Metastatic Melanoma Treated by Immunotherapy: Discovering Prognostic Markers from Radiomics Analysis of Pretreatment CT with Feature Selection and Classification

Gulnur Ungan[1], Anne-Flore Lavandier[2], Jacques Rouanet[3], Constance Hordonneau[2], Benoit Chauveau[2], Bruno Pereira[4], Louis Boyer[2], Jean-Marc Garcier[2,5], Sandrine Mansard[3], Adrien Bartoli[1], Benoît Magnin[1,2,5 §]

1 EnCoV, Institut Pascal, UMR 6602 CNRS, Université Clermont Auvergne, 28 place Henri Dunant, 63000 Clermont-Ferrand, France
2 Department of Medical Imaging, CHU Clermont-Ferrand, 1 place Lucie Aubrac, 63100 Clermont-Ferrand, France
3 Dermatology Department, CHU Clermont-Ferrand, 1 place Lucie Aubrac, 63100 Clermont-Ferrand, France
4 Biostatistics Unit, DRCI, CHU Clermont-Ferrand, 58 rue Montalembert, 63000 Clermont-Ferrand, France
5 Anatomy Department, Université Clermont Auvergne, 28 place Henri Dunant, 63000 Clermont-Ferrand, France

§ Corresponding author:
Benoît Magnin
Orcid: 0000-0002-4246-5688
Radiologie et Imagerie Médicale Estaing
CHU Clermont - Ferrand
1, place Lucie Aubrac
63003 Clermont-Ferrand Cedex 1
Phone : 0473750244
Fax: 0473750239
bmagnin@chu-clermontferrand.fr

Purpose

Immunotherapy has dramatically improved the prognosis of patients with metastatic melanoma (MM). Yet, there is a lack of biomarkers to predict if a patient will benefit from immunotherapy. Our aim was to create radiomics models on pretreatment computed tomography (CT) to predict overall survival (OS) and treatment response in patients with MM treated with anti-PD1 immunotherapy.

Methods

We performed a monocentric retrospective analysis of 503 metastatic lesions in 71 patients with 46 radiomics features extracted following lesion segmentation. Predictive accuracies for OS <1 year vs >1 year and treatment response vs no response were compared for five feature selection methods (Sequential Forward Selection, Recursive, Boruta, Relief, Random Forest) and four classifiers (Support Vector Machine (SVM), Random Forest, K Nearest Neighbour, Logistic Regression (LR)) used with or without SMOTE data augmentation. A 5-fold cross-validation was performed at the patient level, with a tumour-based classification.

Results

The highest accuracy level for OS prediction was obtained with 3D lesions (0.91), without clinical data integration, when combining Boruta feature selection and the LR classifier. The highest accuracy for treatment response prediction was obtained with 3D lesions (0.88), without clinical data integration, when combining Boruta feature selection, the LR classifier and SMOTE data augmentation. The accuracy was significantly higher concerning OS prediction with 3D segmentation (0.91 vs 0.86) while clinical data integration led to improved accuracy notably in 2D lesions (0.76 vs 0.87) regarding treatment response prediction. Skewness was the only feature found to be an independent predictor of OS (HR (CI 95%) 1.34, *p*-value 0.001).

Conclusion

This is the first study to investigate CT texture parameter selection and classification methods for predicting MM prognosis with treatment by immunotherapy. Combining pretreatment CT radiomics features from a single tumour with data selection and classifiers may accurately predict OS and treatment response in MM treated with anti-PD-1.

KEYWORDS

Metastatic melanoma, Immunotherapy, Texture analysis, Survival, Biomarker

ABBREVIATIONS

CAD	Computer Aided Diagnosis
CI	Confidence Interval
CTLA-4	Cytotoxic T-Lymphocyte-Associated protein 4
HR	Hazard Ratio
KNN	K Nearest Neighbour
LDH	Serum lactate dehydrogenase
MM	Metastatic Melanoma
OS	Overall Survival
PD-1	Program cell Death 1
PFS	Progression-Free Survival
RECIST	Response Evaluation Criteria In Solid Tumours
RF	Random Forest
ROI	Region Of Interest
SFS	Sequential Forward Selection
SMOTE	Synthetic Minority Oversampling TEchnique
SVM	Support Vector Machine

1. INTRODUCTION

Melanoma is a primary skin cancer causing approximately 55,500 deaths annually [1]. Metastatic Melanoma (MM) has the highest mortality rate with an estimated three-year survival rate of between 5% and 32% [2]. An early breakthrough in the treatment of MM involved targeted therapy used to block BRAF and MEK; a treatment available to only 40% of patients, for whom the tumour presents a BRAF V600 mutation [3]. The introduction of immunotherapy, (CTLA-4 checkpoint inhibitor: ipilimumab, and later PD-1 checkpoint inhibitor: pembrolizumab and nivolumab) associated with long overall survival (OS) in MM [4,5] constituted a second major breakthrough, and has become a common first-line therapy in MM.

Improvement in survival rates, following treatment by checkpoint inhibitor therapy however, remains heterogeneous. Predictors of immunotherapy response are essential as they enable clinicians to evaluate benefits of immunotherapy, to spare patients unnecessary risk of toxicity [6], to select the most suitable targeted therapy and to enrol patients in clinical trials [7].

Predictors in current use include the LDH value, visceral tumour burden (notably the presence of liver lesions), the relative eosinophil count, the relative lymphocyte count and age [8,9]. In addition, visual analysis of CT images performed by a radiologist may provide prognostic predictors such as the number, size and shape of lesions and number of metastatic sites. The heterogeneity of lesions, unquantifiable by the human eye, is reported to be a prognostic factor in some tumours [10]. Texture analysis, a technique used to quantify tumour heterogeneity [11,12], provides an analysis of the relationship between and distribution of pixel grey-levels in the tumour, thus revealing the spatial variation of grey-levels in image patches.

Radiomics involves the extraction of high-dimensional sets of imaging features that characterize intra-tumoural heterogeneity. These features can be used to build models providing clinicians with key information for clinical diagnosis and assessment of prognosis and therapeutic effects. The extracted image features can be combined with other clinical, biological and genomic data, thus increasing the power of decision support systems.

A large body of studies have reported on the benefit of texture analysis in many types of cancer, namely colorectal cancer [13,14], hepatocellular carcinoma [15], Hodgkin lymphoma [11], non small-cell lung cancer [16–18], soft tissue sarcoma [19], oesophageal cancer [20,21] and head and neck cancer [22], and provided information on survival rates, evaluation of treatment response [21,23–27] and histological characterisation of lesions [28–30].

A few studies have reported on the use of Computer Aided Diagnosis (CAD) including radiomics features, in MM. Some studies identified prognostic factors: Smith et al. [31] used radiomics features from 23 CT obtained before and following initiation of treatment by bevacizumab (which is no longer administered), to identify prognostic factors of survival ; Durot et al. [32] analysed average radiomics features on multiple lesions of 31 patients before administering pembrolizumab, and found predictors of OS. Some studies built prognostic models: Schraag et al. [33] extracted 7 first order texture parameters from the largest lesion in 103 patients prior to immunotherapy and built a multivariable Cox regression model with clinical and texture parameters to enable prediction of OS ; Wang et al. [34] selected radiomics features from the largest lesion in 50 patients before immunotherapy and built a model with a Support Vector Machine (SVM, [35]) able to predict first cycle response based on validation of data from 16 patients.

Feature selection is a key step in building a radiomics model. As some radiomics software can extract up to a thousand features [36], dimensionality reduction is crucial to prevent overfitting. The most common methods are logistic regression and the Least Absolute Shrinkage and Selection Operator (LASSO). From the selected features, a regression or a classification model can be built, the most common ones being SVM and Random Forest (RF). Some studies used several feature selection methods and classifiers in other pathologies [37–39]. Our study is the first to study feature selection and classification methods for evaluation of MM prognosis.

The purpose of this study was to compare feature selection and classification methods, from radiomics features taken from contrast-enhanced CT images before initiation of treatment, so as to predict OS and treatment response in patients with MM treated by pembrolizumab or nivolumab.

2. PATIENTS AND METHODS

2.1 Study population

This monocentric retrospective cohort study was approved by our institutional review board (authorization number CRM-1905-008).

All patients treated with anti PD-1 immunotherapy (pembrolizumab or nivolumab, identified from the hospital pharmacy database) between July 2014 and September 2018 for MM were included in the study. Exclusion criteria were: no metastasis visible on CT scans, no delineable lesion, no baseline CT scan performed within 2.5 months before beginning of treatment and a clinical follow-up period lower than one year (unless death occurred during the first year).

Recorded data included age, gender, BRAF mutation status, date of metastatic status, start date of immunotherapy treatment, date of pretreatment CT scan, number of metastatic sites, presence of hepatic, cerebral and lung metastases, number of segmented lesions, 3 months follow-up iRECIST conclusion, death, decision for supportive care or change of treatment.

2.2 Follow-up and endpoints

All patients underwent clinical, biological, and radiological follow-up every 3 months, in accordance with the hospital protocol. The radiological follow-up comprised contrast-enhanced CT scans of the brain, thorax, abdomen and pelvis. In cases of suspected progression, an additional CT was performed 6 weeks later to rule out pseudo progression.

Two endpoints were used for classification. The first endpoint was OS, for which patients were allocated to two groups based on the survival period; lower than one year after treatment initiation and longer than one year. The second endpoint was treatment response, obtained from the first CT scan taken 3 months after treatment initiation using iRECIST. Patients were allocated to two groups; favorable prognosis for stable disease or partial response and unfavorable prognosis for progression.

2.3 CT examination

The majority of CT scans (65/71, 92%) were performed on a 64-section contrast-enhanced CT scanner (Discovery CT 750 HD, GE Healthcare). A volume of 1.5 mL/kg body weight of non-ionic contrast material was injected into the peripheral upper limb vein at a flow rate of 3 $mL.s^{-1}$. Chest, abdominal and pelvic images were obtained at a portal-venous phase (80 s), and cerebral images were obtained at a late phase (5 minutes after injection).

The acquisition parameters were as follows: 100 kVp tube voltage; helical pitch of 1.375; image reconstruction thickness of 2.5 mm. The images were reconstructed using 30% adaptive statistical reconstruction (ASiR, GE Healthcare).

Six CTs were performed on other scanners located outside our hospital. Visual assessment allowed us to ensure the quality of acquisition. The time of injection was checked before inclusion.

2.4 Data segmentation

Lesion segmentation was performed manually with the pretreatment CT scanner, using LIFEx (version 4.00, www.lifexsoft.org) [40] an ISBI-compliant feature extraction platform. Segmentation was performed following consensus by two radiologists: a senior radiologist with 8 years experience in radiology (6 years in oncological imaging) and a radiology resident with 4 years experience.

All segmentable lesions (sufficient size and definition) were segmented on the CT. A 3D ROI corresponded to segmentation of the whole lesion, while the axial slice with the largest area of the segmented lesion was saved as the 2D ROI.

2.5 Feature extraction

For each segmented ROI (3D and 2D ROIs for each lesion), 46 parameters were extracted by LIFEx, (Supplementary Information 1) and divided into 3 categories according to shape (volume, sphericity, compacity), histogram of grey level (skewness, kurtosis, entropy, energy) and texture parameters.

2.6 Quantification of noise of the radiomics features

To quantify the distribution of noise in the data, we evaluated for each radiomics feature separately on 2D and 3D ROIs the coefficient of variation, defined as the ratio between standard deviation and absolute value of the mean value of the feature ($\frac{\sigma}{|\bar{x}|}$).

2.7 Data augmentation, feature selection and classification methods

To adjust for imbalanced datasets, data augmentation was used using the Synthetic Minority Oversampling TEchnique (SMOTE) [41].

A total of 5 different feature selection algorithms were used to select features extracted from CT data, namely Sequential Forward Selection (SFS), Boruta, Relief, Recursive and RF feature selection algorithms.

Additional clinical data included age; sex; previous treatment by other immunotherapy; BRAF status; presence of lymph node, liver, brain, lung, adrenal, spleen, bone or gastrointestinal tract metastasis.

The 4 classification methods used to classify the selected features from CT imagery and clinical data were SVM (using a linear kernel) [35], RF, K-Nearest Neighbour (KNN), and Logistic Regression.

A total of 40 combinations were tested (with or without data augmentation x 5 feature selection x 4 classification methods).

Classification was firstly performed with radiomics features only, and secondly along with clinical data; the features from 3D and 2D ROIs were processed separately.

A 5-fold cross validation algorithm was used for each classification task (OS and treatment response). For each model, data was split into a training set (80% of the patients) and a test data set (the remaining 20% of the patients). The split between different folds was done on the patient level. Tumours in the test data set were then classified, resulting in a tumour based classification. The split train-test procedure was repeated 5 times and allowed calculation of mean values for accuracy, sensitivity and specificity.

In addition to accuracy, sensitivity and specificity values were used in performance analysis.

2.8 Influence of feature selection and classification methods on performance

To quantify the influence of feature selection methods, we evaluated the mean performance on all combinations (classifiers, with or without data augmentation) of each feature selection method.

The same evaluation was made to quantify the influence of classifiers.

The Shapiro test rejected the normality of the distribution of accuracy. Hence, non-parametric tests were used. The Kruskal Wallis test was used to search for significant differences of accuracy between the 5 feature selection methods first, and the 4 classifiers then.

In case of significant differences, a pairwise comparison was made with the Wilcoxon test, applying Bonferroni's correction for multiple comparisons.

2.9 Quantification of fit

In order to evaluate the models' fit, we calculated the train and test accuracies for each model, and evaluated the train/test accuracy ratio.

2.10 Statistical tests

We used the Cox proportional hazards regression model to assess the association between survival parameters and covariates of interest. The proportional-hazard hypothesis was tested using the Schoenfeld test and results were expressed as hazard-ratios and 95% confidence intervals (95% CI). All statistical tests were two-sided, with *p*-values under 0.05 considered statistically significant.

Data extraction, feature selection and classification were performed with Python using a custom script.

2.11 Inter-observer reproducibility

The assessment of intra-observer reproducibility was based on an initial analysis of 10% of segmentations following random selection, and on a repeat analysis, blinded to the first, performed at a six month interval. Assessment included comparison of features from both segmentations, extracted using LIFEx and an estimation of the Lin concordance correlation. The results were analysed according to conventional rules defined in the literature [42,43]: 0-0.2 (negligible agreement), 0.2-0.4 (low/weak agreement), 0.4-0.6 (moderate agreement), 0.6-0.8 (substantial/good agreement) and >0.8 (strong agreement).

3. **RESULTS**

3.1 **Patient characteristics**

Of 79 eligible patients, 8 were excluded (5 with difficult-to-define lesions, 2 without metastasis, 1 without pretreatment CT evaluation). A total of 71 patients (41 men, 30 women) of median age 66 years, (interquartile range 34-90) were included.

The main baseline patient clinical and radiological characteristics are given in Table 1.

A total of 906 lesions (503 3D lesions and 403 2D lesions; minimum 1 per patient, mean 7 and maximum 31) were segmented. 35% (25/71) of patients presented with oligo metastatic lesions (less than 3 metastatic lesions), 38% (27/71) had less than 10 lesions and 27% (19/71) more than 10, 35% (25/71) presented with hepatic lesions, 31% (22/71) with cerebral lesions and 55% (39/71) with pulmonary lesions.

The main follow-up data are given in Table 2, with a mean follow-up of 882 days.

3.2 **Reproducibility**

Lin's concordance correlation coefficient was >0.8 for 92% of the features while for the remaining features, the coefficient was ≤ 0.20, 0.21-0.40, 0.41 -0.60, and 0.61-0.80 for 2%, 0%, 4%, and 2% of the features, respectively.

3.3 **Overall survival prediction**

The best 3 combinations for 2D and 3D lesions with and without clinical data are shown in Table 3. The performance of all combinations are presented in Supplementary Information 2.

For 2D radiomics features, the best results were obtained using Recursive feature selection combined with logistic regression classification and SMOTE data augmentation (Acc 0.86, Sen 0.7, Spe 0.69).

For 3D radiomics features, the best results were obtained using Boruta feature selection and logistic regression classification (Acc 0.91, Sen 0.79, Spe 0.39).

3.4 Treatment response prediction

The best 3 combinations for 2D and 3D lesions without and with clinical data are shown in Table 4.

The best results were obtained for 2D radiomics features and clinical data integration with Random Forest selection and SVM classification (Acc 0.87, Sen 0.44, Spe 0.8).

For 3D radiomics features, the best combination was obtained using Recursive feature selection combined with logistic regression classification and SMOTE, with clinical data integration (Acc 0.83, Sen 0.65, Spe 0.6).

3.5 Overall survival analysis

Cox proportional hazard models were calculated in 3 ways: as a whole radiomics covariate (shape, histogram and texture features), histogram features only and texture features only.

For whole radiomics covariates, significant features were shape sphericity (p=0.012), GLZLM-LZE (p=0.041), GLZLM-LZHGE (p=0.037), GLZLM-ZLNU (p=0.003) but CI was non-significant.

Concerning histogram features (Table 5), skewness was significantly correlated with OS (p=0.012, CI 95%=1.07-1.7).

Concerning texture features (Table 5), the p-values were significant for GLRLM-SRE (p=0.022), NGLDM-contrast (p=0.032), GLZM-LZE (p=0.037), GLZM-LZHGE (p=0.041) and GLZLM-ZLNM (p=0.011, but only NGLDM-contrast (p=0.032, CI 95%=1.9×10^{-3} - 0.7×10^{-1}) had a significant CI.

3.6 Quantification of noise of the radiomics features

The median coefficient of variation was 1.076 for 2D features and 0.634 for 3D features. The complete set of values is given in Supplementary Information 3.

3.7 Influence of feature selection and classification methods on performance

The statistics of performance on all classifications with one feature selection method or with one classifier are given in Supplementary Information 4 (Tables and Figures 4.1 and 4.2).

There was no difference between the mean accuracies of all feature selection methods (Kruskal Wallis test, p=0.635).

There was a statistical difference between the mean accuracies of classifiers (Kruskal Wallis test, $p=1.3\times10^{-10}$). Pairwise comparison showed that LR and SVM each had a significantly higher accuracy than RF and KNN (Supplementary Information 4, Table 4.3).

3.8 Quantification of fit

The mean training error was 0.18 for the 12 best OS prediction models and for the 12 best response prediction models (Supplementary Information 5, Tables 5.1 and 5.2)

The mean ratio of train/test accuracy was 1.01 for the 12 best OS prediction models and 1.06 for the 12 best response prediction models (Supplementary Information 5, Tables 5.1 and 5.2). The mean ratios of train/test accuracy and the values of train and test accuracy on all classifications are given in Supplementary Information 5, Table 5.3 and 5.4.

4. DISCUSSION

We investigated the performance of different radiomics models as a prognostic tool to predict OS and treatment response, in patients with metastatic melanoma treated by anti PD-1, on pretreatment CT images. We combined 5 feature selection methods with 4 classification methods with or without SMOTE data augmentation on any segmentable lesion, resulting in a tumour based classification. The accuracy of the ten best classification methods for predicting OS up to and beyond one year, and treatment response was found to be good (>0.80).

To date, only 4 studies have reported on the prognosis of patients with MM, based on radiomics parameters. Smith et al. [31] used radiomics features from a CT obtained before treatment and modifications in the features from a CT taken after the initiation of treatment, to identify prognostic factors of survival. This study was however based on a small number of patients (23), the use of a treatment which is no longer administered (bevacizumab) and data recording at 3 months after initiation of treatment.

Durot et al. [32] investigated the association of pre-treatment CT scan texture parameters with OS and progression-free survival, in patients treated with pembrolizumab. The model was built using LASSO penalized Cox regression from 5 histogram features. They found that skewness values above −0.55 at coarse texture scale were significantly associated with both lower OS and lower PFS. The study however has several limitations. Firstly, the low number of patients (31 compared to 71 in our study) and a limited number of lesions per patient (5 maximum). They reported on 74 lesions in total compared to 906 in our study. Secondly, lesion contours concerned single axial sections only rather than the whole tumour in 3D which impedes assessment of tumour heterogeneity and contour replication. Thirdly, few texture parameters were extracted (compared to 46 in our study) with reporting of only average parameter values and including values from other organs. Fourthly, Durot et al used RECIST 1.1 to establish treatment response, without taking into account the pseudo progression phenomenon. Hodi et al. [44] noted that RECIST 1.1 may lead to underestimating responses in 15% of patients and results in early discontinuation of treatment. Finally, the absence of a validation process (validation cohort or cross-validation) weakens the strength of the main result, as the threshold of skewness coarse texture scale was determined and evaluated on the same population. Yet, the only radiomics feature we found to be significantly correlated with OS was skewness, as in the study by Durot et al. [32].

Schraag et al. [33] extracted 7 first order texture parameters from the largest lesion in 103 patients prior to immunotherapy. Their model, built on clinical and texture parameters with a multivariable Cox regression, enabled the prediction of OS (C-index 0.716) but texture parameters did not allow the prediction of treatment response.

In a recent publication, Wang et al. [34] extracted 497 radiomics features from the largest lesion in 50 patients prior to immunotherapy. On the basis of a selection of features by T-test and redundancy, their model using SVM was shown to predict first cycle response in a validation cohort of 16 patients (accuracy 75%), but without survival predictions.

The majority of the above-mentioned studies recorded data from one lesion only (usually the largest) to predict OS or treatment response, with the exception of Durot et al. who reported an average value of various lesions. By performing a tumour-based classification, our study enabled us to evaluate the ability to predict OS or treatment response from any single lesion of a patient, regardless of its size.

Moreover, all these studies used only one model to assess the prognosis in MM. Other studies have compared radiomics model performance regarding prediction of clinical event occurrence in several other cancers by using various feature selection and classification methods: lung [39,45,46], preoperative differentiation of sacral chordoma and sacral giant cell tumour [38], head and neck cancer [47], rectal cancer [37]. Palmar et al. noted that the choice of classification method is the major factor driving the performance variation [47].

The present study compares a total of 160 combinations (5 feature selection methods, 4 classification methods, Smote Data Augmentation, for 2D and 3D lesions, integration or not of clinical data, response therapy and OS), using methods shown in previous studies to provide the best performance. This is the first study to investigate texture parameter selection and classification methods for predicting MM prognosis with treatment by immunotherapy.

The highest performance for OS prediction (accuracy 0.95) was found when combining Recursive feature selection with a logistic regression classifier, while for treatment response (accuracy 0.90) this was found when combining SFS selected features, a RF classifier and integration of clinical data.

For all methods, 3D segmentation provided better results than 2D segmentation (12% accuracy increase). This supports evidence reported by Ortiz-Ramon et al. [29] and Ng et al. [13]. Clinical data integration led to a greater increase in accuracy for 2D features, than for 3D features, notably concerning prediction of treatment response.

No combination of feature selection and classification method emerged clearly as the best for the different data (2D vs 3D, with or without clinical data integration, treatment response or survival). The reasons for the variability of results depending on the "pipeline" (*i.e.*, combination of feature selection, data augmentation and classifier) are difficult to investigate and rarely fully addressed in the literature. Yet, a few articles try to address the point, among which are the articles by Parmar et al. [45,47] which attempt to quantify the impact of the methods on the results. They however do not give a detailed explanation about the characteristics that may explain the differences between different methods. The other articles attempting to address this point simply recall general properties of the methods [38].

Generally speaking, the model's performance can first be explained by a possible unusually good or bad model fitting. The repetition of experiments, and the 5-fold cross validation make sure the model fitting is not a special lucky or unlucky case. Second, the model

performance can be explained by the model's lack of fit. Our data concerning the fit on the training data show that the model did not underfit, as the mean train error was 0.18 for the 12 best OS prediction model and 0.18 for the 12 best response prediction model (Supplementary Information 5, Supplementary Tables 5.1 and 5.2). Lastly, the performance can be explained by the model's expressivity. To assess the model's expressivity, we computed the train/test ratio, which was 1.01 for the 12 best OS prediction model and 1.06 for the 12 best response prediction model (Supplementary Information 5, Supplementary Tables 5.1 and 5.2), showing the model's ability to generalise and its sufficient expressivity.

However, the influence of feature selection methods appears to be moderate in our models, as the performances of all feature selection methods are similar, without any significant difference (Supplementary Information 4, Table and Figure 4.1). It has been indeed noted that the influence of feature selection methods can vary depending on the type of cancer. Namely, feature selection has high impact in lung cancer and twice less in head and neck cancer [47], hence leaving the possibility that feature selection can have little impact on some cancer types, including melanoma, as per our findings.

SMOTE data augmentation had a mixed effect on the results, positive for some combinations and negative for others. It is however worth of note that 11 of the 12 best results for prediction of treatment response involved SMOTE data augmentation.

Concerning the classifier, LR and SVM emerged as the two best classifiers: out of the 24 best results, 12 were performed with LR and 6 with SVM. Moreover, mean performances on all classifications of those two classifiers were significantly better than those of the 2 other classifiers (Supplementary Information 4, Table and Figure 4.2).

Therefore, in our study, the variability depends mainly on the choice of the classifier, the 2 best being LR and SVM. LR is a supervised machine learning classification algorithm that does not make any assumptions regarding the distribution of independent variables. It is a commonly used classifier, providing average performances in classification tasks [48].

SVM is a supervised machine learning algorithm. It aims to find the best hyperplane to split a dataset into two classes. It is often reported to be more robust than LR, with a lower risk of overfitting. It is one of the most used classifiers, and appears to be one of the best classifiers, notably in disease prediction studies [48]. However, its behaviour depends on the type of kernel used; a linear kernel was used in our study. It can be said that linear SVM and LR have

similar behaviours to find a margin between different classes. Moreover, Musa et. al. demonstrated that SVM and LR can work similarly for different scenarios such as balanced and unbalanced datasets [49].

It has been shown that LR consistently performs with a higher overall accuracy as compared to RF when increasing the variance of the noise data [50]. The median coefficient of variation was high (1.076 for 2D features and 0.634 for 3D features), showing noise in our data, hence explaining the improved performances of LR compared to RF. The degree of noise can also explain why LR outperforms SVM in our data [51]. Concerning the performances of KNN, one of the reasons explaining the lack of performances is the fact that the data were not normalised or rescaled.

The limitations of this study include, firstly, that 6 pretreatment CT were performed on a different CT scan, requiring, prior to inclusion, visual assessment to ensure scan quality and time of injection. Secondly, this study was a retrospective monocentric study that, despite the large number of analysed lesions (503 in 2D and 403 in 3D), relied on a relatively small number of patients. We were therefore unable to split the population into training and validation cohorts, with cross-validation ensured by a 5-fold cross-validation algorithm. Our patient number remains however larger than that of most previous studies and reflects the relative rareness of the disease. Finally, our model could be improved by including more biological or genetic features such as LDH level, which was not possible due to insufficient patient data.

Our study involved the use of 5 commonly used feature selection and 4 classifier methods with encouraging results. Future research is required to evaluate the performance of more complex classification methods such as those built on deep learning.

5. CONCLUSION

Our study showed that the combination of CT texture analysis, data selection and classification algorithms may accurately predict treatment response and overall survival for patients starting anti-PD-1 immunotherapy for metastatic melanoma.

FUNDING

The authors confirm that no funding was received for this research.

ACKNOWLEDGEMENT

The authors would like to thank Helen Braund for language editing.

REFERENCES

[1] D. Schadendorf, A.C.J. van Akkooi, C. Berking, K.G. Griewank, R. Gutzmer, A. Hauschild, A. Stang, A. Roesch, S. Ugurel, Melanoma, The Lancet. 392 (2018) 971–984. https://doi.org/10.1016/S0140-6736(18)31559-9.

[2] X. Song, Z. Zhao, B. Barber, A.M. Farr, B. Ivanov, M. Novich, Overall survival in patients with metastatic melanoma, Curr. Med. Res. Opin. 31 (2015) 987–991. https://doi.org/10.1185/03007995.2015.1021904.

[3] J. Larkin, P.A. Ascierto, B. Dréno, V. Atkinson, G. Liszkay, M. Maio, M. Mandalà, L. Demidov, D. Stroyakovskiy, L. Thomas, L. de la Cruz-Merino, C. Dutriaux, C. Garbe, M.A. Sovak, I. Chang, N. Choong, S.P. Hack, G.A. McArthur, A. Ribas, Combined vemurafenib and cobimetinib in BRAF-mutated melanoma, N. Engl. J. Med. 371 (2014) 1867–1876. https://doi.org/10.1056/NEJMoa1408868.

[4] C. Robert, J. Schachter, G.V. Long, A. Arance, J.J. Grob, L. Mortier, A. Daud, M.S. Carlino, C. McNeil, M. Lotem, J. Larkin, P. Lorigan, B. Neyns, C.U. Blank, O. Hamid, C. Mateus, R. Shapira-Frommer, M. Kosh, H. Zhou, N. Ibrahim, S. Ebbinghaus, A. Ribas, KEYNOTE-006 investigators, Pembrolizumab versus Ipilimumab in Advanced Melanoma, N. Engl. J. Med. 372 (2015) 2521–2532. https://doi.org/10.1056/NEJMoa1503093.

[5] C. Robert, G.V. Long, B. Brady, C. Dutriaux, M. Maio, L. Mortier, J.C. Hassel, P. Rutkowski, C. McNeil, E. Kalinka-Warzocha, K.J. Savage, M.M. Hernberg, C. Lebbé, J. Charles, C. Mihalcioiu, V. Chiarion-Sileni, C. Mauch, F. Cognetti, A. Arance, H. Schmidt, D. Schadendorf, H. Gogas, L. Lundgren-Eriksson, C. Horak, B. Sharkey, I.M. Waxman, V. Atkinson, P.A. Ascierto, Nivolumab in previously untreated melanoma without BRAF mutation, N. Engl. J. Med. 372 (2015) 320–330. https://doi.org/10.1056/NEJMoa1412082.

[6] J. Martin-Liberal, T. Kordbacheh, J. Larkin, Safety of pembrolizumab for the treatment of melanoma, Expert Opin. Drug Saf. 14 (2015) 957–964. https://doi.org/10.1517/14740338.2015.1021774.

[7] S.M. Hiniker, H.T. Maecker, S.J. Knox, Predictors of clinical response to immunotherapy with or without radiotherapy, J. Radiat. Oncol. 4 (2015) 339–345. https://doi.org/10.1007/s13566-015-0219-2.

[8] B. Weide, A. Martens, J.C. Hassel, C. Berking, M.A. Postow, K. Bisschop, E. Simeone, J. Mangana, B. Schilling, A.M. Di Giacomo, N. Brenner, K. Kähler, L. Heinzerling, R. Gutzmer, A. Bender, C. Gebhardt, E. Romano, F. Meier, P. Martus, M. Maio, C. Blank, D. Schadendorf, R. Dummer, P.A. Ascierto, G. Hospers, C. Garbe, J.D. Wolchok, Baseline Biomarkers for Outcome of Melanoma Patients Treated with Pembrolizumab, Clin. Cancer Res. 22 (2016) 5487–5496. https://doi.org/10.1158/1078-0432.CCR-16-0127.

[9] A. Nosrati, K.K. Tsai, S.M. Goldinger, P. Tumeh, B. Grimes, K. Loo, A.P. Algazi, T.D.L. Nguyen-Kim, M. Levesque, R. Dummer, O. Hamid, A. Daud, Evaluation of clinicopathological factors in PD-1 response: derivation and validation of a prediction scale for response to PD-1 monotherapy, Br. J. Cancer. 116 (2017) 1141–1147. https://doi.org/10.1038/bjc.2017.70.

[10] S.-X. Rao, D.M. Lambregts, R.S. Schnerr, R.C. Beckers, M. Maas, F. Albarello, R.G. Riedl, C.H. Dejong, M.H. Martens, L.A. Heijnen, W.H. Backes, G.L. Beets, M.-S. Zeng, R.G. Beets-Tan, CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy?, United Eur. Gastroenterol. J. 4 (2016) 257–263. https://doi.org/10.1177/2050640615601603.

[11] B. Ganeshan, K.A. Miles, S. Babikir, R. Shortman, A. Afaq, K.M. Ardeshna, A.M. Groves, I. Kayani, CT-based texture analysis potentially provides prognostic information complementary to interim fdg-pet for patients with hodgkin's and aggressive non-hodgkin's lymphomas, Eur. Radiol. 27 (2017) 1012–1020. https://doi.org/10.1007/s00330-016-4470-8.

[12] V. Verma, C.B. Simone, S. Krishnan, S.H. Lin, J. Yang, S.M. Hahn, The Rise of Radiomics and Implications for Oncologic Management, JNCI J. Natl. Cancer Inst. 109 (2017). https://doi.org/10.1093/jnci/djx055.

[13] F. Ng, B. Ganeshan, R. Kozarski, K.A. Miles, V. Goh, Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival, Radiology. 266 (2013) 177–184. https://doi.org/10.1148/radiol.12120254.

[14] K.A. Miles, B. Ganeshan, M.R. Griffiths, R.C.D. Young, C.R. Chatwin, Colorectal cancer: texture analysis of portal phase hepatic CT images as a potential marker of survival, Radiology. 250 (2009) 444–452. https://doi.org/10.1148/radiol.2502071879.

[15] S. Mulé, G. Thiefin, C. Costentin, C. Durot, A. Rahmouni, A. Luciani, C. Hoeffel, Advanced Hepatocellular Carcinoma: Pretreatment Contrast-enhanced CT Texture

Parameters as Predictive Biomarkers of Survival in Patients Treated with Sorafenib, Radiology. 288 (2018) 445–455. https://doi.org/10.1148/radiol.2018171320.

[16] K.A. Miles, How to use CT texture analysis for prognostication of non-small cell lung cancer, Cancer Imaging Off. Publ. Int. Cancer Imaging Soc. 16 (2016) 10. https://doi.org/10.1186/s40644-016-0065-5.

[17] S.Y. Ahn, C.M. Park, S.J. Park, H.J. Kim, C. Song, S.M. Lee, H.P. McAdams, J.M. Goo, Prognostic value of computed tomography texture features in non-small cell lung cancers treated with definitive concomitant chemoradiotherapy, Invest. Radiol. 50 (2015) 719–725. https://doi.org/10.1097/RLI.0000000000000174.

[18] B. Ganeshan, E. Panayiotou, K. Burnand, S. Dizdarevic, K. Miles, Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival, Eur. Radiol. 22 (2012) 796–802. https://doi.org/10.1007/s00330-011-2319-8.

[19] K. Hayano, F. Tian, A.R. Kambadakone, S.S. Yoon, D.G. Duda, B. Ganeshan, D.V. Sahani, Texture Analysis of Non-Contrast-Enhanced Computed Tomography for Assessing Angiogenesis and Survival of Soft Tissue Sarcoma, J. Comput. Assist. Tomogr. 39 (2015) 607–612. https://doi.org/10.1097/RCT.0000000000000239.

[20] B. Ganeshan, K. Skogen, I. Pressney, D. Coutroubis, K. Miles, Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival, Clin. Radiol. 67 (2012) 157–164. https://doi.org/10.1016/j.crad.2011.08.012.

[21] C. Yip, D. Landau, R. Kozarski, B. Ganeshan, R. Thomas, A. Michaelidou, V. Goh, Primary esophageal cancer: heterogeneity as potential prognostic biomarker in patients treated with definitive chemotherapy and radiation therapy, Radiology. 270 (2014) 141–148. https://doi.org/10.1148/radiol.13122869.

[22] H. Zhang, C.M. Graham, O. Elci, M.E. Griswold, X. Zhang, M.A. Khan, K. Pitman, J.J. Caudell, R.D. Hamilton, B. Ganeshan, A.D. Smith, 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. 269 (2013) 801–809. https://doi.org/10.1148/radiol.13130110.

[23] S. Ramella, M. Fiore, C. Greco, E. Cordelli, R. Sicilia, M. Merone, E. Molfese, M. Miele, P. Cornacchione, E. Ippolito, G. Iannello, R.M. D'Angelillo, P. Soda, A radiomic approach for adaptive radiotherapy in non-small cell lung cancer patients, PloS One. 13 (2018) e0207455. https://doi.org/10.1371/journal.pone.0207455.

[24] S.J. Ahn, J.H. Kim, S.J. Park, J.K. Han, Prediction of the therapeutic response after FOLFOX and FOLFIRI treatment for patients with liver metastasis from colorectal cancer using computerized CT texture analysis, Eur. J. Radiol. 85 (2016) 1867–1874. https://doi.org/10.1016/j.ejrad.2016.08.014.

[25] F. Tian, K. Hayano, A.R. Kambadakone, D.V. Sahani, Response assessment to neoadjuvant therapy in soft tissue sarcomas: using CT texture analysis in comparison to tumor size, density, and perfusion, Abdom. Imaging. 40 (2015) 1705–1712. https://doi.org/10.1007/s00261-014-0318-3.

[26] M. Ravanelli, D. Farina, M. Morassi, E. Roca, G. Cavalleri, G. Tassi, R. Maroldi, Texture analysis of advanced non-small cell lung cancer (NSCLC) on contrast-enhanced computed tomography: prediction of the response to the first-line chemotherapy, Eur. Radiol. 23 (2013) 3450–3455. https://doi.org/10.1007/s00330-013-2965-0.

[27] T. Knogler, K. Thomas, K. El-Rabadi, E.-R. Karem, M. Weber, W. Michael, G. Karanikas, K. Georgios, M.E. Mayerhoefer, M. Marius Erik, Three-dimensional texture analysis of contrast enhanced CT images for treatment response assessment in Hodgkin lymphoma: comparison with F-18-FDG PET, Med. Phys. 41 (2014) 121904. https://doi.org/10.1118/1.4900821.

[28] H.C. Kniep, F. Madesta, T. Schneider, U. Hanning, M.H. Schönfeld, G. Schön, J. Fiehler, T. Gauer, R. Werner, S. Gellissen, Radiomics of Brain MRI: Utility in Prediction of Metastatic Tumor Type, Radiology. 290 (2019) 479–487. https://doi.org/10.1148/radiol.2018180946.

[29] R. Ortiz-Ramón, A. Larroza, S. Ruiz-España, E. Arana, D. Moratal, Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study, Eur. Radiol. 28 (2018) 4514–4523. https://doi.org/10.1007/s00330-018-5463-6.

[30] B. Ganeshan, V. Goh, H.C. Mandeville, Q.S. Ng, P.J. Hoskin, K.A. Miles, Non-small cell lung cancer: histopathologic correlates for texture parameters at CT, Radiology. 266 (2013) 326–336. https://doi.org/10.1148/radiol.12112428.

[31] A.D. Smith, M.R. Gray, S.M. del Campo, D. Shlapak, B. Ganeshan, X. Zhang, W.E. Carson, Predicting Overall Survival in Patients With Metastatic Melanoma on Antiangiogenic Therapy and RECIST Stable Disease on Initial Posttherapy Images Using CT Texture Analysis, Am. J. Roentgenol. 205 (2015) W283–W293. https://doi.org/10.2214/AJR.15.14315.

[32] C. Durot, S. Mulé, P. Soyer, A. Marchal, F. Grange, C. Hoeffel, Metastatic melanoma: pretreatment contrast-enhanced CT texture parameters as predictive biomarkers of survival in patients treated with pembrolizumab, Eur. Radiol. 29 (2019) 3183–3191. https://doi.org/10.1007/s00330-018-5933-x.

[33] A. Schraag, B. Klumpp, S. Afat, S. Gatidis, K. Nikolaou, T.K. Eigentler, A.E. Othman, Baseline clinical and imaging predictors of treatment response and overall survival of patients with metastatic melanoma undergoing immunotherapy, Eur. J. Radiol. 121 (2019) 108688. https://doi.org/10.1016/j.ejrad.2019.108688.

[34] Z. Wang, L. Mao, Z. Zhou, L. Si, H. Zhu, X. Chen, M. Zhou, Y. Sun, J. Guo, Pilot Study of CT-Based Radiomics Model for Early Evaluation of Response to Immunotherapy in Patients With Metastatic Melanoma, Front. Oncol. 10 (2020) 1524. https://doi.org/10.3389/fonc.2020.01524.

[35] C. Cortes, V. Vapnik, Support-Vector Networks, Mach. Learn. 20 (1995) 273–297. https://doi.org/10.1023/A:1022627411411.

[36] H. Chu, Z. Liu, W. Liang, Q. Zhou, Y. Zhang, K. Lei, M. Tang, Y. Cao, S. Chen, S. Peng, M. Kuang, Radiomics using CT images for preoperative prediction of futile resection in intrahepatic cholangiocarcinoma, Eur. Radiol. 31 (2021) 2368–2376. https://doi.org/10.1007/s00330-020-07250-5.

[37] B. Badic, R. Da-ano, K. Poirot, V. Jaouen, B. Magnin, J. Gagnière, D. Pezet, M. Hatt, D. Visvikis, Prediction of recurrence after surgery in colorectal cancer patients using radiomics from diagnostic contrast-enhanced computed tomography: a two-center study, Eur. Radiol. (2021). https://doi.org/10.1007/s00330-021-08104-4.

[38] P. Yin, N. Mao, C. Zhao, J. Wu, C. Sun, L. Chen, N. Hong, 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. 29 (2019) 1841–1847. https://doi.org/10.1007/s00330-018-5730-6.

[39] W. Sun, M. Jiang, J. Dang, P. Chang, F.-F. Yin, Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis, Radiat. Oncol. 13 (2018) 197. https://doi.org/10.1186/s13014-018-1140-9.

[40] C. Nioche, F. Orlhac, S. Boughdad, S. Reuzé, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. Frouin, I. Buvat, LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity, Cancer Res. 78 (2018) 4786–4789. https://doi.org/10.1158/0008-5472.CAN-18-0125.

[41] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, J. Artif. Intell. Res. 16 (2002) 321–357. https://doi.org/10.1613/jair.953.

[42] D.G. Altman, Practical Statistics for Medical Research, CRC Press, 1990.

[43] C.B. Terwee, S.D.M. Bot, M.R. de Boer, D.A.W.M. van der Windt, D.L. Knol, J. Dekker, L.M. Bouter, H.C.W. de Vet, Quality criteria were proposed for measurement properties of health status questionnaires, J. Clin. Epidemiol. 60 (2007) 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.

[44] F.S. Hodi, W.-J. Hwu, R. Kefford, J.S. Weber, A. Daud, O. Hamid, A. Patnaik, A. Ribas, C. Robert, T.C. Gangadhar, A.M. Joshua, P. Hersey, R. Dronca, R. Joseph, D. Hille, D. Xue, X.N. Li, S.P. Kang, S. Ebbinghaus, A. Perrone, J.D. Wolchok, Evaluation of Immune-Related Response Criteria and RECIST v1.1 in Patients With Advanced Melanoma Treated With Pembrolizumab, J. Clin. Oncol. 34 (2016) 1510–1517. https://doi.org/10.1200/JCO.2015.64.0391.

[45] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, H.J.W.L. Aerts, Machine Learning methods for Quantitative Radiomic Biomarkers, Sci. Rep. 5 (2015) 13087. https://doi.org/10.1038/srep13087.

[46] S. Hawkins, H. Wang, Y. Liu, A. Garcia, O. Stringfield, H. Krewer, Q. Li, D. Cherezov, R.A. Gatenby, Y. Balagurunathan, D. Goldgof, M.B. Schabath, L. Hall, R.J. Gillies, Predicting malignant nodules from screening CTs, J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer. 11 (2016) 2120–2128. https://doi.org/10.1016/j.jtho.2016.07.002.

[47] C. Parmar, P. Grossmann, D. Rietveld, M.M. Rietbergen, P. Lambin, H.J.W.L. Aerts, Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer, Front. Oncol. 5 (2015) 272. https://doi.org/10.3389/fonc.2015.00272.

[48] M. Fernandez-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?, (n.d.) 49.

[49] A.B. Musa, Comparative study on classification performance between support vector machine and logistic regression, Int. J. Mach. Learn. Cybern. 4 (2013) 13–24. https://doi.org/10.1007/s13042-012-0068-x.

[50] K. Kirasich, T. Smith, B. Sadler, Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets, 1 (2018) 25.

[51] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inform. Decis. Mak. 19 (2019) 281. https://doi.org/10.1186/s12911-019-1004-8.

Characteristics	*Results*
Nivolumab (n (%))	*32 (45)*
Pembrolizumab (n (%))	*39 (55)*
Men (n (%))	*41 (58)*
Women (n (%))	*30 (42)*
Mean age (mean, (min-max) yr)	*66 (34-90)*
BRAF mutation	
Yes (n (%))	*23 (32)*
No (n (%))	*48 (68)*
Time between CT and treatment (mean, (min-max) days)	*12.96 (0 - 73)*
Spread of lesions	
Lymph node (n (%))	*35 (50)*
Brain (n (%))	*22 (31)*
Lung (n (%))	*39 (55)*
Liver (n (%))	*25 (35)*
Adrenal gland (n (%))	*16 (23)*
Spleen (n (%))	*6 (8)*
Bone (n (%))	*11 (15)*
Number of segmented lesions	*906*
3D (n (%))	*503 (56)*
2D (n (%))	*403 (44)*
2D segmented lesions volumes (mL)	
2D minimal volume	*0.002*
2D median volume	*0.284*
2D maximal volume	*115.817*
3D segmented lesions volumes (mL)	
3D minimal volume	*0.007*
3D median volume	*1.608*
3D maximal volume	*530.73*

Table 1: Patients' baseline characteristics (n=71)

Characteristics	*Results*
Clinical follow up (mean, (min-max) days)	*882 (15-42299)*
Death at the end of follow up	
Yes (n (%))	*46 (65)*
No (n (%))	*25 (35)*
Median survival	
OS (median (min -max), days)	*502 (15-1356)*
< 12 months (n (%))	*27 (38)*
<6 months (n (%))	*14 (20)*
6-12 months (n (%))	*13 (18)*
>12 months (n (%))	*44 (62)*
12-18 months (n (%))	*17 (25)*
18-24 months (n (%))	*12 (16)*
>24 months (n (%))	*15 (22)*
PFS (median (min-max), days)	*166 (43-1330)*
< 12 months (n (%))	*50 (70)*
<6 months (n (%))	*34 (48)*
6-12 months (n (%))	*16 (22)*
>12 months (n (%))	*21 (30)*
12-18 months (n (%))	*13 (18)*
18-24 months (n (%))	*3 (4)*
>24 months (n (%))	*5 (8)*
iRECIST evaluation at 3 months	
Partial response (n (%))	*17 (24)*
Stable disease (n (%))	*16 (22)*
Progression (n (%))	*38 (54)*

Table 2: Follow-up characteristics of the patient (n=71) (PFS: Progression Free Survival)

ROI type	Combination (Classifier + Feature Selection +/- Smote Data Augmentation)	Acc	Sen	Spe
3D	Logistic Regression + Boruta	0.91	0.79	0.39
	Logistic Regression + Boruta + SMOTE	0.88	0.76	0.4
	Logistic Regression + Random Forest	0.88	0.77	0.36
3D + clinical data	SVM + SFS	0.84	0.95	0.6
	SVM + RF + SMOTE	0.78	0.4	0.85
	Logistic Regression + Recursive + SMOTE	0.7	0.86	0.6
2D	Logistic Regression + RF + SMOTE	0.81	0.6	0.75
	RF + Boruta	0.75	0.74	0.48
	Logistic Regression + Boruta + SMOTE	0.74	0.68	0.54
2D + clinical data	Logistic Regression + Recursive + SMOTE	0.86	0.7	0.69
	KNN + Relief + SMOTE	0.87	0.5	0.55
	SVM + Boruta	0.84	0.56	0.62

Table 3: Best OS predictions (Acc, Sen and Spe are mean values of the 5 fold CV, resulting in some cases in Accuracy outside the values of Spe and Sen)

ROI type	Combination (Classifier + Feature Selection +/- SMOTE Data Augmentation	Acc	Sen	Spe
3D	Logistic Regression + Boruta + SMOTE	0.8	0.66	0.88
	SVM + SFS + SMOTE	0.83	0.46	0.81
	Logistic Regression + SFS + SMOTE	0.81	0.46	0.66
3D + clinical data	KNN + Relief + SMOTE	0.75	0.72	0.6
	Logistic Regression + Recursive + SMOTE	0.83	0.65	0.6
	RF + RF + SMOTE	0.74	0.53	0.66
2D	Logistic Regression + Relief + SMOTE	0.81	0.66	0.85
	Logistic Regression + RF + SMOTE	0.81	0.6	0.75
	RF + Recursive + SMOTE	0.76	0.68	0.57
2D + clinical data	SVM + RF	0.87	0.44	0.8
	SVM + Relief + SMOTE	0.78	0.46	0.88
	RF + Recursive + SMOTE	0.76	0.68	0.57

Table 4: Best treatment response predictions (Acc, Sen and Spe are mean values of the 5 fold CV, resulting in some cases in Accuracy outside the values of Spe and Sen)

Histogram parameters	HR	CI 95%	*p* value
Skewness	1.34	1.07-1.7	0.012*
Kurtosis	0.99	0.92-1.1	0.684
Entropy_log10	1.48	0.34-6.3	0.6
Energy	4.16	0.22-78.4	0.342
Texture parameters	**HR**	**CI 95%**	***p* value**
GLRLM_SRE	1.5	1.1-2.1	0.022*
GLRLM_LRE	14	$1.3\ 10^{-3}$ -$1.6\ 10^{-5}$	0.578
GLRLM_HGRE	1.3	0.05-34	0.87
GLRLM_SRHGE	1	1-1	0.308
GLRLM_LRHGE	1	1-1	0.839
GLRLM_GLNU	1	1-1	0.992
GLRLM_RLNU	1	1-1	0.757
GLRLM_RP	1	1-1	0.418
NGLDM_Coarseness	0.99	0.54-1.8	0.974
NGLDM_Contrast	0.038	0.0019-0.76	0.032*
GLZLM_SZE	0.00075	$6.5\ 10^{-8}$-8.7	0.132
GLZLM_LZE	1	1-1	0.037*
GLZLM_HGZE	1	1-1	0.381
GLZLM_SZHGE	1	1-1	0.247
GLZLM_LZHGE	1	1-1	0.041*
GLZLM_GLNU	1	1-1	0.807
GLZLM_ZLNU	1	1-1	0.011*
GLZLM_ZP	1	1-1	0.181

Table 5: Univariate Cox proportional hazard models of histogram and texture parameters for survival analysis

* significant difference ($p<0.05$)

Supplementary Information 1

Supp Table 1.1: 46 radiomic features extracted from LIFEx

Radiomic features	Brief explanation
First order features: Shape	
Volume (mL)	Volume of ROI in mL
Volume (#vx)	Volume of ROI in number of voxels
Sphericity	Sphericity of the volume. 1 for a perfect sphere
Compacity	Compactness of ROI
First order features: Histogram	
Conventional_TLG (mL)	Total lesion glycolysis inside the ROI
Skewness	Asymmetry of the grey-level distribution in the histogram
Kurtosis	Shape of the grey-level distribution (peaked or flat) relative to a normal distribution
Entropy_log10	Randomness of the distribution
Entropy_log2	Randomness of the distribution
Energy	Uniformity of the distribution
minValue	Minimum pixel value of the ROI
meanValue	Average of pixel values
stdValue	Standard deviation of pixel values
maxValue	Maximum pixel value
Second order features	
GLCM	Arrangements of voxel pairs with same grey-level intensity
GLCM_Homogeneity	Homogeneity of grey-level voxel pairs
GLCM_Energy (Uniformity)	Uniformity of grey-level voxel pairs
GLCM_Contrast (Variance)	Local variations in GLCM
GLCM_Correlation	Linear dependency of grey-level voxel pairs
GLCM_Entropy_log10 and Entropy_log2	Randomness of grey-level voxel pairs
GLCM_Dissimilarity	Variation of grey-level voxel pairs
GLZLM	Size of homogeneous zones for each grey-level intensity
GLZLM_SZE	Short-zone emphasis: Distribution of the short homogeneous zones
GLZLM_LZE	Long-zone emphasis: Distribution of the long homogeneous zones
GLZLM_HGZE	High grey-level zone emphasis: Distribution of the high grey-level zones
GLZLM_LGZE	Low grey-level zone emphasis: Distribution of the low grey-level zones
GLZLM_SZHGE; GLZLM_SZLGE	Distribution of the short homogeneous zones with low or high grey-levels
GLZLM_LZHGE; GLZLM_LZLGE	Distribution of the long homogeneous zones with low or high grey-levels
GLZLM_GLNU	Grey-level non-uniformity: Nonuniformity of the grey-levels of the homogeneous zones

GLZLM_ZLNU	Zone length non-uniformity: Nonuniformity of the length of the homogeneous zones
GLZLM_ZP	Zone percentage: Homogeneity of the homogeneous zones
GLRLM	Size of homogeneous runs for each grey-level intensity
GLRLM_SRE	Short-run emphasis: Distribution of the short homogeneous runs
GLRLM_LRE	Long-run emphasis: Distribution of the long homogeneous runs
GLRLM_HGRE	High grey-level run emphasis: Distribution of the high grey-level runs
GLRLM_LGRE	Low grey-level run emphasis: Distribution of the low grel-level runs
GLRLM_SRHGE; GLRLM_SRLGE	Distribution of the short homogeneous runs with low or high grey-levels
GLRLM_LRHGE; GLRLM_LRLGE	Distribution of the long homogeneous runs with low or high grey-levels
GLRLM_GLNU	Grey-level non-uniformity: Nonuniformity of the grey-levels of the homogeneous runs
GLRLM_RP	Run percentage: Homogeneity of the homogeneous runs
GLRLM_RLNU	Run length non-uniformity: Nonuniformity of the length of the homogeneous runs
NGLDM	Difference of grey-level between one voxel and its 26 neighbours in 3 dimensions
NGLDM_Coarseness	Level of spatial rate of change in intensity
NGLDM_Contrast	Intensity difference between neighbouring regions
NGLDM_Busyness	Spatial frequency of changes in intensity

GLCM: Grey-level co-occurence matrix; NGLDM: Neighborhood grey-level different matrix; GLZLM: Grey-level zone length matrix; GLRLM: Grey-level run length matrix

Supplementary Information 2

Supp Table 2.1 : OS prediction with 2D features without clinical data (Acc, Sen and Spe are mean values of the 5 fold CV, resulting in some cases in Accuracy outside the values of Spe and Sen)

SFS= Sequential Forward Selection
KNN= K nearest Neighbour
RF= Random Forest
SVM= Support Vector Machine

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
2D	OS	KNN	Boruta	SMOTE	0.58	0.5	0.46
2D	OS	KNN	Boruta		0.58	0.63	0.29
2D	OS	KNN	Random Forest	SMOTE	0.69	0.54	0.61
2D	OS	KNN	Random Forest		0.68	0.79	0.29
2D	OS	KNN	Recursive	SMOTE	0.53	0.28	0.63
2D	OS	KNN	Recursive		0.63	0.62	0.4
2D	OS	KNN	Relief	SMOTE	0.63	0.52	0.51
2D	OS	KNN	Relief		0.64	0.74	0.27
2D	OS	KNN	SFS	SMOTE	0.7	0.62	0.54
2D	OS	KNN	SFS		0.64	0.75	0.26
2D	OS	Logistic Regression	Boruta	SMOTE	0.74	0.68	0.54
2D	OS	Logistic Regression	Boruta		0.64	0.95	0.018
2D	OS	Logistic Regression	Random Forest	SMOTE	0.81	0.6	0.75
2D	OS	Logistic Regression	Random Forest		0.7	0.98	0.09
2D	OS	Logistic Regression	Recursive	SMOTE	0.63	0.57	0.49
2D	OS	Logistic Regression	Recursive		0.62	0.84	0.11
2D	OS	Logistic Regression	Relief	SMOTE	0.63	0.21	0.9
2D	OS	Logistic Regression	Relief		0.65	0.98	0
2D	OS	Logistic Regression	SFS	SMOTE	0.54	0	1
2D	OS	Logistic Regression	SFS		0.63	0.21	0.9
2D	OS	RF	Boruta	SMOTE	0.71	0.65	0.52
2D	OS	RF	Boruta		0.75	0.74	0.48

2D	OS	RF	Random Forest	SMOTE	0.6	0.62	0.35
2D	OS	RF	Random Forest		0.68	0.81	0.26
2D	OS	RF	Recursive	SMOTE	0.76	0.68	0.57
2D	OS	RF	Recursive		0.71	0.75	0.39
2D	OS	RF	Relief	SMOTE	0.65	0.51	0.57
2D	OS	RF	Relief		0.75	0.84	0.35
2D	OS	RF	SFS	SMOTE	0.69	0.6	0.54
2D	OS	RF	SFS		0.69	0.8	0.26
2D	OS	SVM	Boruta	SMOTE	0.63	0.78	0.2
2D	OS	SVM	Boruta		0.62	0.92	0.02
2D	OS	SVM	Random Forest	SMOTE	0.77	0.59	0.44
2D	OS	SVM	Random Forest		0.66	1	0
2D	OS	SVM	Recursive	SMOTE	0.64	0.81	0.19
2D	OS	SVM	Recursive		0.61	0.909	0.01
2D	OS	SVM	Relief	SMOTE	0.63	0.21	0.9
2D	OS	SVM	Relief		0.66	1	0
2D	OS	SVM	SFS	SMOTE	0.66	1	0
2D	OS	SVM	SFS		0.61	0.9	0.03

Supp Table 2.2: OS prediction with 2D features with clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
2D	OS	KNN	Boruta	SMOTE	0.59	0.62	0.65
2D	OS	KNN	Boruta		0.75	0.7	0.5
2D	OS	KNN	Random Forest	SMOTE	0.45	0.4	0.44
2D	OS	KNN	Random Forest		0.49	0.41	0.24
2D	OS	KNN	Recursive	SMOTE	0.59	0.6	0.45
2D	OS	KNN	Recursive		0.72	0.6	0.7
2D	OS	KNN	Relief	SMOTE	0.87	0.5	0.55
2D	OS	KNN	Relief		0.75	0.86	0.3
2D	OS	KNN	SFS	SMOTE	0.61	0.4	0.6
2D	OS	KNN	SFS		0.42	0.85	0.12
2D	OS	Logistic Regression	Boruta	SMOTE	0.76	0.83	0.5
2D	OS	Logistic Regression	Boruta		0.8	0.89	0.1
2D	OS	Logistic Regression	Random Forest	SMOTE	0.68	0.76	0.72

2D	OS	Logistic Regression	Random Forest		0.7	0.5	0.5
2D	OS	Logistic Regression	Recursive	SMOTE	0.75	0.1	0.97
2D	OS	Logistic Regression	Recursive		0.61	1	0
2D	OS	Logistic Regression	Relief	SMOTE	0.59	0.59	0.59
2D	OS	Logistic Regression	Relief		0.76	0.75	0.2
2D	OS	Logistic Regression	SFS	SMOTE	0.8	0.7	0.6
2D	OS	Logistic Regression	SFS		0.7	0.6	0.4
2D	OS	RF	Boruta	SMOTE	0.56	0.79	0.57
2D	OS	RF	Boruta		0.38	0.45	0.13
2D	OS	RF	Random Forest	SMOTE	0.44	0.53	0.28
2D	OS	RF	Random Forest		0.3	0.8	0.03
2D	OS	RF	Recursive	SMOTE	0.47	0.43	0.33
2D	OS	RF	Recursive		0.57	0.7	0.4
2D	OS	RF	Relief	SMOTE	0.4	0.18	0.83
2D	OS	RF	Relief		0.42	1	0.22
2D	OS	RF	SFS	SMOTE	0.46	0.8	0.33
2D	OS	RF	SFS		0.42	0.66	0.2
2D	OS	SVM	Boruta	SMOTE	0.54	0.79	0.12
2D	OS	SVM	Boruta		0.84	0.56	0.62
2D	OS	SVM	Random Forest	SMOTE	0.71	0.23	0.92
2D	OS	SVM	Random Forest		0.9	1	0
2D	OS	SVM	Recursive	SMOTE	0.77	0.87	0.2
2D	OS	SVM	Recursive		0.66	0.75	0.25
2D	OS	SVM	Relief	SMOTE	0.63	0.5	0.5
2D	OS	SVM	Relief		0.75	0.57	0.4
2D	OS	SVM	SFS	SMOTE	0.48	0.25	0.77
2D	OS	SVM	SFS		0.77	0.83	0.7

Supp Table 2.3: OS prediction with 3D features without clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
3D	OS	KNN	Boruta	SMOTE	0.62	0.43	0.6
3D	OS	KNN	Boruta		0.65	0.46	0.6
3D	OS	KNN	Random Forest	SMOTE	0.71	0.52	0.6

3D	OS	KNN	Random Forest		0.65	0.44	0.66
3D	OS	KNN	Recursive	SMOTE	0.68	0.47	0.66
3D	OS	KNN	Recursive		0.63	0.46	0.54
3D	OS	KNN	Relief	SMOTE	0.7	0.56	0.45
3D	OS	KNN	Relief		0.64	0.54	0.33
3D	OS	KNN	Relief		0.64	0.54	0.33
3D	OS	KNN	SFS	SMOTE	0.4	0.5	0.5
3D	OS	KNN	SFS		0.56	0.34	0.7
3D	OS	Logistic Regression	Boruta	SMOTE	0.88	0.76	0.4
3D	OS	Logistic Regression	Boruta		0.91	0.79	0.39
3D	OS	Logistic Regression	Random Forest	SMOTE	0.72	0.47	0.75
3D	OS	Logistic Regression	Random Forest		0.88	0.77	0.36
3D	OS	Logistic Regression	Recursive	SMOTE	0.51	0.81	0.5
3D	OS	Logistic Regression	Recursive		0.95	0.87	0.3
3D	OS	Logistic Regression	Relief	SMOTE	0.75	0.8	0.16
3D	OS	Logistic Regression	Relief		0.86	0.98	0
3D	OS	Logistic Regression	SFS	SMOTE	0.43	0.38	0.69
3D	OS	Logistic Regression	SFS		0.94	1	0
3D	OS	RF	Boruta	SMOTE	0.64	0.44	0.63
3D	OS	RF	Boruta		0.73	0.86	0.32
3D	OS	RF	Random Forest	SMOTE	0.74	0.53	0.66
3D	OS	RF	Random Forest		0.7	0.75	0.26
3D	OS	RF	Recursive	SMOTE	0.56	0.28	0.55
3D	OS	RF	Recursive		0.44	0.75	0.36
3D	OS	RF	Relief	SMOTE	0.54	0.59	0.44
3D	OS	RF	Relief		0.62	0.75	0.46
3D	OS	RF	SFS	SMOTE	0.66	0.42	0.57
3D	OS	RF	SFS		0.6	0.62	0.44
3D	OS	SVM	Boruta	SMOTE	0.35	0.03	0.96
3D	OS	SVM	Boruta		0.73	0.98	0
3D	OS	SVM	Random Forest	SMOTE	0.5	0.2	0.96
3D	OS	SVM	Random Forest		0.71	0.98	0.02
3D	OS	SVM	Recursive	SMOTE	0.35	0.03	0.96

3D	OS	SVM	Recursive		0.97	0.9	0.24
3D	OS	SVM	Relief	SMOTE	0.89	0.11	0.9
3D	OS	SVM	Relief		0.98	1	0
3D	OS	SVM	SFS	SMOTE	0.98	0.95	0.12
3D	OS	SVM	SFS		0.98	1	0

Supp Table 2.4: OS prediction with 3D features with clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
3D	OS	KNN	Boruta	SMOTE	0.68	0.55	0.55
3D	OS	KNN	Boruta		0.7	0.79	0.41
3D	OS	KNN	Random Forest	SMOTE	0.82	0.64	0.4
3D	OS	KNN	Random Forest		0.54	0.84	0.25
3D	OS	KNN	Recursive	SMOTE	0.61	0.6	0.6
3D	OS	KNN	Recursive		0.8	0.79	0.3
3D	OS	KNN	Relief	SMOTE	0.72	0.62	0.57
3D	OS	KNN	Relief		0.76	1	0.1
3D	OS	KNN	Relief		0.61	0.58	0.48
3D	OS	KNN	SFS	SMOTE	0.57	0.74	0.46
3D	OS	KNN	SFS		0.64	0.8	0.4
3D	OS	Logistic Regression	Boruta	SMOTE	0.73	0.16	0.88
3D	OS	Logistic Regression	Boruta		0.71	1	0.11
3D	OS	Logistic Regression	Random Forest	SMOTE	0.77	1	0.03
3D	OS	Logistic Regression	Random Forest		0.83	0.56	0.62
3D	OS	Logistic Regression	Recursive	SMOTE	0.7	0.86	0.6
3D	OS	Logistic Regression	Recursive		0.71	1	0.11
3D	OS	Logistic Regression	Relief	SMOTE	0.63	0.89	0.05
3D	OS	Logistic Regression	Relief		0.88	1	0
3D	OS	Logistic Regression	SFS	SMOTE	0.69	0.91	0.66
3D	OS	Logistic Regression	SFS		0.7	1	0
3D	OS	RF	Boruta	SMOTE	0.56	0.79	0.57
3D	OS	RF	Boruta		0.38	0.45	0.13
3D	OS	RF	Random Forest	SMOTE	0.44	0.53	0.28

3D	OS	RF	Random Forest		0.3	0.8	0.03
3D	OS	RF	Recursive	SMOTE	0.47	0.43	0.33
3D	OS	RF	Recursive		0.57	0.7	0.4
3D	OS	RF	Relief	SMOTE	0.4	0.18	0.83
3D	OS	RF	Relief		0.42	1	0.22
3D	OS	RF	SFS	SMOTE	0.46	0.8	0.33
3D	OS	RF	SFS		0.42	0.66	0.2
3D	OS	SVM	Boruta	SMOTE	0.65	0.51	0.57
3D	OS	SVM	Boruta		0.9	0.6	0.5
3D	OS	SVM	Random Forest	SMOTE	0.78	0.4	0.85
3D	OS	SVM	Random Forest		0.78	0.97	0.15
3D	OS	SVM	Recursive	SMOTE	0.79	0.95	0.01
3D	OS	SVM	Recursive		0.7	0.2	1
3D	OS	SVM	Relief	SMOTE	0.7	0.2	0.8
3D	OS	SVM	Relief		0.76	1	0
3D	OS	SVM	SFS	SMOTE	0.37	0.03	0.97
3D	OS	SVM	SFS		0.84	0.95	0.6

Supp Table 2.5: Therapy response prediction with 2D features without clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
2D	Response	KNN	Boruta	SMOTE	0.58	0.42	0.63
2D	Response	KNN	Boruta		0.85	0.56	0.62
2D	Response	KNN	Random Forest	SMOTE	0.61	0.51	0.45
2D	Response	KNN	Random Forest		0.51	0.5	0.47
2D	Response	KNN	Recursive	SMOTE	0.62	0.47	0.53
2D	Response	KNN	Recursive		0.71	0.2	0.9
2D	Response	KNN	Relief	SMOTE	0.68	0.48	0.77
2D	Response	KNN	Relief		0.65	0.73	0.42
2D	Response	KNN	SFS	SMOTE	0.62	0.47	0.53
2D	Response	Logistic Regression	Boruta	SMOTE	0.8	0.2	0.9
2D	Response	Logistic Regression	Boruta		0.74	0.17	0.75
2D	Response	Logistic Regression	Random Forest	SMOTE	0.8	0.23	0.88
2D	Response	Logistic Regression	Random Forest		0.73	0.09	0.89
2D	Response	Logistic Regression	Recursive	SMOTE	0.86	0.8	0.42

2D	Response	Logistic Regression	Recursive		0.64	0.45	0.77
2D	Response	Logistic Regression	Relief	SMOTE	0.81	0.66	0.85
2D	Response	Logistic Regression	Relief		0.7	0.3	0.83
2D	Response	Logistic Regression	SFS	SMOTE	0.72	0.2	0.7
2D	Response	Logistic Regression	SFS		0.7	0.57	0.73
2D	Response	RF	Boruta	SMOTE	0.65	0.43	0.59
2D	Response	RF	Boruta		0.44	0.142	0.83
2D	Response	RF	Random Forest	SMOTE	0.59	0.25	0.8
2D	Response	RF	Random Forest		0.34	0.48	0.54
2D	Response	RF	Recursive	SMOTE	0.91	0.89	0.24
2D	Response	RF	Recursive		0.42	0.29	0.5
2D	Response	RF	Relief	SMOTE	0.52	0.37	0.61
2D	Response	RF	Relief		0.4	0.32	0.74
2D	Response	RF	SFS	SMOTE	0.62	0.48	0.57
2D	Response	RF	SFS		0.38	0.125	0.9
2D	Response	SVM	Boruta	SMOTE	0.74	0.54	0.63
2D	Response	SVM	Boruta		0.64	0.1	1
2D	Response	SVM	Random Forest	SMOTE	0.6	0.33	0.79
2D	Response	SVM	Random Forest		0.65	0.1	0.9
2D	Response	SVM	Recursive	SMOTE	0.66	0.25	0.82
2D	Response	SVM	Recursive		0.82	0	1
2D	Response	SVM	Relief	SMOTE	0.77	0.4	0.88
2D	Response	SVM	Relief		0.69	0	1
2D	Response	SVM	SFS	SMOTE	0.65	0.44	0.72
2D	Response	SVM	SFS		0.42	0	1

Supp Table 2.6: Therapy response prediction with 2D features with clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
2D	Response	KNN	Boruta	SMOTE	0.54	0.47	0.48
2D	Response	KNN	Boruta		0.59	0.64	0.5
2D	Response	KNN	Random Forest	SMOTE	0.57	0.407	0.57
2D	Response	KNN	Random Forest		0.65	1	0
2D	Response	KNN	Recursive	SMOTE	0.83	0.45	0.62
2D	Response	KNN	Recursive		0.59	0.01	0.85

2D	Response	KNN	Relief	SMOTE	0.5	0.61	0.46
2D	Response	KNN	Relief		0.5	0.5	0.55
2D	Response	KNN	SFS	SMOTE	0.55	0.42	0.48
2D	Response	Logistic Regression	Boruta	SMOTE	0.59	0.25	0.76
2D	Response	Logistic Regression	Boruta		0.71	0.3	0.8
2D	Response	Logistic Regression	Random Forest	SMOTE	0.7	1	0
2D	Response	Logistic Regression	Random Forest		0.58	0.25	0.81
2D	Response	Logistic Regression	Recursive	SMOTE	0.8	0.1	0.85
2D	Response	Logistic Regression	Recursive		0.75	0.25	0.82
2D	Response	Logistic Regression	Relief	SMOTE	0.71	0.26	0.81
2D	Response	Logistic Regression	Relief		0.72	0.3	0.82
2D	Response	Logistic Regression	SFS	SMOTE	0.77	0.3	0.9
2D	Response	Logistic Regression	SFS		0.8	0.37	0.84
2D	Response	RF	Boruta	SMOTE	0.71	0.65	0.52
2D	Response	RF	Boruta		0.75	0.74	0.48
2D	Response	RF	Random Forest	SMOTE	0.7	0.8	0.5
2D	Response	RF	Random Forest		0.68	0.81	0.26
2D	Response	RF	Recursive	SMOTE	0.76	0.68	0.57
2D	Response	RF	Recursive		0.71	0.75	0.39
2D	Response	RF	Relief	SMOTE	0.65	0.51	0.57
2D	Response	RF	Relief		0.75	0.84	0.35
2D	Response	RF	SFS	SMOTE	0.69	0.6	0.54
2D	Response	RF	SFS		0.69	0.8	0.26
2D	Response	SVM	Boruta	SMOTE	0.64	0.37	0.85
2D	Response	SVM	Boruta		0.55	0.01	0.88
2D	Response	SVM	Random Forest	SMOTE	0.88	0.27	0.92
2D	Response	SVM	Random Forest		0.87	0.44	0.8
2D	Response	SVM	Recursive	SMOTE	0.51	0.3	0.78
2D	Response	SVM	Recursive		0.54	0.01	0.85
2D	Response	SVM	Relief	SMOTE	0.78	0.46	0.88
2D	Response	SVM	Relief		0.47	1	0
2D	Response	SVM	SFS	SMOTE	0.75	0	1
2D	Response	SVM	SFS		0.57	0.2	0.9

Supp Table 2.7: Therapy response prediction with 3D features without clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
3D	Response	KNN	Boruta	SMOTE	0.8	0.45	0.72
3D	Response	KNN	Boruta		0.6	0.375	0.57
3D	Response	KNN	Random Forest	SMOTE	0.57	0.52	0.57
3D	Response	KNN	Random Forest		0.75	0.2	0.83
3D	Response	KNN	Recursive	SMOTE	0.62	0.57	0.75
3D	Response	KNN	Recursive		0.7	0.52	0.75
3D	Response	KNN	Relief	SMOTE	0.68	0.45	0.72
3D	Response	KNN	Relief		0.73	0.38	0.7
3D	Response	KNN	SFS	SMOTE	0.78	0.9	0.5
3D	Response	KNN	SFS		0.57	0.48	0.54
3D	Response	Logistic Regression	Boruta	SMOTE	0.8	0.66	0.88
3D	Response	Logistic Regression	Boruta		0.76	0.4	0.83
3D	Response	Logistic Regression	Random Forest	SMOTE	0.69	0.68	0.65
3D	Response	Logistic Regression	Random Forest		0.88	0.2	0.8
3D	Response	Logistic Regression	Recursive	SMOTE	0.55	0.52	0.66
3D	Response	Logistic Regression	Recursive		0.61	0	1
3D	Response	Logistic Regression	Relief	SMOTE	0.76	0.67	0.32
3D	Response	Logistic Regression	Relief		0.74	0.5	0.57
3D	Response	Logistic Regression	SFS	SMOTE	0.81	0.46	0.66
3D	Response	Logistic Regression	SFS		0.69	0.24	0.92
3D	Response	RF	Boruta	SMOTE	0.71	0.28	0.77
3D	Response	RF	Boruta		0.69	0.31	0.72
3D	Response	RF	Random Forest	SMOTE	0.47	0.275	0.705
3D	Response	RF	Random Forest		0.6	0.3	0.58
3D	Response	RF	Recursive	SMOTE	0.31	0.51	0.54
3D	Response	RF	Recursive		0.42	0.35	0.48
3D	Response	RF	Relief	SMOTE	0.51	0.6	0.66
3D	Response	RF	Relief		0.94	0.14	0.92
3D	Response	RF	SFS	SMOTE	0.71	0.38	0.62
3D	Response	RF	SFS		0.79	0.66	0.52
3D	Response	SVM	Boruta	SMOTE	0.72	0.19	0.92

3D	Response	SVM	Boruta		0.62	0	1
3D	Response	SVM	Random Forest	SMOTE	0.66	0.78	0.2
3D	Response	SVM	Random Forest		0.75	0	1
3D	Response	SVM	Recursive	SMOTE	0.74	0	0.92
3D	Response	SVM	Recursive		0.52	0	1
3D	Response	SVM	Relief	SMOTE	0.8	0.1	0.82
3D	Response	SVM	Relief		0.85	0	1
3D	Response	SVM	SFS	SMOTE	0.83	0.46	0.81
3D	Response	SVM	SFS		0.7	0.12	0.95

Supp Table 2.8: Therapy response prediction with 3D features with clinical data

ROI	AIM	Classifier	Selection	SMOTE ?	Acc	Sen	Spe
3D	Response	KNN	Boruta	SMOTE	0.52	0.33	0.51
3D	Response	KNN	Boruta		0.51	0.48	0.4
3D	Response	KNN	Random Forest	SMOTE	0.66	0.32	0.7
3D	Response	KNN	Random Forest		0.78	0.2	0.8
3D	Response	KNN	Recursive	SMOTE	0.62	0.47	0.48
3D	Response	KNN	Recursive		0.63	0.2	0.69
3D	Response	KNN	Relief	SMOTE	0.75	0.72	0.6
3D	Response	KNN	Relief		0.78	0.12	0.94
3D	Response	KNN	SFS	SMOTE	0.76	0.41	0.7
3D	Response	KNN	SFS		0.6	0.2	0.7
3D	Response	Logistic Regression	Boruta	SMOTE	0.75	0.48	0.65
3D	Response	Logistic Regression	Boruta		0.9	0.3	0.9
3D	Response	Logistic Regression	Random Forest	SMOTE	0.58	0.44	0.52
3D	Response	Logistic Regression	Random Forest		0.53	0.1	0.95
3D	Response	Logistic Regression	Recursive	SMOTE	0.83	0.65	0.6
3D	Response	Logistic Regression	Recursive		0.78	0.3	0.8
3D	Response	Logistic Regression	Relief	SMOTE	0.7	0.77	0.3
3D	Response	Logistic Regression	Relief		0.79	0.12	0.86
3D	Response	Logistic Regression	SFS	SMOTE	0.87	0.33	0.74
3D	Response	Logistic Regression	SFS		0.68	0.01	0.94
3D	Response	RF	Boruta	SMOTE	0.64	0.44	0.63

3D	Response	RF	Boruta		0.73	0.86	0.32
3D	Response	RF	Random Forest	SMOTE	0.74	0.53	0.66
3D	Response	RF	Random Forest		0.7	0.75	0.26
3D	Response	RF	Recursive	SMOTE	0.74	0.62	0.43
3D	Response	RF	Recursive		0.44	0.75	0.36
3D	Response	RF	Relief	SMOTE	0.54	0.59	0.44
3D	Response	RF	Relief		0.62	0.75	0.46
3D	Response	RF	SFS	SMOTE	0.66	0.42	0.57
3D	Response	RF	SFS		0.6	0.62	0.44
3D	Response	SVM	Boruta	SMOTE	0.57	0.25	0.67
3D	Response	SVM	Boruta		0.8	0	1
3D	Response	SVM	Random Forest	SMOTE	0.74	0.2	0.9
3D	Response	SVM	Random Forest		0.7	0	1
3D	Response	SVM	Recursive	SMOTE	0.53	0.98	0.1
3D	Response	SVM	Recursive		0.69	0.01	0.81
3D	Response	SVM	Relief	SMOTE	0.74	0.59	0.35
3D	Response	SVM	Relief		0.71	0	1
3D	Response	SVM	SFS	SMOTE	0.89	0.01	0.9
3D	Response	SVM	SFS		0.88	0.3	0.9

Supplementary Information 3

Table 3.1: Coefficient of variation ($\frac{\sigma}{|\underline{x}|}$). of radiomics features

Radiomics Feature	2D coefficient of variation	3D coefficient of variation
minValue	21.90	3.75
meanValue	0.94	1.08
stdValue	0.88	0.78
maxValue	0.51	0.54
HISTO_Skewness	60.41	4.43
HISTO_Kurtosis	0.28	0.81
HISTO_Entropy_log10	0.24	0.23
HISTO_Entropy_log2	0.24	0.23
HISTO_Energy	0.62	0.61
GLCM_Homogeneity	0.55	0.38
GLCM_Energy	1.64	1.63
GLCM_Contrast	3.58	2.80
GLCM_Correlation	0.70	0.36
GLCM_Entropy_log10	0.55	0.61
GLCM_Entropy_log2	0.55	0.36
GLCM_Dissimilarity	0.90	0.85
GLRLM_SRE	0.52	0.31
GLRLM_LRE	1.08	0.63
GLRLM_LGRE	2.81	0.40
GLRLM_HGRE	3.22	0.30
GLRLM_SRLGE	0.50	0.43
GLRLM_SRHGE	0.52	0.31
GLRLM_LRLGE	1.70	0.63
GLRLM_LRHGE	1.08	0.63
GLRLM_GLNU	3.73	4.45
GLRLM_RLNU	4.04	4.26
GLRLM_RP	0.53	0.31
NGLDM_Coarseness	1.53	1.33
NGLDM_Contrast	1.15	1.78
NGLDM_Busyness	2.59	1.60
GLZLM_SZE	0.55	0.36
GLZLM_LZE	4.90	4.43
GLZLM_LGZE	2.87	0.40
GLZLM_HGZE	0.50	0.30
GLZLM_SZLGE	4.34	0.53
GLZLM_SZHGE	0.56	0.36
GLZLM_LZLGE	4.86	4.48
GLZLM_LZHGE	4.94	4.40
GLZLM_GLNU	1.97	4.47
GLZLM_ZLNU	2.42	4.42
GLZLM_ZP	0.79	0.87

Median	**1.076**	**0.634**

Table 3.2: Mean and standard value (SD) of radiomic features

	2D		3D	
Radiomics Feature	**Mean**	**SD**	**Mean**	**SD**
minValue	3.9309	86.1002	-29.024	108.8098
meanValue	62.50549	58.5625	57.4985	61.9848
stdValue	22.05224	19.4946	26.7891	20.8367
maxValue	120.1881	61.2225	131.7911	70.8837
HISTO_Skewness	-0.00776	0.4688	-0.1296	0.5747
HISTO_Kurtosis	3.04419	0.8395	3.4344	2.7804
HISTO_Entropy_log10	0.85633	0.2095	0.925	0.2088
HISTO_Entropy_log2	2.84465	0.696	3.0727	0.6937
HISTO_Energy	0.18295	0.1131	0.158	0.0956
GLCM_Homogeneity	0.42394	0.234	0.4585	0.1733
GLCM_Energy	0.03809	0.0625	0.04	0.065
GLCM_Contrast	5.60013	20.0311	10.6954	29.8961
GLCM_Entropy_log10	0.37249	0.2591	1.5538	0.5623
GLCM_Correlation	1.29573	0.7082	0.35	0.2148
GLCM_Entropy_log2	4.30432	2.3527	5.1618	1.868
GLCM_Dissimilarity	1.32653	1.1922	1.8596	1.5728
GLRLM_SRE	0.6668	0.3454	0.8023	0.245
GLRLM_LRE	1.93866	2.0868	1.7452	1.0997
GLRLM_LGRE	7.94E-05	2.23E-04	8.19E-05	3.24E-05
GLRLM_HGRE	6.74E-05	2.17E-04	10594.4566	3182.0921
GLRLM_SRLGE	9262.53537	4647.2292	7.15E-05	3.11E-05
GLRLM_SRHGE	7697.86466	4025.1031	9209.5776	2898.8417
GLRLM_LRLGE	1.79E-04	3.04E-04	1.54E-04	9.76E-05
GLRLM_LRHGE	22384.51334	24141.3585	20081.1129	12740.7599
GLRLM_GLNU	104.41274	389.0792	566.0545	2521.7686
GLRLM_RLNU	562.60987	2270.5234	3169.6877	13504.7678
GLRLM_RP	0.63276	0.3328	0.7705	0.2405
NGLDM_Coarseness	0.06395	0.0976	0.016	0.0213
NGLDM_Contrast	0.02827	0.0325	0.0866	0.1543

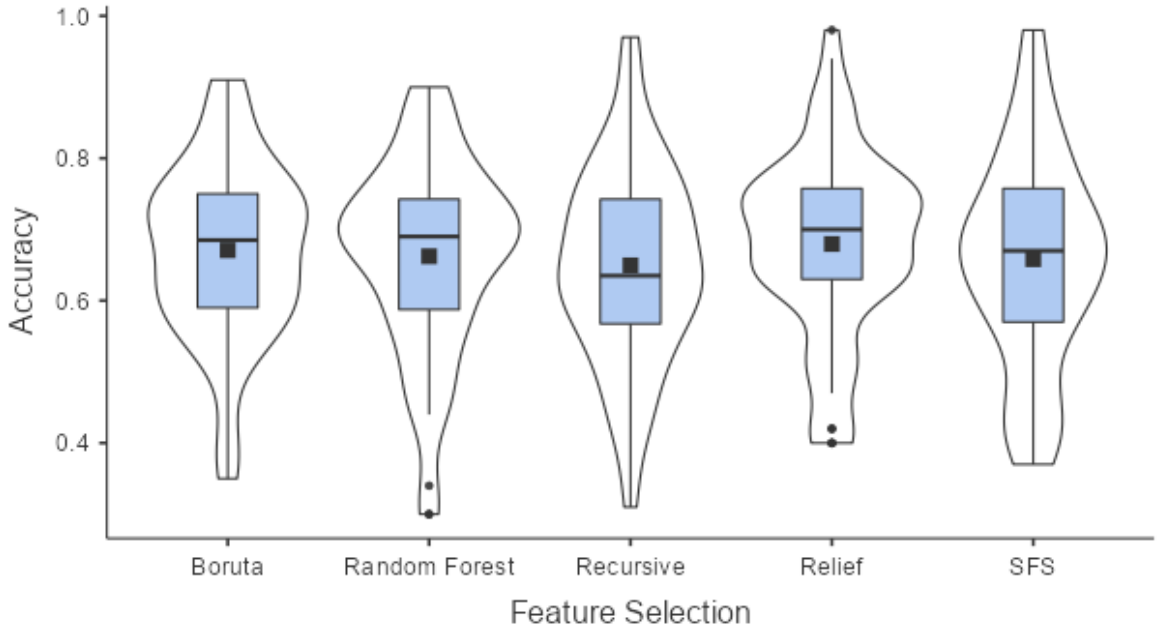
NGLDM_Busyness	1.15815	2.9946	2.6829	4.2944
GLZLM_SZE	0.45316	0.2508	0.5249	0.1898
GLZLM_LZE	1355.38957	6637.224	8078.6047	35828.3955
GLZLM_LGZE	7.97E-05	2.29E-04	8.23E-05	3.30E-05
GLZLM_HGZE	9270.86164	4654.5501	10575.6073	3193.705
GLZLM_SZLGE	4.86E-05	2.11E-04	4.75E-05	2.54E-05
GLZLM_SZHGE	5232.65207	2919.8812	5979.3646	2156.8144
GLZLM_LZLGE	0.11794	0.5731	0.6992	3.1326
GLZLM_LZHGE	1.56E+07	7.70E+07	9.35E+07	4.11E+08
GLZLM_GLNU	17.3782	34.1841	46.0707	206.1363
GLZLM_ZLNU	61.54293	149.164	199.1607	879.8972
GLZLM_ZP	0.25801	0.2046	0.1627	0.1412

Supplementary Information 4

Supp Table 4.1:
Descriptive statistics of performances by feature selection method.

Feature selection method	Acc		Sen		Spe	
	Mean	SD	Mean	SD	Mean	SD
Boruta	0.671	0.123	0.514	0.262	0.563	0.260
RF	0.662	0.136	0.530	0.280	0.524	0.305
Recursive	0.650	0.138	0.515	0.302	0.544	0.273
Relief	0.680	0.130	0.554	0.293	0.521	0.302
SFS	0.658	0.149	0.517	0.295	0.570	0.285

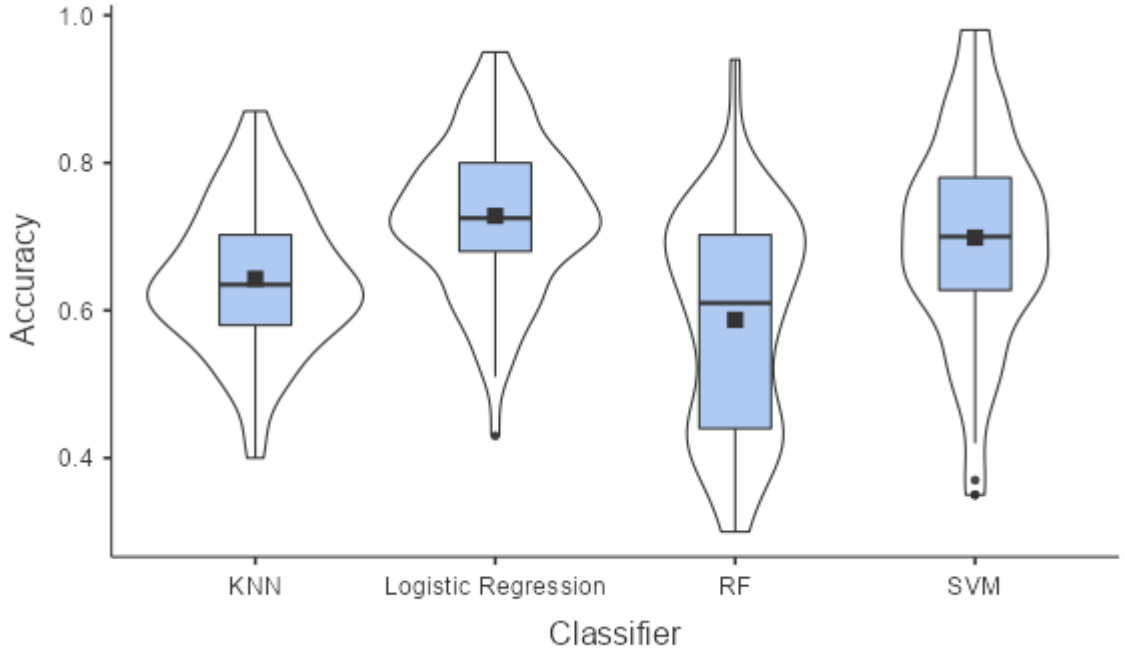
Supp Figure 4.1 : Violin plot of performances by feature selection method



Supp Table 4.2: Descriptive statistic of performances by classifier

Classifier	Acc		Sen		Spe	
	Mean	SD	Mean	SD	Mean	SD
KNN	0.643	0.098	0.529	0.191	0.529	0.178
LR	0.728	0.103	0.545	0.309	0.571	0.317
RF	0.587	0.146	0.585	0.214	0.468	0.195
SVM	0.699	0.142	0.446	0.375	0.609	0.381

Supp Figure 4.2 : Violin plot of performances by classifier



Supp Table 4.3 : Pair wise comparison of accuracy by classifier (Wilcoxon test, Bonferroni's correction for multiple comparison, *=significant difference)

	KNN		
LR	2.86E-06*	LR	
RF	0.218	6.54E-09*	RF
SVM	0.01*	1	5.35E-05*

Supplementary Information 5

Supp Table 5.1: Test and train accuracies of the 12 best models in our study for OS prediction (mean over the 5-fold cross validation)

ROI type	Combination (Classifier + Feature Selection +/- Smote Data Augmentation)	Train Acc	Train error	Test Acc	Train/test Acc ratio
3D	Logistic Regression + Boruta	0.8	0.2	0.91	0.88
	Logistic Regression + Boruta + SMOTE	0.9	0.1	0.88	1.02
	Logistic Regression + Random Forest	0.75	0.25	0.88	0.85
3D + clinical data	SVM + SFS	0.9	0.1	0.84	1.07
	SVM + RF + SMOTE	0.8	0.2	0.78	1.03
	Logistic Regression + Recursive + SMOTE	0.75	0.25	0.7	1.07
2D	Logistic Regression + RF + SMOTE	0.85	0.15	0.81	1.05
	RF + Boruta	0.8	0.2	0.75	1.07
	Logistic Regression + Boruta + SMOTE	0.8	0.2	0.74	1.08
2D + clinical data	Logistic Regression + Recursive + SMOTE	0.9	0.1	0.86	1.05
	KNN + Relief + SMOTE	0.77	0.23	0.87	0.89
	SVM + Boruta	0.85	0.15	0.84	1.01
			Mean Train Error 0.18		**Mean Ratio 1.01**

Supp Table 5.2: Test and train accuracies of the 12 best models in our study for response prediction (mean over the 5-fold cross validation)

ROI type	Combination (Classifier + Feature Selection +/- SMOTE Data Augmentation	Train Acc	Train Error	Test Acc	Train/test Acc ratio
3D	Logistic Regression + Boruta + SMOTE	0.88	0.12	0.8	1.1
	SVM + SFS + SMOTE	0.8	0.2	0.83	0.96
	Logistic Regression + SFS + SMOTE	0.85	0.15	0.81	1.05
3D + clinical data	KNN + Relief + SMOTE	0.83	0.17	0.75	1.11
	Logistic Regression + Recursive + SMOTE	0.9	0.1	0.83	1.08
	RF + RF + SMOTE	0.8	0.2	0.74	1.08
2D	Logistic Regression + Relief + SMOTE	0.9	0.1	0.81	1.11
	Logistic Regression + RF + SMOTE	0.8	0.2	0.6	1.33
	RF + Recursive + SMOTE	0.75	0.25	0.76	0.99
2D + clinical data	SVM + RF	0.9	0.1	0.87	1.03
	SVM + Relief + SMOTE	0.7	0.3	0.78	0.9
	RF + Recursive + SMOTE	0.7	0.3	0.76	0.92
			Mean Train Error 0.18		**Mean ratio 1.06**

Supp Table 5.3: Mean train/test accuracy ratio on all classifications by each classifier for OS prediction

Classifier	Mean train/test Acc ratio
KNN	1.04
LR	1.01
RF	1.07
SVM	0.96

Supp Table 5.4: Mean train/test accuracy ratio on all classification by each classifier for response prediction

Classifier	Mean train/test Acc ratio
KNN	1.02
LR	1.01
RF	1.05
SVM	0.95