scientific reports

Check for updates

OPEN

Prognostic prediction in soft-tissue sarcomas using deep learning and digital pathology of tumor and margin areas

Audrey Michot[1,9,13], Van-Linh Le[2,13], Jean-Michel Coindre[3,13], Valérie Velasco[3], Malika Soussi[3], Nouria Mesli[4], Antoine Italiano[5], Maud Toulmonde[5], Axel Le Cesne[6], Sylvie Bonvalot[7], Lucile Vanhersecke[3], Charles Honoré[8], Carine Ngo[4], François Le Loarer[3,9,13], Olivier Saut[10,13] & Amandine Crombé[9,11,12,13]✉

The histological FNCLCC grade is the primary prognostic factor in soft-tissue sarcoma (STS) but fails to fully capture high risk patients. This study aimed to develop and validate a deep learning (DL) model to predict metastatic relapse-free survival (MFS) using digital hematoxylin and eosin-stained whole-slide images. A retrospective analysis was conducted on 308 STS patients from two cancer centers, divided into a training cohort (149 patients) and two independent validation cohorts (64 and 95 patients). Supervised multi-instance learning convolutional neural network models were trained on distinct tumor regions—center (C), periphery (P), and margins (R)—to optimize predictive performance. Univariable analysis showed DL models using tumor center (DL-C), periphery (DL-P), and their combination (DL-CP) were consistently associated with MFS across cohorts, while models incorporating margins (DL-R and DL-CPR) demonstrated less reliable associations. Multivariable Cox regression confirmed that high risk scores from DL models were independent predictors of MFS. The DL-CP model outperformed FNCLCC grading in prognostic accuracy, with c-indices ≥ 0.74 in validation cohorts. Adding tumor margin information did not improve predictions.DL models focusing on tumor center and periphery provide superior prognostic value in STS, offering a streamlined, effective approach for digital pathology-based risk stratification.

Keywords Deep learning, Artificial intelligence, Digital pathology, Soft-tissue sarcoma, Histologic grading, Prognosis prediction

Abbreviations

C-index	Harrell concordance index
CI	Confidence interval
CNN	Convolutional neural network
CV	Cross validation
DDLPS	Dedifferentiated liposarcoma
DL-C	Deep learning score taking the tumor centrum
DL-CP	Deep learning score taking the tumor centrum and periphery as inputs
DC-CPR	Deep learning score taking the tumor centrum and periphery and R margin as inputs
DL-P	Deep learning score taking the tumor periphery as input
DL-R	Deep learning score taking the R margin as input

[1]Department of Surgical Oncology, Institut Bergonie, Bordeaux, France. [2]Department of Data sciences, Institut Bergonie, Bordeaux, France. [3]Department of Pathology, Institut Bergonie, Bordeaux, France. [4]Department of Pathology, Gustave Roussy, Villejuif, France. [5]Department of Oncology, Institut Bergonie, Bordeaux, France. [6]Department of Oncology, Gustave Roussy, Villejuif, France. [7]Department of Surgical Oncology, Institut Curie, Paris, France. [8]Department of Surgical Oncology, Gustave Roussy, Villejuif, France. [9]Bordeaux Institute of Oncology, BRIC U1312, INSERM, Institut Bergonié, Université de Bordeaux, 33000 Bordeaux, France. [10]INRIA, CNRS, Bordeaux, France. [11]Department of Radiology, Bordeaux University Hospital, Bordeaux, France. [12]Department of Radiology, Institut Bergonie, Bordeaux, France. [13]Audrey Michot, Van-Linh Le, Jean-Michel Coindre, François Le Loarer, Olivier Saut and Amandine Crombé have equally contributed to the work. ✉email: crombeamandine2@gmail.com

FNCLCC	French federation of cancer centers
HES	Hematoxylin and eosin staining
HR	Hazard ratio
LFS	Local recurrence-free survival
LGFMS	Low grade fibromyxoid sarcoma
LMS	Leiomyosarcoma
M-RC/LPS	Myxoid-round cell liposarcoma
MFS	Metastatic relapse-free survival
MPNST	Malignant peripheral nerve sheath tumors
OS	Overall survival
P-LPS	Pleomorphic liposarcoma
STS	Soft tissue sarcoma
SS	Synovial sarcoma
UPS	Undifferentiated pleomorphic sarcoma

Soft tissue sarcomas (STS) are the most frequent sarcomas in adults representing approximately 3000–5000 new patients each year in France[1]. Predicting relapse after initial curative surgery in STS patients is challenging due to the heterogeneity of sarcomas in terms of clinical, radiological and histological presentations. The most important prognostic feature is the histological sarcoma grade according to the French federation of cancer centers (FNCLCC), which relies on three criteria: histological differentiation, mitotic count and tumor necrosis[2]. The FNCLCC 3-tier grade is particularly associated with the risk of metastatic relapse, which is the major determinant of patient prognosis[3,4]. However, the FNCLCC grade is imperfectly reproducible from one pathologist to another and provides limited assistance in tumors classified as intermediate grade II, which accounts for 40% of cases, particularly when dealing with small biopsy specimens.

Computational pathology is opening new possibilities for analyzing histological slides[5–7] that may help overcome the shortcomings of the histological grade. Deep learning (DL), notably convolutional neural networks (CNNs) dedicated to computer vision and taking scanned histological slides as input are increasingly used to develop predictive models (or signatures) for various purposes, including cancer subtyping, correlations with mutations of interest, treatment response prediction and prognostication[8–10]. Validated radiomics model can predict the histological type and grade of retroperitoneal sarcomas with excellent performance[11]. In sarcomas, recent studies have demonstrated the potential of deep learning on digital Hematoxylin, Eosin, Saffron-stained (HES) slides from gastrointestinal stromal tumors to accurately predict patient outcomes as well as *PDGFRA* and *KIT* mutational status[12]. DL models have also been developed to identify the five most common subtypes of sarcomas and predict disease-specific survival in leiomyosarcoma patients[6]. Furthermore, DL models trained on HES slides from rhabdomyosarcoma in children and young adults have shown high diagnostic accuracy in predicting molecular subtypes (*PAX3/7-FOXO1* fusion, *RAS*, *MYOD1* and *TP53* alterations) with better results for the prediction of event-free and overall survival compared to molecular and clinical models[13].

Therefore, our aim was to develop a to develop a novel risk score (first continuous, then categorized as appropriate into high risk or low risk) for survival prediction, derived from deep learning analysis of digitalized HES-stained slides, that could meaningfully challenge the prognostic utility of the FNCLCC histological grade. As the management and biology of peripheral sarcomas (located in the trunk walls and limbs) and intra-truncular sarcomas (located in the abdomen and mediastinum) are distinct, we focused our study on retrospective cohorts of patients affected with peripheral sarcomas managed homogeneously in our sarcoma centers[14,15]. Additionally, previous studies have predominantly focused on characteristics of a single tumor area selected within the central tumor bulk, often referred to as the tumor center, despite evidence of the heterogeneous nature of tumors[16,17]. In this study, we aimed to simultaneously investigate several areas of the tumor bulk to implement different DL models, exploring whether tumor areas other than the tumor center, such as the tumor margins, may help refine patient prognosis prediction.

Material and methods

Study design and patients

This two-center study was approved by the institutional review boards of Bergonié Institute (Bordeaux, France) and Gustave Roussy Institute (Villejuif, France), two French sarcoma reference centers, part of the French national network for the diagnosis and treatment of sarcomas (Netsarc+). All methods were performed in accordance with the relevant guidelines and regulations including the Declaration of Helsinki. The need for written informed consent with an opt-out mechanism was waived by the Ethics committee of Bergonié Institute (Bordeaux, France) because of its retrospective nature using pseudonymized data.

The study involved two initial populations, as illustrated in the study flowchart (Supplementary Fig. SF1). The first cohort included all consecutive patients from Bergonié Institute registered in the French 'Base Clinico-Biologique' (BCB) sarcoma database (https://conticabase.sarcomabcb.org/connect), which is approved by the National Committee for Protection of Personal Data (CNIL, no. 910390), between January 1st, 1990 and December 1st, 2020. Inclusion criteria encompassed adult patient (>18 years old) with newly diagnosed non-metastatic primary tumors, located in the limbs or trunk walls, who underwent upfront curative surgery in a sarcoma center, and had available tissue samples for HES slide digitalization.

Exclusion criteria were: atypical lipomatous tumors (which are tumors of intermediate malignancy), patient who received neoadjuvant treatments posing a risk of denaturing the tissue sample, and patient with no available follow-up. The cohort consisted of *213* patients who were randomly allocated into a Training cohort (*70% [149/213*]) and a validation cohort (Validation-1, *30% [64/213*]).

A second population of *95* patients from Gustave Roussy Institute, fulfilling the same inclusion criteria for a period of inclusion between January 1st, 1998 and December 1st, 2016 was used as a second independent validation cohort (Validation-2). Hence, three distinct cohorts were studied: a Training cohort from Bergonié Institute, an internal Validation-1 cohort (or testing set) also from Bergonié Institute, and an external Validation-2 cohort from Gustave Roussy. We choose this approach to ensure a rigorous evaluation of model generalizability at multiple levels. Using a separate internal Validation-1 cohort from the same institution as the training set allowed us to assess the model ability to generalize to unseen patients from the same clinical setting, while mitigating the risk of overfitting to the training data. The external Validation-2 cohort from Gustave Roussy was used to further evaluate the model robustness across institutions, encompassing potential differences in clinical practice, population characteristics, or image acquisition.

The primary endpoint was metastatic relapse-free survival (MFS), defined as the time (in months) from curative surgery to the occurrence of a distant metastatic relapse or last patient follow-up. In accordance with prior sarcoma studies[3,4], patients who died without experiencing a metastatic relapse were censored at the time of death. Although non systematic, this cause-specific definition of MFS was chosen to allow direct comparability with earlier literature and to maintain consistency with Cox regression analyses, while acknowledging that death represents a competing risk for the event of interest.

Other endpoints included overall survival (OS) and local relapse-free survival (LFS), defined as the time elapsed from curative surgery to death and local relapse, respectively. The clinical and follow-up data of all included patients were updated for the purpose of this study.

Data collection

We extracted the following variables from the BCB sarcoma database: age at diagnosis, histological type (further categorized as in the Sarculator nomogram, i.e., leiomyosarcoma (LMS), dedifferentiated liposarcoma, myxoid/round cell liposarcoma [M/RC-LPS], malignant peripheral nerve sheath tumor [MPNST], myxofibrosarcoma, synovial sarcoma (SS), undifferentiated sarcoma, vascular sarcoma and others)[18] and FNCLCC histological grade performed on the complete surgical specimen (all reviewed by referent pathologists in sarcoma). Regarding M/RC-LPS, according to the last WHO classification, myxoid and round cell liposarcomas belong to the same histotype and the distinction between both subtype may be subjective. In fact myxoid liposarcomas were graded 1 or 2 according to the presence of necrosis or not, and round cell liposarcomas were graded 2 or 3 according to the presence of necrosis and to the mitotic index. The data collection also comprised tumor location (categorized as upper limb, lower limb or trunk wall), depth (categorized as superficial [i.e., located entirely above the superficial muscular fascia and not invading it], deep [i.e., located beneath the superficial muscular fascia or invading it] or deep and superficial [i.e., when the tumor extends across the fascia, involving both superficial and deep compartments]) and size (in mm), surgical margins (categorized as: R0 or R1), adjuvant chemotherapy and radiotherapy, dates of curative surgery, local relapse, metastatic relapse and death or last follow-up.

Review and annotation of histology slides

For all included patients, all slides from the surgical specimen were entirely reviewed by senior pathologists (F.L.L., C.N., J.M.C.), with a total of *2* to *87* slides per case (mean: *21.6* and median: *18*) and *1* to *46* slides with tumor (mean: *11.7* and median: *10*). At least one slide per case was selected for having a representative view of tumor centrum, tumor periphery and peritumoral tissue, and *1* to *4* slides were selected per case with a total of *407* slides. When multiple slides were available, we selected the most cellular one with the highest number of mitoses, as we do when determining the FNCLCC grade. Tumor centrum (C) was defined as the tumor tissue located far from the periphery of the tumor and at least at a distance of 10 mm from its margin for the smallest tumors. Tumor periphery (P) was defined as the tumor tissue between the margin and 10 mm from this margin. Peritumoral tissue (R) was defined as the margin non-tumoral tissue in direct continuity with the tumor area. Figure 1A depicts the demarcation of each area reported on whole slide images (WSI). These HES slides were digitalized using a Hamamatsu Nanozoomer S360 scanner at 40-fold magnification (Hamamatsu, Japan). The TOOLKIT software was used for pseudonymization. The digital slides were annotated with the NDP.view2 software (Hamamatsu, Japan) Annotation of the 3 areas (C, P, R) was done on digital slides with the annotation tool of the Hamamatsu Nanozoomer Series scanner and subsequently submitted for DL analysis. For tumors with infiltrative margins, immunohistochemistry was used to precisely defined the boundary between the tumor and normal tissues, particularly with dedifferentiated liposarcomas (HMGA2 and MDM2 antibodies). However, defining the margins between apparently normal (R) and tumor tissues (P) could be challenging for undifferentiated pleomorphic sarcoma (UPS) and myxofibrosarcoma given the lack of specific markers for these histological subtypes, with particular difficulty in the case of myxofibrosarcoma due to their often infiltrative nature. Among the 31 patients with myxofibrosarcoma, 9 were well circumscribed, 12 showed focal infiltration, and 10 were diffusely infiltrative. In the latter cases, we considered the peripheral tissue to be normal (R) when histological examination did not reveal definite tumor cells, and defined the peripheral tumor tissue (P) as the transitional zone containing both tumor and normal tissue. We acknowledge that this method may introduce a part of subjectivity, which could affect consistency.

The cohort included different tumor histotypes, with 6 sub-groups retained *in fine* with UPS, myxofibrosarcoma, LMS, M/RC-LPS, other liposarcoma (dedifferentiated and pleomorphic) and others with SS,, low grade fibromyxoid sarcoma (LGFMS), and other rarer entities (malignant solitary fibrous tumours, extraskeletal myxoid chondrosarcoma, rhabdomyosarcoma, MPNST, extraskeletal osteosarcoma, clear cell sarcoma, alveolar soft tissue sarcoma, angiosarcoma and epithelioid sarcoma).

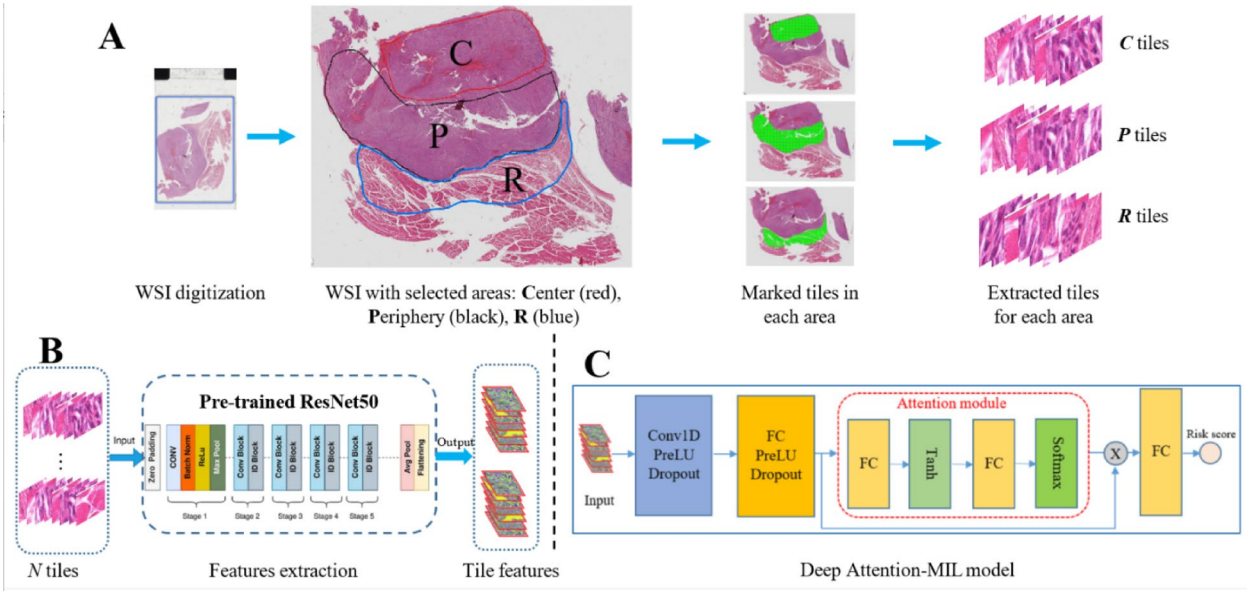

Fig. 1. The workflow of three-stage to predict metastasis risk by using DL model: (**A**) Demarcation of each area reported on whole slide images (WSI). Ct : central of tumor, P: periphery of tumor, R: adjacent non tumoral tissue. The pre-processing WSIs to extract the tiles and these features; (**B**) Features extraction by using a pre-trained DL extractor (e.g. ResNet50), (**C**) Deep Attention-MIL model to predict the risk of metastasis for each patient.

Development of DL models from digital WSI

We used a three-stage procedure (Fig. 1B) to generate the risk score for each patient based on different areas of digitized tumor tissue and their combinations: (i) we extracted image tiles (patches) from selected areas of the HE slides; (ii) we computed imaging features from each tile using a pre-trained deep learning model, e.g., ResNet50[19]; and (iii) we derived an attention-based deep learning model to predict the metastasis risk score for each patient using imaging features from the tiles. These steps are described in detail below.

Tile extraction

We first extract the tiles associated with each area of interest (from the center of the tumor to the R margins). These zones were delineated by pathologists with more than 10 years of experience. The extraction step was performed according to the delineated areas on the image at 40X magnification[9]. The original image is divided into non-overlapping tiles with a size of 224 × 224 (W × H) pixels. Based on the proportion of tissue, the tiles are divided into 4 groups: group A consists of the tiles that contain more than 80% tissue, group B includes the tiles that contain more than 10% and less than 80% tissue, group C includes the tiles that contain less than 10% tissue, and group D includes the tiles that contain no tissue. In this work, only tiles containing of at least 10% tissue were considered (i.e. groups A, B). The number of tiles depends on the size of the selected area and can vary from a few tens of thousands to several hundred thousand tiles. The average number of tiles per patient in the Bergonié and Gustave Roussy cohorts is around 90 K and 67 K respectively. The average number of tiles per area (C, P and R) per patient in the Bergonié cohort corresponds to 38 K, 35 K, 18 K tiles; these values in the Gustave Roussy cohort are 32 K, 23 K, 12 K for areas C, P and R respectively.

Feature extraction

This step is performed using a pre-trained deep learning model, ResNet50[19], to capture 2048 features from each tile. Therefore, for each slide, we obtain a matrix of N (tiles) × 2048 (features) (where *N* is the number of extracted tiles of the slide). Since it was not feasible to use all tiles from an entire slide due to computational constraints, we instead randomly sampled a subset of 10,000 tiles per epoch for training the models[9].

Generating DL scores for all patients[20]

To generate the DL scores for each patient based on the input tiles of each area and realistic incremental combinations of these areas, we used attention-based deep multiple instance learning[21]. Our model uses the extracted features from the tiles (of the WSI) and outputs the risk score for each patient. This model can be broken down into three parts consists of the layers before the attention module to aggregate the features of each tile; (ii) the second part is an attention module[21], which provides the relative importance weights for each tile and aggregates the tile features to obtain features at the slide level; (iii) the layers in the last part are used to predict the risk of metastasis for each patient (Fig. 1C). The model was trained to predict MFS as a time-to-event outcome using a survival-based deep learning framework. Right-censored data (i.e., patients without metastasis at last follow-up) were incorporated through a loss function adapted for censored observations, such as the negative partial log-likelihood from the Cox proportional hazards model. This allows the model to learn from

both censored and uncensored patients without biasing survival estimates. The risk scores were normalized to a fixed range of [− 1, 1] to harmonize the results of different datasets. The details of the DL model and its implementation details are described in Supplementary Method SM1. The algorithm has been trained on Bergonié cohort using different tumor areas (C, P, R) and various combinations to produce five DL models. More precisely, these models produced five DL risk scores: DL-C (for tumor centrum alone), DL-P (for the tumor periphery alone), DL-R (for tumor margin R), DL-CP (combination of tumor centrum and periphery) and DL-CPR (combination of all areas). The DL models were then applied on the Validation cohorts to generate the corresponding DL scores. The normalization of DL scores to the fixed range [− 1, 1] was performed using the minimum and maximum DL score values calculated from the Training cohort to ensure consistency. Finally, the DL scores and the median score on the Training cohort were used to discriminate patients (in each cohort) into 'low risk' (< median DL score in the training cohort) and 'high risk' (≥ median DL score in the Training). Similarly, for binarization, we computed the median of each DL score in the Training cohort and used these thresholds to dichotomize the scores in the Validation cohorts.

Comprehensive histological analysis of the DL score

To better understand the outputs of the best-performing DL model, we selected a total of 20 patients: 5 patients with the lowest DL scores and 5 patients with the highest DL scores from the Training cohort, and similarly, 5 patients with the lowest DL scores and 5 patients with the highest DL scores from the Validation-1 and Validation-2 cohorts. Next, we extracted 50 tiles per patient, which were retrospectively analyzed by a senior pathologist with 45 years of experience in sarcoma pathology (J.M.C.), blinded to any patient data or model output. The pathologist analyzed these 20 patients and collected the following variables: tumor cellularity (categorized as: < 10%, 10–50% or > 50%), tumor stroma (categorized as: absent, chondroid, fibromyxoid, fibrosis or myxoid), main cell type (categorized as: epitheloid, pleomorph, round or spindle cells), atypia (categorized as: no or very mild [0], moderate [1] or severe [2]), hyperchromasia (categorized as: absent or present), mitosis (categorized as: absent, 1 mitosis or 2 mitosis), necrosis (categorized as absent or present), red blood cells (categorized as absent or present), tumor differentiation (categorized as: absent, chondroid or smooth muscle), tumor infiltration (categorized as: absent or present [by lymphocytes ± plasmocytes]) and vessels (categorized as absent or present).

Statistical analysis

Statistical analyses were performed with R (v4.1.0, The R foundation for Statistical Computing, Vienna, Austria). All tests were two-tailed. Significance was set at $p < 0.05$. Survival analysis utilized the 'survival', 'pec', and 'survminer' packages[22].

Descriptive and exploratory analysis

Descriptive statistics were presented for categorical variables as numbers and percentages, and for numeric variables as mean ± standard deviation, or median with range (minimum–maximum) and interquartile range (Q1–Q3). In the entire population, correlations between continuous DL scores and tumor size were assessed using the Spearman rank test. Associations between continuous DL scores and histologic types and grade were investigated using one-way analysis of variance (ANOVA-1) or the Friedman rank test with post-hoc Tukey or Mann–Whitney tests (depending on Shapiro-Wild normality test), corrected for multiple comparisons with the Benjamini–Hochberg adjustment.

Univariable MFS analyses including competing risk analysis

Since deaths may preclude the occurrence of metastatic relapse, cumulative incidence functions (CIF) for distant metastasis were estimated, treating death without metastasis as a competing event. Gray's test was used to compare CIFs between DL-score groups and FNCLCC grades. This approach was used solely for descriptive visualization of metastatic incidence under competing risks for the main study objective and did not modify the Cox regression analyses. Classical Kaplan–Meier curves for MFS were also drawn for the five dichotomized DL-scores and all covariables in the Training, Validation-1 and Validation-2, and difference in survivals were tested with the log-rank test. Univariable Cox regressions estimated hazard ratios (HRs) with 95% confidence intervals (CIs). A subgroup analysis for grade II STS patients was also conducted.

Multivariable analyses

We then designed a multivariable analysis similar to that of the Sarculator nomogram to investigate the prognostic significance of the five DL scores and compared them with an analogous model incorporating the FNCLCC grade. For each dichotomized score and grade (I vs. II–III, as per the latest ESMO guidelines)[14], multivariable Cox regressions including size (continuously), age (continuously), histologic type (categorical variable with M/RC-LPS as the reference category) and the score of interest were trained in the Training cohort. The models were then applied on Validation-1 and Validation-2, and their performances in those two validation cohorts were estimated with the Harrell concordance index (c-index), which ranges from 0 (worst possible) to 1 (perfect model)[23]. Pairwise c-index comparisons were conducted via bootstrapping on 1000 replicates ('boot' package). Calibration curves were plotted for all multivariable models[22].

Other survival outcomes

The Kaplan–Meier curves for LFS and OS were generated for the five dichotomized DL-scores in the Training, Validation-1 and Validation-2 cohorts, complemented with log-rank tests.

Histological features associated with high and low risk DL scores
Associations between categorical histological features and the low- and high-risk groups were evaluated using Chi-square tests, both across the entire tile dataset and stratified by tile location.

Results

Patient and tumor characteristics in the three cohorts

Patient characteristics are summarized in Table 1. In the Training cohort, 75/149 patients (50.3%) were women, compared to 23/64 (35.9%) in Validation-1 and 37/95 (28.9%) in Validation-2 ($P = 0.0781$, Chi-square test). The median age was 67 years (Q1–Q3: 52–79) in Training, 52.5 years (Q1–Q3: 37.5–75) in Validation-1, and 58 years (Q1–Q3: 44.5–72) in Validation-2 ($P = 0.0011$, ANOVA-1). Tumor size averaged 96.1 ± 57 mm in Training, 96 ± 52.4 mm in Validation-1, and 84 ± 44 mm in Validation-2 ($P = 0.1720$, ANOVA-1). The proportion of patients with FNCLCC grade III sarcomas was 70/149 (47%) in Training, 27/64 (42.2%) in Validation-1, and 52/95 (54.7%) in Validation-2 ($P = 0.3110$, Chi-square test). Metastatic relapses were observed in 45/149 (30.2%) patients in Training (5 year MFS probability: 67.3 months, 95%CI 59.6–76.1), 20/64 (31.3%) in Validation-1 (5 year MFS probability: 70.3 months, 95%CI 59.1–83.6), and 33/95 (34.7%) in Validation-2 (5 year MFS probability: 63.8 months 95%CI 54.3–75.1) (Supplementary Table ST1). There were no significant differences in MFS probabilities among the three cohorts ($P = 0.8000$, log-rank test).

Understanding the DL risk scores

All continuous and binarized DL risk scores exhibited strong associations with both the FNCLCC grade and histologic types (all P values < 0.0001, ANOVA-1 and Chi-square tests), demonstrating higher scores in leiomyosarcomas and the undifferentiated sarcoma group, as well as in grade III tumors (Supplementary Table ST2, Supplementary Fig. SF2). However, no association was observed between the DL risk scores and tumor size. Additionally, significant correlations were found between the risk scores of DL-C and DL-R (Spearman rho = 0.258, 95%CI 0.146–0.358, $P < 0.0001$), as well as between the risk scores of DL-P and DL-R (Spearman rho = 0.305, 95%CI 0.202–0.403, $P < 0.0001$).

Univariable survival analyses

Figure 2 depicts the CIF curves for DL-C, DL-P, DL-R, DL-CP, DL-CPR and grade in the 3 cohorts with corresponding P values according to the Gray test (Supplementary Fig. SF3 represents the corresponding classical Kaplan–Meier curves). Univariable analysis for the dichotomized DL risk scores and clinical and histological covariables are detailed in Table 2. In the Training cohort, all DL risk scores showed significant associations with MFS according to log-rank tests (P value range: 0.0004 [for DL-C] to < 0.0001 [for all other scores]) and Gray's test (P value range: 0.0004 [for DL-C] to < 0.0001 [for all other scores]). However, DL-R groups did not exhibit associations with MFS in Validation-1 (log-rank $P = 0.0979$ and Gray's test $P = 0.0935$) or in Validation-2 cohort (log-rank $P = 0.3696$ and Gray's test $P = 0.7485$). Moreover, the DL-CPR score did not demonstrate associations with MFS in Validation-2 (log-rank $P = 0.0747$ and Gray's test $P = 0.3622$). In contrast, all other combinations revealed significant associations between high-risk groups and lower MFS. Regarding histologic grade, it was associated with MFS in Training and Validation-2 (log-rank $P = 0.0122$ and $P = 0.0002$, respectively; Gray's test $P = 0.0238$ and 0.0005, respectively), but not in Validation-1 (log-rank $P = 0.0712$ and Gray's test $P = 0.0565$).

Subgroup analysis in patients with grade II tumors

There were 112/308 (36.4%) grade II patients in total. Among these patients, 25/112 (22.3%) experienced metastatic relapses, with 54 patients in Training (including 12 [22.2%] metastatic relapses), 24 patients in Validation-1 (including 7 [29.2%] metastatic relapses) and 28 patients in Validation-2 (including 6 [21.4%] metastatic relapses). Univariable survival analysis depending on the DL risk score is given in Supplementary Table ST3 and the Kaplan–Meier curve in Supplementary Fig. SF4. No significant differences were observed with dichotomized DL-R and DL-CPR scores across all cohorts. Regarding the model DL-P, a significantly lower MFS probability was found in the high-risk group in the Training and Validation-1 cohorts ($P = 0.0014$ and $P = 0.0176$, respectively, log-rank tests).

For DL-C and DL-CP risk scores, a similar association was only evident in Training ($P = 0.0199$ and $P = 0.0038$, respectively) but not in Validation-1 and Validation-2. Notably, in Validation-2, no significant associations were found regardless of the DL risk score tested.

Multivariable analysis

The results for all multivariable modeling with HRs for each input variable are detailed in Table 3. Each DL risk score emerged as an independent predictor of MFS. Notably, the Cox regression trained on the Training cohort revealed lower survival rates in the high-risk group across all DL risk scores. Specifically, the HR values were as follows: 3.28, 95% CI 1.56–6.89 ($P < 0.0001$) for DL-C; 23.41, 95% CI 6.98–78.53 ($P < 0.0001$) for DL-P; 3.74, 95% CI 1.76–7.96 ($P < 0.0001$) for DL-R; 8.82, 95% CI 3.39–22.96 ($P < 0.0001$) for DL-CP; and 6.86, 95% CI 2.91–16.18 ($P < 0.0001$) for DL-CPR. The FNCLCC grade (with its 3 categories) did not exhibit a significant association with MFS in the Training cohort ($P = 0.5692$ for grade II and $P = 0.1021$ for grade III, respectively, with grade I as the reference). However, when combining grade II and grade III (with grade I as the reference), a significant and independent association with MFS was observed (HR = 1.87, 95% CI 1–1.05, $P = 0.0444$), prompting the adoption of this combination for subsequent analyses.

Comparing the predictive performances, in Validation-1, a highest c-index was reached with the DL-CPR model (c-index = 0.786) followed by DL-C model (c-index = 0.759). In Validation-2, the highest c-index was reached with the DL-C model (c-index = 0.741) followed by DL-CP (c-index = 0.698) and DL-CPR models (c-index = 0.698). Notably, the c-indices of a similar model substituting risk scores with grade were 0.739 in

Characteristics	All patients (*N* = 308)	Training cohort (*n* = 149)	Validation-1 cohort (*n* = 64)	Validation-2 cohort (*n* = 95)
Sex				
Female	135/308 (43.8%)	75/149 (50.3%)	23/64 (35.9%)	37/95 (38.9%)
Male	173/308 (56.2%)	74/149 (49.7%)	41/64 (64.1%)	58/95 (61.1%)
Age at diagnosis (years)				
Mean +/− SD	59.4 +/− 19.3	63.4 +/− 18.4	53.5 +/− 21.1	57.2 +/− 18.2
Median [Q1–Q3] (range)	61 [46.5–76] (14–94)	67 [52–79] (16–94)	52.5 [37.5–75] (14–90)	58 [44.5–72] (17–92)
Tumor size (mm)				
Mean +/- SD	92.4 +/− 52.5	96.1 +/− 57	96 +/− 52.4	84 +/− 44
Median [Q1–Q3] (range)	80 [50–120] (10–310)	80 [60–120] (20–310)	82.5 [57.5–130] (13–275)	80 [50–115] (10–230)
Tumor location				
Lower limb	185/308 (60.1%)	93/149 (62.4%)	40/64 (62.5%)	52/95 (54.7%)
Trunk wall	82/308 (26.6%)	43/149 (28.9%)	13/64 (20.3%)	26/95 (27.4%)
Upper limb	41/308 (13.3%)	13/149 (8.7%)	11/64 (17.2%)	17/95 (17.9%)
Tumor depth				
Superficial	36/308 (11.7%)	17/149 (11.4%)	4/64 (6.2%)	15/95 (15.8%)
Deep	206/308 (66.9%)	97/149 (65.1%)	44/64 (68.8%)	65/95 (68.4%)
Superficial and deep	66/308 (21.4%)	35/149 (23.5%)	16/64 (25%)	15/95 (15.8%)
Histologic type§				
Undifferentiated sarcoma	86/308 (27.9%)	42/149 (28.2%)	9/64 (14.1%)	35/95 (36.8%)
Myxofibrosarcoma	31/308 (10.1%)	19/149 (12.8%)	7/64 (10.9%)	5/95 (5.3%)
Leiomyosarcoma	47/308 (15.3%)	18/149 (12.1%)	9/64 (14.1%)	19/95 (20%)
Myxoid/round cell liposarcoma	33/308 (10.7%)	19/149 (12.8%)	8/64 (12.5%)	6/95 (6.3%)
Dedifferentiated liposarcoma	38/308 (12.3%)	17/149 (11.4%)	11/64 (17.2%)	11/95 (11.6%)
Other sarcoma	67/308 (23.7%)	34/149 (22.8%)	20/64 (31.2%)	13/95 (13.7%)
MPNST	9/308 (2.9%)	3/149 (2%)	1/64 (1.6%)	5/95 (5.3%)
Synovial sarcoma	13/308 (4.2%)	5/149 (3.4%)	6/64 (9.4%)	2/95 (2.1%)
Vascular sarcoma	2/308 (0.6%)	2/149 (1.3%)	0/64 (0%)	0/95 (0%)
Other sarcoma	43/308 (14%)	24/149 (16.1%)	13/64 (20.3%)	6/95 (6.3%)
FNCLCC grade				
I	47/308 (15.3%)	25/149 (16.8%)	13/64 (20.3%)	9/95 (9.5%)
II	112/308 (36.4%)	54/149 (36.2%)	24/64 (37.5%)	34/95 (35.8%)
III	149/308 (48.4%)	70/149 (47%)	27/64 (42.2%)	52/95 (54.7%)
Surgical margins				
R0	188/306 (61.4%)	83/149 (55.7%)	33/64 (51.6%)	72/93 (77.4%)
R1	118/306 (38.6%)	66/149 (44.3%)	31/64 (48.4%)	21/93 (22.6%)
Adjuvant chemotherapy				
No	263/308 (85.4%)	124/149 (83.2%)	52/64 (81.2%)	87/95 (91.6%)
Yes	45/308 (14.6%)	25/149 (16.8%)	12/64 (18.8%)	8/95 (8.4%)
Adjuvant radiotherapy				
No	77/308 (25%)	35/149 (23.5%)	11/64 (17.2%)	31/95 (32.6%)
Yes	231/308 (75%)	114/149 (76.5%)	53/64 (82.8%)	64/95 (67.4%)

Table 1. Characteristics of the study population. *FNCLCC* French federation of cancer centers, *MPNST* malignant peripheral nerve sheath tumor, *Q* quartile, *SD* standard deviation. § Histological type are categorized according to the grouping of the Sarculator nomogram. In details, dedifferentiated liposarcoma corresponds to pleomorphic and dedifferentiated liposarcoma. Undifferentiated sarcoma corresponds to undifferentiated pleomorphic sarcoma, undifferentiated round cell sarcoma, undifferentiated sarcoma (not otherwise specified) and undifferentiated spindle cell sarcoma. Vascular sarcoma corresponds to epitheloid hemangioendothelioma and angiosarcoma. Other sarcoma corresponds to synovial sarcoma, alveolar soft part sarcoma, clear cell sarcoma, soft-tissue dedifferentiated chondrosarcoma, epitheloid sarcoma, extraskeletal myxoid chondrosarcoma, high risk solitary fibrous tumors, low grade fibromyxoid sarcoma, malignant rhabdoid tumor, malignant tenosynovial giant cell tumor, and fibrosarcoma.

Validation-1 and 0.723 in Validation-2, with no significant difference observed against DL-based models. Similarly, no significant differences were noted when comparing the DL models against the FNCLCC grade (Supplementary Table ST4). Detailed c-index comparisons and calibration curves for all models across the three cohorts are illustrated in Fig. 3.

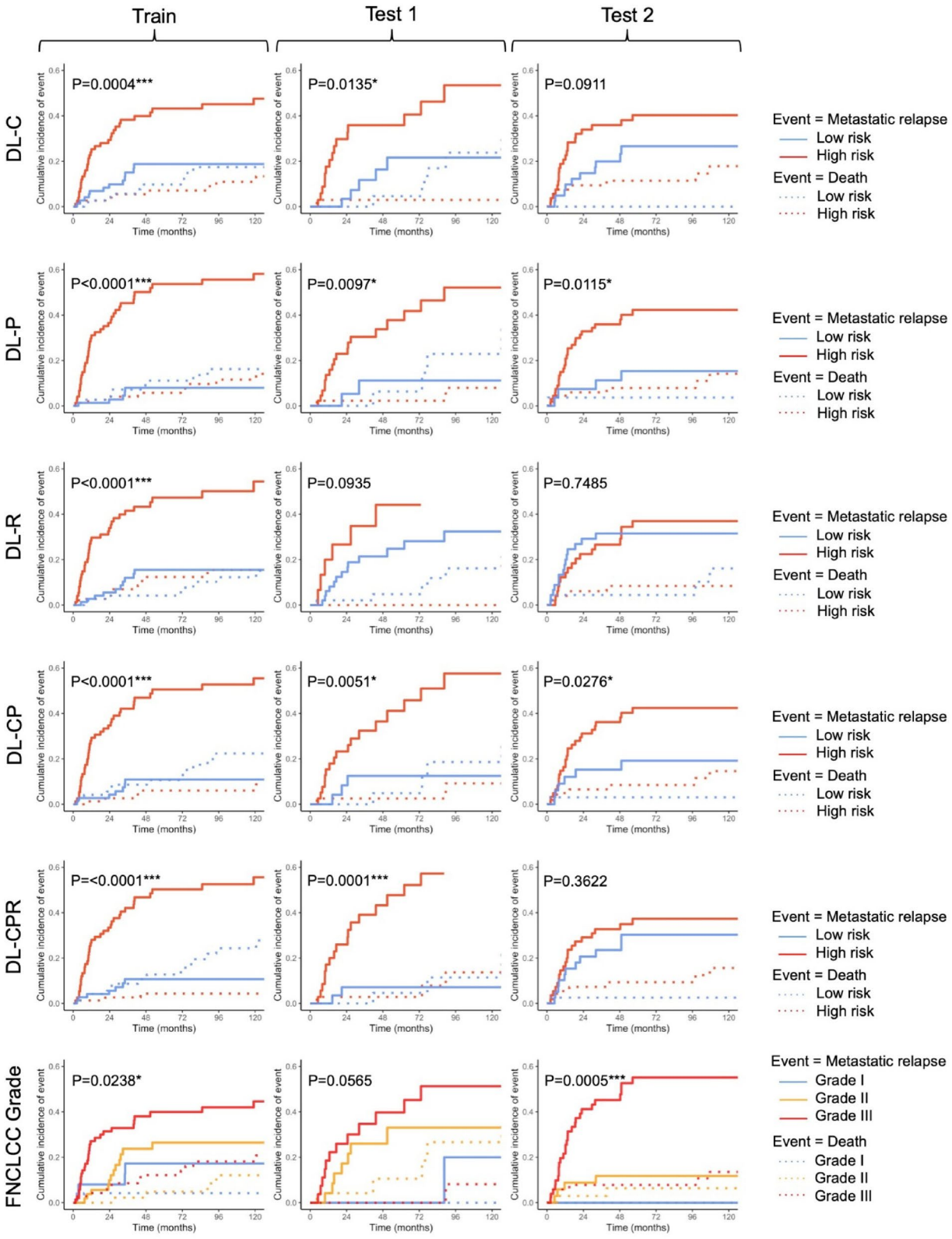

Fig. 2. Cumulative incidence functions (CIF) curve for metastatic relapse (solid lines) and death without prior metastasis (dotted lines) according to the five deep learning (DL) scores and FNCLCC grade, shown separately in the Training, Validation-1, and Validation-2 cohorts. For each panel, the y-axis indicates the cumulative probability of the event of interest over time since curative surgery, accounting for death as a competing risk. DL scores were based on the tumor centrum (C), periphery (P) and surrounding tissues (R) and their incremental combinations (CP, CPR). *P* values from Gray's tests for differences in metastatic relapse-free survival are reported: $^{*}P<0.05$, $^{**}P<0.005$, $^{**}P<0.001$.

Characteristics	Training cohort				Validation-1 cohort				Validation-2 cohort			
	5y MFS probability	Log rank *P* value	HR (95%CI)	Cox *P* value	5y MFS probability	Log rank *P* value	HR (95%CI)	Cox *P* value	5y MFS probability	Log rank *P* value	HR (95%CI)	Cox *P* value
DL-C risk score												
Low risk	80 (71–91)	**0.0004*****	–	–	78 (62–98)	**0.0121***	–	–	73 (60–89)	**0.0456***	–	–
High risk	55 (44–68)		3.12 (1.61–6.05)	**0.0007*****	63 (48–82)		3.39 (1.23–9.33)	**0.0182***	56 (43–72)		2.11 (1–4.46)	0.0505
DL-P risk score							15.45 (3.42–70)	**0.0004*****				
Low risk	92 (85–99)	<**0.0001*****	–	–	89 (75–100)	**0.0112***	–	–	84 (71–100)	**0.0099***	–	–
High risk	44 (34–58)		10.58 (4.17–27)	<**0.0001*****	61 (47–80)		5.41 (1.25–23.34)	**0.0238***	55 (44–70)		3.63 (1.27–10.4)	**0.0159***
DL-R risk score												
Low risk	84 (75– 94)	<**0.0001*****	–	–	75 (62–89)	0.0979	–	–	69 (55–87)	0.3696	–	–
High risk	50 (39– 64)		4.85 (2.39–9.84)	<**0.0001*****	56 (34–91)		2.14 (0.85–5.38)	0.1060	61 (49–76)		1.41 (0.67–2.96)	0.3717
DL-CP risk score												
Low risk	88 (80– 97)	<**0.0001*****	–	–	88 (75–100)	**0.005***	–		80 (67–96)	**0.0228***	–	–
High risk	48 (38–62)		6.49 (2.9–14.55)	<**0.0001*****	58 (42–79)		4.92 (1.44–16.83)	**0.0112***	55 (43–70)		2.69 (1.11–6.55)	**0.0286***
DL-CPR risk score												
Low risk	89 (81–97)	<**0.0001*****	–	–	93 (84–100)	<**0.0001*****	–	–	100 (100–100)	0.0749	–	–
High risk	49 (38–62)		6.3 (2.81–14.12)	<**0.0001*****	51 (35–73)		10.27 (2.38–44.36)	**0.0018****	61 (51–73)		*Inf (0—Inf)*	*0.9971*
Grade												
I	82 (68–100)	**0.0122***	–	–	100 (100–100)	0.0712	–	–	100 (100–100)	**0.0002*****	–	–
II	73 (61–88)		1.4 (0.45–4.34)	0.5610	65 (46–91)		4.86 (0.60–39.61)	0.1396	88 (78–99.76)		*Inf (0—Inf)*	*0.9971*
III	58 (47–72)		3.08 (1.08–8.78)	**0.035***	60 (44–83)		7.53 (0.98–58.02)	0.0525	41 (28–58)		*Inf (0—Inf)*	*0.9968*
Sex												
Female	68 (58–81)	0.6427	–	–	82 (67–100)	0.6228			68 (55–86)	0.8821		
Male	66 (55–79)		1.15 (0.64–2.06)	0.6430	64 (50–83)		1.27 (0.49–3.31)	0.6236	61 (49–76)		0.95 (0.47–1.91)	0.8833
Age (continuous, years)	–	–	1.03 (1.01–1.05)	**0.0032****	–	–	1.02 (1–1.05)	0,0696			1.02 (1—1.04)	**0.0366***
Tumor size (continuous, mm)	–	–	1.01 (1.01–1.01)	<**0.0001*****	–	–	1.01 (1–1.01)	0,1373			1 (1–1.01)	0.4296
Histologic types§												
Liposarcoma,myxoid	73 (53–100)	0.4298	–	–	89 (71–100)	0.3475			89 (71–100)	0.2015		
Leiomyosarcoma	31 (11–88)		3.06 (0.92–10.2)	0.0690	50 (27–93)		6.79 (0.82–56.47)	0.0764	48 (29–80)		7.06 (0.89–56)	0.0641
Liposarcoma	72 (55–96)		1.41 (0.38–5.24)	0.6124	83 (58–100)		3.27 (0.34–31.48)	0.3051	86 (63–100)		1.4 (0.09–22.51)	0.8102
Myxofibrosarcoma	62 (41–92)		1.58 (0.45–5.6)	0.4794	86 (63–100)		1.43 (0.09–22.94)	0.7989	75 (43–100)		2.64 (0.16–42.2)	0.4931
Other	68 (53–87)		1.5 (0.47–4.8)	0.4903	68 (49–93)		3.77 (0.45–31.45)	0.2198	53 (33–85)		6.09 (0.77–48)	0.0870
Undifferentiated sarcoma	75 (63–90)		1.36 (0.44–4.21)	0.5978	58 (31–100)		4.31 (0.44–41.81)	0.2079	64 (49–84)		4.56 (0.59–35)	0.1455
Tumor depth												
Continued												

Characteristics	Training cohort				Validation-1 cohort				Validation-2 cohort			
	5y MFS probability	Log rank *P* value	HR (95%CI)	Cox *P* value	5y MFS probability	Log rank *P* value	HR (95%CI)	Cox *P* value	5y MFS probability	Log rank *P* value	HR (95%CI)	Cox *P* value
Superficial	100 (100–100)	**0.0148***	–	–	25 (5–100)	**0.0323***			93 (80–100)	0.0993		
Deep	67 (57–78)		Inf (0—Inf)	0.9959	72 (59–88)		0.21 (0.06–0.78)	**0.0198***	60 (49–74)		6.23 (0.84–46)	0.0727
Superficial and deep	54 (38–76)		Inf (0—Inf)	0.9958	80 (63–100)		0.22 (0.05–1.01)	0.0519	54 (33–89)		7.44 (0.89–62)	0.0634
Tumor location												
Lower limb	65 (56–77)	0.8252	–	–	71 (57–87)	0.729			68 (56–83)	0.3533		
Trunk wall	72 (59–88)		0.81 (0.40–1.61)	0.5423	68 (46–100)		1.26 (0.41–3.9)	0.6844	53 (36–79)		1.54 (0.7–3.35)	0.2803
Upper limb	67 (45–100)		0.89 (0.31–2.53)	0.8303	76 (51–100)		0.64 (0.14–2.83)	0.555	68 (49–96)		1.78 (0.73–4.38)	0.2073
Surgical margins												
R0	78 (69–89)	**0.0023****	–	–	75 (61– 92)	0.8078			64 (53–77)	0.9956		
R1	53 (42–68)		2.48 (1.35–4.54)	**0.0032****	65 (48–87)		1.12 (0.46–2.68)	0.8079	61 (42–89)		1 (0.45–2.24)	0.9955
Adjuvant chemotherapy												
No	65 (56–75)	0.1727	–	–	68 (55– 84)	0.6544			65 (55–76)	0.6543		
Yes	79 (64–97)		0.55 (0.23–1.31)	0.1789	75 (54–100)		0.78 (0.25–2.37)	0.6552	54 (26–100)		1.27 (0.44–3.65)	0.6564
Adjuvant radiotherapy												
No	64 (50–83)	0.3059	–	–	44 (17–100)	0.4530			68 (53–88)	0.9391		
Yes	69 (60–79)		0.71 (0.37–1.37)	0.3083	74 (62–87)		0.66 (0.22–1.99)	0.4563	62 (51–76)		0.97 (0.46–2.05)	0.9416

Table 2. Univariable survival analysis for metastatic relapse-free survival in the three cohorts. *CI* confidence interval, *DL* deep learning risk score (based on *C* tumor centrum, *P* tumor periphery and *R* tumor surrounding tissues, and their combination), *HR* hazard ratio, *no*. number. *: $P < .05$, **: $P < 0.005$, ***: $P < 0.001$. Significant results are in bold. §: As in the Sarculator nomogram, myxoid/round cells liposarcoma are the reference level.

Retrospective histological review of the DL model output

Since the DL models incorporating both centrum and periphery tiles consistently showed associations with MFS across all three cohorts, we focused our analysis on the DL-CP model. Supplementary Table ST5 presents the associations between the DL-CP low and high risk groups and the histological features, analyzed across the entire tile dataset, as well as separately for tiles from the tumor periphery and centrum. Several significant associations were observed. The DL-CP high-risk group was characterized by higher tumor cellularity in both the centrum ($P = 0.0002$) and periphery ($P = 0.0002$), a predominance of pleomorphic cell types in the centrum ($P = 0.0129$) and periphery ($P = 0.0038$), and more pronounced atypia in both regions (centrum: $P = 0.0029$; periphery: $P = 0.0036$). Conversely, myxoid or fibromyxoid stroma was more commonly seen in the DL-CP low-risk group, both in the centrum ($P = 0.0054$) and periphery ($P = 0.0043$). Additionally, increased mitotic activity was noted in the tumor periphery of the high risk group ($P = 0.0357$). Representative tiles from the DL-CP low and high risk groups are shown in Fig. 4.

Other outcomes

We provide the results of the models for both LFS and OS in Supplementary Tables ST6 and ST7, respectively. A notable association with LFS was observed solely with the DL-P risk score in Training (log-rank $P = 0.0126$). In Training, all DL risk scores demonstrated significant associations with OS (log-rank *P* value range: < 0.0001 [for DL-P and DL-R] to 0.0057 [for DL-C]). Conversely, in Validation-1, only the DL-CPR score exhibited a significant association with OS (log-rank $P = 0.0009$), while neither the other scores nor the FNCLCC grade showed significance. Similarly, in Validation-2, the DL-P score displayed a significant association with OS (log-rank $P = 0.0123$), whereas the other scores and the FNCLCC grade did not.

Discussion

The prognostication and risk stratification of patients with STS pose significant challenges. In cases of newly-diagnosed locally-advanced STS, patient outcomes are primarily linked to the occurrence of metastatic relapses, heavily influenced by the FNCLCC histologic grade. This grade dictates treatment strategies, including anthracyclines-based chemotherapy and radiotherapy, in addition to curative surgery[14]. However, accurately

Predictors	HR	*P* value	c-index in Training	c-index in Validation-1	c-index in Validation-2
DL-C model					
Size (mm)	1.01 (1.01–1.02)	< **0.0001*****	0.793 (0.732–0.854)	0.759 (0.643–0.876)	0.741 (0.662–0.821)
Histologic type (ref: M/RC-LPS)					
Leiomyosarcoma	1.66 (0.42–6.59)	0.4679			
Liposarcoma	0.26 (0.06–1.15)	0.0760			
Myxofibrosarcoma	0.56 (0.13–2.35)	0.4298			
Other	1.41 (0.39–5.11)	0.5987			
Undifferentiated sarcoma	0.54 (0.15–1.98)	0.3547			
DL-C High risk	3.28 (1.56–6.89)	**0.0018****			
Age (years)	1.03 (1–1.05)	**0.0416***			
Chemotherapy (yes)	0.63 (0.24–1.67)	0.3583			
Radiotherapy (yes)	0.57 (0.28–1.16)	0.1201			
DL-P model					
Size (mm)	1.02 (1.01–1.02)	< **0.0001*****	0.851 (0.803–0.900)	0.74 (0.635–0.845)	0.667 (0.573–0.760)
Histologic type (ref: M/RC-LPS)					
Leiomyosarcoma	0.43 (0.09–2.05)	0.2926			
Liposarcoma	0.1 (0.02–0.5)	**0.0051***			
Myxofibrosarcoma	0.16 (0.03–0.8)	**0.0253***			
Other	0.52 (0.13–2.2)	0.3769			
Undifferentiated sarcoma	0.12 (0.03–0.54)	**0.0054***			
DL-P High risk	23.41 (6.98–78.5)	< **0.0001*****			
Age (years)	1.01 (0.99–1.04)	0.2782			
Chemotherapy (yes)	0.62 (0.23–1.67)	0.3451			
Radiotherapy (yes)	0.52 (0.26–1.04)	0.0658			
DL-R model					
Size (mm)	1.01 (1.01–1.02)	< **0.0001*****	0.790 (0.733–0.847)	0.690 (0.564–0.817)	0.664 (0.567–0.760)
Histologic type (ref: M/RC-LPS)					
Leiomyosarcoma	2.6 (0.67–10.03)	0.1656			
Liposarcoma	0.46 (0.12–1.83)	0.2695			
Myxofibrosarcoma	1.19 (0.31–4.53)	0.8036			
Other	1.95 (0.56–6.81)	0.2958			
Undifferentiated sarcoma	0.92 (0.27–3.12)	0.8904			
DL-R (binary)	3.74 (1.76–7.96)	**0.0006*****			
Age (years)	1.02 (1–1.04)	0.1094			
Chemotherapy (yes)	1.01 (0.39–2.63)	0.9865			
Radiotherapy (yes)	0.46 (0.22–0.95)	**0.0347***			
DL-CP model					
Size (mm)	1.01 (1.01–1.02)	< **0.0001*****	0.819 (0.761–0.877)	0.74 (0.623–0.856)	0.698 (0.604–0.791)
Histologic type (ref: M/RC-LPS)					
Leiomyosarcoma	0.57 (0.13–2.51)	0.4605			
Liposarcoma	0.22 (0.05–0.97)	**0.0458***			
Myxofibrosarcoma	0.26 (0.06–1.17)	0.0793			
Other	0.87 (0.23–3.26)	0.8379			
Undifferentiated sarcoma	0.22 (0.06–0.89)	**0.0342***			
DL-CP High risk	8.82 (3.39–22.96)	< **0.0001*****			
Age (years)	1.03 (1.01–1.06)	0.0068*			
Chemotherapy (yes)	0.68 (0.25–1.8)	0.4324			
Radiotherapy (yes)	0.79 (0.38–1.64)	0.5224			
DL-CPR model					

Continued

Predictors	HR	*P* value	c-index in Training	c-index in Validation-1	c-index in Validation-2
Size (mm)	1.01 (1.01–1.02)	< **0.0001*****	0.804 (0.738–0.870)	0.786 (0.689–0.883)	0.698 (0.620–0.775)
Histologic type (ref: M/RC-LPS)					
Leiomyosarcoma	1.76 (0.46–6.67)	0.4082			
Liposarcoma	0.55 (0.14–2.16)	0.3927			
Myxofibrosarcoma	0.69 (0.18–2.61)	0.5808			
Other	1.99 (0.56–7.06)	0.2860			
Undifferentiated sarcoma	0.58 (0.17–1.96)	0.3809			
DL-CPR High risk	6.86 (2.91–16.18)	< **0.0001*****			
Age (years)	1.02 (1–1.05)	0.0693			
Chemotherapy (yes)	0.79 (0.3–2.06)	0.6271			
Radiotherapy (yes)	0.54 (0.27–1.07)	0.0758			
Grade-based model					
Size (mm)	1.01 (1.01–1.02)	< **0.0001*****	0.764 (0.699–0.828)	0.739 (0.621–0.856)	0.723 (0.652–0.794)
Histologic type (ref: M/RC-LPS)					
Leiomyosarcoma	2.61 (0.71–9.65)	0.1493			
Liposarcoma	0.4 (0.1–1.69)	0.2136			
Myxofibrosarcoma	1.09 (0.28–4.28)	0.8981			
Other	2.21 (0.66–7.44)	0.2002			
Undifferentiated sarcoma	0.86 (0.25–2.96)	0.8093			
FNCLCC Grade (ref: grade I)	1.87 (0.59–5.96)	0.2908			
Age (years)	1.02 (1–1.05)	**0.0444***			
Chemotherapy (yes)	0.71 (0.27–1.81)	0.4697			
Radiotherapy (yes)	0.61 (0.3–1.23)	0.1689			

Table 3. Results and performances of the multivariable modeling involving the deep learning risk scores and the FNCLCC grade. *c-index* concordance index, *DL* deep learning risk score (based on C: tumor centrum, P: tumor periphery and R: tumor surrounding tissues, and their combination), *HR* hazard ratio, *M/RC-LPS* myxoid/round cells liposarcoma, *ref* reference. Hazard ratio and c-index are given with 95% confidence intervals. *: $P<0.05$, **: $P<0.005$, ***: $P<0.001$. Significant results are in bold.

assessing the FNCLCC grade requires considerable expertise in STS pathology, which is hindered by the disease's relative rarity and the scarcity of expert pathologists. This can result in prolonged delays before diagnosis, grading, and referral to specialized centers. Therefore, there is a pressing need for precise and reproducible automated histological tools to assist pathologists. In this study, we introduce an original deep learning pipeline leveraging digital pathology, pre-trained CNN, and MIL. Our objectives were twofold: (i) to predict MFS and (ii) to explore whether incorporating normally appearing surrounding tissues (R areas) in the HES digitalized slice could enhance performance compared to standard assessments focusing solely on the tumor center and periphery (C and P areas) or histological grading. Utilizing one training cohort and two independent validation cohorts from two of the three French sarcoma reference centers all annotated by expert pathologists, our approach revealed that including the R area did not improve the performance of DL models already utilizing the C and P areas. Moreover, we found that DL models could outperform models based on grading assessed by senior pathologists in predicting MFS.

First, we observed significant associations between the DL risk scores evaluated on C alone, P alone and C + P and MFS in the two independent validation cohorts. However, no significant associations were observed for R alone and C + P + R. Specifically, the DL-R risk score was not linked to MFS in Validation-1 and Validation-2, while the DL-CPR risk score showed no association with MFS in Validation-2. In parallel, the histological grade was not significantly associated with MFS in Validation-1, but in Validation-2. Despite this finding, the population remained representative of the typical demographic seen in sarcoma studies, with the majority of patients aged over 50 years, presenting with large tumors exceeding 5 cm, and half of the cases exhibiting high histological grade III, resulting in a considerable 30% risk of metastatic relapse at 5 years. Moreover, excepted the grade, the clinical and pathological covariables linked to lower MFS were consistent with previous findings, namely older age, larger tumor size, deep-seated or in-between sarcomas, and R1 surgical margins[24].

Secondly, we developed supervised survival models inspired by the methodology employed in the construction of the Sarculator nomogram for primary non-metastatic tumors, utilizing Cox regression models[18]. These regressions included patient age, tumor size, and histotype, with simplified categorization due to the small sample sizes in vascular sarcoma and MPNST. Additionally, we included adjuvant chemotherapy and radiotherapy as potential impactors on patient outcomes[14]. The coefficients of the multivariable DL models obtained in the Training cohort indicated that all DL risk scores were independent predictors of MFS, whereas the histological grade was not associated with MFS.

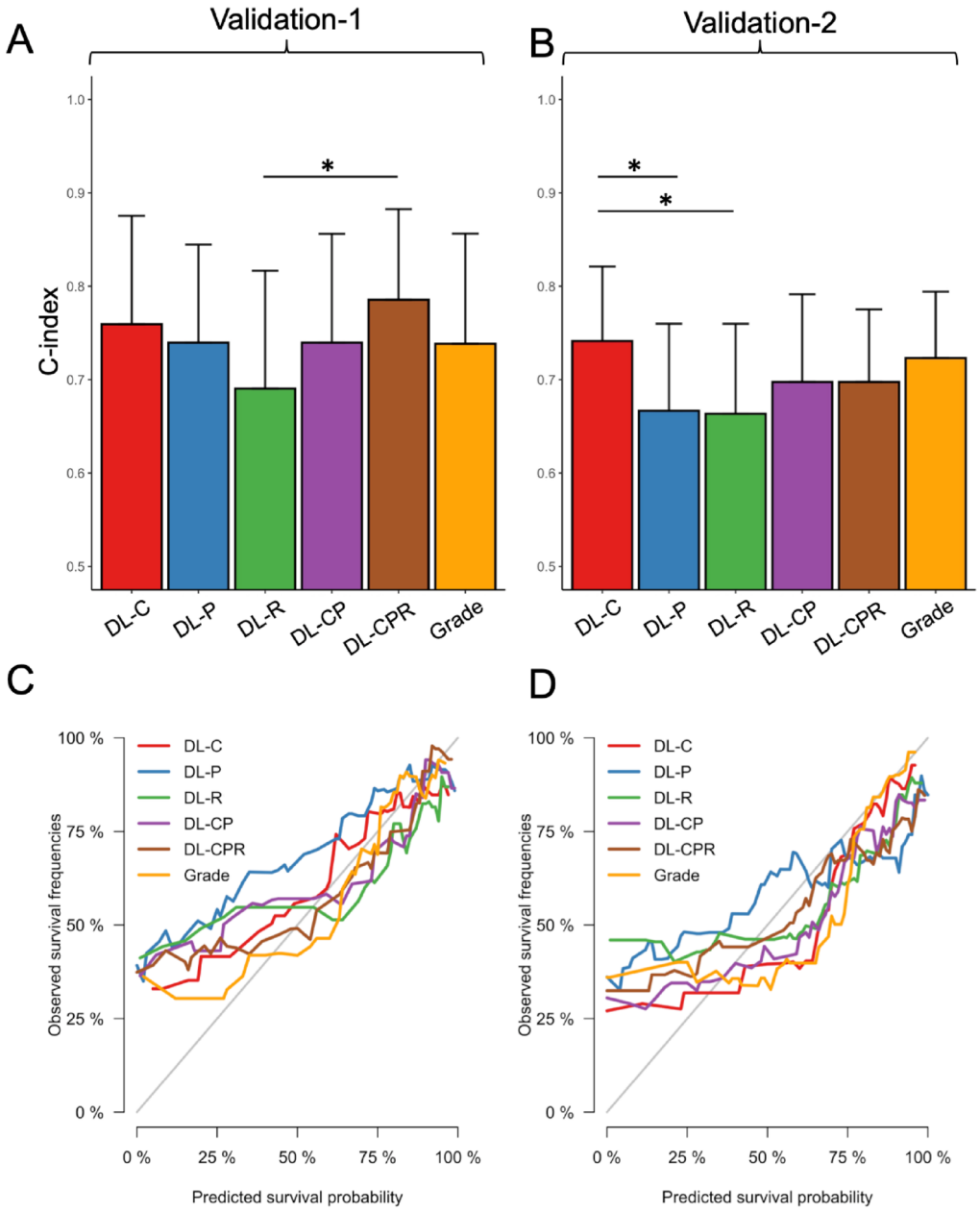

Fig. 3. Results of the multivariable modeling in the validation cohorts. Barchart of the concordance index (c-index) with 95% confidence intervals (95%CIs) in Validation-1 (**A**) and Validation-2 (**B**). *: $P < .05$. Calibration plot of the DL and grade-based models in Validation-1 (**C**) and Validation-2 (**D**).

Thirdly, according to the c-index, the DL-CPR model exhibited the highest prognostic performance in Validation-1 (c-index = 0.786), followed by DL-C. In Validation-2, the highest prognostic performance was achieved with DL-C (c-index = 0.741), followed by DL-CP and DL-CPR. Notably, the DL model based on DL-C risk score provided the best overall performance in both Validation-1 and Validation-2, outperforming the models based on FNCLCC grade, although the comparisons did not reach statistical significance. It is worth noting that the DL models exhibited a decrease in performance from Training to Validation-1 and Validation-2, particularly the DL-CPR, DL-CP, and DL-P models, with a c-index decrease > 0.1, indicating potential overfitting in the modeling process. This finding was not unexpected given the complexity of the DL models and the numerous hyperparameters involved in training. Importantly, while additional covariables might have bolstered the c-indices, we chose to limit their inclusion in the modeling and to assess our DL risk scores and subsequent models against a grade-based model akin to the Sarculator. Hence, the c-indices of the Sarculator nomogram in external validation ranged between 0.65 and 0.75, i.e., closely mirroring the performances of our benchmark grade-based model and slightly lower than the DL-C model in Validation-1 and Validation-2[18,25].

Overall, these findings underscore the potential utility of the DL-C and DL-CP models as a prognostic tool for identifying patients at heightened risk of metastatic relapse, thus guiding the allocation of more aggressive

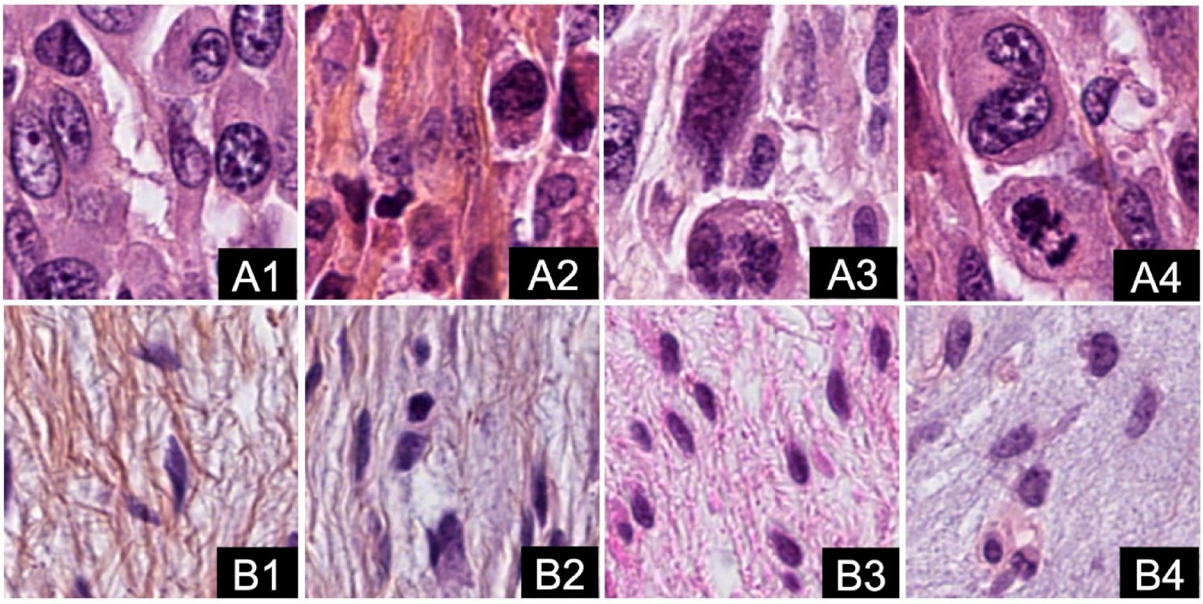

Fig. 4. Typical examples of histological tiles from the high risk (**A**) and low risk (**B**) group according to the deep learning model trained on the tumor centrum and periphery. The most predictive tiles for the high risk group showed high cellular density (A1) or fibrous stroma (A2), pleomorphic cells (A3 and A4), nuclear atypia (A1 to A4) and mitoses (A4) while tiles with low risk showed low cellular density with fibromyxoid (B1) or myxoid (B2) stroma, spindle cells with no nuclear atypia and no mitoses (B1 to B4).

local and systemic therapies, as well as intensified monitoring. Moreover, DL models offer robustness by consistently providing the same prediction when presented with identical input images. In contrast, the inter-observer reproducibility of the FNCLCC grade among senior pathologists from both centers in the Validation-2 cohort yielded a weighted Kappa of 0.480 (95%CI 0.333–0.644, $P < 0.0001$—data not shown), indicating only fair agreement, a finding consistent with previous studies[26]. Hence, a main advantage of DL models lies in their reproducibility and elimination of inter-observer variability associated with FNCLCC grading. Moreover, the DL approach offers practical utility in settings with limited access to expert sarcoma pathologists, such as low-resource or non-specialized centers. By enabling automated and consistent risk assessment from routine histology slides, the model could help reduce disparities in prognostic evaluation. These strengths position the DL models as complementary or alternative tool to existing nomograms. Hence, future studies prospectively could compare the prognostic value of the Sarculator nomogram with FNCLCC grade, the Sarculator coupled with a DL score, and an end-to-end DL model, to better define their respective clinical utilities.

The advent of digital pathology, coupled with the widespread scanning of tissue slices in extensive sarcoma databases, coincides with a scarcity of sarcoma pathology experts and the emergence of telemedicine. In this context, such tools could assist in pre-labeling the HES slice from each newly-diagnosed patient, streamlining the process for pathologists during secondary reviews and confirmatory analyses.

Previous studies have consistently supported these findings, albeit in different cancer types or with smaller sample sizes. For instance, Foersch et al. developed a DL model utilizing CNN to predict disease-specific survival in a cohort of 85 leiomyosarcoma patients, demonstrating high diagnostic accuracy and superior predictive performance compared to histological grading[6]. Similarly, other DL models have been tailored to specific histological subtypes such as synovial sarcoma[27] and rhabdomyosarcoma[28]. Milewski et al. recently investigated DL models in a cohort of 321 rhabdomyosarcoma patients, revealing strong discrimination in overall survival and event-free survival, surpassing existing molecular and clinical models, although quantitative comparison metrics specific to survival analysis were not utilized by the authors[13].

Future researches should include validating the DL-C and DL-CP models in independent prospective cohorts to assess its utility in clinical decision-making processes compared to unaided pathology reviews. Integration of digital immunostaining alongside HES slides could potentially enhance characterization of the tumor microenvironment. Given the focus of our study on peripheral STS, further investigations are warranted for visceral sarcomas, leveraging pre-trained DL pipelines through transfer learning. Moreover, dissecting the characteristics of the most crucial tiles in accurately predicting outcomes could provide valuable insights for enhancing future prognostic models. Herein, our retrospective review of the histological characteristics of representative tiles from the DL-CP high risk and low risk groups yielded findings consistent with established histological markers of STS aggressiveness, namely: increased cellularity, higher mitotic activity, and distinctive stromal patterns and cell types. Furthermore, it is also important to note that our study was not designed to identify the most accurate DL model for predicting patient outcomes from digital slides, nor to systematically benchmark various optimized CNN architectures. Rather, our primary objectives were to demonstrate the feasibility of generating a prognostic deep histological grade, and to determine which combinations of tiles

from the tumor core, periphery, and peritumoral region would yield the highest prognostic performance for this deep grade. Accordingly, if any benchmarking was performed, it pertained to the comparison between the conventional FNCLCC grade, DL-C, DL-P, DL-R, DL-CP and DL-CPR. However, current expansions of our work include alternative encoders, notably CONtrastive learning from Captions for Histopathology, UNI2 and CTransPath. Lastly, an important future direction would be the development of DL models tailored to individual histological subtypes of STS, rather than applying a single model across all subtypes. However, this approach was not feasible in our study due to the limited sample sizes available for each subtype across the three cohorts. Even for the most prevalent histological type—undifferentiated sarcoma—only 42 patients were included in the Training cohort, raising concerns about the reliability and generalizability of subtype-specific models. Moreover, one of our objectives was to critically assess the performance of the FNCLCC grade, which is universally applied across all histological subtypes in clinical practice. In this context, developing DL models on a similarly heterogeneous population was a deliberate and clinically consistent choice. Nonetheless, future studies with larger, subtype-enriched datasets may enable the development of more refined, histotype-specific models.

Our study has limitations. First, it was a retrospective study with a limited study population for a deep learning framework, though it was the largest population regarding the use digital pathology and AI for predicting MFS in STS patients. Hence, some histotypes were under-represented and gathered in the 'Other' group. Second, while the study reported associations between DL risk scores and clinical and pathological variables, the underlying mechanisms driving these associations may not be fully elucidated. Enhancing model interpretability and transparency could improve the clinical adoption and trustworthiness of DL-based risk prediction models. Third, the DL model performances could have been enhanced by including other 'omics' data (including gene-expression or radiomics), as it has already been shown than the Sarculator nomogram and gene expression signature (such as CINSARC) are complementary and potentiate each other[29,30]. Fourth, the selection of digital slides by pathologists for training and validating the DL models could introduce a sampling bias, potentially impacting model generalizability. Future studies should assess the reproducibility and robustness of DL models with respect to inter-observer variability in slide selection, to determine how different slide sampling strategies influence model performance. Fifth, herein, patients who died without experiencing metastatic relapse were censored at the time of death, However, we acknowledge that this approach may introduce a competing risk, as death precludes the observation of relapse. Future studies could consider competing risk methods to further refine prognostic assessment. Sixth, This study was exploratory in nature, and no multiple comparison corrections were applied to the evaluation of DL scores. However, all findings were independently validated in two independent validation cohorts, limiting the risk of false-positive results. Seventh, several factors likely contributed to the lower and variable prognostic performance of the DL models in the validation cohorts, particularly the discrepancy in c-indices between Validation-1 and Validation-2. These include differences in patient characteristics, histological subtype distribution, and technical variations in slide preparation, staining, or scanning protocols across centers and time periods. Importantly, the relatively small size of the validation cohorts likely impacted statistical power and model stability. As this is a proof-of-concept study, further validation on larger, multicenter datasets will be essential before clinical implementation. Lastly, another limitation of our study is the handling of death as a censoring event in the definition of MFS, whereas death may also be considered a competing risk for metastatic relapse. We adopted this cause-specific approach to ensure comparability with prior sarcoma studies and to maintain consistency with Cox regression analyses[3,4]. Nevertheless, competing risk methods were additionally applied for descriptive purposes, which confirmed the robustness of our findings.

In conclusion, our study underscores the potential of the DL-C model as a robust prognostic tool for identifying STS patients at heightened risk of metastatic relapse, aiding pathologists, guiding treatment allocation and monitoring strategies. Adding the surrounding tissues from HES digitalized slide did not improve the model performance. Despite limitations such as retrospectivity, limited representation of certain histotypes and lack of other 'omics' data, our findings contribute to the growing body of evidence supporting the utility of digital pathology and deep learning in oncology and sarcoma.

Data availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the clinical and confidential nature of the material but can be made available from the corresponding author on reasonable request.

Received: 13 December 2024; Accepted: 17 September 2025
Published online: 04 November 2025

References

1. Ducimetière, F. et al. Incidence rate, epidemiology of sarcoma and molecular biology. Preliminary results from EMS study in the Rhône-Alpes region. *Bull. Cancer* **97**, 629–641 (2010).
2. Trojani, M. et al. Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. *Int. J. Cancer* **33**, 37–42 (1984).
3. Coindre, J. M. 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. *J. Clin. Oncol.* **14**, 869–877 (1996).
4. 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* **91**, 1914–1926 (2001).
5. van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: The path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
6. Foersch, S. et al. Deep learning for diagnosis and survival prediction in soft tissue sarcoma. *Ann. Oncol.* **32**, 1178–1187 (2021).
7. Verghese, G. et al. Computational pathology in cancer diagnosis, prognosis, and prediction—present day and prospects. *J. Pathol.* **260**, 551–563 (2023).

8. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
9. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
10. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
11. Arthur, A. et al. A CT-based radiomics classification model for the prediction of histological type and tumour grade in retroperitoneal sarcoma (RADSARC-R): A retrospective multicohort analysis. *Lancet Oncol.* **24**, 1277–1286 (2023).
12. Fu, Y. et al. Deep learning predicts patients outcome and mutations from digitized histology slides in gastrointestinal stromal tumor. *NPJ. Precis. Oncol.* **7**, 71 (2023).
13. Milewski, D. et al. Predicting molecular subtype and survival of rhabdomyosarcoma patients using deep learning of H&E images: A report from the children's oncology group. *Clin. Cancer Res.* **29**, 364–378 (2023).
14. Gronchi, A. et al. Soft tissue and visceral sarcomas: ESMO-EURACAN-GENTURIS clinical practice guidelines for diagnosis, treatment and follow-up☆. *Ann. Oncol.* **32**, 1348–1365 (2021).
15. Blay, J.-Y. et al. Surgery in reference centers improves survival of sarcoma patients: A nationwide study. *Ann. Oncol.* **30**, 1143–1153 (2019).
16. Fisher, R., Pusztai, L. & Swanton, C. Cancer heterogeneity: Implications for targeted therapeutics. *Br. J. Cancer* **108**, 479–485 (2013).
17. Wu, C. et al. Identification of tumor antigens and immune subtypes for the development of mRNA vaccines and individualized immunotherapy in soft tissue sarcoma. *Cancers* **14**, 448 (2022).
18. 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**, 671–680 (2016).
19. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90.
20. Le, V.-L. *et al.* A deep attention-multiple instance learning framework to predict survival of soft-tissue sarcoma from whole slide images. In *Cancer Prevention Through Early Detection: Second International Workshop, CaPTion 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 12, 2023, Proceedings* 3–16 (Springer-Verlag, Berlin, Heidelberg, 2023). https://doi.org/10.1007/978-3-031-45350-2_1.
21. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based deep multiple instance learning. Preprint at https://doi.org/10.48550/arXiv.1802.04712 (2018).
22. Mogensen, U. B., Ishwaran, H. & Gerds, T. A. Evaluating random forests for survival analysis using prediction error curves. *J. Stat. Softw.* **50**, 1–23 (2012).
23. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
24. Casali, P. G. et al. Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **29**, 51–67 (2018).
25. Callegaro, D., Miceli, R., Mariani, L., Raut, C. P. & Gronchi, A. Soft tissue sarcoma nomograms and their incorporation into practice. *Cancer* **123**, 2802–2820 (2017).
26. Italiano, A. et al. Effect of adjuvant chemotherapy on survival in FNCLCC grade 3 soft tissue sarcomas: a multivariate analysis of the French Sarcoma Group Database. *Ann. Oncol.* **21**, 2436–2441 (2010).
27. Han, I., Kim, J. H., Park, H., Kim, H.-S. & Seo, S. W. Deep learning approach for survival prediction for patients with synovial sarcoma. *Tumour Biol.* **40**, 1010428318799264 (2018).
28. Zhang, X. et al. Deep Learning of Rhabdomyosarcoma Pathology Images for Classification and Survival Outcome Prediction. *Am J Pathol* **192**, 917–925 (2022).
29. Chibon, F. et al. Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat. Med.* **16**, 781–787 (2010).
30. Crombé, A. et al. Gene expression profiling improves prognostication by nomogram in patients with soft-tissue sarcomas. *Cancer Commun. (Lond.)* **42**, 563–566 (2022).

Acknowledgements

This work was supported by a grant from the Fondation Bergonié and the charity Au fil d'Oriane of Lycée Saint Elme (Arcachon, France), collaboration with GSF-GETO and interSARC.

Author contributions

Conceptualization: AM, JMC, LVL, OS, AC; Data Curation: AM, JMC, VV, MS, NM, AI, MT, ALC, SB, CH, CN, FLL.; Formal Analysis: AM, JMC, VV, MS, NM, AI, MT, ALC, SB, CH, CN, FLL; Funding Acquisition: JMC, OS; Investigation: all authors; Methodology: AM, JMC, LVL, OS, AC; Project Administration: JMC, OS; Resources: JMC, OS; Software: LVL, OS, AC; Supervision: JMC, OS, AC; Validation : all authors; Visualization: AC, LVL, JMC; Writing—Original Draft Preparation: AC, AM, LVL, JMC; Writing—Review & Editing: all authors.

Declarations

Conflicts of interest

The authors declare no potential conflicts of interest related to this work.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-20804-1.

Correspondence and requests for materials should be addressed to A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025