Thesis

Study on Quick Prediction of Dose Volume Statistics in Proton Beam Therapy using Deep Learning

(深層学習を用いた陽子線治療における線量体積統計量の迅速予測に関する研究)

03/24/2022

Hokkaido University

Sira Jampa-ngern

Thesis

Study on Quick Prediction of Dose Volume Statistics in Proton Beam Therapy using Deep Learning

(深層学習を用いた陽子線治療における線量体積統計量の迅速予測に関する研究)

03/24/2022

Hokkaido University

Sira Jampa-ngern

Table of Contents

Presented Paper List and Conference Presentation List

Part of this study was presented in the following paper:

1. Sira Jampa-ngern, Keiji Kobashi, Shinichi Shimizu, Seishin Takao, Keiji Nakazato, Hiroki Shirato
   Prediction of liver D_{mean} for proton beam therapy using deep learning and contour-based data augmentation
   Journal of Radiation Research, 62(6), 1120-1129 (2021)

Part of this study was presented in the following academic conferences:

1. Sira Jampa-ngern, Keiji Kobashi, Shinishi Shimizu
   Feasibility Study on DVH Estimation by Machine Learning from Small Data towards Simplified Model-based Approach for Selective Use of Proton Therapy
   The 118th Congress on Japan Society of Medical Physics (JSMP), 12-14 September 2019, Fukui, Japan
2. Sira Jampa-ngern, Keiji Kobashi, Shinichi Shimizu
   Study of EUD estimation using machine learning from small data as pre-screening tool prior to MBA for PBT patient selection
   The 121st Scientific Meeting of the Japan Society of Medical Physics (JSMP), 15-18 April 2021, Yokohama, Japan

Introduction

Proton beam therapy (PBT) is effective for cancer treatment due to advantageous physical characteristics which allows minimizing unwanted doses and limits complication outcomes from radiation when compared with X-ray therapy (XRT) including intensity modulated X-ray therapy (IMXT) and stereotactic irradiation (STI). However, the treatment costs of PBT are much higher than IMXT and STI (Loeffler and Durante, 2013), and in many countries, the access of PBT systems is very limited. For this reason, the patient selection for PBT is important from the aspect of social healthcare economics.

A biophysical model for normal tissue complication probability (NTCP) of organs at risk (OAR) has been proposed for the selection of PBT versus XRT prescribing the same dose to the cancer. However, uncertainties in the parameters of the model-based approach still limits its wide spread usage (Prayongrat *et al.*, 2018). In the real world including our institution, the evaluation for PBT is often assessed by the dose volume statistics such as the mean dose (D_{mean}) of OAR rather than NTCP. For example, in a conventionally fractionated XRT, liver D_{mean} < 28 - 32 Gy for 5% of classical radiation-induced liver diseases has been often used as the dose constrains based on the clinical evidence (Marks *et al.*, 2010). Applying 1.1 as the relative biological effectiveness (RBE) of PBT in the clinical decision (Chen *et al.*, 2018), if the liver D_{mean} is lower than the threshold in PBT but not in XRT, the PBT is often recommended as a radiotherapy option for the cancer board for liver cancer in our institution.

This raises the problem that the physician has to wait overly long before being able to select the treatment option for cancer patients in daily clinical practice. The physicians are required to prepare the results of the liver D_{mean} calculations for to the cancer board within a few hours or few days after the visit of the patient or the inquiry from a referring hospital. Three-dimensional (3D) radiotherapy treatment planning (RTP) using computed tomography (CT) requires considerable time and resources for both PBT and XRT since contouring of many OARs and intensity modulation has become the standard of care for many diseases (Cilla *et al.*, 2021). The development and availability of a simple quantitative clinical decision support tool for the selection of PBT presenting less of a burden will be helpful to overcome these barriers.

Recently, a machine learning including deep learning (DL) techniques has become widely used in the medical field. In the radiation oncology area, deep learning is applied for many tasks including image segmentation and detection, image phenotyping and radiomic signature discovery, clinical outcome prediction, image dose quantification, dose-response modeling, radiation adaption, and image generation (Boldrini *et al.*, 2019). The DL-based automatic 3DRTP for XRT is proposed for situations with relatively large sets of data. Chen et al., Nguyen et al., and Ibragimov et al. have used 80 patients, 88 patients, and 125 patients, respectively (Ibragimov *et al.*, 2018; Chen *et al.*, 2019; Nguyen *et al.*, 2019). However, the number of training images in oncology studies are criticized as still too few since it is known that about 5,000 labeled examples are required as training input to train a convolution neural network (CNN) effectively, as a role of thumb (Goodfellow, Bengio and Courville, 2016). To develop CNN models of the 3DRTP for PBT, it will be more difficult to collect sufficient numbers of data sets than XRT studies because of the small number of patients who have been (and are being) treated with PBT.

The theme of this thesis is the development of a simple quantitative clinical decision support tool for the selection of PBT with less of a burden. This thesis is composed of Chapter 1 and chapter 2 to solve the problems in the selection of PBT. The first chapter is the prediction of liver D_{mean} for PBT using CNN and contour-based data augmentation using our in-house programs. The second chapter is about the DL with architecture of pre-trained models to improve our in-house program.

Abbreviation Table

The abbreviations used in the body text and figures are as follows:

2DP	2-dimension profiling
3DP	3-dimension profiling
BP	Bragg peak
CDA	Contour-based data augmentation
CNN	Convolution neural network
Conv	Convolution
CT	Computer tomography
CTOI	CT of interest
CTV	Clinical target volume
dDVH	Differential dose volume histogram
DICOM	Digital Imaging and Communications in Medicine
DL	Deep learning
D_{max}	Maximum dose
D_{mean}	Mean dose
D_{min}	Minimum dose
DVS	Dose volume statistic
DWConv	Depthwise convolution
EUD	Equivalent uniform dose
GTV	Gross tumor volume
IMPT	intensity modulated proton beam therapy

IMXT	intensity modulated X-ray therapy
MBA	Model-based approach
MPE	Mean percentage error
MRE	Mean relative error
MSE	Mean square error
NTCP	Normal tissue complication probability
OAR	Organ at risk
PBT	Proton Beam Therapy
PC	Personal computer
PDD	Percentage depth dose
PTV	Planning target volume
ReLU	Rectified linear units
RILD	radiation-induced liver diseases
ROI	Region of interest
RT	Radiation therapy
SDP	Simple dose Prediction
SOBP	Spread-out-Bragg-peak
STI	Stereotactic irradiation
TCIA	The Cancer Imaging Archive
TCP	Tumor control probability
VOI	Volume of interest
XRT	X-ray therapy
β	Regression coefficient

Chapter 1

1.1 Introduction

There are two subjects addressed in this chapter, the first subject is the investigation about our proposal of a simple dose prediction tool (SDP). We developed a prototype of the SDP which can predict the liver D_{mean} in PBT for liver cancer is possible based on CNN. The liver D_{mean} was selected since the liver can be regarded as a parallel structure organ in which D_{mean} is known to be a reasonable clinical and biological metric (Dawson *et al.*, 2002). The physicians are not required to wait for a precise 3DRTP comparison when they only require an approximate predicted liver D_{mean} for rapid presentation to the cancer board or in discussions with the health insurance society. Compared to 3DRTP, the SDP requires fewer resources to predict the liver D_{mean}.

The second subject is the proposal of contour-based data augmentation (CDA) before the generation of the CNN to predict the outcome, the liver D_{mean}. In the CDA, contours from a small number of patients are required to generate the CNN. Virtual CTV are artificially embedded in the liver for data augmentation of the contours of the CTV and corresponding liver D_{mean} are used as labels for the data sets. We term this approach the CDA approach in this study. In the training process with a big amount of conventional CT image data, the memory of computers is often insufficient during the training of the network. This can be mitigated by reduction of input data size using the CDA approach. Another advantage of the CDA approach is the rapidity of the model training and data loading. The working hypothesis of this study is that deep learning and CDA can quickly predict the liver D_{mean} for PBT without increasing the input data and time for the generation of the CNN.

1.2 Experiment Method

1.2.1 Concept and workflow of the Simple dose prediction (SDP) tool

A conceptual flow chart to utilize the SDP is shown in **Figure 1**. A simple contouring software package is required for the physician to do the contouring of the CTV and OARs on the CT images. Information about the size (mm) of a pixel of the two-dimensional (2D) image, the number of pixels in one image (for example, 512 x 512), and the slice thickness of the original image are required. The contours are input into a personal computer where the SDP has been installed. The SDP can be installed in the personal computer where the contouring has been performed. The contours are then used to estimate the liver D_{mean} based on the simple dose calculation algorithm described below. The predicted liver D_{mean} will then directly be shown on the screen. The contours are visualized on the screen with pre-determined homogeneous grayscales using the bitmap format and the physician can visually check the appropriateness of the contours input to the SDP.

In a real-world clinic, the angles to treat a type of tumors are often determined by treatment plan protocols in the particular hospital. For example, 2 beam with almost the same beam angles are used for liver cancers in our institution (data not shown). In the SDP, therefore, several different combinations of two angles can be pre-determined for liver cancer. The program directly predicts the liver D_{mean} for the pre-determined combinations of portal numbers and angles. The physician can use the liver D_{mean} to select the best combination of portals based on the threshold on the liver D_{mean} which is also pre-determined by the hospital or health authorities of the country. In this study, a patient with a liver cancer is assumed to visit a clinic with CT images in the DICOM format. We have selected 3 organs as the candidate OARs for contouring: liver, spinal cord, and skin.

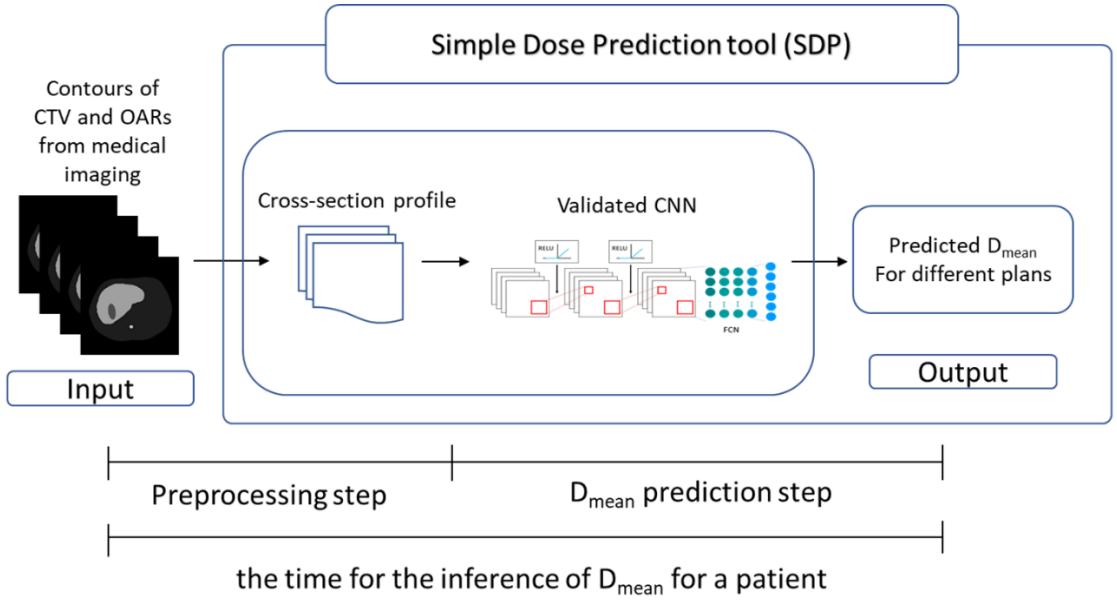

Figure 1. Flowchart of simple dose prediction (SDP) tool with the time to establish the inference of liver mean dose (D_{mean}) for a patient.

1.2.2 Image data sets and Contour-based Data Augmentation (CDA) before DL

In this study, CT images of actual patients with a single liver tumor were obtained from The Cancer Imaging Archive (Clark *et al.*, 2013; Erickson *et al.*, 2016). There were 52 data sets which were labeled 'liver cancer'. A radiologist and a radiation technologist have seen the images and removed 35 data sets of magnetic resonance images, multiple metastatic liver cancers, and diffuse tumors which are not good candidates for radiotherapy. Therefore, the remaining 17 data sets of CT images were available to be used in this study. The CT slice thickness and side length range was 2.5-8.0 mm and 320-400 mm, respectively. The CTV margin for gross tumor volume was assumed to be 0.0 cm. The contours of the CTV, liver, spinal cord, and skin surface in the CT images were provided by MIM (MIM Software Inc).

To increase the data for the deep learning process, CDA was used in the training and validation (**Figure 2**). Virtual CTVs were generated and placed inside the contours of the liver as ellipsoids at various principal semi-axes and orientations for the data augmentation (**Figure 3**). The part of the virtual CTV which is outside the liver boundary was removed. For each patient data set, 199 virtual CTVs of different sizes and locations,

were generated making it 200 CTVs including the original and virtual CTVs for one patient.

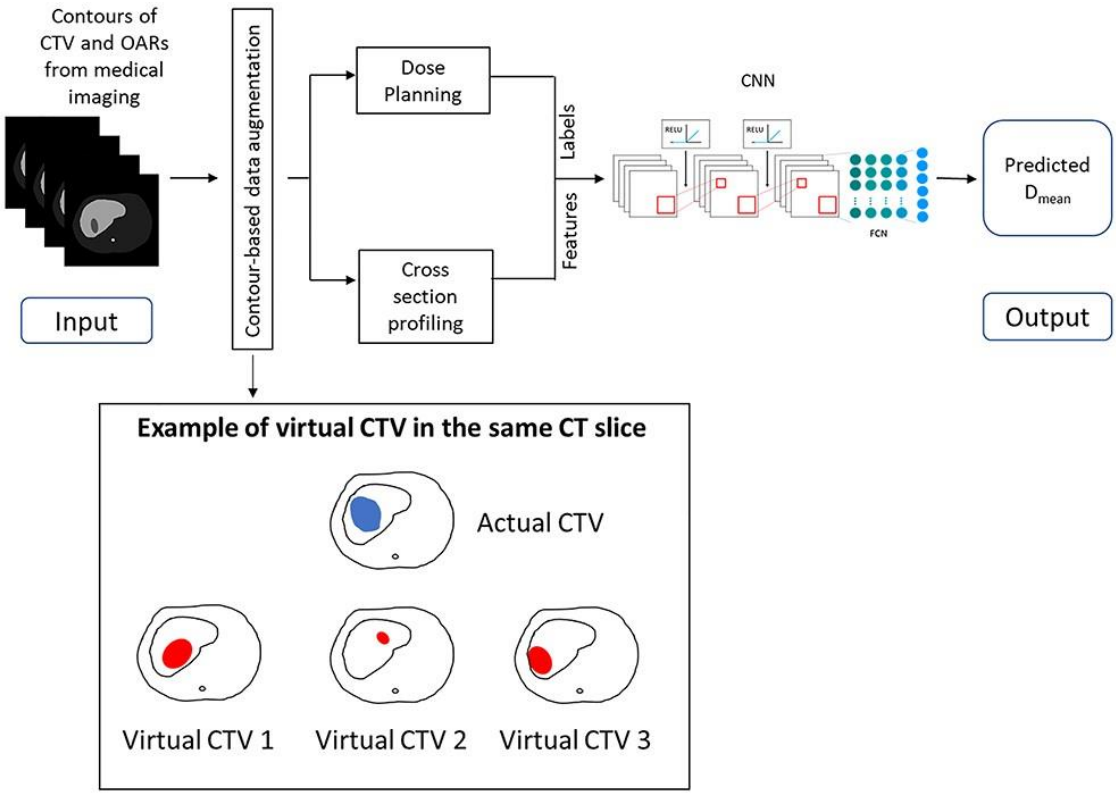

Figure 2. Flow chart for the training of the CNN in the SDP tool using CDA. CTV: clinical target volume, OAR: organ at risk, CNN: convolution neural network.

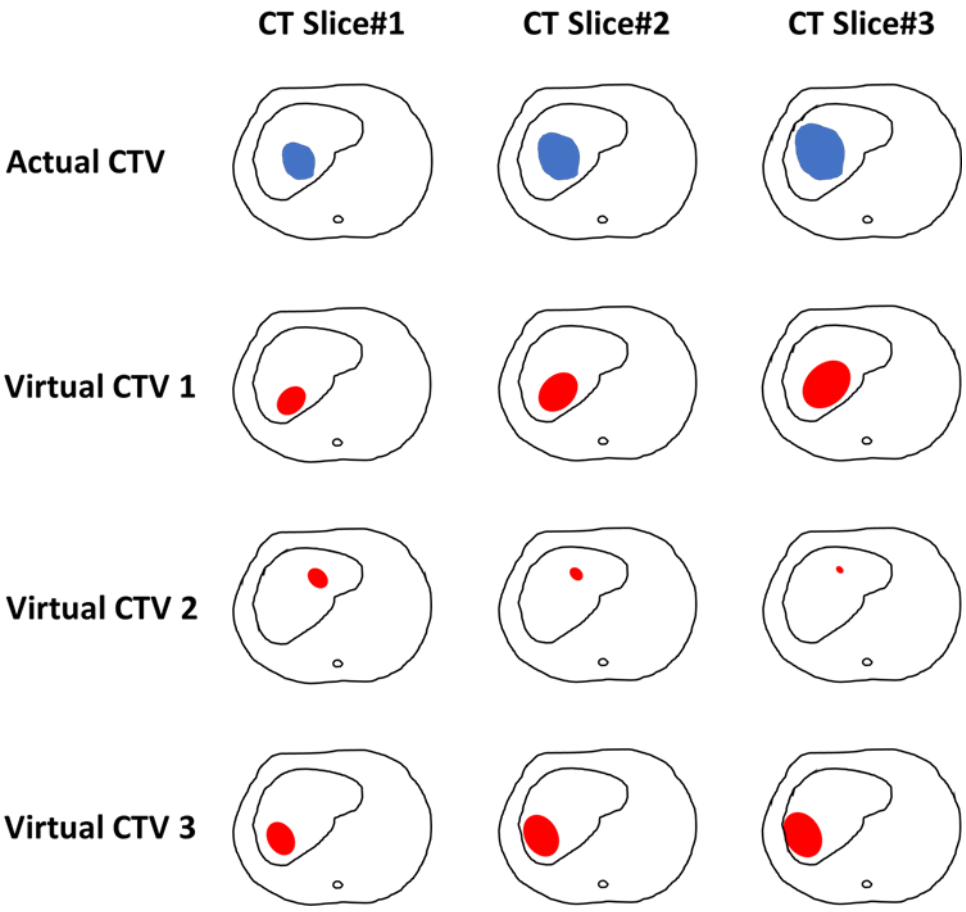

Figure 3. The example of creating and embedding of virtual CTVs inside the actual patient contour in contour-based data augmentation (CDA).

In the 17 data sets of patients with a single liver tumor obtained from TCIA (Clark *et al.*, 2013; Erickson *et al.*, 2016), we assumed that the minimum number of data sets for the test is 10 (Chen *et al.*, 2019; Nguyen *et al.*, 2019) and randomly selected 10 data sets from the 17 data sets available **(Table 1)**. The mean of the diameters of the CTV in the 10 patients for tests was 3.317 cm (0.707-7.213 cm) (**Table 3)**.

Table 1. CT slices parameter of the 10 patients used in test data set and the total time required to predict the liver D_{mean} for 6 plans altogether (by simultaneous model).

Patient	Total no. of CT slices	No. of liver slices	No. of tumor slices	Slice thickness (cm)	Slide range (cm)	Pre-processing time (sec)	D_{mean} estimation time (sec)	Total time (sec)
Test 1	46	32	7	0.50	36.00	8.47	4.40	12.87
Test 2	95	70	9	0.30	38.00	13.34	4.51	17.85
Test 3	46	19	3	0.50	36.00	6.27	4.39	10.66
Test 4	60	20	14	0.80	35.60	6.94	4.46	11.40
Test 5	90	74	5	0.25	40.00	12.40	4.33	16.73
Test 6	106	69	12	0.25	35.00	12.56	4.63	17.19
Test 7	86	42	36	0.50	39.00	10.15	4.52	14.67
Test 8	62	34	7	0.68	35.00	6.91	4.39	11.30
Test 9	97	60	16	0.25	40.00	12.72	4.49	17.21
Test 10	58	55	28	0.50	32.00	6.34	4.58	10.92
Mean	74.60	47.50	13.70	0.45	36.66	9.61	4.47	14.08
SD	22.50	20.83	10.60	0.19	2.56	2.94	0.09	2.96

Abbreviation: D_{mean} = Mean dose, SD = Standard deviation

The remaining 7 data sets were used for the training and validation of the prediction model. The validation dataset is used to prevent overfitting and selected as the best performance of the prediction model. No data sets for tests were used in the training and validation. The median volume of the CTV was 11 (range: 2-508) cm^3 and the median of the liver volume was 1202 (896-2299) cm^3 in the 7 patients for training or validation.

1.2.3 Investigation of appropriate parameter and data sets before DL

To access the best performance of our CDA method we investigated an appropriate 1) number of virtual CTVs by CDA, 2) ratio of training to validation data sets, 3) number of OARs, 4) number for parameter K, and 5) Dose prediction of 6 D_{mean}; simultaneously or separately. The MRE was used as a measure of the error between the predicted liver D_{mean} and the planned liver D_{mean} which was predicted from the model trained by the different conditions. The condition with the lowest MRE was used as the appropriate one in the deep learning.

1) Number of virtual CTVs by CDA

To investigate the appropriate number of CTVs (virtual CTVs + actual CTV) for training and validation, several different numbers of CTVs, at 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 250, and 300 were used without changing the other parameters. The ratio of training to validation data sets was 5:2 and the training OAR input is only the CTV with the liver. The trained model by the different numbers of CTVs was tested by 10 actual patients. The MRE was used as a measurement of the discrepancy between the predicted liver D_{mean} and the planned liver D_{mean} which was predicted from the model trained by the different numbers of CTVs. The appropriate number of CTVs was selected at the point of convergence of MRE with the number of CTVs.

2) Ratio of training to validation data sets

To investigate the appropriate ratio of data sets of patients for training to for validation, we changed the ratio (training: validation) from 1:6, 2:5, 3:5, 4:3, 5:2, 6:1, to 7:0 without changing the other parameters. The training OAR input was for the liver only (with the CTVs) and the number of virtual CTVs were 199 for each actual patient. The model was trained and tested by 10 actual patient data sets. The MRE was used as a measure of the error between the predicted liver D_{mean} and the planned liver D_{mean} which was predicted from the model trained by different ratios in the training and validation. The condition with the lowest MRE was used as the appropriate ratio.

3) Number of OARs

To investigate the appropriate number of OARs in deep learning, the model was trained with different conditions of OARs without changing the other parameters, 1) 3 OARs including skin, liver, and spinal cord, 2) 2 OARs including skin and liver, 3) 1 OAR including only the skin, and 4) 1 OAR including only the liver. The ratio of training to validation data sets was 5:2 and the number of CTVs (virtual CTVs + actual CTV) were 200 for each actual patient. The model trained by different conditions was tested by 10 actual patient data sets. The MRE was used as a measure of the error between the predicted liver D_{mean} and the planned liver D_{mean} which was predicted from the model trained by the different conditions of the OAR(s). The

condition with the lowest MRE was used as the appropriate number of OARs in the deep learning.

4) Number for parameter K

To investigate the appropriate number for parameter K, the models were trained with different numbers of parameter K from 4-15 without changing the other parameters. The ratio of training to validation data sets was 5:2, the training OAR input was only livers and the CTVs, and the number of CTVs (virtual CTVs + actual CTV) was 200. The trained models by different parameter K were tested by 10 actual patients' data. The appropriate parameter K was selected at the convergence of MRE with the parameter K.

5) Dose prediction of 6 D_{mean} simultaneously or separately

We also investigated which is better to use 6 different plans simultaneously in one model (simultaneous model) or separately in 6 different models (Separate model).

1.2.4 Dose Planning and Calculation of the liver D_{mean}

As the dose calculation algorithm, a simple 2-dimensional percentage depth dose program for PBT was used in the in-house RTP in this study as follows.

A simple analytical approximation for depth dose distribution of a spread-out-Bragg-peak (SOBP) proton beam is introduced by Bortfeld (Bortfeld and Schlegel, 1996) and used in this study. The SOBP is achieved by superposition of elementary Bragg peak depth-dose curves.

The equation of depth-dose curve of a monoenergetic broad beam at depths between d = 0 cm and $d = R$ cm, $D_{BP}(d)$, can be approximated as in the following equation. The Bragg curve, $D_{BP}(d)$, is in units of dose per incident particle fluence ($Gycm^2$) as shown in equation (1).

$$D_{BP}(d) = -\frac{1}{\rho}\frac{dE}{dd} = \frac{1}{\rho p \alpha^{1/p}(R-d)^{1-1/p}} \quad (1)$$

where, ρ is the density of the body and p is an exponent of the range-energy relation,

and equal to 1.5. The proportionality factor, α, is equal to 1.9e-3. The Bragg peaks at different positions R is superimposed to the SOBP with weighting factors, W(R) for the Bragg peak using equation (2). The superposition of Bragg peaks results in a flat SOBP of height D_0 within an interval [d_a, d_b]. The SOBP resulting from convolving the individual Bragg peak with the weighting function is expressed using equation (3). **Figure 4** shows the relative depth-dose distribution assuming D_0 as 1.0.

$$W(R) = \begin{cases} \rho D_0 \frac{p \sin\left(\frac{\pi}{p}\right) \alpha^{1/p}}{\pi (d_b - R)^{1/p}} & : for\ d_a \leq R < d_b \\ 0 & : for\ R < d_a\ or\ R > d_b \end{cases} \quad (2)$$

$$D_{SOBP}(d) = \int_{d_a}^{d_b} W(R) D_{BP}(d, R) dR \quad (3)$$

The estimation of 3D D_{mean} of CTV and OARs from 2D D_{mean} uses simple mathematics. For the calculation of the D_{mean} as the 3D D_{mean}, the following equation (4) is used.

$$3DD_{mean} = \sum_{i=1}^{T} 2\text{DDmean(i)} \times \text{F(i)} \quad (4),$$

where T is the total number of CT slices which contain the CTV or OAR of interest, and $F(i)$ is the ratio of i-th slice volume (slice area times slice thickness) to the summation of that of the T slices. The $2DD_{mean(i)}$ is the mean dose of the ROI of the CTV or OAR in the i-th CT slice.

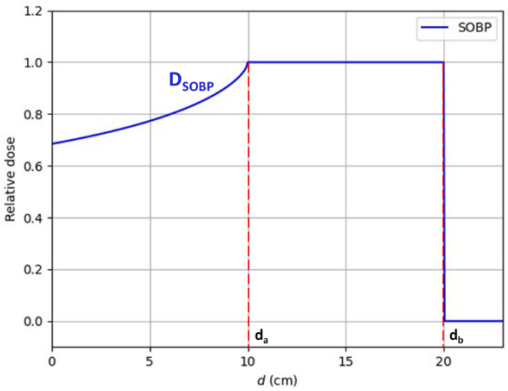

Figure 4. Relationship between the relative dose and depth in the case that [d_a, d_b] is equal to [10, 20] cm generated by equation 1-3.

All of the densities in the body were assumed to be 1.0. No consideration was given to the lateral spread in the calculations. The planning treatment volume (PTV) margin was 1.0 cm for the CTV in the transaxial plane with no margin for the cranio-caudal direction, for simplicity. The size and shape of fields and energies of PBT were determined to cover the PTV. The "2D D_{mean}" for the CTV and 3OARs in each CT slice were calculated from the differential dose volume histogram (dDVH) using the in-house RTP. The 3D $D_{mean,}$ for the CTV and 3OARs in the 3D space were calculated using a simple estimate from the 2D planning as described above.

For simplicity, 6 different combinations of 2-portals were assumed to be pre-determined as the treatment plan protocol for the liver PBT in this study. Plan A is anterior-posterior (AP), plan B is anterior-right (AR), plan C is anterior-left (AL), plan D is posterior-right (PR), plan E is posterior-left (PL), and plan F is right-left (RL), respectively (**Figure 5**).

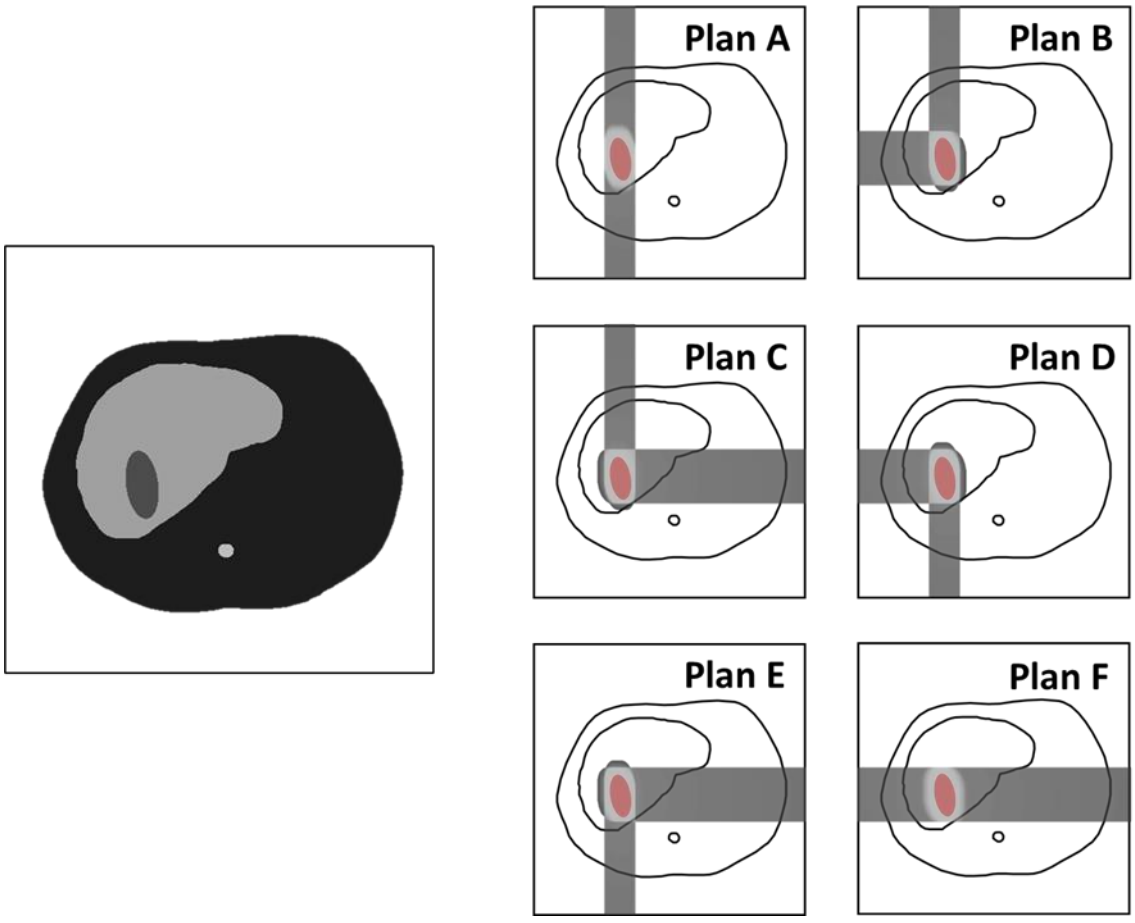

Figure 5. The dose distribution in a slice of CT contour in 6 different plans, Plan A-F, generated by in-house RTP.

Thus, 6 sets of dose distributions were generated for one set of CTV after CDA and 3OARs. The mean and standard deviations of the planned 3D $D_{mean,}$ for CTVs and for the normal liver (liver – CTV), were 72.40±1.71 Gy(RBE) and 5.11±4.67 Gy(RBE) respectively in the 10 patients for tests. The mean and standard deviations of the planned 3D $D_{mean,}$ for CTVs and for the normal liver (liver – CTV), were 73.43±1.40 Gy(RBE) and 4.63±4.36 Gy(RBE) respectively in the 7 patients for training or validation.

1.2.5 Cross-section Profiling before DL

To extract the features of the CTV and OARs in terms of its location, size, and shape, the following procedure is performed before the deep learning. We used only CT images where tumors are available in deep learning. The first pixel in the first column and row in each CT slice is defined as the location (0,0) and the last pixel in the last column and row is the location (n, n) in the CT slice. A two-dimensional region of interest (ROI) in each CT slice is used as the data set (**Figure 6)**. To extract the feature of an ROI, we used horizontal and vertical segmentation of the ROI. In either segmentation, the ROI is divided into K segments with uniform intervals. Each segment of the ROI in the CT slice is expressed by the centroid location (G_x, G_y) in the CT slice and the corresponding partial area (A) of the ROI in the segment. If we have used 10 as the number for K in this study, the matrix size for one ROI in the CT slice is, therefore, (number of ROI) x 2 (horizontal and vertical) x 3 (G_x, G_y, A) x K = number of ROI x 60. We term the extracted features of CTV and OARs in each CT slice as 2-dimension profiling (2DP).

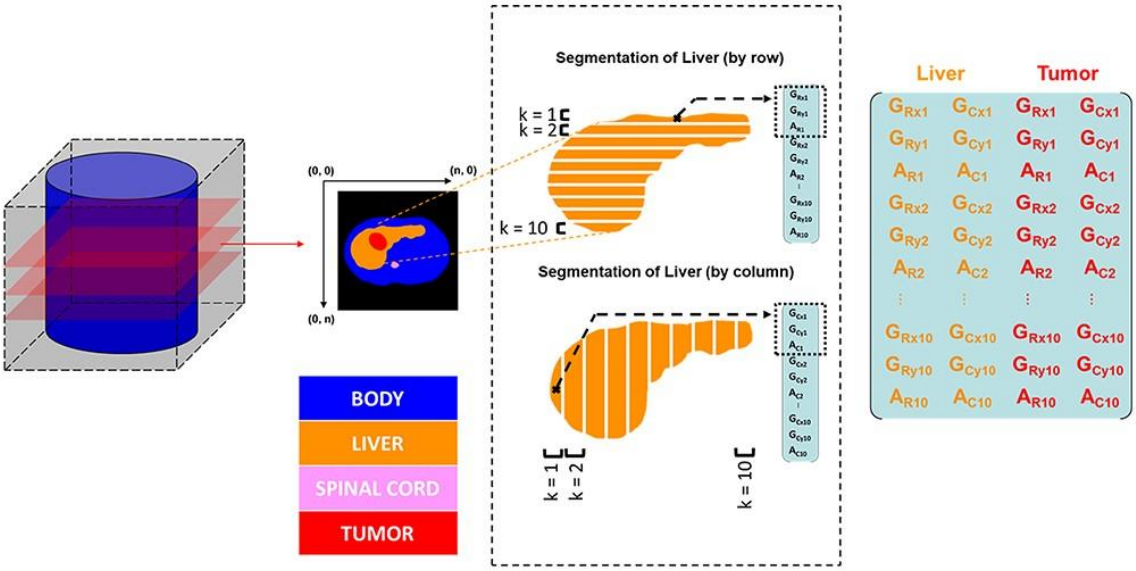

Figure 6. Profiling of the ROI from a CT slice. k: the total number of segmentations for profiling, G_{Rxi}; The x-coordinate of the gravity center of the ith segment of liver in the horizonal segmentation. G_{Ryi}; The y-coordinate of the gravity center of the ith segment of liver in the horizonal segmentation. A_{Ri}; The corresponding partial area of the ith segment of liver. G_{Cxj}; The x-coordinate of the gravity center of the jth segment of liver in the vertical segmentation. G_{Cyj}; The y-coordinate of the gravity center of the jth segment of liver in the horizonal segmentation. A_{Cj}; The corresponding partial area of the jth segment of liver. The parameters i and j are from 1 to 10 when we use k = 10.

1.2.6 Data Labeling

Labeling of the data set was performed before the training and validation of the CNN (**Figure 2**). After the 3D liver D_{mean} is calculated using the in-house RTP, the data set for training and validation is labeled with the planned 3D liver D_{mean}. We used the data sets of the ROIs in all CT slices for the 3D dose calculation. A total of 6 different treatment plans (A – F) were made for each CTV and the data sets were labeled with the 6 liver D_{mean}. Flowchart of preprocessing, model training, and prediction in shown in **Figure 7**.

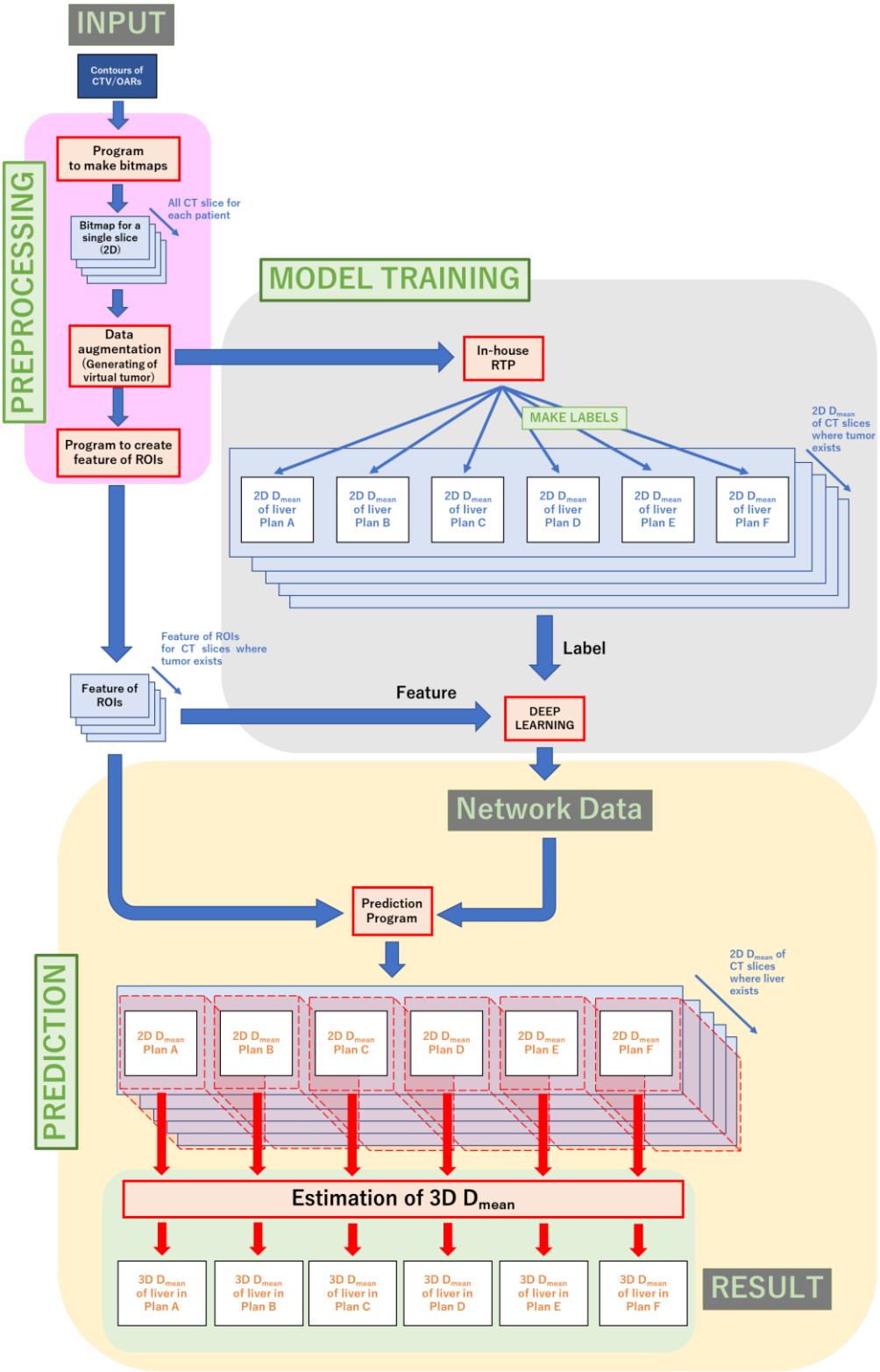

Figure 7. Flowchart of preprocessing, model training, and prediction.

1.2.7 Convolutional Neural Network

A convolutional neural network (CNN) was made on PyTorch. The input is the set of contours of the ROIs (2DP) in section 1.2.5 and the output is the set of 6 liver D_{mean} **(Figure 8)**. No additional information including none for the beam arrangement or the Hounsfield units of the CT for planning were given during the training of the CNN. The CNN architecture is based on a basic CNN compound which has been used for digit classification (Gu *et al.*, 2018). The ADAM optimizer ($\alpha = 10^{-4}$, $\beta 1 = 0.9$, $\beta 2 = 0.999$) with a weight decay of 0.0001 was used for the training (Kingma and Ba, 2015). Rectified linear units (ReLU) were used as the activation function (Nair and Hinton, 2010). The mean square error (MSE, L2 loss) was selected as the loss function. All networks were trained using an NVIDIA Quadro P5000.

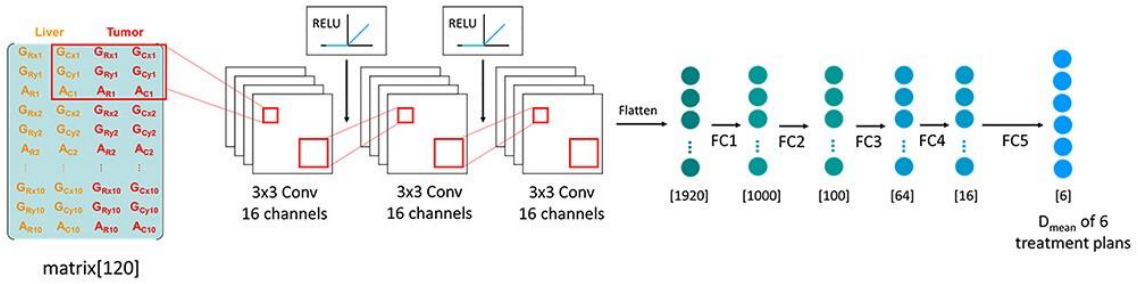

Figure 8. Illustrated outline of the CNN. Conv: convolution layer, FC1 – FC 5: Fully-connected layer 1 to 5, D_{mean}; liver mean dose.

The maximum epoch number was set to 200 epochs with an early stopping strategy. The early stopping strategy lets the training process continue as long as the training and validation error is decreased. The appropriate model was saved at the epoch, where the lowest validation loss occurred, to avoid overfitting.

1.2.8 Measurement of accuracy and time required to predict using SDP

The accuracy of the predictions of the liver D_{mean} using the validated model were tested by using independent data sets of 10 actual patients. The number of unlabeled data sets was 10 for each beam combination in the test. The accuracy of the predicted 3D liver D_{mean} was investigated by comparison with the liver D_{mean} which had been calculated using the in-house RTP. We termed the 3D liver D_{mean} calculated using the in-house RTP as the planned liver D_{mean} (ground truth) in this study. The mean relative error (MRE) is used as a measurement of the discrepancy between the predicted liver D_{mean} and the

planned liver D_{mean} which is predicted from the validated model. The regression coefficient (β) is used to describe the relationship between the predicted and planned liver D_{mean}. The p-value of a two-tailed student's t-test at a 0.05 significance level was used to assess the difference between the means of the predicted liver D_{mean} and the means of the planned liver D_{mean} of 6 different treatment plans in the 10 patients tested.

The CT images of the 10 patients in the tests were re-evaluated using VQA (Hitachi, Tokyo), which is a commercially available 3DRTP that has been used in PBT for actual patients with liver cancer in our institution. Intensity modulated proton beam therapy (IMPT) was simulated using the actual beam data from PROBEAT-RT (Hitachi, Tokyo), which is a commercially available PBT used in our institution. The CT number was used for the calculations without assuming the density as 1.0. The CTV, PTV, and prescription dose were the same as for the in-house RTP. The beam angles were selected by a physician without knowledge of the recommendation and the predicted D_{mean} by the SDP. The liver D_{mean} was calculated using the VQA and compared with that calculated using the in-house RTP and SDP for the 10 patients in the test.

We measured the time required to train the CNN for the liver D_{mean} of all of the 6 different treatment plans. We also measured the time for the inference of liver D_{mean} using the SDP.

1.3 Experiment Results

1.3.1 Investigation of appropriate parameter and data sets before DL

The appropriate parameters and data sets for deep learning are investigated in section 1.2.3 of the experimental method. The MRE was used as a measure of the error between the predicted liver D_{mean} and the planned liver D_{mean} which was predicted from the model trained by the different conditions.

1) The appropriate number of CTVs (virtual CTVs + actual CTV)

To investigate the appropriate number of CTVs (virtual CTVs + actual CTVs) for training and validation, several different numbers of CTVs, at 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 250, and 300 were used without changing the other parameters. The test data set is 10 actual patients. The MRE between the planned and predicted D_{mean} were shown on the y-axis of **Figure 9(a)** with the number of CTV on the x-axis. The result shows that the MRE was reduced and converged at the number of virtual CTV around 200. This result ensures that the number of virtual CTVs at 199 is appropriate to use with the 200 CTVs (199 virtual CTVs + 1 actual CTVs = 200 CTVs) for training. **Table 2** shows the MRE between the planned D_{mean} and the predicted D_{mean} of models trained by different number of CTVs.

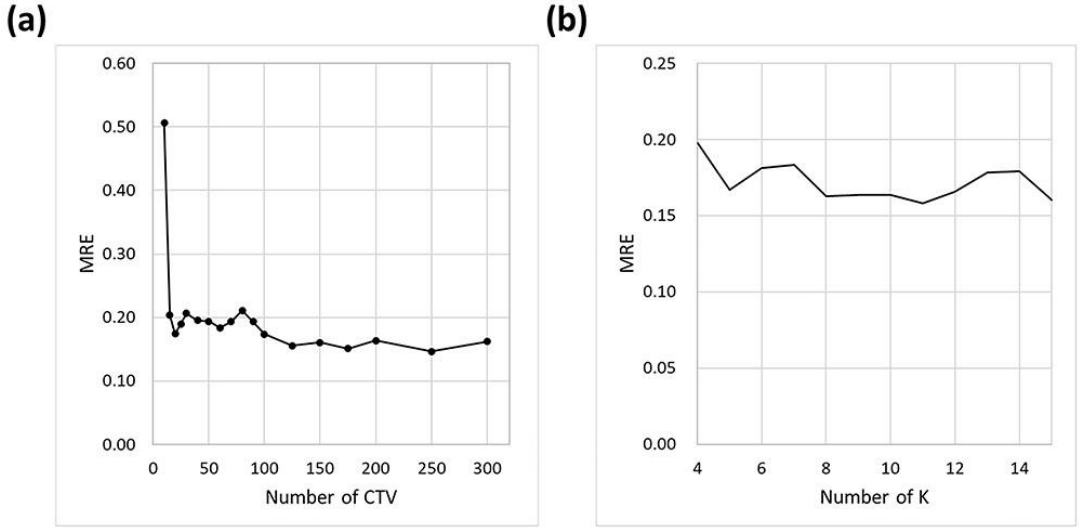

Figure 9 (a) The MRE between planned liver mean dose (D_{mean}) and predicted liver D_{mean} in model trained by different number of virtual tumors by CDA. (b) The MRE between planned liver D_{mean} and predicted liver D_{mean} in model trained by different number of parameter K.

Table 2. The MRE between planned D_{mean} and predicted liver D_{mean} in model trained by different number of virtual tumors by contour-based data augmentation.

Number of CTVs (Actual CTV + Virtual CTV)	**MRE**
10	0.5059
15	0.2033
20	0.1743
25	0.1897
30	0.2065
40	0.1954
50	0.1940
60	0.1838
70	0.1936
80	0.2112
90	0.1933
100	0.1739
125	0.1554
150	0.1608
175	0.1509
200	0.1637
250	0.1466
300	0.1622

Abbreviation: CTV = Clinical target volume, MRE = Mean Relative Error

2) The appropriate ratio of training to validation data sets

To investigate the appropriate ratio of data sets of patients for training and for validation, we changed the ratio (training: validation) from 1:6, 2:5, 3:4, 4:3, 5:2, 6:1, to 7:0 without changing other parameters. The MRE between the predicted and the planned liver D_{mean} was 0.2402, 0.2053, 0.2342, 0.1914, 0.1637, 0.1917, and 0.1766 respectively. The MRE was the lowest at 5:2. To investigate whether the predictive performance increases with more patients in the training, the number of patients in the training data set was changed from 1 to 5 keeping the number of patients for validation at 2. The MRE between the predicted and the planned liver D_{mean} of 1, 2, 3, 4, and 5 training data sets was 0.2334, 0.2143, 0.2235, 0.1959, and 0.1637, respectively showing that MRE improved with the number of training data sets. Based on this, we decided to use 5 patients for training and 2 patients for validation in the following. The patients were randomly selected for training and validation. The characteristics of the data sets for training and validation are shown in **Table 3**. The number of CT slices used for training and validation was 52 (37 of 5 patients for training and 15 of 2 patients for validation). As a result, the number of labeled data sets was 10,400 (200 CTV x 52 CT slices).

Table 3. Diameter of CTVs after data augmentation and liver volume determination. The characteristics of the data sets for the training, validation, and test.

Data sets	CTV			Liver		
	Number of patients	Number of CTVs	Equivalent diameter (cm) mean (range)	Number of patients	Number of livers	Volume (cm^3) median (range)
Training	5	1000	4.82 (1.44-11.42)	5	5	1202 (896-2036)
Validation	2	400	5.38 (1.73-12.58)	2	2	1624 (948-2299)
Test	10	10	3.31 (0.707-7.213)	10	10	2005 (637-3559)

Abbreviation: CTV = clinical target volume

3) The appropriate number of OARS

To determine the influence of the number of OARs in the deep learning, the model was trained with the following different conditions without changing other parameters; 1) 3 OARs including skin, liver, and spinal cord, 2) 2 OARs including skin and liver, 3) 1 OAR including only skin, and 4) 1 OAR including only the liver. The MRE between the planned D_{mean} and predicted D_{mean}, which was trained by different numbers of OARs, were 1) 0.1832, 2) 0.1841, 3) 0.2126, and 4) 0.1637, respectively. Since the MRE from the model trained by CTV and only liver as the OAR was the lowest, we used the model trained by only CTV and liver in the following study.

4) The appropriate number of parameter K

To find the appropriate number of K, we have trained the model with different numbers of parameter K from 4-15. The test data set is the 10 actual patients. The MRE between the planned D_{mean} and the predicted D_{mean} of each model were compared in **Figure 9(b**) and **Table 4**. The MRE not substantially different in cases where K is between 8 and 15, confirming that K at 10 is an appropriate number to be used in this model.

5) Dose prediction of 6 D_{mean} simultaneously or separately

The MRE between the planned and predicted liver D_{mean} was 0.1637 in the simultaneous model while 0.1606 in the separated models. There was no statistical difference between the predicted D_{mean} from the simultaneous model and separated models (p=0.2168). In the following analysis, we used the simultaneous prediction model which is a bit faster.

Table 4. The MRE between planned liver D_{mean} and predicted liver D_{mean} in model trained by different number of parameter K.

Number of K	MRE
4	0.1977
5	0.1671
6	0.1813
7	0.1833
8	0.1629
9	0.1636
10	0.1637
11	0.1579
12	0.1656
13	0.1782
14	0.1791
15	0.1600

Abbreviation: MRE = Mean Relative Error

1.3.2 The relationship between the predicted liver D_{mean} by SDP and the planned liver D_{mean}

The model trained by different conditions was tested by 10 actual patient data sets shown in Table 1. The number of CTVs (virtual CTVs + actual CTV) were 200 for each actual patient and the ratio of training to validation data sets was 5:2. Only CTV and liver were used in the model training. The parameter K was determined to be 10. The simultaneous model was used. Using these appropriate numbers and conditions, the trained model saved at epoch 98, in which validation loss is the lowest. The training and validation loss

during model training was shown in Figure 10.

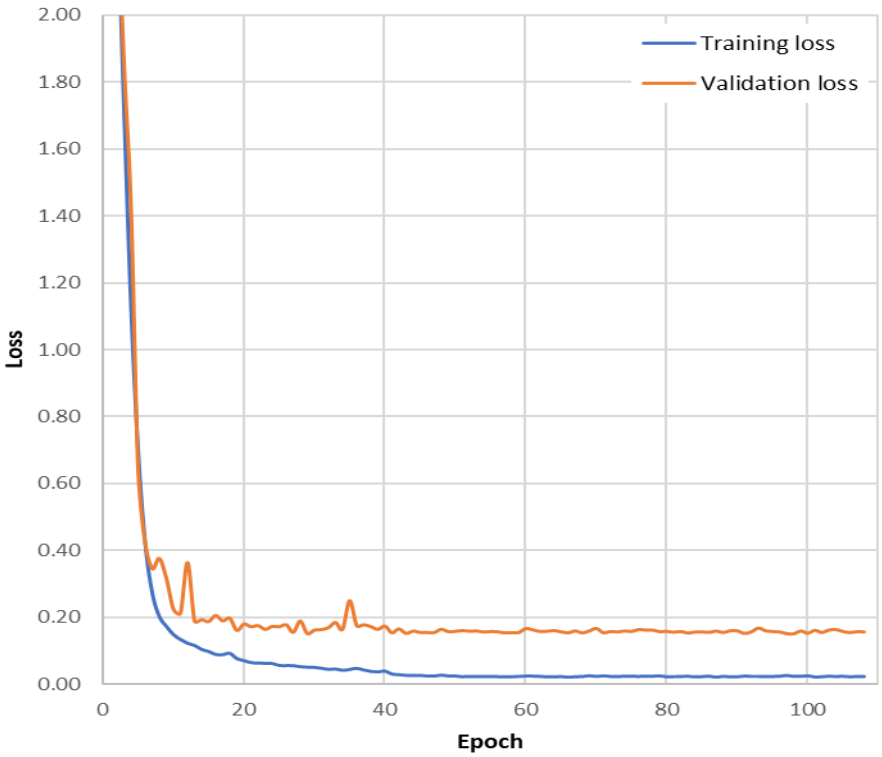

Figure 10. The training and validation loss over the epochs during model training.

Consequently, the relationship between the predicted D_{mean} by SDP and the planned liver D_{mean} calculated by in-house RTP is shown in **Figure 11 (a)** (10 patients x 6 plans = 60 points in plot)**.** The MRE between the predicted and the planned liver D_{mean} is 0.1637 and for the β it is 0.9455. There was no significant difference in the mean of predicted liver D_{mean} and the mean of the planned liver D_{mean} of the 6 different treatment plans for the 10 actual patients tested (**Table 5**). **Table 6** shows a comparison data for the liver D_{mean} calculated using the VQA, in-house RTP, and the predicted liver D_{mean} by the SDP. The predicted results using separated models is also shown in **Table 6**.

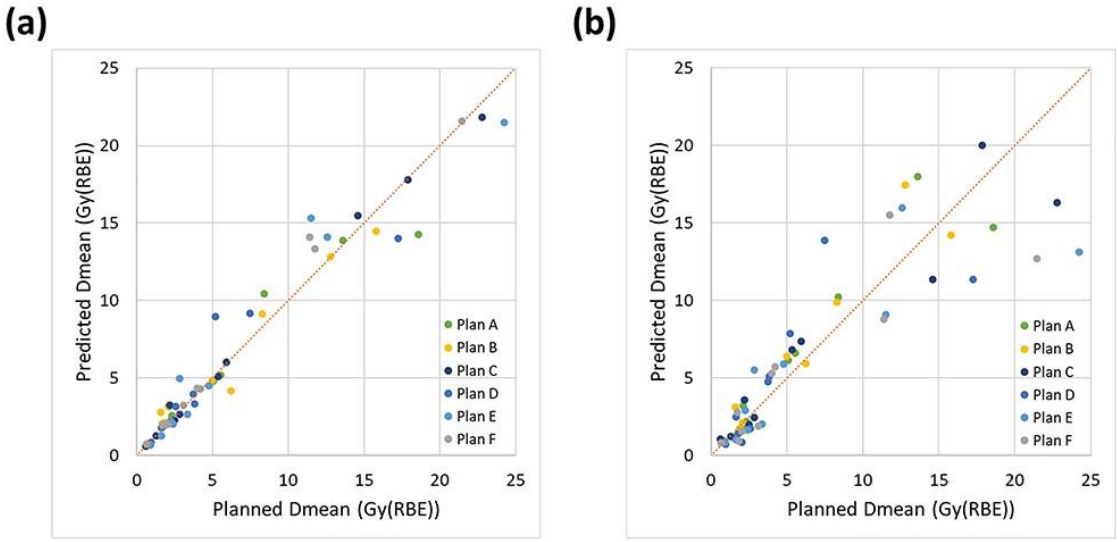

Figure 11. Plots of the planned liver mean dose (Dmean) and predicted liver Dmean with the line of identity ($\beta = 1$). (a) model trained with CDA, (b) model trained without CDA.

Table 5. Differences between the mean of the planned D_{mean} and the mean of the predicted D_{mean} of 6 different treatment plans A to F for the 10 patients in the test.

Patient	Plan	Liver D_{mean} (Gy(RBE))		ΔD_{mean} (Planned-Predicted) (Gy(RBE))	p-value
		Planned	Predicted		
Test 1	Mean	4.876	4.699	0.176	0.137
	SD	0.787	0.902	0.244	
Test 2	Mean	2.129	2.082	0.047	0.644
	SD	0.300	0.134	0.234	
Test 3	Mean	2.710	2.927	-0.216	0.429
	SD	0.469	0.326	0.616	
Test 4	Mean	12.678	13.516	-0.839	0.059
	SD	3.340	2.762	0.842	
Test 5	Mean	0.783	0.721	0.062	0.308
	SD	0.157	0.091	0.133	
Test 6	Mean	1.907	2.527	-0.620	0.038
	SD	0.301	0.625	0.544	
Test 7	Mean	20.020	17.942	2.078	0.027
	SD	3.312	4.058	1.645	
Test 8	Mean	1.661	1.654	0.007	0.952
	SD	0.253	0.420	0.282	
Test 9	Mean	4.535	4.556	-0.021	0.971
	SD	1.233	0.462	1.334	
Test 10	Mean	9.893	12.231	-2.337	0.007
	SD	3.283	3.057	1.322	

Abbreviation: D_{mean} = Mean dose, ΔD_{mean} (Planned-Predicted) = Differences between the mean of the planned D_{mean} and the mean of the predicted D_{mean}, SD = standard deviation

Table 6. The comparison among three-dimensional radiation treatment planning system commercially available, in-house radiation treatment planning system, and the simple dose prediction tool in the beam angles and liver $D_{mean.}$

Patient	Beam Angles (degree)	Prescribed dose at CTV (Gy(RBE))	Liver -GTV (Gy (RBE))			
			3DRTP (CT number)	In-house RTP (ρ = 1.0)	Prediction by SDP using simultaneous model (ρ = 1.0)	Prediction by SDP using separated models (ρ = 1.0)
Test 1	270, 180	72	4.316	3.837	3.340	4.191
Test 2	270, 180	73	2.217	1.756	1.886	1.824
Test 3	270, 0	70	3.892	2.086	3.215	3.089
Test 4	270, 180	74	7.885	7.480	9.189	9.027
Test 5	270, 0	70	1.196	0.669	0.750	0.817
Test 6	270, 0	73	2.062	1.579	2.796	2.747
Test 7	270, 0	74	14.137	15.808	14.453	13.120
Test 8	270, 0	75	2.144	1.680	2.055	1.613
Test 9	270, 0	72	5.171	6.223	4.184	4.781
Test 10	270, 0	71	7.729	8.287	9.144	9.194

Abbreviation: D_{mean} = Mean dose, CTV = clinical target volume, GTV = gross target volume, RBE = relative biological effectiveness, 3DRTP = three-dimensional radiation treatment planning system, RTP = radiation treatment planning system, ρ = density, SDP = simple dose prediction tool.

To investigate whether CDA was effective to improve the predictive performance, results without CDA were compared with those with CDA. **Figure 11(b)** plots the relationship between the planned and predicted liver D_{mean} predicted from the model trained without CDA. The MRE between the planned and predicted liver D_{mean} is 0.3211 while in the model which was trained with CDA is 0.1637. There was statistically significant difference between the MRE from model trained without and with CDA by two-tailed student's t-test ($p<0.001$). Therefore, the CDA was shown to be effective to improve the predictive performance.

The mean percentage error (MPE), which is defined as the ((predicted D_{mean} - planned D_{mean})/ planned D_{mean})*100, has also been evaluated. The MPE between the planned liver D_{mean} and the predicted liver D_{mean} is 6.60% ± 23.52%.

The time required for training and validation of the CNN was 1 hour and 23 minutes for all of the 6 treatment plans. The mean time required for the inference of the liver D_{mean} of the 6 different treatment plans for a patient using the SDP was 4.47±0.09 seconds in 10 actual patient data sets in the tests as shown in **Table 1**.

1.4 Discussion

We have generated a CNN for the prediction of the liver D_{mean} in 6 different treatment plans from the contouring of only CTV and liver. The propose of the model is to predict D_{mean} of the liver which is a regression problem. The CNN architecture is flexible for both regression and classification problems. And once the model is trained, the prediction time is extremely short. There are numerous variants of CNN architectures. However, the basic components of these are similar to each other. We follow the well-known principle of 'Occam's razor" in which simpler theories are preferable to more complex ones but there could be more sophisticated architectures of CNN which predict more accurately.

The time required to predict the D_{mean} by the SDP was short enough to be of use in the clinic. The time to do the contouring of the CTV and 3OARs will depend on the tool being used in a clinic. We have not detailed this part further in this study because it is outside of scope of the study. Conventional tools for manual contouring for CT images in the DICOM format are widely available at reasonable cost. Using sophisticated tools for automatic contouring would be quicker but more expensive. A clinic can decide what kind of contouring tool is to be used depending on the needs of the clinic.

Studies using machine learning in the selection of PBT are becoming more common. Kouwenberg et al. have proposed the use of machine learning and automated treatment planning for pre-selection of patients for IMPT. It was found to reduce the volume of a formal workload-intensive model based selection procedure with a negative outcome by 67% (Kouwenberg *et al.*, 2021). They concluded that such a system could help to screen larger patient populations and avoid unnecessary delays in the start of radiotherapy for head & neck cancer patients. They used 45 patients and trained a Gaussian naïve Bayes classifier. Although they achieved a large reduction in the workload in the selection process of IMPT, the number of data sets in their study may still be too small and the excellent prediction may have been specific to the institutional protocol. Recently, Guerreiro et al. have used a CNN model to predict the D_{mean} of CTV and OARs by PBT for pediatric abdominal tumors (Guerreiro *et al.*, 2021). They included 80 patients: 48 patients for training, 12 patients for validation, and 20 patients for testing with a 5-fold cross validation procedure. They reported that the MPE between the planned and predicted D_{mean} of all OARs was - 0.3% ± 2.9% a variation which is superior to our results.

However, they need precise 3D volumes and 3D dose distributions as the input for the CNN, something that requires a very large data volume, high performance computer, and large resources. Our approach requires only 2D contours of the liver and CTV as the input, it requires fewer resources, and is cost-effective to predict liver D_{mean} directly. Since there are no compatible approaches with the same cost-effectiveness as ours, our approach will likely become regarded as a benchmark in this category. If only gross estimations of the liver D_{mean}, is required for the selection of PBT by the cancer board, we think our approach is reasonably accurate as can be seen in Figure 11(a). Chen et al. have published a deep learning prediction of patient-specific DVH of IMXT for nasopharyngeal cancer using 153 cases for training and 27 cases for testing (Chen *et al.*, 2021). Except for the optic organs, their CNN model performs better than or is comparable to conventional 3DRTP with the mean difference in the proportion of points of interest 3.59% ± 7.78% with 14.5 hours for the training of 16 volume of interests and less than 1 minute to generate DVHs for one patient. The excellent results of these studies are encouraging but again the precise prediction methods require the creation of large and tumor-specific databases to account for multicenter treatment planning diversity.

The shortcomings of this study are as follows. We used a simple 2D treatment planning of PBT and not the 3DRTP which is used for actual patients. The benefit in simplicity and cost of the SDP can be attributed to the simple RTP, at least to some extent. The dose distribution must be changed by the assumption of the density as 1.0 throughout the body in in-house RTP. We compared the accuracy of the in-house RTP and commercially available 3DRTP and found that the difference in D_{mean} is acceptable as long as the purpose of the SDP is for an approximate estimate to select the PBT as a possible treatment of choice in a busy clinic. However, further improvement is obviously required for more precise predictions of the liver D_{mean}. Secondly, the liver may be one of the most straightforward OAR to apply with our concept of SDP. For other treatment sites such as the head and neck, the accuracy of the prediction may be lower since the relationship between CTV and OARs are more complicated. Thirdly, from the viewpoint of the physician, because the contouring of CTV and OARs still requires a significant amount of time, the method that does not require contour information is preferable. Auto-contouring can be a key technology for this residual problem since a recent study showed

that the average time was reduced from 108 minutes in manual contouring to 10 minutes by auto-contouring ($p < 0.001$) (Fung *et al.*, 2020). Also, it is well known that tools based on deep learning could have an overfitting problem. In this study, the data set was spilt into training and validation data sets to detect possible overfitting and the model with the lowest validation error was selected. We have not used cross-validation but that could have made the prediction better.

To summarize, we developed a SDP to predict a liver D_{mean}, in which a physician is only required to do contouring of the CTV and liver in the clinic, and assessed its accuracy and usability. The SDP is cost-effective and usable for an approximate estimation of liver D_{mean} in the clinic, however, the accuracy should be improved further if we need the accuracy of liver D_{mean} to be compatible with 3DRTP. The same concept as our SDP can simply be applied to the dose distribution of STI and IMXT. After the development of the CNN for these treatments, we will be able to compare the D_{mean} or other dose volume statistics of the PBT and XRT. It will be meaningful to examine whether the D_{mean} can be changed to other dose volume statistics such as the D_{max} of the OARs which have a serial structure or the D_{min} of the CTV, depending on the requirements in different situations.

Chapter 2. Attempts to improve the SDP

2.1 Introduction

We have developed the SDP tool in Chapter 1 and published. Because the speed of improvement in the DL is rapid, we have continuously attempted to improve the accuracy and speed of the SDP further.

One idea is to change the cross-section profiling method before DL. In chapter 1, the independent two-dimensional CT images are used to train the model in deep learning. The input of model was 2DP which were explained in section 1.2.5 of chapter 1. However, the three-dimensionally reconstructed CT for profiling may be an option to improve the accuracy of prediction model. In this chapter, the 3D cross-section profiling (3DP) method was constructed and investigated about its effectiveness.

Another idea is to use architecture of pre-trained models in the deep learning. In the chapter 1, we used a rather, but the accuracy may be improved if we used the contour image with more complex architecture of prediction model. In this chapter, the pre-trained models used are Alexnet, VGG-net, GoogLeNet, Inception net V3, Resnet, Wide Resnet, Densenet, and Shufflenet V2, which are neural network for image classification in Pytorch. We applied the contour-based data augmentation (CDA) approach, which were the generation of virtual CTV before model training in DL, to pre-trained model-based prediction model training. The feature to train the model is contour bitmap image of CTV and 3 OARs (skin surface, liver, and spinal cord) and corresponding liver D_{mean} are used as labels for the data sets.

2.2 Experimental Method

2.2.1 Three-dimension profiling (3DP)

We call the profiling used in chapter 1 as two-dimensional profiling (2DP) and the profiling using three-dimensionally reconstructed CT as three-dimensional profiling (3DP). In 3DP, CTV and OARs in 3D space are used for profiling. In the 3D space, the first voxel in the first column, row and layer is defined as location (0, 0, 0) and the last voxel was defined as (n, n, m), where n and m represent the side length and the total thickness in mm of the CT images. We term the CT images to be used for the profiling as CTOI (CT of interest). The CTOI contains only CT slices where liver exists. The CTV and OARs in the CTOI are divided into J cross-sections perpendicular to cranio-caudal direction with similar intervals. When the number of contours of the CTV or OARs is equal or more than J, the contours are grouped into J groups of cross-sections. When the number of contours of the CTV or OARs are fewer than J, virtual cross-sections are created by interpolation. We have termed the CTV and OAR in the CTOI as the volume of interest (VOI) in this study.

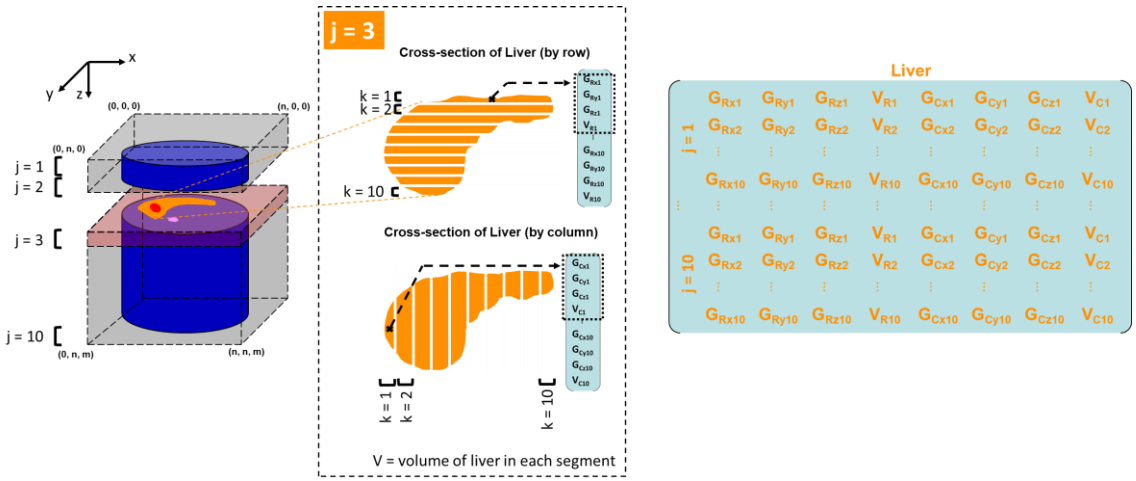

Figure 12. The matrix of 3DP of an OAR, liver (J = 10, K=10). The example segmented image of liver at j = 3 was shown.

The **Figure 12** shows the 3DP of liver. A cross-section of the VOI is divided into K segments with uniform intervals with 2 types of segmentation: one horizontal and the

other vertical. To do profiling of one VOI in the CTOI in 3D space, 2(horizontal and vertical) x K x J segments are generated. We have used 10 as the number for K as well as for J in this study. In the 3DP, each segment of the cross-section of a VOI can be expressed by the centroid location (G_x, G_y, G_z) of the segment in the 3D space and the corresponding partial volume of VOI in segment (V).

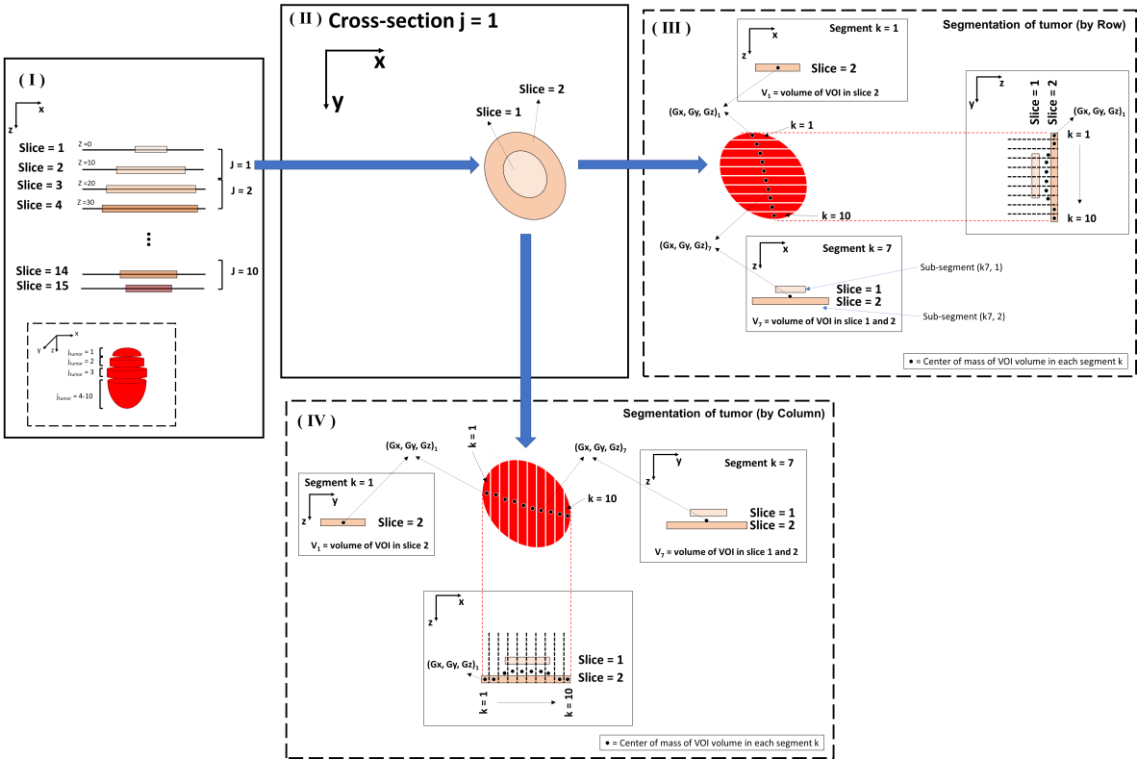

Figure 13. The segmentation process in 3DP for virtual tumor (J=10, K=10). See test.

The segmentation process in 3DP is a little complexed one and illustrated in **Figure 13**. **(I)** Assuming that the contours of the VOI exist in CT slices from 1 to 15 of the CTOI with a slice thickness of 10 mm. The CT slices are group into J cross-sections, which equals to 10 sections in this study. **(II)** The cross-section j=1 consists of the corresponding partial volume of the CTV in CT slice = 1 and 2. **(III)** The calculation for the center of the segment (G_x, G_y, G_z) and the corresponding partial volume (V) of the CTV of the segment is explained for cross-section j=1 as an example. The view is from the top of z-axis of cross-section j=1 of the CTV. Cross-section j =1 is divided into K segments horizontally (K=10). The corresponding part of the CTV in the CT slice=1 and that in CT slice=2 is divided horizontally to be the sub-segment. At the segment [j=1, k=1], the

corresponding sub-segment is in CT slice 2 but not in CT slice 1. The center of this segment and the corresponding partial volume (V) of CTV is calculated using the sub-segment of CTV in slice 2. On the other hand, at segment [j=1, k=7] as an example, the corresponding sub-segments are in both CT slice 1 and CT slice 2. Then the centroid location of this segment is the center of the mass calculated from the sub-segment in CT slice 1 (k=7, 1) and the sub-segment in CT slice 2 (k=7, 2). The corresponding partial volume (V) of CTV of this segment (j=1, k=7) is calculated from a summation of the sub-segments in slice 1 and in slice 2. **(IV)** Cross-section j=1 is also divided vertically into k segments and treated similar to **(III)**.

The detailed flowchart of the preprocessing, model training, and prediction using 3DP is shown in Figure 14.

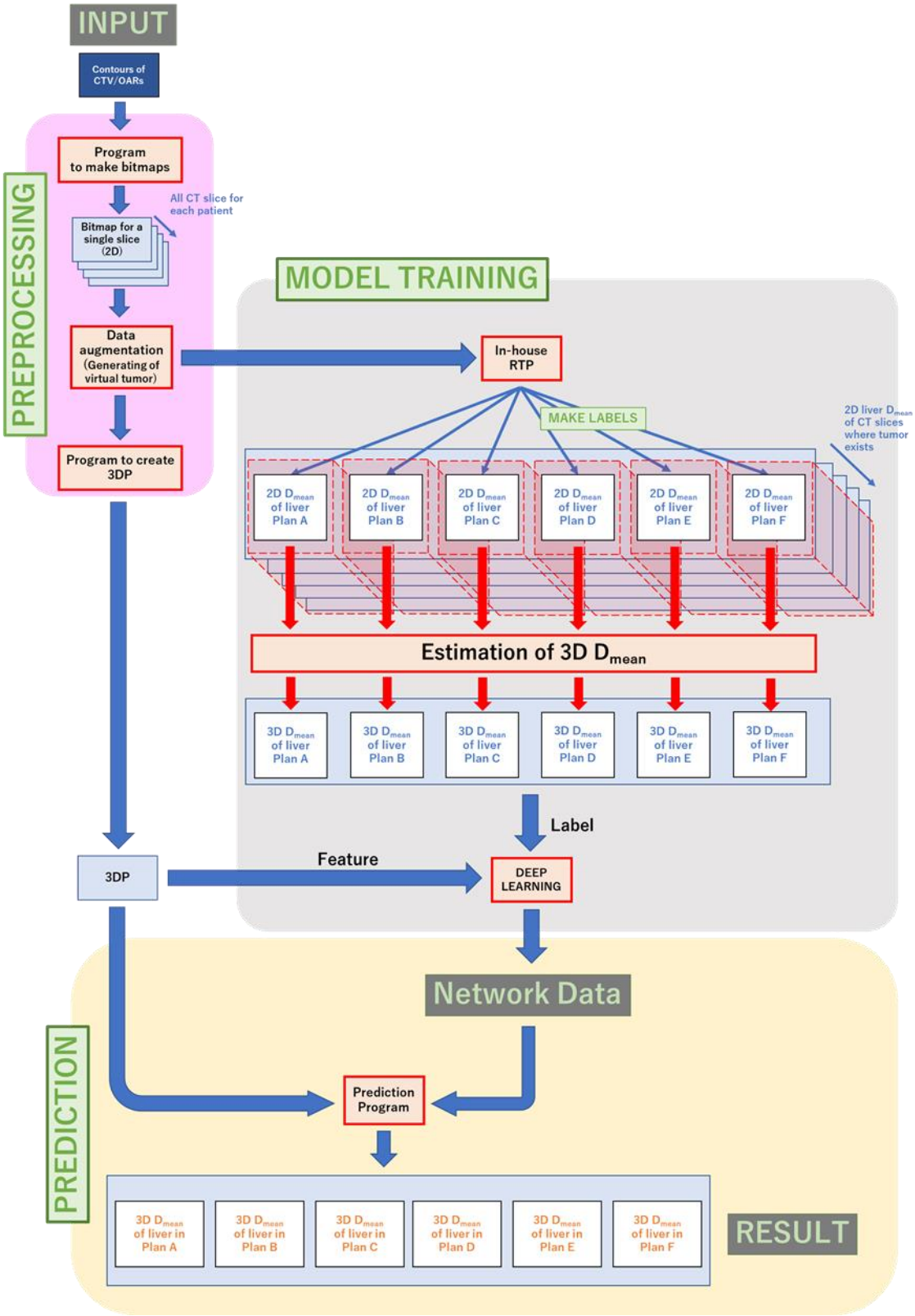

Figure 14 Flowchart of the preprocessing, model training, and prediction using 3DP as input.

2.2.2 Deep learning with architecture of pre-trained models

The pre-trained models for image classification in Pytorch package, including Alexnet (Krizhevsky, 2014), VGG-net (Simonyan and Zisserman, 2015), GoogLeNet (Szegedy *et al.*, 2015), Inception net V3 (Szegedy *et al.*, 2016), Resnet (He *et al.*, 2016), Wide Resnet (Zagoruyko and Komodakis, 2016), Densenet (Huang *et al.*, 2017), and Shufflenet V2 (Ma *et al.*, 2018), were used as the D_{mean} prediction model. The last layer of models was fine-tuned to be fully connected layer with 6 outputs, which were liver D_{mean} of 6 different treatment plans. A conceptual flow chart to utilize the SDP with pre-trained model architecture is shown in **Figure 15**. **Figure 16-23** show the architecture of liver D_{mean} prediction model based on each pre-trained model. The loss function was changed from cross-entropy in the original model, which is appropriate for classification problem, to the mean square error (MSE, L2 loss), which is appropriate for regression problem. The ADAM optimizer ($\alpha= 10^{-4}$, $\beta1= 0.9$, $\beta2= 0.999$) with a weight decay of 0.0001 was used for the training (Kingma and Ba, 2015). The input of pre-trained model-based architecture prediction models is two-dimensional contour bitmap images, which were generated by the DICOM-RT structure of the CTV and OARs in SDP tool. In the preprocessing process, the contour bitmap images were resized to 224x224 and were normalized for model training. However, for Inception net V3, the contour bitmap images were resized to 299x299, as the requirement of the model input size. The number of virtual CTVs by CDA were 199 for each actual patient in training and validation process. **Figure 24** shows the detailed of flowchart of the preprocessing, model training, and prediction of pre-trained model-based prediction model.

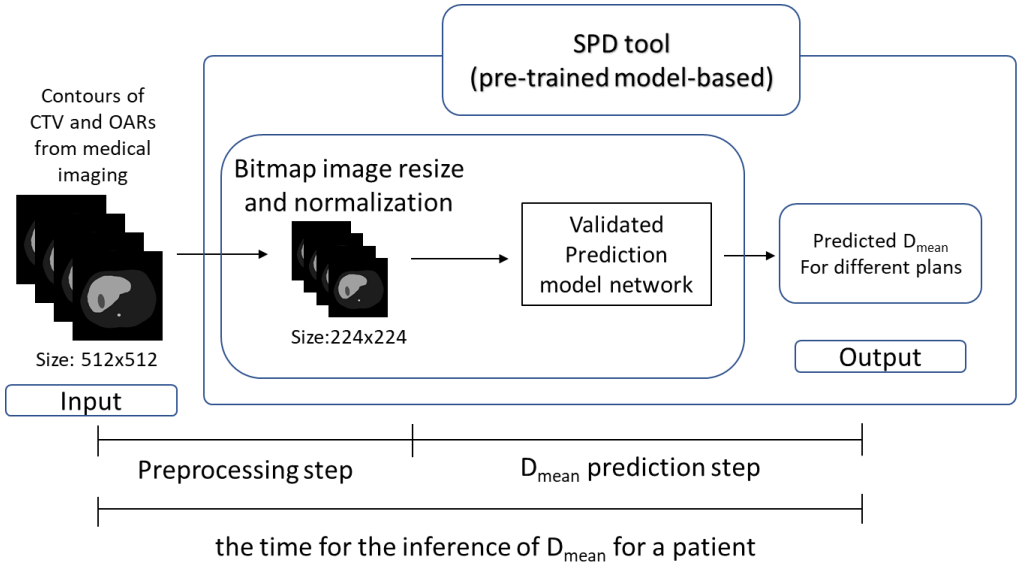

Figure 15. Flowchart of simple dose prediction (SDP) tool with pre-trained model-based structure of liver mean dose (D_{mean}) prediction model.

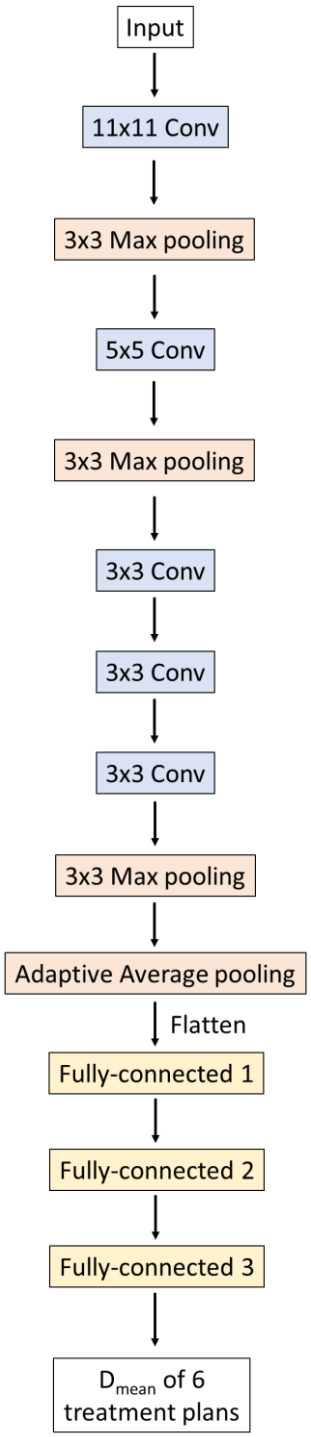

Figure 16. Alexnet-based liver D_{mean} prediction model architecture

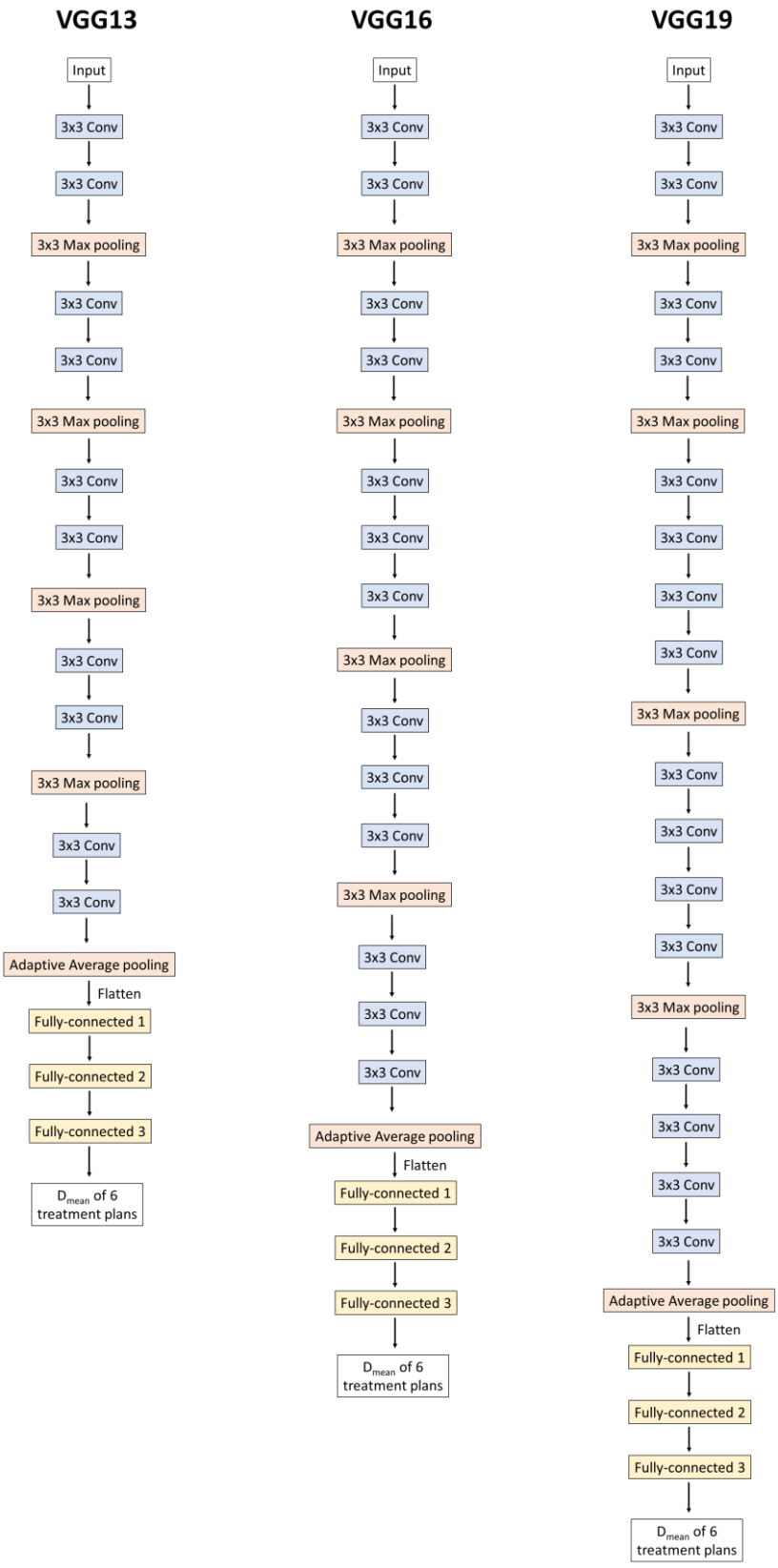

Figure 17. VGGnet-based liver D_{mean} prediction model architecture

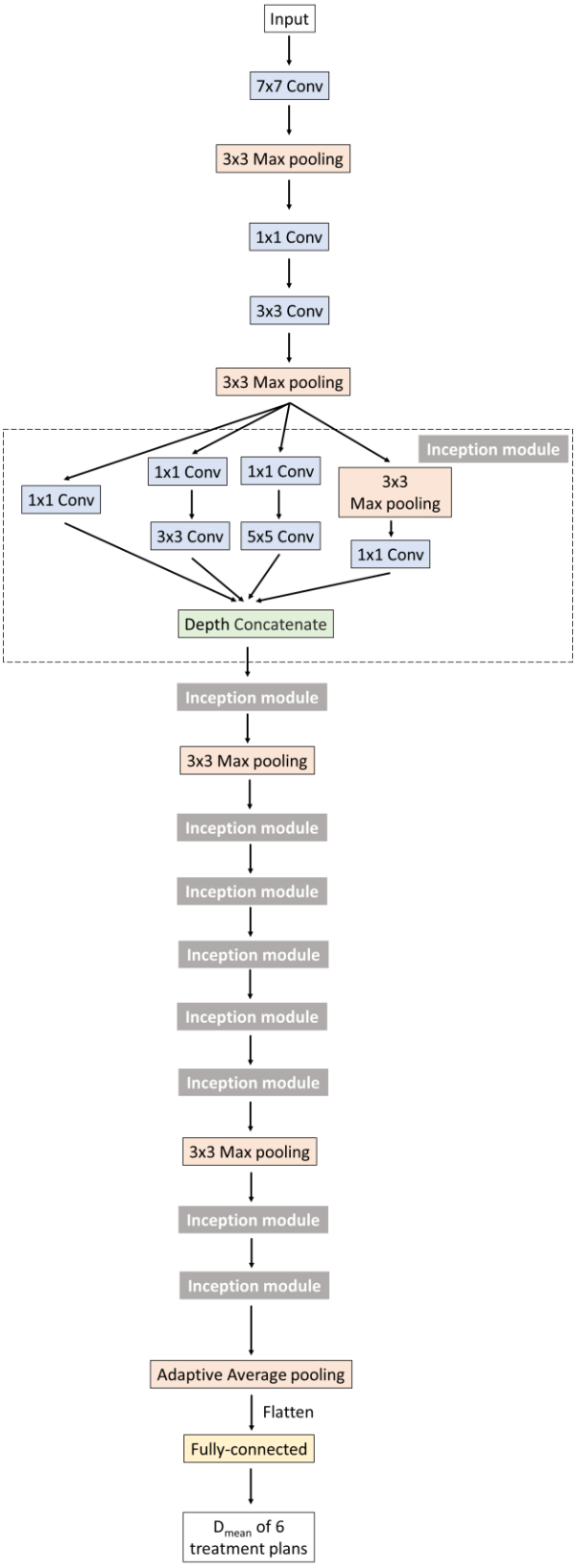

Figure 18. GoogLenet-based liver D_{mean} prediction model architecture.

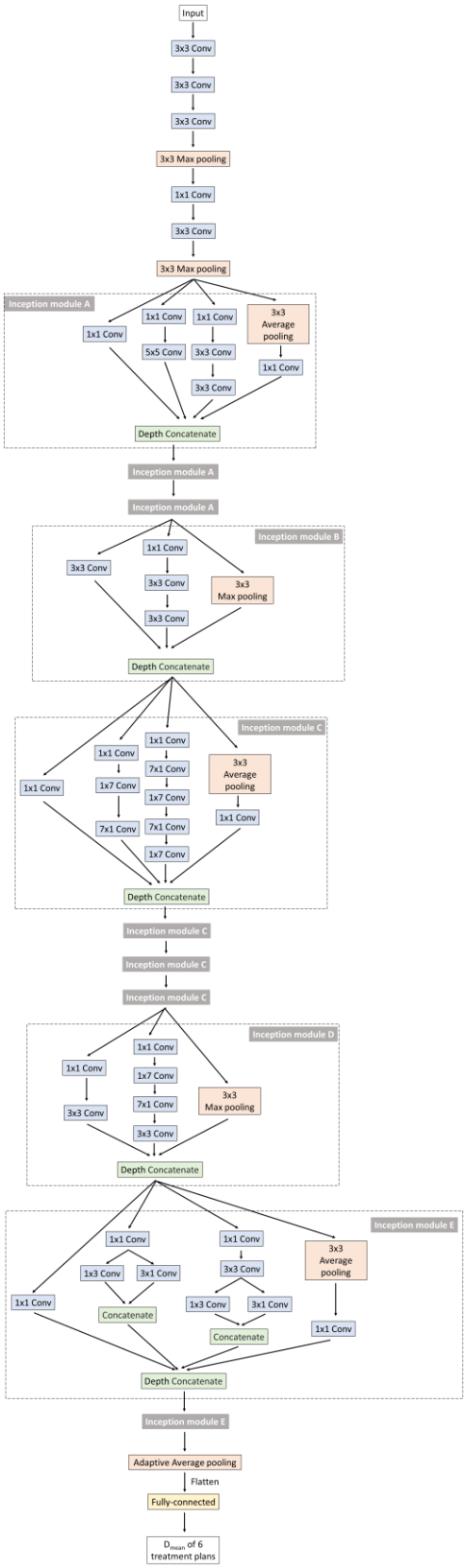

Figure 19. Inception net V3-based liver D_{mean} prediction model architecture

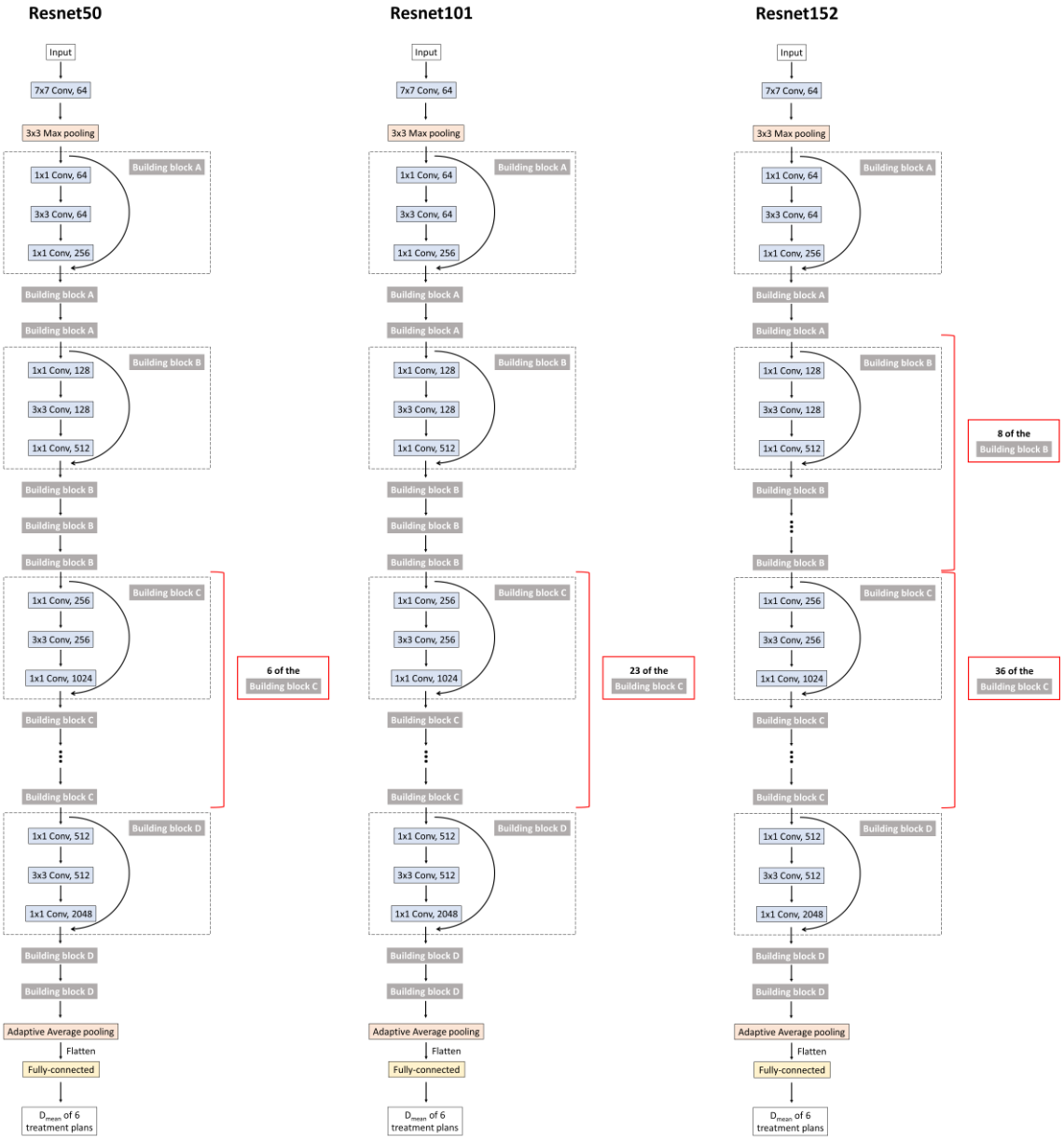

Figure 20. Resnet-based liver D_{mean} prediction model architecture

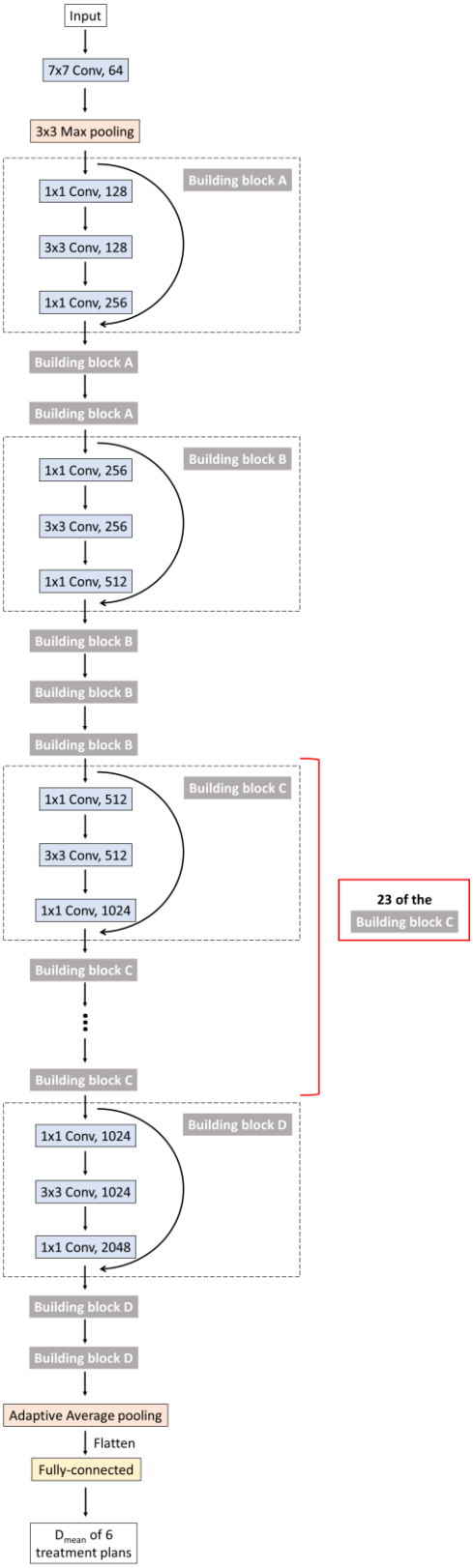

Figure 21. Wide Resnet-based liver D_{mean} prediction model architecture

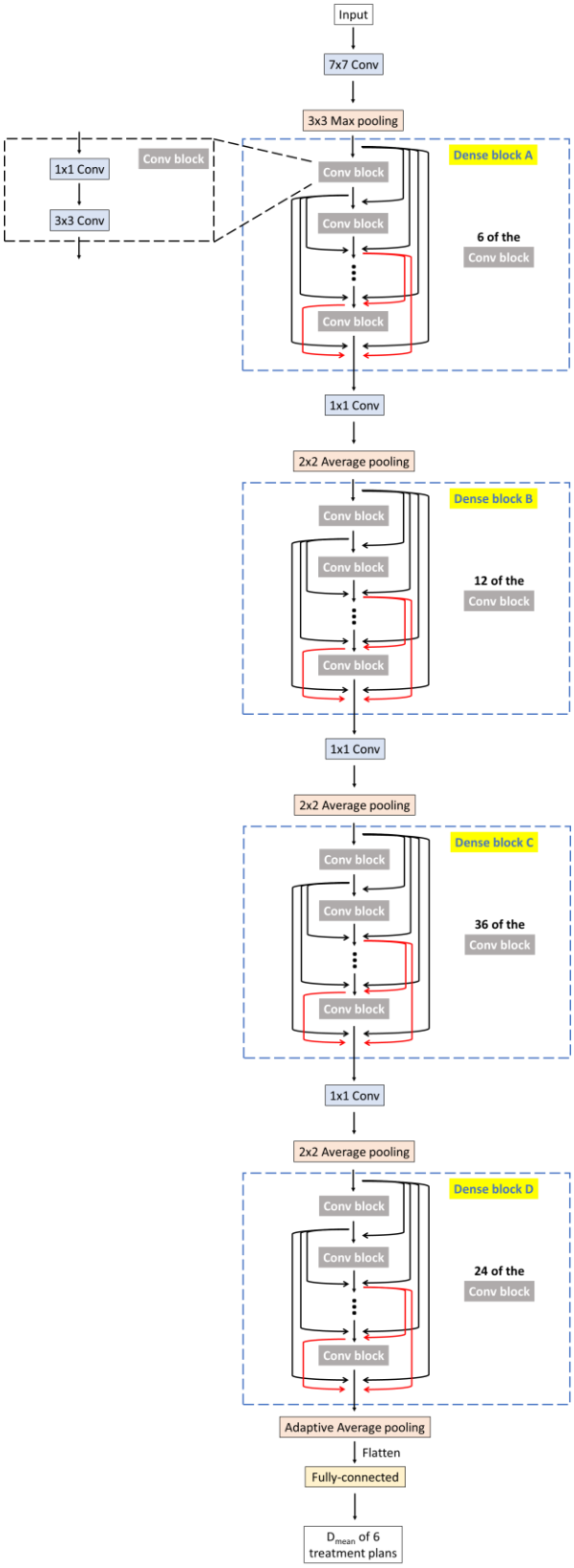

Figure 22. Densenet-based liver D_{mean} prediction model architecture

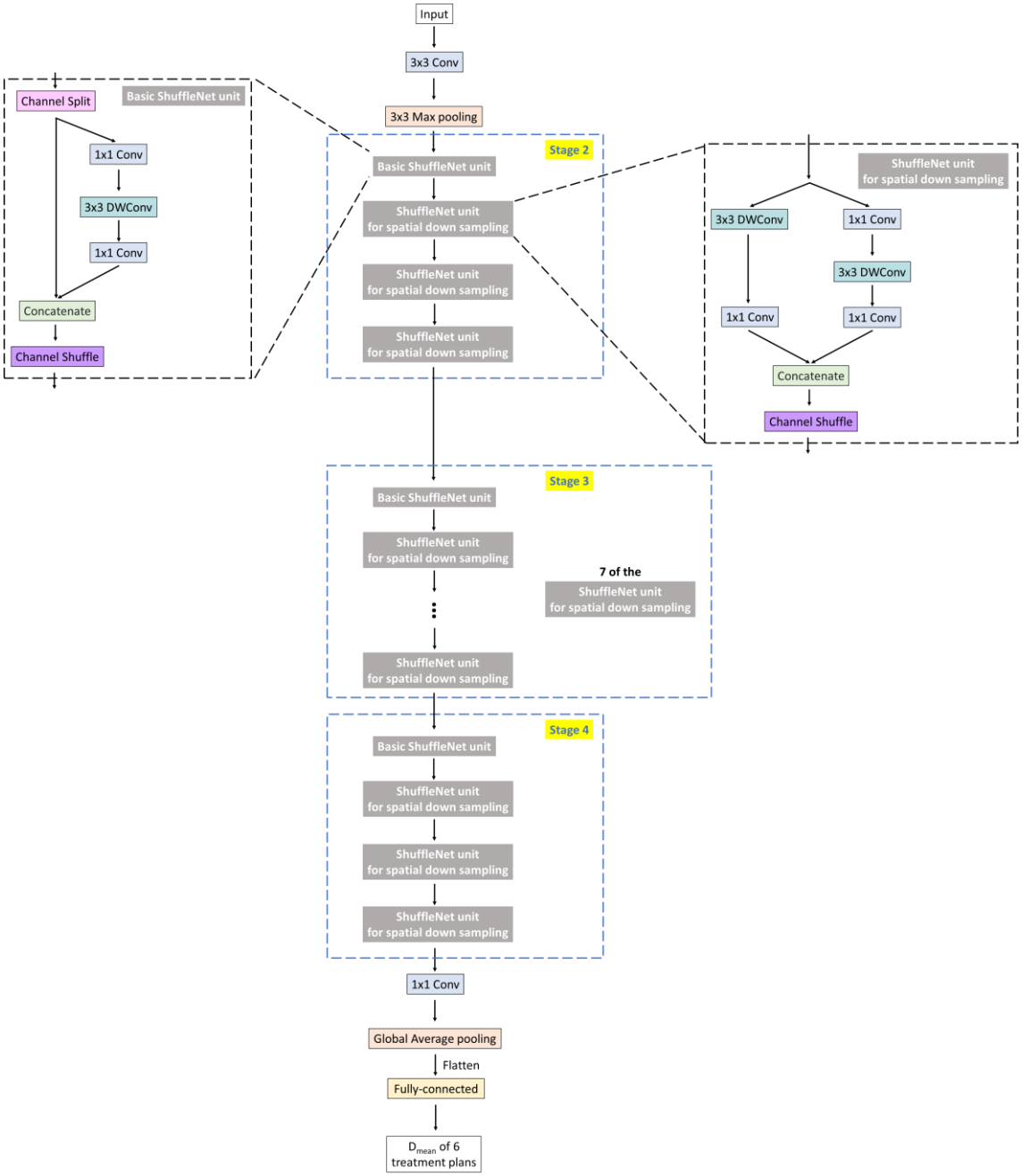

Figure 23. Shufflenet V2-based liver D_{mean} prediction model architecture.

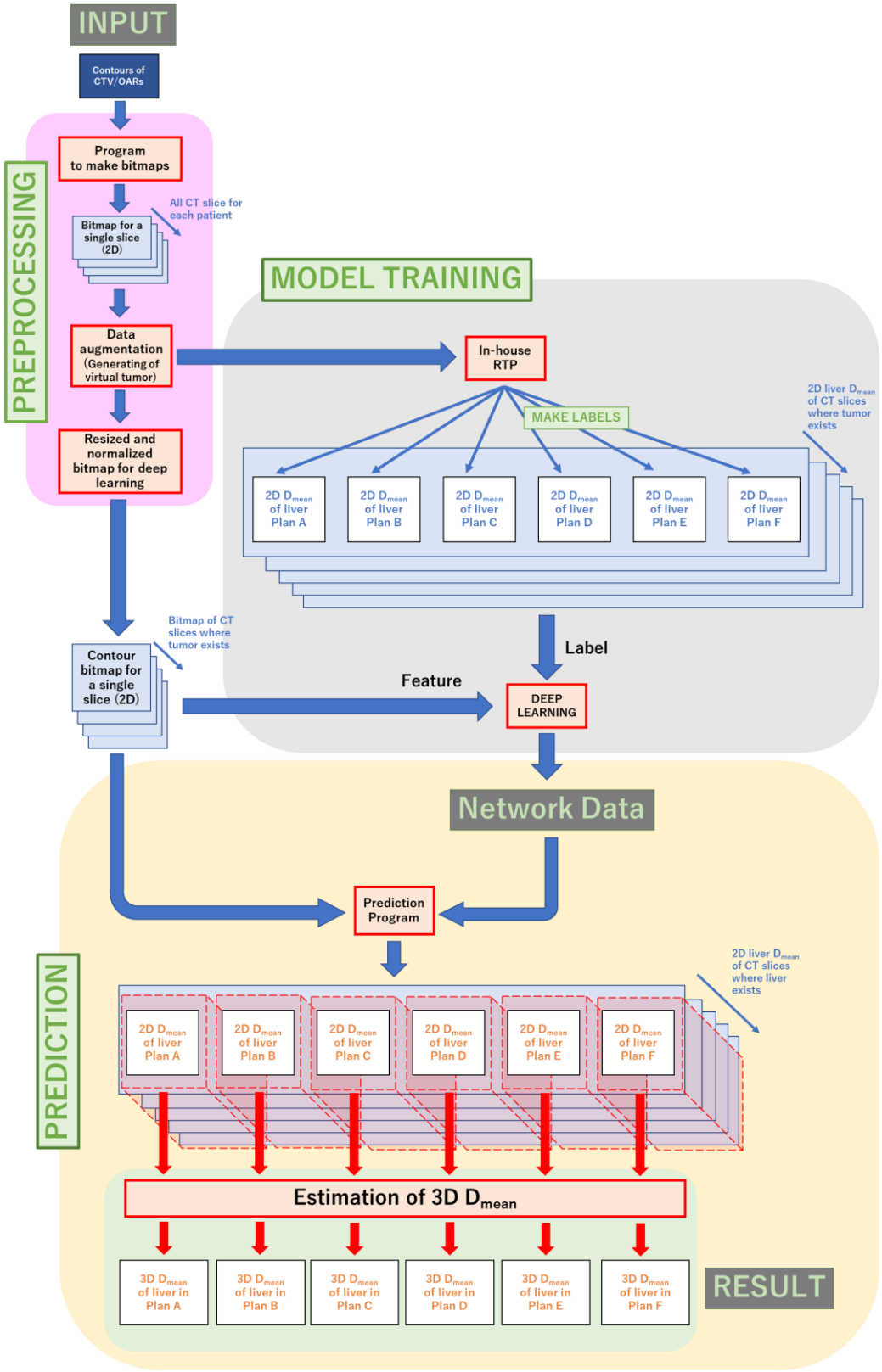

Figure 24. Flowchart of preprocessing, model training, and prediction of pre-trained model-based prediction model.

2.3 Experimental Results

2.3.1 Three-dimension profiling (3DP)

In model trained by 3DP as input, the model was tested by 10 actual patients as shown in **Table 1** in the chapter 1. The MRE between the predicted and the planned liver D_{mean} is 0.4611 and for the β it is 1.1070. The plot of relationship between predicted D_{mean} and planned D_{mean} was shown in **Figure 25**.

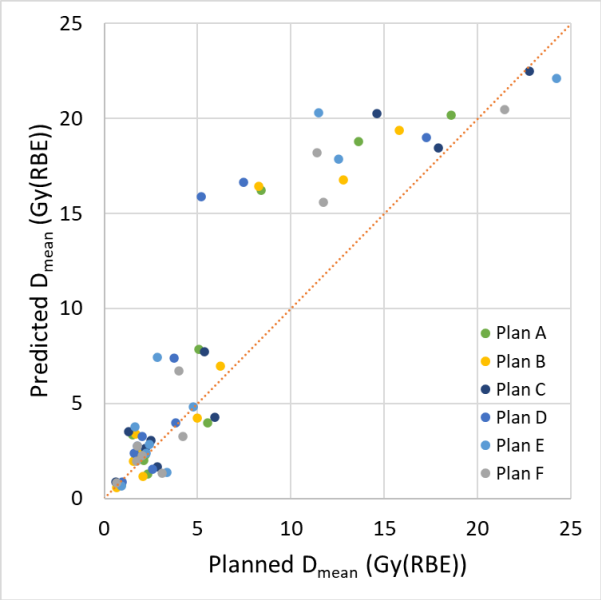

Figure 25. Plots of the planned liver mean dose (D_{mean}) and predicted liver D_{mean} with the line of identity (β = 1) using the model trained by 3DP with contour-based data augmentation.

2.3.2 Deep learning with architecture of pre-trained models

In pre-trained model-based architecture models, the model were tests by 10 actual patients as shown in **Table 1** in the chapter 1. The MRE between planned and predicted liver D_{mean} which was predicted by each pre-trained model architecture, and the training time of pre-trained model-based prediction model are summarized in Table 7 and 8.

In Alexnet-based prediction model trained with CDA, the MRE between the predicted and planned liver D_{mean} is 0.2363 and for the β it is 0.8455. The MRE is 0.2807 in case of model trained without CDA. Comparing with 2DP-based prediction model, the MRE between planned and predicted liver D_{mean} is moderately higher. Moreover, the time required to trained Alexnet-based model is 3 hours and 47 minutes, which is

approximately 2.7 times longer than model trained by 2DP.

In VGG net-based prediction model, there are 3 different models depending on the number layers including VGG13, VGG16, and VGG19. The MRE and β between the predicted liver D_{mean} which was predicted by VGG13-based model trained with CDA, and the planned liver D_{mean} is 0.1539 and 0.8975. The MRE is 0.5910 when we trained the model without CDA. In the VGG16-based model trained with CDA, the MRE between predicted and planned liver D_{mean} is 0.1017 and the β is 0.9948. Without CDA in VGG16-based model, the MRE is 0.2472. In VGG19-based model trained with CDA, the MRE between predicted and planned liver D_{mean} is 0.1338 and the β is 0.9831. In VGG19-based model trained without CDA, the MRE is 0.3967.

In GoogLenet-based prediction model, the MRE and β between predicted and planned liver D_{mean} is 0.0797 and 0.9747, respectively, for model trained with CDA. In model trained without CDA, the MRE between predicted and planned liver D_{mean} is 0.3133. When we compare between GoogLenet-based prediction model and 2DP based CNN model, in model trained with CDA, the MRE between planned and predicted liver D_{mean} is significantly reduced by using GoogLenet-based model. It means that the GoogLenet-based prediction model have more accuracy to predict liver D_{mean} than those 2DP based CNN model. However, the training time for GoogLenet based prediction model was 4 hours and 55 minutes, which was approximate 3.5 times of the training of 2DP based CNN model. In separated models, in which the 6 liver D_{meam} are predicted separately, the MRE between predicted and planned liver D_{mean} is 0.0989 and for β it is 0.9990. The time required for train the 6 separated models are 22 hours and 57 minutes, which is approximately 4.6 times of simultaneous model of GoogLenet-based prediction model. Figure 26 shows the relationship between the predicted D_{mean} and the planned liver D_{mean}.

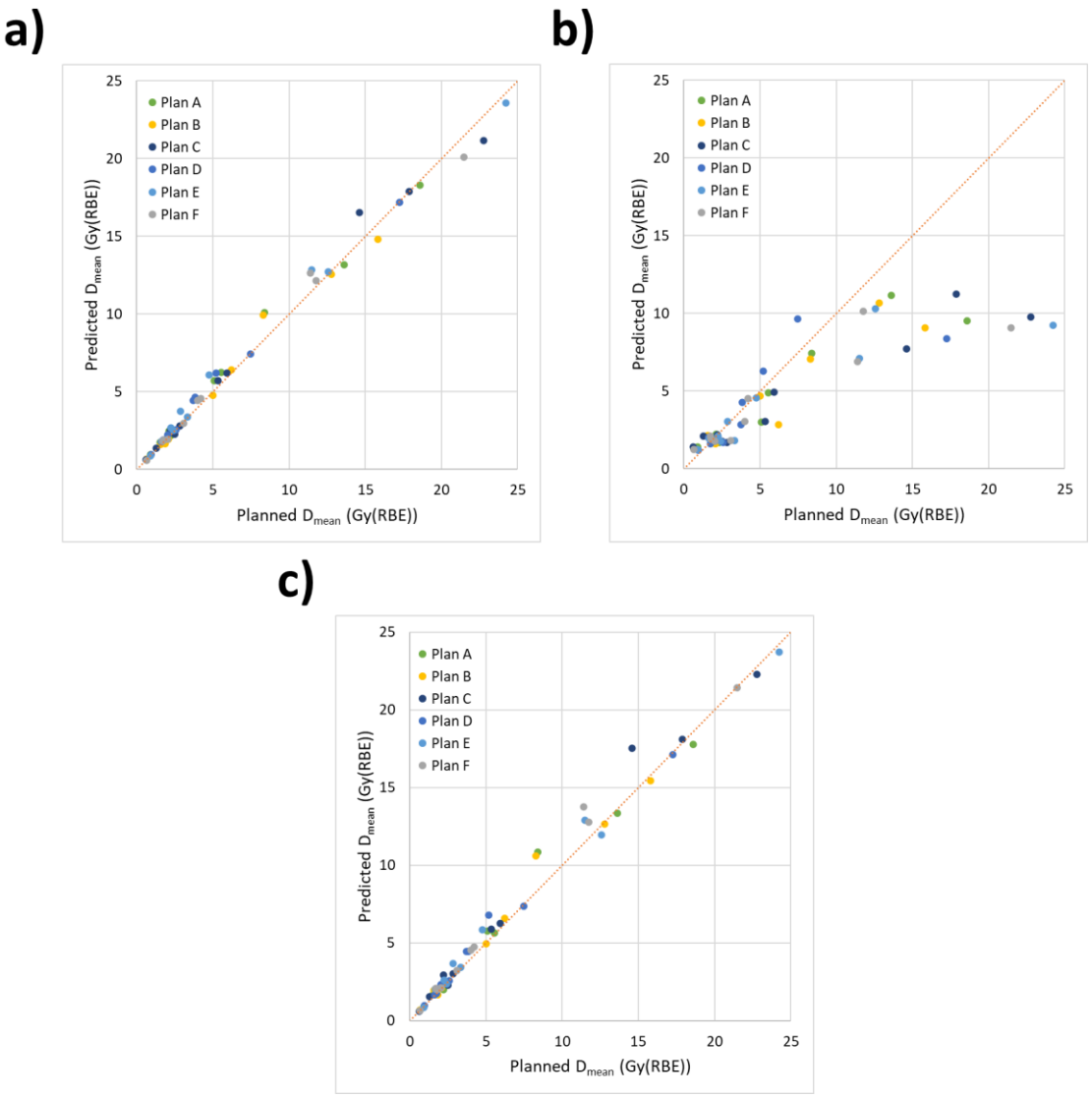

Figure 26 Plots of the planned liver mean dose (D_{mean}) and predicted liver D_{mean}, which was predicted by GoogLenet-based prediction model, with the line of identity ($\beta = 1$). (a) simultaneous model with contour-based data augmentation, (b) simultaneous model without contour-based data augmentation, (c) separated models with contour-based data augmentation.

In Inception net V3-based liver D_{mean} prediction model, the MRE between the predicted liver D_{mean}, which was predicted by model trained with CDA, and the planned liver D_{mean} is 0.1061 and for the β it is 1.0127. The MRE of those without CDA is 0.3676.

In Resnet-based prediction models, there are 3 different models depending on the number layers including Resnet50, Resnet101, and Resnet152. The MRE between the predicted liver D_{mean}, which was predicted by Resnet50, Resnet101, and Resnet152-based model trained with CDA, and the planned liver D_{mean} are 0.1022, 0.0924, and 0.1104, respectively, and the β are 1.0370, 0.9930, and 0.9944. In Resnet-based prediction model trained without CDA, the MRE are 0.4748, 0.3352, and 0.4824 for Resnet50, Resnet101, and Resnet152.

In Wide Resnet-based prediction model, the MRE between the predicted and planned liver D_{mean} is 0.0815 for model trained with CDA and for the β it is 0.9858. In Wide Resnet-based prediction model trained without CDA, the MRE between predicted and planned liver D_{mean} is 0.4997.

In Densenet-based prediction model, the MRE between planned and predicted D_{mean}, for model trained by CDA, is 0.0912 and for β it is 0.9858. In case pf Densenet-based liver D_{mean} prediction model trained without CDA, the MRE between planned and predicted D_{mean} is 0.3456.

In Shufflenet V2-based liver D_{mean} prediction model trained with CDA, the MRE between planned and predicted liver D_{mean} is 0.0901 and the β is 0.9871. For the Shufflenet V2-based model trained without CDA, the MRE between planned and predicted liver D_{mean} is 0.3494.

Table 7. The summary of MRE between planned liver D_{mean} and predicted liver D_{mean}, which were predicted by different model structure as simultaneous model, with and without CDA, and separated models.

Model Structure	MRE between planned and predicted liver D_{mean}		
	CDA	Non-CDA	Separated models
Alexnet	0.2363	0.2807	0.2330
VGG13	0.1539	0.5910	0.1316
VGG16	0.1017	0.2472	0.1100
VGG19	0.1338	0.3967	0.1215
GoogLenet	0.0797	0.3133	0.0989
Inception net V3	0.1061	0.3676	0.1156
Resnet50	0.1022	0.4748	0.0970
Resnet101	0.0924	0.3352	0.0912
Resnet152	0.1104	0.4824	0.1046
Wide Resnet101	0.0815	0.4997	0.0956
Densenet	0.0912	0.3456	0.0814
Shufflenet V2	0.0901	0.3494	0.9869

Abbreviation: MRE = Mean relative error, CDA = Contour-based data augmentation.

Table 8. The summary of training time of pre-trained model-based structure as simultaneous model and separated models.

Model Structure	Training time	
	Simultaneous model	Separated models
Alexnet	3 hours 47 minutes	16 hours 29 minutes
VGG13	8 hours 58 minutes	50 hours 7 minutes
VGG16	10 hours 6 minutes	65 hours 14 minutes
VGG19	11 hours 20 minutes	65 hours 17 minutes
GoogLenet	4 hours 55 minutes	22 hours 57 minutes
Inception net V3	8 hours 55 minutes	46 hours 13 minutes
Resnet50	6 hours 4 minutes	34 hours 20 minutes
Resnet101	9 hours 3 minutes	54 hours 51 minutes
Resnet152	12 hours 17 minutes	70 hours 45 minutes
Wide Resnet101	15 hours 39 minutes	95 hours 21 minutes
Densenet	12 hours 59 minutes	55 hours 49 minutes
Shufflenet V2	4 hours 21 minutes	19 hours 25 minutes

2.4 Discussion

2.4.1 Three-dimension profiling (3DP)

Despite my expectation, comparing to the results of 2DP shown in Chapter 1, there was no improvement by using the 3DP. Considering the complexity of the 3DP comparing to 2DP, we decided not to use the 3DP for SDP tool.

2.4.2 Deep learning with architecture of pre-trained models

Comparing to the result in Chapter 1, the most of prediction model based on pre-trained model architecture have higher accuracy to predict liver D_{mean}, except for only Alexnet-based model. The GoogLenet-based prediction model seem to have the highest accuracy in liver D_{mean} prediction of 10 patients in the test data set. In the GoogLenet structure, their consists of 22 layers, which part of these layers is 9 inception modules. This inception module is a neural network that leverages feature detection at different scales through convolutions with different filters and reduced the computational cost of training an extensive network through dimensional reduction (Szegedy *et al.*, 2015). The present study showed that GoogLenet-based model can predict liver D_{mean} with the highest accuracy with shorter training time than other pre-trained model-based liver D_{mean} prediction models such as VGGnet, Resnet, and Densenet. In fact, the accuracy of GoogLenet-based model is still inferior to the research of Guerreiro et al., which reported that the MPE between the planned and predicted D_{mean} of all OARs in pediatric abdominal tumors was -0.3% ± 2.9% (Guerreiro *et al.*, 2021). The MPE between the planned and predicted liver D_{mean}, which was predicted by our GoogLenet model was 5.08% ± 9.47%. However, my model input is contour bitmap which required less performance computer and resource when compared to their work which use precise 3D volumes and 3D dose distributions.

The MRE between planned liver D_{mean} and predicted liver D_{mean}, which was predicted by VGG16, Inception net V3, Resnet101, Wide Resnet 101, Densenet, and Shufflenet V2 are 0.1017, 0.1061, 0.0924, 0.0815, 0.0912, and 0.0901, respectively. The results of MRE of various pre-trained model-based structures are not significantly different each other. Litjens et al have reviewed many papers in medical image analysis and discussed about the key aspects of successful deep learning methods that the architecture is not the most important factor for a good result (Litjens *et al.*, 2017). There are many researchers that

obtained good result without improved the deep neural network structure but focusing on other aspects, for example data preprocessing, augmentation techniques, or model hyper-parameter optimization. Therefore, the key point to improve the liver D_{mean} prediction model may not be only these pre-trained model-based structure, but others. More work is required from these aspects in future.

Summary and Conclusion

1. We developed an SDP tool to predict a liver D_{mean}, in which a physician is only required to do contouring of the CTV and OARs (skin surface, liver, and spinal cord), and assessed its accuracy and usability.
2. The SDP tool is cost-effective and usable for an approximate estimation of liver D_{mean}, however, the accuracy should be improved further in case of we need the accuracy to be compatible with 3DRTP.
3. The same concept as our simple dose prediction (SDP) tools, which are deep learning models trained with our contour-based data augmentation (CDA), can be applied to the dose distribution of STI and IMXT.
4. The pre-trained model-based liver D_{mean} prediction model have high accuracy in liver D_{mean} prediction, especially in GoogLenet-based prediction model. The model can predict liver D_{mean} from only ROI Contour Module of CTV and OARs in DICOM RT structure as input.

It is important to conclude that an SDP tool using deep learning can predict liver D_{mean} directly with much shorter time than conventional 3DRTP. It is notable that requirement for the prediction of liver D_{mean} using an SDP tool is only DICOM RT structure set of CTV and 3 OARs (skin surface, liver, and spinal cord). It is also important to stress that the CDA is useful technique in deep learning to develop the SDP tool. We proved that the result of liver D_{mean} prediction model trained with CDA technique is more accurate than those without CDA technique in case of small training data sets. The CDA technique is useful to solve the problem of lacking data set in model training. Direct liver D_{mean} prediction has potential to use in the busy clinic. The accuracy of predicted results of our main method described in Chapter I is acceptable but not perfectly accurate. Deep learning with architecture of pre-trained models described in Chapter II can improve the prediction accuracy without increasing the time for calculation in SDP.

As a possible research development of the new findings obtained in this study, the concept of CDA can be applied to other deep learning for radiation dose prediction. For example, doses for other OARs and CTV will be able to be predicted. Also, any other organs in

the body, such as lung and pancreas, can be the target of new research. Using additional biological assumptions, prediction of normal tissue complication probability (NTCP) or tumor control probability (TCP) is good targets of research to be investigated. The accuracy of prediction can be improved not only by the model architecture but other factors such as techniques of data preprocessing, augmentation, or model hyper-parameter optimization in deep learning method. Research for these factors is also important step forwards to improve the accuracy and speed of prediction of dose distribution in precise PBT and any other radiotherapies.

Acknowledgement

I would like to use this opportunity to express my gratitude to everyone who supported me during the completion of this thesis.

First, I would like to thank my thesis advisors, Prof. Hiroki Shirato, M.D., Prof. Shinichi Shimizu, M.D., and Assoc. Prof. Keiji Kobashi, Ph.D. for their helpful guidance, encouragement and support. The door to their office was always open whenever I ran into a trouble spot or had a question about my research or writing.

I would also like to thank Asst. Prof. Seishin Takao, Ph.D. and Mr. Keiji Nakazato, M.S., for providing radiation treatment plan data in this study. Without their participation, this research could not have been successfully conduced.

I would also like to thank Mr. Izuru Ohtake and Mr. Atsushi Omori for their supporting in computational advising and material.

I would also like to thank Dr. Yusuke Nomura, Ph.D. from Department of Radiation Oncology, Stanford University, Stanford, CA, United States of America, who took time out to hear my ideas and give valuable suggestions.

Moreover, I am very thankful to staff of the Global Center for Biomedical Science and Engineering (GCB), Hokkaido university and staff of the Office of Educational Affairs, Royal Thai Embassy Tokyo for their support, assistance, and encouragement.

Most importantly, I must express my profound gratitude to my family for providing me with their unfailing support throughout many years of study as well as through the process of researching and writing this thesis. Finally, I would like to thank the Office of the Civil Service Commission along with the Ministry of Science and Technology, Thailand for funding support as a fully scholarship during my study period.

Citated references

Boldrini, L., Bibault, J. E., Masciocchi, C., Shen, Y. and Bittner, M. I. (2019) 'Deep Learning: A Review for the Radiation Oncologist', *Frontiers in Oncology*, 9(October). doi: 10.3389/fonc.2019.00977.

Bortfeld, T. and Schlegel, W. (1996) 'An analytical approximation of depth - dose distributions for therapeutic proton beams', *Physics in Medicine and Biology*. {IOP} Publishing, 41(8), pp. 1331–1339. doi: 10.1088/0031-9155/41/8/006.

Chen, X., Men, K., Li, Y., Yi, J. and Dai, J. (2019) 'A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning', *Medical Physics*, 46(1), pp. 56–64. doi: 10.1002/mp.13262.

Chen, X., Men, K., Zhu, J., Yang, B., Li, M., Liu, Z., Yan, X., Yi, J. and Dai, J. (2021) 'DVHnet: A deep-learning-based prediction of patient-specific dose volume histograms for radiotherapy planning', *Medical Physics*. doi: 10.1002/mp.14758.

Chen, Y., Grassberger, C., Li, J., Hong, T. S. and Paganetti, H. (2018) 'Impact of potentially variable RBE in liver proton therapy', *Physics in Medicine and Biology*. IOP Publishing, 63(19). doi: 10.1088/1361-6560/aadf24.

Cilla, S., Deodato, F., Romano, C., Ianiro, A., Macchia, G., Re, A., Buwenge, M., Boldrini, L., Indovina, L., Valentini, V. and Morganti, A. G. (2021) 'Personalized automation of treatment planning in head-neck cancer: A step forward for quality in radiation therapy?', *Physica Medica*. Elsevier Ltd, 82(December 2020), pp. 7–16. doi: 10.1016/j.ejmp.2020.12.015.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. and Prior, F. (2013) 'The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository', *Journal of Digital Imaging*, 26(6), pp. 1045–1057. doi: 10.1007/s10278-013-9622-7.

Dawson, L. A., Normolle, D., Balter, J. M., McGinn, C. J., Lawrence, T. S. and Ten Haken, R. K. (2002) 'Analysis of radiation-induced liver disease using the Lyman NTCP model', *International Journal of Radiation Oncology*Biology*Physics*, 53(4), pp. 810–821. doi: https://doi.org/10.1016/S0360-3016(02)02846-8.

Erickson, B. J., Krik, S., Lee, Y., Bathe, O., Kearns, M. and Gerdes, C. (2016) *Radiology Data from The Cancer Genome Atlas Liver Hepatocellular Carcinoma [TCGA-LIHC] collection. The Cancer Imaging Archive.* doi: http://doi.org/10.7937/K9/TCIA.2016.IMMQW8UQ.

Fung, N. T. C., Hung, W. M., Sze, C. K., Lee, M. C. H. and Ng, W. T. (2020) 'Automatic segmentation for adaptive planning in nasopharyngeal carcinoma IMRT: Time, geometrical, and dosimetric analysis', *Medical Dosimetry*, 45(1), pp. 60–65. doi: https://doi.org/10.1016/j.meddos.2019.06.002.

Goodfellow, I. J., Bengio, Y. and Courville, A. (2016) *Deep Learning*, *Nature*. Cambridge, MA, USA: MIT Press. Available at: http://www.deeplearningbook.org.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. and Chen, T. (2018) 'Recent advances in convolutional neural networks', *Pattern Recognition*. Elsevier Ltd, 77, pp. 354–377. doi: 10.1016/j.patcog.2017.10.013.

Guerreiro, F., Seravalli, E., Janssens, G. O., Maduro, J. H., Knopf, A. C., Langendijk, J. A., Raaymakers, B. W. and Kontaxis, C. (2021) 'Deep learning prediction of proton and photon dose distributions for paediatric abdominal tumours', *Radiotherapy and Oncology*. The Authors, 156, pp. 36–42. doi: 10.1016/j.radonc.2020.11.026.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem, pp. 770–778. doi: 10.1109/CVPR.2016.90.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K. Q. (2017) 'Densely connected convolutional networks', *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

Ibragimov, B., Toesca, D., Chang, D., Yuan, Y., Koong, A. and Xing, L. (2018) 'Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT', *Medical Physics*, 45(10), pp. 4763–4774. doi: 10.1002/mp.13122.

Kingma, D. P. and Ba, J. L. (2015) 'Adam: A method for stochastic optimization', *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track*

Proceedings, abs/1412.6, pp. 1–15.

Kouwenberg, J., Penninkhof, J., Habraken, S., Zindler, J., Hoogeman, M. and Heijmen, B. (2021) 'Model based patient pre-selection for intensity-modulated proton therapy (IMPT) using automated treatment planning and machine learning', *Radiotherapy and Oncology*. The Author(s), 158, pp. 224–229. doi: 10.1016/j.radonc.2021.02.034.

Krizhevsky, A. (2014) 'One weird trick for parallelizing convolutional neural networks'. Available at: http://arxiv.org/abs/1404.5997.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B. and Sánchez, C. I. (2017) 'A survey on deep learning in medical image analysis', *Medical Image Analysis*. Elsevier B.V., 42(December 2017), pp. 60–88. doi: 10.1016/j.media.2017.07.005.

Loeffler, J. S. and Durante, M. (2013) 'Charged particle therapy--optimization, challenges and future directions.', *Nature reviews. Clinical oncology*. England, 10(7), pp. 411–424. doi: 10.1038/nrclinonc.2013.79.

Ma, N., Zhang, X., Zheng, H. T. and Sun, J. (2018) 'Shufflenet V2: Practical guidelines for efficient cnn architecture design', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11218 LNCS, pp. 122–138. doi: 10.1007/978-3-030-01264-9_8.

Marks, L. B., Yorke, E. D., Jackson, A., Ten Haken, R. K., Constine, L. S., Eisbruch, A., Bentzen, S. M., Nam, J. and Deasy, J. O. (2010) 'Use of Normal Tissue Complication Probability Models in the Clinic', *International Journal of Radiation Oncology Biology Physics*, 76(3 SUPPL.). doi: 10.1016/j.ijrobp.2009.07.1754.

Nair, V. and Hinton, G. (2010) 'Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair', in *Proceedings of ICML*, pp. 807–814.

Nguyen, D., Long, T., Jia, X., Lu, W., Gu, X., Iqbal, Z. and Jiang, S. (2019) 'A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning', *Scientific Reports*, 9(1), pp. 1–10. doi: 10.1038/s41598-018-37741-x.

Prayongrat, A., Umegaki, K., Van Der Schaaf, A., Koong, A. C., Lin, S. H., Whitaker, T., McNutt, T., Matsufuji, N., Graves, E., Mizuta, M., Ogawa, K., Date, H., Moriwaki, K., Ito, Y. M., Kobashi, K., Dekura, Y., Shimizu, S. and Shirato, H. (2018) 'Present developments in reaching an international consensus for a model-based approach to particle beam therapy', *Journal of Radiation Research*, 59(March), pp. i72–i76. doi: 10.1093/jrr/rry008.

Simonyan, K. and Zisserman, A. (2015) 'Very deep convolutional networks for large-scale image recognition', *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) 'Going deeper with convolutions', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) 'Rethinking the Inception Architecture for Computer Vision', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

Zagoruyko, S. and Komodakis, N. (2016) 'Wide Residual Networks', *British Machine Vision Conference 2016, BMVC 2016*, 2016-Septe, pp. 87.1-87.12. doi: 10.5244/C.30.87.