

# Attention-based Multiple Instance Learning for Survival Prediction on Lung Cancer Tissue Microarrays

Jonas Ammeling<sup>1</sup>, Lars-Henning Schmidt<sup>2</sup>, Jonathan Ganz<sup>1</sup>, Tanja Niedermair<sup>3</sup>,  
 Christoph Brochhausen-Delius<sup>3</sup>, Christian Schulz<sup>4</sup>, Katharina Breininger<sup>5</sup>,  
 Marc Aubreville<sup>1</sup>

<sup>1</sup>Technische Hochschule Ingolstadt, Ingolstadt, Germany

<sup>2</sup>Medical Department IV, Pulmonary Medicine and Thoracic Oncology, Klinikum Ingolstadt, Ingolstadt, Germany

<sup>3</sup>Institute of Pathology, University of Regensburg, Regensburg, Germany

<sup>4</sup>Department of Internal Medicine II, University Hospital Regensburg, Regensburg, Germany

<sup>5</sup>Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[jonas.ammeling@thi.de](mailto:jonas.ammeling@thi.de)

**Abstract.** Attention-based multiple instance learning (AMIL) algorithms have proven to be successful in utilizing gigapixel whole-slide images (WSIs) for a variety of different computational pathology tasks such as outcome prediction and cancer subtyping problems. We extended an AMIL approach to the task of survival prediction by utilizing the classical Cox partial likelihood as a loss function, converting the AMIL model into a nonlinear proportional hazards model. We applied the model to tissue microarray (TMA) slides of 330 lung cancer patients. The results show that AMIL approaches can handle very small amounts of tissue from a TMA and reach similar C-index performance compared to established survival prediction methods trained with highly discriminative clinical factors such as age, cancer grade, and cancer stage.

## 1 Introduction

Despite advances in medicine over the past two decades, lung cancer remains the leading cause of cancer-related deaths worldwide [1]. Survival prediction methods are used for predicting time-to-event outcomes such as cancer recurrence or death and play an important role in clinical decision making in oncology. Recent survival prediction methods used tissue samples from whole sections [2, 3] to make predictions about a patient’s prognosis. However, with the trend toward minimally invasive biopsy techniques in the treatment of lung cancer, approaches that can process smaller amounts of tissue are needed [4]. TMAs provide the opportunity to explore such methods because they typically contain only a very small amount of tissue for each patient from the original tumor sample [5]. Bychkov et al. [6] used a combination of convolutional and recurrent architectures to predict five-year survival as a binary classification problem from TMA cores of colon cancer patients. However, recurrent architectures are more complex than attention-based approaches, which have been shown to be successful for predicting patient prognosis on whole-slide images (WSI) [2, 7]. To investigate the applicability of attention-based methods to TMA slides, we extended an AMIL approach originally

developed for weakly supervised classification to the survival prediction task similar to [2] by converting the classification problem into a regression problem by optimizing the modified partial Cox likelihood [8] as our loss function. We applied the model to TMA slides from patients with non-small cell lung cancer (NSCLC) and compared performance with established survival prediction methods trained on prognostically discriminative tabular data. The full code is available online<sup>1</sup>.

## 2 Materials and Methods

### 2.1 Data

Tissue samples, clinicopathologic features, and follow-up information were collected and examined from 379 NSCLC patients in the thoracic department of St. Georgs Klinikum in Ostercappeln, Germany. Clinical TNM staging (including clinical examination, CT scans, sonography, endoscopy, MRI, bone scan) was performed based on UICC/AJCC recommendations. The definite tumor staging was carried out post-surgically by pathological exploration. The histologic grade was determined according to WHO criteria on the original whole section tumor sample. The follow-up time was computed from the date of histological diagnosis to death or censored at the date of last contact. The TMAs were constructed by sampling with a 0.6 mm core needle from the most representative parts of the original tumor block. Three cores per patient were punched from the paraffin-embedded tumor block and assembled into TMA blocks. Multiple (at most 3x) four-micrometer-thick sections were cut from the TMA blocks, stained with hematoxylin and eosin (H&E), and digitized with a whole-slide scanner (Pannoramic P1000, 3D HISTECH Ltd., Budapest, Hungary) at a resolution of 0.24  $\mu\text{m}/\text{pixel}$ . The TMA cores were linked to a unique patient identifier and extracted from the whole-slide TMA image. Due to missing values in the tabular data and loss of tissue sections from the TMA blocks, 49 cases were excluded from the analysis. Thus, data from 330 NSCLC patients was included for further analysis.

### 2.2 Image Processing

To compare the use of the patient-level label for different tissue amounts, all cores from the same patient were processed once individually and stitched together once horizontally to create a new patient-level TMA (Fig. 1). All images were processed using the public CLAM [9] repository for WSI analysis. After tissue segmentation, non-overlapping patches of size  $256 \times 256$  were extracted at full resolution. Then, a ResNet50 model, pre-trained on ImageNet, was used to convert each patch into a 1024-dimensional feature vector by spatial average pooling after the third residual block.

### 2.3 Attention-based Multiple Instance Learning

In the context of multiple instance learning, each image is viewed as a collection of  $M$  patches or instances, known as a bag, with a corresponding bag-level label associated

---

<sup>1</sup>[https://github.com/DeepPathology/Cox\\_AMIL](https://github.com/DeepPathology/Cox_AMIL)

with it. After image processing, each bag is represented by the patch-level embeddings  $\mathbf{H} \in \mathbb{R}^{M \times C}$ , where  $C$  is the feature dimension from the ResNet50 model. The model consists of three components as shown in Figure 1: the projection layer  $f_{\text{proj}}$ , the attention module  $f_{\text{attn}}$ , and the prediction layer  $f_{\text{pred}}$ . The projection layer  $f_{\text{proj}}$  is a fully-connected layer with weights  $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{512 \times C}$  (all bias terms are implied for notational convenience) that maps the patch-level embedding into a more compact, dataset-specific 512-dimensional feature space. The attention module  $f_{\text{attn}}$  learns to assign a score to each patch based on its contribution to the patient's predicted prognosis. In particular,  $f_{\text{attn}}$  consists of three fully connected layers with weights  $\mathbf{U} \in \mathbb{R}^{256 \times 512}$ ,  $\mathbf{V} \in \mathbb{R}^{256 \times 512}$ , and  $\mathbf{Z} \in \mathbb{R}^{1 \times 256}$ . After projection, the attention score  $a_m$  for the  $m$ -th patch embedding  $\mathbf{h}_m \in \mathbb{R}^{512}$ , is computed by [9]:

$$a_m = \frac{\exp\{\mathbf{Z}(\tanh(\mathbf{V}\mathbf{h}_m^T) \odot \text{sigm}(\mathbf{U}\mathbf{h}_m^T))\}}{\sum_{m=1}^M \exp\{\mathbf{Z}(\tanh(\mathbf{V}\mathbf{h}_m^T) \odot \text{sigm}(\mathbf{U}\mathbf{h}_m^T))\}} \quad (1)$$

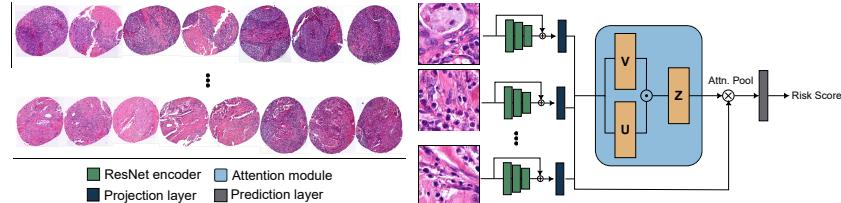
The computed attention scores for each patch are then used as weight coefficients to aggregate the patch-level embeddings into the bag representation  $\mathbf{h}_{\text{bag}} \in \mathbb{R}^{512}$  via attention-pooling [9] by:

$$\mathbf{h}_{\text{bag}} = \sum_{m=1}^M a_m \mathbf{h}_m \quad (2)$$

The final prediction layer  $f_{\text{pred}}$  is a single fully-connected layer with weights  $\mathbf{W}_{\text{pred}} \in \mathbb{R}^{1 \times 512}$  with a single output node and linear activation. Let  $\theta$  represent all weights of the network, then the model can be described as a function  $h_\theta : \mathbb{R}^{M \times C} \mapsto \mathbb{R}$ , where the output is a patient's hazard log risk score, described in more detail in the next section.

## 2.4 Loss Function

The right-censored patient-level survival data for the  $i$ -th patient consist of the triple  $(t_i, \delta_i, x_i)$  and represent the observed time, binary censoring status ( $\delta = 0$ , death not observed), and image data, respectively. Censoring is assumed to be non-informative, such that for a given  $x_i$ , survival time and censoring are independent. Let  $t_1 < t_2 < t_D$  be the ordered event times. The risk set  $\mathcal{R}(t_i)$  is defined as the set of all individuals



**Fig. 1.** Data stitching and overall architecture. Left: tissue-cores of the same patient stitched together horizontally. Right: overview of the attention-based multiple instance learning algorithm for survival prediction.

still in the study at a time immediately preceding  $t_i$ . The loss function is derived from the common Cox proportional hazards model where the hazard function of the form  $\lambda(t | x) = \lambda_0(t) \cdot e^{h(x)}$  is composed of the hazard baseline function  $\lambda_0(t)$ , and a risk score  $r(x) = e^{h(x)}$ . Our proposed model estimates a patient's log-risk score  $\hat{h}_\theta(x)$  parameterized by the weights of the network  $\theta$  such that the modified Cox partial likelihood becomes [8]:

$$L(\theta) = \prod_{i:\delta=1} \frac{\hat{r}_\theta(x_i)}{\sum_{j \in \mathcal{R}(t_i)} \hat{r}_\theta(x_j)} = \prod_{i:\delta=1} \frac{\exp(\hat{h}_\theta(x_i))}{\sum_{j \in \mathcal{R}(t_i)} \exp(\hat{h}_\theta(x_j))} \quad (3)$$

The loss function which is minimized by the network is then obtained from the average of the negative partial log likelihood with regularization as shown below:

$$l(\theta) = -\frac{1}{N} \sum_i \delta_i \left( \hat{h}_\theta(x_i) - \log \sum_{j \in \mathcal{R}(t_i)} \exp(\hat{h}_\theta(x_j)) \right) + \lambda \cdot \|\theta\|_1, \quad (4)$$

where  $N$  is the number of patients with an observable event and  $\lambda$  is the  $l_1$  regularization parameter. Intuitively, the loss function penalizes discordance between the scores of higher-risk and lower-risk patients. Similar loss functions have been used previously by other authors [7, 10]. Details about the training and hyperparameter settings are online<sup>1</sup>.

## 2.5 Baselines

The performance of the proposed model was compared against three established survival analysis methods based on tabular data. The tabular data used to train these baseline methods consist of patient characteristics (age, sex, and smoking status) and clinical characteristics (cancer stage and grade). It should be noted that the latter require elaborate determination and are commonly accepted to be highly prognostic. The first baseline method was a classical linear Cox proportional hazards (CPH) model [8]. The second was a random survival forest (RSF) [11], a non-proportional, flexible and robust alternative to the classical CPH model. The third baseline method was DeepSurv [10], a modern nonlinear, deep learning-based CPH model. Moreover, the proposed model was compared with a MIL method using classical max-pooling, and with the performance of an AMIL model trained on each patient core individually, using the respective patient-level label.

## 2.6 Evaluation

To evaluate and compare the predictive performance on the survival data, we performed a 10-fold cross-validation and measured Harrell's concordance index (C-index) [12]. The C-index indicates how well a model predicts the ranking of patients' death times, where large values of  $\hat{h}_\theta(x_i)$  should be associated with small values of  $t_i$  and vice versa. A value of  $C = 0.5$  corresponds to the average C-index of a random model, whereas  $C = 1$  corresponds to a perfect association. Patient stratification was assessed by assigning patients to either a high-risk or a low-risk group based on the median of the predicted

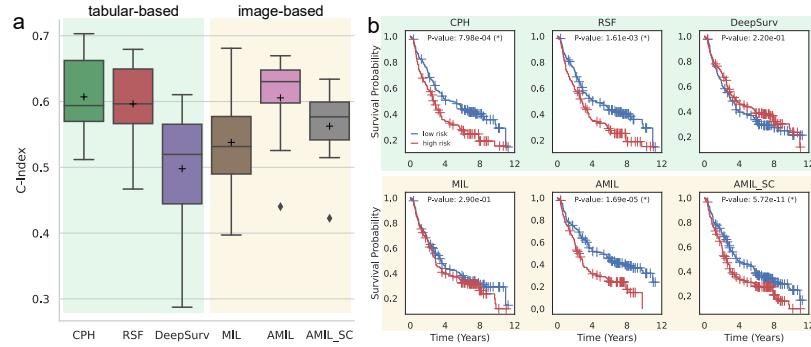
risk score. Kaplan-Meier curves were constructed by pooling the risk predictions of the test folds and plotting them against their survival time. Logrank tests were performed to check for statistically significant differences ( $P\text{-value} < 0.05$ ) between the two survival distributions.

### 3 Results

The cross-validated C-index values and the Kaplan-Meier curves for the patient stratification results are shown in Figure 2. The classical CPH model achieved the overall best performance with an average  $C\text{-index}$  of  $0.61 \pm 0.07$  ( $P\text{-value}: 7.98 \times 10^{-4}$ ) and a median  $C\text{-index}$  of 0.59. Closely similar, the RSF model achieved an average  $C\text{-index}$  of  $0.60 \pm 0.06$  ( $P\text{-value}: 1.61 \times 10^{-3}$ ) and a median  $C\text{-index}$  of 0.60. The DeepSurv method performed the worst among all methods with an average  $C\text{-index}$  of  $0.50 \pm 0.1$  ( $P\text{-value}: 2.20 \times 10^{-1}$ ) and a median  $C\text{-index}$  of 0.53. Among the image-based methods, the classical MIL model with max-pooling performs the worst with an average  $C\text{-index}$  of  $0.54 \pm 0.08$  ( $P\text{-value}: 2.90 \times 10^{-1}$ ) and a median  $C\text{-index}$  of 0.53. The AMIL model performed the best among the image-based methods with an average  $C\text{-index}$  of  $0.61 \pm 0.07$  ( $P\text{-value}: 1.69 \times 10^{-5}$ ) and a median  $C\text{-index}$  of 0.63. The AMIL model trained on individual cores performed worse than the AMIL model trained using the patient label for all cores together, with an average  $C\text{-index}$  of  $0.56 \pm 0.06$  ( $P\text{-value}: 5.72 \times 10^{-11}$ ) and a median  $C\text{-index}$  of 0.58.

### 4 Discussion

The results show that the proposed model performed similarly well to established survival prediction methods based on tabular data, such as CPH and RSF, by extracting features directly from the small amount of tissue provided by the TMA. Attention pooling proved to be very effective in this regard for aggregating these features to estimate the patient's risk score. However, the experiments also showed that the prognostically



**Fig. 2.** a) Cross-validated C-index performance. + indicates mean value. b) Kaplan-Meier curves for patient stratification results on the pooled validation splits.

relevant information is not necessarily evenly distributed across a patient's tissue cores when selected from the original sample, as indicated by lower average C-index when trained on individual cores. Thus, performance is highly dependent on tissue core selection, and it is recommended to use all available tissue cores. Nevertheless, the results show that attention-based MIL approaches are applicable to TMA slides where only a small amount of tissue per patient is available to perform reliable patient stratification, as shown by the statistically significant logrank test results. Therefore, in the context of the trend toward minimally invasive biopsy procedures in lung cancer treatment, these methods could become a part of reliable decision support systems in the future.

**Acknowledgement.** J. A. acknowledges funding by the Bavarian Institute for Digital Transformation (Project ReGInA).

## References

1. Wang M, Herbst RS, Boshoff C. Toward personalized treatment approaches for non-small-cell lung cancer. *Nat Med.* 2021;27:1345–56.
2. Chen RJ, Lu MY, Williamson DF, Chen TY, Lipkova J, Noor Z et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell.* 2022;40:865–878.e6.
3. Vale-Silva LA, Rohr K. Long-term cancer survival prediction using multimodal deep learning. *Sci Rep.* 2021;11:13505.
4. Coley SM, Crapanzano JP, Saqi A. FNA, core biopsy, or both for the diagnosis of lung carcinoma: Obtaining sufficient tissue for a specific diagnosis and molecular testing. *Cancer Cytopathol.* 2015;123:318–26.
5. Schmidt LH, Biesterfeld S, Kümmel A, Faldum A, Sebastian M, Taube C et al. Tissue microarrays are reliable tools for the clinicopathological characterization of lung cancer tissue. *Anticancer res.* 2009;29:201–9.
6. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE, Verrill C et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep.* 2018;8:1–11.
7. Yao J, Zhu X, Jonnagaddala J, Hawkins N, Huang J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med Image Anal.* 2020;65:101789.
8. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Series B Stat Methodol.* 1972;34:187–202.
9. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng.* 2021;5:555–70.
10. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol.* 2018;18:24.
11. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat.* 2008;2(3):841–60.
12. Harrell FE, Calif RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. *JAMA.* 1982;247:2543–6.