Two-Year Overall Survival Prediction in NSCLC Patients Using Pre-Treatment CT Images and Deep Neural Networks: A Multicentric Study

Zahra Khodabakhshi[1] ZAHRA.KHODABAKHSHI@USZ.CH
[1] *Department of Radiation Oncology, University Hospital Zurich, University of Zurich, Switzerland*
Isaac Shiri[2] ISAAC.SHIRILORD@UNIGE.CH
[2] *Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211, Geneva 4, Switzerland*
Habib Zaidi[2] HABIB.ZAIDI@HCUGE.CH
Nicolaus Andratschke[1] NICOLAUS.ANDRATSCHKE@USZ.CH
Stephanie Tanadini-Lang[1] STEPHANIE.TANADINI-LANG@USZ.CH

Editors: Under Review for MIDL 2022

Abstract

We propose a deep learning (DL)-based predictive model for NSCLC patients by developing a combined 2D and 3D model to include all 3D tumour information in CT images without losing spatial information. We enrolled 363 histopathological proven patients from 5 different centres with 2 year overall survival. Tumour region with size of 128×128×Z including background of tumors were extracted. 2D networks to construct 3D combination CNNs network were implemented. Feature extraction was performed in each 2D slices separately and then final layer of network were averaged to train in 3D volume of tumor. Different architecture including Xception, VGG, ResNet, Inception, and DensNet were implemented in this approach. Data of three centre (257) were used for train/validation and two centres were hold for external test set (106 patients). Considering differnt parameters parameters, VGG had highest performance with precision, recall, AUC, and accuracy of 0.72, 0.83, 0.78, and 0.75, respectively. There was no statistically significant difference between different models (delong test p-value < 0.05). Notwithstanding high variability across different datasets including geographic, ethnicity and CT scanner, image acquisition and image reconstruction, proposed models performed very well on different centres

Keywords: Deep Learning, NSCLC, Prognostic Model, Lung

1. Introduction

Lung cancer is leading cause of cancer death and NSCLC is its most prevalent subtype. Due to heterogeneity of NSCLC different patients have different response to different type of treatments. Development of prognostic model for cancer patients provides insight to make proper decision for treatment, toward personalize medicine. Although, different clinical prognostic models have been developed including models based on the tumour node metastasis (TNM), clinical information, and genenomics data, however; the performance of these models is highly limited due to not considering information regarding heterogeneity of tumours. Tomographic medical imaging could potentially provides more information regarding heterogeneity of tumours which could be mine by using ML/DL algorithms(Tomaszewski and Gillies, 2021). In conventional ML based prognostic modelling, hand crafted radiomics feature are extracted from gross tumour volume (GTV) and then based on these features,

ML modelling perform by using feature selection and classification or time-to-event analysis step. Radiomics based ML approaches have some draw backs, including non-robust modelling due to need exact GTV boundary, not including peritumoral or background information, and multiple steps which are time consuming and need fine tuning. All radiomics process including feature extraction, feature selection and classifications could be performed on only one package by using DL algorithms. DL algorithms could potentially address aforementioned challenges of radiomics . In current study we proposed DL-based prognostication model in NSCLC patients. We developed combined 2D and 3D modelling to include all 3D tumour information in CT images without losing special resolution and tested these developed models using multicentric CT dataset

2. Methods

In current study we enrolled 363 histopathological proven NSCLC Patients from 5 different centres gathered from The Cancer Imaging Archive (TCIA)(Clark et al., 2013). All patients CT images were acquired before any treatments. Continues value of survival times (calculated from start of the treatment) were dichotomized by 2-year cut-off point. We excluded right-censored patients (alive patients with follow up less than 2 years(Hosny et al., 2018). We extracted bounding box of each tumour separately by using segmentation provided by human observer and find largest box in datasets and 15 pixels were added in each axis to preserve the tumour boundary information in large tumours. We applied largest bounding box on each tumour separately considering the centre of mass of each lesion. Finally, we extracted a tumour region with size of 128×128×Z which Z is different for each different patient, and we didn't performed any resizing in Z direction to preserve image spatial resolution. For DL models, feature extraction was performed in each 2D slices separately using same architecture and then final layer of network were averaged to train in 3D volume of tumor. Different architecture of networks including Xception, VGG (Version 16), ResNet (Version 50V2), Inception (Version V2), and DensNet (Version 169) were implemented in this approach. Then we optimize each network by using different depth of each network, and feature vector. Data of three centre (257) were used for train/validation and two centres were hold for external test set (106 patients). The predictive power of each model was assessed by using the area under the receiver operating characteristic (AUC - ROC), precision, recall, and accuracy. All evaluation metrics reported on external validation set which was unseen during training process.Training was performed by using Adam optimizer, initial learning rate of 10-5 and decay 0.05 for 300 iterations. Training and validation loss were observed to avoid overfitting and early stopping were preformed after 15 epochs without changing in validation loss in maximum 1000 defined epoch. DeLong tests were performed on AUC to compare different models.

3. Results

In term of accuracy, VGG had highest performance of 0.75. In term of precision and recall, Xception and VGG had highest value of 0.77 and 0.83, respectively. Considering all four parameters, VGG had highest performance with precision, recall, AUC, and accuracy of 0.72, 0.83, 0.78 and 0.75, respectively. Figure 1 depict ROC curve for different models using 1000 bootstrapping separately and altogether for comparing different models. There were no statistically significant difference between different models(delong test p-value < 0.05).