

Image-based Survival Analysis for Lung Cancer Patients using CNNs

Christoph Haarbuerger¹, Philippe Weitz¹, Oliver Rippel, Dorit Merhof

Institute of Imaging and Computer Vision, RWTH Aachen University, Germany

Abstract

Traditional survival models such as the Cox proportional hazards model are typically based on scalar or categorical clinical features. With the advent of increasingly large image datasets, it has become feasible to incorporate quantitative image features into survival prediction. So far, this kind of analysis is mostly based on radiomics features, i.e. a fixed set of features that is mathematically defined a priori. In order to capture highly abstract information, it is desirable to learn the feature extraction using convolutional neural networks. However, for tomographic medical images, model training is difficult because on one hand, only few samples of 3D image data fit into one batch at once and on the other hand, survival loss functions are essentially ordering measures that require large batch sizes. In this work, we show that by simplifying survival analysis to median survival classification, convolutional neural networks can be trained with small batch sizes and learn features that predict survival equally well as end-to-end hazard prediction networks. Furthermore, we demonstrate that adding features from a fine-tuned convolutional neural network improves the predictive accuracy of Cox models that otherwise only rely on radiomics features. Moreover, we propose survival label noise as a means of data augmentation for deep image based survival analysis.

Keywords: Survival Analysis, Convolutional Neural Networks, Lung Cancer

1. Introduction

The medical image computing (MIC) community has been influenced strongly by advancements in machine learning and computer vision. Public availability of large annotated datasets such as from the BraTS and LUNA challenges [1, 2] has highly improved applicability and reproducibility of deep learning in MIC. As a result, the state of the art in computer aided diagnosis and detection as well as segmentation of medical images is currently dominated by convolutional neural networks (CNNs) [3]. A MIC subfield that has not seen such a strong benefit from these methods yet is survival analysis (i.e. prognosis) based on medical images. Survival analysis has been influenced mostly from biostatistics, i.e. statistical modeling based on non-image data. Motivated by the recent success of radiomics [4], there has been increasing interest in image-based survival analysis.

1.1. Survival Analysis

Survival analysis refers to the study of the time-to-event data for an individual or the study of the distribution of those times for a cohort. Especially in clinical research, survival analysis is broadly applied as it allows the quantitative comparison of subgroups of a patient cohort e.g. to evaluate the response to different treatment options and to identify the key factors for a response. Typical events in a medical context are death, disease incidence or relapse from remission. Common regression

modeling strategies usually cannot be applied to survival data since while for each patient, a time-to-event is specified, those events are qualitatively different. Because in medical trials, it is common for patients to become lost to follow-up, for some patients, the time indicated is the time-to-event, while for others, it is the time until dropping out of the study. This is referred to as *right-censoring* and indicated by the event indicator δ_i that equals 1 if the event occurred and 0 for censoring.

A common approach to survival analysis is the prediction of hazards λ from which a survival time can be obtained. The most broadly applied model for hazard prediction is the *Cox proportional hazards model* [5]. This model predicts patient-individual hazards λ_i based on covariates x_i with

$$\lambda(t|x)_i = \lambda_0(t) \cdot \exp(\beta^T x_i). \quad (1)$$

Here, $\lambda_0(t)$ is the baseline hazard that is equal for all patients and that parametrizes the model together with β .

An essential property of Eqn. 1 is the *proportional hazards assumption*, that states that the impact of covariates is not time dependent, since $\beta \neq f(t)$. While this assumption might prove incorrect, it significantly simplifies survival analysis if the primary objective is not the estimation of survival times but rather to assess the significance of individual covariates, e.g. a treatment variable. This assessment is adjusted for the confounding factors of the model. When the quantification of relative risks is the main goal, it is not necessary to parametrize $\lambda_0(t)$. For this reason, the Cox model is often referred to as *semi-parametric*. Since all time dependencies are eliminated from the model if $\lambda_0(t)$ is not parametrized, hazard prediction becomes essentially an ordering task. If $S(x_i)$ denotes the survival time of patient i ,

Email address: christoph.haarbuerger@lfb.rwth-aachen.de
(Christoph Haarbuerger)

¹These authors contributed equally

two observations are correctly ordered if

$$S(x_i) > S(x_j) \rightarrow \lambda(x_i) < \lambda(x_j). \quad (2)$$

If this holds true for the predicted hazards of two observations, they are referred to as *concordant*. Correspondingly, the most broadly applied metric in survival analysis is the *concordance index* or *c-index*, which is defined as

$$C = \frac{\# \text{ concordant pairs}}{\# \text{ possible pairs}} \in [0, 1]. \quad (3)$$

Here, only pairs with a maximum of one censored sample are possible, where the time-to-event is higher for the censored observation. Therefore, the c-index is a unit-less metric that quantifies the accuracy of pairwise ordering of observations as predicted by a model, with 1 indicating a completely correct order and 0.5 random order. Its numeric expression can therefore be interpreted the same way as the area under the receiver operating curve (AUROC), although c-indices that are currently achieved in state-of-the-art models are usually much lower than common AUROC values [6].

For several reasons, image-based survival analysis has not yet fully benefitted from recent advancements in deep neural networks: Training data is typically censored and cannot be handled properly by classification or regression approaches. Therefore, the most widely-used loss functions and network architectures are not applicable. Moreover, the standard evaluation measure in survival analysis, the *concordance index*, is an *ordering measure* that can be hard to interpret, especially when combined with batch-wise gradient descent methods. Datasets are even harder to collect than in computer-aided diagnosis or segmentation because laborious follow-up, potentially over years, is required.

1.2. Related Work and Contributions

Most approaches to image-based survival analysis performed a large-scale image feature extraction and feature selection, followed by a linear combination of the selected features in a Cox model [4, 7, 8, 9].

The first applications of neural networks to survival analysis date back to the 1990s [10], but neural networks have not become a common replacement for the Cox model since. Recently, modern neural networks were employed for survival analysis based on non-image data in [11, 12, 13], significantly outperforming traditional methods such as Cox models. However, these models did not incorporate a trainable image feature extraction as needed for image-based survival prediction. In [14], convolutional neural networks (CNNs) were first utilized for end-to-end trainable image feature extraction and survival analysis based on pathology images. This model was further extended in [15] to capture information from whole-slide images. The method proposed in [16] can perform survival analysis based on both pathology images and scalar clinical data by maximizing correlation between clinical and CNN features. To our best knowledge, there is no literature on survival analysis based on trainable image features from *tomographic* images so far. In [17], features from a CNN trained

for RGB image classification were extracted for survival prediction based on magnetic resonance images. However, in this work the CNN was not actually trained on tomographic medical images but only used as a fixed feature extractor. Tomographic medical image data is especially challenging to combine with survival prediction networks: On one hand, due to high dimensionality, tomographic medical images require small batch sizes during training to fit into GPU memory. On the other hand, the loss function typically used in survival analysis, the Cox partial log likelihood loss, is an ordering measure for which large sample sizes are required.

We aim to address this issue by transferring features learned by a classification problem to survival analysis without losing performance. Moreover, we propose survival label noise for data augmentation that models uncertainty in time-to-event ground truth data and proves effective for training CNNs for survival analysis. Finally, we propose a method to combine radiomics and learned CNN features that enforce the CNN to learn features that are both discriminative and not covered by the radiomics feature set. All methods are evaluated on a publicly available dataset of computed tomography (CT) images of non-small-cell lung cancer (NSCLC) patients and corresponding survival labels. We show that our method can outperform the approach presented in [4].

2. Material and Methods

2.1. The Lung1 Dataset

The *Lung1* data set is publicly available at *The Cancer Imaging Archive (TCIA)* [18, 19] and consists of 422 NSCLC patients. It was acquired at the Department of Radiation Oncology, MAASTRO Clinic Maastricht (The Netherlands). For 318 of the 422 patients, segmentations of the tumor are publicly available from TCIA. The voxel dimensions are constant across all patients with 1 mm both in sagittal and coronal direction and 3 mm in axial direction. An example of an axial slice from the *Lung1* dataset is provided in Fig. 1.



Figure 1: Axial slice of NSCLC patient with a survival time of 72 months. The segmented ROI is indicated in red.

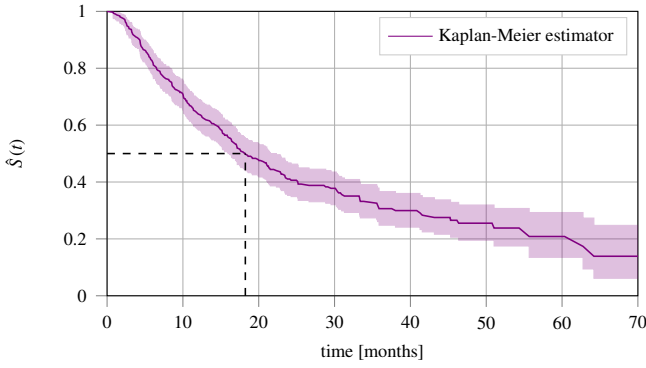


Figure 2: Kaplan-Meier estimate of the survival time in months for the entire patient cohort of 313 patients. $\hat{S}(t)$ indicates the percentage of patients that are still alive at time t . The dashed line indicates the median survival time at $\hat{S}(t) = 0.5$ at $t = 18.2$ months. It is notable that the median survival time does not refer to the median of all event times but to the time at which the event has occurred for 50 % of the cohort.

We excluded patient 176 since the segmentation mask appears to be corrupted. Patients 72, 249, 256 and 269 were also excluded since their metastasis stage M (TNM cancer staging) indicates tumors in distant organs that may potentially corrupt a survival analysis for NSCLC. The remaining 313 patients are in cancer stages I-IIIb. Of these 313 patients, 205 (65.5 %) died, while 108 (34.5 %) were censored. The median survival time determined by a Kaplan-Meier estimator as depicted in Fig. 2 is 18.2 months.

2.2. Baseline - Cox Proportional Hazards Model with Radiomics Features

As a baseline method for comparison with the proposed methods, we utilize a Cox proportional hazards model that performs hazard prediction using image features. Based on the segmentation masks that are provided with the dataset, 18 statistics features, 15 shape features and 73 texture features based on Gray-Level-Cocurrence-Matrix, Gray-Level-Runlength-Matrix, Gray-Level-Size-Zone-Matrix, Gray-Level-Difference-Matrix as well as Neighbourhood-Gray-Tone-Difference-Matrix were extracted using PyRadiomics [20] and a bin width of 20.

We deployed a forward feature selection that iteratively adds the feature with the next-highest bootstrapped univariate c-index to the feature set, unless its monotone correlation with a feature that is already in the feature set is higher than a threshold as proposed in Algorithm 1.

The only hyperparameter of this approach is the number of features in the set of selected features. To obtain the specified number of features, the feature selection process is repeated with a higher threshold value for the Spearman correlation if the desired number of features is not reached in the first iteration.

2.3. CNN for Hazard Prediction

In this approach, a ResNet18 [21] is pretrained on the ImageNet dataset [22] for classification of natural RGB images. The input weights of the three RGB channels are replicated such that 25

Data: features per observation, event indicator and survival outcome

Result: feature set

fit univariate CPH model to each feature

bootstrap to obtain mean of the univariate c-index $c(i)_{mean}$

discard all features except for the best k

while $n_{selected} < n_{desired}$ **do**

 pick next feature with the highest $c(i)_{mean}$

 get highest correlation coefficient $\rho(i)$ to any feature in the output set

while $i \neq k$ **do**

if $\rho(i) \geq \rho_{lim}$ **then**

 add feature to output set

end

end

 increase ρ_{lim}

end

Algorithm 1: Feature selection algorithm to maximize explanatory value and minimize correlation of features. For median survival classification, the univariate Cox model is replaced by a logistic regression model and the c-index by the corresponding AUROC value.

CT slices centered around the slice containing the most tumor tissue can be utilized as input for the model. To accommodate the entire section of the CT slices around the patient, central patches comprising 260×260 pixels around the tumor centroid are extracted. In order to adjust the ResNet18 architecture for the problem at hand, the following modifications of the architecture are performed: The 7×7 average pooling kernels are replaced by global average pooling, which makes the transition between convolutional and fully connected layers independent of the size of the consequently larger feature maps. The CNN is used in two feature extraction approaches:

1. **CNN features:** Extract features by finetuning pretrained ResNet18 as listed in top right of Fig. 3.
2. **Multimodal features:** Concatenate radiomics features selected as explained in Section 2.2 with ResNet18 features. This approach is sketched by considering both blocks at top of Fig. 3.

After image feature extraction, hazard prediction is performed in two variants:

1. **Direct hazard prediction:** In this setup, hazard prediction is performed by the trainable layers listed as "Prediction" in Fig. 3. With Eqn. 1, the prediction layer can be interpreted as the term βx_i , where β corresponds to the weights of the layer and x_i to the activations of the previous layer instead of covariates. Optimizing the final fully connected layer is equivalent to the maximum likelihood estimation of β when fitting Cox models. Consequently, the resulting network can perform hazard prediction similar to [11, 15]. It is notable that while the linearity assumptions of the Cox proportional hazards model only applies to the final neural network layer, the proportional hazards assumption extends to the predicted hazard.

2. **Cox hazard prediction:** Perform hazard prediction by a Cox model and use CNN only for feature extraction after fine-tuning it with the negative partial log-likelihood. In this case, the "Hazard Prediction" part in Fig. 3 is replaced by a Cox model. Features are selected from the radiomics features and all activations of the fully-connected layers.

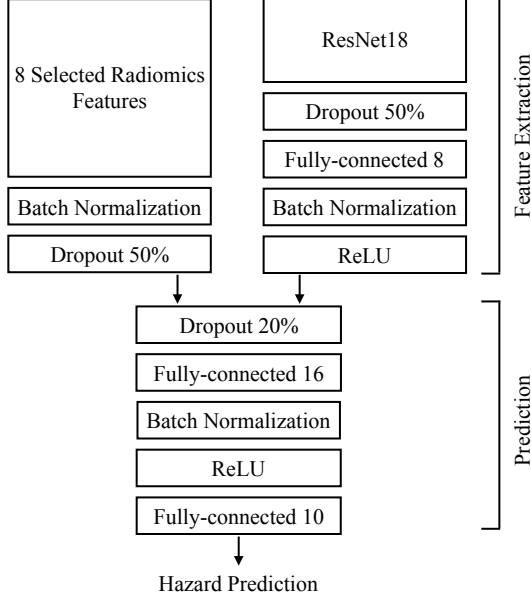


Figure 3: Model schematic for hazard prediction or classification based on both radiomics and CNN features.

The CNN is trained using cox negative partial log likelihood loss

$$\log L(\beta) = \sum_{T_i \text{ uncensored}} \left[\beta^T x_i - \log \left(\sum_{T_j \geq T_i} \exp(\beta^T x_j) \right) \right]. \quad (4)$$

However, training deep neural networks using this loss function is problematic: Typically, training is performed using stochastic gradient descent or closely related methods. This works well for classification and regression problems, but when minimizing an ordering measure as in this case, the ordering problem becomes easier to solve if the batch size is small. This is especially problematic when working with tomographic medical images and very deep network architectures because in that case, batch size *must* be set very small in order to fit both model and data into GPU memory. It is not uncommon to work with batch sizes as small as one [23], with which no gradient for Eqn. 4 can be computed.

2.4. CNN for Median Survival Classification

In order to overcome the shortcomings of ordering measures as loss functions it is desirable to modify the problem formulation in a way that allows training with a batch size of one. We propose to formulate the hazard prediction problem as a classification problem. This allows to train CNNs in a more standard setup. The learned features can then be utilized either by a Cox model or for direct hazard prediction. Therefore, in this

approach we classify whether a patient's survival time exceeds the median survival time. This has the additional advantage of ensuring balanced classes. The class of a patient is then defined as 1 if the patient lived past the median survival time and 0 otherwise. To incorporate censored observations for median survival classification, each patient i is assigned a weight w_i for a binary cross-entropy loss. With the median survival time $T_{0.5}$, the survival time T_i for patient i and the corresponding event indicator δ_i , the weights for the loss are computed according to

$$w_i = \begin{cases} 1 & \text{if } T_{0.5} \leq T_i \\ \delta_i & \text{if } T_{0.5} > T_i \end{cases}. \quad (5)$$

This weight assumes a value of 0 for all patients censored before the median survival time and 1 otherwise.

If survival analysis was formulated as a regression task and all censored patients were therefore excluded, this would distort the distribution of survival times significantly. This is because typically, most censorings occur towards higher survival times, when patients are lost to follow-up because the study ends. Excluding these observations would therefore exclude precisely those patients that did not die of the disease in question. Since patients that are lost to follow-up before the median survival time can fairly be assumed to be lost due to technical reasons, excluding these patients does not alter the distribution of survival times significantly.

Another major advantage of median survival classification is that the corresponding survival models can be trained with a batch size of one sample. With hazard losses such as the partial negative log-likelihood or the simoid, log-sigmoid and exponential lower bound losses from [24], the loss can only be computed across a whole batch. This is because the loss of an ordering task can only be meaningfully evaluated relatively between samples and therefore with batches that contain a sufficient number of observations.

Our complete setup looks as follows: The CNN features from median survival classification are concatenated to the radiomics features before feature selection. Then, a Cox proportional hazards model is fitted based on selected radiomics and CNN features to predict a hazard that allows the calculation of a c-index.

2.5. Time-to-Event Label Noise for Data Augmentation

As typical for medical image datasets, the *Lung1* data set is relatively small for training deep neural networks, data augmentation is therefore applied heavily. Specifically, the CT slices are randomly cropped, rotated, mirrored, altered with additive Gaussian noise and deformed with elastic transformations. Moreover, windows of 25 consecutive CT slices used for prediction are randomly shifted around the slice that contains the most tumor tissue. This extends random 2D translations to tomographic slices.

Apart from these commonly applied data augmentation techniques, *label noise* is a promising augmentation strategy for survival data. In classification, the rationale behind label smoothing is that cross-entropy based loss functions enforce a classifier response during training that maximizes the prediction

likelihood for the correct class while minimizing all other classifier activations. Misclassification between similar classes is penalized the same way as misclassification between dissimilar classes, which leads to overfitting [25].

In contrast to classification labels, survival labels are *inherently* noisy: Survival time is defined as the temporal distance between diagnosis and event (death, recurrence). The time of diagnosis is typically not the same as the time of outbreak of a disease but a function of symptoms, time until appointment with a specialist and many other random variables. Moreover, occurrence of event is essentially a random process that is influenced by treatment, lifestyle, comorbidities and other aspects. A medical image and clinical covariates only contain an incomplete subset of priors for modeling that process. Therefore, in survival analysis, ordering neighboring samples incorrectly should not be penalized the same way as incorrectly ordering more distant samples. We address this issue by adding artificial noise to survival labels as follows: In every training epoch, the survival label for patient i is drawn from a probability distribution $G_i(\mu_i, \sigma_i)$, e.g. a uniform or normal distribution. Here, μ_i denotes the mean of the distribution and σ_i its standard deviation. We choose σ_i to be a percentage of the survival time of the individual patient, where the exact value of that percentage can be determined through a hyperparameter search. This way, for patients with a short survival time, the noise introduced is lower, while the added uncertainty for patients with a long survival time is higher. For overall survival predictions, μ_i should be chosen as the time of death as reported. For time-to-recurrence labels, it might be more appropriate to center a skewed distribution at a time prior to the reported survival time, since recurrence always occurs before the diagnosis of recurrence.

3. Experiments

For evaluation purposes, the data set is split into 100 random splits with a relative test set size of 25 %, which corresponds to 81 patients. Another 15 % of the data is held out as a validation set for hyperparameter tuning. The random splits are stratified based on the event indicator. Despite the higher amount of required model fits, 100 random splits appear to be a more reliable approach to assessing the predictive accuracy of survival models than cross-validation. While cross-validation is an appropriate evaluation method for classification or segmentation tasks, the c-index is a relative measure for predictive accuracy between individual hazard predictions. Increasing the number of combinations of patients in different test sets therefore allows for more meaningful interpretations of the c-indices achieved. For a fair comparison with the previous state-of-the-art, we evaluate the approach from [4], a linear combination of four specific radiomics features in a Cox model, on the exact same data. This comparison is biased in favour of the approach in [4] since the features were selected on the *Lung1* data set, part of which is now the test set. The c-index of this approach is 0.609.

Tab. 1 lists the results for the different models proposed. The c-index of the baseline Cox model with features selected according to Alg. 1 without deep features is 0.615. The highest

c-index achieved is 0.623 for a Cox model fitted with a selection of deep features and radiomics features from a hazard prediction CNN without concatenated radiomics features within the neural network. This score is virtually identical to the corresponding c-index of 0.623 of a Cox model fitted with deep features from a median survival classification CNN and radiomics features. The corresponding c-indices for Cox models with deep features from a multi-modal hazard prediction network and median survival classification network are lower with 0.62 and 0.622, respectively. Direct hazard predictions from a neural network with radiomics features (multi-modal) and without are less precise with c-indices of 0.613 and 0.585 respectively.

Besides the c-indices, AUROC values for median survival classification are provided in Tab. 2. The baseline model for classification is a logistic regression model. The features for this model are selected with Algorithm 1, but the Cox model is replaced by a logistic regression model to obtain univariate AUROC values instead of c-indices. The baseline AUROC obtained with this model is 0.585. For this model, radiomics features were selected as described in Section 2.2. The best AUROC of 0.585 for the logistic regression model is achieved with a combination of radiomics features with CNN features. The median survival classification model based on ResNet reaches an AUROC of 0.598 without and 0.636 with radiomics features, which is the highest AUROC achieved in this study.

In another experiment, we evaluated the impact of survival label noise as described in Section 2.5. To this end, we performed the same experiments as described earlier in this section with additional uniform 50 % noise on event times. Results for this experiment are provided in Tab. 3. While the c-index of the CNN without radiomics features increases slightly to 0.588, all other models worsen with this additive noise.

4. Discussion

The best c-index achieved in our approach is higher than the baseline, however it is lower than the c-index of 0.65 reported in [4]. Since the approach in their work is very similar to our baseline approach, we hypothesize that the difference is mainly due to the different amount of available training data. While data for 422 patients could be used for training in [4], our study relies on a training set containing 232 patients only. Aerts et al. do not provide a c-index for the *Lung1* data set because they assessed performance on a separate test set for which segmentation masks are not publicly available. Therefore, we implemented their approach in this work for a fair comparison on the same data. Both the baseline model as well as the CNN models outperform the Cox model presented by Aerts et al.

The relatively high variance that can be observed for all models, even the linear and deterministic Cox model from [4] with four fixed features, indicates that the variance is not due to the models but rather to different generalization properties inherent to different random splits. This is also an indication that the number of samples in the data set is insufficient. Furthermore, it prevents a meaningful assessment of the statistical significance of the differences in c-index e.g. with a Kolmogorow-Smirnow test. [16] report similarly high variances for their experiments

Model	C-Index	
	Cox hazard prediction	Direct hazard prediction
Radiomics + Cox (Aerts et al.) [4]	0.609 \pm 0.041	-
Radiomics + Cox (baseline)	0.615 \pm 0.037	-
Hazard prediction CNN	0.623 \pm 0.039	0.585 \pm 0.044
Multi-modal hazard prediction CNN	0.620 \pm 0.039	0.613 \pm 0.04
Median survival CNN	0.623 \pm 0.04	-
Multi-modal median survival CNN	0.622 \pm 0.038	-

Table 1: Hazard prediction results for proposed models and baseline method. Reported c-indices refer to mean and standard deviations over 100 stratified random splits on the dataset. The results in the left column were obtained by combination of extracted (CNN-) features in a Cox model. Results in the right column were achieved by direct hazard prediction by respective CNNs. The hazard prediction CNN is trained using negative cox partial log likelihood loss as explained in Section 2.3. Median survival CNN is trained as a classifier as explained in Section 2.4. Multimodal refers to a combination of CNN and radiomics features as outlined in Fig. 3.

Model	AUROC
Radiomics + logistic regression (baseline)	0.585 \pm 0.051
Median survival CNN	0.598 \pm 0.067
Multi-modal median survival CNN	0.636 \pm 0.057

Table 2: Median survival classification results of overall survival at $T_{0.5}$. Reported AUROCs refer to mean and standard deviations for the test performance over 100 stratified random splits on the dataset. Median survival CNN refers to a ResNet18 CNN that performs binary classification. Multi-model median survival CNN additionally incorporates radiomics features as outlined in Fig. 3.

Model	Cox hazard prediction	Direct hazard prediction
Hazard prediction CNN	0.608 \pm 0.043	0.588 \pm 0.045
Multi-modal hazard prediction CNN	0.601 \pm 0.045	0.602 \pm 0.044
Median survival CNN	0.594 \pm 0.042	-
Multi-modal median survival CNN	0.595 \pm 0.04	-

Table 3: Results with additional survival label noise, given by uniform distribution with 50 % noise around the event times.

with pathology images. This underlines that the acquisition of larger survival data sets would be beneficial for more reliable evaluation of model performance, not only for CNNs but also Cox models or survival forests.

Cox models with deep features and radiomics features outperform CNNs with concatenated radiomics features, which can be attributed to several causes. First, overfitting might impact the CNN stronger than the Cox model despite the high drop-out for the radiomics features since there are significantly more parameters even after the concatenation than in the corresponding Cox model. While this can be expected to influence generalization, an even more compelling second reason might be that the Cox models are fitted with 25 % more data relatively to the training set for the CNN models. This is because the fine tuning of hazard CNNs proved to be highly volatile, such that a validation set is always required for early stopping, not only for hyperparameter optimization. Considering the smaller number of training examples, the CNNs actually perform quite well in relation to the Cox models. If the segmentations for the remaining NSCLC data sets are ever to be made publicly available, it would be interesting to study whether the CNN models scale

better with an increase in training set size than the Cox models.

The fact that AUROC and c-index values obtained are in the same range indicates that the median classification task is not a particularly well posed classification problem. This is to be expected since transforming the prediction of a continuous survival time into a binary classification is unlikely to yield classes that are easily separable. Furthermore, we acknowledge that the equivalent performance of hazard prediction and median survival classification might be due to generally low c-indices on this data set. For the low c-indices typically achieved for NSCLC studies, the predicted hazards might simply not be more precise than a median survival classification. Nevertheless, the simplifications due to reformulation as a classification problem show promising results and could increase the accessibility of survival analysis to deep learning techniques, especially if the results can be repeated on larger data sets with higher c-indices. Furthermore, the technique opens up survival analysis to high dimensional image data such as 3D+t MRI or CT images that require small batch sizes when combined with CNNs.

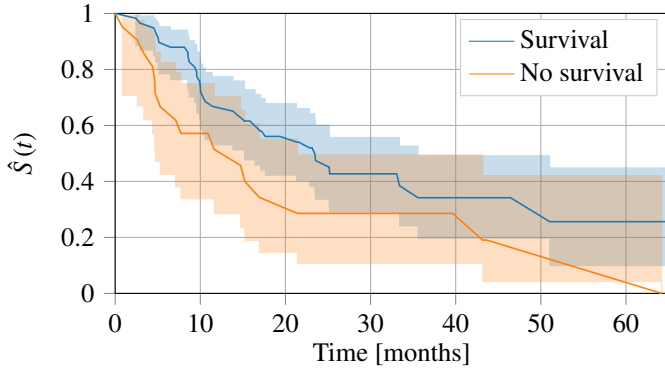


Figure 4: Stratification of patients from the NSCLC data set for median survival classification as predicted by the multi-modal median survival classifier. The split used has an AUROC value of 0.635 and is the closest to the mean of all AUROC values of the 100 splits of 0.636.

Fig. 4 shows the Kaplan-Meier estimator for the split with an AUROC value of 0.635 that is closest to the mean of 0.636. While this stratification divides the survival functions reasonably well for some shorter survival times, confidence intervals overlap strongly for longer survival times. With a test set of 78 samples and median survival classes being approximately balanced, this yields only ca. 39 samples per class, which further decreases with time as patients are lost to follow-up or die. Consequently, the confidence bounds are wide and this effect should be re-evaluated on a larger dataset in the future.

The label noise experiments reveal low predictive resolution of the predictions. If additive variances as high as 50 % of the actual label have no impact on the c-indices, this might either indicate that label noise has little effect on the model training since the ground truth labels are in fact as noisy as hypothesized, or that a difference in predictive performance can only be noticed for higher c-index data sets. Therefore, this method needs to be re-evaluated on data sets with a higher mean c-index as well. Another limitation of the CNN-based methods is the limited interpretability. While the model from [4] consists of only four mathematically-defined features and a Cox model, which is straightforward to interpret, CNNs are much harder to interpret.

5. Conclusion

We presented a method for survival prediction based on tomographic medical images. Our method can leverage trainable CNN features from CT image data, capturing abstract image information as well as clinical features in a single model. We show that by simplifying survival analysis to median survival classification, CNNs can be trained with small batch sizes and learn features that predict survival equally well as end-to-end hazard prediction networks and outperform the previous radiomics approach. This is a crucial step towards large scale image-based survival analysis that will allow survival prediction for more complex image data such as 3D+t images in the future.

Acknowledgements

No funding to declare.

References

References

- [1] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, K. Van Leemput, The multimodal brain tumor image segmentation benchmark (BRATS), *IEEE Transactions on Medical Imaging* 34 (2015) 1993–2024.
- [2] A. A. A. Setio, A. Traverso, T. de Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. van der Gugten, P. A. Heng, B. Jansen, M. M. de Kaste, V. Kotov, J. Y.-H. Lin, J. T. Manders, A. Sññora-Mengana, J. C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C. M. Schaefer-Prokop, E. T. Scholten, L. Scholten, M. M. Snoeren, E. L. Torres, J. Vandemeulebroucke, N. Walasek, G. C. Zuidhof, B. van Ginneken, C. Jacobs, Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge, *Medical Image Analysis* 42 (2017) 1 – 13.
- [3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, *Medical Image Analysis* 42 (2017) 60 – 88.
- [4] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, *Nature Communications* 5 (2014).
- [5] D. R. Cox, Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (1972) 187–220.
- [6] F. Harrell, *Regression Modelling Strategies*, 2 ed., Springer International Publishing, 2015. doi:10.1007/978-3-319-19425-7.
- [7] S. Leger, A. Zwanenburg, K. Pilz, F. Lohaus, A. Linge, K. Zöphel, J. Kotzerke, A. Schreiber, I. Tinhofer, V. Budach, A. Sak, M. Stuschke, P. Balermipas, C. Rödel, U. Ganswindt, C. Belka, S. Pigorsch, S. E. Combs, D. Mönnich, D. Zips, M. Krause, M. Baumann, E. G. C. Troost, S. Lück, C. Richter, A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling, *Scientific Reports* 7 (2017) 13206.
- [8] A. L. Simpson, A. Doussot, J. M. Creasy, L. B. Adams, P. J. Allen, R. P. DeMatteo, M. Gönen, N. E. Kemeny, T. P. Kingham, J. Shia, W. R. Jarnagin, R. K. G. Do, M. I. D’Angelica, Computed tomography image texture: A noninvasive prognostic marker of hepatic recurrence after hepatectomy for metastatic colorectal cancer, *Annals of Surgical Oncology* 24 (2017) 2482–2490.
- [9] M. A. Attiyeh, J. Chakraborty, A. Doussot, L. Langdon-Embry, S. Mainarich, M. Gönen, V. P. Balachandran, M. I. D’Angelica, R. P. DeMatteo, W. R. Jarnagin, T. P. Kingham, P. J. Allen, A. L. Simpson, R. K. Do, Survival prediction in pancreatic ductal adenocarcinoma by quantitative computed tomography image analysis, *Annals of Surgical Oncology* 25 (2018) 1034–1042.
- [10] K. Liestbl, P. K. Andersen, U. Andersen, Survival analysis and neural nets, *Statistics in Medicine* 13 (1994) 1189–1200.
- [11] J. Katzman, U. Shaham, J. Bates, A. Cloninger, T. Jiang, Y. Kluger, DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network, 2016. arXiv:1606.00931.

- [12] M. Luck, T. Sylvain, H. Cardinal, A. Lodi, Y. Bengio, Deep Learning for Patient-Specific Kidney Graft Survival Analysis, 2017. [arXiv:1705.10245](https://arxiv.org/abs/1705.10245).
- [13] C. Lee, W. R. Zame, J. Yoon, M. von der Schaar, DeepHit: A deep learning approach to survival analysis with competing risks, 2017. URL: http://medianetlab.ee.ucla.edu/papers/AAAI_2018_DeepHit.pdf.
- [14] X. Zhu, J. Yao, J. Huang, Deep convolutional neural network for survival analysis with pathological images, in: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016, pp. 544–547. doi:10.1109/BIBM.2016.7822579.
- [15] X. Zhu, J. Yao, F. Zhu, J. Huang, Wsisa: Making survival prediction from whole slide histopathological images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6855–6863. doi:10.1109/cvpr.2017.725.
- [16] J. Yao, X. Zhu, F. Zhu, J. Huang, Deep correlational learning for survival prediction from multi-modality data, in: Medical Image Computing and Computer Assisted Interventions (MICCAI), 2017, pp. 406–414. doi:10.1007/978-3-319-66185-8_46.
- [17] J. Lao, Y. Chen, Z.-C. Li, Q. Li, J. Zhang, J. Liu, G. Zhai, A deep learning-based radiomics model for prediction of survival in glioblastoma multi-forme, Scientific Reports 7 (2017) 10353.
- [18] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The cancer imaging archive (TCIA): Maintaining and operating a public information repository, Journal of Digital Imaging 26 (2013) 1045–1057.
- [19] H. J. W. L. Aerts, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, P. Lambin, Data from NSCLC-radiomics, in: The Cancer Imaging Archive, 2015. doi:10.7937/K9/TCIA.2015.PFOM9REI.
- [20] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer Research 77 (2017) e104–e107.
- [21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. doi:10.1109/cvpr.2016.90.
- [22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision 115 (2015) 211–252.
- [23] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2016, pp. 424–432. doi:10.1007/978-3-319-46723-8_49.
- [24] H. Steck, B. Krishnapuram, C. Dehing-oferije, P. Lambin, V. C. Raykar, On ranking in survival analysis: Bounds on the concordance index, in: Advances in Neural Information Processing Systems 20, Curran Associates, 2008, pp. 1209–1216.
- [25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, 2015. [arXiv:1512.00567](https://arxiv.org/abs/1512.00567).