**Title:**

A Time-Dependent Radiomic Artificial Intelligence Model to Predict Survival of Patients with Liver Metastases

**Authors:**

**Institutions:**

School of Medicine, Queen’s University

Memorial Sloan Kettering Cancer Center  
School of Computing Queen’s University

Department of Biomedical ad Molecular Sciences, Queen’s University

**Corresponding Author:**

Ricky Hu, rhu@qmed.ca

**Funding Statement:**

**Conflict of Interest Statement:**

**Author Contributions:**

Ricky Hu: Methodology, software development, data analysis, visualization, and writing.

Ishita Chen: Conceptualization, data acquisition, data processing, data analysis, and writing.

Jacob Peoples: Methodology, software development, data analysis, and writing.

Jean-Paul Salemeh: Data analysis and writing

Amber Simpson: Supervision, conceptualization, methodology, data analysis, and writing.

Marsha Reyngold: Supervision, data acquisition, data processing, data analysis, and writing.  
Paul Romesser: Supervision, data acquisition, data processing, data analysis, and writing.

**Acknowledgements:**

**Abstract:**

**Background:**

Prognostic assessment of colorectal liver metastases is essential for guiding patient management. Existing assessment criteria involve scoring from clinicopathological variables with limitations on accuracy and generalizability. Computed tomography (CT) contains detailed liver texture information which may be analyzed by artificial intelligence (AI) to provide quantitative prediction of survival capable of modeling nonlinear variable dependencies.

**Methods**

A time-dependent AI survival model was built by first extracting 108 radiomic features from patient CT scans and predicting survival with the random survival forest algorithm (RSF). Accuracies were measured by the average concordance index (C-index) and integrated Brier score (IBS). This was repeated with different segmentations of the liver scan and with standard radiotherapy clinical variables as input to the RSF. Feature importance scores were computed by comparing perturbation error rates to identify predictive features.

**Results**

The AI radiomics model achieved a C-index of 0.76 ± 0.06, greater than previous studies. IBS for all variations of the models were below the 0.25 threshold for random prediction, indicating that a viable signal exists within radiomic analysis. The most consistently predictive feature was kurtosis (feature importance: 0.383 ± 0.035). Utilization of dosimetric variables achieved good accuracies (C-index: 0.67 ± 0.08), with max dose observed as the most predictive variables (feature importance: 3.113 ± 0.138).

**Conclusions**

The AI radiomics model achieves high survival prediction accuracy, providing support that radiomic analysis of CT scans combined with complex AI analysis may aid in clinical decision making that requires prognostic prediction.

**Main Text:**

**Introduction:**

**Background:**

Patients with colorectal cancer develop colorectal liver metastases (CLM) in approximately 50% of cases [1], with a 12-month recurrence rate of approximately 40%, with variability in resistance to treatment options such as radiotherapy [2]. Accurate prognostic prediction of patients with CLM can assist in guiding personalized treatment aiming to maximize quality of life.

**Existing Work:**

There are several existing methods that aim to predict survival of patients with liver metastases. A common method is to select clinically relevant variables, perform a multivariate Cox proportional hazard regression, and select the variables with the highest hazard ratio to create a scoring system. We identified 10 studies which implement this method [1-10]. Taking one of the more recent of the studies as an example, Brudvik et al. [10] utilized 6 clinicopathological characteristics, (node-positive tumor, disease-free interval, multiple metastases, metastases greater than 50mm, carcinoebryonic antigen level, and RAS mutation status) in a Cox hazard model, then selected three variables with a p-value of under 0.05 as part of a 3-point system (node-positive tumor, metastases greater than 50mm, and RAS mutation). The scoring system achieved a concordance index (C-Index) of 0.66 (95% CI: 0.56-0.75) for predicting recurrence free survival when tested on their own dataset.

There are several opportunities to improve on existing methods. First, the Cox model involves an exponential hazard function in the following form:

Where *h0(t)* is the baseline hazard function when covariates are zero, *βi* is the i-th parameter optimized by the partial likelihood function of variable *xi* and *h(t)* is the hazard function at time *t* [11]. The limitation is that variables in the model are parameterized in a linear combination, and the Cox model may not sufficiently characterize nonlinear dependencies between the variables.

Secondly, the existing studies use Cox regression to extract significant variables (p<0.05) and use a manual scoring system to map the variables to a prediction. The limitation is that this is a linear combination of information susceptible to nonlinear confounding of other variables. Scoring systems that rely on integer scores for convenience are potentially unoptimized as thresholds for risk stratification may not be exactly at an integer value.

Thirdly, the studies were restricted in analyzing only clinicopathological variables. As tumor progression results in changes in liver tissue, there may be observable structural changes in the liver associated with survival. Rahmim et al. [12] utilized volumetric and heterogeneity measurements from positron emission tomography scans of 52 patients as input variables to a Cox hazard regression. Metabolic tumor volume (HR = 1.44-5.12) and total lesion glycolysis (HR = 1.38-4.98) were found to be significant and highly correlated with survival.

Wang et al. [9] evaluated nine different survival prediction scoring systems and found six to result in a concordance index (C-index) crossing the 0.50 threshold in its 95% confidence interval, with their own scoring system resulting in a C-index of 0.642 (95% CI: 0.570–0.713) using KRAS mutation and "poor chemotherapy response" as input to their Cox model.

Thus, there exist current research gaps in implementing an automated nonlinear survival model that does not require manual score assignments, that considers the usage of imaging features, and that can be implemented with an algorithm that does not construct the model on the full dataset.

**Motivation**

Two motivating ideas have been developed for this work: the usage of radiomic imaging features as input data and modelling survival with machine learning.

Computed tomography (CT) has been standard of care for characterizing tumors by a trained expert. Radiomic features are computed attributes of a CT image which quantitatively characterize shape, intensity statistics, and gray-level relationships within the anatomy of interest. Differences in radiomic features have observed for different prognoses of liver metastases with Creasy et al. [13] and Simpson et al [14] observing that increased homogeneity was associated with increased risk of hepatic recurrence. Characterizing CT scans quantitatively with computer-aided methods such as radiomics provides the benefits of automated, consistent, and detailed observations of shape and texture, which can be used in the construction of predictive models.

To address the challenge of modelling variables which may not necessarily be linearly related, nonlinear and automated methods are required. Artificial intelligence (AI) methods have shown potential in survival prediction in previous studies [14,15] A major strength of AI methods are their automated iterative approach, where parameters are repeatedly optimized. This allows for the initialization of complex model architecture which may be more suitable to model interdependent variables than a linear model.

The main contribution of this work is the development of an automated survival prediction system involving the usage of CT liver scans as input data into a machine learning prediction model.

**Materials and Methods**

**Patient Demographics:**

Patient demographics and baseline clinical variables are displayed in Table 1.

**Data Collection and Equipment:**

Data for a total of N=97 patients undergoing radiation treatment for CLM were retrospectively obtained from Memorial Sloan Kettering Cancer Center (MSKCC), including pre-treatment CT scans, dosimetric data, and outcome data. Right-censored time-to-event data was collected for two outcomes related to progression of disease, namely progression anywhere in the liver and local progression in the treated tumor. The study was approved by the MSKCC research ethics board (ID: 16-328). Liver and tumor volumes were segmented by radiation oncologists at MSKCC, as is standard of care. The AI model was programmed in Python, utilizing the Pyradiomics and Pysurvival libraries.

**Problem Definition**

The task for the AI model was to predict the primary endpoint, defined as time until local tumor progression. To accomplish this, an AI survival prediction model, visualized in Figure 1, has been developed, consisting of an offline training component and a real-time prediction component. The input to the training component is a set of liver CT scans. Radiomic features are extracted from the liver and/or tumor volumes and used to train a survival model, which learns to predict a survival time interval for the patients in the training dataset. After the model is trained, new patient CT scans can be used as an input to the finalized model to compute a real-time survival prediction. The training stage contains three main components: radiomic feature extraction, feature selection, and random survival forest modelling, visualized in Figure 2.

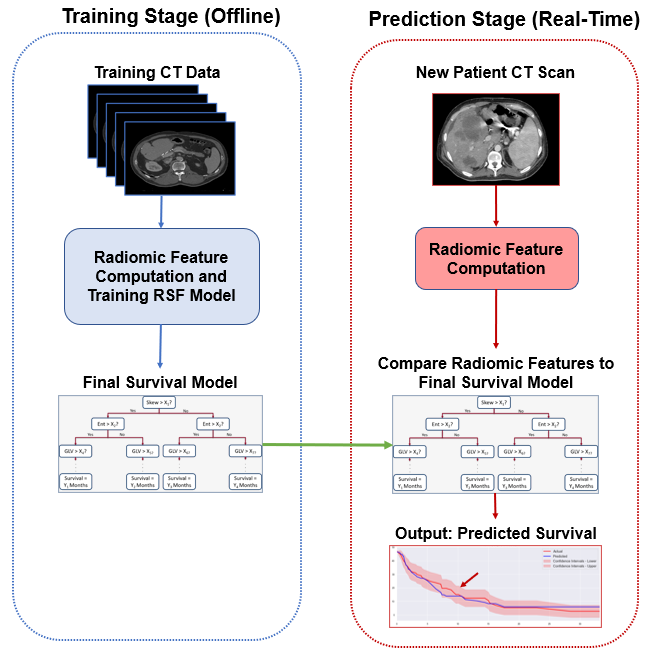


Figure 1: A visualization of the survival prediction system. The system contains two stages. The first is a training stage, where a survival model is built by training a random survival forest from radiomic features extracted from a set of computed tomography liver scans. Once the survival model has been built, it can be exported to a real-time prediction environment, where liver scans of new patients can be fed as input to the survival model to obtain a predicted survival for the new patient. In this way, most of the computation required is done beforehand to build the model and prediction can occur in real-time for new patients.

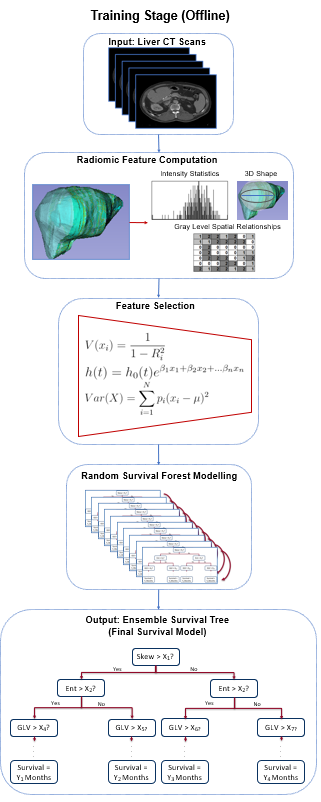


Figure 2: Steps involved in training the prediction model. The computed tomography scans of the training set are first processed by extracting the liver volumes and computing radiomic features on the volumes. Feature selection is used to filter irrelevant or redundant variables, and the remaining features are used to build a decision tree using the random survival forest algorithm. The final output is an optimized ensemble survival tree which predicts survival given a set of radiomic features.

**Radiomic Feature Extraction**

In the first stage to train the model, 108 radiomic features were computed for a liver volume extracted from a CT scan. This includes computations related to shape, intensity statistics, gray-level co-occurrence matrices, gray level run-length matrices, gray level dependence matrices, and gray tone difference matrices. A full list of radiomic features is available in the Supplemental Table 1. A visualization of radiomic feature extraction is displayed in Figure 3.

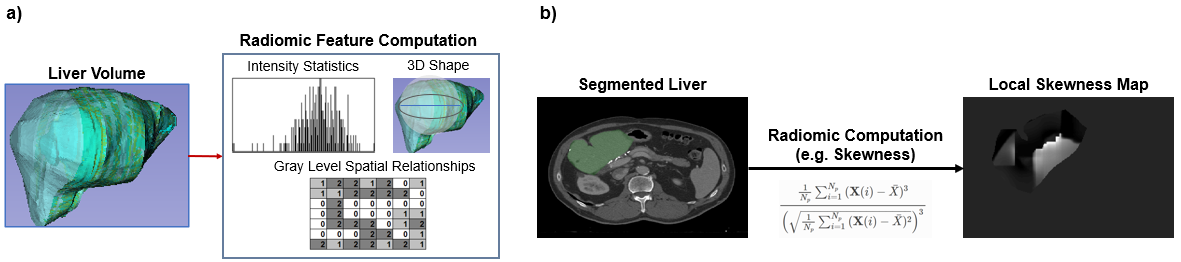


Figure 3: Visualization of radiomic feature extraction. a) A liver volume is extracted from the computed tomography scan and features relating to the intensity statistics, shape, and gray level matrices of the live volume are computed. b) An example of one such radiomic feature, skewness, from a liver slice. The skewness generated is a local map of skewness. In the final radiomic computation, one singular value is computed for the entire volume.

**Feature Selection**

Retaining all 108 radiomic features would likely result in overfitting due to inclusion of irrelevant or redundant features. Hence, several feature selection strategies have been employed. First, variance thresholding was performed to remove any feature with no variance. Next, redundant features were removed using a variance inflation factor threshold of 10 as an indicator of collinearity [16].

If the number of samples are not sufficient for the dimensionality of the feature space, then overfitting is likely to occur [17]. To further reduce the feature set we therefore scored the relevance of the remaining features using the hazard ratios predicted for each variable in a Cox proportional hazards model [18], and removed less relevant features until the ratio of features to samples was less than 1:10. This additional step is dependent on the sample size and with sufficient samples in future studies may be unnecessary.

**Random Survival Forest Model**

To predict survival from the filtered feature set, the random survival forest (RSF) algorithm was used [19]. The RSF algorithm operates by creating an ensemble decision tree with nodes representing features with a threshold value. The features used for the nodes and the threshold values were adjusted with each pass of patient data to maximize the log-rank test given by:

Where at time *ti*, *Ei*is the number of events at time *ti, Ei,j* is the number of events at a daughter node *j, Yi* is the number of patients with an events or at risk at time *ti,* and *Yi,j*is the number of patients with an event or at risk at a daughter node. The cumulative hazard was calculated by the Nelson-Aalen estimator [20]. A detailed outline of the algorithm is available in the Supplemental Information section. Hyperparameters, including maximum number of trees, maximum number of patients for a terminal node, and maximum depth, were optimized with a gridsearch algorithm that occurs within a cross-validation loop. After optimization, feature importances were computed by the difference between the error rate and the perturbed error rate for that feature.

**Validation**

A 5-fold cross-validation scheme was used to provide multiple estimates of the performance of the model. In this type of scheme, the dataset is shuffled and partitioned into five approximately equal sized subsets. Each subset is then used as a test data set for a model trained on the union of the remaining four subsets, giving five separate models trained and evaluated using the same overall approach. The concordance index (C-index) and integrated Brier score (IBS) were used as accuracy metrics to evaluate the models in each fold. The final reported accuracy was computed by averaging the C-index and IBS over the 5 k-folds.

Ablation analysis was performed to investigate the performance of the model when adjustments to individual components were made. First, we defined 6 different feature sets:

1. Non-imaging and non-dosimetric clinical data: baseline patient variables not related to dosimetric information or tumor geometry from CT imaging
2. Dosimetric clinical data: variables related to dosage
3. Imaging clinical data: variables related to tumor geometry measure in CT imaging
4. Radiomics: tumor volume: radiomic features computed from the tumor volume only
5. Radiomics: liver parenchyma: radiomic features computed from the liver parenchyma only
6. Raciomis: Liver parenchyma + tumor: : radiomic features computed from the union of the tumor volume and liver parenchyma

UEach feature set was used to build a RSF survival model twice, once with the outcome being local progression and once with the outcome being any hepatic progression, resulting in 12 models. This was repeated with and without the feature selection stage, resulting in a total of 24 models to be evaluated.

**Results**

The averaged cross-validation accuracies are summarized in Table 2. Samples of the predicted survival and IBS curves compared to the ground truth are visualized in Figure 4. The results of feature importance computation are summarized in Table 3.

|  |  |  |
| --- | --- | --- |
| **Input Features** | **Concordance Index** | **Integrated Brier Score** |
| **(No Feature Selection, Local Progression as Outcome)** | | |
| Non-Imaging Clinical Data | 0.66 ± 0.10 | 0.17 ± 0.02 |
| **Imaging Clinical Data** | **0.74 ± 0.21** | **0.15 ± 0.02** |
| Dosimetric Clinical Data | 0.62 ± 0.07 | 0.16 ± 0.04 |
| Radiomics: Tumor Volume | 0.61 ± 0.07 | 0.16 ± 0.02 |
| Radiomics: Liver Parenchyma | 0.61 ± 0.07 | 0.17 ± 0.01 |
| Radiomics: Liver Parenchyma + Tumor | 0.69 ± 0.10 | 0.22 ± 0.03 |
|  | | |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| **(With Feature Selection, Local Progression as Outcome)** | | |
| Non-Imaging Clinical Data | 0.58 ± 0.08 | 0.21 ± 0.05 |
| Imaging Clinical Data | 0.62 ± 0.07 | 0.19 ± 0.01 |
| Dosimetric Clinical Data | 0.67 ± 0.08 | 0.21 ± 0.05 |
| Radiomics: Tumor Volume | 0.65 ± 0.10 | 0.18 ± 0.02 |
| Radiomics: Liver Parenchyma | 0.63 ± 0.07 | 0.21 ± 0.01 |
| **Radiomics: Liver Parenchyma + Tumor** | **0.76 ± 0.06** | **0.20 ± 0.03** |
|  | | |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| **(With Feature Selection, Any Hepatic Progression as Outcome)** | | |
| Non-Imaging Clinical Data | 0.57 ± 0.04 | 0.14 ± 0.01 |
| Imaging Clinical Data | 0.55 ± 0.03 | 0.14 ± 0.03 |
| Dosimetric Clinical Data | 0.62 ± 0.08 | 0.13 ± 0.02 |
| Radiomics: Tumor Volume | 0.61 ± 0.07 | 0.16 ± 0.03 |
| Radiomics: Liver Parenchyma | 0.52 ± 0.01 | 0.13 ± 0.05 |
| **Radiomics: Liver Parenchyma + Tumor** | **0.72 ± 0.05** | **0.13 ± 0.03** |

Table 2: A summary of accuracy results for each input combination to the model with the standard deviation range. The combination that achieved the highest accuracy was using radiomic features from the entire liver and tumor volume. The variances for the models decreased after feature selection. However, there is still overlap within the confidence intervals of the accuracies.

We note that nearly all input dataset variations resulted in a C-index greater than 0.50 within uncertainty ranges, with the except of utilizing radiomic features of liver parenchyma only to predict any hepatic progression and utilizing non-imaging clinical data to predict local progression. The greatest accuracy occurred when utilizing the entire liver parenchyma and tumor volume for radiomic analysis. Utilizing only dosimetric data for predicting freedom of local progression resulted in a C-index of 0.67 ± 0.08.



Figure 4: Comparison of the prediction survival (blue) compared to the actual survival (red), including the survival curves (top) and the integrated Brier scores (bottom) of the a) best model, using liver parenchyma plus tumor volume radiomics and the b) worst model using non-imaging clinical variables. The integrated Brier scores are below the 0.25 threshold for random prediction, indicated significant prediction of survival. However, the confidence intervals are quite large, which can be refined further by training the model with a larger dataset.

The IBS of all models were below 0.25, which is the standard threshold to not be a random prediction. Although the best model’s survival curves on average were similar to the actual survival, there is a large confidence interval range, reflected in the variance of the C-index measurements.

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Neighboring gray tone difference matrix strength | 0.874 ± 0.643 |
| Gray level size matrix large area low gray level emphasis | 0.567 ± 0.192 |
| Kurtosis | 0.383 ± 0.035 |
| Neighboring gray tone difference matrix busyness | 0.316 ± 0.069 |
| Gray level co-occurrence matrix cluster prominence | 0.286 ± 0.040 |
| Skewness | 0.181 ± 0.092 |
| Gray level size matrix small area low gray level emphasis | 0.297 ± 0.114 |
| Flatness | -0.048 ± 0.114 |

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| Max dose | 3.113 ± 0.138 |
| Carcinoembryonic antigen at radiotherapy | 1.435 ± 0.099 |
| Lesion 1 dimension 1 | 1.413 ± 0.086 |
| Lesion 1 dimension 2 | 1.351 ± 0.124 |
| Planning target volume | 1.033 ± 0.108 |
| Number of liver lesions | 0.862 ± 0.093 |
| Other sites at radiotherapy | 0.105 ± 0.050 |
| Lines | 0.081 ± 0.057 |
| Pump before radiotherapy | -0.181 ± 0.011 |

Table 3: The feature importance for the filtered radiomic features from the best model and the clinical variables used. The importances of the clinical variables were more consistent than the radiomic variables, indicating that there is a clearer discrepancy between which features are the most predictive. Further validation, such as with a larger dataset, may assist in reducing the variance of the radiomic feature importances.

The most predictive radiomic feature was the neighboring gray tone difference matrix strength, though with a large variance in importance over the 5 k-folds. The features that were assessed were all filtered features, hence, the majority of the radiomic features extracted were not used in modelling nor were their importances computed. Kurtosis was the feature most consistently important. The clinical variables were much more consistent in predictivity over the different test sets, with the maximum dosage having the greatest feature importance.

**Discussion**

The IBS of every dataset combination was below the threshold of 0.25, indicating that the predictions by the RSF model is non-random and information-gaining [21]. An additional strength of our study is that our method was cross-validated 5 times, rather than testing on the same data used for modelling in order to prevent inflated accuracies due to overfitting.

Without feature selection, although the survival model performed with an average C-index greater than 0.60, there was a large variance in performance across the cross-validation folds. Given that the variance was much lower for the models using feature selection, it is possible that the large variance was a product of overfitting due to an overly high-dimensional feature space given the sample size.

As discussed in the Existing Work section, most prior studies reported a C-index lower than 0.60 when tested on dataset that was not used in creating the prediction model, with one study achieving a C-index of 0.64 [9]. Utilizing clinical data only resulted in prediction C-indices within this range, though we did not have access to three of the variables in the model by Wang et al. [9]. In future studies, a fairer comparison between manual scoring systems and automated RSF methods could be done by ensuring that all clinical variables used to produce the score are available to the RSF.

~~Our models predicting local progression or any hepatic progression performed similarly when trained on only clinical variables. However, the models trained on features extracted from the liver parenchyma only resulted in near random prediction regardless of feature selection. This was expected as radiomic features that ignored the tumor was unable to predict recurrence locally near the tumor.~~

The best performing model for prediction of any hepatic recurrence was that relying on radiomic features extracted from the union of the liver parenchyma and tumor, with feature selection enabled, achieving a C-index of 0.76 ± 0.06. The C-index accuracy is greater than the 95% confidence interval of any previous studies to our knowledge. This method was second best at predicting local progression in terms of mean C-index (0.72 ± 0.05), falling only behind the model trained on clinical imaging features with no feature selection (0.74 ± 0.21), however, the latter model’s confidence interval entirely contains that of the former. The comparatively high performance of the model using the tumor and parenchyma may suggest that there is predictive textural information within the liver parenchyma and tumor volume that may be investigated further to understand structural changes leading to a different prognosis. This is consistent with findings from Simpson et al. [14], who observed that increased homogeneity of liver texture is associated with recurrence. The low accuracy when utilizing imaging clinical variables only for predicting any hepatic progression was expected, as the imaging variables were related to the target tumor volume.

The accuracy of the RSF has two major implications. First, radiomic features provide correlations to tumor progression. Changes in liver texture visualized on a CT image can be quantified with features associated with progression. Although subjective labelling results in inter-observer error in segmentation of the volume of interest, the extraction radiomic features is quantitative and automated compared to manual analysis of the volume, reducing variability. The regions of interest can be analyzed in future studies, such as with histological analysis, to investigate what changes have occurred in the liver that are either protective or susceptible to recurrence. For instance, kurtosis measures the weight of the tails in a distribution and with kurtosis being a predictive feature, there may be structural changes associated with survival when a liver has extreme hyperintense and hypointense regions. The second major implication is machine learning methods such as our RSF model can perform prediction tasks that have been historically difficult with linear and manual scoring methods. This supports the potential for machine learning to be used in the future to aid clinical decision making.

**Limitations**

A limitation of the study is the sample size. The model was able to achieve a C-index greater than previous studies, however, further validation with a diverse patient population from different centers for instance can evaluate the generalizability of the model with greater statistical power. A wide uncertainty in C-index values indicates that when tested on a different subset of patients, the model accuracy is not consistent. In our study, we limited the cross-validation to 5 k-folds. With a larger sample size, higher-order cross-validation, such as with 10 k-folds, may be conducted to have more testing sets so that evaluation of the model has more statistical power. Feature selection may not be necessary with sufficient samples as with sufficient computational power, a random forest can in theory filter variables that have low predictive information gain [22]. Future studies may also include patients before and after radiotherapy, as texture in CT scans may change after treatment.

Although the model accuracy supports the idea that the radiomic features are predictive, we are limited in the medical interpretation of these features. For instance, high skewness indicates that the intensities in the region of interest are not symmetrically distributed. This may indicate inhomogeneous interactions of electromagnetic radiation with the liver tissue, but the cause of the inhomogeneity is not well understood. Hence, although our study can assess the performance of a survival model, the features determined to be predictive require future studies to understand how they can be descriptive of the pathophysiology of metastases.

**Conclusion**

We have developed a time-dependent survival prediction model utilizing radiomic features from CT scans and an AI random survival forest. The model was able to achieve C-indices greater than previous studies utilizing radiomic features from the liver parenchyma and tumor volume. The resulting model may be applied in future studies to aid in prognostic decision making. Radiomic features determined to be predictive may be investigated in the future to understand structural changes reflected in radiomic observations in the CT scan to provide new directions for clinician analysis of liver texture.

**References:**

Leung U, Gönen M, Allen PJ, Kingham TP, DeMatteo RP, Jarnagin WR, D'Angelica MI. Colorectal Cancer Liver Metastases and Concurrent Extrahepatic Disease Treated With Resection. *Ann Surg*. 2017;265:158–165.

Jones RP, Jackson R, Dunne DF, et al. Systematic review and meta-analysis of follow-up after hepatectomy for colorectal liver metastases. *Br J Surg.* 2012;99:477–486.

Fong Y, Fortner J, Sun RL, Brennan MF, Blumgart LH. Clinical score for predicting recurrence after hepatic resection for metastatic colorectal cancer: analysis of 1001 consecutive cases. *Ann Surg.* 1999;230(3):309-321. doi:10.1097/00000658-199909000-00004.

Iwatsuki S, Dvorchik I, Madariaga JR, et al. Hepatic resection for metastatic colorectal adenocarcinoma: a proposal of a prognostic scoring system. *J Am Coll Surg.* 1999;189(3):291-299. doi:10.1016/s1072-7515(99)00089-7.

Konopke R, Kersting S, Distler M, et al. Prognostic factors and evaluation of a clinical score for predicting survival after resection of colorectal liver metastases. *Liver Int*. 2009;29(1):89-102. doi:10.1111/j.1478-3231.2008.01845.x.

Nagashima I, Tadahiro T, Matsua K, et al. A new scoring system to classify patients with colorectal liver metastases: proposal of criteria to select candidates for hepatic resection. *J Hepatobiliary Pancreat Surg.* 2004;11(2):79-83.

Nordlinger B, Guiguet M, Vaillant JC, et al. Surgical resection of colorectal carcinoma metastases to the liver. A prognostic scoring system to improve case selection, based on 1568 patients. Association Française de Chirurgie. *Cancer*. 1996;77(7):1254-1262.

Imai K, Allard MA, Castro Benitez C, et al. Nomogram for prediction of prognosis in patients with initially unresectable colorectal liver metastases. *Br J Surg*. 2016;103(5):590-599. doi:10.1002/bjs.10073.

Sasaki K, Morioka D, Conci S, et al. The Tumor Burden Score: A New "Metro-ticket" Prognostic Tool For Colorectal Liver Metastases Based on Tumor Size and Number of Tumors. *Ann Surg.* 2018;267(1):132-141. doi:10.1097/SLA.0000000000002064.

Rees M, Tekkis PP, Welsh FK, O'Rourke T, John TG. Evaluation of long-term survival after hepatic resection for metastatic colorectal cancer: a multifactorial model of 929 patients. *Ann Surg.* 2008;247(1):125-135. doi:10.1097/SLA.0b013e31815aa2c2.

Brudvik KW, Jones RP, Giuliante F, et al. RAS Mutation Clinical Risk Score to Predict Survival After Resection of Colorectal Liver Metastases*. Ann Surg*. 2019;269(1):120-126. doi:10.1097/SLA.0000000000002319.

Lin DY, Wei LJ. The robust inference for the cox proportional hazards model. *J Am Stat Assoc*. 1989;84:1074–1078.

Wang K, Liu W, Yan XL, Li J, Xing BC. Long-term postoperative survival prediction in patients with colorectal liver metastasis. *Oncotarget*. 2017;8(45):79927-79934. doi:10.18632/oncotarget.20322.

Rahmim A, Bak-Fredslund KP, Ashrafinia S, et al. Prognostic modeling for patients with colorectal liver metastases incorporating FDG PET radiomic features*. Eur J Radiol*. 2019;113:101-109. doi:10.1016/j.ejrad.2019.02.006.

Creasy JM, Cunanan KM, Chakraborty J, et al. Differences in Liver Parenchyma are Measurable with CT Radiomics at Initial Colon Resection in Patients that Develop Hepatic Metastases from Stage II/III Colon Cancer. *Ann Surg Oncol.* 2021;28(4):1982-1989. doi:10.1245/s10434-020-09134-w.

Simpson, AL., Doussot, A, Creasy, JM, et al. Computed tomography image texture: a noninvasive prognostic marker of hepatic recurrence after hepatectomy for metastatic colorectal cancer. *Ann Surg Oncol. 2017;24*(9): 2482-2490.

Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ. Deep learning-based survival prediction of oral cancer patients. *Sci Rep*. 2019;9(1):6994. doi:10.1038/s41598-019-43372-7.

Wang W, Liu W. Integration of gene interaction information into a reweighted random survival forest approach for accurate survival prediction and survival biomarker discovery. Sci Rep. 2018;8(1):13202. doi:10.1038/s41598-018-31497-0.

Salmeron R, Garcıa CB, Garcıa J. Variance Inflation Factor and Condition Number in multiple linear regression. *J Stat Comput Simul*. 2018;88(12):2365-2384. Doi:10.1080/00949655.2018.1463376.

Liu R, Gillies DF. Overfitting in linear feature extraction for classification of high-dimensional image data. *Pattern Recognit.* 2016;53(C):73–86. doi:10.1016/j.patcog.2015.11.015.

Bourgon R, Gentleman R, Huber W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(21), 9546–9551. https://doi.org/10.1073/pnas.0914005107.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2:841–860.

Colosimo EA, Ferreira FF, Oliveira MD, Sousa CB. Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. *J Stat Comput Simul.* 2002;72:299–308.

Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2.

Menze BH, Kelm BM, Masuch R, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*. 2009;10:213. doi:10.1186/1471-2105-10-213.

**Supplemental Information:**

**Random Survival Forest Algorithm:**

To build a survival tree that predicts survival from an input vector of radiomic features, the following steps are taken:

1. Select *N* samples from the dataset.
2. For each sample *i = 1, 2, … N,* initialize a binary decision tree with max depth *D.*
3. At each node, iterate through set of features *X = {x­1­, x2, … xN}* and its range of feature values *S = {Smin, Smax}* to select feature *xi* and a threshold split value *si* such that:

Where *L(x,s)* is the log rank test such that

Where at time *ti*, *Ei*is the number of events at time *ti, Ei,j* is the number of events at a daughter node *j, Yi* is the number of patients with an events or at risk at time *ti,* and *Yi,j*is the number of patients with an event or at risk at a daughter node *j*

1. Continue to grow children nodes unless the children node has no more than *M* surviving samples, where *M* is a user-defined hyperparameter
2. Calculate the cumulative hazard function for the decision tree with the Nelson-Aalen estimator:

Where *p* is a patient in the set of *M* patients in set *P = {p1,p2, … pM}*, *q* is a node in the set of *N* nodes in set *Q = {q1,q2, … qN}*, *Ep,q* is the number of events at time *tp,q*, and *Yp,q* is the number of patients with an event or at risk at time *tp,q*.

1. Repeat steps 1-5 *K* times to create *K* separately initialized trees, where *K* is a user-defined hyperparameter.
2. Average the cumulative hazard function over all trees to compute the ensembled cumulative hazard.

**List of Radiomic Features**

Supporting Table 1 shows a full list of radiomic features extracted from a liver volume.

|  |  |
| --- | --- |
| **Feature Category** | **Features** |
| First Order Statistics | Energy  Total Energy  Entropy  Minimum  10th Percentile  90th Percentile  Maximum  Mean  Median  Interquartile Range  Range  Mean Absolute Deviation  Robust Mean Absolute Deviation  Root Mean Squared  Standard Deviation  Skewness  Kurtosis  Variance  Uniformity |
| 3D Shape | Mesh Volume  Voxel Volume  Surface Area  Surface Area to Volume Ratio  Sphericity  Compactness  Spherical Disproportion  Maximum 3D Diameter  Maximum 2D Diameter (Axial)  Maximum 2D Diameter (Coronal)  Maximum 2D Diameter (Sagittal)  Major Axis Length  Minor Axis Length  Least Axis Length  Elongation  Flatness |
| Gray level Co-occurrence Matrix | Autocorrelation  Joint Average  Cluster Prominence  Cluster Shade  Cluster Tendency  Contrast  Correlation  Difference Average  Difference Entropy  Difference Variance  Difference Average  Joint Energy  Joint Entropy  Informational Correlation  Inverse Difference Moment  Inverse Difference Moment Normalized  Inverse Difference  Inverse Difference Normalized  Inverse Variance  Maximum Probability  Sum Average  Sum Entropy  Sum of Squares |
| Gray Level Size Zone Matrix | Small Area Emphasis  Large Area Emphasis  Gray Level Non-Uniformity  Gray Level Non-Uniformity Normalized  Size-Zone Non-Uniformity  Size-Zone Non-Uniformity Normalized  Zone Percentage  Gray Level Variance  Zone Variance  Zone Entropy  Low Gray Level Zone Emphasis  High Gray Level Zone Emphasis  Small Area Low Gray Level Emphasis  Small Area High Gray Level Emphasis  Large Area Low Gray Level Emphasis  Large Area High Gray Level Emphasis |
| Gray Level Run Length Matrix | Short Run Emphasis  Long Run Emphasis  Gray Level Non-Uniformity  Gray Level Non-Uniformity Normalized  Run Length Non-Uniformity  Run Length Non-Uniformity Normalized  Run Percentage  Gray Level Variance  Run Variance  Run Entropy  Low Gray Level Run Emphasis  High Gray Level Run Emphasis  Short Run Low Gray Level Emphasis  Short Run High Gray Level Emphasis  Long Run Low Gray Level Emphasis  Long Run High Gray Level Emphasis |
| Gray Level Dependence Matrix | Small Dependence Emphasis  Large Dependence Emphasis  Gray Level Non-Uniformity  Dependence Non-Uniformity  Dependence Non-Uniformity Normalized  Gray Level Variance  Dependence Variance  Dependence Entropy  Low Gray Level Emphasis  High Gray Level Emphasis  Small Dependence Low Gray Level Emphasis  Small Dependence High Gray Level Emphasis  Large Dependence Low Gray Level Emphasis  Large Dependence High Gray Level Emphasis |
| Neighboring Gray Tone Difference Matrix | Coarseness  Contrast  Busyness  Complexity  Strength |

Supporting Table 1: A list of radiomic features extracted from a liver volume. The features include computations related to the statistics, shape, and gray-level relationships of the image.