Development and validation of survival prognostic models for head and neck cancer patients using machine learning and dosiomics and CT radiomics features

a multicentric study

Mansouri, Zahra; Salimi, Yazdan; Amini, Mehdi; Hajianfar, Ghasem; Oveisi, Mehrdad; Shiri, Isaac; Zaidi, Habib

Published in:
Radiation Oncology

DOI:
10.1186/s13014-024-02409-6

Publication date:
2024

Document version:
Final published version

Document license:
CC BY

Citation for pulished version (APA):
Mansouri, Z., Salimi, Y., Amini, M., Hajianfar, G., Oveisi, M., Shiri, I., & Zaidi, H. (2024). Development and validation of survival prognostic models for head and neck cancer patients using machine learning and dosiomics and CT radiomics features: a multicentric study. *Radiation Oncology*, *19*, Article 12. https://doi.org/10.1186/s13014-024-02409-6

Mansouri *et al. Radiation Oncology* (2024) 19:12
https://doi.org/10.1186/s13014-024-02409-6

Radiation Oncology

RESEARCH Open Access

Development and validation of survival prognostic models for head and neck cancer patients using machine learning and dosiomics and CT radiomics features: a multicentric study

Zahra Mansouri[1†], Yazdan Salimi[1†], Mehdi Amini[1], Ghasem Hajianfar[1], Mehrdad Oveisi[2], Isaac Shiri[1] and Habib Zaidi[1,3,4,5*]

Abstract

Background This study aimed to investigate the value of clinical, radiomic features extracted from gross tumor volumes (GTVs) delineated on CT images, dose distributions (Dosiomics), and fusion of CT and dose distributions to predict outcomes in head and neck cancer (HNC) patients.

Methods A cohort of 240 HNC patients from five different centers was obtained from The Cancer Imaging Archive. Seven strategies, including four non-fusion (Clinical, CT, Dose, DualCT-Dose), and three fusion algorithms (latent low-rank representation referred (LLRR),Wavelet, weighted least square (WLS)) were applied. The fusion algorithms were used to fuse the pre-treatment CT images and 3-dimensional dose maps. Overall, 215 radiomics and Dosiomics features were extracted from the GTVs, alongside with seven clinical features incorporated. Five feature selection (FS) methods in combination with six machine learning (ML) models were implemented. The performance of the models was quantified using the concordance index (CI) in one-center-leave-out 5-fold cross-validation for overall survival (OS) prediction considering the time-to-event.

Results The mean CI and Kaplan-Meier curves were used for further comparisons. The CoxBoost ML model using the Minimal Depth (MD) FS method and the glmnet model using the Variable hunting (VH) FS method showed the best performance with CI = 0.73 ± 0.15 for features extracted from LLRR fused images. In addition, both glmnet-Cindex and Coxph-Cindex classifiers achieved a CI of 0.72 ± 0.14 by employing the dose images (+ incorporated clinical features) only.

[†]†Zahra Mansouri and Yazdan Salimi contributed equally to this manuscript.

*Correspondence:
Habib Zaidi
habib.zaidi@hcuge.ch

Full list of author information is available at the end of the article

Conclusion Our results demonstrated that clinical features, Dosiomics and fusion of dose and CT images by specific ML-FS models could predict the overall survival of HNC patients with acceptable accuracy. Besides, the performance of ML methods among the three different strategies was almost comparable.

Keywords Head and neck cancer, Machine learning, Survival analysis, Radiomics, Dosiomics

Introduction

Head and neck cancers (HNCs) account for around 5% of all malignancies, with 931,931 new cases and 467,125 (almost half of the incidences) deaths worldwide reported in 2020 [1]. The standard recommended treatment for HNC patients includes surgery and radiation therapy (RT) as adjuvant or concurrent with chemotherapy [2–4]. These patients' leading causes of treatment failure and death are locoregional recurrences and distant metastasis [5, 6], which can affect overall survival. Although some prognostic factors, such as tumor location, age, stage, and human papillomavirus (HPV) status, are beneficial for prognostication, these patients still present with very poor prognoses [7, 8]. Even patients with similar prognostic factors may have different ultimate outcomes [9]. For those patients who undergo radiotherapy, dose delivery to the different parts of the tumor (necrotic and hypoxic) can be insufficient or non-uniform, which might influence tumor recurrence or residuals, leading to metastases and affecting patients' outcomes. 3D dose maps obtained from treatment planning, contain information about the uniformity or inhomogeneity of dose distribution which can be predictive. Traditionally, this information is summarized into dose-volume histograms (DVHs), which proved to have limited predictive value [10]. As such, developing a reliable prognostic analysis and outcome prediction algorithm based on information from dose distributions in more effective way is an essential step in assisting personalized decision-making and treatment strategies.

The use of radiomics analysis as a noninvasive, fast, and cost-efficient approach to extract various image-based quantitative features has proven to be valuable for patient prognosis and outcome prediction modeling [4, 11, 12]. Radiomics has played an essential role in characterizing the internal structures of tissues, e.g., intratumor inhomogeneities that are becoming more widely recognized as a related factor in HNC prognosis [13–15]. Several studies have shown that using multi-modality fusion-based radiomic features from different medical imaging modalities, such as CT, MRI, and PET can significantly improve the predictive power of radiomics for other cancer types [16–19].

Implementing the radiomics concept on 3D dose distributions (called Dosiomics [20]) provided an opportunity to use the valuable predictive information hidden in the 3D dose distributions more effectively than DVHs.

While most studies investigated normal tissue complication prediction ability of dosiomics [20–29], few of them used radiomics and dosiomics for prognosis or outcome prediction [8, 30]. For instance, Lee et al. [30] used the Radiomic and Dosiomic features to predict weight loss in lung cancer patients after RT. They demonstrated that this analysis could improve the power of predicting weight loss as a prognostic factor and developing personalized treatment planning. Wu et al. [8] established a prediction model using radiomic and dosiomic features for locoregional recurrence in HNC patients who had received intensity-modulated radiation therapy (IMRT) and revealed that dosiomics improves the prognostic results.

However, to the best of our knowledge, previous studies barely integrated the dose distribution with one of the imaging modalities to predict prognosis or treatment outcome for HNCs using fusion-based features. In a recent study by Cai et al. [31], they trained a model for overall survival prediction and used different fusions of CT and dose distributions, reporting that fusion models outperformed single-modality models.

This study aimed to investigate the value of radiomic and dosiomic features extracted from GTVs on CT images as the primary modality used in RT treatment planning and dose distributions, in addition to image fusion of CT and absorbed doses for the prediction of survival in HNC patients who received IMRT. Moreover, utilizing multiple combinations of machine learning algorithms and feature selection methods, we explored the optimal combination suitable for our purposes. We considered the prediction of the overall survival of patients after treatment as the endpoint.

Materials and methods

The overall workflow of the current study is shown in Fig. 1.

Study population

A total of 240 patients with HNC obtained from the "Head-Neck-PET-CT" [32–34] and "HNSCC [35, 36]" databases archived in The Cancer Imaging Archive (TCIA) open access repository [32]. The "Head-Neck-PET-CT" database included data from four different centers, i.e., CHUM, CHUS, HGJ, and HMR, with 298 patients, whereas "HNSCC" included 627 patients. After excluding patients with incomplete data in terms of pre-treatment CT images, radiotherapy planning dose and

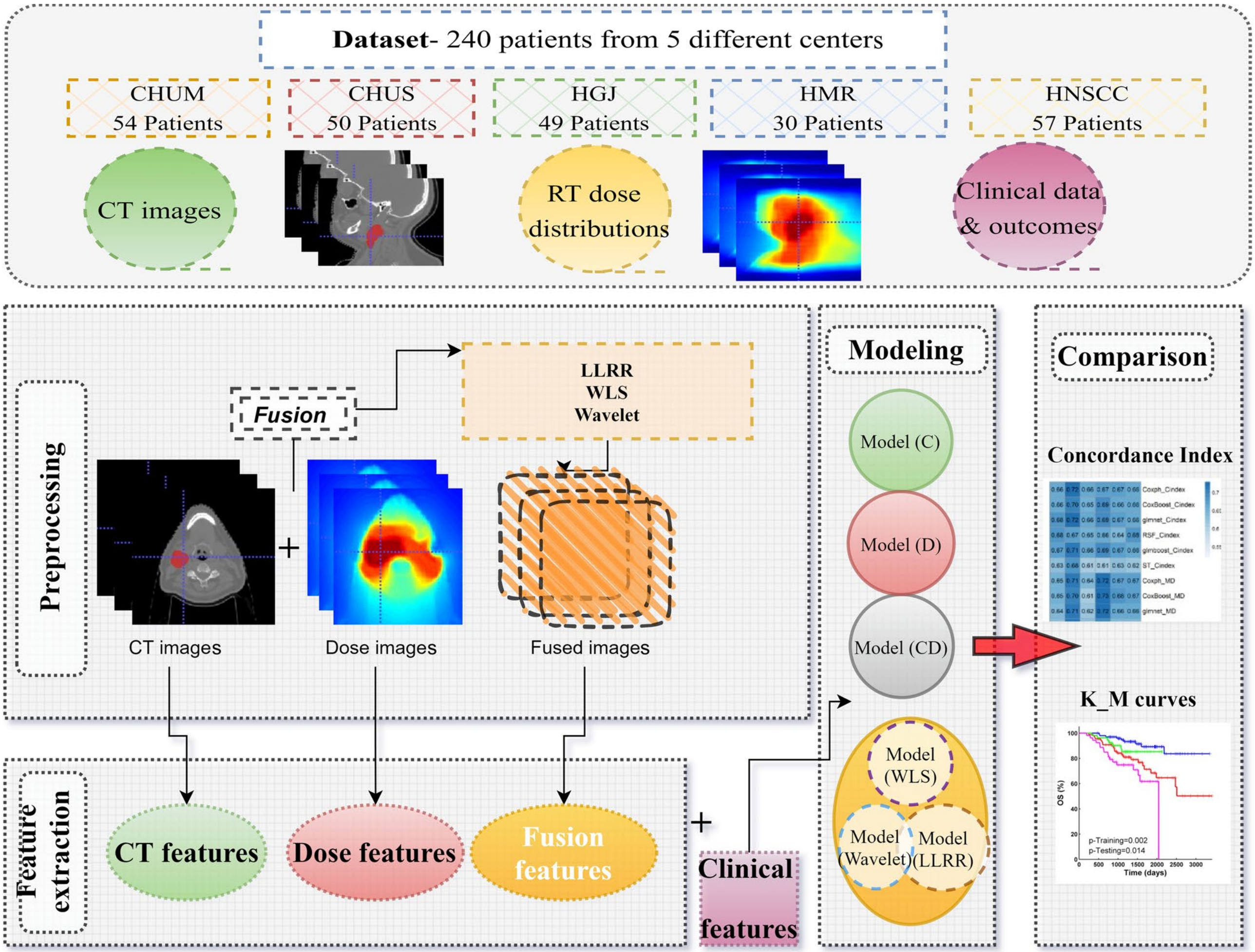


Fig. 1 The flowchart adopted in this study protocol (radiomics, dosiomics, and three different fusion algorithms) combines dose and CT information to predict vital status and overall survival (time to event) for head and neck cancer patients

outcome data, especially vital status, only 183 patients (54 CHUM, 50 CHUS, 49 HGJ, and 30 HMR) along with the other 57 patients were collected from "Head-Neck-PET-CT" and "HNSCC," respectively, (in total 238 cases). The clinical characteristics of the analyzed patients are listed in Table 1. Overall survival of patients is defined as the time from diagnosis to the date of the last follow-up considered as the endpoint of this study. As evident in Table 1, the range of time from diagnosis to the last follow-up was 350–1806 days (average 1190 days), 245–2001 days (average 1189 days), 361–2119 days (average 1277 days), 194–2136 days (average 1195 days), 193–3542 days (average 2235 days) for CHUM, CHUS, HGJ, HMR and, HNSCC, respectively. For overall survival, our observation object was GTV which was contoured on CT images and stored in DICOM format retrieved from online datasets.

The metrics of age, sex, T-Stage, N-Stage, and TNM-group, primary tumor site, treatment type, and outcome were compared between the datasets with one way ANOVA test. P-values less than 0.05 were considered statistically significant.

Preprocessing

All preprocessing procedures were performed in MatLab IBM (The Math Works Inc, *MATLAB*. Version 2020b) software. The dose distributions were registered on the axial CT images according to the location tag stored in the DICOM header. Then the GTV area was extracted and utilized for the next steps.

Image fusion

To suppress any plausible bias in the results due to the selection of a specific image fusion model, three different publicly available algorithms were utilized to fuse CT images and dose maps. These included a technique based on 3D discrete wavelet transform, referred to as wavelet fusion (WF), one using visual saliency map (VSM) and weighted least square optimization, referred to as WLS,

Table 1 Characteristics of patients included in this study protocols

Characteristics	CHUM	CHUS	HGJ	HMR	HNSCC	p-value
Total patients	54	50	49	30	57	-
Sex (M/F No.)	38/16	32/18	40/9	24/6	48/9	0.089
(%)	70/30	64/36	82/18	80/20	84/16	
Average Age (year)	62.1 ± 8.5	63.2 ± 11.6	62.2 ± 9.9	67.9 ± 9.9	56.7 ± 8.7	< 0.001
No. of patients	34	33	30	24	20	
>=60 years	20	17	19	6	37	
< 60						
T stage (No.)	8	4	9	1	11	0.046
T1	22	20	13	14	19	
T2	15	17	20	5	16	
T3	5	1	5	6	4	
T4	0	6	0	1	0	
T4a	0	2	0	2	0	
T4b	4	0	2	1	0	
Tx						
N stage	3	18	10	5	10	< 0.001
N0	7	4	10	4	9	
N1	38	25	0	0	2	
N2	0	0	6	0	2	
N2a	0	0	15	10	24	
N2b	0	0	7	9	8	
N2c	6	3	1	1	2	
N3	0	0	0	0	0	
N3a	0	0	0	1	0	
N3b						
TNM stage	0	2	1	0	2	< 0.001
I	1	7	2	2	2	
II	52	9	19	5	13	
III	0	0	0	0	0	
IV	0	0	0	0	0	
V	0	0	0	0	0	
VI	0	27	25	17	38	
IVA	0	5	1	5	2	
IVB	0	0	1	1	0	
IIB	1	0	0	0	0	
IV1A						
Primary site	4	0	2	0	2	< 0.001
Unknown	47	36	34	14	47	
Oropharynx	2	3	5	4	2	
Nasopharynx	0	10	6	7	4	
Larynx	1	1	2	5	2	
Hypopharynx						
Therapy	3	15	3	5	9	< 0.001
Radiation only	51	35	46	25	26	
CHRT	0	0	0	0	5	
Surgery + RT	0	0	0	0	17	
Surgery + CHRT						
RT modality	IMRT/TOMO	IMRT	VMAT	IMRT	IMRT	
Total prescribed dose (Gy, (median ± SD)	70 ± 1.6	69 ± 21	71 ± 2	68 ± 3.1	70 ± 2.3	
Outcome	50/4	41/9	44/5	16/14	40/17	< 0.001
VS(alive/dead)						
Time (days)*	1190	1189	1277	1195	2235	
Average	350	245	361	194	193	
Min	1806	2001	2119	2136	3542	
Max						

CHRT: Chemoradiation, VS: Vital Status, *diagnosis to last follow-up and finally, a fusion method based on latent low-rank representation referred as LLRR.

For the wavelet fusion [19], volumes (3D CT and dose maps) were first decomposed up to one level utilizing the wavelet basis function *symlet8* as a 3D discrete wavelet transform. Following the decomposition of volumes, which led to eight wavelet coefficient sub-bands for each volume, corresponding sub-bands were averaged to obtain a single set of fused sub-bands. Finally, fused wavelet coefficients underwent inverse 3D discrete wavelet transform to reconstruct the fused images.

For WLS fusion [37], first, a unique multi-scale decomposition (MSD) technique, including two filters, namely, a Gaussian filter and a rolling guidance filter (RGF), were applied to input images to decompose them into base and detail layers. With this specific MSD, information of the specific scales is maintained, and the voids near the edges reduce. For the fusion of the based layers, an enhanced VSM-based technique is used that suppresses the residual low-frequency information in based layers leading to better contrast and improved general visual appearance of the fused images. Detailed layers are merged by the state-of-the-art WLS optimization method, which captures more details and less noise. In the final step, fused, based and detailed layers are integrated to achieve the fused scan. The default parameters used by [37] were adopted in our study.

For LLRR fusion [38], input scans were first fed into latent low-rank representation to decompose into two parts: the low-rank part, i.e., global structure, and the saliency part, i.e., detailed local structures. A weighted average strategy was used to fuse the corresponding low-rank parts to capture more edge information. Saliency parts were simply summed. The final step included the integration of the fused low-rank and saliency parts. In this fusion also, we used the default parameters presented by [38]. All image processing and image fusions were performed in Matlab®.

Feature extraction

All images were interpolated to an isotropic voxel spacing of 1×1×1 mm^3 prior to feature extraction, first to standardize the voxel size over images from different scanners/centers and second to preserve the rotational invariance characteristic of the texture features. In addition, the intensity levels inside ROIs were discretized to a 64-level grayscale to make the feature calculation tractable. A feature extraction package based on MATLAB®, known as the Standardized Environment for Radiomics Analysis (SERA)[1] [39], was used for feature calculation. SERA agrees with guidelines from Image Biomarker Standardization Initiative (IBSI) [40]. This package was previously evaluated in multi-center standardization studies for improved feature reproducibility and robustness [40, 41]. Overall, 215 features per modality were extracted, i.e. 215 for CT, 215 for dose, and 215 for each fusion method. The feature set included 29 shape (namely morphological), 50 first-order (namely statistical, histogram and intensity histogram) and 136 three-dimensional texture features calculated using GLCM, GLRLM, GLSZM, GLDZM, NGTDM, and NGLDM matrices. Besides, 7 clinical features (age, sex, primary tumor site, T staging, N staging, TNM staging, treatment modalities) were also included alongside the other features to construct the prediction models. Noteworthy, one of the implemented strategies (Dual-CT-Dose) involved concatenating the radiomic (n=215) and dosiomic (n=215) features. This concatenated group, along with the inclusion of 7 clinical features, yielded a total of 437 features. Subsequently, feature selection was performed as described below. The details of the extracted features can be found in Supplementary Table 1.

Feature selection

We utilized five distinct feature selection (FS) algorithms, namely C-Index, Minimal Depth (MD) [42, 43], Variable hunting (VH) [42, 43], Variable Importance (VH. VIMP) [42, 43], and Mutual Information (MI) [44] to identify appropriate features. The Concordance index for each feature was calculated using the C-Index FS method—a hybrid approach employing a filter and a wrapper based on univariate Cox proportional hazard regression. This calculation was performed after eliminating features in pairs with a Spearman's rank correlation coefficient (rho) less than 0.9. Features with rho greater than 0.9 were retained for further analysis and subjected to a univariate Cox proportional hazard model. The top ten features demonstrating optimal performance (highest mean C-index) were selected through 100 repetitions using bootstrap resampling.

In MD, a method based on random survival forest, the features were sorted by depth, and those closer to the root node, indicating higher predictive power, were chosen. The top 10 features with minimal depth were selected. Notably, the number of features equal to 10 was an arbitrary choice based on the most predictive features (depending on the feature selection approach).

For other FS methods, such as VH and VH.VIMP, both model-based FS techniques utilizing the Random Survival Forest (RSF) model, the data were randomly divided into train and test sets. RSF was applied to the train set, and random features were selected based on the minimal depth threshold. The initial model was constructed using these selected features, with continiously adding the features until the importance of the joint variable is stabilized. This process was iterated 50 times, and the features with the highest frequency of occurrence were selected. It is noteworthy that

[1] https://github.com/ashrafinia/SERA.

the process for VH and VH.VIMP is identical, except for VH.VIMP, where variable importance is utilized for feature ordering, whereas VH relies on the minimal depth threshold—a method slower than VH.VIMP [42, 43, 45]. MI represents a completely parallelized implementation designed for computing the Mutual Information Matrix. The calculation of MI involved a linear approximation based on Pearson's or Spearman's correlation between two columns. To assess the correlation between survival data, Somers' Dxy index was employed [44].

In supplementary Table 2, we provide detailed names of the features selected throughout the five-fold cross-validation.

Time-to-event survival models machine learning and hyperparameter optimization

In this study, we assessed the performance of six machine learning algorithms that can manage the continuous time-to-event survival data. The models are listed below:

1. Cox model fitted the by likelihood-based boosting (CoxBoost) [46];
2. Random survival forest (RSF) [47];
3. Cox proportional hazard (Cox PH) [48];
4. Gradient boosting with a component-wise linear model (glmboost) [49];
5. Lasso and Elastic-Net regularized generalized linear model (glmnet) [50];
6. Survival tree (ST) [51].

Feature selection, all ML model training, model evaluation, and hyperparameters tuning were implemented in the MLR package in R programming language, version 3.6.2. The hyperparameters were tuned for all ML methods (except Coxph) using grid search. The details of hyperparameter settings and the R packages used in this study are mentioned in supplementary Table 3. The hyperparameter setting was guided by the C-index as a performance metric, calculated by 3-fold cross-validation in the training dataset. Figure 2 describes the feature selection, model training, hyperparameters tuning, and model evaluation.

Model evaluation

This study used a one-leave-center-out strategy for model building and testing with a hold-out external validation set to build a generalizable model across the variability of centers, scanners, acquisition, reconstruction and treatment parameters. The feature normalization function was transformed from the training to the test set, meaning that the same normalization was implemented for features in both train and test dataset. After selecting features and optimized hyperparameters for ML models, ML models were tested on the hold-out external test dataset utilizing bootstrapping resampling with 1000 repetitions. This step was repeated five times, where in each time, one of the five centers selected as the test set and the remaining four centers as the training set. The results of the one-leave-center-out scheme were reported for all centers on average and for each center separately.

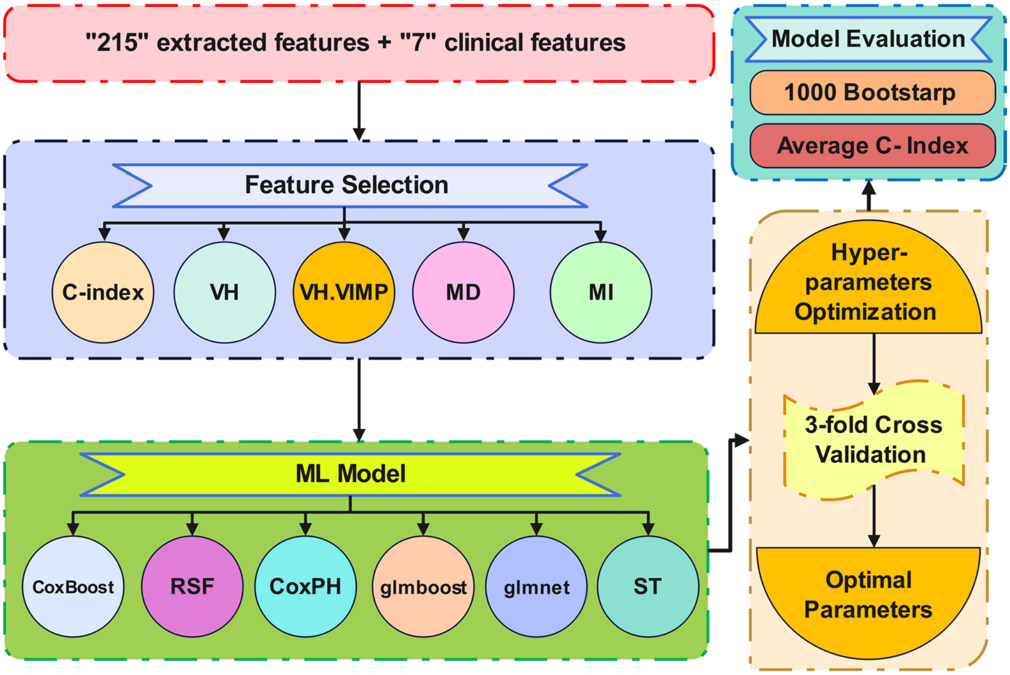
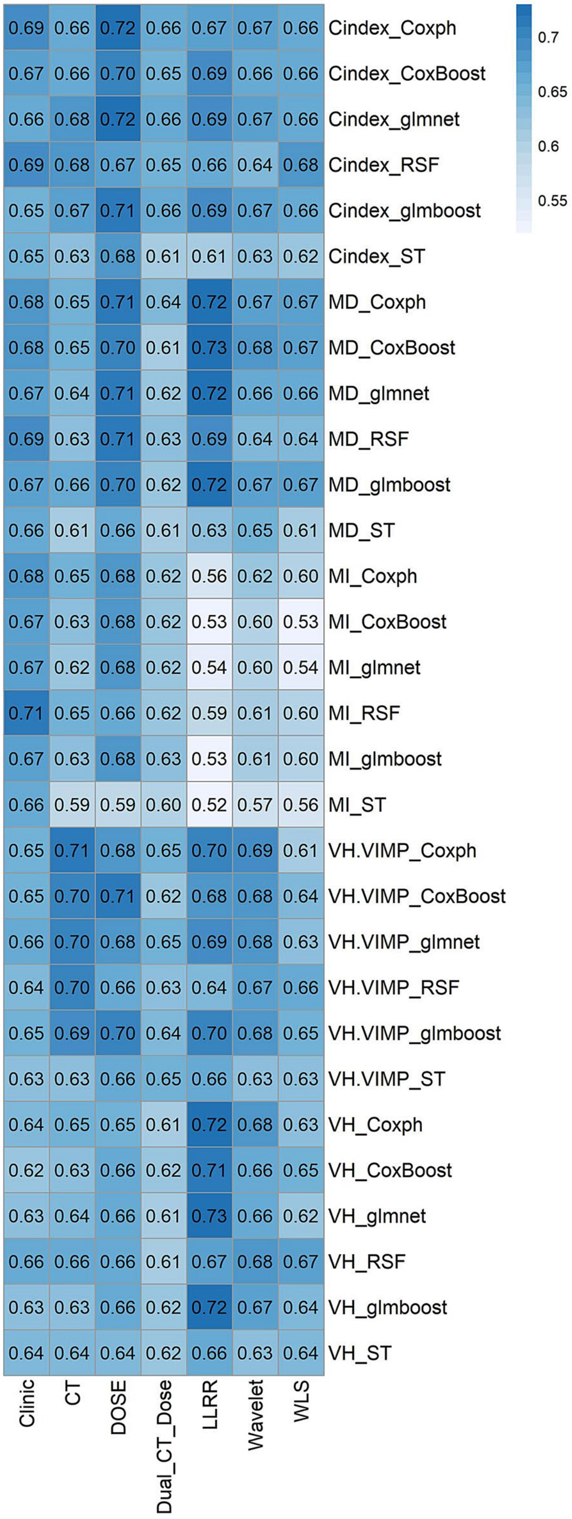


Fig. 2 Flowchart describing feature selection, model training, model tuning, and model evaluation



	Clinic	CT	DOSE	Dual_CT_Dose	LLRR	Wavelet	WLS
Cindex_Coxph	0.69	0.66	0.72	0.66	0.67	0.67	0.66
Cindex_CoxBoost	0.67	0.66	0.70	0.65	0.69	0.66	0.66
Cindex_glmnet	0.66	0.68	0.72	0.66	0.69	0.67	0.66
Cindex_RSF	0.69	0.68	0.67	0.65	0.66	0.64	0.68
Cindex_glmboost	0.65	0.67	0.71	0.66	0.69	0.67	0.66
Cindex_ST	0.65	0.63	0.68	0.61	0.61	0.63	0.62
MD_Coxph	0.68	0.65	0.71	0.64	0.72	0.67	0.67
MD_CoxBoost	0.68	0.65	0.70	0.61	0.73	0.68	0.67
MD_glmnet	0.67	0.64	0.71	0.62	0.72	0.66	0.66
MD_RSF	0.69	0.63	0.71	0.63	0.69	0.64	0.64
MD_glmboost	0.67	0.66	0.70	0.62	0.72	0.67	0.67
MD_ST	0.66	0.61	0.66	0.61	0.63	0.65	0.61
MI_Coxph	0.68	0.65	0.68	0.62	0.56	0.62	0.60
MI_CoxBoost	0.67	0.63	0.68	0.62	0.53	0.60	0.53
MI_glmnet	0.67	0.62	0.68	0.62	0.54	0.60	0.54
MI_RSF	0.71	0.65	0.66	0.62	0.59	0.61	0.60
MI_glmboost	0.67	0.63	0.68	0.63	0.53	0.61	0.60
MI_ST	0.66	0.59	0.59	0.60	0.52	0.57	0.56
VH.VIMP_Coxph	0.65	0.71	0.68	0.65	0.70	0.69	0.61
VH.VIMP_CoxBoost	0.65	0.70	0.71	0.62	0.68	0.68	0.64
VH.VIMP_glmnet	0.66	0.70	0.68	0.65	0.69	0.68	0.63
VH.VIMP_RSF	0.64	0.70	0.66	0.63	0.64	0.67	0.66
VH.VIMP_glmboost	0.65	0.69	0.70	0.64	0.70	0.68	0.65
VH.VIMP_ST	0.63	0.63	0.66	0.65	0.66	0.63	0.63
VH_Coxph	0.64	0.65	0.65	0.61	0.72	0.68	0.63
VH_CoxBoost	0.62	0.63	0.66	0.62	0.71	0.66	0.65
VH_glmnet	0.63	0.64	0.66	0.61	0.73	0.66	0.62
VH_RSF	0.66	0.66	0.66	0.61	0.67	0.68	0.67
VH_glmboost	0.63	0.63	0.66	0.62	0.72	0.67	0.64
VH_ST	0.64	0.64	0.64	0.62	0.66	0.63	0.64

Fig. 3 Heat map of C-index for each strategy and ML in combination with feature selection methods

C-indices were reported along with the standard deviations (±SD) among five scores from leave-one-center-out cross validation. With this methodology, we used all data sets as training and external hold-out test sets, which revealed model generalizability.

Kaplan-Meier curves were created for the best models, the cut-off criteria for risk stratification were the median of the risk scores calculated by the models, meaning the individuals witha risk score above or equal to and below the median were categorized into high-risk and low-risk groups, respectively. The log-rank test was used to calculate the p-values.

Results

Study population Anova test results

The ANOVA p-values were less than 0.05 when comparing all age, sex, T-Stage, N-Stage, and TNM-group, primary tumor site, treatment type, and outcome metrics among the five datasets, confirming that there is statistically significant difference amongst the population included in the cohort in terms of the considered metrics. The p-values are indicated in Table 1.

Overall comparison between the different strategies

Figure 3 depicts the heat map of the mean C-index for each strategy of radiomics, dosiomics, and fusion-based methods amongst all 30 combinations of feature selection and ML models. More information is detailed in supplementary Fig. 1, where the heatmaps are denominated with the name of the hold-out center (selected as test set).

Figure 4 shows the median value of the CIs among all models and strategies. Moreover, The mean and SD of C-indices are also listed in Supplementary Table 4. The violon plots of Fig. 5 depict the distribution of mean C-indices for each strategy separately.

To evaluate the significance of differences among the different strategies, we performed Friedman test followed by Nemenyi post-hoc test as the P-value for the Friedman test was <0.05. Supplementary Table 5 summarizes the results of Nemenyi post-hoc test, which showed significant difference among the different strategies.

Comparicon of different machine learning models

For survival prediction, two model and feature selection methods, i.e., CoxBoost-MD (C-index=0.73±0.15) and glmnet-VH (C-index=0.73±0.15) for LLRR, achieved the highest performance relative to other features. In a comparison of the top 10% C-indices among all non-fusion strategies (Clinical, CT, Dose, and Dual-CT-Dose), the dose strategy (dosiomics) showed the highest values of C-indices (0.7–0.72). The comparison among fusion-based strategies (LLRR, Wavelet, WLS) revealed the highest values for LLRR (0.7–0.73). The minimum CI

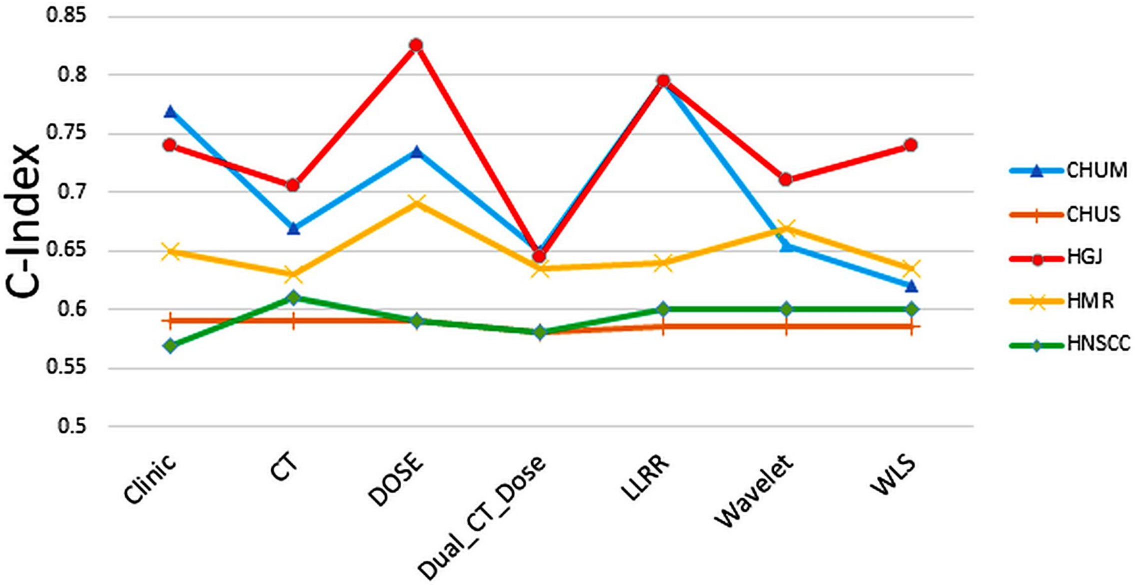


Fig. 4 The median value of C-indices among all strategies and all models for CHUM (blue), CHUS (orange), HGJ (red), HMR (yellow), and HNSCC (Green)

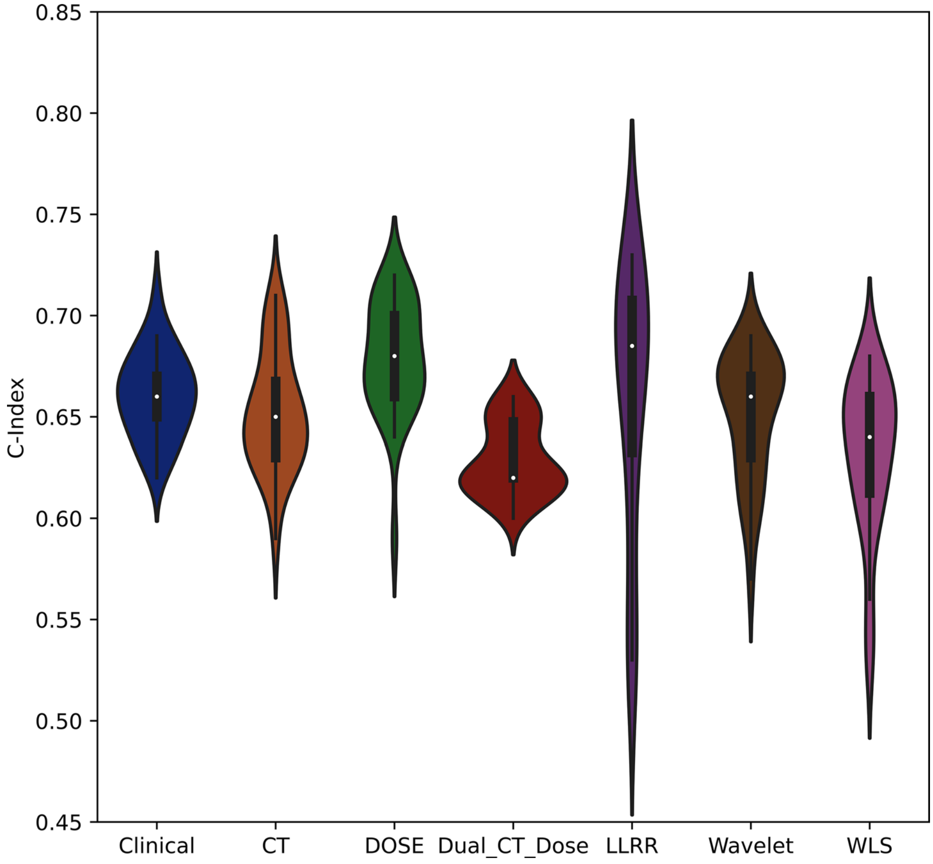


Fig. 5 The violon plots of mean C-index for radiomic and dosiomic and three different fusion-based strategies for survival prediction

values among all models and all strategies were from ST_MI (0.52–0.6), except for WLS, in which the minimum CI was from CoxBoost-MI (0.53) but still among the minimum values, CI of ST-MI was significantly low (0.56).

Table 2 shows the highest mean C-index for the best machine learning and FS combinations and their corresponding strategy. All models have been assessed in one leave center out cross-validation. The corresponding CI

Table 2 Best combinations of feature selection and machine learning methods (highest mean CIs) for each strategy. Standard deviations of CIs are also provided here and in the supplementary Table 3. Slash "/"separated SDs related to their corresponding feature selections

ML model	Feature selection method	Highest CI ± SD	Strategy
Coxph	CI/ MD/ VH	0.72 ± 0.14 / 0.15 / 0.15	DOSE/LLRR/LLRR
CoxBoost	MD	0.73 ± 0.15	LLRR
glmnet	VH	0.73 ± 0.15	LLRR
RSF	MD	0.71 ± 0.11	DOSE
glmboost	MD/ VH	0.72 ± 0.15 / 0.14	LLRR

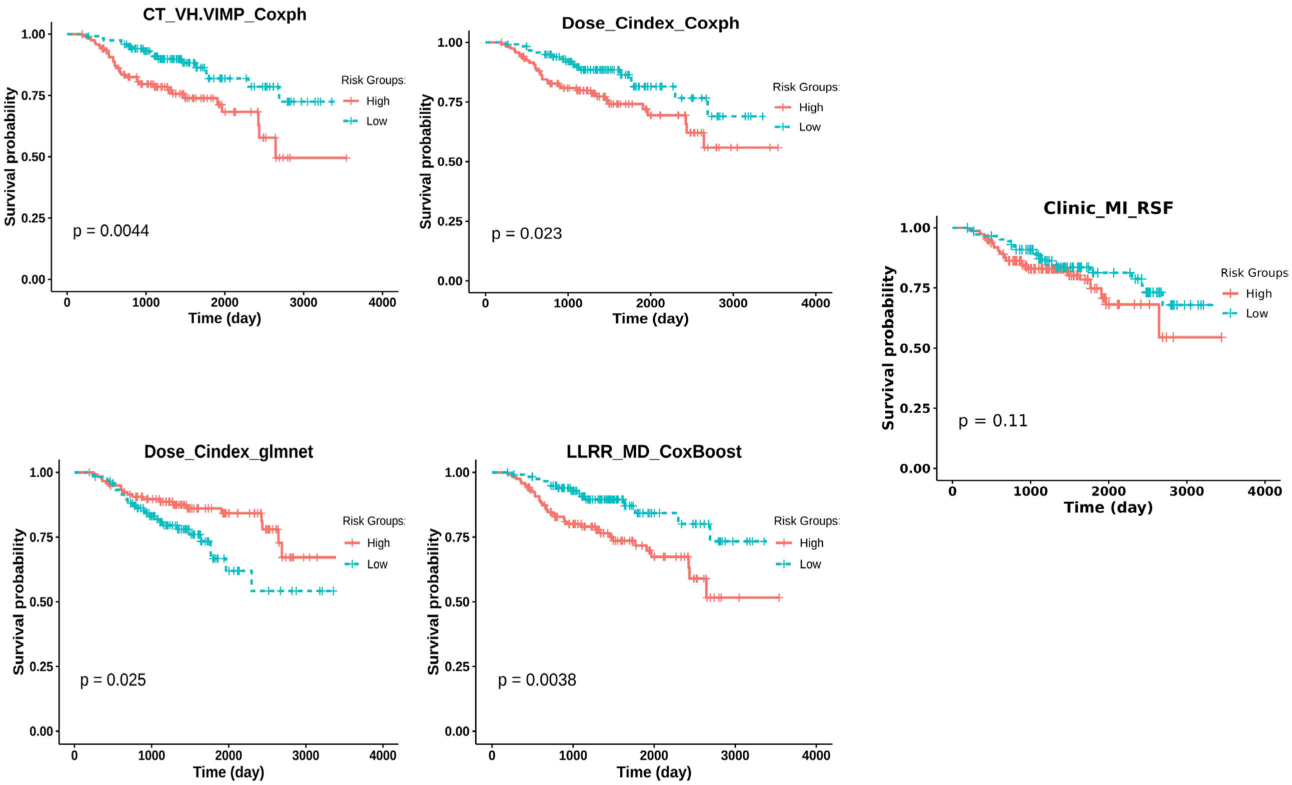


Fig. 6 Kaplan-Meier curves of the high- and low-risk groups for the five best models by clinical, Dose and LLRR fusion strategy. The cut-off criteria for risk stratification were median. A log-rank test was used to calculate the p-values

values for each center are reported in supplementary Fig. 2 in separate heat maps, and the top 10% C-indices for any strategy are reported in Supplementary Table 6. The models with the highest CI values were selected as more efficient models, and the K-M curves were applied to them for further comparison. Figure 6 illustrates the K-M curves for the best models, showing statistically significant power in dividing groups into low and high risks. The other K-M curves for the best models for all strategies are shown in Supplementary Fig. 3.

Comparison of different feature selection methods

A comparison between CI values in terms of evaluation of the feature selection methods performance showed that the higher values of CIs obtained from MD (0.73±0.15 and 0.71±0.11 for LLRR and Dose, respectively) and VH methods (0.73±0.15 for LLRR). VH.VIMP method for dosiomics only (Dose) and radiomics only (CT) showed similar performance (0.71±0.11). Among the fusion-based strategies, LLRR had the highest value (0.7±0.12). In the C-index method, the most potent CI among non-fusion strategies belonged to dose image selected features, i.e., dosiomics (0.72±0.14). In contrast, LLRR (0.69±0.13) had the highest C-index value amongst the fusion-based methods.

Discussion

In this study, we considered the image-based fusion of CT images and 3D dose distributions and integrate the concepts of radiomics and dosiomics by comprehensively comparing different machine learning and feature selection combinations to predict the overall survival of HNC patients after radiotherapy treatment. For this purpose, we used three different fusion algorithms to fuse the CT and dose maps. Then, corresponding extracted features were driven into the 30 different combinations of feature selections and machine learning algorithms to explore the value of dosiomics and fusion-based features. We also compared the results with single modality radiomics models. Our results demonstrated that the 3D dose distribution included valuable information highly correlated with overall survival prediction in HNC patients. Moreover, fusion (especially with the LLRR algorithm) of the dose distribution with CT images can improve some prediction models' performance and accuracy. However, the fusion approach provided a slightly more accurate performance than the dosiomics approach alone.

This study was conducted under the assumption that in image-level fusion, the incorporation of neighboring voxel values in the fused images creates a novel texture that offers increased prognostic value compared to the interpretation of voxel values in an individual image alone (e.g. HU in CT or Gy in dose map), which may lack a clear physical or biological meaning.

Five independent datasets were used in this study. We carried out a one-way ANOVA statistical test on our cohort. According to the p-values reported in Table 1, there is a significant heterogeneity in population characteristics, such as sex and TNM staging, indicating that our models presented a robust behavior against the heterogeneity of characteristics in the cohort, indicating the generizability of the models. Moreover, the effect of cycling in hold-out dataset on the performance of the models was investigated during the one-center-leave-out procedure. It is worth emphasizing that our models were consistently trained incorporating clinical features. Specifically, the features derived from CT, Dose, Dual-CT-Dose, and fusion-based strategies were not exclusively image- or dose-based. Instead, clinical features were consistently integrated. In addition, we conducted analyses with models utilizing only clinical features to establish a baseline assessment. The results indicated that clinical strategies achieved a performance comparable to Dose and fusion levels in terms of the C-index. However, as depicted in Fig. 4, the median values of Dose and LLRR in the folds exhibited superior performance compared to the clinical approach. Furthermore, the clinical strategy did not perform effectively in stratifying high and low-risk cases based on the survival curves illustrated in Fig. 6.

It should be noted that we compared our results with models established by Wu et al. [8]. However, they investigated local recurrence prognostic models, whereas our models predicted the overall survival (OS), and as such, the results are not directly comparable. Still, this was the most similar study involving Dosiomics results in HNC patients. By comparing our results to Wu et al. [8] models, our established models outperformed (CI=0.66 vs. 0.54) for CT with the Coxph-Cindex model and (CI=0.72 vs. 0.66) with Coxph-CIndex for Dose. However, we implemented much more comprehensive ML-FS (6×5) methods and models evaluated in the multi-center strategy. The VIMP model for CT has shown an even higher CI (0.71). We also achieved a CI of 0.72 for the Dose model with glmnet-Cindex. Our training method has benefited from the one-leave-center-out scheme. However, Wu et al. [8] used the data from two centers for training and the other two centers for testing their models.

Moreover, in this study, we fused CT images and Dose distributions. By combining the radiomics and dosiomics data into a single fused image, the performance of specific models could be improved (CI increasing to 0.73 in glmnet-VH and CoxBoost-MD for the LLRR fusion method). Compared with a fusion-based radiomics study by Lv et al. [18], our results have shown a higher prognostic performance for OS (CI=0.67 vs. 0.64) in the wavelet fusion strategy. They only implemented the Coxph-Cindex model with different fusion strategies. In contrast, our results for the same fusion strategy demonstrated that other ML-FS models have even better performance (the highest CI of 0.69 was achieved for the Coxph-VH. VIMP model in the wavelet fusion strategy). Besides, they fused PET and CT images of HNC patients. While in this study, we fused CT images and 3D dose distributions to combine their information in a single image. The average CI for the "CT only" strategy in this study with the Coxph-Cindex model was 0.66, whereas the CI of 0.71 was the highest prognostic performance using the Coxph-VH. VIMP model. However, the established model by Lv et al. [18] study achieved a CI of 0.59±0.5.

In comparison with Vallières et al. [33] results, who used a random forest model, our RFS models showed a lower CI for OS (0.7 for CT Via RSF-VH. VIMP and 0.71 for Dose Via RSF-MD vs. 0.75 for CT and 0.76 for PET). It should be noted that the larger number of extracted features in their study (1615 vs. 215+7 in our study) may have influenced their results. In a CT-based radiomics study by Sun et al. [52] in which the effect of ML methods on lung OS was investigated, the highest reported CI for some ML models were similar to some of ours, i.e., 0.674, 0.627, and 0.646 for CoxBoost-Cindex, RSF-PCC and Cox-MI, respectively. By including the dose information (dosiomics) and fusing the dose with CT images, the performance improved compared with the dose and CT-only approach (0.7 for Dose via CoxBoost-Cindex, 0.71, 0.69, for Dose, and LLRR, respectively, via RSF-MD, 0.68 via RSF Cindex and RSF-VH for WLS and Wavelet respectively and 0.68 via Coxph-MI for Dose strategy).

In a study by Lee et al. [30], the authors considered a radiomics and dosiomics strategy to predict the weight loss of lung cancer patients after radiotherapy. To summarize, a review of the dosiomics results demonstrated that our most robust dosiomics models (CI of 0.72, Coxph-Cindex, and glmnet-Cindex) are as predictive as theirs (AUC=0.71). It should be noted that AUC and CI are not directly comparable, and the subjects used in these studies are different.

In general, our models showed a good performance, particularly when using features derived from 3D dose distributions (Dosiomics). According to our findings, the combination of radiomics and Dosiomics into a single fused image (particularly employing the LLRR approach) yielded comparable results to those achieved with Dosiomics alone. However, Dosiomics remains unaffected by differences in CT characteristics. Moreover, the approach is both methodologically and conceptually simpler, requiring less data for collection.On the other hand, an investigation involving larger cohorts may reveal an increased discrimination between Dosiomics and fusion-based models as the hypothesis is that such models would derive greater benefits from a large-scale dataset.

The combination of LLRR fusion and specific feature selection and machine achieved the highest average C-Index. However, this doesn't mean that LLRR fusion always performs better with all feature selections/machines. We suggest that the model achieving the best performance be used. However, if there is a desire to test the model with limited inputs, such as clinical only, CT only or Dose only inputs, the best model according to supplementary Table 6 should be selected and used.

A radiotherapy 3D dose map contains by nature information correlated to outcome prediction. Although dose-volume histograms are the most common tools to display this information, this information is accumulated in existing DVHs, and DVHs are deficient in showing the spatial information [53–55]. Using the DVH-based metrics may even over/underestimate the prediction of therapeutic toxicity in head and neck cancer radiotherapy by up to 50% [56].

This study provided information to tailor a subset of feature selection and ML algorithms for overall survival prediction modeling in HNC patients. While this study involved five independent centers and multiple treatment modalities, a major limitation was the limited sample size and the usage of non-harmonized feature-sets. To mitigate the limited size, we utilized a one center leave-out strategy and averaged the results to enhance the reproducibility across different centers. However, further studies using larger databases and implementing more robust harmonization methods are still required. We showed that Dosiomics could more robustly explain the features within the 3D dose maps. Moreover, the results were extended to the fusion of treatment planning CT and radiotherapy dose distributions. Overall survival prediction based on radiomics and dosiomics and the fusion of these two images can be helpful in the decision-making process and personalized treatment.

Conclusion

We proposed a comprehensive framework for the development and validation of time-to-event overall survival models (cross-combination of feature selection and ML) for clinical as baseline, single (CT, Dose), multi-modality (CT-dose) and fusion models (Wavelet, LLRR, WLS). We also investigated the best combination of feature selection and machine learning model reported for each strategy. Our results demonstrated the potential of clinical, radiomics applied on CT, dosiomics derived from radiotherapy 3D dose distributions, and three different fusion strategies in overall survival prediction of head and neck cancer patients. Our results support the superiority of dosiomics in identifying prognoses associated with overall survival. The fusion-based models showed a comparable result to dosimics tending to improve the results of overall survival prediction probably by training models with large-scale datasets.

Abbreviations

CT	Computed Tomography
PET	positron emission tomography
GTV	gross tumor volumes
Dosiomics	dose distributions
HNC	Head and neck cancer
FS	feature selection
ML	machine learning
CI	concordance index
OS	overall survival
MD	Minimal Depth
VH	Variable hunting
LLRR	latent low-rank representation referred
RT	radiation therapy
DVH	dose volume histogram
IMRT	intensity-modulated radiation therapy
TCIA	The Cancer Imaging Archive
WF	wavelet fusion
VSM	visual saliency map
MSD	multi-scale decomposition
RGF	rolling guidance filter

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13014-024-02409-6.

Supplementary Material 1

Acknowledgements

This work was supported by Swiss National Science Foundation under grant SNSF 320030_176052 and by the Distinguished Professor Program of Óbuda University, Budapest Hungary.

Author contributions

Design of the work: ZM and YS. Acquisition, interpretation and analysis: MA, GH, MO, IS, HZ. Funding and supervision: HZ. Interpretation of data. Drafted or substantively revision of the work: All authors. The authors read and approved the final manuscript.

Funding
Open access funding provided by Óbuda University.

Data availability
The data are available from The Cancer Imaging Archive (TCIA) open access repository.

Declarations

Ethics approval and consent to participate
This study used data from The Cancer Imaging Archive (TCIA) open access repository.

Consent for publication
The authors consent to publish the manuscript in its current form.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva, Switzerland
[2]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada
[3]Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
[4]Department of Nuclear Medicine, University of Southern Denmark, Odense, Denmark
[5]University Research and Innovation Center, Óbuda University, Budapest, Hungary

Received: 11 April 2023 / Accepted: 17 January 2024
Published online: 22 January 2024

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Cancer J Clin. 2021;71(3):209–49.
2. Grossberg A, Mohamed A, Elhalawani H, Bennett W, Smith K, Nolan T, et al. Data from head and neck cancer CT atlas. The Cancer Imaging Archive. 2017;10:K9.
3. Global Burden of Disease, Cancer C, Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, et al. Global, Regional, and National Cancer incidence, mortality, years of Life Lost, Years lived with disability, and disability-adjusted life-years for 32 Cancer groups, 1990 to 2015: a systematic analysis for the global burden of Disease Study. JAMA Oncol. 2017;3(4):524–48.
4. 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):1–9.
5. Blanchard P, Baujat B, Holostenco V, Bourredjem A, Baey C, Bourhis J, et al. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): a comprehensive analysis by tumour site. Radiother Oncol. 2011;100(1):33–40.
6. FH T, CYW C. Radiomics AI prediction for head and neck squamous cell carcinoma (HNSCC) prognosis and recurrence with target volume approach. BJR| Open. 2021;3(1):20200073.
7. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med. 2010;363(1):24–35.
8. Wu A, Li Y, Qi M, Lu X, Jia Q, Guo F, et al. Dosiomics improves prediction of locoregional recurrence for intensity modulated radiotherapy treated head and neck cancer cases. Oral Oncol. 2020;104:104625.
9. Jia WH, Huang QH, Liao J, Ye W, Shugart YY, Liu Q, et al. Trends in incidence and mortality of nasopharyngeal carcinoma over a 20–25 year period (1978/1983–2002) in Sihui and Cangwu counties in southern China. BMC Cancer. 2006;6(1):178.
10. Deasy JO, Niemierko A, Herbert D, Yan D, Jackson A, Ten Haken RK, et al. Methodological issues in radiation dose-volume outcome analyses: summary of a joint AAPM/NIH workshop. Med Phys. 2002;29(9):2109–27.
11. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441–6.
12. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. Physica Med. 2017;38:122–39.
13. Lv W, Yuan Q, Wang Q, Ma J, Jiang J, Yang W, et al. Robustness versus disease differentiation when varying parameter settings in radiomics features: application to nasopharyngeal PET/CT. Eur Radiol. 2018;28(8):3245–54.
14. Marur S, Forastiere AA, editors. Head and neck squamous cell carcinoma: update on epidemiology, diagnosis, and treatment. Mayo Clinic Proceedings; 2016: Elsevier.
15. Mroz EA, Tward AD, Hammon RJ, Ren Y, Rocco JW. Intra-tumor genetic heterogeneity and mortality in head and neck cancer: analysis of data from the Cancer Genome Atlas. PLoS Med. 2015;12(2):e1001786.
16. Mu W, Qi J, Lu H, Schabath M, Balagurunathan Y, Tunali I, et al. editors. Radiomic biomarkers from PET/CT multi-modality fusion images for the prediction of immunotherapy response in advanced non-small cell lung cancer patients. Medical Imaging 2018: Computer-Aided Diagnosis; 2018: International Society for Optics and Photonics.
17. Riyahi S, Choi W, Liu C-J, Nadeem S, Tan S, Zhong H, et al. Quantification of local metabolic tumor volume changes by registering blended PET-CT images for prediction of pathologic tumor response. Perinatal, and Paediatric Image Analysis: Springer;: Data Driven Treatment Response Assessment and Preterm; 2018. pp. 31–41.
18. 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. 2020;24(8):2268–77.
19. Vallieres 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(14):5471–96.
20. Liang B, Yan H, Tian Y, Chen X, Yan L, Zhang T, et al. Dosiomics: extracting 3D spatial features from dose distribution to predict incidence of radiation pneumonitis. Front Oncol. 2019;9:269.
21. Gabryś HS, Buettner F, Sterzing F, Hauswald H, Bangert M. Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of xerostomia. Front Oncol. 2018;8:35.
22. Guo Y, Jiang W, Lakshminarayanan P, Han P, Cheng Z, Bowers M, et al. Spatial Radiation dose influence on Xerostomia Recovery and its comparison to Acute incidence in patients with Head and Neck Cancer. Adv Radiat Oncol. 2020;5(2):221–30.
23. Han P, Lakshminarayanan P, Jiang W, Shpitser I, Hui X, Lee SH, et al. Dose/Volume histogram patterns in salivary gland subvolumes influence xerostomia injury and recovery. Sci Rep. 2019;9(1):3616.
24. Jiang W, Lakshminarayanan P, Hui X, Han P, Cheng Z, Bowers M, et al. Machine learning methods uncover radiomorphologic dose patterns in salivary glands that predict xerostomia in patients with head and neck cancer. Adv Radiation Oncol. 2019;4(2):401–12.
25. Lakshminarayanan P, Jiang W, Robertson S, Cheng Z, Han P, Bowers M, et al. Radio-Morphology: Parametric shape-based features for Outcome Prediction in Radiation Therapy. Int J Radiat Oncol Biol Phys. 2018;102(3):212.
26. Nakatsugawa M, Cheng Z, Goatman K, Lee J, Robinson A, Choflet A, et al. Radiomic analysis of salivary glands and its role for predicting xerostomia in irradiated head and neck cancer patients. Int J Radiat Oncol Biol Phys. 2016;96(2):217.
27. Rossi L, Bijman R, Schillemans W, Aluwini S, Cavedon C, Witte M, et al. Texture analysis of 3D dose distributions for predictive modelling of toxicity rates in radiotherapy. Radiother Oncol. 2018;129(3):548–53.
28. Sheikh K, Lee SH, Cheng Z, Lakshminarayanan P, Peng L, Han P, et al. Predicting acute radiation induced xerostomia in head and neck Cancer using MR and CT Radiomics of parotid and submandibular glands. Radiat Oncol. 2019;14(1):1–11.
29. Liang B, Tian Y, Chen X, Yan H, Yan L, Zhang T, et al. Prediction of Radiation Pneumonitis with dose distribution: a convolutional neural network (CNN) based model. Front Oncol. 2019;9:1500.
30. Lee SH, Han P, Hales RK, Voong KR, Noro K, Sugiyama S, et al. Multi-view radiomics and dosiomics analysis with machine learning for predicting acute-phase weight loss in lung cancer patients treated with radiotherapy. Phys Med Biol. 2020;65(19):195015.
31. Cai C, Lv W, Chi F, Zhang B, Zhu L, Yang G et al. Prognostic generalization of multi-level CT-dose fusion dosiomics from primary tumor and lymph node in nasopharyngeal carcinoma. Med Phys. 2022.

32. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26(6):1045–57.
33. Vallieres M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts H, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017;7(1):10117.
34. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Khaouam N, et al. Data from head-neck-PET-CT. The Cancer Imaging Archive. 2017;10:K9.
35. Grossberg AJ, Mohamed AS, Elhalawani H, Bennett WC, Smith KE, Nolan TS, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Sci data. 2018;5(1):1–10.
36. MICCAI/MD. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. Sci Data. 2017;4:170077.
37. Ma J, Zhou Z, Wang B, Zong H. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. Infrared Phys Technol. 2017;82:8–17.
38. Li H, Wu X-J. Infrared and visible image fusion using latent low-rank representation. arXiv preprint arXiv:180408992. 2018.
39. Ashrafinia S. Quantitative nuclear medicine imaging using advanced image reconstruction and radiomics. Johns Hopkins University; 2019.
40. Zwanenburg A, Vallieres M, Abdalah MA, Aerts H, Andrearczyk V, Apte A, et al. The image Biomarker Standardization Initiative: standardized quantitative Radiomics for High-Throughput Image-based phenotyping. Radiology. 2020;295(2):328–38.
41. McNitt-Gray M, Napel S, Jaggi A, Mattonen SA, 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(2):118–28.
42. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. J Am Stat Assoc. 2010;105(489):205–17.
43. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random survival forests for high-dimensional data. Sci J. 2011;4(1):115–32. Statistical Analysis and Data Mining: The ASA Data.
44. De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. mRMRe: an R package for parallelized mRMR ensemble feature selection. Bioinformatics. 2013;29(18):2365–8.
45. Kursa MB, Rudnicki WR. Feature selection with the Boruta Package. J Stat Softw. 2010;36(11):13.
46. Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. Bioinformatics. 2009;25(7):890–6.
47. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics. 2008;2(3):841–60.
48. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. The Annals of Statistics. 1982:1100–20.
49. 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(1–2):3–35.
50. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39(5):1.
51. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC press; 1984.
52. Sun W, Jiang M, Dang J, Chang P, Yin F-F. Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. Radiat Oncol. 2018;13(1):1–8.
53. Mijnheer B, Battermann J, Wambersie A. What degree of accuracy is required and can be achieved in photon and neutron therapy? Radiother Oncol. 1987;8(3):237–52.
54. Wittkämper F, Mijnheer B. Dose intercomparison at the radiotherapy centers in the Netherlands. 3. Characteristics of electron beams. Radiother Oncol. 1993;27(2):156–63.
55. Yang K, Tian J, Zhang B, Li M, Xie W, Zou Y, et al. A multidimensional nomogram combining overall stage, dose volume histogram parameters and radiomics to predict progression-free survival in patients with locoregionally advanced nasopharyngeal carcinoma. Oral Oncol. 2019;98:8591.
56. Ibragimov B, Toesca D, Chang D, Yuan Y, Koong A, Xing L. Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT. Med Phys. 2018;45(10):4763–74.

Publisher's Note