ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ANNALS OF ONCOLOGY driving innovation in oncology

ORIGINAL ARTICLE

Deep learning for diagnosis and survival prediction in soft tissue sarcoma

S. Foersch[1*], M. Eckstein[2], D.-C. Wagner[1], F. Gach[1], A.-C. Woerl[1,3], J. Geiger[1,3], C. Glasner[1,3], S. Schelbert[1], S. Schulz[1], S. Porubsky[1], A. Kreft[1], A. Hartmann[2], A. Agaimy[2] & W. Roth[1]

[1]Institute of Pathology, University Medical Center Mainz, Mainz; [2]Institute of Pathology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen; [3]Institute of Computer Science, Johannes Gutenberg University Mainz, Mainz, Germany

Available online 15 June 2021

Background: Clinical management of soft tissue sarcoma (STS) is particularly challenging. Here, we used digital pathology and deep learning (DL) for diagnosis and prognosis prediction of STS.

Patients and methods: Our retrospective, multicenter study included a total of 506 histopathological slides from 291 patients with STS. The Cancer Genome Atlas cohort (240 patients) served as training and validation set. A second, multicenter cohort (51 patients) served as an additional test set. The use of the DL model (DLM) as a clinical decision support system was evaluated by nine pathologists with different levels of expertise. For prognosis prediction, 139 slides from 85 patients with leiomyosarcoma (LMS) were used. Area under the receiver operating characteristic (AUROC) and accuracy served as main outcome measures.

Results: The DLM achieved a mean AUROC of 0.97 (±0.01) and an accuracy of 79.9% (±6.1%) in diagnosing the five most common STS subtypes. The DLM significantly improved the accuracy of the pathologists from 46.3% (±15.5%) to 87.1% (±11.1%). Furthermore, they were significantly faster and more certain in their diagnosis. In LMS, the mean AUROC in predicting the disease-specific survival status was 0.91 (±0.1) and the accuracy was 88.9% (±9.9%). Cox regression showed the DLM's prediction to be a significant independent prognostic factor ($P = 0.008$, hazard ratio 5.5, 95% confidence interval 1.56-19.7) in these patients, outperforming other risk factors.

Conclusions: DL can be used to accurately diagnose frequent subtypes of STS from conventional histopathological slides. It might be used for prognosis prediction in LMS, the most prevalent STS subtype in our cohort. It can also help pathologists to make faster and more accurate diagnoses. This could substantially improve the clinical management of STS patients.

Key words: artificial intelligence, deep learning, soft tissue sarcoma, prognosis prediction, clinical decision support system, digital pathology

INTRODUCTION

Soft tissue sarcomas (STSs) are malignant tumors of mesenchymal origin, representing a highly heterogeneous group with numerous different subtypes. While STSs are rare compared to other tumor entities, their incidence has steadily increased over the last decade and they are associated with significant morbidity and mortality.[1,2] This holds true for young adults in particular, where certain sarcoma subtypes are among the leading causes of death.[3] Precise classification is critically important as these tumors vary substantially in their biological behavior, their clinical prognosis, and their response to therapy. Pathology plays a pivotal role in the clinical management of STS, yet high level of diagnostic expertise is often limited to a few reference centers. This might contribute to substantially delayed diagnosis which has a significant negative impact on the lives of patients with STS.[4]

One potential solution to this could be the adoption of artificial intelligence (AI) in the pathological management of STS. While other disciplines have already undergone almost complete digitalization, pathologists still heavily rely on analog technologies such as benchtop microscopes, glass slides, or written reports. This is despite the fact that recently AI techniques have been introduced into the field and have performed remarkably well in a number of diagnostic, prognostic, and predictive tasks.[5-7] For example, we were recently able to predict the molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides alone using deep learning (DL).[8] Apart from preliminary reports, these techniques have rarely been used in the extraordinarily difficult context of diagnosis or

*Correspondence to: Dr Sebastian Foersch, University Medical Center Mainz, Institute of Pathology, Langenbeckstr. 1, 55131 Mainz, Germany. Tel: +0049-6131-17-5144; Fax: +0049-6131-17-477305

E-mail: sebastian.foersch@unimedizin-mainz.de (S. Foersch).

prognosis prediction of STS. In the current study, we utilized state-of-the-art DL approaches to correctly diagnose five of the most common subtypes of STS. Furthermore, we compared our best model's performance with the performance of pathology experts with varying experience in general pathology. Additionally, we investigated the most common subtype of leiomyosarcomas (LMSs) more thoroughly and were able to differentiate between patients with a good or bad prognosis based on their tumor's histomorphology alone using AI. Lastly, we explored different visualization techniques to identify recurring microscopic features associated with different prognosis.

PATIENTS AND METHODS

Patient cohorts

Five major entities as defined by the recent World Health Organization Classification of Tumors were included in the study [dedifferentiated liposarcoma (DDLPS), LMS, myxofibrosarcoma (MFS), synovial sarcoma not otherwise specified (SS), and undifferentiated pleomorphic sarcoma (UPS)]. Two cohorts of patients with STS were used. The first cohort served as the basis for training of the neural network and validation to determine performance metrics. It consisted of 240 patients of The Cancer Genome Atlas Sarcoma (TCGA SARC) cohort for which we downloaded the diagnostic hematoxylin–eosin (H&E)-stained whole slide through the Genomic Data Commons portal (https://portal.gdc.cancer.gov/). All initial pathology reports as well as clinicopathological and survival data [disease-specific survival (DSS)] were gathered from Cancer Genome Atlas Research Network[9] and www.cbioportal.org. In cases with conflicting information, they were thoroughly re-evaluated and discussed again using all available information to reach a final diagnosis. This was then used as a ground truth. A comprehensive quality assessment excluded slides with large folds, no tumor tissue, and/or scans that were out of focus. A total of 456 slides were used for tile generation. A second, multicenter cohort of 51 patients was used as an additional external test set. This consisted of patients from the Comprehensive Cancer Center Erlangen Metropole Region Nuremberg Sarcoma cohort (CCC-EMN) diagnosed between 2007 and 2016 at the University Clinic Hospital Erlangen or in the capacity as Sarcoma Reference Center (41 patients) and the Mainz cohort diagnosed between 2017 and 2020 at the University Medical Center Mainz (10 patients). Generation of the cohort was approved by the ethical review board of the Friedrich-Alexander-University Erlangen-Nürnberg (ref. no. 3755) and the ethical committee of the medical association of the State of Rhineland-Palatinate [ref. no. 837.031.15(9799)] and was in accordance with the Declaration of Helsinki.

Scanning and preprocessing

TCGA slides were digitalized at various institutions participating in the TCGA consortium. Slides from the second cohort were scanned using a Hamamatsu Nanozoomer Series scanner (Hamamatsu Photonics, Hamamatsu, Japan) at 40-fold magnification. Slides were thoroughly annotated by a pathology expert. Annotation describes the process in which a polygonal region of interest was drawn around the tumor area. Smaller image tiles (1024 px^2) were then generated from these annotations. A tile size of 1024 px^2 was chosen to compromise between the negative effect of tiles being too small and nuclear features still being detectable to the system. To minimize the potential bias introduced by scanning and staining at different institutions and achieve more robust results, all images were normalized to an external reference image of a different dataset using structure-preserving color normalization as proposed by Vahadane et al.[10] and Anand et al.[11] and as established in our laboratory.[8] Different types of augmentation were randomly applied, such as flipping, mirroring, contrast/saturation/brightness changes, and progressive sprinkles (Supplementary Figure S1, available at https://doi.org/10.1016/j.annonc.2021.06.007).

Deep learning algorithm

For the classification experiments, a standard Densely Connected Convolutional Network (DenseNet121) was used.[12] The fast.ai function lr_find() was used to identify an adequate learning rate which was 0.0001 in our experiments. Cross-entropy loss served as the loss function, Adam was used as an optimizer, and a momentum of 0.95 was chosen. The default batch size was 32. After classification of the tiles, the slide of a given patient was assigned to the STS subtype, which was predicted for the weighted majority of all tiles of that respective image. This means that before deciding on the majority, the prediction certainty for every tile was used as an additional factor together with the total number of tiles for each subtype. Classification markup was carried out using our previously reported sliding window approach.[8] In short, a squared grid matching the tile size is placed on the whole slide image and each resulting sub-image is classified by the network in inference mode. This part of the grid is then color-coded either according to which class was called or according to the probability of the majority class for the whole slide. Class activation maps (CAMs) were established as proposed by Zhou et al.[13]

Statistical analysis

Training and validation on the TCGA cohort were carried out patient-wise using stratified random permutations cross-validation ('shuffle & split') with a 90% (training) to 10% (validation) ratio. The results were confirmed in up to three similar independent experiments, each reaching comparable conclusions. Performance metrics included recall (sensitivity), true-negative rate (specificity), precision, F1 score, and area under the receiver operating characteristic (AUROC). The mean AUROC either of multiple classes or as a summary of cross-validation for each individual class was calculated using micro- and macro-averaging.[14] For each analysis, the values' distribution

and variances were tested. Unpaired *t*-test was used when two individual groups with normal distribution and equal variances were analyzed. Welch's correction was added when variances were significantly different. Mann–Whitney *U* test was used when two individual groups with non-normal distribution were analyzed. Log-rank test was used to compare the survival of two or more groups. Univariable and multivariable Cox regression was used for prognosis analyses. Pearson's *r* was used for correlation evaluation. If not indicated otherwise, ± standard deviation is given. When *P* values were <0.05, the differences between the groups compared were considered statistically significant ($P \geq 0.05$: not significant, $*P = 0.01$-0.05, $**P = 0.001$-0.01, $***P = 0.0001$-0.001, $****P < 0.0001$). For annotations and image preprocessing, QuPath open source software[15] was used. All DL experiments were done in Python using PyTorch/fast.ai or TensorFlow/Keras. Our algorithms were developed utilizing open access material and tutorials, such as 'Practical Deep Learning for Coders' by Jeremy Howard, PyImageSearch by Adrian Rosebrock, and others.

RESULTS

Clinical management of STS cases often involves multiple subspecialties (internal medicine, surgery, radiology, pathology, etc.) and a typical case (liposarcoma) is shown in Figure 1A. Among the five most frequent STS subtypes are DDLPS, LMS, MFS, SS, and UPS. Morphology of specific cases is displayed in Figure 1B, showing the high histomorphologic heterogeneity of STS. Characteristics of the TCGA cohort are shown in Figure 1C. Kaplan–Meier analysis revealed that DDLPS had the worst prognosis (median survival of 66.9 months), while SS showed the most favorable DSS (median survival not reached). Local tumor extent as indicated by T stage also demonstrated to be of prognostic importance (Figure 1D).

A total of 365 011 tiles were generated from annotations of 456 slides from 240 patients with a mean tile count of 1520.88 (±1704.11) per patient. For the diagnostic experiments, up to 75 randomly selected tiles per patient were used for training resulting in 17 919 tiles for these experiments (to keep computing times within a reasonable timeframe). Validation was done on all tiles of the respective patient. Cross tables after a total of 50 epochs of training are displayed in Figure 2A. These represent a summary of six times stratified random permutations cross-validation. Mean classification accuracy was 79.9% (±6.1%) with the best model achieving up to 87.5% (21/24 cases). The AUROC for all classes combined was 0.95 (±0.02) using micro-averaging and 0.97 (±0.01) using macro-averaging (Figure 2B). The AUROC was 0.91 (±0.03) for DDLPS, 0.97 (±0.02) for LMS, 0.98 (±0.02) for MFS, 1.00 (±0.00) for SS, and 0.92 (±0.03) for UPS (Figure 2C). Additional

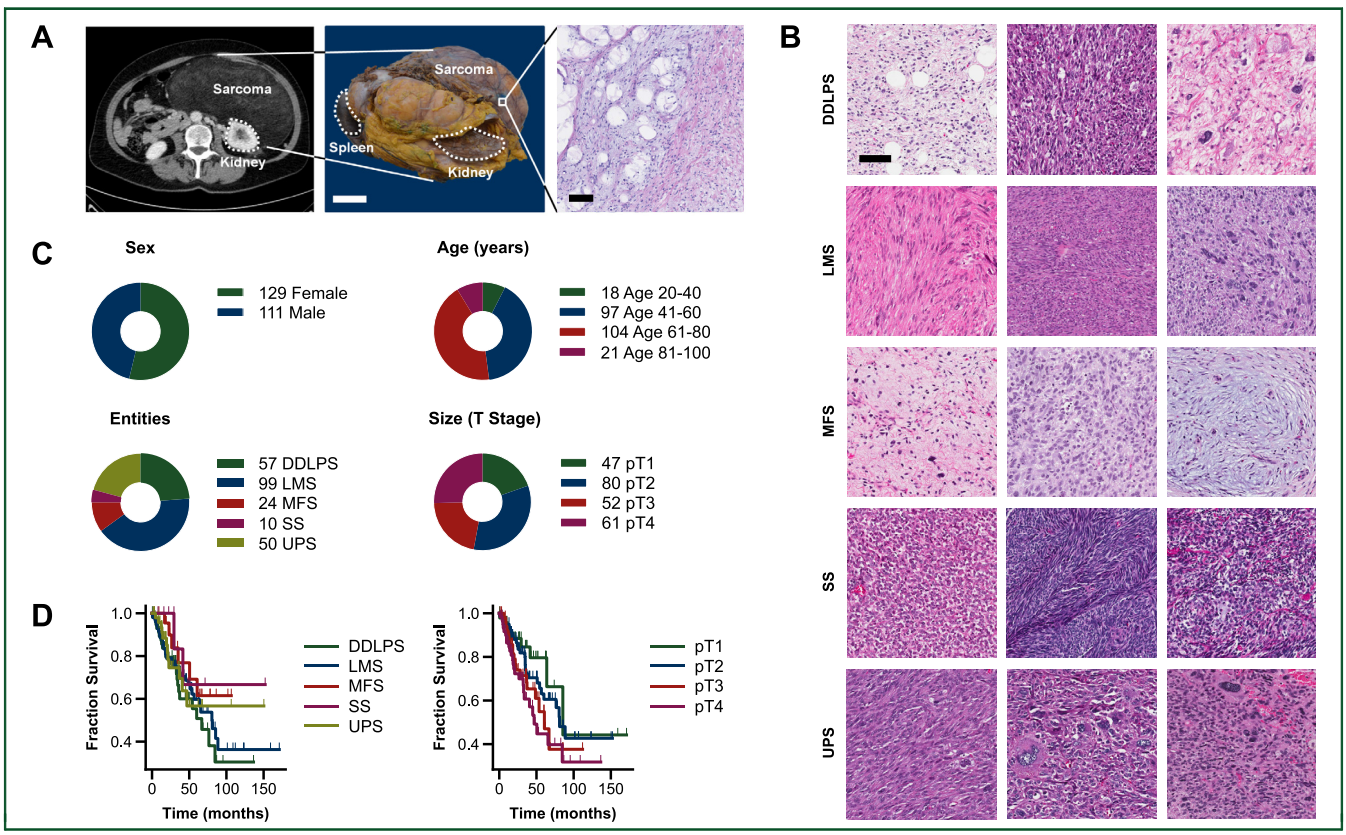

Figure 1. Clinical examples and patient characteristics.
(A) Typical clinical appearance of a soft tissue sarcoma (liposarcoma) upon radiological, macroscopical, and histopathological examination. Left scale bar: 5 cm, right scale bar: 100 μm. (B) Representative histopathological images of the five different subtypes evaluated [dedifferentiated liposarcoma (DDLPS), leiomyosarcoma (LMS), myxofibrosarcoma (MFS), synovial sarcoma (SS), undifferentiated pleomorphic sarcoma (UPS)]. Scale bar: 100 μm. (C) Clinicopathologic features of the cohort. (D) Kaplan–Meier curves stratified by histopathological subtype (left panel) and tumor size (as shown by T stage) (right panel).

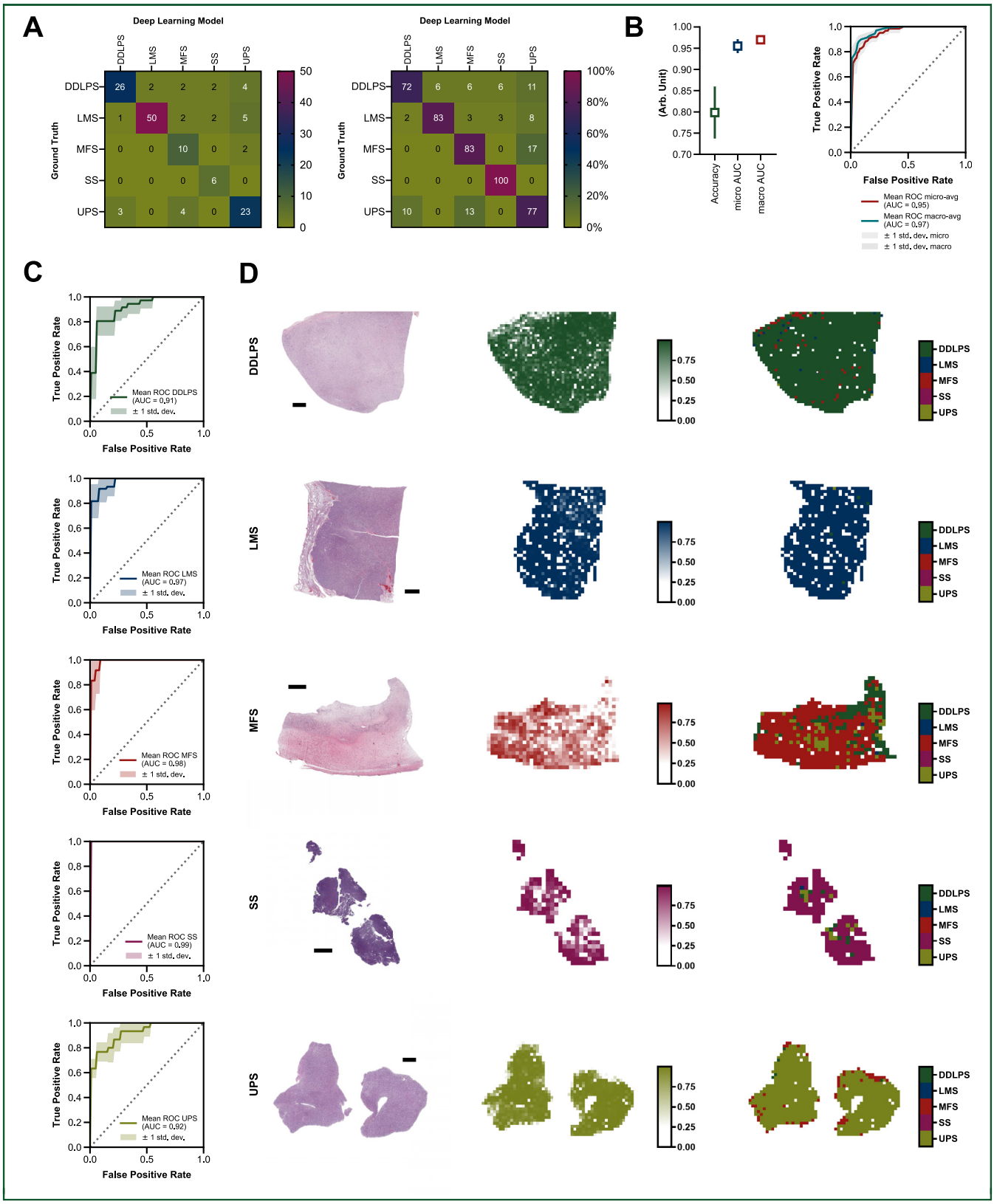

Figure 2. Classification of soft tissue sarcoma (STS) using deep learning.
(A) Cross tables as a summary of all patients evaluated during stratified random permutations cross-validation. Absolute (left) and relative (right) values are shown. (B) Accuracy and area under the receiver operating characteristic using micro- and macro-averaging. Respective receiver operating characteristic (ROC) curves with micro- and macro-average are also shown (right). (C) ROC curves for each STS subtype. (D) Sliding window visualization of a representative specimen for each STS subtype. The left panel shows the hematoxylin—eosin input image. The middle panel shows the prediction probability ('certainty') for the class that was assigned to the whole image on a tile-by-tile basis. The right panel shows the classification on a tile-by-tile basis, where five different colors represent the subtype with the highest classification probability for each tile. A tumor is classified as the subtype weighted majority of all tiles of that particular case. Scale bar: 2 mm. AUC, area under the curve; DDLPS, dedifferentiated liposarcoma; LMS, leiomyosarcoma; MFS, myxofibrosarcoma; std. dev., standard deviation; SS, synovial sarcoma; UPS, undifferentiated pleomorphic sarcoma.

https://doi.org/10.1016/j.annonc.2021.06.007

performance measures and tile-based receiver operating characteristic (ROC) curves can be found in Supplementary Figure S2, available at https://doi.org/10.1016/j.annonc.2021.06.007. A sliding window approach was used to visualize classification results of all subgroups as well as the classification probability of the most prevalent subgroup on representative slides (Figure 2D). Next, we wanted to confirm our results on an additional, external multicenter test set and used a total of 51 cases of the CCC-EMN/Mainz cohort. Frequency distribution of the different subtypes was similar to the TCGA cohort (Supplementary Figure S3A, available at https://doi.org/10.1016/j.annonc.2021.06.007) and we were able to achieve an accuracy of 78.4% on these unseen cases. The AUROC was 0.93 using micro-averaging and 0.94 using macro-averaging. Cross tables and additional ROC curves of this experiment can be found in Supplementary Figure S3B-D, available at https://doi.org/10.1016/j.annonc.2021.06.007.

As STSs are among the most difficult entities to diagnose, we wanted to evaluate the ability of pathology experts with different levels of experience to correctly identify representative slides of the five most common subtypes. To this end, an online tool was set up and participants were shown the original, unprocessed H&E slides either with or without the diagnosis of our deep learning model (DLM). Experience in general pathology ranged from 2.5 months to 25 years. The experiment started with a short self-assessment, followed by the diagnostic part. We used 24 cases of the best run during stratified random permutations cross-validation. Cases were randomly split and the first 12 slides were presented without any additional information, while the last 12 slides were presented with the prediction of the DLM as text, the model's accuracy, as well as markup images similar to those seen in Figure 3D. The time per slide was recorded and additional information on how the rater would proceed in a clinical setting was queried. While the mean accuracy of all raters was 46.3% (±15.5%) without the support of the DLM, it improved significantly ($P < 0.0001$) to 87.1% (±11.1%). This was on a par with the model's own performance which was 87.5% (Figure 3A and D). Figure 3B shows performance metrics including ROC measurements. A potential integration into the routine pathology workflow as a clinical decision support system using a tablet can be seen in Figure 3C. The mean time from opening the file to logging in the diagnosis decreased significantly ($P = 0.01$) from 89.3 s (±25.4 s) to 43.2 s (±42.9 s) with the support of the DLM. The biggest performance increase could be observed for those raters with low pathology experience. Furthermore, with the assistance participants were significantly ($P = 0.026$) more confident in their diagnoses [confidence from −0.25 (±0.61) to 0.65 (±0.32)] and the hypothetical number of downstream analyses (immunohistochemistry, molecular pathology, reference pathology consultation) decreased almost 20% from 163 to 132 (Figure 3D). Interestingly, while the most experienced pathologists showed the highest accuracy of all raters on the cases without the help of the DLM, their improvement (ΔAccuracy) upon computer assistance was the lowest (Figure 3E). There was a significant inverse correlation between a rater's accuracy improvement and their experience in years. For two out of the three senior-level experts, the processing time increased and the diagnostic certainty suffered when being offered a diagnosis suggestion by the DLM.

Next, we wanted to investigate whether the DL approach could be used to predict other clinically relevant information, rather than merely subtyping of STS entities. To this end, we chose the most frequent subtype, LMS, and trained another neural network on the 2-year DSS status, meaning if a given patient was alive or dead after 2 years. These data could be gathered from 85 patients and up to 300 randomly selected tiles from 139 slides were used for training resulting in 23 663 tiles for these experiments. Validation was also done on up to 300 randomly selected tiles of the respective patient. Cross tables are displayed in Figure 4A. These represent a summary of six times stratified random permutations cross-validation. Mean classification accuracy was 88.9% (±9.9%) with the best model achieving up to 100% (9/9 cases). The AUROC for all classes combined was 0.91 (±0.084) using micro-averaging and 0.91 (±0.098) using macro-averaging (Figure 4B and C). Additional performance measures and tile-based ROC curves can be found in Supplementary Figure S4, available at https://doi.org/10.1016/j.annonc.2021.06.007. A sliding window approach was used to visualize classification results of all subgroups as well as the classification probability of the most prevalent subgroup on representative slides (Figure 4D). Training on the survival status of the other, non-LMS subtypes resulted in a worse prognosis prediction when compared to the LMS subtype with an AUROC of 0.80 (±0.116) using micro-averaging and 0.71 (±0.196) using macro-averaging (Figure 4E). We then carried out Kaplan−Meier analyses after grouping all validation patients according to the prediction of the DLM. Log-rank test showed a significant difference in survival ($P < 0.0001$) (Figure 4F). Upon univariable Cox regression analysis, La Fédération nationale des centres de lutte contre le cancer (FNCLCC) grade [$P = 0.017$, hazard ratio (HR) 2.6, 95% confidence interval (CI) 1.2-5.5], T stage ($P = 0.020$, HR 1.6, 95% CI 1.1-2.5), R status ($P = 0.03$, HR 2.8, 95% CI 1.1-7.3), and the DLM ($P = 0.0002$, HR 5.3, 95% CI 2.2-13) were significant prognostic factors for DSS in the LMS cohort. Interestingly, whether samples were provided from a large institution (>5 cases to the whole cohort) or a small institution (<5 cases to the whole cohort) did not have an effect on the DSS (Supplementary Figure S5A-C, available at https://doi.org/10.1016/j.annonc.2021.06.007) and was not a significant factor upon Cox regression (Supplementary Figure S5D, available at https://doi.org/10.1016/j.annonc.2021.06.007 and Figure 4G). Multivariable Cox regression analysis showed the DLM's prediction to be the only remaining significant independent prognostic factor ($P = 0.008$, HR 5.5, 95% CI 1.6-19.7) in these patients (Figure 4G).

Next, we wanted to identify the histomorphological features that were most relevant to the model and thus might be associated with DSS in LMS. To make sure that classification results are based on similar, recurring

image features, we used t-distributed Stochastic Neighbor Embedding (t-SNE) as a non-linear dimensionality reduction method. When examining the output vectors of the DLM before the final classification layer with this method, clear clusters could be identified. Visualizing 300 random tiles per patient, they clustered well according to their DSS status (Figure 5A). Figure 5B depicts two representative patients and the classification by the DLM. Here we could observe that tiles 'falsely' classified also tended to cluster and somewhat coalesced between the groups. We then used CAMs to highlight the particular image regions that were most responsible for calling a certain class. After systematic evaluation of the 25 tiles with the highest prediction probability per patient, several features could be identified (Figure 5C and D). Most notably, densely packed, small, monomorphic nuclei as well as lymphoid infiltrates were associated with better survival, while atypical, large, pleomorphic nuclei with prominent intercellular matrix were more frequently observed in patients with poor prognosis. Furthermore, intratumoral hemorrhage and a higher number of tumor-associated vessels were also frequently found in patients with shorter survival. Next, we wanted to confirm these findings by additional methods. Therefore, we used classical image analysis where numerical parameters served as surrogates for the descriptive features found by CAMs. These experiments confirmed our initial findings as the mean nuclear area, the mean distance triangle area, and the eosin-positive cell count were all significantly associated with either good or bad prognosis (Figure 5E).

Lastly, semi-supervised learning techniques such as multiple instance learning (MIL) have recently been proposed to achieve better results in histopathological classification tasks than supervised transfer learning. To see whether this

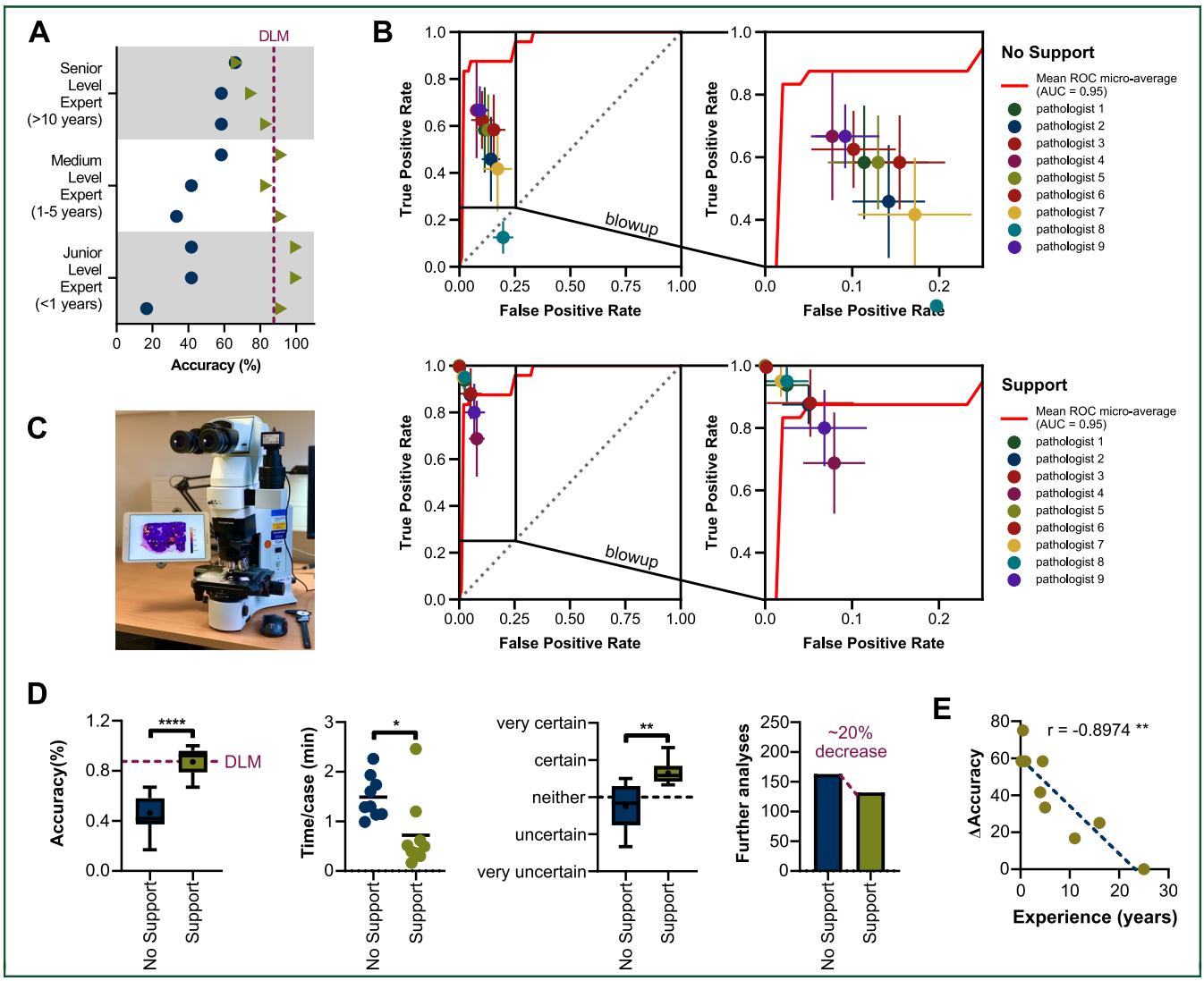

Figure 3. Clinical decision support using deep learning.
(A) Accuracy improvement of nine human raters with different levels of expertise in soft tissue sarcoma pathology. Measurements without (blue circle) and with (green triangle) the support of the deep learning model (DLM) are displayed. Junior-level expert means less than a year of experience in general pathology, medium-level expert means between 1 and 5 years of experience in general pathology, and senior-level expert means >10 years of experience in general pathology. (B) Receiver operating characteristic (ROC) curve of the DLM and sensitivity/specificity measurements of the human raters. Results without (top) and with (bottom) the support are shown. The right panel represents a blowup of the box in the left panel. (C) Potential setup within the clinical pathology routine using a mobile device. (D) Comparison of accuracy, time per slide, diagnostic certainty, and potential downstream analyses without and with the support of the DLM. (E) Correlation between accuracy improvement (ΔAccuracy) and experience in years. AUC, area under the curve. *P = 0.01-0.05, **P = 0.001-0.01, ****P < 0.0001.

https://doi.org/10.1016/j.annonc.2021.06.007

holds true for survival prediction in LMS, we established an MIL pipeline similar to that of Campanella et al.[16] Using the same six times stratified random permutations cross-validation setup as in the experiments described above, MIL did not improve classification performance with a mean accuracy of 87.0% (±12.99%) and a mean AUROC of 0.84 (±0.07) upon micro-averaging and 0.75 (±0.17) upon macro-averaging (Supplementary Figure S6, available at https://doi.org/10.1016/j.annonc.2021.06.007).

DISCUSSION

Despite continuous advances in basic and translational research, STSs still pose a significant clinical challenge.[4] There are major obstacles that are particularly relevant: (i) Due to their relatively low prevalence, there is an apparent lack of expertise among physicians in general and pathologists in particular. Reproducibility of STS diagnosis is poor across pathologists who are not seeing these lesions regularly and there is a high intra- and interobserver variability among 'non-sarcoma' pathologists.[17] (ii) This can partly contribute to a significant delay in diagnosis. As reviewed by Soomers et al., the average diagnostic interval for STS is ~41.2 weeks and can range from 4.3 weeks to 94.6 weeks.[18] Although pathologists are only partly to blame for this postponement, it can have disastrous effects on the patients' outcome and a number of studies reported a longer time interval to be associated with a worse overall survival.[19-21] Furthermore, delay in diagnosis is the most

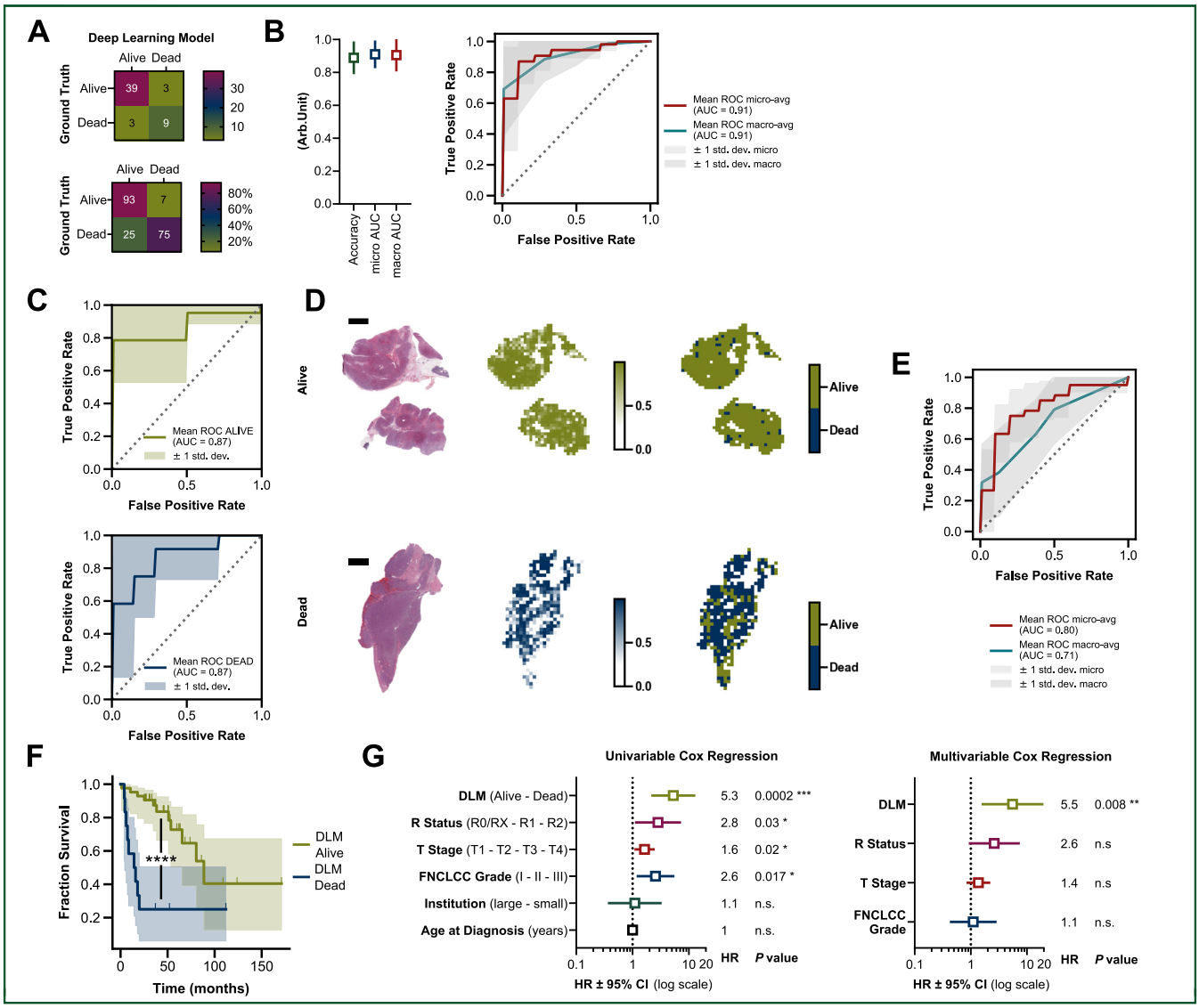

Figure 4. Survival prediction of leiomyosarcoma using deep learning.
(A) Cross tables as a summary of all patients evaluated during stratified random permutations cross-validation. Absolute (top) and relative (bottom) values are shown. (B) Accuracy, area under the receiver operating characteristic using micro- and macro-averaging. Respective receiver operating characteristic (ROC) curves with micro- and macro-average are also shown (right). (C) ROC curves for the prediction of dead and alive patients. (D) Sliding window visualization of a representative specimen for each group. The left panel shows the hematoxylin–eosin input image. The middle panel shows the prediction probability ('certainty') for the class that was assigned to the whole image on a tile-by-tile basis. The bottom panel shows the classification on a tile-by-tile basis, where the different colors represent the decision with the highest classification probability for each tile. A tumor is classified as the weighted majority of all tiles of that particular case. Scale bar: 2 mm. (E) ROC curves for non-leiomyosarcoma (LMS) subtypes. (F) Kaplan–Meier curve of all 54 validation patients grouped by the prediction of the deep learning model (DLM). (G) Univariable and multivariable Cox regression of known risk factors in LMS. AUC, area under the curve; CI, confidence interval; HR, hazard ratio; n.s., not significant; std. dev., standard deviation; n.s., $P \geq 0.05$, *P = 0.01-0.05, **P = 0.001-0.01, ***P = 0.0001-0.001, ****$P < 0.0001$.

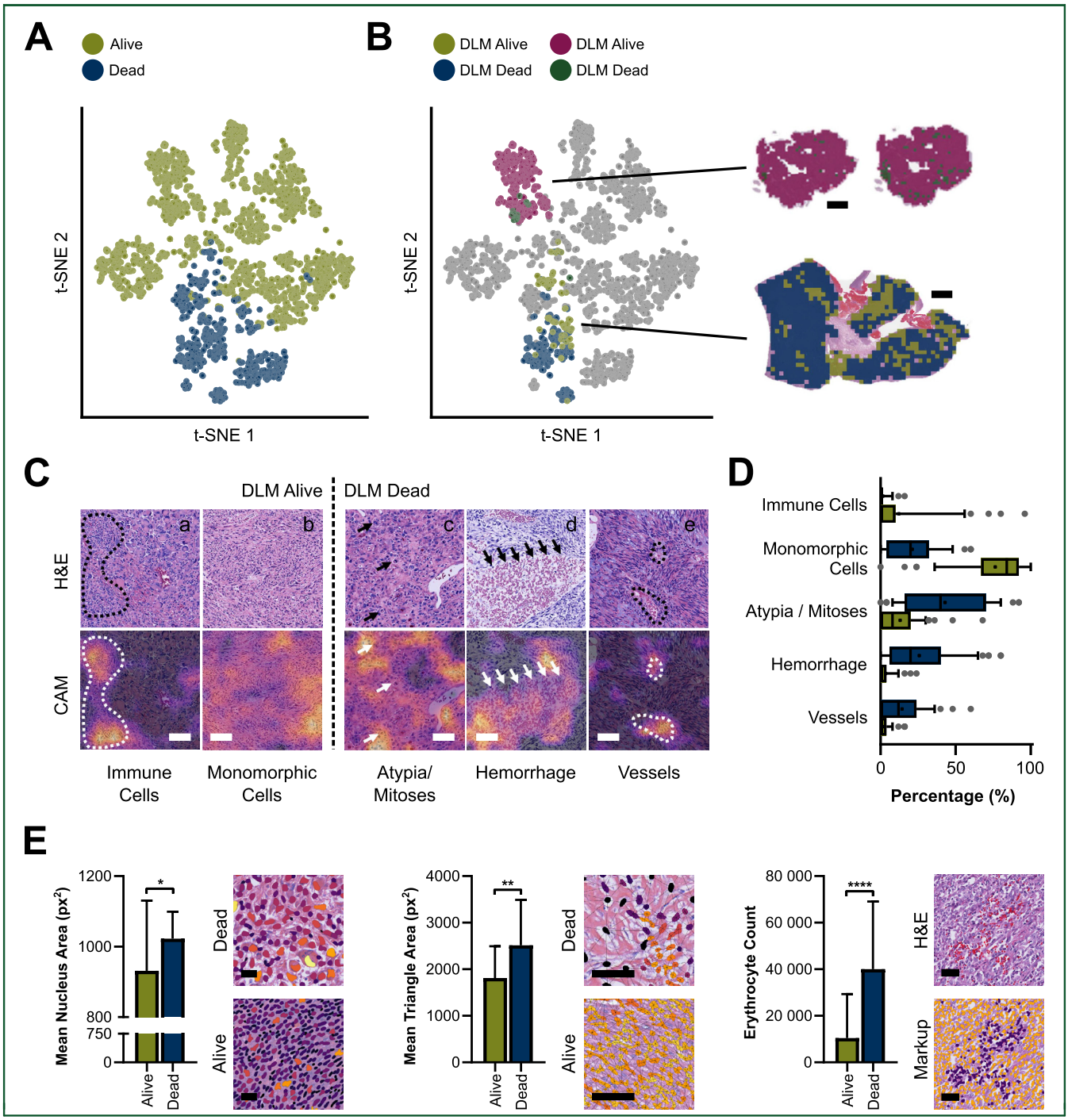

Figure 5. Visualization of histopathological features relevant for the deep learning model (DLM).
(A) t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis of the neural network's output layers of one representative fold during stratified random permutations cross-validation. The top 25 classified tiles per patient were analyzed. Tiles of deceased patients are shown in blue, and tiles of patients who are alive are displayed in green. (B) The same t-SNE plot as in A. Here the DLM predictions for the tiles of two representative patients are visualized. Scale bar: 5 mm. (C) Class activation maps revealing recurrent histopathological features which are most relevant to the DLM: (a) immune cell infiltrates (dotted line, associated with prediction 'Alive'), (b) densely packed, monomorphic cells (associated with prediction 'Alive'), (c) atypical, large, separated nuclei and mitoses (arrows, associated with prediction 'Dead'), (d) intratumoral hemorrhage (arrows, associated with prediction 'Dead'), and (e) tumor vasculature (dotted line, associated with prediction 'Dead'). Scale bar: 90 µm. (D) Quantification of these features in the top 25 classified tiles per patient. Box plots of deceased patients are shown in blue, and box plots of patients who are alive are displayed in green. Here the range between the 10th and the 90th percentile is shown. (E) Confirmation of these features with additional methods of image analysis. (Left) Mean nuclear area is significantly higher in deceased patients. Scale bar: 100 px. (Middle) Mean triangle area upon Delaunay clustering, as a measurement of cell density is significantly smaller in living patients. Scale bar: 250 px. (Right) Erythrocyte count as a surrogate for intratumoral hemorrhage and tumor vasculature is significantly higher in deceased patients. Scale bar: 200 px. H&E, hematoxylin—eosin. *P = 0.01-0.05, **P = 0.001-0.01, ****P < 0.0001.

common cause for complaints and litigation in STS patients.[22] For these and other reasons, it is recommended that STS patients are managed by an experienced multidisciplinary team in a specialized sarcoma center.[23] (iii) Apart from diagnosis, treatment of STS can also be highly challenging—particularly in advanced tumor stages not

https://doi.org/10.1016/j.annonc.2021.06.007

suitable for curative surgical intervention. Here, overall response rates remain low with 12%-24%.[24] Since the histologic subtype of STS is one of the most important independent prognostic and predictive factors, there is a desperate need for histology-specific treatment algorithms.[25]

In the current study, we present a novel approach to address these serious challenges using AI for histopathological diagnosis and survival prediction in STS. The DLM was able to correctly identify the five most common subtypes of STS with an accuracy of up to 87.5% and a macro-averaged AUROC of up to 0.98. When pathology experts with varying experience in sarcoma pathology were assisted in their diagnosis by a recommendation of the DLM, their diagnostic accuracy increased significantly from 46.3% to 87.1%. Interestingly, human raters with little experience benefitted the most from the DL support while some of the senior-level experts showed doubt when provided with computer assistance. While pathologists usually do not diagnose STS from H&E alone, our approach could nonetheless be used to better distribute STS diagnostic expertise among non-reference pathology institutions. It could also contribute to shortening the diagnostic interval as we demonstrated that turnaround time decreased significantly when additional support by the DLM was provided. Additionally, the number of hypothetical downstream analyses could also be reduced by almost 20%, potentially decreasing the diagnostic delay even further while also saving costs. This could substantially improve the clinical management of STS by shortening the interval between the very first consultation and the beginning of the treatment. Furthermore, such a system could be used to increase the level of sarcoma expertise among smaller medical centers, potentially reaching even more patients with STS. When trained on the DSS in patients with LMS, the DLM was even able to accurately predict the 2-year survival status and the model's outcome was a significant independent prognostic factor, outperforming known risk factors such as high FNCLCC grade or incomplete surgical resection. This could contribute to better identify high-risk patients and thus improve treatment and surveillance of LMS. Using various sophisticated visualization techniques, we were able to 'reverse engineer' the histomorphological features associated with better or worse prognosis, broadening our understanding of the underlying tumor biology, while at the same time making our approach more transparent.

This is one of the first studies to investigate the use of AI in the pathological management of STS. There is a rapidly growing body of work, trying to use machine learning algorithms to improve diagnosis and derive other clinically important information from conventional histopathological slides. But while classification accuracy has steadily increased, there is still room for improvement.[7,26,27] This is particularly relevant in histopathological diagnoses that are highly challenging, such as STS. AUROC has been shown to overestimate a model's performance[28,29] and reporting of other metrics is often incomplete. Even for the multiclass classification problem of detecting the most common STS subtypes, we achieved some of the highest performance measurements when compared with some of the most recent studies.[30] Other semi-supervised learning approaches such as MIL could provide further improvements, but these techniques heavily depend on large sample numbers. This can be particularly challenging for rare tumors such as STS. Consequently, MIL did not improve performance in our setup. So, while semi-supervised learning techniques might outperform supervised transfer learning when high sample numbers are available, this does not hold true in a setting with limited available training data.

FUNDING

This work was supported by the Federal Ministry of Education and Research [grant number 16SV8167]; the Stage-I-Program of the University Medical Center Mainz (no grant number); the Mainz Research School of Translational Biomedicine (TransMed) (no grant number); and the Manfred-Stolte-Foundation (no grant number).

DISCLOSURE

The authors have declared no conflicts of interest.

DATA SHARING

Code and data examples will be made available under https://github.com/AGFoersch/DeepLearningSarcoma after acceptance. The trained models, the complete source code as well as the underlying datasets can also be provided upon reasonable request to the corresponding author.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin*. 2018;68(1):7-30.
2. Toro JR, Travis LB, Hongyu JW, Zhu K, Fletcher CDM, Devesa SS. Incidence patterns of soft tissue sarcomas, regardless of primary site, in the Surveillance, Epidemiology and End Results program, 1978-2001: an analysis of 26,758 cases. *Int J Cancer*. 2006;119(12):2922-2930.
3. Pastore G, Peris-Bonet R, Carli M, Martínez-García C, de Toledo JS, Steliarova-Foucher E. Childhood soft tissue sarcomas incidence and survival in European children (1978-1997): report from the Automated Childhood Cancer Information System project. *Eur J Cancer*. 2006;42(13):2136-2149.
4. Gamboa AC, Gronchi A, Cardona K. Soft-tissue sarcoma in adults: an update on the current state of histiotype-specific management in an era of personalized medicine. *CA Cancer J Clin*. 2020;70(3):200-229.
5. Skrede O-J, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet*. 2020;395(10221):350-360.
6. Pantanowitz L, Quiroga-Garza GM, Bien L, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet Digit Health*. 2020;2(8):e407-e416.
7. Yamashita R, Long J, Longacre T, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol*. 2021;22(1):132-141.
8. Woerl A-C, Eckstein M, Geiger J, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. *Eur Urol*. 2020;78(2):256-264.
9. Cancer Genome Atlas Research Network. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*. 2017;171(4):950-965.e28.

10. Vahadane A, Peng T, Albarqouni S, et al. Structure-preserved color normalization for histological images. In: *Proceedings of the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. Brooklyn, NY: IEEE; 2015:1012-1015.
11. Anand D, Ramakrishnan G, Sethi A. Fast GPU-enabled color normalization for digital pathology. In: 2019 International Conference on Systems, Signals and Image Processing (IWSSIP). 2019:219-224.
12. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. 2017:2261-2269.
13. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV: IEEE; 2016:2921-2929.
14. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009;45(4):427-437.
15. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep*. 2017;7:1-27.
16. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301-1309.
17. Arbiser ZK, Folpe AL, Weiss SW. Consultative (expert) second opinions in soft tissue pathology analysis of problem-prone diagnostic situations. *Am J Clin Pathol*. 2001;116:473-476.
18. Soomers V, Husson O, Young R, Desar I, Van der Graaf W. The sarcoma diagnostic interval: a systematic review on length, contributing factors and patient outcomes. *ESMO Open*. 2020;5:e000592.
19. Bandyopadhyay D, Panchabhai TS, Bajaj NS, Patil PD, Bunte MC. Primary pulmonary artery sarcoma: a close associate of pulmonary embolism-20-year observational analysis. *J Thorac Dis*. 2016;8(9):2592-2601.
20. Nakamura T, Matsumine A, Matsubara T, Asanuma K, Uchida A, Sudo A. The symptom-to-diagnosis delay in soft tissue sarcoma influence the overall survival and the development of distant metastasis. *J Surg Oncol*. 2011;104(7):771-775.
21. Ferrari A, Miceli R, Casanova M, et al. The symptom interval in children and adolescents with soft tissue sarcomas. *Cancer*. 2010;116(1):177-183.
22. Mesko NW, Mesko JL, Gaffney LM, Halpern JL, Schwartz HS, Holt GE. Medical malpractice and sarcoma care—a thirty-three year review of case resolutions, inciting factors, and at risk physician specialties surrounding a rare diagnosis. *J Surg Oncol*. 2014;110(8):919-929.
23. Blay JY, Honoré C, Stoeckle E, et al. Surgery in reference centers improves survival of sarcoma patients: a nationwide study. *Ann Oncol*. 2019;30(7):1143-1153.
24. Linch M, Miah AB, Thway K, Judson IR, Benson C. Systemic treatment of soft-tissue sarcoma—gold standard and novel therapies. *Nat Rev Clin Oncol*. 2014;11(4):187-202.
25. Brennan MF, Antonescu CR, Moraco N, Singer S. Lessons learned from the study of 10,000 patients with soft tissue sarcoma. *Ann Surg*. 2014;260(3):416-421. discussion 421-422.
26. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med*. 2019;25(7):1054-1056.
27. Echle A, Grabsch HI, Quirke P, et al. Clinical-grade detection of microsatellite instability in colorectal tumors by deep learning. *Gastroenterology*. 2020;159(4):1406-1416.e11.
28. Copas JB. Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika*. 2002;89(2):315-331.
29. Liao P, Wu H, Yu T. ROC curve analysis in the presence of imperfect reference standards. *Stat Biosci*. 2017;9(1):91-104.
30. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer*. 2020;1:789-799.