

A Comparative Study of Radiomics and Deep-Learning Approaches for Predicting  
Surgery Outcomes in Early-Stage Non-Small Cell Lung Cancer (NSCLC)

by

Haozhao Zhang

Graduate Program of Medical Physics  
Duke Kunshan University and Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Chunhao Wang, Co-Supervisor

\_\_\_\_\_  
Fang-Fang Yin, Co-Supervisor

\_\_\_\_\_  
David Huang

Thesis submitted in partial fulfillment of  
the requirements for the degree of  
Master of Science in the Graduate Program of  
Medical Physics in the Graduate School of  
Duke Kunshan University and Duke University

2022

ABSTRACT

A Comparative Study of Radiomics and Deep-Learning Approaches for Predicting  
Surgery Outcomes in Early-Stage Non-Small Cell Lung Cancer (NSCLC)

by

Haozhao Zhang

Graduate Program of Medical Physics  
Duke Kunshan University and Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Chunhao Wang, Co-Supervisor

\_\_\_\_\_  
Fang-Fang Yin, Co-Supervisor

\_\_\_\_\_  
David Huang

An abstract of a thesis submitted in partial  
fulfillment of the requirements for the degree of  
Master of Science in the Graduate Program of  
Medical Physics in the Graduate School of  
Duke Kunshan University and Duke University

2022

Copyright by  
Haozhao Zhang  
2022

## Abstract

**Purpose:** To compare radiomics and deep-learning (DL) methods for predicting NSCLC surgical treatment failure.

**Methods:** A cohort of 83 patients undergoing lobectomy or wedge resection for early-stage NSCLC from our institution was studied. There were 7 local failures and 16 non-local failures (regional and/or distant). Gross tumor volumes (GTV) were contoured on pre-surgery CT datasets after 1mm<sup>3</sup> isotropic resolution resampling. For the radiomics analysis, 92 radiomics features were extracted from the GTV and z-score normalizations were performed. The multivariate association between the extracted features and clinical endpoints were investigated using a random forest model following 70%-30% training-test split. For the DL analysis, both 2D and 3D model designs were executed using two different deep neural networks as transfer learning problems: in 2D-based design, 8x8cm<sup>2</sup> axial fields-of-view (FOVs) centered within the GTV were adopted for VGG-16 training; in 3D-based design, 8x8x8 cm<sup>3</sup> FOVs centered within the GTV were adopted for U-Net's encoder path training. In both designs, data augmentation (rotation, translation, flip, noise) was included to overcome potential training convergence problems due to the imbalanced dataset, and the same 70%-30% training-test split was used. The performances of the 3 models (Radiomics, 2D-DL, 3D-DL) were tested to predict outcomes including local failure, non-local failure, and disease-free survival.

Sensitivity/specificity/accuracy/ROC results were obtained from their 20 trained versions.

**Results:** The radiomics models showed limited performances in all three outcome prediction tasks. The 2D-DL design showed significant improvement compared to the radiomics results in predicting local failure (ROC AUC =  $0.546 \pm 0.056$ ). The 3D-DL design achieved the best performance for all three outcomes (local failure ROC AUC =  $0.768 \pm 0.051$ , non-local failure ROC AUC =  $0.683 \pm 0.027$ , disease-free ROC AUC =  $0.694 \pm 0.042$ ) with statistically significant improvements from radiomics/2D-DL results.

**Conclusions:** 3D-DL execution outperformed the 2D-DL in predicting clinical outcomes after surgery for early-stage NSCLC. By contrast, classic radiomics approach did not achieve satisfactory results.

## **Dedication**

I would like to dedicate this thesis work to my grandpa who passed away one month ago for his endless love and unconditional encouragement supporting me throughout my graduate career.

# Contents

Abstract .....	iv
Dedication .....	vi
List of Figures .....	ix
Acknowledgements .....	x
1. Introduction .....	1
1.1 Background .....	1
1.2 Non-small Cell Lung Cancer .....	3
1.3 Radiomics .....	5
1.4 Deep Learning.....	7
1.5 Characteristics of Medical Data.....	10
1.6 Motivation and Objective.....	11
2. Materials and Methods.....	12
2.1 Data Description.....	13
2.2 Radiomics .....	14
2.2.1 Radiomics Feature Extraction.....	14
2.2.2 Random Forest.....	16
2.2.3 Summary of Radiomics Implementation .....	18
2.3 Deep Learning.....	20
2.3.1 Deep Learning with Keras .....	20
2.3.2 Deep Learning Mechanism .....	21

2.3.3 Convolution Neutral Network.....	23
2.3.4 Transfer Learning.....	24
2.3.4.1 Feature Extraction.....	25
2.3.4.2 Fine-tuning.....	25
2.3.5 Data Augmentation.....	26
2.3.6 2D Deep Learning Execution.....	30
2.3.7 3D Deep Learning Execution.....	32
2.3.8 Summary of Deep Learning Execution.....	34
2.4 Model Evaluation.....	37
3. Results.....	38
4. Discussion.....	43
5. Conclusion.....	48
References.....	49

## List of Figures

Figure 1: Schematic diagram of surgery for NSCLC.....	5
Figure 2: The relationship of AL, ML, and DL. ....	7
Figure 3: Radiomic features used in this analysis, color-coded by their class.....	15
Figure 4: Schematic of a random forest model .....	17
Figure 5: Workflow of radiomics implementation.....	19
Figure 6: The deep-learning software and hardware stack.....	21
Figure 7: Relationship between the network, layers, loss function, and optimizer.....	23
Figure 8: Concept of transfer learning.....	26
Figure 9: Data augmentation.....	29
Figure 10: 2D DL execution.....	31
Figure 11: 3D DL execution.....	33
Figure 12: Overall workflow of 2D & 3D execution.....	36
Figure 13: ROC curves and quantitative results .....	40
Figure 14. The CC matrix of saliency maps of 3D-DL model.....	42

## Acknowledgements

I would first like to express my great gratitude to my advisor Dr. Chunhao Wang and appreciation to Dr. Fang-Fang Yin, for providing me this unique opportunity to participate in the research work of deep learning and radiomics, thanks for their expert guidance, unconditional encouragement, and endless support in all aspects. Under their supervision, I learned how to find a research problem, utilize the resource, investigate the possible solution, and finally present the results.

Great thanks to Dr. David Huang and Dr. James Bowsher for their assistance, encouragement, and support all the time.

Sincere thanks to Zhenyu Yang, Zongsheng Hu, Jingtong Zhao, Zeyu Zhang, Ke Lu, and Nan Li for their generous help in the coding, data processing, and data analysis.

Also, thanks to Class 2022 for support of each other in the past two years.

Finally, I would like to thank family for life-long love and support.

# **1. Introduction**

## **1.1 Background**

Lung cancer, originating from the bronchial mucosa or glands of the lungs, is caused by unchecked growth, and spread of some cells from the lungs, which is currently the second most common cancer worldwide. Global cancer statistics showed that there were 2.2 million cancer diagnoses and 1.8 million deaths in 2020, of which lung cancer accounted for about one in ten (11.4%) of diagnosed cancers and one in five (18.0%) of all deaths.<sup>[1]</sup> The etiology of lung cancer has not yet been fully investigated, but smoking is recognized as the most common cause of lung cancer. In China, lung cancer ranks first and second in incidence in males and females, respectively.<sup>[2]</sup> Although several developed countries such as the United States, in recent decades, as the smoking rate of men has decreased significantly, lung cancer rate has fallen to the second place in the incidence of lung cancer among men in the United States <sup>[2]</sup>. However, lung cancer remains the leading cause of cancer death in America <sup>[1-3]</sup>. By 2022, it was estimated that about 350 patients will die of lung cancer per day <sup>[3]</sup>. So far, researchers have not found any drugs, vitamins, herbal remedies, or alternative medicines that can help prevent lung cancer. For the treatment of lung cancer, the corresponding pathological type and clinical stage should be identified first, and the overall status of the patient should be comprehensively assessed, and multiple methods should be weighed for comprehensive treatment. These

results in alleviating the symptoms of patients, improving their quality of life, and prolonging the prognosis and survival time.

There are two main types of lung cancer diagnosed <sup>[4]</sup>: non-small cell lung cancer (NSCLC, accounting for 85% of all lung cancer cases) and small cell lung cancer (SCLC). SCLC is a malignant epithelial tumor consisting of small cells of lung tissue <sup>[5]</sup>. Most patients with SCLC presented with extensive-stage disease and poor prognosis, with median overall survival only 9.5 months <sup>[6]</sup>. Compared with patients undergoing SCLC or advanced-stage NSCLC, survival for those with early-stage NSCLC is promising and increasing (i.e., NSCLC 2-year relative survival increased from 34% in 2009 to 42% in 2016) <sup>[7]</sup> with advances in diagnostic and surgical procedures, such as pathologic staging, video assisted thoracoscopic surgery, and stereotactic body radiation therapy <sup>[8]</sup>. To sum up, small cell lung cancer progresses rapidly and metastasizes early, relying on chemotherapy or radiotherapy; non-small cell lung cancer is often limited in disease, and surgical operations are often used in combination with radiotherapy and chemotherapy.

At the same time, under the wave of artificial intelligence (AI) in the past decade, the vigorous development of medical imaging and computer assisted diagnosis (CAD) has led to more and more promising research progress in the clinical application and retrospective research of medical image data. <sup>[9]</sup>

The radiomics method has been able to quantify the pixel and spatial distribution relationships of most medical images, fully exploit the hidden information in the images

that cannot be observed by the naked eye and supplement the traditional medical imaging knowledge system. Radiomics also extensively incorporates analytical modeling tools such as statistics and machine learning to further aid clinical decision-making. <sup>[10-12]</sup>

Deep learning has effectively improved the performance of various machine learning tasks by automatically learning from large sample data to obtain excellent feature expressions. It has been widely used in many fields such as signal processing, computer vision and natural language processing. The analysis of medical images based on deep learning is currently a research hotspot in the field of AI, in which deep learning methods have been applied to the entire process of medical image processing, feature analysis and other fields. Multiple tasks such as image reconstruction, lesion detection, image segmentation, image registration, and computer-aided diagnosis have been achieved through deep learning, based on commonly used clinical X-ray, ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI). <sup>[12-14]</sup>

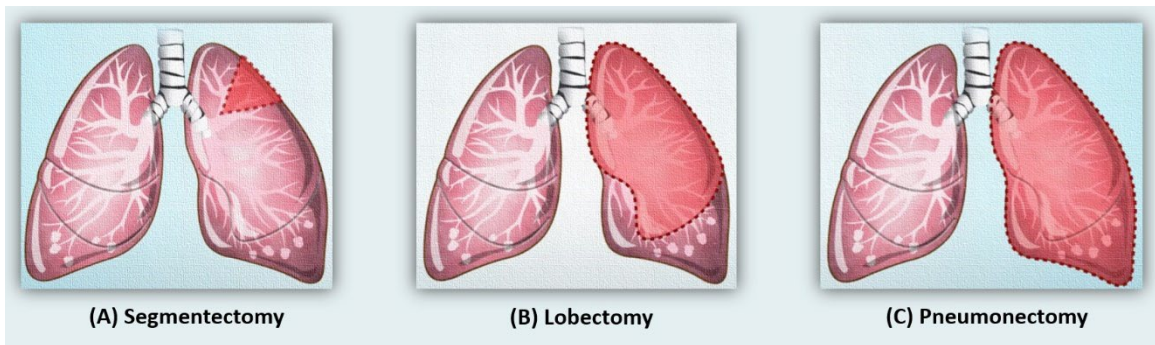
## ***1.2 Non-small Cell Lung Cancer***

NSCLC is the most common lung cancer, which affects 8 of every 10 people with lung cancer. The most common subtypes of NSCLC are adenocarcinoma (accounting for 40% of all lung cancers), squamous cell (epidermoid) carcinoma (occupying 25–30% of all lung cancer cases), and large cell (undifferentiated) carcinoma (comprising around 5–10% of lung cancers). Adenocarcinoma is the most common type of lung cancer, which is mostly peripheral type (referring to lung cancers that occur in the subsegmental bronchi).

Adenocarcinoma is prone to local infiltration, early metastasis through blood, and easy involvement of the pleura to cause pleural effusion. Squamous cell carcinoma progresses slowly and metastasize late, which makes the surgical treatment of squamous cell carcinoma more opportunities. Large-cell cancer cells have a high degree of malignancy, but still metastasize later than small-cell cancer, and have a greater chance of controlling the disease by surgical resection. <sup>[15-17]</sup>

Lung biopsy is the gold standard for diagnosing lung cancer. Chest CT can further verify the location and scope of lesions and can also roughly distinguish benign from malignant. It is an important method for diagnosing lung cancer. Among them, low-dose spiral CT plays an increasingly significant role in the early diagnosis of lung cancer and is gradually applied to early lung cancer screening. MRI, PET-CT are also suitable for judging the metastasis of lung cancer brain, lymph nodes, bones, and other tissues. According to the analysis of the results of imaging examination and laboratory examination, combined with clinical disease progression, NSCLC is divided into stages I, II, III, and IV <sup>[15]</sup>. Stage I is an early stage, which means that the tumor is in the lung tissue and has not yet metastasized. Stage II belongs to the middle stage, which means that the cancer cells have metastasized to the lymph nodes near the hilum of the lungs, but the degree of infiltration is not high. Stage III is an advanced stage, which means that the cancer cells have further metastasized to the mediastinum or extrapulmonary lymph nodes. Stage IV is an advanced stage, which refers to the occurrence of pleural metastasis,

pleural effusion, or multiple metastases throughout the body, such as liver, brain, and bone. [4,15] Surgery is the first and most important method for the treatment of lung cancer. [4,8,15,17] It is suitable for all patients with early stage, middle stage, and a small number of patients with advanced non-small cell lung cancer. The 5-year survival rate of lung cancer patients after surgery is 30%-44%, and the mortality rate during surgery is 1%-2%. Lobectomy combined with systematic lymph node dissection is currently the standard surgical procedure for lung cancer. [7,8,15]



**Figure 1: Schematic diagram of surgery for NSCLC.**

Depending on how far the cancer has spread, types of procedures involving surgery (shown in Figure 1.) may result in the removal of part of a lung lobe (called a segmentectomy or wedge resection); an entire lobe (called a lobectomy); or the whole lung (called a pneumonectomy).

### **1.3 Radiomics**

With the development of imaging technology, CT, PET/CT, MRI, and other detection methods play an increasingly vital role in cancer diagnosis, treatment, and

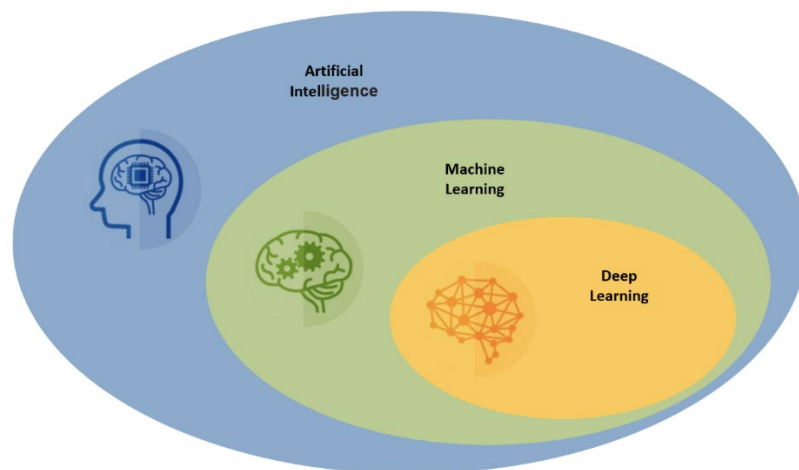
efficacy evaluation. Nowadays, imaging examination can not only obtain the morphological characteristics of the tumor, but also interpret part of the biological characteristics of the tumor through functional imaging, hypoxia imaging and other means. The fusion of digital image processing techniques and machine learning methods in the analysis of medical images gave birth to the concept of radiomics. Radiomics is a high-throughput quantitative analysis method, which extracts hundreds of quantitative image features from regions of interest in medical images such as CT, MRI, and PET, and then screens and analyzes these features. <sup>[19-20]</sup>

Radiomics integrates digital image processing, statistics, machine learning and other methods to perform quantitative, high-throughput analysis of medical images, and extract more clinical significance from routine imaging examinations. From the perspective of image processing, medical images can be viewed as a two-dimensional matrix. Each element in the matrix corresponds to a voxel in CT, PET-CT, and MRI examinations, and the values of the matrix elements are the values obtained from these imaging examinations. Applying methods developed in the field of digital image processing, hundreds of quantitative features can be extracted from these images. Integrating the concept of radiomics, screening and analyzing these features can fully excavate the hidden information in medical images. The clinical significance that cannot be observed by the naked eye is now improved, the utilization efficiency of imaging examinations is improved, and its clinical significance and value are expanded. <sup>[10-11]</sup>

The combination of radiomics and machine learning has been demonstrated in several retrospective studies based on CT images in clinical outcomes (e.g., patient survival and recurrence) and in quantifying pulmonary function. However, few studies have focused on the application of radiomics for clinical outcome prediction and modeling in early-stage NSCLC surgical patients, let alone a comprehensive comparative study with other methods, such as deep learning. [21-22]

## **1.4 Deep Learning**

Machine learning (ML), deep learning (DL), and artificial intelligence (AI) have been the subject of much media hype over the past few years, appearing in countless articles in the fields of computing, engineering, and medicine. AI is a comprehensive field that attempts to automate intellectual tasks normally performed by humans, including not only ML and DL, but also methods that do not involve any learning. The relationship of AI, ML, and DL is shown in the Figure 2. [23-24]



**Figure 2: The relationship of AI, ML, and DL.**

Machine learning is systematically trained, not explicitly programmed. Feed many samples related to a task into a machine learning system, and it will find statistical structure in those examples, and ultimately rules to automate the task.

Deep learning is a subfield of machine learning: it is a novel approach to learning representations from data, emphasizing learning from successive layers that correspond to increasingly meaningful representations. The "deep" in "deep learning" does not refer to the deeper understanding gained using this method, but to a series of successive representation layers. The number of layers in the data model is called the depth of the model. Modern deep learning models typically contain dozens or even hundreds of successive representation layers, all automatically learned from training data. In contrast, other machine learning methods tend to focus on learning only one or two layers of data representation and are therefore sometimes referred to as shallow learning. In deep learning, these hierarchical representations are almost always learned through models called neural networks. The structure of the neural network is stacked layer by layer. The term neural network comes from neurobiology, however, while some of the core concepts of deep learning are partly inspired by our understanding of the brain, deep learning models are not models of the brain. There is no evidence that the brain's learning mechanisms are the same as those used by modern deep learning models. Briefly, deep learning is a mathematical framework for learning representations from data. <sup>[24]</sup>

In recent years, deep learning has continued to make noteworthy progress, mainly due to the continuous improvement of computing power and the continuous increase in the amount of available data, as well as the continuous improvement of deep learning models and their algorithms [25]. Its essence is to build a multi-hidden layer machine learning model, use massive sample data for training, learn more accurate features, and finally improve the accuracy of classification or prediction. The characteristic of deep learning, which learns hierarchical features from data, makes it very suitable for discovering complex structures in high-dimensional data. The remarkable success of deep learning in the field of computer vision has inspired many scholars at home and abroad to apply it to medical image analysis. Applying deep learning to solve medical image analysis tasks is a development trend in this field [26]. Many experts have summarized, commented, and discussed the research status and problems of deep learning in medical image analysis. Deep Learning (DL) algorithms, such as Convolutional Neural Networks (CNN), which can automatically extract disease diagnostic features hidden in medical image data, are now being widely used to analyze medical images. [13-14]

Deep learning has made some progress in the survival prediction of high-grade glioma [27] based on MRI images and the prediction of COVID-19 based on CT images [28]. Regarding NSCLC, most of the existing studies are retrospective analyses of prognostic outcomes based on images of SBRT patients combined with radiomics. [9-12,21] *Xu et al* utilized serial CT images to study stage III patients undergoing chemotherapy and

surgery.<sup>[29]</sup> However, there are not many studies focusing on the prediction of clinical prognosis of patients with early-stage NSCLC surgery by means of deep learning. Therefore, a focus of this study is also to conduct a comparative study through deep learning schemes of 2D and 3D modes.

### ***1.5 Characteristics of Medical Data***

With the advancement of medical technology and the rapid penetration of big data in the health field, many new discoveries and methods have been promoted in the past few years. Biological and clinical data at different scales are generated and collected at unprecedented speed and scale for health management. In terms of disciplines, big data has a key role in promoting the development of the following biomedical fields: bioinformatics, clinical informatics, imaging informatics and public health informatics. Big data can be used to solve professional problems in medical imaging and optimize processes. The application of big data in the field of medical imaging business is also being widely explored. Because imaging examinations have been digitized, there is a natural possibility of big data processing.<sup>[30]</sup>

Although the total amount of medical imaging data is huge today, only an exceedingly small part can be integrated, interpreted, and analyzed. The difficulties in dealing with massive data research lie in the huge amount of data, too many data sources, inconsistent data formats and flawed data. flood the database. Most medical data have an uneven distribution of observations because only a small percentage of patients

experience health problems. Therefore, depending on the proportion of positive and negative samples, imbalanced data may need to be preprocessed, as traditional algorithms tend to treat few observations as noise. In addition, in the process of big data processing, statistical theory and machine learning technology are very important, but medical imaging experts have limited knowledge of IT technology, and IT technical experts are not easy to understand the essence of medical problems, and interdisciplinary talents are scarce. <sup>[31-32]</sup>

## ***1.6 Motivation and Objective***

For patients with NSCLC, lobectomy is recognized as one of the primary therapeutic plans based on overall survival status and evaluation of lung function <sup>[8]</sup>. In multiple medical image-based oncology studies, radiomics and DL methods reveal immense potential in feature extraction and malignant prediction <sup>[9,12]</sup>. However, with the same amount of imbalanced scale training data as in this study, it was not clear which method is better at predicting clinical outcomes based on pre-operative radiography. This work compared two computational approaches, radiomics and deep learning (DL), in outcome prediction of early-stage non-small cell lung cancer (NSCLC) surgery using pre-procedure CT image. This work is a comprehensive comparison study about NSCLC surgery outcome prediction.

## 2. Materials and Methods

*The following methods were conducted in accordance with relevant guidelines and regulations. Retrospective data collection and analysis was completed with approval from the Duke University Health System Institutional Review Board.*

Pretreatment X-ray CT images were acquired on several different CT or PET/CT (Discovery STE, Discovery 690, Discovery CT750 HD, LightSpeed Xtra etc.) and reconstructed using a filtered back projection reconstruction algorithm. Lesions were identified on pre-treatment computed tomography scans under free-breathing conditions. For each image, the gross tumor volume (GTV) was manually segmented using commercially available contouring software (ARIA Eclipse, Varian). A single physician with experience in interpretation and manual delineation of lung lesions performed all contouring. No pre-defined directions regarding display settings (e.g., window/level, thresholding, pixel representation) were used or specified. Lesion size and appearance were cross-checked against prior radiological interpretation to ensure appropriate delineation from lymph vascular structures.

Radiomics feature extraction process and original image preprocess including resampling and rescaling were carried by MATLAB software (MathWorks, Natick, MA, USA); The radiomics- based ML execution and DL pipelines were realized with Python 3.8.5 implanted in Kears with TensorFlow 2.7.0 on workstation equipped with Radeon 5900 32G CPU and NAVIDA GTX 3080 16GB GPU.

## **2.1 Data Description**

83 NSCLC patients undergoing lobectomy or wedge resection from our institution were retrospectively studied with IRB approval. Each patient received a pre-operative CT scan, and the GTV was contoured by experienced radiologists. According to the recording of last follow-up, sixty-two patients were defined as disease-free survival without any recurrence; among the rest of twenty-one patients suffered recurrence, including 7 local failures and 16 non-local failures (2 of them suffered both). Three outcome prediction schemes were executed based on three diagnosis categories (local failure/non-local failure/failure).

Patient specific outcomes for surgical patients ( $n = 83$ ) were scored based on the following categories based on follow-up CT, PET/CT, or other pathological information. A score of 1 was given to patients who experienced the specific type of recurrence and a score of 0 was given to those who did not experience this recurrence. Local failures were identified based on radiographic criteria (recurrence along the surgical suture line). Besides, regional failures meant disease recurring in regional lymph nodes; distant failures were deemed as the occurrence of distant metastatic disease. For patients undergoing sub lobar resection, the primary clinical end point was free from local failure, which is defined as cancer recurrence along the surgical suture line.

Specifically, the treatment outcomes fell into these specific categories:

**Failure** ( $F \in \{0, 1\}$ ): Cancer recurrence following treatment ( $n = 21$ )

**Local failure** (LF  $\in \{0, 1\}$ ): recurrence along the surgical suture line ( $n = 7$ )

**Non-local failure** (nLF  $\in \{0, 1\}$ ): Either regional or distant failure ( $n = 16$ )

**Regional failure** (RF  $\in \{0, 1\}$ ): recurrence in regional lymph nodes ( $n = 8$ )

**Distant failure** (DF  $\in \{0, 1\}$ ): development of metastatic disease ( $n = 14$ ).

Three basic prediction categories were divided according to the outcomes above:

(A) local failure prediction (7 local failure vs 62 disease-free).

(B) non-local failure prediction (16 local failure vs 62 disease-free).

(C) disease-free prediction (62 disease-free vs 21 failure).

## 2.2 Radiomics

### 2.2.1 Radiomics Feature Extraction

After the GTV's were segmented, the resulting CT volume and segmentation object was exported into MATLAB for analysis. For each GTV, ninety-two radiomic features were extracted as potential biomarkers for recurrence of NSCLC following treatment. This resulted in the given feature spaces each defined by a matrix of the form,

$$\mathcal{F} = (f_{i,j}) \in \mathbb{R}^{92 \times 83} \quad (1)$$

where,  $f_{i,j}$  denotes the  $i^{th}$  radiomic feature measured on the image of the  $j^{th}$  patient.

Each radiomic expression profile (i.e., each column vector of  $\mathcal{F}$ ) was designed to collectively capture tumor morphology (i.e., shape and size), intensity (i.e., first order image statistics), and texture (i.e., spatially encoded patterns of image intensity at both

fine and coarse length-scales). Texture features were averaged over thirteen unique directions to approximate a rotationally invariant system. Equation 1 was zero-mean centered and expressed as a unit-normalized z-score.

A complete list of features used in this study has shown in Figure 3. These features are separated into the following categories:

Intensity		Fine Texture		Coarse Texture	
#	Feature Name	#	Feature Name	#	Feature Name
1	Mean	39	Auto Correlation	61	Short Run Emphasis
2	Variance	40	Cluster Prominence	62	Long Run Emphasis
3	Skewness	41	Cluster Shade	63	Gray Level Non-uniformity
4	Intensity histogram kurtosis	42	Cluster Tendency	64	Gray Level Non-uniformity
5	Median	43	Contrast	65	Run Length Non-uniformity
6	Minimum grey level	44	Correlation	66	Run Length Non-uniformity
7	10th percentile	45	Differential Entropy	67	Run Percentage
8	90th percentile	46	Dissimilarity	68	Low Gray Level Run Emphasis
9	Maximum grey level	47	Joint Energy / Angular Second	69	High Gray Level Run Emphasis
10	Interquartile range	48	Joint Entropy	70	Short Run Low Gray Level
11	Range	49	Homogeneity 1 / Inverse	71	Short Run High Gray Level
12	Mean absolute deviation	50	Homogeneity 2 / Inverse	72	Long Run Low Gray Level
13	Robust mean absolute deviation	51	Info Measure Correlation 1	73	Long Run High Gray Level
14	Median absolute deviation	52	Info Measure Correlation 2	74	Grey Level Variance
15	Coefficient of variation	53	Inverse Difference Moment	75	Run Length Variance
16	Quartile coefficient of dispersion	54	Inverse Difference Normalized	76	Run Entropy
17	Energy	55	Inverse Variance	77	Small Zone Emphasis
18	Root mean square	56	Joint maximum	78	Large Zone Emphasis
19	Intensity histogram mean	57	Sum Average	79	Gray Level Non-uniformity
20	Intensity histogram variance	58	Sum Entropy	80	Gray Level Non-uniformity
21	Intensity histogram skewness	59	Sum Variance	81	Size Zone Non-uniformity
22	Intensity histogram kurtosis	60	Joint Variance	82	Size Zone Non-uniformity
23	Intensity histogram median			83	Zone Percentage
24	Intensity histogram minimum grey			84	Low Gray Level Size Emphasis
25	Intensity histogram 10th			85	High Gray Level Size Emphasis
26	Intensity histogram 90th			86	Small Size Low Gray Level
27	Intensity histogram maximum			87	Small Size High Gray Level
28	Intensity histogram interquartile			88	Large Size Low Gray Level
29	Intensity histogram range			89	Large Size High Gray Level
30	Intensity histogram mean			90	Gray Level Variance
31	Intensity histogram robust mean			91	Zone Size Variance
32	Intensity histogram median			92	Zone Size Entropy
33	Intensity histogram coefficient of				
34	Intensity histogram quartile				
35	Intensity histogram entropy				
36	Intensity histogram uniformity				
37	Maximum histogram gradient				
38	Minimum histogram gradient				

**Figure 3: Radiomic features used in this analysis, color-coded by their class (Red: 18 first order intensity features; Yellow: 20 histogram-based features; Blue: 22**

**GLCOM-based features; Orange: 16 GLRLM-based features; Green: 16 GLSZM-based features)**

**Intensity based features:** Intensity based features are defined based on the gray level histogram of each image. This can be considered a probability density function that describes the frequency of different gray levels in the image. They are understood to measure the density characteristics of a tumor. <sup>[33]</sup>

**Fine Texture Features:** Fine texture features focus on the small-scale heterogeneity of a tumor within the high-resolution structure. They are calculated from the Gray Level Co-occurrence Matrix of the image. <sup>[34]</sup>

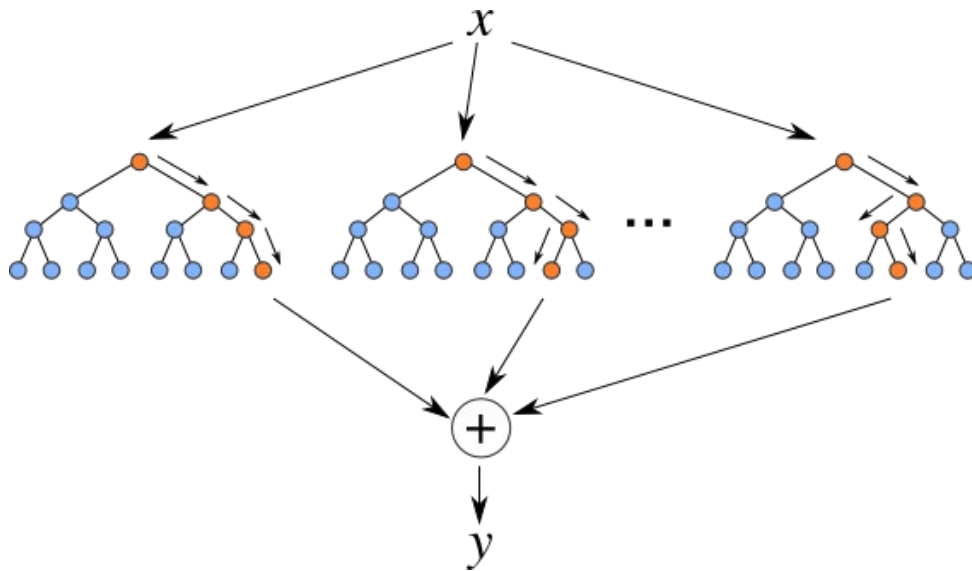
**Coarse Texture Features:** Coarse texture features capture the larger scale heterogeneity of a tumor within the low-resolution structure. They are calculated from the Gray-Level Run Length Matrix and Gray-Level Size Zone Matrix. The first measures the frequency of run-lengths, or the size of a set of consecutive pixels with the same intensity. The second provides information on the size of homogeneous zones for each gray level. <sup>[35]</sup>

Feature extraction was calculated using in-house radiomics software developed in MATLAB.

### **2.2.2 Random Forest**

A decision tree is a flowchart-like structure that can classify input data points or predict an output value based on a given input. Visualization and interpretation of decision trees are simple. In the first decade of the 21st century, decision trees learned

from data began to attract widespread attention from researchers. By 2010, decision trees were often more popular than kernel methods. In particular, the random forest algorithm introduced a robust and practical method for learning decision trees by first building many decision trees and then integrating their outputs together. [36]



**Figure 4: Schematic of a random forest model with multiple included decision trees. The output from each tree is averaged to get a singular output for the model.**

Random Forest is a supervised learning algorithm that uses ensemble learning method for classification. Random forests are suitable for a wide variety of problems - it is almost always an ideal algorithm for any shallow machine learning task. A random forest is a meta estimator that fits several classifying decision trees on various sub-samples of the dataset and uses averaging to improve the accuracy of predictive and control over-fitting. In this study, the number of trees in the forest: 100; Depth of the decision tree: 1-3.

To assess the availability of the radiomics model, 3 categories ((A) local failure prediction; (B) non-local failure prediction; and (C) disease-free prediction) were conducted following 70%-30% training-test data sample split.

### **2.2.3 Summary of Radiomics Implementation**

In the radiomics implementation, ninety-two radiomics features (shown in Figure 3) were extracted from GTV of each patient (18 first order intensity features, 20 histogram-based features, 22 GLCOM-based features, 16 GLRLM-based features, and 16 GLSZM-based features) in a 3D fashion using 64 gray levels. The first-order intensity feature is derived directly from the raw image. For the rest histogram-based features and other second-order features, the features were calculated from the discretized image.

For the extracted the feature space, the z-score normalization was performed to normalize each feature value with respect to the mean value. The multivariate association between the extracted features and clinical endpoints were investigated using a random forest (RF) machine learning algorithm following 70%-30% training-test data sample split (shown in Figure 5).

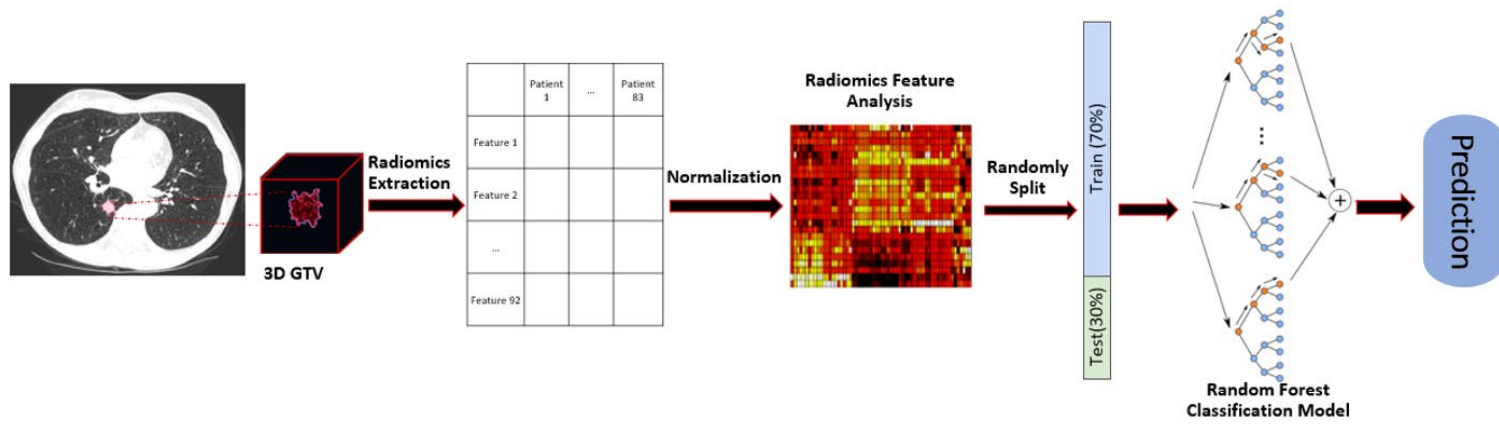


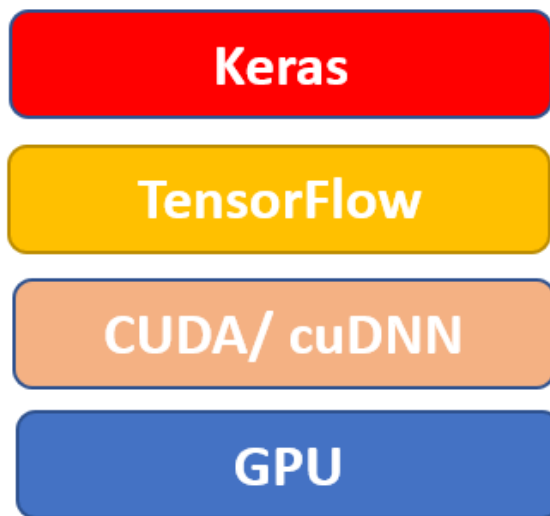
Figure 5: Workflow of radiomics implementation

## **2.3 Deep Learning**

The main idea behind deep learning (DL) was to prediction clinical outcome using convolutional neural networks (CNNs) based on CT images without manual image processing. We studied both 2D and 3D DL executions as transferred learning schemes. To make better predictions, 2D DL and 3D DL models were developed using transfer learning of CNNs based on pretrained VGG-16 architecture and the down sampling part of U-Net network, respectively. Primary data were resampled to the CT image voxel size of  $1 \times 1 \times 1 \text{ mm}^3$ . The DL trainings were implemented in Kears with TensorFlow 2.7.0 on workstation equipped with Radeon 5900 32G CPU and NAVIDA GTX 3080 16GB GPU, and the same 7:1:2 training-validation-test split was used.

### **2.3.1 Deep Learning with Keras**

The implementation of this project is based on Keras platform. (<https://keras.io>). Keras is a deep-learning framework for Python that provides a convenient way to define and train almost any kind of deep-learning model. Keras is a model-level library, providing high-level building blocks for developing deep-learning models. With TensorFlow, Keras can run seamlessly on GPUs. TensorFlow packages a well-optimized library of deep learning operations called the NVIDIA CUDA Deep Neural Network Library (cuDNN).<sup>[37]</sup>



**Figure 6: The deep-learning software and hardware stack**

### **2.3.2 Deep Learning Mechanism**

In deep learning, these hierarchical representations are almost always learned through models called neural networks. The structure of the neural network is stacked layer by layer. What each layer does with the input data in the neural network is stored in the layer's weights, which are essentially strings of numbers. In technical terms, the transformations implemented by each layer are parameterized by their weights.

The weights are sometimes called the parameters of the layer. Learning in this case means finding a set of weight values for all layers of the neural network that will allow the network to correctly map each example input to its target. But here is the thing: a deep neural network can have tens of millions of parameters. Finding the correct values for all

parameters can be a daunting task, especially considering that changing the value of one parameter affects the behavior of all other parameters. To control the output of a neural network, you need to be able to measure the distance between the output and the desired value. This is the task of a neural network loss function, also known as the objective function. The input to the loss function is the network predicted value and the true target value (i.e., what do you want the network to output), and then a distance value is calculated to measure how well the network is working in this example. The basic trick of deep learning is to use this distance value as a feedback signal to fine-tune the weight value to reduce the loss value corresponding to the current example. <sup>[37]</sup>

This adjustment is done by the optimizer, which implements the so-called backpropagation algorithm, which is the core algorithm of deep learning. The network with the smallest loss whose output value is as close as possible to the target value is the trained network. The relationship of these four is in Figure 7: multiple layers are linked together to form a network that maps input data to predicted values. The loss function then compares these predictions to the target, resulting in a loss that measures how well the network's predictions match the expected outcome. The optimizer uses this loss value to update the weights of the network.

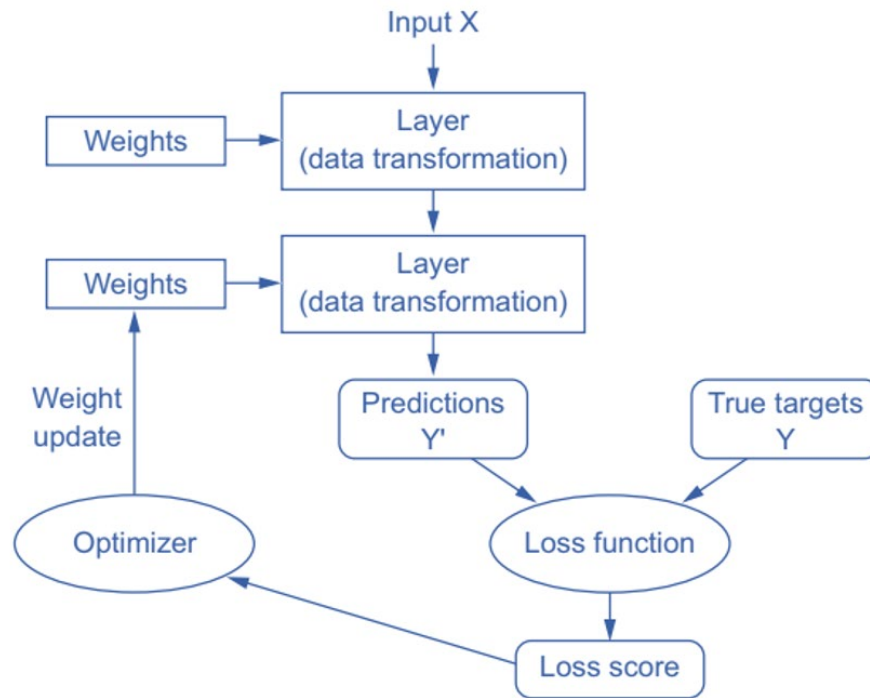


Figure 7: Relationship between the network, layers, loss function, and optimizer

### 2.3.3 Convolution Neutral Network

The fundamental difference between densely connected layers and convolutional layers is that Dense layers learn global patterns from the input feature space while convolutional layers learn local patterns. For images, what is learned is the patterns found in small 2D (or 3D) windows of the input image. Convolutional neural networks are highly data efficient when processing images (because the visual world is fundamentally translation invariant), and it requires fewer training samples to learn data representations that generalize. Convolutional neural networks can learn spatial hierarchies of patterns. The first convolutional layer will learn smaller local patterns (like edges), the second convolutional layer will learn larger patterns composed of the features of the first layer,

and so on. This allows convolutional neural networks to efficiently learn increasingly complex and abstract visual concepts (since the visual world fundamentally has a spatial hierarchy). [38]

CNNs used for deep learning tasks can be classified according to the dimension of convolution kernel in network structure. 2D CNNs utilize 2D convolutional kernels to realize different features extraction. 3D CNN solves this problem by using 3D convolutional kernels for the volume patch of 3D. The input of 2D CNNs is a single slice, which results in 2D CNNs not being able to take advantage of the context of adjacent slices. However, voxel information from adjacent areas of ROI may be useful in the process of deep learning. This disadvantage can be remedied by 3D CNNs to some extent.

### **2.3.4 Transfer Learning**

CNNs typically outperform in a larger dataset than a smaller one. Transfer learning can be useful in those applications of CNNs where the dataset is not large. The concept of transfer learning is shown Figure 8.

A pretrained network is a saved network that has been previously trained on a large dataset (usually a large-scale image classification task). If this raw dataset is large enough and general enough, the spatial hierarchy of features learned by the pretrained network can effectively serve as a general model of the visual world. These features can be used for a variety of different computer vision problems, even if these latest problems involve classes that are completely different from the original tasks. This portability of

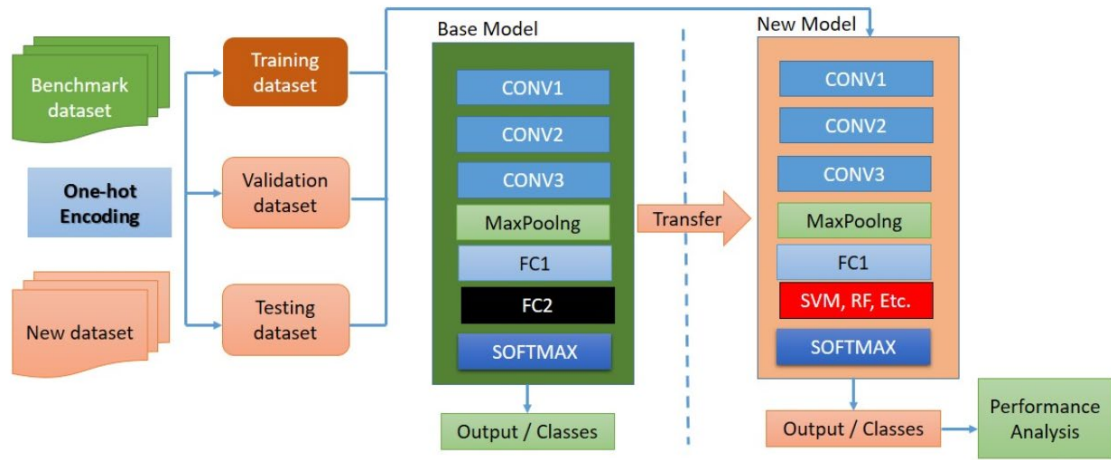
learned features across different problems is an important advantage of deep learning over many earlier shallow learning methods, and it makes deep learning highly effective for small data problems. There are two ways to use a pretrained network: feature extraction and fine-tuning. <sup>[37][39]</sup>

#### **2.3.4.1 Feature Extraction**

Feature extraction is to extract interesting features from new samples using the representations learned by the previous network. These features are then fed into a new classifier, trained from scratch. As mentioned earlier, a convolutional neural network for image classification consists of two parts: first a series of pooling and convolutional layers, and finally a densely connected classifier. The first part is called the convolutional base of the model. For convolutional neural networks, feature extraction is to take the convolutional base of the previously trained network, run new data on it, and then train a new classifier on the output.

#### **2.3.4.2 Fine-tuning**

In addition to feature extraction, fine-tuning can be used to apply some previously learned data representations from existing models to recent problems. This method can further improve the performance of the model. For a frozen model base used for feature extraction, fine tuning refers to "thawing" the top layers and adding the thawed layers and new additions. It is called fine-tuning because it only slightly adjusts the more abstract representation of the model being reused.



**Figure 8: Concept of transfer learning**

The main idea behind deep learning (DL) in this study was to prediction clinical outcome using convolutional neural networks (CNNs) based on CT images without manual image processing. We studied both 2D and 3D DL executions as transferred learning schemes. By this mean, we do not need to build a whole new network because it is time-consuming but would not achieve ideal result. Transfer learning simplify this process and also deal with the deficiency of small dataset

### **2.3.5 Data Augmentation**

For deep learning tasks in the medical physic field, the major problem is that researchers cannot collect enough diverse medical image data. The available public medical data are either of a small amount or have similar patterns.

The main problem on small imbalanced-data sets is overfitting. Data augmentation is a powerful way to reduce overfitting when processing image data. Overfitting problems describe the phenomenon that the trained model could best predict

the results on the training data, but it cannot demonstrate a comparable result on the testing data. This indicates that the trained models lost their generalization and has the tendency to complicate the solutions. To overcome this problem, we could both increase the data amount and diversity, and optimize the model architecture (in this study, a dropout layer was added between the first two Dense layers, which would be demonstrated in DL execution part). In deep learning, the data augmentation, regularization, dropout, and batch normalization are usually used to resolve this problem.

To expand the data amount as well as the data diversity, data augmentation strategies were commonly used such as image translation, flipping, scaling, and cropping, adding noise.<sup>[30-32]</sup> Even though it is simple for naked eye, data augmentation is a powerful way to reduce overfitting when processing image data.

In both 2D and 3D DL execution, data augmentation methods that includes rotation, translation, flipping, and noise level adjustment (Figure 9) were included to enhance data sample utilization and minimize convergence problem of using original imbalanced dataset. After augmentation, the scale of three basic prediction categories were enlarged and balanced.

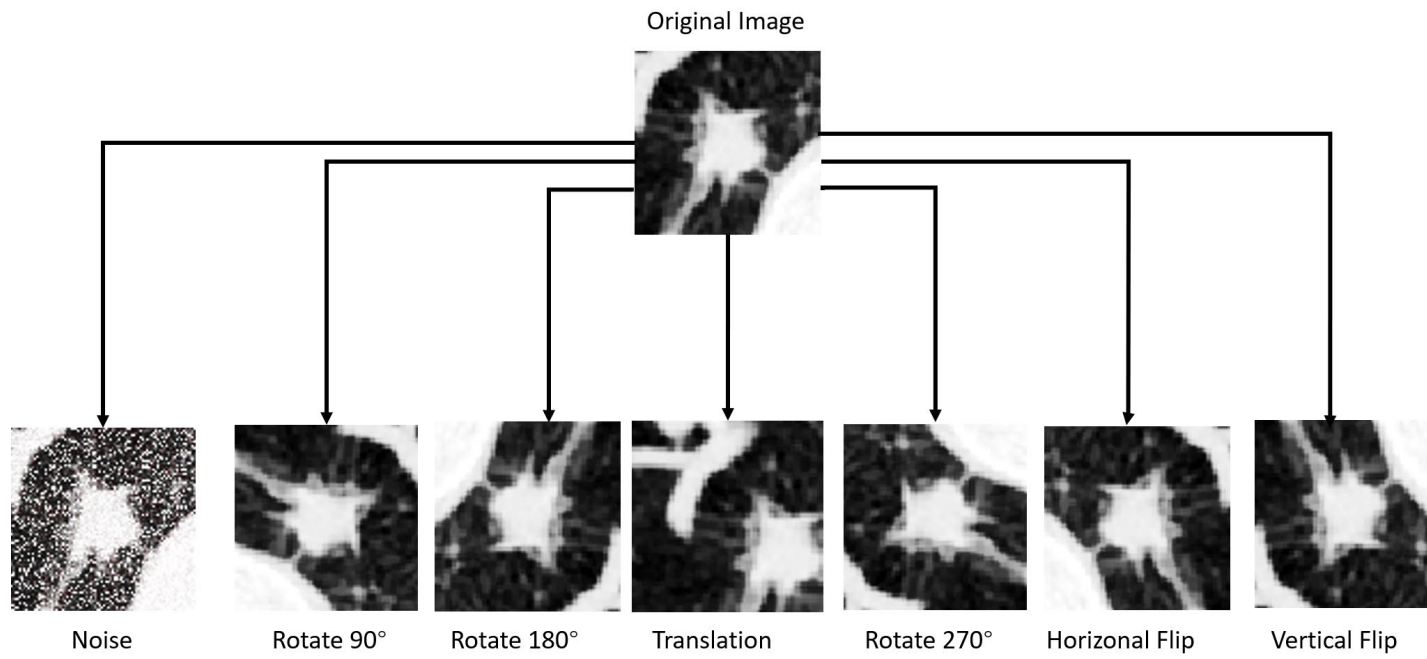
(A) local failure prediction: **336 local failure** (7\*48) vs **310 disease-free** (62\*5).

(B) non-local failure prediction: **768 non-local failure** (16\*48) vs **744 disease-free** (62\*12).

(C) disease-free prediction: **1008 disease-free** (21\*48) vs **930 failure** (62\*15).

In this study, we only chose adding noise/rotate/flip and translation (and their combination forms). Because medical imaging has specific representation, shearing/cropping/ scaling or other techniques would destroy this meaning. Only the changes mentioned above can realize expand data diversity and keep the medical representations according to our pre-research.

To sum up, data augmentation was performed to increase the diversity and number of datasets. This strategy solves the issue of imbalanced dataset and avoid the overfitting problem to some extent. Seven main augmentation strategies and their combinations (after rotating then translation) were adopted to expand the original dataset by a factor of 48 (in local failure/non-local failure/failure group). 3D operation cannot be shown in figure, but with the same principle to make data augmentation.



**Figure 9: Data augmentation**

### 2.3.6 2D Deep Learning Execution

In the 2D DL execution (Figure 10), an 8x8 cm<sup>2</sup> axial CT field-of-view (FOV) centered at GTV was adopted to represent patient-specific image input. To match the input size of pre-trained 2D network, images were broadcast to 3 channels with the dimension of 80 × 80. That is, the input of the neural network is a three-channel image with an 80×80×3 shape size following the dataset specification, while the output is categorical binary label vectors, i.e., [1,0] and [0,1], which correspond to short-term and long-term survival groups, respectively. To deal with relatively small data size in this work, the convolutional base loaded the weights that were pre-trained on ImageNet as a transfer learning scheme. This model comprised a classic VGG-16<sup>[40]</sup> convolutional base and a stack of Dense layers, and the convolutional base was loaded by weights that were pre-trained using ImageNet <sup>[41]</sup>.

For each convolutional layer, the filter size was 3×3 with padding and stride of 1. Max-pooling was performed over a 2×2-pixel window, with a stride of 2. The self-defined dense classifier connects with convolutional base and consists of five Dense layers with size of 1024, 1024, 512, 256, and 3, respectively. The output was a binary diagnosis label. To avoid the occurrence of overfitting, a dropout layer was added between the first two Dense layers. Soft-max activation was used in output layer. The last three convolutional layers were set as free parameters for training.

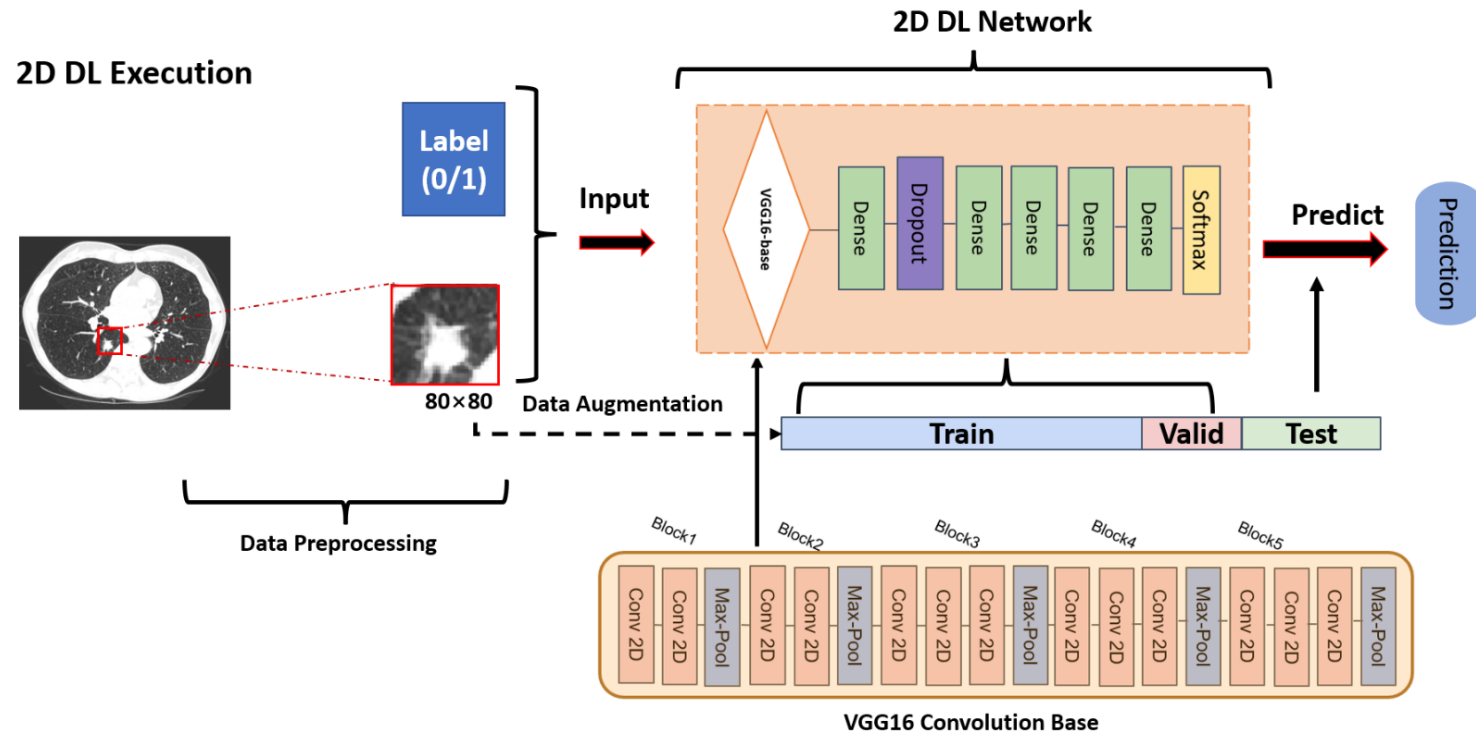


Figure 10: 2D DL execution

### 2.3.7 3D Deep Learning Execution

2D CNNs not being able to take advantage of the context of adjacent slices. However, voxel information from adjacent areas of ROI may be useful in the process of deep learning. This disadvantage can be remedied by 3D CNNs to some extent. In the 3D DL execution (Figure 11), an 8x8x8 cm<sup>3</sup> CT FOV centered at GTV was adopted to represent patient-specific image input. As a transfer learning problem, the convolutional base was loaded with weights from pre-trained results that were derived from a medical image segmentation training work. The DL model was developed as the encoding path of classic U-Net <sup>[42]</sup>. The convolutional base consisted of four convolutional blocks. Each convolutional block was stacked by 2 or 3 3D convolutional layers and a 3D max-pooling layer.

In each convolutional layer, the filter size was 3×3×3 with padding and stride of 1. Max-pooling was performed over a 2×2×2-pixel window with a stride of 2.

The dense part was the stack of two dense layers with a size of 1024 and 2. Soft-max activation was used in the output layer.

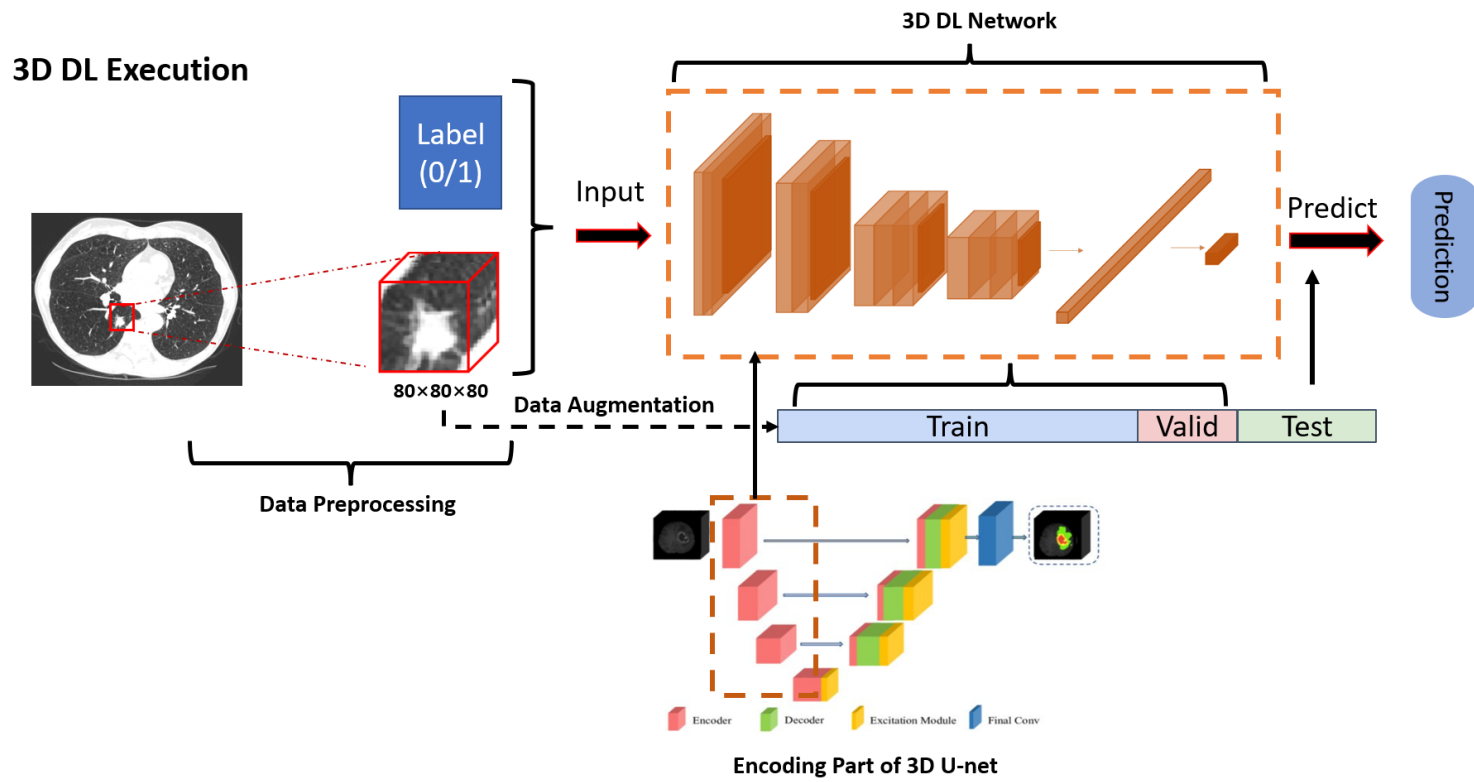


Figure 11: 3D DL execution

### 2.3.8 Summary of Deep Learning Execution

In this subsection, we summarize three crucial functions that are core components of the compilation step during 2D and 3D DL training. <sup>[37]</sup>

The role of the activation function is to convert the input of the upper layer of the neural network through the nonlinear transformation of the neural network layer, and then obtain the output through the activation function. Soft-max activation was used in the output layer in this study. The sigmoid function is the most used activation function in traditional neural networks and was once regarded as the core of neural networks. Although soft-max and sigmoid can be transformed into the same mathematical expression in the case of binary classification, it does not mean that the two have the same meaning, and the input and output of the two are different. The result obtained by sigmoid is "the probability of being assigned to the correct category and the probability of not being assigned to the correct category", and the result obtained by soft-max is "the probability of being assigned to the correct category and the probability of being assigned to the wrong category". In other words, in the classification problem performed by soft-max regression, the classes are mutually exclusive, that is, an input can only be classified into one class. Therefore, soft-max is suitable for prediction in this study, specifically recurrence or disease-free.

The purpose of the loss function is to measure the difference between the predicted value of the neural network's output and the actual value, so that the network is heading

in the right direction. The loss function we chose in this study is categorical-crossentropy. It is used to measure the distance between two probability distributions, where the two probability distributions are the probability distribution of the network output and the true distribution of the labels. By minimizing the distance between these two distributions, the network is trained so that the output is as close to the true labels as possible.

The optimization function is the mechanism for updating the network based on the training data and the loss function. The adam function, an algorithm that performs first-order gradient optimization on a stochastic objective function, is used in this study. It is characterized by high computational efficiency and low memory requirements.

Overall workflow of both 2D and 3D DL executions is shown in Figure 12. In both 2D and 3D DL execution, data augmentation methods that includes rotation, translation, flipping, and noise level adjustment were included to enhance data sample utilization and minimize convergence problem of using original imbalanced dataset. The DL trainings were implemented in TensorFlow 2.7.0 based in a NAVIDA GTX 3080 16GB GPU, and the same 7:1:2 training-validation-test split was used.

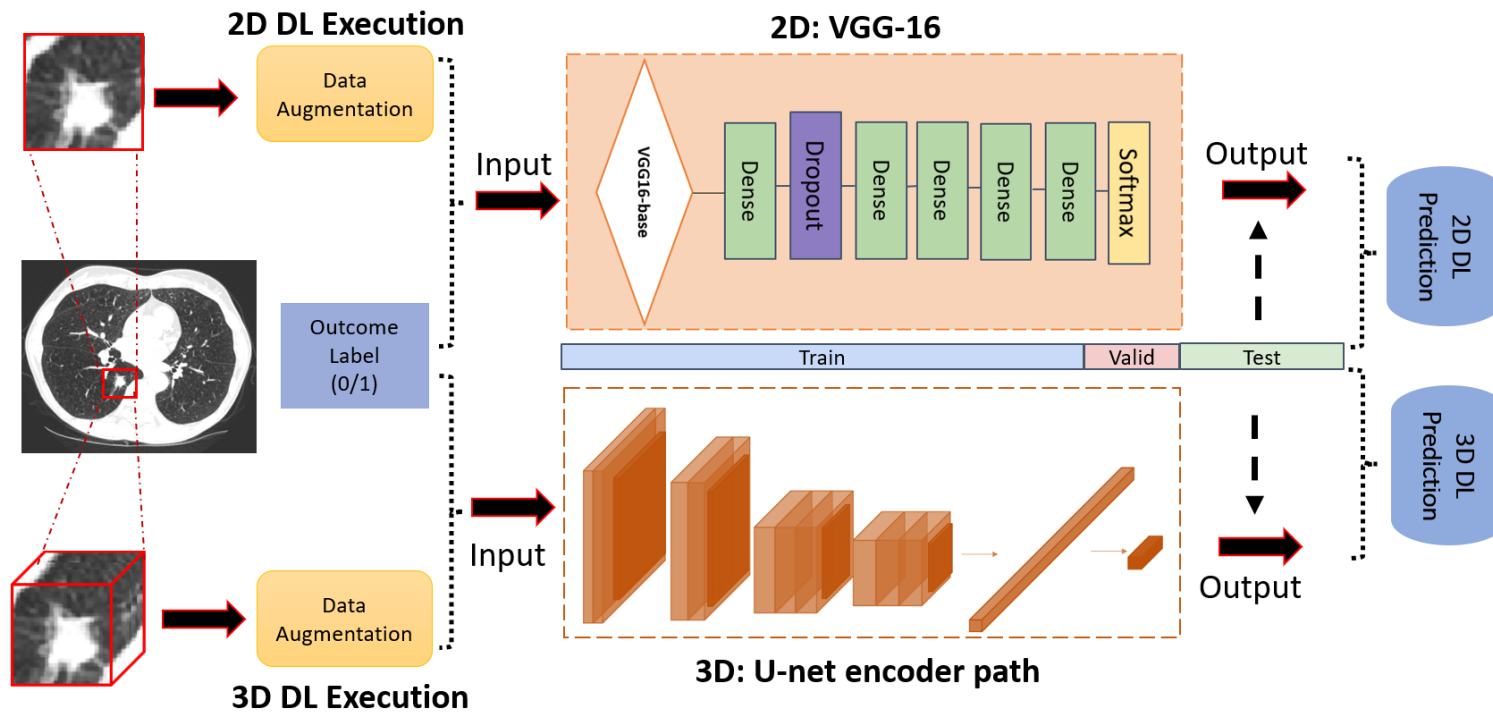


Figure 12: Overall workflow of 2D & 3D execution

## **2.4 Model Evaluation**

Taking together, three outcome prediction tasks, including local failure prediction, non-local failure prediction, and disease-free prediction, were studied. In each task, three different models (radiomics, 2D-DL, and 3D-DL) were trained, and each model was trained in 20 versions with random training-validation assignments.

Sensitivity, specificity, accuracy, and ROC results of different models were compared. Wilcoxon signed-rank test was used with a significance level of 0.05 when applicable.

Finally, the grouped cross-correlation (CC) matrix of saliency map in 3D-DL model, which evaluates DL spatial attention pattern, was analyzed to study potential survival-specific CT image inherent patterns.

### 3. Results

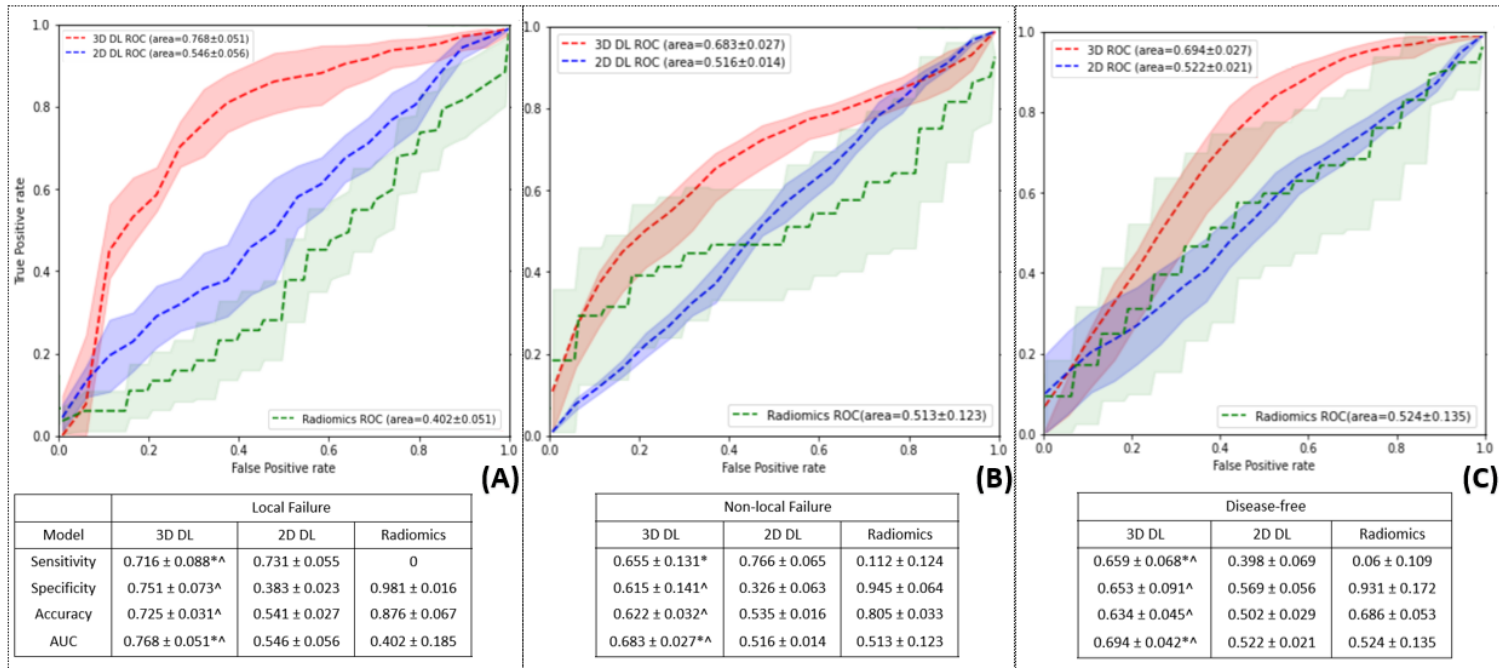
Sensitivity, specificity, accuracy, and ROC results of different models were compared. Figure 13 summarizes the models' performances in all three prediction tasks. Figure 13 (A), (B), and (C) represented the prediction of local failure, non-local failure, and disease-free, in which green curve demonstrating the ROC of radiomics model, blue referring 2D DL model, and red meaning 3D DL.

In general, the radiomics approach showed limited prediction performances. Due to the limitation in its capability of handling imbalanced dataset distribution, prediction accuracy results from the radiomics approach were unreliable with non-effective sensitivity/specificity results following sample size distribution. As a robust model descriptor, ROC results demonstrated that DL approaches achieved better prediction performances than the radiomics approach in this study.

Wilcoxon signed-rank test was used with a significance level of 0.05 when applicable. In DL execution, 3D model achieved prominent improvements in all statistics than 2D model with significance ( $P < 0.05$ ). While 2D-DL design only showed significant improvement from radiomics models in local failure prediction (ROC AUC =  $0.546 \pm 0.056$ ), 3D-DL design achieved the best performance in all three prediction tasks (local failure ROC AUC =  $0.768 \pm 0.051$ , non-local failure ROC AUC =  $0.683 \pm 0.027$ , non-failure ROC AUC =  $0.694 \pm 0.042$ ). The highest accuracy achieved by 3D-DL was found in local failure prediction ( $0.725 \pm 0.031$ ). Although the accuracy rate can judge the total accuracy rate, it

cannot be used as a good indicator to measure the results in the case of imbalanced samples. Due to the imbalance of samples, the results with high accuracy in radiomics model were not dependable. That is, if the sample is not balanced, the accuracy will be lost.

Still, the radiomics approach showed limited prediction performances in this study.



**Figure 13: ROC curves and quantitative results (sensitivity, specificity, accuracy, AUC)**  
**(A) local failure prediction; (B) non-local failure prediction; and (C) disease-free prediction.**

**Green: radiomics; Blue: 2D-DL; Red: 3D-DL.**

**\*: significant improvement from radiomics results; ^: significant improvements from 2D-DL results**

To further evaluate the 3D DL model, the grouped cross-correlation (CC) matrix of saliency map in 3D-DL model, was analyzed to study potential survival-specific CT image inherent patterns. The saliency map is defined as the gradient of output class activation over the input image, which indicates how important each pixel is with respect to the final classification results of the neural network.

Figure 14 illustrates the CC matrix results of saliency map in all three prediction tasks. The x and y axis represents the patients in (A) local failure group: 7 local failure cases vs 62 disease-free cases; (B) non-local failure group: 16 non-local failure cases vs 62 disease-free cases; (C) disease-free group: 21 failure cases vs 62 disease-free cases.

As seen, higher CC values were found within each patient group (mean CC value of 6 groups = 0.14), while cross-group CC values were lower (mean CC value = 0.07). This observation suggests that the designed 3D DL model successfully extracted survival-specific latent patterns in CT that cannot be visually captured by human readers.

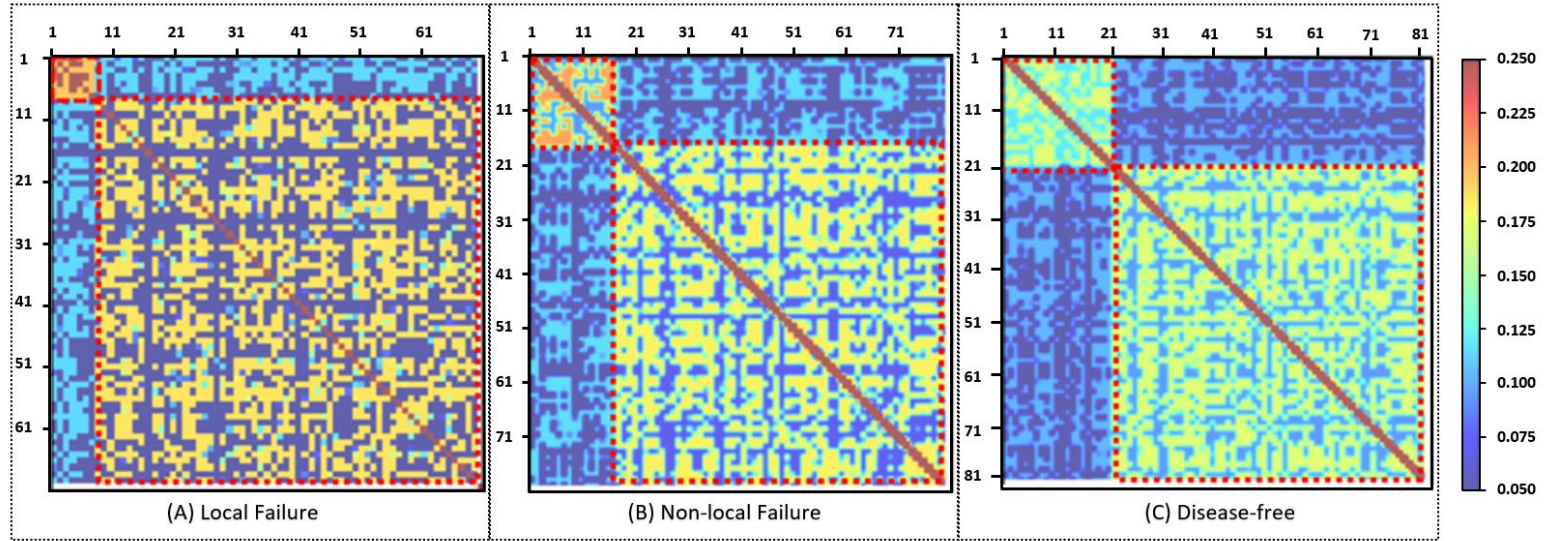


Figure 14. The CC matrix of saliency maps of 3D-DL model. The x and y axis represents the patients in each group.

## 4. Discussion

The types of lung cancer surgery can be generally divided into two categories: local resection and extended resection. Local resection refers to surgical procedures that remove less than one lobe, including segmentectomy and wedge resection. Mainly suitable for small, frail, poor lung function or very early lung cancer. Relatively speaking, extended resection refers to the surgical resection beyond one lobe, including bronchial sleeve lobectomy, pulmonary artery sleeve lobectomy, and total lobectomy. Compared with local excision, extended excision may lead to higher incidence of complications and mortality, which may lead to low long-term survival rate and poor quality of life of patients to a certain extent. Under the premise of strictly grasping the surgical indications, how to weigh the scope of surgical resection has become a tradeoff for the treatment plan. Especially for NSCLC patients, the control of local recurrence is an important indicator for prognosis evaluation. Despite the improvement of treatment methods in recent years, about 10-20% of surgical patients experience recurrence after surgery. This is largely attributable to the lack of precise boundaries between the stages of NSCLC, resulting in limited resection or incomplete lymph node dissection; on the other hand, if the resection is too large, it may lead to the above complications and increased mortality. [7-8]

This study compared radiomics and deep learning for the prediction of postoperative outcomes in patients with early-stage NSCLC surgery, and initially

concluded that 3D DL has obvious advantages in the learning of small-sample imbalanced datasets. Through the transfer learning-based scheme, coupled with fine-tuning technology and data augmentation, the problems of data set imbalance and over-fitting are solved. If there is enough large sample data to train the model, the prediction accuracy will further increase. This means that pretreatment CT images can be potentially predicted to cancer recurrence or not by means of deep learning. This result will assist the implementation of surgical options, such as expanding or reducing the scope of resection, combined with radiotherapy and chemotherapy. In this way, under the premise of ensuring the quality of life of the patient, the cancerous lesions can be clearly identified to the greatest extent possible, and the progression and metastasis of NSCLC can be prevented, thereby improving the survival rate of the patient.<sup>[15]</sup>

In most current real-world applications, 2D execution of model design is likely to be favored for its lighter model design and less computational load (such as being able to run and complete model training on the CPU). In contrast, methods employing 3D CNNs are computationally expensive and can only be attempted to run when a GPU is available. It is absolutely hard to run on a CPU. <sup>[43]</sup> It is worth mentioning that 2D-DL design only showed significant improvement from radiomics models in local failure prediction, whereas 3D-DL scenarios achieved the best performance in all three prediction tasks (shown in Figure 13). This result suggests that 3D-based model design is believed to be

superior to 2D-based model design with additional volumetric information. With additional high-throughput computation capability, deep neural network may extract more sufficient radiography information based on the 3D patch.

As for radiomics method, although it is of great value in tumor diagnosis, staging, and prognosis, there are still many challenges before the related results can be translated into clinical applications: there is still a lack of uniform standards for scanning parameters and reconstruction algorithms for imaging equipment from different manufacturers. Even with the same equipment, the differences in the injection time and amount of contrast agent, the thickness of the scanning layer, and the convolution kernel will have a potential impact on the calculation of features. Radiomics uses quantitative methods to calculate the characteristics of lesions, and any quantitative research requires a set of standardized and standard processes and quality control systems. Due to the short development time of radiomics in the medical field, there are no strict standards for reference in lesion segmentation, feature calculation, screening, statistical analysis, and predictive modeling, resulting in discrepancies between the research results of different centers or even the same center. varying degrees of difference. <sup>[44]</sup> Therefore, establishing a consensus among radiomics experts and standardizing the analysis process of radiomics is an urgent problem to be solved. Although some studies have shown that 3D radiomics features are more robust compared to 2D (single slice) measurements, factors such as tumor volume,

noise characteristics, and image resolution significantly affect radiomics analysis. <sup>[45]</sup> This may also partly explain the unsatisfactory performance of the radiomics model in this study.

What is more, the basic biological meaning of radiomics features and deep features of deep learning has not been clarified in the published literature, which makes the interpretability of radiomics features and deep features not strong, hindering their further development. Therefore, only by clarifying and mastering the biological meanings contained in these features or being able to explain the meanings, as well as the related influencing factors, can the research of radiomics and deep learning reach a new level. At present, most of the research on radiomics and deep learning is retrospective analysis, and the relevant research results still need to be evaluated by a large-scale multi-center prospective study. Our analysis was performed in the context of a retrospective study, on a relatively small cohort of early-stage lung cancer patients. Generalization of our findings to patients with other respiratory conditions remains unclear and warrants future investigation.

In addition, it should be noted imbalanced datasets and inconsistent data storage methods are the biggest constraints that restrict radiomics and deep learning in the task of training clinical outcome prediction models. To overcome the difficulties of medical imaging big data work, norms, technologies, and tools should be developed to adapt to

the standardized storage, processing, and mining of imaging information. To make better use of medical imaging big data, it is necessary to improve the transparency and convenience of data to users, improve data utilization efficiency and data quality, and conduct quantitative and structured analysis and mining of image data. This requires extensive multidisciplinary and multidisciplinary collaboration and is also an important link in transforming these models into clinical outcomes.

## 5. Conclusion

This work is a comprehensive comparison study about NSCLC surgery outcome prediction, which compared two computational approaches, radiomics and deep learning (DL), in outcome prediction of early-stage non-small cell lung cancer (NSCLC) surgery using pre-procedure CT image.

Results in this study showed that 3D-DL execution outperformed the 2D-DL in predicting clinical outcomes after surgery for early-stage NSCLC. By contrast, classic radiomics approach did not achieve satisfactory results.

The current work could be extended for potential clinical applications to optimize surgical strategies prior to the procedure, such as margin expansion, to maximize outcome of individual patients as a personalized therapy of early-stage NSCLC.

## References

- [1] Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 71.3 (2021): 209-249.
- [2] Wei, Fang, et al. "Trend analysis of cancer incidence and mortality in China." *Science China Life Sciences* 60.11 (2017): 1271-1275.
- [3] Siegel, Rebecca L., et al. "Cancer statistics, 2022." *CA: a cancer journal for clinicians* (2022).
- [4] Molina, Julian R., et al. "Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship." *Mayo clinic proceedings*. Vol. 83. No. 5. Elsevier, 2008.
- [5] Johnson, Bruce E., et al. "Small cell lung cancer: Clinical Practice Guidelines in Oncology™." *JNCCN Journal of the National Comprehensive Cancer Network* 4.6 (2006): 602-622.
- [6] Franco, Fernando, et al. "Epidemiology, treatment, and survival in small cell lung cancer in Spain: Data from the Thoracic Tumor Registry." *PloS one* 16.6 (2021): e0251761.
- [7] Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 71.3 (2021): 209-249.
- [8] Ackerson, Bradley G., et al. "Stereotactic body radiation therapy versus sublobar resection for stage I NSCLC." *Lung Cancer* 125 (2018): 185-191.
- [9] Astaraki, Mehdi, et al. "A Comparative Study of Radiomics and Deep-Learning Based Methods for Pulmonary Nodule Malignancy Prediction in Low Dose CT Images." *Frontiers in oncology* 11 (2021): 737368-737368.
- [10] Limkin, E. J., et al. "Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology." *Annals of Oncology* 28.6 (2017): 1191-1206.
- [11] Chen, Bojiang, et al. "Radiomics: an overview in lung cancer management—a narrative review." *Annals of Translational Medicine* 8.18 (2020).

- [12] Avanzo, Michele, et al. "Radiomics and deep learning in lung cancer." *Strahlentherapie und Onkologie* 196.10 (2020): 879-887.
- [13] Suzuki, Kenji. "Overview of deep learning in medical imaging." *Radiological physics and technology* 10.3 (2017): 257-273.
- [14] Zhou, S. Kevin, et al. "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises." *Proceedings of the IEEE* (2021).
- [15] Zappa, Cecilia, and Shaker A. Mousa. "Non-small cell lung cancer: current treatment and future advances." *Translational lung cancer research* 5.3 (2016): 288.
- [16] Johnson, Bruce E., et al. "Small cell lung cancer: Clinical Practice Guidelines in Oncology™." *JNCCN Journal of the National Comprehensive Cancer Network* 4.6 (2006): 602-622.
- [17] Rosell, Rafael, and Niki Karachaliou. "Optimizing lung cancer treatment approaches." *Nature reviews Clinical oncology* 12.2 (2015): 75-76.
- [18] Howlader, Nadia, et al. "The effect of advances in lung-cancer treatment on population mortality." *New England Journal of Medicine* 383.7 (2020): 640-649.
- [19] Lambin, Philippe, et al. "Radiomics: extracting more information from medical images using advanced feature analysis." *European journal of cancer* 48.4 (2012): 441-446.
- [20] Kumar, Virendra, et al. "Radiomics: the process and the challenges." *Magnetic resonance imaging* 30.9 (2012): 1234-1248.
- [21] Lafata, Kyle J., et al. "Association of pre-treatment radiomic features with lung cancer recurrence following stereotactic body radiation therapy." *Physics in Medicine & Biology* 64.2 (2019): 025007.
- [22] Lafata, Kyle J., et al. "An exploratory radiomics approach to quantifying pulmonary function in CT images." *Scientific reports* 9.1 (2019): 1-9.
- [23] Janiesch, Christian, Patrick Zschech, and Kai Heinrich. "Machine learning and deep learning." *Electronic Markets* 31.3 (2021): 685-695.

- [24] Sarker, Iqbal H. "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions." *SN Computer Science* 2.6 (2021): 1-20.
- [25] Shen, Dinggang, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis." *Annual review of biomedical engineering* 19 (2017): 221-248.
- [26] Wells III, William M. "Medical image analysis—past, present, and future." *Medical Image Analysis* 33 (2016): 4-6.
- [27] Han, W., et al. "Deep transfer learning and radiomics feature prediction of survival of patients with high-grade gliomas." *American Journal of Neuroradiology* 41.1 (2020): 40-48.
- [28] Loey, Mohamed, Gunasekaran Manogaran, and Nour Eldeen M. Khalifa. "A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images." *Neural Computing and Applications* (2020): 1-13.
- [29] Xu, Yiwen, et al. "Deep learning predicts lung cancer treatment response from serial medical imaging." *Clinical Cancer Research* 25.11 (2019): 3266-3275.
- [30] Luo, Jake, et al. "Big data application in biomedical research and health care: a literature review." *Biomedical informatics insights* 8 (2016): BII-S31559.
- [31] Maheshwari, Satyam, R. C. Jain, and R. S. Jadon. "A review on class imbalance problem: Analysis and potential solutions." *International journal of computer science issues (IJCSI)* 14.6 (2017): 43-51.
- [32] Saini, Manisha, and Seba Susan. "Deep transfer with minority data augmentation for imbalanced breast cancer dataset." *Applied Soft Computing* 97 (2020): 106759.
- [33] Aerts, Hugo JWL, et al. "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach." *Nature communications* 5.1 (2014): 1-9.
- [34] Haralick, Robert M., Karthikeyan Shanmugam, and Its' Hak Dinstein. "Textural features for image classification." *IEEE Transactions on systems, man, and cybernetics* 6 (1973): 610-621.
- [35] Tang, Xiaoou. "Texture information in run-length matrices." *IEEE transactions on image processing* 7.11 (1998): 1602-1609.

- [36] Belgiu, Mariana, and Lucian Drăguț. "Random forest in remote sensing: A review of applications and future directions." *ISPRS journal of photogrammetry and remote sensing* 114 (2016): 24-31.
- [37] Gulli, Antonio, and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.
- [38] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." *2017 international conference on engineering and technology (ICET)*. Ieee, 2017.
- [39] Rahman, Tawsifur, et al. "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray." *Applied Sciences* 10.9 (2020): 3233.
- [40] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
- [41] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.
- [42] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.
- [43] Yu, Juezhao, et al. "2D CNN versus 3D CNN for false-positive reduction in lung cancer screening." *Journal of Medical Imaging* 7.5 (2020): 051202.
- [44] Shi, Liting, et al. "Radiomics for response and outcome assessment for non-small cell lung cancer." *Technology in cancer research & treatment* 17 (2018): 1533033818782788.
- [45] Roy, Sudipta, et al. "Optimal co-clinical radiomics: Sensitivity of radiomic features to tumour volume, image noise and resolution in co-clinical T1-weighted and T2-weighted magnetic resonance imaging." *EBioMedicine* 59 (2020): 102963.