# Towards Explainable End-to-End Prostate Cancer Relapse Prediction from H&E Images Combining Self-Attention Multiple Instance Learning with a Recurrent Neural Network

**Esther Dietrich**[1] ESTHER.DIETRICH@ZMNH.UNI-HAMBURG.DE
**Patrick Fuhlert**[1] PATRICK.FUHLERT@ZMNH.UNI-HAMBURG.DE
**Anne Ernst**[1] ANNE.ERNST@ZMNH.UNI-HAMBURG.DE
**Guido Sauter**[2] G.SAUTER@UKE.DE
**Maximilian Lennartz**[2] M.LENNARTZ@UKE.DE
**H. Siegfried Stiehl**[3] HANS-SIEGFRIED.STIEHL@UNI-HAMBURG.DE
**Marina Zimmermann**[1,4*] MARINA.ZIMMERMANN@ZMNH.UNI-HAMBURG.DE
**Stefan Bonn**[1*] SBONN@UKE.DE
[1] *Institute of Medical Systems Biology, Center for Biomedical AI (bAIome)* [2] *Institute of Pathology* [4] *III. Department of Medicine – University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
[3] *Department of Informatics, Universität Hamburg, Hamburg, Germany*

## Abstract

Clinical decision support for histopathology image data mainly focuses on strongly supervised annotations, which offers intuitive interpretability, but is bound by expert performance. Here, we propose an explainable cancer relapse prediction network (eCaReNet) and show that end-to-end learning without strong annotations offers state-of-the-art performance while interpretability can be included through an attention mechanism. On the use case of prostate cancer survival prediction, using 14,479 images and only relapse times as annotations, we reach a cumulative dynamic AUC of 0.78 on a validation set, being on par with an expert pathologist (and an AUC of 0.77 on a separate test set). Our model is well-calibrated and outputs survival curves as well as a risk score and group per patient. Making use of the attention weights of a multiple instance learning layer, we show that malignant patches have a higher influence on the prediction than benign patches, thus offering an intuitive interpretation of the prediction. Our code is available at www.github.com/imsb-uke/ecarenet.

**Keywords:** Computational Pathology, Gated Recurrent Units, Multiple Instance Learning, Prostate Cancer, Recurrent Neural Network, Self Attention, Survival Prediction, Explainability

## 1. Introduction

In recent years deep learning has greatly improved the performance in computer vision tasks for medical applications (Esteva et al., 2021). In particular, decision support systems for cancer treatment in the field of computational pathology are emerging (Abels et al., 2019). Many systems rely on physicians' annotations like treatment decisions, manual annotation of tissue regions or patient classification according to a staging system (Bulten et al., 2020). This strong supervision limits the models' performance through subjectivity and thus ambiguity of the ground truth, and emphasizes the need for quantifiable labels that are independent of the physician, such as time to disease-related death or relapse. Difficulties arise as such labels are relatively weak (a single survival time per patient), thus requiring a large dataset for training, and include censored cases as not all patients relapse or die of the disease. A survival model, unlike a classification model, models the patient's survival over time and can include censored patients.

The aim of this work is to show that high predictive power can be achieved when training end-to-end only with quantitative patient relapse times, while having a majority of censored cases. We focus on prostate cancer as a use case, which was the cancer with the second most new cases in men worldwide in 2020 (Ferlay et al., 2020). Instead of predicting

[*] Equally contributing last authors

Gleason grades to estimate cancer severity (Gleason and Mellinger, 1974), which are highly controversial, often revised and suffer from interobserver variability of up to 81% (Egevad et al., 2012), we use time to biochemical recurrence (BCR) as annotation. This is defined as a significant rise in prostate specific antigen (PSA) level in the blood after prostatectomy. As input, digitized hematoxylin and eosin (H&E) stained tissue microarray (TMA) spots are used, of which a dataset containing 14,479 images is available.

To the best of our knowledge, we are the first to propose an explainable end-to-end deep learning model to predict BCR over time after prostatectomy from TMA spots. We introduce a novel network based on self-attention (Rymarczyk et al., 2021), attention-based multiple instance learning (MIL, Ilse et al. (2018)) and recurrent neural networks (RNNs, Rumelhart and McClelland (1987)) for survival prediction, called eCaReNet (explainable cancer relapse prediction network). With an AUC of 0.78 on the validation set (0.77 on the test set) we achieve state-of-the-art results, while assuring calibration. Through evaluation of attention weights of the MIL layer, we further show that our model weights malignant patches higher than benign patches. In general, our approach is applicable to various cancer and non-cancer histopathology survival prediction problems.

This work is organized as follows. Section 2 gives an overview on related work. In Section 3, the available data is described. The details of our model can be found in Section 4, followed by a discussion of the results in Section 5, including a comparison to benchmark models and a pathologist.

## 2. Related Work

Image-based decision support systems often aim at reproducing classification systems used in clinical practice (Bulten et al., 2020). Only after classification Arvaniti et al. (2018) and Nagpal et al. (2019) correlate findings to patient survival. Disadvantages are time-consuming annotations and label quality limited by the annotator. However, this classification allows for an improved interpretability of progression prediction. If only a weak label for a whole image is available, MIL approaches as proposed by Ilse et al. (2018) can be applied to analyze which image regions have the highest influence on a model's prediction. Couture et al. (2018) for example integrate MIL methods in their model for risk of recurrence prediction from image patches, which is also a clinical score. Especially when analyzing whole slide images, MIL approaches are often used. Yao et al. (2020) predict a single risk score per patient with attention-based MIL, while (Campanella et al., 2019) use MIL for binary tumor classification. Lu et al. (2021) extend MIL to multi-class classification by implementing multiple attention branches. Image regions relevant for the diagnosis are indicated by high attention weights.

Human performance can be improved upon if disease progression is modeled based on patient outcome directly. A binary classification of whether a patient has a relapse before a certain point in time is often applied, but is diagnostically less conclusive (Duanmu et al., 2020; Yamamoto et al., 2019). Wulczyn et al. (2020) treat survival prediction as a multi-class problem with the goal to correctly classify the interval of the event and output a risk score.

To predict relapse probability over time, a survival analysis model can be used. One option is the Cox model (Cox, 1972), where the linear part can be replaced by a neural network, as proposed in DeepSurv by Katzman et al. (2018) and its counterpart for images DeepConvSurv by Zhu et al. (2016). Especially in histopathology, often a complex feature extraction step is applied prior to the Cox model (Yao et al., 2020; Tang et al., 2019; Zhang et al., 2021). Furthermore, the Cox model is limited by the proportional hazard assumption, which enforces hazard rates to be constant over time. Xiao et al. (2020) and Vale-Silva and Rohr (2021) avoid the proportional hazard assumption as well as annotation-expensive preprocessing steps by developing end-to-end deep learning models. The latter however also include electronic health record and omics data to improve performance and neither includes an explainability mechanism, treating the model as a black box. Ren et al. (2019) and Giunchiglia et al. (2018) include recurrent layers to model time dependency, but only use patient electronic health records. A different approach is applied by Yala et al. (2021), who predict risk of cancer over time from mammography images by converting the prediction into a classification across multiple time points.

We propose a novel framework named eCaReNet for explainable end-to-end relapse prediction by exploiting the advantages of different works.

## 3. Dataset

Two datasets were provided by the local pathology department. All images in our datasets show prostate tissue obtained after prostatectomy, during which the patient's prostate is removed. Multiple tissue samples are then taken with a hollow needle, resulting in tissue cores of 0.6 mm diameter each. Arranged in a TMA, multiple samples from multiple patients are stained at once with H&E and digitized afterwards. Small differences, or biases, in staining intensity between TMA blocks arise due to e.g. staining times (Parsons and Grabsch, 2009).

The survival dataset (see Table 1) comprises 39 digitized TMAs with 129 to 522 images each. For these images, besides the time to BCR and the censoring status, the integrative quantitative (IQ-) Gleason and International Society of Urological Pathology (ISUP) scores (Sauter et al., 2018; Egevad et al., 2012) of the whole prostate are labeled (for details on Gleason grading see Appendix A.1). In this context it is important to note that the Gleason scores are based on the whole prostate, while in our dataset the image per patient only represents a small part of prostate tissue. Since the TMA spot image can only cover a very small part of the prostate, and annotations for individual images are missing, it is possible that a given image is not fully representative of a patient's disease status and outcome. In order to remove the most extreme of those cases, per-image Gleason scores are predicted and compared to the annotated overall Gleason score. Images that are predicted as non-cancerous – but have a high overall IQ Gleason annotation, a PSA value $> 4 \frac{ng}{ml}$ and a relapse within 2 years – are removed from the dataset, as these are considered unrepresentative and expected to reduce the model's generalizability. Other discrepancies between the images and relapse times are left unchanged.

Images from all but one TMA are shuffled and split into training, validation and test sets, stratified by prostate Gleason annotation. One TMA is left out as a separate test set to evaluate model performance on a set with a unique staining bias. Table 1 summarizes the number of images per dataset split. The distribution of event times for censored and uncensored patients in the training and validation datasets is shown in Figure A.1.

We pretrained our model on a second, smaller dataset (Gleason dataset), which also contains TMA spots, but is annotated with image-level Gleason and ISUP scores (see Figure C.1A). It includes 1930 images in the training, 417 in the validation and 419 in the test set.

For clinical practice, an estimate of the relapse time is of interest prior to prostatectomy. Since tissue samples obtained through needle biopsy are visually similar to post-prostatectomy tissue cores, the dataset can simulate such biopsies. The Gleason labels are annotated following the convention for biopsies. Preprocessing and data augmentation steps are detailed in Appendix A.3.

Table 1: Overview of number of images for each split in the survival dataset. 80% of patients are censored. $c = 0$:uncensored, $c = 1$:censored, valid.:validation.

| | training | valid. | test | single TMA test |
|---|---|---|---|---|
| $c=0$ | 1965 | 445 | 429 | 36 |
| $c=1$ | 8023 | 1698 | 1742 | 141 |
| total | 9988 | 2143 | 2171 | 177 |

## 4. Methods

### 4.1. Survival prediction

The following is derived according to Kvamme et al. (2019). For a patient with relapse at time $t^*$, the probability to be event-free at time $t$ is modeled via the survival function

$$S(t) = P(t^* > t). \tag{1}$$

The risk of the event to occur in the interval at time $t + \Delta t$, given that it did not occur until time $t$, is expressed with the hazard rate

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq t^* < t + \Delta t | t^* \geq t)}{\Delta t}. \tag{2}$$

A well-established method for modeling survival functions of individual patients via the hazard rate is the Cox model (Cox, 1972). It is limited by the proportional hazard assumption, which assumes that the hazard is constant over time and equal for all patients, therefore not allowing for crossing survival curves.

In order not to be constrained by this assumption, we directly model the individual hazard functions with a neural network. The time is discretized into time intervals $t_j \in (t_0, ..., t_k)$ and discrete versions of survival and hazard functions are defined as

$$S(t_j) = P(t^* > t_j), \tag{3}$$

$$h(t_j) = P(t^* = t_j | t^* > t_{j-1}), \tag{4}$$

$$S(t_j) = \prod_{k=0}^{j} (1 - h(t_k)). \tag{5}$$

The survival function is a monotonically decreasing function, as can be seen from Equation 5.

An important characteristic of survival data is censoring. Not all patients in the dataset experience an event, either because they are lost to follow-up, their event occurs after the end of documentation or they never relapse. These patients are right-censored and here $t^*$ is not the time of the event, but the last observed time without any event.

## 4.2. Model

As a base model for our proposed survival prediction an InceptionV3 network (Szegedy et al., 2015), pretrained on the ImageNet dataset (Russakovsky et al., 2015), is chosen, while replacing the last layers to perform survival prediction as described below. We chose InceptionV3 as it achieved best results in our experiments. We include two preceding steps (4.2.1 and 4.2.2), before training our survival model eCaReNet in a third step. Figure C.1 shows an overview of the presented models and which datasets these are trained on.

### 4.2.1. $\mathrm{M}_{\mathrm{ISUP}}$

In the first step, we additionally pretrain the InceptionV3 model to adapt it to our histopathology domain. Our model $\mathrm{M}_{\mathrm{ISUP}}$ takes images from the Gleason dataset as input (Figure C.1A), downsized with bilinear interpolation to $1024 \times 1024$ pixels, and classifies these into one out of six classes (benign or one of 5 malignant ISUP classes). During training, a cross-entropy loss is used. For training details and results, see Appendix B.

### 4.2.2. $\mathrm{M}_{\mathrm{Bin}}$

In the second step, a binary classification model $\mathrm{M}_{\mathrm{Bin}}$ is used to predict relapse within 2 years on the survival dataset (Figure C.1B). 2 years was chosen, as it lies close to the median (26.8 months) of the relapse times (44% of relapses earlier than 2 years). For this, we took the model $\mathrm{M}_{\mathrm{ISUP}}$ and modified the output to 2 classes. The input image is resized to $1024 \times 1024$ pixels as in $\mathrm{M}_{\mathrm{ISUP}}$ and a cross-entropy loss is applied during training. As opposed to the first step, the prediction per image is saved and used in the third step, which is the survival prediction model eCaReNet, shown in Figure 1.

### 4.2.3. eCaReNet

Each image of the survival dataset is cut into square, non-overlapping patches as input to eCaReNet (64 patches with $256 \times 256$ pixels each, see also Section 5). As this model predicts the hazard over time, one output node per time interval is needed. We chose 28 intervals to cover a time span of 7 years with intervals of 3-months length, covering the 90% of relapses that occur prior to 7 years. For eCaReNet, only the first 4 inception blocks of $\mathrm{M}_{\mathrm{ISUP}}$ are used to reduce overfitting. The following global average pooling layer reduces the dimensionality. Then a self-attention block, as proposed by Rymarczyk et al. (2021), models the influence of each patch across all other patches. Next, the aforementioned binary classification is concatenated with the output vector of the self-attention layer. This concatenated vector is repeated 28 times to model the discrete time intervals. The current time step is concatenated to each of these vectors. A gated recurrent unit (GRU) layer (Cho et al., 2014) models the temporal dependency of the hazard rate in the output, as proposed by Ren et al. (2019). At the end, an attention-based MIL-layer weights the predictions per patch and outputs a prediction per image, as proposed in Ilse et al. (2018).

An individual survival curve per patient is obtained through Equation 5. Using the normalized area under the survival curve, the patient's overall risk is estimated. Since a large area under the survival curve indicates a low risk $r$ and vice versa, the normalized area is subtracted from one:

$$r = 1 - \frac{1}{t_k} \sum_{i=1}^{k} S(t_i) \cdot |t_i - t_{i-1}|, \tag{6}$$

with the last interval $k$ at time $t_k$ (based on the survival time prediction in Xiao et al. (2020)). Since the risk score is a single numerical value between 0 and 1, it eases comparison among patients.

As proposed by Kvamme et al. (2019), during training a maximum likelihood loss is optimized. It differs for censored ($c = 1$) and uncensored ($c = 0$) patients with the observed event time $t^*$. For uncensored patients, the loss $L_u$ can be defined by the

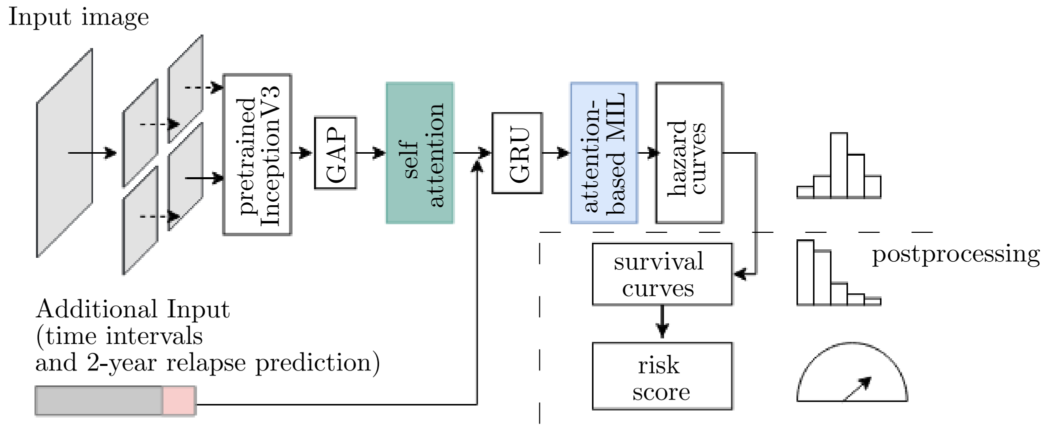


Figure 1: The survival model eCaReNet consists of a pretrained InceptionV3 base model, followed by global average pooling (GAP), a self-attention layer, a recurrent layer with gated recurrent units (GRU) and multiple instance learning (MIL) to combine results per patch. As output, a survival curve as well as a risk score are calculated per patient. The input image needs to be cut into regular patches. As additional information, a prediction whether the relapse occurs in the first two years is used and a time grid is included. The influence of the colored parts is evaluated in Section 5.

known hazard, whereas for censored patients the loss $L_c$ can be defined with the survival function:

$$L_u = \sum_{c=0}[log(h(t^*)) + \sum_{t_i:t_i<t^*} log(1-h(t_i))], \tag{7}$$

$$L_c = \sum_{c=1}[\sum_{t_i:t_i\leq t^*} log(1-h(t_i))] \tag{8}$$

$$= \sum_{c=1}[log(S(t^*))], \tag{9}$$

$$L = \alpha L_u + (1-\alpha L_c). \tag{10}$$

Both censored and uncensored patients' losses are linearly combined and equally weighted with $\alpha = 0.5$. As labels, the survival and hazard are defined as described in Section 4.1. Since only a discrete event time is known, from Equation 3 it follows that $S(t_j) = 1 \; \forall \; t_j < t^*$ and $S(t_j) = 0 \; \forall \; t_j \geq t^*$. For the hazard function, $h(t_j) = 0 \; \forall \; t_j < t^*$ and $h(t^*) = 1$ hold. The hazard function is not defined for $t_j > t^*$.

### 4.3. Metrics

To evaluate a survival model, both discrimination and calibration need to be considered. Discrimination estimates whether patients are ranked in the correct order, whereas calibration measures how well the predicted survival curves match with the ground truth.

To evaluate discrimination, we use the cumulative dynamic area under the curve (c/d AUC),

$$\begin{aligned} \text{c/d AUC}(t) = P(S_i(t) < S_j(t) | t_i^* \leq t, t_j^* > t) \\ + 0.5 P(S_i(t) = S_j(t) | t_i^* \leq t, t_j^* > t), \end{aligned} \tag{11}$$

which is further integrated over time and weighted by the Kaplan-Meier estimate to account for censored and uncensored patients. Details can be found in Blanche et al. (2019). With the c/d AUC the order of the patients' survival probabilities are compared at multiple discrete time points $t$. Censored patients are only comparable to patients with a known survival time that is shorter than the time of censoring. Perfect order results in a measure of 1 (Blanche et al., 2019). To improve readability, we refer to the c/d AUC as AUC in the following.

In the literature, the concordance, or c-index, is more commonly used (Blanche et al., 2019). This measure also ranges between 0 and 1, with 1 being perfect discrimination. However, the c-index is not a proper scoring rule, meaning that the underlying data generation process does not necessarily give the best score (Gerds and Kattan, 2021). The Brier score combines calibration and discrimination (Brier, 1950), as it measures the mean squared error between the ground truth survival curve and the predicted survival curve. A model that reaches a Brier score below 0.25 is considered to be meaningful (Gerds and Kattan, 2021).

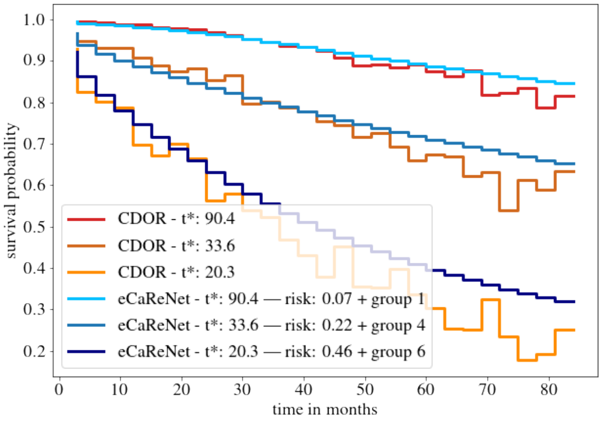


Figure 2: Example survival curves for 3 patients of the survival test dataset, predicted with eCaReNet (blue) and CDOR (orange-red). Both models predict the order of the patients correctly. The survival curves predicted by our model are monotonically decreasing, in contrast to CDOR. $t^*$: time of BCR in months. For eCaReNet, also the predicted risk and risk group are indicated.

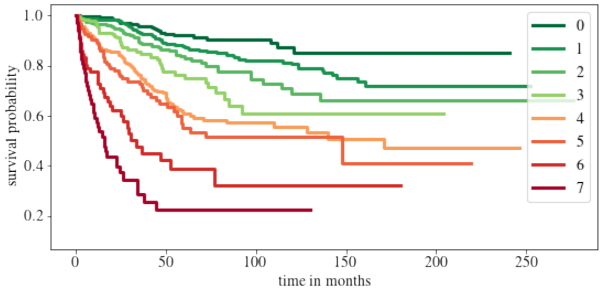


Figure 3: Kaplan-Meier curves for separate risk groups on the test set. The majority of groups separate well, only the log-rank tests between groups 2/3 as well as 4/5 fail with p-values 0.07 and 0.31 respectively.

In an ideal case, the predicted survival curve would be compared to the true survival probability over time, but this cannot be observed. To evaluate how meaningful the resulting survival curves for single patients are, the d-calibration is introduced by Haider et al. (2020). The idea behind this is to verify that the predicted survival probability at time $t$ matches the true probability of surviving up to time $t$. The d-calibration is calculated by comparing the number of patients that relapse while having a certain predicted survival probability to the expected number. D-calibration is measured with a chi-square test, that needs to pass. For details, see Appendix D.
The Brier score evaluates individual patients' predictions, while the other metrics are only applicable to whole populations. These are thus best used for comparisons between model performances on the same data, not across datasets (Gerds and Kattan, 2021).

## 5. Experiments and Results

All models are trained with the Nadam optimizer (Dozat, 2016) on the training set and the model with the smallest loss $L$ on the validation set is chosen for evaluation. 5 training runs are performed per model with different random seeds for weight initialization to avoid initialization bias. The models are implemented in Tensorflow 2.1 in Python 3.6. Training is performed on an NVIDIA Quadro RTX 8000 GPU with 48 GB memory. As the focus of this work is on survival prediction, the results for the pretraining on ISUP scores are provided in Appendix B.

### 5.1. Benchmark

As benchmark, we compare eCaReNet to architectures and loss functions proposed in the literature as well as to an expert pathologist's annotations. Results are summarized in Table 2, where higher AUC and c-index, but lower Brier score indicate better model performance. For d-calibration, only pass or failure of the chi-square test is indicated. We start by comparing eCaReNet to two models proposed in the literature. First, we retrain our pretrained $\mathrm{M_{ISUP}}$ (see Section 4.2.1) with the Cox loss and output as proposed in DeepConvSurv by Zhu et al. (2016). To do so, the output needs to be reduced to only one node. That model reaches an AUC of 0.69 (c-index of 0.65) on the validation and 0.71 (0.64) on the test set. The test for d-calibration fails and also the Brier score of 0.305 (0.296 on the test set) indicates a non-calibrated model. As second comparison, we train $\mathrm{M_{ISUP}}$ with the output structure and loss proposed in censoring-aware deep ordinal regression (CDOR) by Xiao et al. (2020). That model reaches an AUC of 0.77 on the validation and 0.78 on the test set and a c-index of 0.73 for both sets, but also fails in terms

Table 2: Benchmark of our proposed approach. Values are the mean of 5 training runs with the standard deviation in parentheses. For d-calibration only failure (f) or pass (p) is indicated.

| **Validation set** | **AUC** | **Brier** | **C-index** | **D-calibration** |
|---|---|---|---|---|
| ISUP | **0.78** | - | **0.75** | - |
| DeepConvSurv (Zhu et al., 2016) | 0.69 (0.0207) | 0.305 (0.0146) | 0.65 (0.0173) | f |
| CDOR (Xiao et al., 2020) | 0.77 (0.0089) | 0.111 (0.0014) | 0.73 (0.0046) | f |
| **eCaReNet** | **0.78** (0.0041) | **0.107** (0.0004) | **0.75** (0.0016) | **p** |
| **Test set** | | | | |
| ISUP | **0.80** | - | **0.76** | - |
| DeepConvSurv | 0.71 (0.0232) | 0.296 (0.0227) | 0.64 (0.0132) | f |
| CDOR | **0.78** (0.0005) | 0.110 (0.0001) | 0.73 (0.0003) | f |
| **eCaReNet** | 0.77 (0.0048) | **0.109** (0.0006) | **0.74** (0.0037) | **p** |

of calibration. Furthermore, the resulting survival curves are not monotonically decreasing, therefore biologically unreasonable (see Figure 2).

Compared to both previously described models, eCaReNet shows the best performance for all measures on the validation set (AUC 0.78, Brier score 0.107, c-index 0.75) and passes the chi-square test for d-calibration. On the test set it also obtains the best Brier score (0.109) and c-index (0.74) and passes the chi-square test for d-calibration. CDOR performs best on the test set's AUC. In contrast to CDOR, eCaReNet outputs monotonically decreasing survival functions (see Figure 2).

In addition, we assign individual patients to 8 risk groups, to enable a relative ranking as detailed in Appendix E. Risk groups further allow the evaluation of Kaplan-Meier curves (Kaplan and Meier, 1958) with a log-rank test, which is common in survival analysis (Li et al., 2015). Kaplan-Meier curves are calculated for the risk groups on the training, validation and test datasets. Overall, we can show that the risk groups stratify well on all sets. The results for the test set are shown in Figure 3, where five out of seven log-rank tests pass ($p < 0.05$).

Furthermore, we compare eCaReNet to annotations of an expert pathologist. In clinical practice, pathologists do not estimate relapse times for patients directly, but assign a Gleason score. We compare eCaReNet's discrimination power to the assigned ISUP scores, since a higher ISUP score corresponds to an increased risk of relapse. eCaReNet reaches on par performance in terms of AUC and c-index with the pathologist's annotations on the validation set (AUC 0.78 and c-index 0.75). Only on the test set, the ISUP annotation shows higher AUC and c-index. In contrast to our model that uses a single TMA spot image per patient, for the ISUP annotation the whole prostate tissue was available, giving Gleason-based survival prediction an advantage over model-based prediction.

## 5.2. Comparison of model adaptations

In the following, eCaReNet (see Figure 1) is adapted to evaluate which parts contribute most to model discrimination power and calibration (see Table 3). As base model $\mathrm{M}_{\mathrm{base}}$, the first 4 blocks of an InceptionV3 model, pretrained on the ImageNet dataset, are extended with a GRU layer for survival prediction. The following adaptations are included gradually. As first adaptation ($\mathrm{M}_{\mathrm{pretr}}$), a retraining of the InceptionV3 on Gleason classes as described in $\mathrm{M}_{\mathrm{ISUP}}$ is applied (see Section 4.2.1). The next adaptation is model $\mathrm{M}_{\mathrm{MIL}}$, which processes image patches. Here, an attention-based MIL layer is added to the previous model (blue part in Figure 1). For model $\mathrm{M}_{\mathrm{MIL\text{-}Bin}}$, the binary relapse prediction in $\mathrm{M}_{\mathrm{Bin}}$ from Section 4.2 is added (red part in Figure 1). The last evaluated model is eCaReNet, where additionally a self-attention layer is included (green part in

Table 3: Comparison of model adaptations. Values are the mean of 5 training runs with the standard deviation in parentheses for the validation (Valid.) and test sets. When models $M_{\mathrm{ISUP}}$ or $M_{\mathrm{Bin}}$, or MIL or self-attention (sa) layers are included, it is indicated with a dot (●). Best results are marked in bold. MIL=multiple instance learning, Bin=including binary relapse prediction from $M_{\mathrm{Bin}}$. For d-calibration (D-cal.) only failure (f) or pass (p) is indicated.

| **Valid. set** | $\mathbf{M_{ISUP}}$ | **MIL** | $\mathbf{M_{Bin}}$ | **sa** | **AUC** | **Brier** | **C-index** | **D-cal.** |
|---|---|---|---|---|---|---|---|---|
| $M_{\mathrm{base}}$ | | | | | 0.74 (0.0042) | 0.116 (0.0038) | 0.72 (0.0008) | **p** |
| $M_{\mathrm{pretr}}$ | ● | | | | 0.76 (0.0018) | 0.109 (0.0005) | 0.73 (0.0023) | **p** |
| $M_{\mathrm{MIL}}$ | ● | ● | | | 0.76 (0.0004) | 0.109 (0.0000) | 0.74 (0.0000) | **p** |
| $M_{\text{MIL-Bin}}$ | ● | ● | ● | | 0.77 (0.0012) | **0.107** (0.0003) | 0.74 (0.0026) | **p** |
| **eCaReNet** | ● | ● | ● | ● | **0.78** (0.0041) | **0.107** (0.0004) | **0.75** (0.0016) | **p** |
| **Test set** | | | | | | | | |
| $M_{\mathrm{base}}$ | | | | | 0.74 (0.0054) | 0.115 (0.0007) | 0.71 (0.0031) | **p** |
| $M_{\mathrm{pretr}}$ | ● | | | | 0.76 (0.0031) | 0.110 (0.0004) | 0.73 (0.0018) | **p** |
| $M_{\mathrm{MIL}}$ | ● | ● | | | 0.76 (0.0002) | 0.110 (0.0000) | **0.74** (0.0003) | **p** |
| $M_{\text{MIL-Bin}}$ | ● | ● | ● | | **0.77** (0.0011) | **0.109** (0.0003) | **0.74** (0.0022) | **p** |
| **eCaReNet** | ● | ● | ● | ● | **0.77** (0.0048) | **0.109** (0.0006) | **0.74** (0.0037) | **p** |

Figure 1), to account for inter-patch influences. For model $M_{\mathrm{MIL}}$ and $M_{\text{MIL-Bin}}$, the experiments showed best results using 16 patches of size 512 × 512 pixels, whereas 64 patches with 256 × 256 pixels each lead to best results when including self-attention.

All results are summarized in Table 3. It can be seen that the pretraining on histopathology images has a positive effect on all metrics. Adding MIL further improves the discrimination on the validation dataset. Best results are achieved when adding the 2-year relapse prediction and self-attention, reaching performance rivaling that of expert pathologists (AUC validation set: 0.78, test set: 0.77). However, the model with self-attention shows a slightly higher variance in the results than $M_{\text{MIL-Bin}}$. It is concluded that the inter-dependence of patches does not add much additional information to the prediction, as both model versions with and without self-attention show similar performance on the test set. The Brier score is similar for all models and best also for the variants that include a binary relapse prediction. For all our models, the d-calibration chi-square test passes, assuring calibration. Furthermore, the models generalize well, as there is only a slight performance drop when evaluating on the test set. Evaluation of the results on the separate test set, only containing a single TMA, also results in AUC scores of 0.74-0.76 for all adaptations.

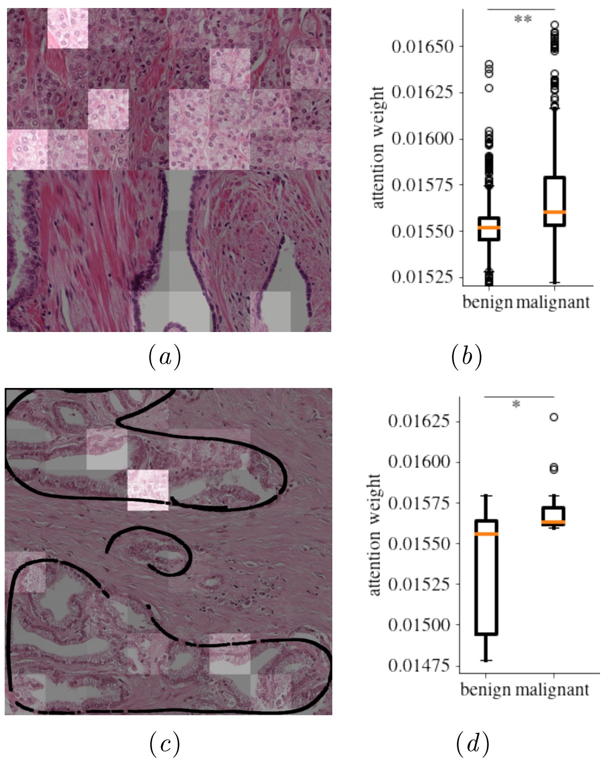
### 5.3. Evaluation of attention weights

To apply a model in clinical practice, an accurate performance on test data is not sufficient. Physicians can only benefit from a support system if the decisions can be explained and interpreted, with the terms ‘explainability’ or ‘interpretability’ having many different and non-standardized meanings in the literature (Barredo Arrieta et al., 2020). In this paper, we include explainability by computing the attention weights of the MIL layer and showing which image regions have the highest influence on the prediction. It is expected that malignant patches show higher attention weights than patches with benign tissue.

In a first analysis we use the Gleason dataset to create an artificial dataset for which the annotation per image patch is known. Each image in this dataset combines one image showing benign tissue and one image with malignant Gleason grade 5 tissue by stitching half of each together (see example in Figure 4($a$)). For each image, the attention weights per patch are extracted from the MIL layer of eCaReNet. In the example it can be seen that the upper, malignant part, receives the highest attention weights, while in the benign tissue only relatively bright regions are highlighted. This may be because white regions correspond to glands, which are an important structure to distinguish benign from malignant tissue (see also Figure F.1). A boxplot of the attention weights of all 12 example images is shown in Figure 4(*b*). The attention weights for malignant patches are significantly higher than for benign patches. The original images that were stitched together are neither part of the training nor of the validation or test sets and give an unbiased estimate of importance.

Another experiment was conducted on the survival dataset. From each TMA, one image was randomly chosen from both the validation and test sets of the survival dataset, while maintaining the overall data distribution with respect to the ISUP grades, relapse time and censoring status. An expert pathologist marked tumor regions in each image, enabling us to compare this to the attention weights per patch. A patch is counted as tumorous if 66% of it lie within the marked tumor region. Figure 4(*c*) shows that all highlighted patches lie within the tumor area, however not all patches in the tumor area receive a high attention weight. Figure 4(*d*) shows the results on all images showing tumor tissue drawn from the test set. Patches marked as tumor show on average higher attention weights than non-tumor image patches.

Overall, both experiments provide strong evidence that eCaReNet focuses on tumor regions, thus human interpretable explanations are provided.



(*a*) (*b*)

(*c*) (*d*)

Figure 4: In the example images, lighter patches indicate higher attention weights. Boxplots (b) and (d) show results over all example images per experiment. (a) An example image of the first experiment, with the lower part being benign and the upper part being malignant. (c) An example image with the tumor region marked in black. Patches with highest attention weights all lie within the tumor area. For both experiments, the attention weights for malignant patches are significantly higher than for benign patches ($* : p < 0.05$, $** : p < 0.01$).

## 6. Conclusion

We developed an end-to-end deep learning model for predicting prostate cancer patients' time to relapse using only images as data source. By directly utilizing time to relapse as ground truth, we could show that detailed annotations are not necessary for training, but are useful for pretraining on a small dataset. eCaReNet reaches the same AUC on the validation set that can be reached with ISUP scores, annotated by an expert pathologist on the whole prostate, while our model only uses a single image per patient.

By including explainability in our model, we tackle a major drawback of end-to-end systems. With an attention module, we open up the black box and showed in two experiments that the model weights malignant patches significantly higher than benign patches.

Since eCaReNet only requires pairs of histopathology images and physician-independent labels, it is generalizable and can be applied to other use cases and end points, like time to disease-related death.

Future work includes a more detailed analysis of the model in terms of explainability. Further improvements are expected when including multiple images per patient or adding additional information, e.g. about the patient's age, PSA value or family history.

## Acknowledgments

The authors would like to thank Pierre Machart for his contributions especially during planning and creation of this project.

## References

E. Abels, L. Pantanowitz, F. Aeffner, M. D. Zarella, J. van der Laak, M. M. Bui, V. N.P. Vemuri, A. V. Parwani, J. Gibbs, E. Agosto-Arroyo, A. H. Beck, and C. Kozlowski. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *The Journal of Pathology*, 249(3):286–294, 2019.

E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rüschoff, and M. Claassen. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Nature Scientific Reports*, 8(1): 1–11, 2018.

A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

P. Blanche, M. W. Kattan, and T. A. Gerds. The c-index is not proper for the evaluation of t -year predicted risks. *Biostatistics*, 20(2):347–357, 2019.

G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78 (1):1–3, 1950.

W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. De Bel, B. van Ginneken, J. van Der Laak, C. Hulsbergen-van de Kaa, and G. Litjens. Automated deep-learning system for Gleason grading of prostate cancer using biopsies : a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020.

G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing – EMNLP 2014*, pages 1724–1734, 2014.

H. D. Couture, J. S. Marron, C. M. Perou, M. A. Troester, and M. Niethammer. Multiple instance learning for heterogeneous images: Training a cnn for histopathology. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 254–262, Cham, 2018. Springer International Publishing.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

T. Dozat. Incorporating Nesterov Momentum into Adam. In *ICLR Workshop*, pages 2013–2016, 2016.

H. Duanmu, P. B. Huang, S. Brahmavar, S. Lin, T. Ren, J. Kong, F. Wang, and T. Q. Duong. Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer Using Deep Learning with Integrative Imaging , Molecular and Demographic Data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 242—-252. Springer International Publishing, 2020.

L. Egevad, R. Mazzucchelli, and R. Montironi. Implications of the International Society of Urological Pathology Modified Gleason Grading System. *Archives of Pathology & Laboratory Medicine*, 136 (4), 2012.

A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher. Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4(1):1–9, 2021.

J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray. Global cancer observatory: Cancer today. Lyon, France: International agency for research on cancer. https://gco.iarc.fr/today/home(accessed: 17.06.21), 2020.

T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis.* John Wiley & Sons Inc., 2005.

T. Gerds and M. Kattan. *Medical Risk Prediction: With Ties to Machine Learning.* CRC Press, Jan 2021.

E. Giunchiglia, A. Nemchenko, and M. van der Schaar. Rnn-surv: A deep recurrent model for survival analysis. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 23–32. Springer International Publishing, 2018.

D. F. Gleason and G. T. Mellinger. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *Journal of Urology*, 111(1):58–64, 1974.

D. J. Grignon. Prostate cancer reporting and staging: needle biopsy and radical prostatectomy specimens. *Modern Pathology*, 31:96–109, 2018.

H. Haider, B. Hoehn, S. Davis, and R. Greiner. Effective Ways to Build and Evaluate Individual Survival Distributions. *Journal of Machine Learning Research*, 21:1–63, 2020.

M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, Jul 2018.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18:1–12, 2018.

H. Kvamme, Ø. Borgan, and I. Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20: 1–30, 2019.

H. Li, D. Han, Y. Hou, H. Chen, and Z. Chen. Statistical inference methods for two crossing survival curves: A comparison of methods. *PLoS ONE*, 10 (1):1–18, 2015.

M. Y. Lu, D. F.K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

K. Nagpal, D. Foote, Y. Liu, P.-H. C. Chen, E. Wulczyn, F. Tan, N. Olson, J. L. Smith, A. Mohtashamian, J. H. Wren, G. S. Corrado, R. MacDonald, L. H. Peng, M. B. Amin, A. J. Evans, A. R. Sangoi, C. H. Mermel, J. D. Hipp, and M. C. Stumpe. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *npj Digital Medicine*, 2(48):1–10, 2019.

OpenCV. Open source computer vision library. https://github.com/opencv/opencv-python, 2015.

N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

M. Parsons and H. Grabsch. How to make tissue microarrays. *Diagnostic Histopathology*, 15(3):142–150, 2009.

K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu. Deep Recurrent Survival Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4798–4805, 2019.

Reuven Rubinstein. The Cross-Entropy Method for Combinatorial and Continuous Optimization. *Methodology And Computing In Applied Probability*, 1(2):127–190, 1999.

D. E. Rumelhart and J. L. McClelland. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, pages 318–362, 1987.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

D. Rymarczyk, A. Borowa, J. Tabor, and B. Zieliński. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1720–1729, 2021.

G. Sauter, T. Clauditz, S. Steurer, C. Wittmer, F. Büscheck, T. Krech, F. Lutz, M. Lennartz,

L. Harms, L. Lawrenz, C. Möller-Koop, R. Simon, F. Jacobsen, W. Wilczak, S. Minner, M. C. Tsourlakis, V. Chirico, S. Weidemann, A. Haese, T. Steuber, G. Salomon, M. Matiu, E. Vettorazzi, U. Michl, L. Budäus, D. Tilki, I. Thederan, D. Pehrke, B. Beyer, C. Fraune, C. Göbel, M. Heinrich, M. Juhnke, K. Möller, A. A. A. Bawahab, R. Uhlig, H. Huland, H. Heinzer, M. Graefen, and T. Schlomm. Integrating Tertiary Gleason 5 Patterns into Quantitative Gleason Grading in Prostate Biopsies and Prostatectomy Specimens. *European Urology*, 73(5):674–683, 2018.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition – CVPR 2015*, 07-12-June:1–9, 2015.

B. Tang, A. Li, B. Li, and M. Wang. CapSurv: Capsule Network for Survival Analysis With Whole Slide Pathological Images. *IEEE Access*, 7:26022–26030, 2019.

L. A. Vale-Silva and K. Rohr. Long-term cancer survival prediction using multimodal deep learning. *Nature Scientific Reports*, 11(1):1–12, 2021.

E. Wulczyn, D. F. Steiner, Z. Xu, A. Sadhwani, H. Wang, I. Flament-Auvigne, C. H. Mermel, P.-H. C. Chen, Y. Liu, and M. C. Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS ONE*, 15(6):1–18, 2020.

L. Xiao, J.-G. Yu, Z. Liu, J. Ou, S. Deng, Z. Yang, and Y. Li. Censoring-Aware Deep Ordinal Regression for Survival Prediction from Pathological Images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 449–458, 2020.

A. Yala, P. G. Mikhael, F. Strand, G. Lin, K. Smith, Y. L. Wan, L. Lamb, K. Hughes, C. Lehman, and R. Barzilay. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578):1–12, 2021.

Y. Yamamoto, T. Tsuzuki, J. Akatsuka, M. Ueki, H. Morikawa, Y. Numata, T. Takahara, T. Tsuyuki, K. Tsutsumi, R. Nakazawa, A. Shimizu, I. Maeda, S. Tsuchiya, H. Kanno, Y. Kondo, M. Fukumoto, G. Tamiya, N. Ueda, and G. Kimura. Automated acquisition of explainable knowledge from unannotated histopathology images. *Nature Communications*, 10(1), 2019.

J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65, 2020.

L. Zhang, D. Dong, L. Zhong, C. Li, C. Hu, X. Yang, Z. Liu, R. Wang, J. Zhou, and J. Tian. Multi-focus Network to Decode Imaging Phenotype for Overall Survival Prediction of Gastric Cancer Patients. *IEEE Journal of Biomedical and Health Informatics [published online ahead of print]*, (June 8):1–1, 2021.

X. Zhu, J. Yao, and J. Huang. Deep convolutional neural network for survival analysis with pathological images. In *IEEE International Conference on Bioinformatics and Biomedicine – BIBM 2016*, pages 544–547. IEEE, 2016.

# Appendix A. Dataset and preprocessing

## A.1. Prostate cancer grading

If prostate cancer is suspected, the amount of tumor and its grade are first estimated with a biopsy (Grignon, 2018). Among different treatment options, a prostatectomy might be chosen. For both biopsy and prostatectomy, the tissue is graded according to the Gleason grading system (Gleason and Mellinger, 1974). The tumor is stratified into five Gleason patterns. The Gleason score is defined as the sum of two patterns (in biopsy the most common and the worst, in prostatectomy the two most common patterns). As there is controversy about the grading system, the International Society of Urological Pathology (ISUP) decided on a scoring system that combines different Gleason pattern combinations into five groups (Egevad et al., 2012). However, there is no consensus yet on how to include possible tertiary patterns and also the percentages of the Gleason patterns in the tumor are neglected. Therefore, Sauter et al. (2018) introduced a more differentiated score, the Integrated Quantitative (IQ) Gleason score. In this work, we focus on the ISUP scores, as they are used in most other studies.

## A.2. Dataset distribution

In total, our survival dataset contains 17,230 images with prostate tissue, of which 60 images that either contain little to no tissue or are of poor quality (e.g. artifacts in the image) are omitted. Additionally, 3,624 patients with unknown relapse time or censoring status are excluded. 709 patients fall under the filtering criterion described in Section 3. Since multiple exclusion criteria may apply to one patient, 14,479 patients remain in the final survival dataset. The dataset distribution is shown in Figure A.1 for the training and validation sets. 90% of uncensored events occur prior to 7 years, 44% prior to 2 years.

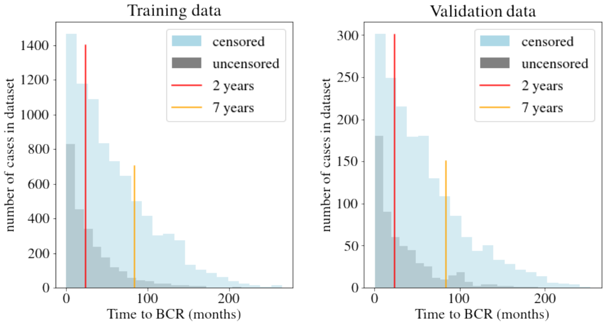


Figure A.1: Distribution of censored and uncensored cases in training and validation sets. The red and orange lines indicate 24 and 84 months after prostatectomy.

## A.3. Preprocessing

The images are all square but of different sizes (2490 × 2490 to 3181 × 3181 pixels), cut out from digitized TMA images. Therefore, they consist of circular tissue on white background. Since the white background does not include any information, only a center square of each circular spot is used, which results in images of size 2048 × 2048. The steps are shown in Figure A.2. Using the OpenCV package for Python (OpenCV, 2015), the RGB image A.2($a$) is first converted to grayscale A.2($b$) and binarized with Otsu-thresholding A.2($c$) (Otsu, 1979). Then, an ellipse is fitted A.2($d$) to use its center as center point for the resulting square A.2($e$)-A.2($f$), as not all tissue spots are perfectly round. It is assumed that the information loss at the margins by excluding some tissue is negligible compared to the gain in the foreground to background ratio.

During model training, data augmentation is applied. As data augmentation methods, the images are randomly flipped and rotated. For the survival model, images are further cut into regular, non-overlapping tiles. The continuous time to BCR label is converted to a binary vector of length 28 for 28 time intervals $t_j$. It has value 1 for $t_j < t^*$ and 0 for $t_j \geq t^*$.

# Appendix B. ISUP classification

For the ISUP classification in model $\mathrm{M}_{\mathrm{ISUP}}$, the ISUP score labels are encoded as an ordinal regression, to account for the fact that e.g. classes 2 and 3 are closer than classes 2 and 5. Labels are in the form $y = [l_i]$ for $i = 0...4$ with $l_i = 1$ if $i < c$ else 0 for each class c. For example, class 2 is encoded as $[1, 1, 0, 0, 0]$. The model's output $o$ is converted back to class label by summing all output values $p = \sum_i o_i$ and rounding the result. During training, the cross-entropy loss,

$$l = \sum_C y_c \log(o_c), \tag{12}$$

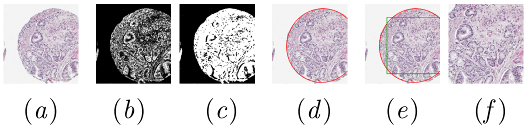


Figure A.2: Preprocessing steps to cut center pieces of the image and remove most of the white background on an example. Starting with the original image (a), it is converted to grayscale (b), then Otsu-thresholding is applied (c). Next, an ellipse is fitted to the tissue spot, here projected back to the RGB image (d) and a center square is cut (e) - (f).

introduced by Rubinstein (1999), is optimized using $C$ classes.

The ISUP classification was trained on 1863 images with 402 validation images of the Gleason dataset (see Figure C.1A). The kappa score on the validation set is 0.85. The confusion matrix is shown in Figure B.1.

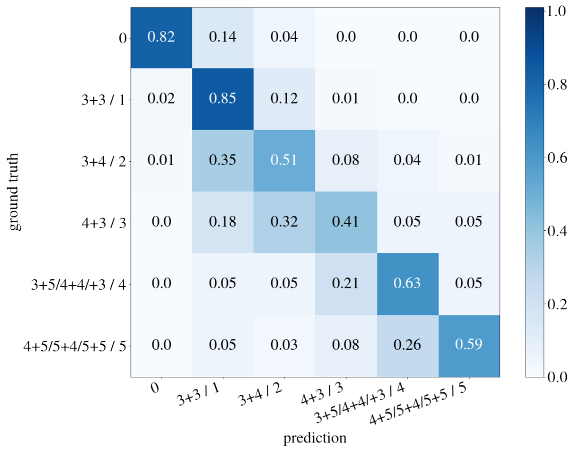


Figure B.1: Confusion matrix for ISUP classification. The axes show the ISUP scores as well as the corresponding possible Gleason grade combinations. Most class confusions are between neighboring classes.

# Appendix C. Model

The complete model eCaReNet is shown in Figure C.1 including the pretraining model $\mathrm{M}_{\mathrm{ISUP}}$ and $\mathrm{M}_{\mathrm{Bin}}$.

## C.1. Model results

The d-calibration for eCaReNet is shown in Figure C.2. The distribution's uniformity has been confirmed with a chi-square test. The AUC of eCaReNet can also be evaluated over time, as shown in Figure C.3.

# Appendix D. D-calibration

In a d-calibrated model, the survival functions per patient can be interpreted as probability of relapse over time. If a survival curve shows 90% survival probability, the patient can trust that only 10% of patients with the same diagnosis experience a relapse at that time point. This also means that 10% of patients should experience their event when the survival probability is between 90 and 100%. Since the same holds for all other intervals (0-10, 10-20, ...), the expected number of events is compared to the true number of events and a chi-square test is used to measure this. Censored patients need to be treated differently from uncensored patients, because their true event time is not known. For details, see (Haider et al., 2020).

# Appendix E. Risk stratification details

For each patient, an individual risk score is calculated with Equation 6. It follows that $r=0$ if $\forall t_j : S(t_j)=1$ and $r=1$ if $\forall t_j : S(t_j)=0$. Risk scores are grouped into classes to enable a relative ranking among patients. In order to assign risk scores to risk groups, intervals need to be defined, for which an exploratory approach is used.

For the selection of the interval limits, multiple possible interval limits are defined and the best combination is chosen as follows. For each possible combination, the patients are assigned to the risk groups and patients within one group are combined in a single Kaplan-Meier curve (Kaplan and Meier, 1958). The resulting Kaplan-Meier curves per risk group are tested for discrimination power with a log-rank test, which is commonly used in survival analysis (Li et al., 2015). Since the proposed model allows for non-proportional hazards and therefore crossing survival

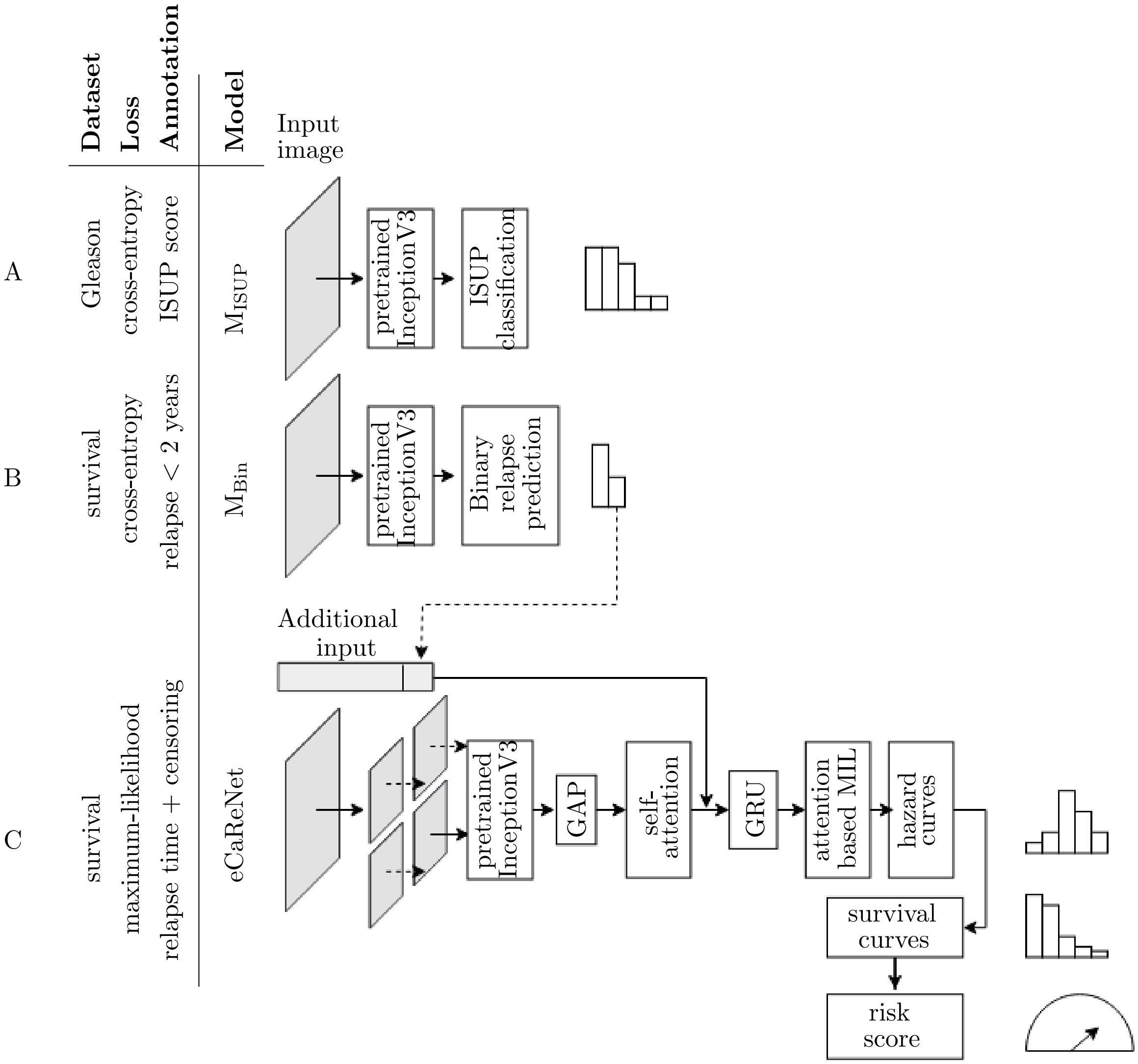


Figure C.1: Overview of the complete model, including the three steps $M_{ISUP}$, $M_{Bin}$ and eCaReNet. On the left, the dataset, the loss and the annotation for training are indicated. ISUP: International Society of Urological Pathology, Bin: Binary, GAP: Global Average Pooling, GRU: Gated Recurrent Unit, MIL: Multiple Instance Learning.

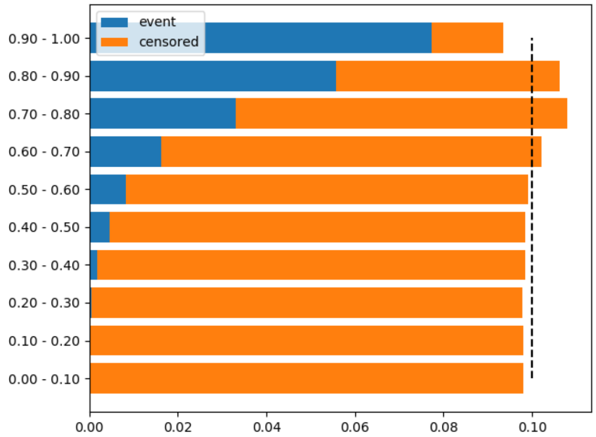


Figure C.2: Resulting d-calibration plot for eCaReNet.

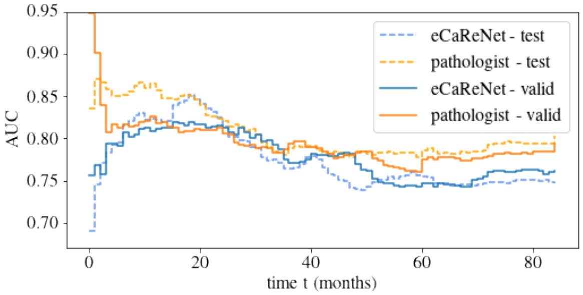


Figure C.3: The resulting AUC of eCaReNet over time. In the time range from 6 to 26 months after prostatectomy, the AUC is higher than 0.8. Overall, eCaReNet performs very similar to the expert pathologist. Only in the first months, where eCaReNet's survival curves are close to 1, it is outperformed by the pathologist, whose predictions are constant over time.

curves, the log-rank test is modified with Fleming-Harrington weights according to Fleming and Harrington (2005). If the test passes, the survival curves stratify well. Multiple combinations of boundaries can give perfectly stratified curves on the training set, which is why the best suited limits are further evaluated on the validation set. The limits with the best results on the validation set are used for final evaluation on the test set. The number of risk groups is also varied in this procedure, since using too few risk groups gives good stratification but is not meaningful for patients, whereas using too many risk groups, the Kaplan-Meier curves cannot separate well any more. In our analysis, using 8 risk groups was the largest number of possible groups that led to the best possible stratification in the training set. The found interval limits are 0.06, 0.12, 0.15, 0.18, 0.3, 0.42 and 0.51. Resulting Kaplan-Meier curves on the test set are shown in Figure 3. While all groups separate well in the training set, one log-rank test fails in the validation set and two tests fail for the test set. As limit for the p-value, 0.05 is chosen.

## Appendix F. Attention example

As an additional example for the experiment described in Section 5.3, Figure F.1 is provided. Also here it is shown that the malignant patches (in the lower part of the image) receive higher attention weights.

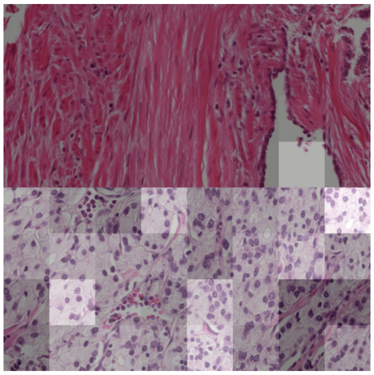


Figure F.1: Another example of attention weights. The lower part is annotated as malignant (Gleason 5), the upper part as benign.