A Deep Attention-Multiple Instance Learning Framework to Predict Survival of Soft-Tissue Sarcoma from Whole Slide Images

Van-Linh Le, Audrey Michot, Amandine Crombé, Carine Ngo, Charles Honoré, Jean-Michel Coindre, Olivier Saut, Francois Le-Loarer

A Deep Attention-Multiple Instance Learning Framework to Predict Survival of Soft-Tissue Sarcoma from Whole Slide Images

Van-Linh Le[1,2,3,8(✉)], Audrey Michot[3,5], Amandine Crombé[1,4,6], Carine Ngo[9], Charles Honoré[10], Jean-Michel Coindre[4,7], Olivier Saut[1,2], and Francois Le-Loarer[3,4,7]

[1] MONC Team, INRIA Bordeaux Sud-Ouest, Talence, France
[2] Bordeaux Mathematics Institute UMR 5251 (IMB), University of Bordeaux, CNRS and Bordeaux INP, Talence, France
van-linh.le@u-bordeaux.fr
[3] Bordeaux Institute of Oncology, BRIC U1312, INSERM, University of Bordeaux, Institute Bergonié, 33000 Bordeaux, France
[4] Faculty of Medicine, University of Bordeaux, Bordeaux, France
[5] Department of Oncological Surgery, Institute Bergonié, Bordeaux, France
[6] Department of Radiology, Pellegrin University Hospital, Bordeaux, France
[7] Department of Pathology, Institute Bergonié, Bordeaux, France
[8] Department of Data and Digital Health, Institute Bergonié, Bordeaux, France
[9] Department of Pathology, Institute Gustave Roussy, Villejuif, France
[10] Department of Oncological Surgery, Institute Gustave Roussy, Villejuif, France

Abstract. Soft-tissue sarcomas are heterogeneous cancers of the mesenchymal lineage that can develop anywhere in the body. A precise prediction of sarcomas patients' prognosis is critical for clinicians to define an adequate treatment plan. In this paper, we proposed an end-to-end Deep learning framework via Multiple Instance Learning (MIL), Deep Attention-MIL framework, for the survival predictions: Overall survival (OS), Metastasis-free survival (MFS), and Local-recurrence free survival (LRFS) of sarcomas patients, by studying the features from Whole Slide Images (WSIs) of their tumors. The Deep Attention-MIL framework consists of three steps: tiles selection from the WSIs to choose the relevant tiles for the study; tiles feature extraction by using a pre-trained deep learning model; and a Deep Attention-MIL model to predict the risk score for each patient via MIL approach. The risk scores outputted from the Deep Attention-MIL model are used to divide the patients into low/high-risk groups and predict survival time. The framework was trained and validated on a local dataset including 220 patients, then it was used to predict the survival for 48 patients in an external validation dataset. The experiments showed the proposed framework yielded satisfactory and promising results and contributed to accurate cancer survival predictions on both the validation and external testing datasets: By using the WSIs

O. Saut and F. Le-Loarer—Co-directed the research.

S. Ali et al. (Eds.): CaPTion 2023, LNCS 14295, pp. 3–16, 2023.
https://doi.org/10.1007/978-3-031-45350-2_1

feature only, we obtained an average C-index (of 5-fold cross-validation) of 0.6901, 0.7179, and 0.6211 for OS, MFS, and LRFS tasks on the validation dataset, respectively. On the external testing dataset, these scores are 0.6294, 0.682, and 0.76 for the three tasks (OS, MFS, LRFS), respectively. By adding the clinical features, these scores have been improved both on validation and external testing datasets. We obtained an average C-index of 0.7835/0.6378, 0.7389/0.6885, and 0.6883/0.7272 for the three tasks (OS, MFS, LRFS) on validation/external testing datasets.

Keywords: Multiple Instance Learning · Deep Attention model · Survival prediction · Soft-tissue sarcoma · Whole Slide Image

1 Introduction

Soft-tissue sarcomas (STS) are heterogeneous malignant tumors developing anywhere in the body. They represent 1% of cancers in adults and 5% in children. STSs have a variable prognosis, their management requires the use of aggressive treatments including debilitating surgeries and/or high-dose chemotherapies. The prognosis of STS is dominated by two events: local recurrence and distant metastasis. The occurrence of metastasis is a major adverse factor for overall survival (OS), but local control of the disease also impacts OS [1]. In most studies, the most significant factor to predict local recurrence is the quality of surgical margins [1], whereas metastasis and overall survival are mostly related to the FNCLCC histological grade [2] which remains to date the most widely used standard to predict survival of sarcoma patients. A clinical nomogram integrating the grade and clinical variables such as patient age and tumor size has improved the prognosis evaluation of sarcomas patients [3].

In addition to biological and clinical information, Whole Slide Images (WSIs) contain information relevant for analyzing the diagnosis and prognosis of cancer, e.g., Overall survival (OS), Metastasis-free survival (MFS), or Local-recurrence free survival (LRFS), as well as prediction of response to treatment. WSIs can indeed assess the tumor growth and morphology in detailed, high resolution. However, capturing cell detail makes the exported image potentially cumbersome to cope with, and analyzing WSIs challenging for several reasons: (1) WSIs may contain a billion difficult pixels to process computationally; (2) a patient could have several WSIs for study, with significant differences in texture and biological structures; (3) we receive an only label at the patient level but different WSIs for diagnosis.

Deep learning has become a current solution for image processing applications comprising pathological image analysis. However, the processing of WSIs is different from usual images due to the massive resolution of this kind of image. One possible solution to overcome this challenge is to consider a weakly supervised method via a Multiple Instance Learning (MIL) approach [4]. In MIL, we split a WSI into non-overlapping tiles (patches). Therefore, a WSI could be considered a bag of tiles. It is not mandatory to analyze all tiles, as some of them may not be relevant for diagnostic detection; therefore, a subset of tiles is selected for the study.

In recent years, Deep Learning via MIL [5] has emerged as a promising way to predict survival in cancer patients by analyzing the WSIs [6,7]. Ilse et al. [6] have proposed to use the attention-MIL for classifying the histopathological images of breast and colon cancers datasets: the model with attention operator outperformed the other operators (e.g., max-pooling MIL or mean-pooling MIL), achieving an average AUC (of 5 folds cross-validation) of 0.775 and 0.968 in breast and colon cancers datasets, respectively. Yao et al. [7] have introduced a combination of a Siamese model [8] and an attention-based one to predict survival based on imaging features. The method used the K-Means algorithm [9] to cluster the imaging features into several phenotypes. Then, the Siamese model was used to extract the features for each phenotype before feeding to the attention module for aggregating the WSI-level feature. Finally, the aggregated WSI feature was processed by two fully-connected layers and outputted the risk score for each patient. The method was applied to lung and colorectal cancer datasets. The C-indexes on lung and colorectal datasets were 0.6963 and 0.652, respectively. Likewise, Pierre et al. [10] have proposed the MesoNet model to predict the OS of mesothelioma patients. To develop their model, they firstly splitted the WSIs into tiles with a size of 224×224 pixels and selected 10K of tiles for analysis due to the limitation of the computation memory. Secondly, the pre-trained ResNet50 [11] was used to extract the features of the tiles. Then, a convolutional one-dimensional was used to generate the score for each tile. Finally, the 10 highest and 10 lowest scores were selected and used as the input for the multi-layer perceptron classifier to provide the scores for each patient. MesoNet has achieved an average C-index of 0.642 and 0.643 on the training (2981 patients) and testing dataset (56 patients), respectively [10].

In this work, we report an end-to-end framework, Deep Attention-MIL, to predict the survival of sarcomas patients. At the heart of our framework is a deep learning model with an attention mechanism for survival prediction via MIL. We evaluate the proposed framework on two datasets originating from two different comprehensive cancer centers in France. In this work, we show that the framework offers satisfactory predictions of the survival probability of the patients compared to the gold standard used sarcomas patients, the FNCLCC histological grade [12] (Sect. 3.2).

2 Methodology

Figure 1 presents the workflows of the proposed framework. It was developed in three phases: firstly, the non-overlapping patches (tiles) were extracted from the WSIs of the patients. Then, a pre-trained deep learning model (e.g., ResNet50 [11]) was used as an encoder to extract the features from the tiles. Finally, the extracted features were fed into the deep learning survival model to predict the risk score for each patient. The risk scores were then taken by a non-parametric estimator (e.g., Kaplan-Meier [13]) to predict the survival probability for the patient.

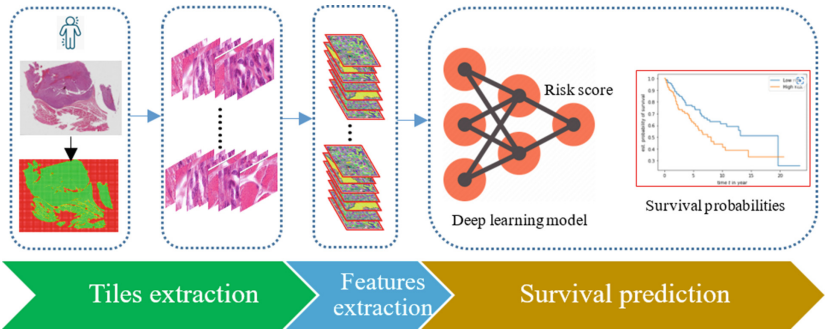

Fig. 1. The workflow of the proposed framework

2.1 Tiles Extraction

A WSI is organized as a pyramid of images with different magnification levels, the image at the highest level (40× of magnification) could have a resolution of $100K \times 100K$ pixels, and it is down-sampled over the magnification levels. Usually, the image at the highest magnification level (40×) is chosen for study. It is worth noting that a large part of the image does not contain any tissue, and is therefore useless for analysis. These regions are discarded during the tiling process. Because of the problem with high resolution of the original image, it is impossible to apply the classical image processing techniques on whole image for pre-processing step. Instead, an image at a lower magnification level (e.g., 12×, scaled image) is used to perform the pre-processing operations (e.g., segmentation, binarization), to select the tissue area, and to determine and mark the location of the interesting tiles. At the end of this stage, the tiles on the original image corresponding to the selected tiles on the scaled images are extracted and used for the study.

As a preferred size from the deep learning models for image classification [11,14], the original image is divided into non-overlapping tiles with a size of 224×224 (W × H) pixels. Based on the fraction of tissue, tiles are classified into 4 groups: Group A consists of the tiles that are composed of more than 80% of tissue, group B contains the tiles which have more than 10% and less than 80% of tissue, group C includes the tiles which have more than 10% of tissue, and group D of that tiles that do not contain tissue. In the context of this work, the tiles composed of more than 10% (from group A and B) of tissue have been used for our analysis. As mentioned, a patient could have several WSIs, even if we only extract tissue tiles, we can still get hundreds of thousands of tiles for each patient.

2.2 Features Extraction

Unlike segmentation and detection tasks in WSIs analysis [15–17], our framework predicts patient-level outcome aggregated from tile-level information. As pointed

out in [18], training patch-based CNNs for weakly supervised learning is very time-consuming (several weeks), we propose to use features from pre-trained models instead of using CNNs to learn features from scratch. Here, for instance, we use a pre-trained ResNet50 [11] on ImageNet [14] as an extractor to extract the features of the tiles. For each tile, 2048 features are considered. Then, the extracted features are concatenated to obtain the features for each patient. The extracted features of WSIs are presented as a matrix of $(N \times 2048)$, where N is the total number of WSI tiles.

2.3 Deep Attention-Multiple Instance Learning Model for Survival Prediction

Figure 2 illustrates the proposed deep learning architecture (Deep Attention-MIL) for survival predictions. The layers of this model can be decomposed into three groups: the first group consists of the layers before the Attention module [6]. These are used to aggregate the features for each tile as well as compute the score for each tile; the second group is an Attention module [6] which outputs the attention score for each tile. These scores present the importance of each tile conducted to the final prediction. The attention score is combined with the corresponding tile features before passing it to the layers in the last group to estimate the risk score for each patient.

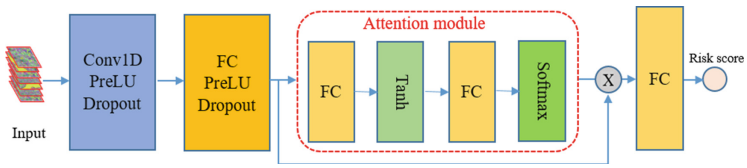

Fig. 2. The proposed Deep Attention-MIL survival model

Our model is derived from other studies [6,7] with modifications to adapt the architecture to our objective. First, we replaced the ReLU activation function with PreLU activation, which is more precise in the decision by making the leakage coefficient a parameter that is learned along with the other network parameters [19]. Then, to prevent overfitting, we added dropout layers [20] at the end of each group of layers, and reduce the number of features in the last layer of the model. It is worth that the final model was obtained after trying different combinations of layers and performing experiments on the same dataset.

Table 1 details the input/output dimensions at each layer/module of the proposed Deep Attention-MIL survival model. The input of the model is the selected features of the patient, organized as a matrix of $(M \times 2048)$ where M is the number of tiles considered. After passing the layers in the first group, the features of each tile are collected and dimension is reduced to 256. These reduced features are inputted to the attention module to output the attention score for each tile

Table 1. The input/output dimensions at each layer of the survival model.

Layer	Input	Output
Conv1D/PreLU/Dropout	$M \times 2048$	$M \times 512$
FC/PreLU/Dropout	$M \times 512$	$M \times 256$
Attention module	$M \times 256$	1×256
FC	1×256	1

$(1 \times M)$. The attention score is multiplied with corresponding tile features to obtain the representation feature for WSI (1×256) which is the input for the last layer in the model. Finally, a linear function learns the representation of WSI to provide the risk score for the patient.

Attention Module: Local representation (two layers before the attention module) encodes features of the tiles, but our model provides the score at the patient level. Therefore, aggregating tile features into patient-level representation is a necessary step. A popular choice would be to use the maximum or the mean operator. Yet, the drawbacks are clear: they are pre-defined and not trainable which might not be adequate for this specific task. A better way to integrate tile information is to leverage an attention mechanism that considers the importance of each tile. In this work, we proposed to use the attention based MIL for aggregation of tile features to obtain the representation of patient-level [6]. It consists of two linear layers combined with Tanh activation functions. A soft-max activation function is placed at the end of the module to compute the attention score for each tile, and ensure that the sum of all attention scores is equal to 1.

3 Experiments and Results

3.1 Dataset Description and Experimental Setups

Datasets: The experiments were carried out on two different clinical cohorts from two comprehensive cancer centers: Institute Bergonié (Bordeaux, France) (IB dataset) and Gustave Rossy (Villejuif, France) (IGR dataset). These two cohorts were extracted from the Sarcoma BCB (https://sarcomabcb.org:connect). The criteria of inclusion included: primary sarcoma, location of trunk walls and limbs, upfront surgical resection, and patient naive of neoadjuvant therapy. The IB dataset consisted of 220 patients with more than 450 WSIs representing at least 2 WSIs per patient. The samples in this dataset were collected from 01/01/1990 to 01/12/2020. The IGR dataset consisted of 48 patients with more than 100 WSIs collected from 01/01/2000 to 01/12/2016. For all included patients, clinical follow-up was updated regarding survival, date of death, occurrence of metastasis and local recurrence. The IB dataset was used to train and validate the models. The IGR dataset was used

as an external validation cohort (testing dataset). Table 2 details the number of patients, number of WSIs, recorded events of patients, location for collecting samples for each dataset.

Ethics: This study was conducted following local ethical guidelines and approved by the institutional research board of our institutions; all cases are recorded in the French expert sarcoma network (NetSarc+) database, which is approved by the National Committee for Protection of Personal Data (CNIL, no. 910390).

Table 2. The details of studied datasets.

Dataset	IB	IGR
No. patients	220	48
No. WSIs	450	105
No. patients alive/dead	133/87	36/12
No. patients non-metastatic/metastatic	148/72	37/11
No. patients non-recurrence/recurrence	188/32	40/8
Location	Institute Bergonié (Bordeaux, France)	Institute Gustave Roussy (Villejuif, France)

Experiments Setup: We evaluated the performance of the model on three survival tasks: Overall survival (OS), Metastasis free survival (MFS), or Local-recurrence free survival (LRFS). For each task, we performed a 5-fold cross-validation on the training dataset (IB). Then, the 5 corresponding models were used to predict the scores for the patients in the external validation set. (IGR). For all three tasks, we reported the C-index and Confidence Interval (CI-95%). The reported C-index in this study was the average C-index of 5-fold cross-validation.

As mentioned in Sect. 2.1, each patient had a hundred thousand tiles. For clear computational reasons, we could not analyze all tiles; therefore, we selected a subset of tiles (M = 10K) for the study. This value is empirically set after trying different values for the number of tiles for each patient. As the output of this step, each patient was represented as a matrix of $(M \times 2048)$, this matrix was used as the input of the survival model.

Implementation Details: The model was implemented in the PyTorch library [21]. The model was trained for 200 epochs using an Adam optimization [22] with a weight decay of 10^{-4}. The learning rate and batch size have been set to 3×10^{-3} and 1, respectively. An early stopping strategy was applied by monitoring the validation loss to avoid over-fitting.

3.2 Experimental Results

In this section, we investigate the performance of our approach. First, we present the model's performances on tile features only for three survival tasks: OS, MFS,

and LRFS. Then, we add more insight to the model by considering some clinical features, compare the two approaches (with and without clinical features). Finally, these results are compared to the results of a Cox model [23], a standard model for survival prediction.

Table 3. The C-index scores (± CI-95%) for OS, MFS, and LRFS tasks on IB validation and IGR testing datasets (with WSI features only).

Dataset	OS	MFS	LRFS
IB	0.6901 (± 0.0388)	0.7179 (± 0.0709)	0.6211(± 0.0537)
IGR	0.6294 (± 0.0153)	0.6820 (± 0.0491)	0.7600(± 0.0184)

Prediction from Tiles Features: Table 3 presents the obtained average C-index (±CI-95%) for each task on each dataset. On the validation set (IB), we obtained the average C-index scores of 0.6901 (±0.0388), 0.7179 (±0.0709), and 0.6211 (±0.0537) for OS, MFS, and LRFS tasks, respectively. On the external validation set (IGR), the average C-index scores on OS and MFS tasks are lower than the validation set, 0.6294 (±0.0153) and 0.682 (±0.0491) for OS and MFS, respectively. However, the C-index on LRFS outperforms the score on the validation set 0.76 (±0.0184). Although the C-index scores are a bit smaller on the test set, the difference is tiny. This could indicate a good generalization ability of our approach to unseen data.

As mentioned, the model provided the risk (event) score for each patient. Then, the risk scores were used to divide the patients into two groups: low-risk and high-risk, using Eq. 1. Finally, an estimator (e.g., Kaplan-Meier [13]) was used to obtain the survival probability of the two groups.

$$dx = f(x) = \begin{cases} 0 & \text{if } Px < PI \\ 1 & \text{if } Px \geq PI \end{cases} \tag{1}$$

where Px is the predicted risk score from the model and PI is the median of the risk scores.

Figure 3 illustrates the survival curves (in 10 years) of low/high-risk patients by using the Kaplan-Meier estimator [13] for each task on the IB validation dataset and the IGR testing dataset. On the IB validation dataset (left column in 3),the statistical information between the two groups was significant ($p < 0.05$), and the predictions of our model were good enough to separate the patients. The prediction curves were compared to the Grade curves (Grade is a gold standard to classify cancer tissues based on their appearance and behavior when viewed under a microscope for helping the doctor know about the aggressiveness of cancer. The grade is usually described using a number from 1 to 3 or 4. The higher the number, the more different the cancer tissues look from normal tissues and the faster they are growing), the curves provided by our scores are the same

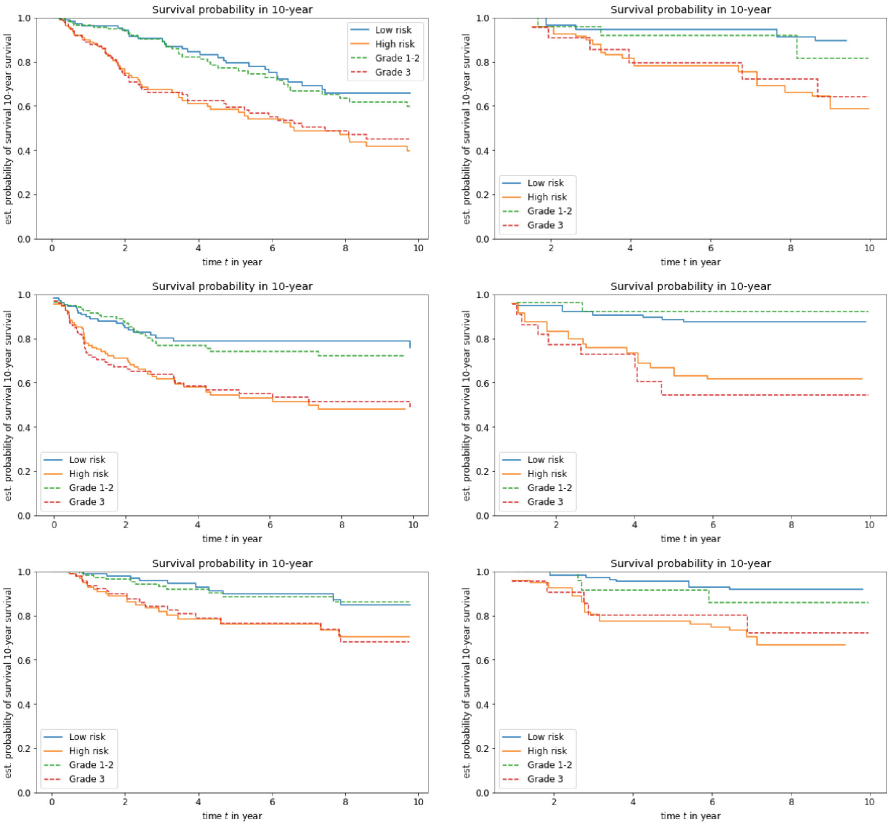

Fig. 3. The survival curves on IB validation (left column) and IGR testing (right column) datasets. From top to bottom: the survival curves on OS, MFS, and LRFS tasks compared to the histological Grade.

level or better in some periods (e.g., from 5 to 6 years). On the external cohort (IGR) (right column in Fig. 3), the stratification for OS and MFS were similar to grade and a bit better for LRFS.

Adding Clinical Features to Imaging: This section presents the results of the enhanced version of the proposed deep learning model. In this version, we consider additional clinical features beside the imaging features. In order to compare to another approach on clinical features, clinicians have selected four clinical features: *age, size of tumor, grade, and histotype*, to add to the tiles features for the prediction of survival.

Table 4 summaries the C-index (± CI-95%) for each task on each dataset. On the validation dataset (IB), the C-index is improved on all three survival tasks: OS - 0.7835 (±0.034), MFS - 0.7389 (±0.034), LRFS - 0.6883 (±0.037). On IGR dataset, the scores for three survival tasks are: OS - 0.6378 (±0.043),

MFS - 0.6885 (±0.033), LRFS - 0.7272 (±0.068). It is the same improvement on OS and MFS tasks. However, the improvement in IGR was not as large as expected from the validation set results, even the score was decreased a little bit on LRFS task. To explain the difference, we hypothesized that we have a bias between the clinical data of the patients from the two centers.

Table 4. The C-index scores (± CI-95%) of OS, MFS, LRFS tasks on IB validation and IGR testing datasets (with WSI and clinical features).

Dataset	OS	MFS	LRFS
IB	0.7835 (± 0.034)	0.7389 (± 0.034)	0.6883 (± 0.037)
IGR	0.6378 (± 0.043)	0.6885 (± 0.033)	0.7272 (± 0.068)

Using the same strategy to evaluate the prediction scores as presented in the previous section, the output scores of the model are used to split the patients into low/high-risk groups. Then, we illustrate the two group curves by using the Kaplan-Meier estimator (Fig. 4). On the IB dataset, the survival curves on OS and MFS tasks are significantly separate, and they are the same level as the grade curves; on the LRFS task, the model met difficulty to split the patients in the first period of 5 years. The survival curves (for three tasks: OS, MFS, and LRFS) on the IGR dataset are not significantly changed compared to the curves without clinical features. It seems that adding the clinical has more effect on the IB (C-index scores) than the IGR dataset.

In addition to the risk score, proposed model also provides the attention score for each tile, which allows us to predict the survival pattern before predicting the risk score. Figure 5 illustrates the top 15 tiles of a patient who had a metastatic relapse, these tiles with high attention scores are the ones affecting the most the model prediction. Among these, 4 interested normal tissue surrounding the tumor, 9 originated from the tumor, and 2 normal tissue far from the tumor.

Comparison with the Cox Model: Cox model [23] is a popular model for the prediction of survival. We have used the Cox model with two objectives: (1) to verify the informative value of tile encoding, is it enough for use as a feature in a survival model? (2) to compare the performance of our framework with a classical method for survival tasks.

Table 5. The C-index scores from Cox model on OS, MFS, LRFS tasks.

Dataset	OS	MFS	LRFS	Nb. features
IB	0.7455	0.6835	0.7216	4 features (clinical)
IGR	0.57	0.7049	0.7381	
IB	0.7591	0.7071	0.7274	5 features (clinical + risk score)
IGR	0.5733	0.74	0.7279	

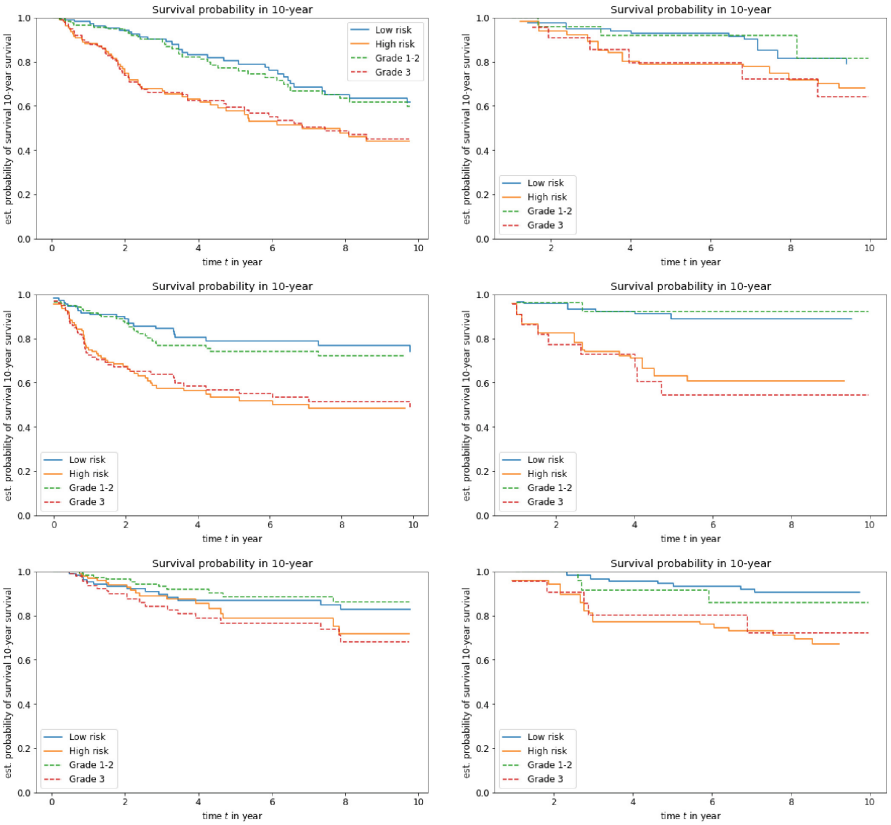

Fig. 4. The survival curves (in 10 years) on IB (left column) and IGR (right column) datasets by using imaging and clinical features. From top to bottom: the survival curves on OS, MFS, LRFS tasks.

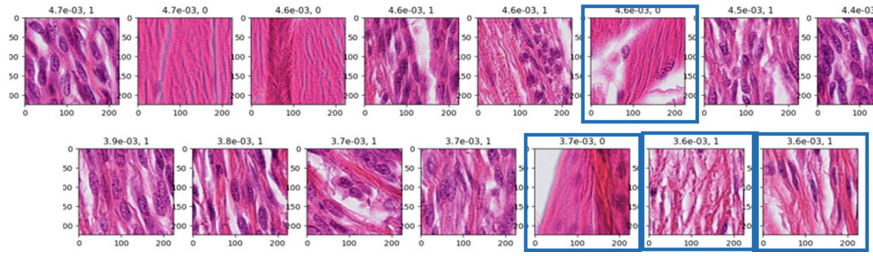

Fig. 5. Top-15 tiles with highest attention score of a metastasic patient from IB dataset. The blue boxes highlights tiles of normal tissue surrounding the tumor tissue. (Color figure online)

The Cox model was trained on the IB dataset and tested on the IGR dataset. We first try the Cox with 4 clinical features. Then, we used the outputs of deep learning model (risk score) on WSI features as the fifth feature on the Cox model. Table 5 shows the C-index scores on three tasks from two datasets (IB/IGR). We see from the results that adding the tile features has improved the performance of the Cox model. On the validation set, the Cox model has obtained a C-index of 0.7455, 0.6835, and 0.7216 for OS, MFS, and LRFS tasks, respectively, by utilizing clinical features. By adding the deep learning risk score, these C-indexes have improved to 0.7591, 0.7071, and 0.7274 for OS, MFS, and LRFS tasks, respectively. In addition, the C-index scores have been improved as well on the testing set. Besides that, in the comparison between the results of Cox model (5 features) and proposed Deep Attention-MIL model, our results on the validation set are better than the Cox model on OS and MFS tasks, while we are close to Cox's result on the LRFS task.

Comparison with Other Deep Learning Survival Models: To have an objective assessment of the performance of the model, we have re-implemented the methods which have described in [10] and [6]. Then, we performed 5-fold cross-validation and reported the average values of the C-index on IB dataset. Table 6 shows the prediction power of the proposed framework compared to the different survival models based on the average C-index scores of 5-fold cross-validation on different survival tasks. We see from the table that the performance of our proposed framework on OS and MFS tasks outperforms the other methods, while we are less than a little bit on the LRFS task compared to the Attention-based MIL average pooling approach. Generally, the proposed framework achieves the best performance among all methods on most of survival tasks.

Table 6. Performance comparison of the proposed framework with other available methods using average C-index scores (± CI-95%) of OS, MFS, and LRS tasks on IB validation set. The **bold** values indicate the best scores for each task.

Method	OS	MFS	LRFS
Deep Attention-MIL (proposed)	**0.6901** (± 0.0388)	**0.7179** (± 0.0709)	0.6211 (± 0.0537)
MesoNet [10]	0.6118 (± 0.0531)	0.6134 (± 0.0283)	0.5881 (± 0.0790)
Attention-based MIL [6] Max-pooling	0.5905 (± 0.0211)	0.6333 (± 0.1061)	0.5255 (± 0.0477)
Attention-based MIL [6] Average-pooling	0.6468 (± 0.0609)	0.6535 (± 0.0671)	**0.6444** (± 0.0454)

4 Discussion and Conclusion

In this paper, we propose a Deep Attention-MIL framework for survival predictions from WSIs. Our objective was to investigate the role of the WSIs for the prognostic tasks along the clinical features. First, we have built a baseline from the Cox model, which took into account the clinical only. Then, the deep learning risk score obtained from WSIs features was added to the analysis of the Cox

model. In this study, we show that imaging adds some fascinating insight into the Cox analysis, and that it can improve the performance of the Cox model. Finally, we propose a fully deep learning framework to combine the tile and clinical features for survival prediction in sarcoma patients. Our framework achieved good performance for various survival prediction tasks, even better results than the Cox model, and comparable to the gold standard for cancer studies. The results have been also compared to several deep survival models, the comparison showed that our framework achieves higher performance than these methods. One should keep in mind that this model may apply to sarcoma patients affected with any sarcoma histotypes developed in the trunk walls and the limbs naive of treatment.

The present study raises some questions that we plan to address in future works. (1) Concerning the feature extraction step, one may replace the current extractor with another extractor that retains a relation with studied images, for example, we are examining a self-supervised learning model which can be downstream to use as an extractor. (2) We plan on analyzing the effect of the origins of the tiles on the tile selection step, in which we are ongoing to automate the classification of the tiles from different regions. (3) One needs to investigate the variability of the WSIs coming from various centers to improve the model's performances and develop an adequate harmonization method.

Acknowledgements. This work was supported by grants from the SIRIC Bordeaux and the Fondation Bergonié. This study is carried out in collaboration between the MONC team, Inria Bordeaux Sud-Ouest, and the BRIC U1312.

References

1. Casali, P.G., et al.: Soft tissue and visceral sarcomas: Esmo-euracan clinical practice guidelines for diagnosis, treatment and follow-up. Ann. Oncol. **29**, iv51–iv67 (2018)
2. Coindre, J.-M., Terrier, P., Bui, N.B., et al.: Prognostic factors in adult patients with locally controlled soft tissue sarcoma: a study of 546 patients from the French federation of cancer centers sarcoma group. Journal of Clinical Oncology **14**(3), 869–877 (1996)
3. Callegaro, D., et al.: Development and external validation of two nomograms to predict overall survival and occurrence of distant metastases in adults after surgical resection of localised soft-tissue sarcomas of the extremities: a retrospective analysis. Lancet Oncol. **17**(5), 671–680 (2016)
4. Herrera, F., et al.: Multiple Instance Learning. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-319-47759-6
5. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1), 31–71 (1997)
6. Ilse, M., Tomczak, J., Welling, W.: Attention-based deep multiple instance learning. In International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
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. **65**, 101789 (2020)

8. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 539–546. IEEE (2005)
9. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, vol. 1, pp. 281–297 (1967)
10. Courtiol, P., et al.: Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat. Med. **25**(10), 1519–1525 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
12. Coindre, J.M., et al.: Predictive value of grade for metastasis development in the main histologic types of adult soft tissue sarcomas: a study of 1240 patients from the French federation of cancer centers sarcoma group. Cancer: Interdisc. Int. J. Am. Cancer Soc. **91**(10), 1914–1926 (2001)
13. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**, 457–481 (1958)
14. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision (IJCV) **115**(3), 211–252 (2015)
15. Courtiol, P., Tramel, E.W., Sanselme, M., Wainrib, G.: Classification and disease localization in histopathology using only global labels: a weakly-supervised approach. preprint arXiv:1802.02212 (2018)
16. Rony, J., Belharbi, S., Dolz, J., Ayed, I.B., McCaffrey, L., Granger, E.: Deep weakly-supervised learning methods for classification and localization in histology images: a survey. arXiv preprint arXiv:1909.03354 (2019)
17. Wang, D., Khosla, A., et al.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)
18. Hou, L., Samaras, D., Kurç, T.M., Gao, Y., Davis, J.E., Saltz, J.: Efficient multiple instance convolutional neural networks for gigapixel resolution image classification, vol. 7, pp. 174–182 (2015). preprint arXiv:1504.07947
19. He, K., Zhang, X., et al.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
21. Paszke, A., et al.: Automatic differentiation in pytorch. In: NIPS-W (2017)
22. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
23. Cox, D.R.: Regression models and life-tables. J. Royal Stat. Soc. Ser. B (Methodological) **34**(2):187–202 (1972)