*, vol. 43, no. 11, pp. 5934–5941, 2016.

[92] A. I. Holm, T. B. Nyeng, D. S. Møller, M. S. Assenholt, R. Hansen, L. Nyvang, et al., "Density calibrated cone beam CT as a tool for adaptive radiotherapy," *Acta Oncol.*, vol. 60, no. 10, pp. 1275–1282, 2021.

[93] R. S. Thing, R. Nilsson, S. Andersson, M. Berg, and M. D. Lund, "Evaluation of CBCT based dose calculation in the thorax and pelvis using two generic algorithms," *Phys. Med.*, vol. 103, pp. 157–165, 2022.

[94] Y. Hu, M. Arnesen, and T. Aland, "Characterization of an advanced cone beam CT (CBCT) reconstruction algorithm used for dose calculation on Varian Halcyon linear accelerators," *Biomed. Phys. Eng. Express*, vol. 8, no. 2, p. 025023, 2022.

[95] S. Elzawawy, D. Alzayat, and A. Darwish, "Dosimetric comparison between coplanar and non-coplanar fields in irradiation of middle and lower lobes lung tumors," *Int. J. Med. Phys. Clin. Eng. Radiat. Oncol.*, vol. 5, no. 2, pp. 130–137, 2016.

[96] S. X. Jiao, M. L. Wang, L. X. Chen, and X. W. Liu, "Evaluation of dose-volume histogram prediction for organ-at risk and planning target volume based on machine learning," *Sci. Rep.*, vol. 11, no. 1, p. 3117, 2021.

[97] A. S. Oinam, L. Singh, A. Shukla, S. Ghoshal, R. Kapoor, and S. C. Sharma, "Dose volume histogram analysis and comparison of different radiobiological models using in-house developed software," *J. Med. Phys.*, vol. 36, no. 4, pp. 220–229, 2011.

[98] K. Inoue, H. Matsukawa, Y. Kasai, K. Edamitsu, K. Matsumoto, Y. Suetsugu, et al., "Difference in target dose distributions between Acuros XB and collapsed cone convolution/superposition and the impact of the tumor locations in clinical cases of stereotactic ablative body radiotherapy for lung cancer," *J. Cancer Res. Ther.*, vol. 19, no. 5, pp. 1261–1266, 2023.

[99] G. A. Failla, T. Wareing, Y. Archambault, and S. Thompson, "Acuros XB advanced dose calculation for the Eclipse treatment planning system," Palo Alto, CA: Varian Medical Systems, 2010, p. 18.

[100] T. Han, J. K. Mikell, M. Salehpour, and F. Mourtada, "Dosimetric comparison of Acuros XB deterministic radiation transport method with Monte Carlo and model-based convolution methods in heterogeneous media," *Med. Phys.*, vol. 38, no. 5, pp. 2651–2664, 2011.

[101] Y. Li, J. Wang, L. Tan, B. Hui, X. Ma, Y. Yan, et al., "Dosimetric comparison between IMRT and VMAT in irradiation for peripheral and central lung cancer," *Oncol. Lett.*, vol. 15, no. 3, pp. 3735–3745, 2018.

[102] Y. Peng, D. Z. Chen, and M. Sonka, "U-net v2: Rethinking the skip connections of u-net for medical image segmentation," in *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 2025, pp. 1–5.

[103] H. Zhang, Y. Yu, and F. Zhang, "Prediction of dose distributions for non-small cell lung cancer patients using MHA-ResUNet," *Med. Phys.*, vol. 51, no. 10, pp. 7345–7355, 2024.