Clinical Oncology 34 (2022) 114–127

Contents lists available at ScienceDirect

Clinical Oncology

journal homepage: www.clinicaloncologyonline.net

Overall Survival Prognostic Modelling of Non-small Cell Lung Cancer Patients Using Positron Emission Tomography/Computed Tomography Harmonised Radiomics Features: The Quest for the Optimal Machine Learning Algorithm

Mehdi Amini [*], Ghasem Hajianfar [†], Atlas Hadadi Avval [‡], Mostafa Nazari [†§], Mohammad Reza Deevband [§], Mehrdad Oveisi [¶], Isaac Shiri [*], Habib Zaidi [*||**††]

[*] *Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland*
[†] *Rajaie Cardiovascular Medical and Research Center, Iran University of Medical Science, Tehran, Iran*
[‡] *School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran*
[§] *Department of Biomedical Engineering and Medical Physics, Shahid Beheshti University of Medical Sciences, Tehran, Iran*
[¶] *Comprehensive Cancer Centre, School of Cancer & Pharmaceutical Sciences, Faculty of Life Sciences & Medicine, Kings College London, London, UK*
[||] *Geneva University Neurocenter, Geneva University, Geneva, Switzerland*
[**] *Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands*
[††] *Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark*

Abstract

Aims: Despite the promising results achieved by radiomics prognostic models for various clinical applications, multiple challenges still need to be addressed. The two main limitations of radiomics prognostic models include information limitation owing to single imaging modalities and the selection of optimum machine learning and feature selection methods for the considered modality and clinical outcome. In this work, we applied several feature selection and machine learning methods to single-modality positron emission tomography (PET) and computed tomography (CT) and multimodality PET/CT fusion to identify the best combinations for different radiomics modalities towards overall survival prediction in non-small cell lung cancer patients.

Materials and methods: A PET/CT dataset from The Cancer Imaging Archive, including subjects from two independent institutions (87 and 95 patients), was used in this study. Each cohort was used once as training and once as a test, followed by averaging of the results. ComBat harmonisation was used to address the centre effect. In our proposed radiomics framework, apart from single-modality PET and CT models, multimodality radiomics models were developed using multilevel (feature and image levels) fusion. Two different methods were considered for the feature-level strategy, including concatenating PET and CT features into a single feature set and alternatively averaging them. For image-level fusion, we used three different fusion methods, namely wavelet fusion, guided filtering-based fusion and latent low-rank representation fusion. In the proposed prognostic modelling framework, combinations of four feature selection and seven machine learning methods were applied to all radiomics modalities (two single and five multimodalities), machine learning hyper-parameters were optimised and finally the models were evaluated in the test cohort with 1000 repetitions via bootstrapping. Feature selection and machine learning methods were selected as popular techniques in the literature, supported by open source software in the public domain and their ability to cope with continuous time-to-event survival data. Multifactor ANOVA was used to carry out variability analysis and the proportion of total variance explained by radiomics modality, feature selection and machine learning methods was calculated by a bias-corrected effect size estimate known as ω^2.

Results: Optimum feature selection and machine learning methods differed owing to the applied radiomics modality. However, minimum depth (MD) as feature selection and Lasso and Elastic-Net regularized generalized linear model (glmnet) as machine learning method had the highest average results. Results from the ANOVA test indicated that the variability that each factor (radiomics modality, feature selection and machine learning methods) introduces to the performance of models is case specific, i.e. variances differ regarding different radiomics modalities and fusion strategies. Overall, the greatest proportion of variance was explained by machine learning, except for models in feature-level fusion strategy.

Author for correspondence: H. Zaidi, Geneva University Hospital, Division of Nuclear Medicine and Molecular Imaging, CH-1211 Geneva, Switzerland. Tel: +41-22-372-7258; Fax: +41-22-372-7169.
E-mail address: habib.zaidi@hcuge.ch (H. Zaidi).

https://doi.org/10.1016/j.clon.2021.11.014

Conclusion: The identification of optimal feature selection and machine learning methods is a crucial step in developing sound and accurate radiomics risk models. Furthermore, optimum methods are case specific, differing due to the radiomics modality and fusion strategy used.

Key words: Fusion; machine learning; non-small cell lung cancer; PET/CT; radiomics

Introduction

Digitally encrypted medical scans captured by different imaging modalities hold anatomical, physiological and metabolic information of underlying tumours, which can be harnessed through 'radiomics'. This emerging image analysis tool converts images into high-dimensional, mineable, quantitative features that are enriched with crucial information regarding tumour phenotype [1]. This unprecedented technology provides new opportunities for precision oncology by tackling inherent limitations of biopsy-based assays (e.g. invasive procedures and inability to characterise the whole tumour volume) for the characterisation of intra-tumour heterogeneity [2,3]. Numerous studies reported promising results regarding the high performance of radiomics models based on different modalities in different cancer types. The results from these studies proved that radiomics models are associated with various outcomes, namely tumour histology [4–6], stage or grade [4,7,8], survival [9–13] and other kinds of clinical outcome [14–16]. Moreover, the correlation between radiomics features and underlying gene expression patterns has been reported in numerous studies [17–19].

Despite the potential of radiomics for clinical applications, multiple challenges still need to be addressed and several improvements should be made to achieve an optimal model. First, the information reflected by a single imaging modality is limited to specific aspects of tumour heterogeneity. For instance, fluorodeoxyglucose (^{18}F-FDG)-positron emission tomography (PET) scanning measures glucose metabolism, reflecting the molecular status of disease, such as cellular proliferation, hypoxia and metabolism [20,21], whereas computed tomography (CT) images provide attenuation coefficients reflecting absorption of X-rays, which convey other information about lesions, such as vascularisation and/or necrosis [22]. Therefore, investigators have taken a further step and are attempting to integrate information from different imaging modalities into a single model through different levels of fusion (i.e. feature-, matrix- and image-level fusion).

Several studies reported improvements in the prognostic performance of multimodality radiomics models developed through the fusion of modalities at the feature [23,24], matrix [25,26] and image level [27,28]. Lv *et al.* [29] and Amini *et al.* [30] compared the performance of PET/CT multimodality radiomics models developed at multiple levels of fusion for survival prognosis of head and neck squamous cell carcinoma (HNSCC) and non-small cell lung carcinoma (NSCLC) patients, respectively. Both studies reported the highest performance from image-level fusion models. The results of these studies indicated that a multimodality radiomics approach has the potential to tackle the lack of comprehensiveness in malignant lesion data in single modalities. Hence, it can provide better characterisation of intra-tumour heterogeneity by extracting more meaningful features.

Another challenge to address is the selection of an optimal statistical algorithm to build the most efficient prognostic/predictive model, which best fits both the specific imaging modality and the clinical end point in use. Machine learning techniques have become the leading strategy in radiomics studies [31]. What makes machine learning algorithms more useful for radiomics analysis compared with standard statistical analysis is their ability to cope with high-dimensional features extracted from a small sample size [32]. They are also capable of capturing complex interactions between features to produce relevant, non-redundant combinations (signatures) and use these signatures to develop robust prognostic/predictive models by learning from underlying data/experience distribution [33]. Although various machine learning and feature selection strategies were devised, using different approaches might yield different results and even contradictory interpretations. Hence, it is crucial to compare the performance of the different combinations of machine learning and feature selection methods to develop the most clinically relevant radiomics model that best fits the modality and the desired outcome.

So far, several studies have carried out such evaluations on single-modality models. These studies investigated the performance of different methods for both classification [34–37] and time-to-event survival analysis [38,39]. Regarding time-to-event analysis, Parmar *et al.* [38,39] applied several algorithms in two separate studies on patients with NSCLC and locally advanced HNSCC and transformed their outcome of interest and overall survival to a binary outcome. Although outcome dichotomisation is a common procedure in stratifying patients into risk groups, it might bias prediction accuracy. Therefore, researchers have recently investigated the effect of different machine learning and feature selection methods on predicting the overall survival of HNSCC [40] and NSCLC [41] patients in a continuous time-to-event manner.

Although several studies reported on the application of different algorithms to single imaging modalities, comprehensive comparisons of the performance of various algorithms applied to multimodality radiomics models are still lacking. Here we report on a thorough framework for performance evaluation of different time-to-event predicting algorithms being applied to single- and multimodality radiomics models. In the presented framework for multimodality models, we considered different feature- and image-level fusion methods. Using the proposed framework,

we assessed the performance of different combinations of seven machine learning and four feature selection techniques as applied to single-modality (CT and PET) and multimodality radiomics model towards overall survival prediction in NSCLC patients. In addition, our retrospective study was conducted on an open dataset, using publicly available fusion, feature selection and machine learning algorithms, to ensure that predictive radiomic models are as robust and reproducible as possible. The proposed framework provides a robust method for selecting a subset of promising feature selection and machine learning algorithms suitable for single- and multimodality radiomics models for continuous (we avoided dichotomising data) time-to-event prediction modelling.

Materials and Methods

Datasets

In this study, two cohorts from two independent institutions, namely Palo Alto Veterans Affairs Healthcare System (VA) and Stanford University School of Medicine (Stanford) were involved. These datasets were adopted from The Cancer Imaging Archive open-access repository [42], originally comprising 211 histologically proven NSCLC patients. Some patients were excluded owing to the high level of noise or presence of artifacts, mis-segmentation and mis-registration errors. Overall, 87 subjects from Stanford and 95 from VA were included. All patients underwent ^{18}F-FDG PET/CT scans, prior to surgical treatment. Image acquisition parameters are reported in [30] and Supplementary Table S1 separately for VA and Stanford datasets.

Radiomics Framework

Our efforts in this study consisted of two main frameworks, namely a radiomics framework and a prognostic modelling framework (PMF). In the radiomics framework, we developed seven different radiomics modalities, including CT and PET for the single-modality strategy, two models from the feature-level fusion strategy and three models from the image-level fusion strategy. The sections below describe our radiomics framework in detail and Figure 1 illustrates the workflow adopted in the radiomics framework.

Image Segmentation

The segmentation of malignant lesions from PET images was carried out using OSIRIX® software [43], whereas CT images were segmented using an automatic region-growing algorithm implemented in the 3D-Slicer package. An experienced radiologist edited/verified both segmentations. Finally, we merged PET and CT masks to reduce segmentation errors and also to ensure that the volume of interest (VOI) is constant over all single- and multimodality models. An OR logic was used to merge the masks, i.e. each voxel was identified as tumour if the corresponding voxels from PET or CT masks were labelled tumour.

Developing Radiomics Modalities Using Different Strategies

Besides having both single-CT and single-PET models (single-modality strategy), we developed another five PET/CT fusion models, from feature- and image-level fusion strategies. To offset the plausible bias owing to the selection of a specific fusion method, we considered two different methods for feature-level fusion and three methods for image-level fusion. The first feature-level fusion technique (concatenating PET and CT features into a single feature set; ConFea) is considered a common multimodal radiomics strategy and concatenates PET and CT features into a single feature set. The second technique (averaging PET and CT features; AvgFea) has features from two modalities averaged.

Previous to image-level fusion, trivial mis-registration of PET and CT images was treated as elaborated in [30]. In addition, both volumes were resampled to the same resolution by interpolating images to isotropic voxel spacing of $2 \times 2 \times 2$ mm^3 using cubic interpolation and an anti-aliasing kernel.

For the image-level fusion strategy, three publicly available techniques, namely wavelet-based fusion (https://github.com/mvallieres/radiomics), guided filtering-based fusion (GFF; https://github.com/funboarder13920/image-fusion-guidedfiltering) and latent low-rank representation fusion (LLRR; https://github.com/hli1221/imagefusion_Infrared_visible_latlrr) were used. The wavelet fusion and GFF methods were used in previous studies for the fusion of medical images, including PET and CT, reporting promising results. The LLRR method was originally designed for the fusion of visible and infrared images, which could be analogous to PET/CT images.

The wavelet fusion method first applies the three-dimensional discrete wavelet transform to decompose CT and PET images using wavelet basis function symlet8. Then, PET and CT corresponding wavelet coefficients (eight pairs of wavelet sub-bands) are averaged to get a single set of fused wavelet coefficients. Finally, a three-dimensional inverse discrete wavelet transform is applied to the fused coefficients to obtain a fused volume.

The GFF method [44] first applies an average filter, conventionally set to 31×31, to PET and CT images to obtain a base layer from the images. Then the base layer is subtracted from the source image to generate a detailed layer. This two-scale decomposition step aims to divide source images into a separated base layer containing large-scale variations of intensities and a detailed layer capturing small-scale details. Then, a guided filtering-based weighted average method is used to fuse corresponding base and detailed layers of source images by making full use of the spatial consistency. Finally, the fused image is obtained by combining the fused base and detailed layers. In this study, we adopted the default parameters of GFF used in [44].

The LLRR method [45] first uses latent low-rank representation to decompose source images into low-rank parts (global structure) and saliency parts (local structures). Then, in order to preserve contour information, the low-rank parts and the saliency parts are fused by a weighted-

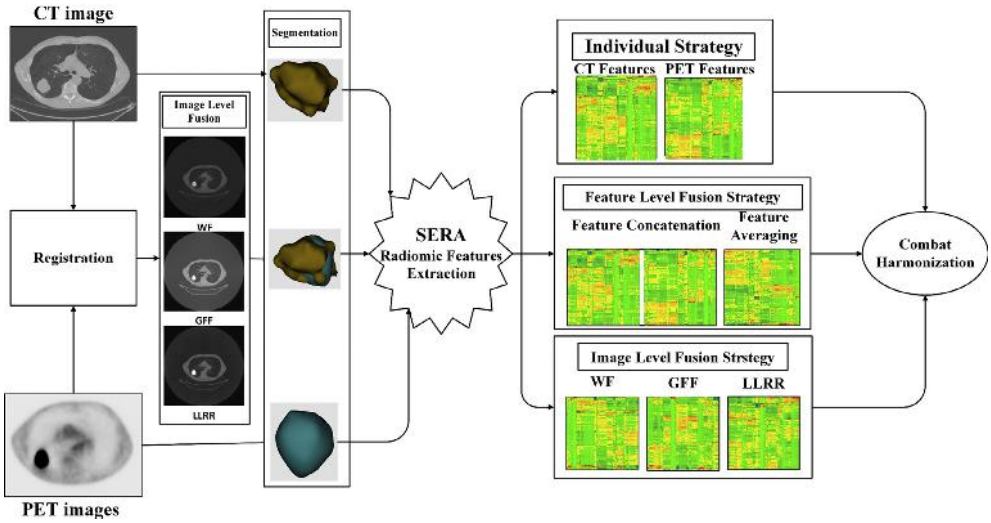

Fig 1. Flowchart of radiomics framework adopted in this study.

average strategy. The fusion of saliency parts is carried out by a simple sum strategy, which is an efficient operation in this fusion framework. Finally, the fused low-rank part and fused saliency part are combined to obtain the fused image. The default parameters for LLRR used in [45] were adopted in this study.

Overall, seven different radiomics models, namely individual CT, ConFea, AvgFea, wavelet fusion, GFF, LLRR and individual PET models were developed in order to identify the best combination of feature selection and machine learning methods for single-modality, and feature- and image-level fusion strategies towards NSCLC overall survival prognosis. All image processing and image fusions were carried out in Matlab® 2020a.

Feature Extraction

Images were interpolated to isotropic voxel spacing of $2 \times 2 \times 2\ mm^3$, so that images from different modalities were standardised to a uniform size, to render texture features rotationally invariant. Moreover, in order to make the calculation of texture features tractable, intensities inside the VOIs were quantised into 64 discretised grey levels. The Standardized Environment for Radiomics Analysis Package (https://github.com/ashrafinia/SERA) [46], a Matlab®-based framework, compliant with the guidelines of the Image Biomarker Standardization Initiative [47], evaluated in the context of multicentre standardisation studies [47,48] for improved features reproducibility, was used to extract a total of 225 features from each VOI. This included 79 first-order (morphological, statistical, histogram and intensity-histogram) features, 136 three-dimensional texture features extracted from GLCM, GLRLM, GLSZM, GLDZM, NGTDM, NGLDM matrices and 10 moment-invariant features.

Feature Harmonisation

Multicentre radiomics studies are challenging as radiomics features are notorious for exhibiting variable sensitivity to the so-called 'batch effect', i.e. differences in scanner model, acquisition protocols and reconstruction settings [49]. Hence, in order to generate consistent robust models using multicentre datasets, harmonisation of features from different datasets prior to pooling them is highly recommended [50]. In this study, we used a well-known harmonisation method, ComBat (https://github.com/Jfortin1/ComBatHarmonization; 'combining batches'). ComBat harmonisation removes batch effects based on an empirical Bayes framework using Bayes estimations for the location-scale parameters, including mean and variance for each variable [51]. As a result, it generates batch-specific transformations to present all data in a common space, devoid of centre effects [52]. ComBat has been reported as an efficient method for eliminating multicentre effects from radiomics features [53,54].

Prognostic Modelling Framework

This study was conducted using two NSCLC cohorts from independent centres. For a comprehensive evaluation of our models, two train/test arrangements were investigated; using VA as training and Stanford as testing (untouched external validation) cohorts (noted as VAS partition) and alternatively vice versa (noted as SVA partition). A PMF was developed to generate radiomics signatures using different feature selection algorithms, train predictive models utilising different machine learning methods, optimise their hyper-parameters and, eventually, evaluate the models by determining their overall survival prognostic performance of NSCLC patients via Harrell's concordance index (C-index). Figure 2 illustrates our four-step PMF, including: (i) feature selection, (ii) machine learning hyper-parameter optimisation, (iii) model training and (iv) model evaluation.

Feature Selection Methods

Four different feature selection methods were used in this study. The first method (referred to as C-index) is a hybrid

feature selection method consisting of a filter followed by a wrapper method based on univariate Cox proportional hazard regression. In this method, Spearman's correlation was used as a measure of redundancy to omit one of each feature pairs with correlation coefficients higher than 0.9. The remaining features were fed into the univariate Cox proportional hazard model with 100 repetitions using bootstrap resampling and the top 10 features with the best performance (highest average C-index) were selected. Moreover, we used three wrapper feature selection methods by implementing random forest variable selection, with tree minimum depth methodology (https://rdrr.io/cran/randomForestSRC/man/var.select.rfsrc.html) [52,53]. There are three different approaches in this method, including minimum depth, variable hunting and variable hunting with variable importance (VH.VIMP). For variable hunting and VH.VIMP, the process was repeated 50 times and the top P ranked variables were selected, where P was the average model size (rounded up to the nearest integer) and variables were ranked by frequency of occurrence.

Machine Learning Methods

The performance of seven machine learning methods was assessed, including Cox proportional hazard [54], Cox model fitted by likelihood-based boosting (CoxBoost) [55], Lasso and Elastic-Net regularised generalised linear model (glmnet) [56], random survival forest (RSF) [57], gradient boosting with component-wise linear model (glmboost) [58], generalised boosted regression model (GBM) [59,60] and survival tree [61]. All feature selection and machine learning methods used in this work were able to cope with continuous time-to-event data. Our selection criterion was public availability to ensure an unbiased and easily reproducible evaluation. The packages of each machine learning method are listed in Table 1.

Hyper-Parameter Optimisation

Hyper-parameter optimisation (machine tuning) was carried out for all machine learning methods (except Cox proportional hazard) using grid search. It was guided by the performance metric, C-index, calculated by three-fold cross-validation in the training dataset. The hyper-parameters and their ranges can be found in Table 1.

Model Evaluation

Subsequently after identification of hyper-parameters, the optimal models were applied to the test cohort with 1000 repetitions using bootstrap resampling. VA and Stanford datasets were cycled as the training and test cohorts, respectively, and the results were averaged. MLR package in R 3.6.2 was used for hyper-parameter optimisation, model training and model evaluation.

Variability Analysis with Multifactor ANOVA

The three main factors affecting the performance of our radiomics-based prognostic models were radiomics modality (seven models, including single- and multimodality), feature selection methods (four algorithms) and machine learning algorithms (seven algorithms). To quantify the variance of C-indices owing to these three factors and their interactions, we used a multifactor ANOVA test. Moreover, the proportion of total variance explained by each factor was calculated by a bias-corrected effect size estimate known as ω^2 [62]. Finally, we carried out a multiple comparison test, once separately on each strategy (individual, feature- and image-level fusion) and once on all models combined, to determine which feature selection/machine learning or combination of these methods provides the most significant result.

Results

Although two separate train/test arrangements (VAS and SVA) were used in this study, some results are only reported for VAS partition as the frequency of events (deaths) in the VA dataset was higher than in the Stanford set (37% versus 26%).

Selected Features

The selected features from each radiomics modality in the VAS partition (columns) by each feature selection method (rows) are presented in Supplementary Table S2. Moreover, we investigated the distribution of features selected by different feature selection methods from different radiomics modalities, within different feature families (morphological, statistical, intensity histogram and intensity volume histogram, and texture features). Figure 3 shows the pie chart of features distribution within different feature families, in each radiomics modality (Figure 3a–g) and from each feature selection method (Figure 3h–k).

Prognostic Performance of Different Feature Selection and Machine Learning Combinations

The prognostic performance of all cross-combinations of feature selection and machine learning methods towards overall survival prediction of NSCLC patients was evaluated for all radiomics models. The performance of all models is reported with C-index in the test (external validation) cohort. Figure 4 shows the performance of all radiomics modalities (columns) trained by all combinations of feature selection and machine learning methods (rows), by averaging the VAS and SVA results. They are also presented showing more details, containing standard deviations and 95% confidence intervals, in Supplementary Table S3. Supplementary Figure S1 shows the C-index of all models separately for VAS and SVA partitions. In the VAS partition, many models within image-level fusion achieved C-indices higher than 0.7 (e.g. LLRR fusion model with minimum depth as the feature selection method and glmboost as the machine learning method, achieved a C-index of 0.73).

Table 2 lists the best combinations of feature selection/machine learning methods for each radiomics modality. The

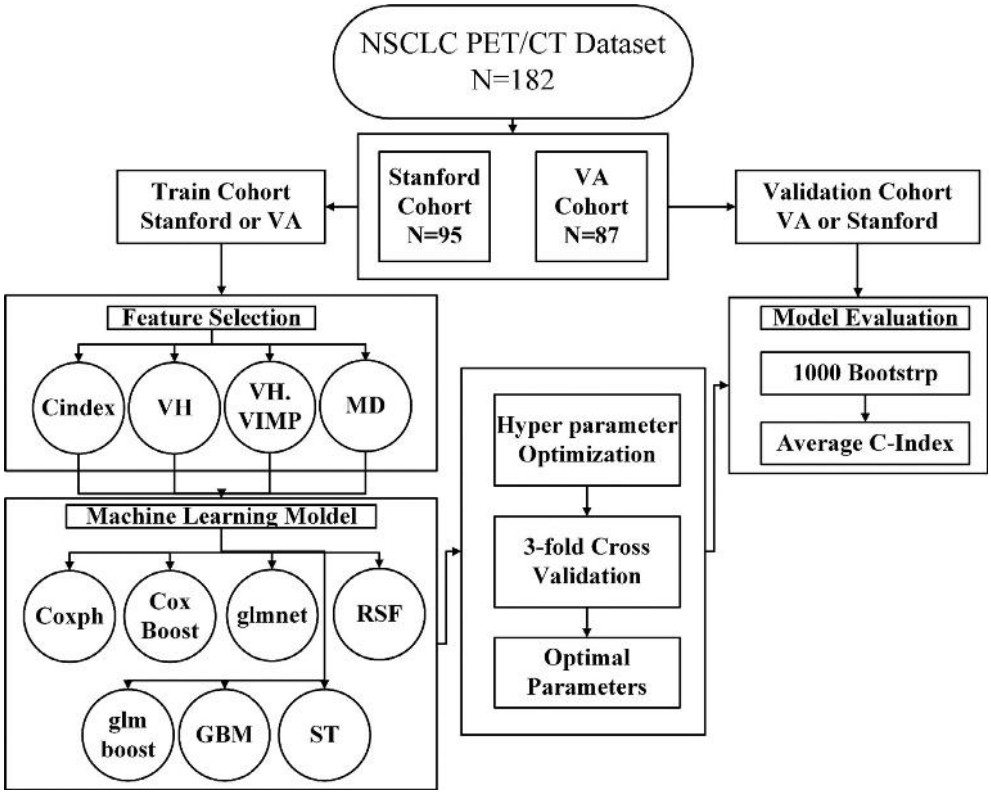

Fig 2. Flowchart of prognostic modeling framework (PMF) employed in this study.

best performance was achieved by minimum depth/glmnet and minimum depth/glmboost applied to the LLRR model (C-index = 0.68) and C-index/glmnet applied to the wavelet fusion model (C-index = 0.68). In addition, in CT-only, PET-only and ConFea radiomics modalities, the best prognostic model was trained with glmnet. The minimum depth feature selection method achieved the best results in CT-only, ConFea, GFF and LLRR radiomics modalities and frequently appeared within the best combinations (Table 2).

In Table 3, the average performance of models with a fixed feature selection or machine learning method (e.g. row: minimum depth; column: LLRR indicates the average C-index of all LLRR models with minimum depth fixed as the feature selection method, and varying machine learning methods). For models in the single-modality strategy, VH.VIMP as the feature selection method (average C-index = 0.61) and glmnet and glmboost as the machine learning methods (average C-index = 0.61) had the highest average,

Table 1
Machine learning methods, corresponding packages and hyper-parameters used in this study

Model	R package	Hyper-parameter
Coxph	Survival	-
CoxBoost	CoxBoost	maxstepno: 50-500
glmnet	glmnet	s: 0.001-0.1 alpha: 0-1
RSF	randomForestSRC	ntree: 100, 500, 1000 mtry: 1-10 nodesize: 1:20 splitrule: logrank, logrankscore
glmboost	mboost	mstop: 50-500
GBM	gbm	n.trees: 100, 500, 1000 interaction.depth: 1-5 n.minobsinnode: 3-5 shrinkage: 0.01,0.05,0.1
ST	rpart	minsplit: 1-20 maxdepth: 1-30

CoxBoost, Cox model fitted by likelihood-based boosting; Coxph, Cox proportional hazard; GBM, generalised boosted regression model; Glmboost; gradient boosting with component-wise linear model; glmnet, Lasso and Elastic-Net regularised generalised linear model; RSF, random survival forest; ST, survival tree.

whereas variable hunting as the feature selection method (average C-index = 0.58) and survival tree as the machine learning method (average C-index = 0.57) had the worst performance. Regarding the feature-level fusion strategy, minimum depth and VH.VIMP as the feature selection methods (average C-index = 0.6) and CoxBoost and glmnet as the machine learning methods (average C-index = 0.61) had the best performance, whereas C-index as the feature selection method (average C-index = 0.59) and RSF as the machine learning method (average C-index = 0.58) had the lowest results. In image-level fusion strategy, minimum depth as the feature selection method (average C-index = 0.63) and CoxBoost and glmnet as the machine learning methods (average C-index = 0.64) reached the highest results, whereas VH.VIMP as the feature selection method (average C-index = 0.6) and survival tree as the machine learning method (0.57) had the worst performance. Finally, considering all models together, minimum depth as the feature selection method (average C-index = 0.61) and glmnet as the machine learning method (average C-index = 0.62) had the best results, whereas survival tree as the machine learning method (average C-index = 0.57) had the worst performance.

Results from Variability Analysis with ANOVA

We applied the ANOVA test once separately on each radiomics modality, once separately on each strategy (single-modality, feature- and image-level fusion) and once on all models together. The corresponding results are shown in Table 4. As Table 4 suggests, the different radiomics modalities and different strategies are sensitive to different factors. For instance, in the CT models, the selection of feature selection method is significantly effective (P-value < 0.005), whereas the machine learning method is not (P-value = 0.099), whereas unlike CT, in PET models, changing the machine learning method has a significant effect (P-value < 0.02) whereas the feature selection method has not (P-value = 0.097). For other radiomics modalities, reference is made to Table 4. Regarding the different radiomics strategies, single-modality models were affected by all three factors; feature-level fusion models were affected by changing the radiomics model and machine learning algorithm, whereas the feature selection method did not have a significant effect on their outcome (P-value = 0.096); image-level fusion models were sensitive to all three factors. Finally, considering all models together, all three factors were significantly effective. The proportion of variance explained by each factor and their interactions are shown in Figure 5. It varied owing to the strategy. However, in general, the most effective factor was the selection of machine learning methods, except for feature-level fusion strategy, in which radiomics modality was the most effective factor.

Figure 6 illustrates the results from a multiple comparison test to identify which feature selection (a, d, g, j), machine learning (b, e, h, k) or combination (c, f, i, l) of methods provides significantly different results in a single-modality strategy (a–c), a feature-level fusion strategy (d–f), an image-level fusion strategy (g–i) and within all models (j–l). As an example, considering the results of all models together, regarding feature selection methods (Figure 6j), the minimum depth feature selection method (marked blue), significantly outperformed variable hunting methods (marked red), whereas it had comparable results with VH.VIMP (marked grey).

Discussion

Here we provide a thorough comparison of overall survival prognostic performance of feature selection and

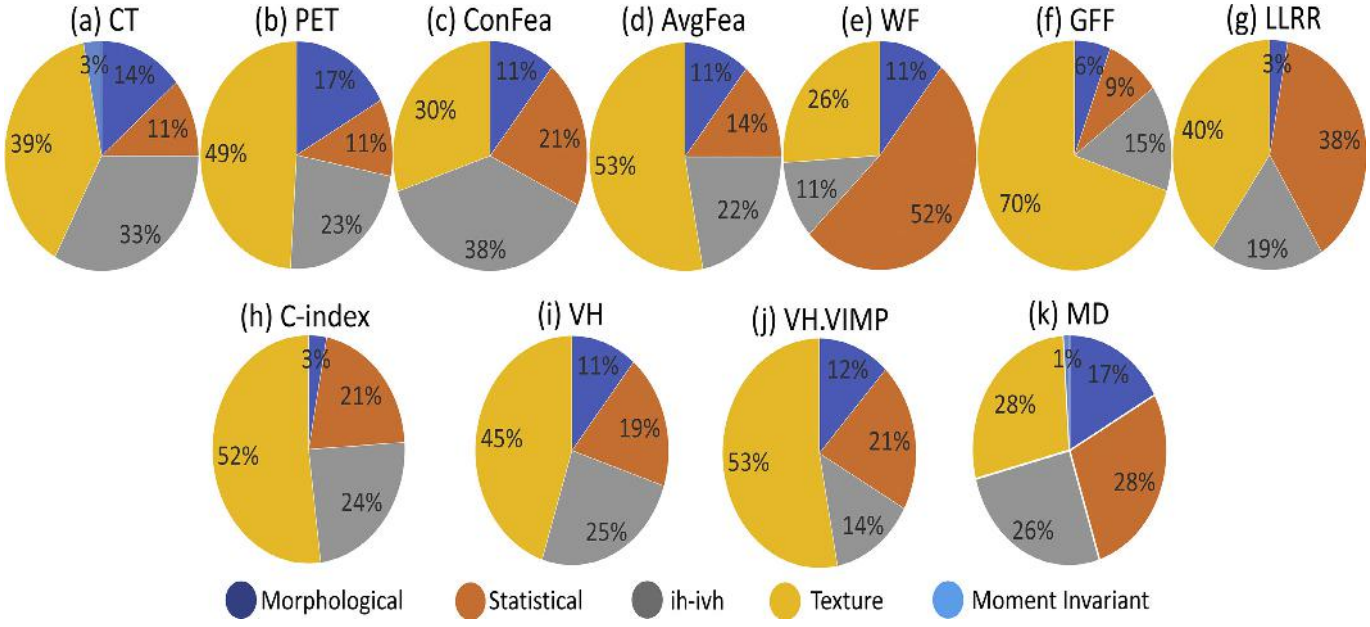

Fig 3. Pie chart of distribution of features selected from different radiomics modalities (a–g), by different feature selection methods (h–k) within feature families.

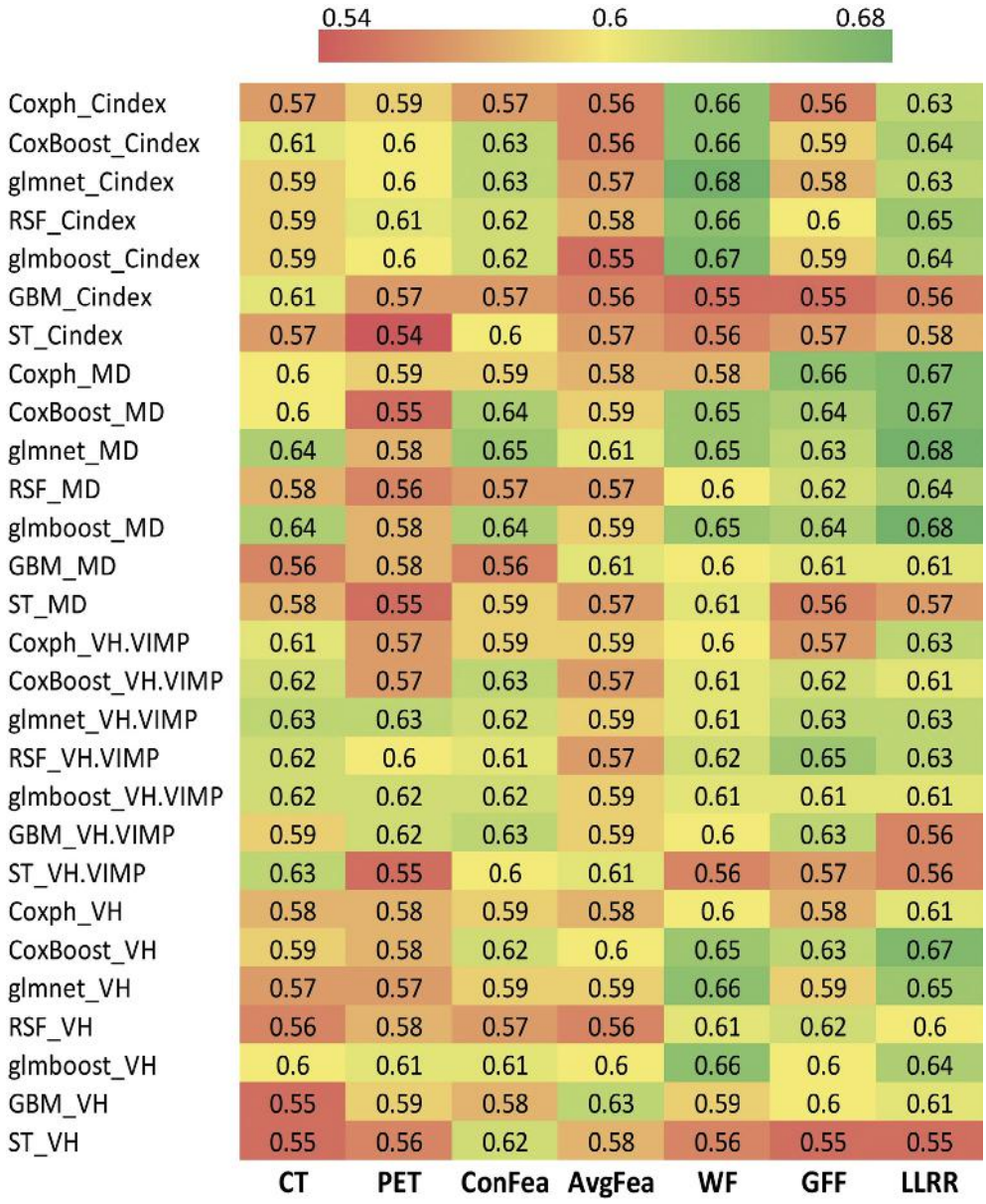

Fig 4. Performance of all radiomics modalities (columns) trained by different combinations of FS and ML methods (rows).

machine learning methods applied to multilevel fusion of multimodality PET/CT radiomics-based models developed for NSCLC patients. In this regard, a cross-combination of four feature selection and seven machine learning methods was applied to single-modality PET and CT models, feature-level fusion AvgFea and ConFea models, and image-level fusion wavelet fusion, GFF and LLRR models.

Overall, feature selection methods can be divided into four categories: filter-based, wrapper-based, embedded-based and hybrid methods (the latter is a combination of previous methods). Previous studies mostly used filter-based feature selection methods as they are computationally efficient. However, filter-based methods often evaluate features individually and ignore the effect of the selected feature subset on the performance of the algorithm. Conversely, wrapper and embedded methods utilise a machine learning algorithm to evaluate feature subsets. In this study, we proposed the C-index method, which is a hybrid feature selection method using Spearman's correlation, as a filter to remove redundant features and then using Cox regression as a measure of relevance. Moreover, three other methods based on the wrapper strategy were used. With respect to machine learning, seven publicly available methods with reported hyper-parameters were used to ensure an unbiased evaluation. All methods were able to handle continuous time-to-event data.

The performance of all models was reported separately (Figure 4) and the best combinations for each radiomics model were reported (Table 2). Minimum depth as the feature selection method and glmnet as the machine learning method were the most frequently present among the best combinations. Moreover, for each radiomics model, each

Table 2
Best combinations of feature selection and machine learning methods for each radiomics modality. C-indices and corresponding standard deviations are also provided. C-indices higher than 0.65 are marked in bold

Model	Feature selection	Machine learning	C-index	Standard deviation
CT	MD	glmnet	0.64	0.052
		glmboost	0.64	0.052
PET	VH.VIMP	glmnet	0.63	0.053
ConFea	MD	glmnet	0.65	0.051
AvgFea	VH	GBM	0.63	0.056
WF	C-index	glmnet	**0.68**	0.063
GFF	MD	Coxph	**0.66**	0.073
LLRR	MD	glmnet	**0.68**	0.068
		glmboost	**0.68**	0.074

AvgFea, averaging PET and CT features; ConFea, concatenating PET and CT features into a single feature set; Coxph, Cox proportional hazard; CT, computed tomography; GBM, generalised boosted regression model; GFF, guided filtering-based fusion; glmnet, Lasso and Elastic-Net regularised generalised linear model; glmboost, gradient boosting with component-wise linear model; LLRR, latent low-rank representation; MD, minimum depth; PET, positron emission tomography; VH, variable hunting; VH.VIMP, variable hunting with variable importance; WF, wavelet fusion.

strategy, and for all models together, the average performance of each feature selection and machine learning method was reported (Table 3). Although the results were case-specific, minimum depth and VH.VIMP as feature selection methods and CoxBoost and glmnet within the machine learning methods had overall higher average results.

We also carried out a variability analysis using the multifactor ANOVA test to assess the effect of changing radiomics modality, feature selection method, machine learning algorithm and their interactions on the performance of prognostic models. As Table 4 suggests, the sensitivity to feature selection and machine learning methods or their interactions, are case-specific, i.e. not only different radiomics modalities but also different strategies (single-modality, feature- and image-level fusion strategies) are sensitive to different factors. This highlights the importance of optimising the models, regarding the strategy that is being considered and, furthermore, the specific radiomics modality that is being used.

We also reported the proportion of total variance contributed by each factor and their interactions in different strategies. Surprisingly, except for the feature-level fusion strategy, the selection of machine learning method generated greater variance even compared with the selection of radiomics modality. For instance, in the single-modality strategy (Figure 5b), 18% of the variance was attributed to the machine learning method, whereas the radiomics modality only contributed to 4.2% of the variance. In other words, the identification of the proper machine learning method may have a greater effect on the outcome than selecting between CT and PET modalities. This difference was even more significant for image-fusion strategy (44.6% of the variance was due to the selection of the machine learning method, whereas fusion type and feature selection method had less than 10% contribution to the total variance).

Multiple comparisons were also carried out to identify the best feature selection/machine learning methods and their combinations, separately for different radiomics

Table 3
Average performance of models with a defined feature selection or machine learning method, separately for each radiomics modality, each strategy, and within all models together. Highest C-indices within each category are marked in bold

Method	CT	PET	Single modality	ConFea	AvgFea	Feature fusion	WF	GFF	LLRR	Image fusion	All
C-index	0.59	**0.59**	0.59	**0.61**	0.56	0.59	**0.63**	0.58	0.62	0.61	0.60
MD	0.6	0.57	0.59	**0.61**	**0.59**	**0.60**	0.62	**0.62**	**0.65**	**0.63**	**0.61**
VH.VIMP	**0.62**	**0.59**	**0.61**	**0.61**	**0.59**	**0.60**	0.6	0.61	0.6	0.60	0.60
VH	0.57	0.58	0.58	0.6	**0.59**	0.60	0.62	0.6	0.62	0.61	0.60
Coxph	0.59	0.58	0.59	0.59	0.58	0.59	0.61	0.59	0.64	0.61	0.60
CoxBoost	**0.61**	0.58	0.60	**0.63**	0.58	**0.61**	0.64	**0.62**	**0.65**	**0.64**	0.61
glmnet	**0.61**	**0.6**	**0.61**	0.62	0.59	**0.61**	**0.65**	0.61	**0.65**	**0.64**	**0.62**
RSF	0.59	0.59	0.59	0.59	0.57	0.58	0.62	**0.62**	0.63	0.62	0.60
glmboost	**0.61**	**0.6**	**0.61**	0.62	0.58	0.60	**0.65**	0.61	0.64	0.63	0.61
GBM	0.58	0.59	0.59	0.59	**0.6**	0.60	0.59	0.6	0.59	0.59	0.59
ST	0.58	0.55	0.57	0.6	0.58	0.59	0.57	0.56	0.57	0.57	0.57

AvgFea, averaging PET and CT features; ConFea, concatenating PET and CT features into a single feature set; CoxBoost, Cox model fitted by likelihood-based boosting; Coxph, Cox proportional hazard; CT, computed tomography; GBM, generalised boosted regression model; GFF, guided filtering-based fusion; Glmboost, gradient boosting with component-wise linear model; glmnet, Lasso and Elastic-Net regularised generalised linear model; LLRR, latent low-rank representation; MD, minimum depth; PET, positron emission tomography; RSF, random survival forest; ST, survival tree; VH, variable hunting; VH.VIMP, variable hunting with variable importance; WF, wavelet fusion.

Table 4
P-values related to the ANOVA test to show the significant effects of different factors on different radiomics modalities, different radiomics strategies and all models together. Non-statistically significant *P*-values are marked in bold

Method	CT	PET	Single modality	ConFea	AvgFea	Feature fusion	WF	GFF	LLRR	Image fusion	All
Radiomics	_	_	0.034	_	_	<0.001	_	_	_	0.001	<0.001
Feature selection	0.002	**0.097**	0.005	**0.485**	0.016	**0.096**	**0.155**	0.004	0.006	0.002	0.003
Machine learning	**0.099**	0.012	0.010	0.016	**0.319**	0.018	0.001	0.013	0.000	<0.001	<0.001
Radiomics + feature selection	_	_	0.040	_	_	**0.052**	_	_	_	0.001	<0.001
Radiomics + machine learning	_	_	**0.248**	_	_	0.020	_	_	_	**0.145**	<0.001
Machine learning + feature selection	_	_	**0.616**	_	_	**0.191**	_	_	_	**0.108**	0.001

AvgFea, averaging PET and CT features; ConFea, concatenating PET and CT features into a single feature set; CT, computed tomography; GFF, guided filtering-based fusion; LLRR, latent low-rank representation; PET, positron emission tomography; WF, wavelet fusion.

strategies, and for all radiomics modalities together (Figure 6). As Figure 6 suggests, not all the radiomics strategies used (single-modality, feature- and image-level fusion) had a specific feature selection or machine learning method that could significantly outperform other methods. For instance, considering feature selection methods in the image-level fusion strategy (Figure 6g), a special feature selection method (minimum depth) significantly outperformed all other methods, whereas in the feature-level fusion strategy (Figure 6d), the results from different feature selection and machine learning methods were comparable to each other (no method significantly outperformed other methods). In other strategies, some feature selection and/or machine learning methods significantly altered performance compared with other specific methods.

Our analysis focused on features selected from different radiomics modalities by different feature selection methods to investigate their distribution within feature families (Figure 3). Overall, the texture family had the greatest proportion and moment invariant and morphological families had the smallest share in feature sets. In the minimum depth feature selection method, which appeared in several best combinations (Table 2), and had the best average performance (Table 3), the features were approximately distributed equally within families (except moment invariant family). It conveys that ideal feature signatures may include equal shares of feature families to

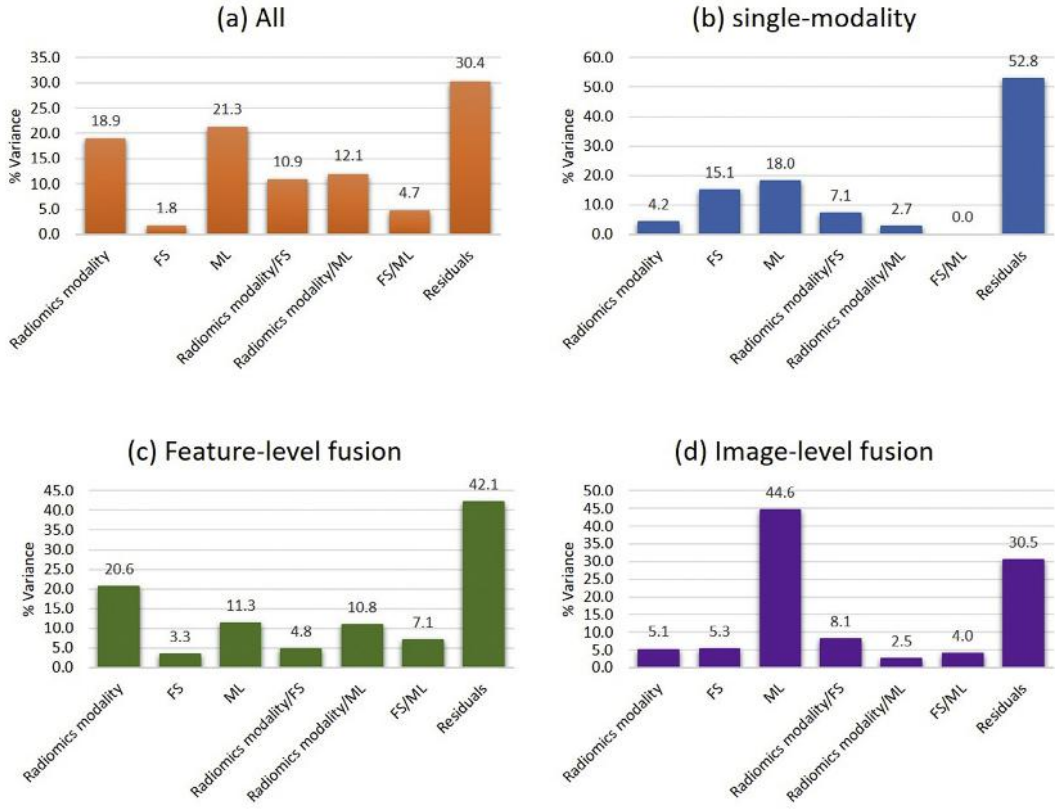

Fig 5. The proportion of variance explained by each factor and their interactions for (a) all models together, (b) single-modality strategy, (c) feature-level fusion strategy, and (d) image-level fusion strategy.

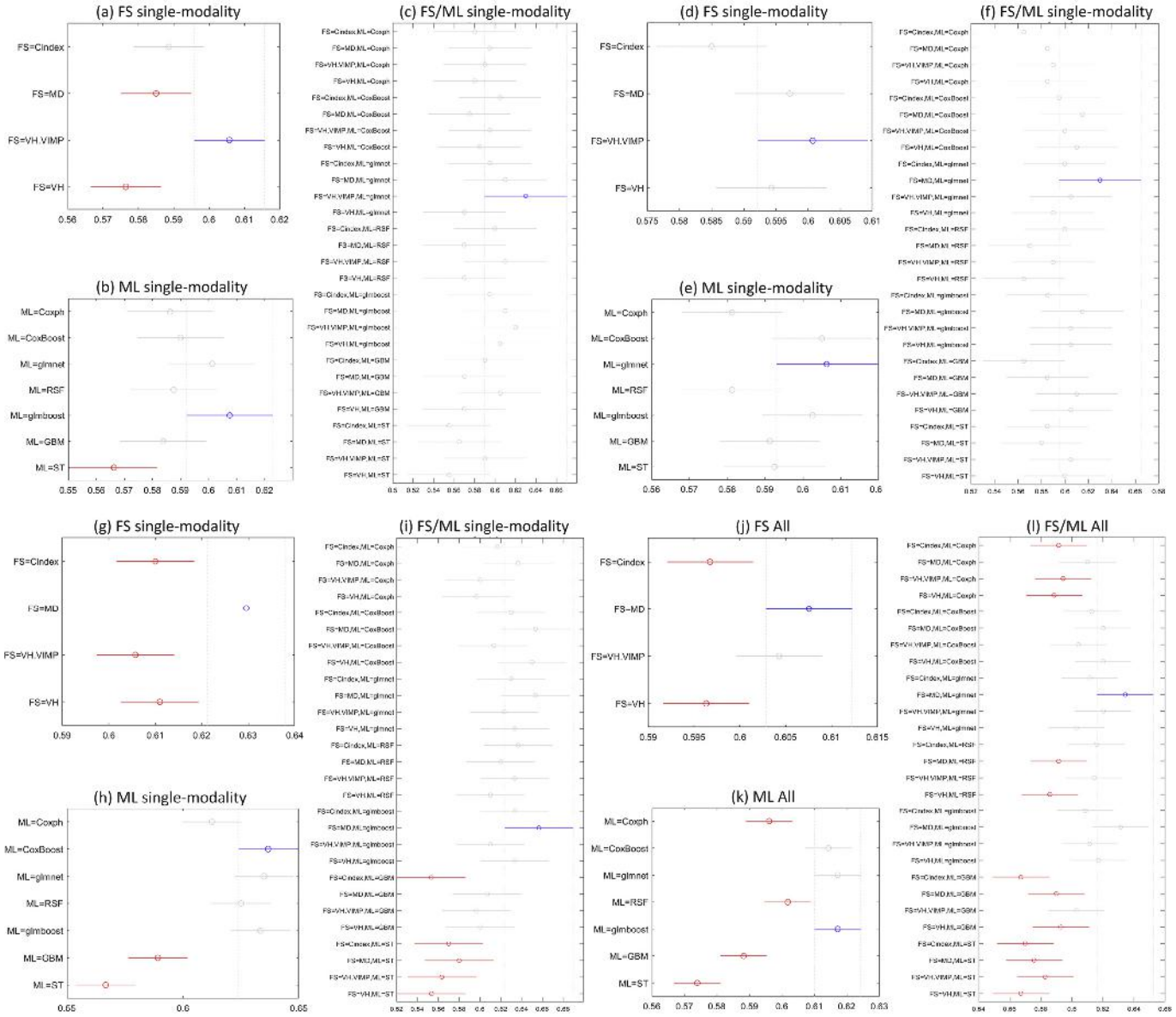

Fig 6. Multiple comparison test identifying which FS (a, d, g, j), ML (b, e, h, k), or combination of FS/ML methods (c, f, i, l) provides significantly different results in single-modality strategy (a, b, c), feature-level fusion strategy (d, e, f), image-level fusion strategy (g, h, i), and all (j, k, l) within models together. The best methods or combinations are marked in blue, methods that are significantly outperformed are marked in red, and methods with comparable results are marked in gray.

exploit all aspects of tumours, including shape, intensity distribution and texture.

In a recent study, Amini *et al.* [30] developed PET/CT multimodality radiomics models for NSCLC survival analysis. They integrated PET and CT data at different fusion levels and developed AvgFea and ConFea for features and wavelets for image-level fusion. For feature selection, they evaluated all possible combinations of top 10 features with the highest performance (identified by univariate Cox proportional hazard) to select the optimum signature and then used multivariate Cox proportional hazard to train their model. In the current study, we added two other image-level fusions to existing models and applied different feature selection and machine learning methods to optimise the performance of the models. Note that although we used the same dataset as in [42], we harmonised the feature sets using ComBat harmonisation to address the plausible batch effect owing to the dual-centric nature of the study. For CT, the performance was improved from a C-index of 0.631–0.64 using minimum depth and glmnet/glmboost (as feature selection and machine learning methods, respectively). Likewise, for PET, the performance increased from 0.591 to 0.63 using VH.VIMP and glmnet. For ConFea, our model using minimum depth and glmnet (C-index = 0.65) dominated their model (C-index = 0.616). For AvgFea, our model based on variable hunting and GBM slightly increased the performance (0.63 versus 0.629). Finally, for image-level fusion,

our models did not result in superior performance. Yet, we achieved comparable results to their wavelet fusion model (C-index = 0.68) with minimum depth and glmnet/glmboost applied to LLRR fusion or C-index and glmnet applied to wavelet fusion. In this work, we tested different feature selection and machine learning methods to improve the outcome. At the same time, we used two additional image-level fusion methods to investigate the effect of image-fusion methods on the performance of image-level fusion models. However, as shown in Figure 5d, compared with the selection of the fusion method, the selection of the machine learning algorithm introduced larger variance in the performance of image-level fusion models.

To the best of our knowledge, this is the first effort extending the comparison of different feature selection and machine learning methods to multilevel fusion of multimodality radiomics models towards NSCLC overall survival prognostication. In a study by Sun *et al.* [41], five feature selection and eight machine learning methods were applied to the radiomic features extracted from CT scans of NSCLC patients to investigate their prognostic performance. They achieved the highest performance using the gradient boosting linear model based on Cox's partial likelihood for the machine learning method and the concordance index as the feature selection method (C-index = 0.68). In comparison, our best model for CT used minimum depth as the feature selection method and glmnet or glmboost as the machine learning method (C-index = 0.64). Moreover, they only considered a single-modality CT model, whereas our study also shed light on multilevel fusion models. Another study by Parmar *et al.* [38] examined 14 feature selection and 12 machine learning methods to assess their performance and stability for predicting the overall survival of NSCLC patients. Our results (reported with C-indices, which preserved the continuous nature of the time-to-event analysis) cannot be compared with their results (reported with area under the curve). In their study, a combination of the Wilcoxon test-based feature selection method and the random forest classification method reached the highest performance (area under the curve = 0.66) with the highest stability against data perturbation. In addition, their study was limited to single-modality CT and they transformed the overall survival (a continuous outcome) to a binary outcome, which might have biased the prediction accuracy.

This study can provide information to tailor a subset of promising feature selection and machine learning algorithms to single- and multimodality radiomics models for overall survival prognostication of NSCLC patients. Moreover, the results were extended to different levels of fusion of PET and CT data. The most prominent limitation of this study was the small sample size.

Conclusion

In this study, we proposed a comprehensive framework for finding the optimum time-to-event prognostic algorithms (feature selection and machine learning), suitable for single- and multimodality radiomics models. For PET/CT radiomics-guided survival prognostication of NSCLC patients, the optimal combination of feature selection and machine learning methods vary considerably depending on the imaging modality (PET, CT or dual-modality PET/CT fusion) they are applied to. However, on average, minimum depth as the feature selection method and glmnet as the machine learning method resulted in a better performance. The proposed framework enabled finding optimum time-to-event prognostic algorithms for application on different single- and multimodality PET/CT radiomics for survival prediction of NSCLC patients.

Conflicts of interest

H. Zaidi reports financial support from Swiss National Science Foundation. H. Zaidi reports a relationship with Swiss National Science Foundation that includes: funding grants.

Acknowledgements

This work was supported by the Swiss National Science Foundation under Grant SNRF 320030_176052.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clon.2021.11.014.

References

[1] Lambin P, Leijenaar RT, Deist TM, Peerlings J, De Jong EE, Van Timmeren J, *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017;14:749–762.
[2] Longo DL. Tumor heterogeneity and personalized medicine. *N Engl J Med* 2012;366:956–957.
[3] Sala E, Mema E, Himoto Y, Veeraraghavan H, Brenton J, Snyder A, *et al.* Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clin Radiol* 2017;72:3–10.
[4] Parmar C, Leijenaar RT, Grossmann P, Velazquez ER, Bussink J, Rietveld D, *et al.* Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Sci Rep* 2015;5:11044.
[5] Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, Miles KA. Non-small cell lung cancer: histopathologic correlates for texture parameters at CT. *Radiology* 2013;266:326–336.
[6] Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, *et al.* Exploratory study to identify radiomics classifiers for lung cancer histology. *Front Oncol* 2016;6:71.
[7] Nazari M, Shiri I, Hajianfar G, Oveisi N, Abdollahi H, Deevband MR, *et al.* Noninvasive Fuhrman grading of clear cell renal cell carcinoma using computed tomography radiomic features and machine learning. *Radiol Med* 2020;125:754–762. https://doi.org/10.1007/s11547-020-01169-z.
[8] Liang C, Huang Y, He L, Chen X, Ma Z, Dong D, *et al.* The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer. *Oncotarget* 2016;7:31401.

[9] Khodabakhshi Z, Amini M, Mostafaei S, Haddadi Avval A, Nazari M, Oveisi M, *et al*. Overall survival prediction in renal cell carcinoma patients using computed tomography radiomic and clinical information. *J Digit Imaging* 2021;34:1086–1098. https://doi.org/10.1007/s10278-021-00500-y.

[10] Pyka T, Bundschuh RA, Andratschke N, Mayer B, Specht HM, Papp L, *et al*. Textural features in pre-treatment [F18]-FDG-PET/CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic radiation therapy. *Radiat Oncol* 2015;10:100.

[11] Wang T, Deng J, She Y, Zhang L, Wang B, Ren Y, *et al*. Radiomics signature predicts the recurrence-free survival in stage I non-small cell lung cancer. *Ann Thorac Surg* 2020;109:1741–1749.

[12] Zhang H, Graham CM, Elci O, Griswold ME, Zhang X, Khan MA, *et al*. Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy. *Radiology* 2013;269:801–809.

[13] Bae S, Choi YS, Ahn SS, Chang JH, Kang S-G, Kim EH, *et al*. Radiomic MRI phenotyping of glioblastoma: improving survival prediction. *Radiology* 2018;289:797–806.

[14] Coroller TP, Grossmann P, Hou Y, Velazquez ER, Leijenaar RT, Hermann G, *et al*. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol* 2015;114:345–350.

[15] Alic L, Niessen WJ, Veenland JF. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review. *PloS One* 2014;9:e110300.

[16] Mostafaei S, Abdollahi H, Dehkordi SK, Shiri I, Razzaghdoust A, Moghaddam SHZ, *et al*. CT imaging markers to improve radiation toxicity prediction in prostate cancer radiotherapy by stacking regression algorithm. *Radiol Med* 2020;125:87–97.

[17] Gevaert O, Mitchell LA, Achrol AS, Xu J, Echegaray S, Steinberg GK, *et al*. Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. *Radiology* 2014;273:168–174.

[18] Shiri I, Maleki H, Hajianfar G, Abdollahi H, Ashrafinia S, Hatt M, *et al*. Next-generation radiogenomics sequencing for prediction of EGFR and KRAS mutation status in NSCLC patients using multimodal imaging and machine learning algorithms. *Mol Imaging Biol* 2020;22:1132–1148. https://doi.org/10.1007/s11307-020-01487-8.

[19] Hajianfar G, Shiri I, Maleki H, Oveisi N, Haghparast A, Abdollahi H, *et al*. Noninvasive O6 methylguanine-DNA methyltransferase status prediction in glioblastoma multiforme cancer using magnetic resonance imaging radiomics features: univariate and multivariate radiogenomics analysis. *World Neurosurg* 2019;132:e140–e161.

[20] O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res* 2015;21:249–257.

[21] Zaidi H, Karakatsanis N. Towards enhanced PET quantification in clinical oncology. *Br J Radiol* 2018;91:20170508.

[22] Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, *et al*. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014;5:1–9.

[23] Jiang C, Kong Z, Liu S, Feng S, Zhang Y, Zhu R, *et al*. Fusion radiomics features from conventional MRI predict MGMT promoter methylation status in lower grade gliomas. *Eur J Radiol* 2019;121:108714.

[24] Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, El Naqa I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. *Radiother Oncol* 2012;102:239–245.

[25] Chaddad A, Daniel P, Desrosiers C, Toews M, Abdulkarim B. Novel radiomic features based on joint intensity matrices for predicting glioblastoma patient survival time. *IEEE J Biomed Health Inform* 2018;23:795–804.

[26] Parekh VS, Jacobs MA. Multiparametric radiomics methods for breast cancer tissue characterization using radiological imaging. *Breast Cancer Res Treat* 2020;180:407–421. https://doi.org/10.1007/s10549-020-05533-5.

[27] Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol* 2015;60:5471.

[28] Zhou H, Jiang J, Lu J, Wang M, Zhang H, Zuo C, *et al*. Initiative, dual-model radiomic biomarkers predict development of mild cognitive impairment progression to Alzheimer's disease. *Front Neurosci* 2019;12:1045.

[29] Lv W, Ashrafinia S, Ma J, Lu L, Rahmim A. Multi-level multi-modality fusion radiomics: application to PET and CT imaging for prognostication of head and neck cancer. *IEEE J Biomed Health Inform* 2019;24:2268–2277.

[30] Amini M, Nazari M, Lord IS, Hajianfar G, Deevband MR, Abdollahi H, *et al*. Multi-level multi-modality (PET and CT) fusion radiomics: prognostic modeling for non-small cell lung carcinoma. *Phys Med Biol* 2012;66:205017.

[31] Rastegar S, Vaziri M, Qasempour Y, Akhash M, Abdalvand N, Shiri I, *et al*. Radiomics for classification of bone mineral loss: a machine learning study. *Diagn Interv Imaging* 2020;101:599–610.

[32] Zaidi H, El Naqa I. Quantitative molecular positron emission tomography imaging using advanced deep learning techniques. *Ann Rev Biomed Eng* 2021;23:249–276.

[33] Nasrabadi NM. Pattern recognition and machine learning. *J Electron Imaging* 2007;16:049901.

[34] Sun P, Wang D, Mok VC, Shi L. Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading. *IEEE Access* 2019;7:102010–102020.

[35] Hamerla G, Meyer H-J, Schob S, Ginat DT, Altman A, Lim T, *et al*. Comparison of machine learning classifiers for differentiation of grade 1 from higher gradings in meningioma: a multicenter radiomics study. *Magn Reson Imaging* 2019;63:244–249.

[36] Yin P, Mao N, Zhao C, Wu J, Sun C, Chen L, *et al*. Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. *Eur Radiol* 2019;29:1841–1847.

[37] Zhang B, He X, Ouyang F, Gu D, Dong Y, Zhang L, *et al*. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Lett* 2017;403:21–27.

[38] Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep* 2015;5:13087.

[39] Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front Oncol* 2015;5:272.

[40] Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zöphel K, *et al*. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Sci Rep* 2017;7:1–11.

[41] Sun W, Jiang M, Dang J, Chang P, Yin F. Effect of machine learning methods on predicting NSCLC overall survival time based on radiomics analysis. *Radiat Oncol* 2018;13:1–8.
[42] Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, *et al*. A radiogenomic dataset of non-small cell lung cancer. *Sci Data* 2018;5:1–9.
[43] Rosset A, Spadola L, Ratib O. OsiriX: an open-source software for navigating in multidimensional DICOM images. *J Digit Imaging* 2004;17:205–216.
[44] Li S, Kang X, Hu J. Image fusion with guided filtering. *IEEE Trans Image Process* 2013;22:2864–2875.
[45] Li H, Wu X-J. *Infrared and visible image fusion using latent low-rank representation*. arXiv preprint arXiv:1804.08992 2018.
[46] Ashrafinia S. *Quantitative nuclear medicine imaging using advanced image reconstruction and radiomics*. Johns Hopkins University; 2019.
[47] Zwanenburg A, Vallières M, Abdalah MA, Aerts HJ, Andrearczyk V, Apte A, *et al*. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 2020;295:328–338.
[48] McNitt-Gray M, Napel S, Jaggi A, Mattonen S, Hadjiiski L, Muzi M, *et al*. Standardization in quantitative imaging: a multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets. *Tomography* 2020;6:118.
[49] Yip SS, Aerts HJ. Applications and limitations of radiomics. *Phys Med Biol* 2016;61:R150.
[50] Zwanenburg A. Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *Eur J Nucl Med Mol Imaging* 2019;46:2638–2655.
[51] Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. *Phys Med Biol* 2020;65:24TR02.
[52] Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random survival forests for high-dimensional data. *Stat Anal Data Min* 2011;4:115–132.
[53] Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *J Am Stat Assoc* 2010;105:205–217.
[54] Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat* 1982;10(4):1100–1120.
[55] Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* 2009;25:890–896.
[56] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw* 2011;39:1.
[57] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008;2:841–860.
[58] Hofner B, Mayr A, Robinzonov N. Schmid M Model-based boosting in R: a hands-on tutorial using the R package mboost. *Comput Stat* 2014;29:3–35.
[59] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997;55:119–139.
[60] Friedman. Stochastic gradient boosting. *Comput Stat Data Anal* 2002;38:367–378.
[61] Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and regression trees*. CRC Press; 1984.
[62] Winkler RL. *Statistics; probability, inference, and decision*. Houghton Mifflin Harcourt School; 1975.