UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Medicina - DIMED

CORSO DI DOTTORATO DI RICERCA IN ONCOLOGIA CLINICA E SPERIMENTALE E IMMUNOLOGIA

Development of a Deep learning-based pipeline to classify Small Round Cells Sarcomas histotypes

Coordinatore: Ch.mo Prof. Stefano Indraccolo

Supervisore: Ch.mo Prof. Angelo Paolo Dei Tos

Dottorando: Dr. Lorenzo Nicolè

2019-2022

INDEX

ABSTRACT……………………………………………………………………..4

1. INTRODUCTION……………………………………………………………5

1.1 Small Round Cell Sarcomas……………………………………………...5

1.1.1 Ewing Sarcoma………………………………………………………5

1.1.2 **Ewing-like Sarcomas**…………………………………………………...7

1.1.3 Synovial Sarcoma……………………………………………………9

1.2 Artificial Intelligence and Deep Learning in digital pathology…………..11

1.3 Work Hypothesis…………………………………………………………12

1.4 Graphical Abstract………………………………………………………13

2. EXPERIMENTS……………………………………………………………...14

2.1 Datasets implementation (WSI management & quality check)………….15

2.2 TilerPath…………………………………………………………………..15

2.3 The classifier……………………………………………………………...16

2.4 Results: training, test & validation……………………………………...17

3. **DISCUSSION & CONCLUSION & PERSPECTIVES** ……………………..19

REFERENCES……………………………………..……………...………22

To my Wife, to my Sons

May the wind under your wings
bear you where the sun sails
and the moon walks…
J.R.R.T.

ABSTRACT

Ewing sarcoma (ES), Ewing-like sarcomas (ELS) and undifferentiated synovial sarcoma (SS) represent the main entities belonging to the family of the Small Round Cell Sarcomas (SRCS), a group of rare, heterogenous and highly aggressive mesenchymal tumors. SRCS are classified according to a specific single gene rearrangement. However, despite specific histological features are strongly correlated with the underlying molecular alteration, morphological overlapping may occur, and combined with their rarity, make the diagnosis challenging especially for non-expert pathologists. Within this context, the spreading of digital pathology and the recent developments of deep learning technologies for image processing, offer new opportunities for analysis, interpretation, and classification of histopathological slides.

In this study, a deep learning-based framework called **DeeRasNET,** was specifically developed to classify hematoxylin and eosin-stained slides of ES, SS, *BCOR* and *CIC* rearranged sarcomas. Accuracy was the main metrics parameter used to evaluate the model performance.

Initially, due to the small size of the datasets implemented for the model training, the classification accuracy for each class of sarcoma resulted low (mean accuracy of 0.6). To increase the performance of the model, we developed a pre-processing semi-automated pipeline comprising an open-source graphical interface unit (called **TilerPath**) with which we managed the tissue whole slide images, selecting interesting tissue areas and performing a quality control of the images used for classifier implementation. By TilerPath uninformative and misleading images were excluded from the model. After pre-preprocessing by Tilerpath, a total of 18193 tiles, selected from 124 digital slides covering all the four histotypes investigated, was used to train and test DeeRasNET. Finally, the scalability of the system was demonstrated on a validation dataset comprising 2706 tiles randomly selected from cases not included into the training and test set.

After quality improvement, the final model showed a strong increase of classification performance, with accuracies ranging from 0.98 to 0.99 among all the sarcoma types.

Both the **TylerPath** and the **DeeRASnet** source code were released as open-source software.

CHAPTER 1

INTRODUCTION

Clinical problem & work Hypothesis

1.1 Small Round Cell Sarcomas

Small round cell sarcoma (SRCS) is a term used to indicate a large group of high aggressive malignant neoplasms that most often occurring in children and young adults. SRCS share a monotonous morphology of small cells with dark round nucleus, inconspicuous nucleoli and scant cytoplasm(Lessnick et al., 2009)(Sbaraglia et al., 2020). This category includes subtypes of sarcomas, and rare histological variants of carcinomas, lymphomas, and melanoma. Small round cell sarcomas (SRCS) include Ewing and Ewing-like sarcomas, alveolar rhabdomyosarcoma, desmoplastic small round cell tumor (DSRCT), poorly differentiated synovial sarcoma (round cell variant), round cell liposarcoma, and undifferentiated round cell sarcoma(Honoré et al., 2020)(Tos, 2018). Interestingly, SRCS are characterized by specific chromosomal translocations associated with an extremely low tumor mutational burden(Chalmers et al., 2017). The differential diagnosis among the SRCS entities remains a challenge for pathologists, and requires careful attention to a combination of clinical examination, imaging consultation, conventional histopathology examination and dedicated molecular pathology(Righi et al., 2019). This study focuses on the main four types of SRCS: Ewing sarcoma (ES), Ewing-like sarcoma (ELS, *CIC* and *BCOR* rearranged types), and monophasic synovial sarcoma (SS).

1.1.1 Ewing Sarcoma

ES accounts for less than 1% of all soft tissue sarcomas, affecting predominantly (approximately 80% of cases) the metaphysis of long bones, with a peak incidence between the first and the second decades(Grünewald et al., 2018). ES in adults tends to occur predominantly in the deep soft tissues of the paravertebral region and of the proximal portions of the lower and upper extremities. Visceral locations, such as kidney, pancreas, and meninges were also documented(Sbaraglia et al., 2020).

The typical molecular alteration that drives ES oncogenesis is the translocation of the *EWSR1 (*or less represented *FUS)* gene with one member of the ETS (avian Erythroblastosis virus Transforming Sequence) family of transcription factors. In 85% of cases the molecular alteration harbored by ESs is the t(11;22)(q24;q12) that encodes for the fusion product EWSR1-FL1. Less represented (about 10%) is the t(21;22)(q22;q12) that encodes for the fusion product EWSR1-ERG. The remaining 5% is covered by the other ETS family members which are *ETV1*, *ETV4* and *FEV* (**see table 1 for more details**).

Cancer Type	Molecular alteration	Gene fusion
Ewing sarcoma	t(11;22)(q24;q12)	*EWSR1-FLI1*
	t(21;22)(q22;q12)	*EWSR1-ERG*
		EWSR1-ETV1orETV4orFEV
Synovial sarcoma	t(X;18)(p11;q11)	*SYT-SSX (SSX2 or SSX4)*
BCOR-rearranged sarcoma	inv(x)(p11;p11)	*BCOR-CCNB3*
	BCOR-ITD	*BCOR-ITD*
	t(10;17)(q23.3;p13.3)	*YWHAE-NUTM2B*
	t(4;x)(p11;q31)	*BCOR-MAML3*
	t(x;22)(p11;q13.2)	*ZC3H7B-BCOR*
CIC-rearranged sarcoma	t(4;19)(q35;q13)	*CIC-DUX4*
	t(10;19)(q26;q13)	*CIC-DUX4*
	t(x;19)(q13;q13.3)	*CIC-FOXO4*
	t(15;19)(q14;q13.2)	*CIC-NUTM1*
	t(10;19)(q23.3;q13)	*CIC-NUTM2B*

Table 1. Genomic alterations of Ewing sarcoma, Synovial sarcoma and Ewing-like sarcomas investigated

The 5-year overall survival in localized disease is currently around 75% whereas in metastatic disease, it drops to approximately 30%(Casali et al., 2018).

Grossly ES appears as large and multilobulated mass, necrosis and/or haemorrhagic areas are frequent. Microscopically, ES is composed of a distinctively monomorphic round cell population, showing vesicular nuclei with finely dispersed chromatin and scant cytoplasm **(Figure 1)**. Cytoplasmic clearing may be observed. Tumor cells forming rosettes can be detected and are traditionally interpreted as evidence of neuroectodermal differentiation. Mitotic activity is usually high. Immunohistochemically strong CD99 membrane immunopositivity it's a characteristic finding in all cases of ES. However, as CD99 is expressed within a variety of mesenchymal tumors, its positivity needs to be evaluated in context with morphology. Importantly, in consideration of its remarkable sensitivity, CD99 immuno-negativity would strongly argue against a diagnosis of ES(Sbaraglia et al., 2020). S-100 protein, CD57, neurofilaments, cytokeratin, and desmin are markers that may be expressed in ES however with no particular diagnostic utility. *FLI-1* and *ERG* expression can be seen in those ES harboring *EWSR1-FLI1* and *EWSR1-ERG* gene fusions, respectively. Recently, expression of *PAX7* has been shown to represent another promising diagnostic tool for those

ES demonstrating a fusion between *EWSR1* and *FLI1*, *ERG* (Charville et al., 2017)(Charville et al., 2019). NKX2–2, a home domain transcription factor involved in neuroendocrine/glial differentiation and a downstream target of *EWSR1-FLI1*, has been reported as an immunohistochemical marker for ES(Hung et al., 2016).

1.1.2 Ewing-like sarcomas

Ewing-like tumors (ELS) comprise three main categories: round cell sarcomas with *EWSR1* gene fusion with non-ETS family members (not discussed in this work), *CIC*-rearranged sarcomas, and *BCOR*-rearranged sarcomas(Sbaraglia et al., 2020).

CIC-rearranged small round cell sarcomas

The most frequent and best characterized subgroup of the family of ELS is the *CIC*-rearranged sarcoma. This tumor occurs mostly in children and young adults, with median age in the second decade. Most tumors arise in the deep soft tissue of the trunk, limbs, or head and neck region, sometimes with secondary bone involvement. Superficial soft tissues are primarily involved in less than 10% of cases. Occurrence as bone primary is extremely rare, whereas visceral location is reported in approximately 10% of cases. The recently reported *CIC-NUTM1* variant seems to exhibit significant tropism for the central nervous system(Le Loarer et al., 2019).

CIC sarcomas generally showed a dismal prognosis with poor response to the standard chemotherapeutic protocols, and most often presents with lung metastasis at onset. The 5-year overall survival is less than 50%(Sbaraglia et al., 2020).

This entity represents an undifferentiated round cell malignancy characterized by the recurrent *CIC* (capicua transcriptional repressor) gene rearrangements. The *CIC* gene is the human homolog of the Drosophila gene Capicua. It encodes a high-mobility group box transcription factor mainly involved in the development of the central nervous system(Lee, 2020). The most common molecular alterations described so far are the translocation t(4;19) (q35;q13) or the t(10;19)(q26;q13), both encoding for fusion product CIC-DUX4 (Richkind et al., 1996). Fusions with non-*DUX4* gene partners (*FOXO4, LEUTX, NUTM1*, and *NUTM2A*) are less common and occur in approximately 5% of cases and are sustained by specific molecular alterations (**Table 1**)(Italiano et al., 2012). Morphologically, *CIC*-rearranged sarcoma appears less monotonous than ES, featuring mild-to-moderate pleomorphism. In particular, areas with vesicular nuclei and distinctive nucleoli are present (**Figure 1**). In rare cases, neoplastic cells assume an epithelioid morphology with occasionally rhabdoid-like cytoplasm or with clear cell change of cytoplasm. Neoplastic cells may also be observed in a lobular growth pattern, with associated fibrous septa. Confluent necrosis represents a frequent finding and myxoid change of the stroma may also be observed. Rarely focal cell spindling can be appreciated. Mitotic activity is typically high. Immunohistochemically, CD99 staining is observed in approximately 85% of cases, however, it is often patchy and lacks the strong, diffuse membranous pattern observed in ES. Nuclear expression of DUX4 is consistently present. ETV4 is diffusely expressed as a consequence of the upregulation of the *ETV4* gene; however, it is not entirely specific. In fact, 10% of ES, rare DRCTs, rhabdomyosarcomas, and melanomas may also show ETV4 nuclear expression(Le Guellec et al., 2016). Immunoreactivity for both n- and c-terminus of WT1, desmin, and S100 has been reported.

BCOR-rearranged small round cell sarcomas

BCOR-rearranged tumors were described for the first time only in 2012 by the Pierron group. This entity accounts for approximately 4% of round cell sarcomas, with a striking male predominance and the peak of incidence in the second decade. (Pierron et al., 2012).

BCOR-rearranged sarcomas tend to occur more frequently in bone than in soft tissues, and involves more frequently the pelvis, the lower limbs, and the paraspinal region. Visceral location is reported but appears to be extremely rare. Importantly, compared to ES and *CIC*-rearranged sarcomas, patients with *BCOR*-rearranged sarcomas seem to have a more indolent clinical behavior(Puls, Niblett, Marland, Gaston, Douis, Mangham, et al., 2014).

The morphologic spectrum of *BCOR-CCNB3* sarcoma is rather broad, with tumors composed of a mixed proliferation of round and spindle cells arranged in sheets or fascicles. In some cases spindling may predominate. Nuclei are angulated and hyperchromatic with finely dispersed chromatin. In the majority of cases, nucleoli are not prominent. Significant variation in cellularity and myxoid change of the stroma is sometimes seen. Small foci of necrosis are commonly seen. Mitotic activity is often very high (**Figure 1**). When compared with the primary tumor, recurrent and metastatic lesions show increased cellularity and higher pleomorphism, occasionally simulating undifferentiated pleomorphic sarcoma. Immunohistochemically, almost all cases exhibit strong and diffuse cyclin B3 (CCNB3) nuclear positivity, with only a few cases showing patchy staining. CCNB3 staining is highly specific although cytoplasmic staining may be seen in several sarcomas including ES and SS(Puls, Niblett, Marland, Gaston, Douis, Chas Mangham, et al., 2014). BCOR immunoreactivity can also be observed but it is less specific than CCNB3. CD99 staining is generally weaker or totally absent. Two-thirds of cases may show SATB2 expression. *BCOR-CCNB3* rearrangements account for 60% of *BCOR* gene alterations. This fusion originates from a paracentric inversion on the X-chromosome and splicing of the end of the *BCOR* coding sequence to the *CCNB3* exon 5 splice acceptor site. The resultant fusion protein is composed of full-length *BCOR*, a transcriptional repressor encoding the Bcl-6 co-repressor, and the C-terminus of *CCNB3*, a cyclin normally expressed in leptotene and zygotene phases of meiosis(Nguyen et al., 2002). In vitro studies suggest that the *BCOR-CCNB3* fusion protein is oncogenic and drives proliferation in this sarcoma. Recently alternative fusion partners of *BCOR* have been identified, including *MAML3* (a member of the mastermind-like family of transcriptional coactivators), and *ZC3H7B* (a zinc-finger CCCH domain-containing protein 7B)(Specht et al., 2016). *BCOR* internal tandem duplication (ITD) has been described in a subgroup of round cell sarcoma of infancy most often involving the soft tissue of the trunk, retroperitoneum, and head and neck region. Despite remarkable clinical and occasionally pathological similarities to ES, gene profiling and single nucleotide polymorphic allele (SNP) array analyses indicate that this new group of tumors is biologically distinct from both ES and *CIC*-rearranged sarcoma.

1.1.3 Synovial sarcoma

SS represents about 10% of all soft tissue sarcoma of the adult. In particular SS affects young adults more frequently within the second-third decade. More common sites affected are limbs and head and neck, and a slight male predominance is reported. The 20% of SS presented at diagnosis as poorly differentiated SS, that often, presents a round small cell morphology.

Poorly differentiated synovial sarcoma has a poor prognosis, with an even higher metastatic rate than conventional forms of SS.

Synovial sarcoma has a characteristic chromosomal translocation, t(X;18)(p11;q11), that results in fusion of the *SS18 (SYT)* gene at chromosome 18 to *SSX* genes, which have two different copies, *SSX1* (SYT-SSX1) and *SSX2* (SYT-SSX2), located in two subregions of chromosome Xp11 (23 and 21, respectively), extremely rare fusion partner is also SSX4(El Beaino et al., 2020). Synovial sarcoma provides a clear example of the correlation that may exist between the fusion transcript type and the tumor phenotype. Interestingly, *SYT-SSX1* fusions are associated with biphasic synovial sarcoma (in both epithelioid and spindle cell elements), whereas the monophasic variant usually contains *SYT-SSX2* fusions(Antonescu et al., 2020). No significant correlations exist between the round cell variant of poorly differentiated synovial sarcoma and a specific transcript subtype(Sbaraglia et al., 2020).

Poorly differentiated synovial sarcoma has three morphologic sub-variants: the most common round cell variant, a large cell epithelioid variant, and a high-grade spindle cell variant(Gazendam et al., 2021). The round cell variant of poorly differentiated synovial sarcoma, compared with other synovial sarcoma variants, more frequently shows necrosis, a high mitotic rate (>10 mitoses/10 high-power fields), vascular invasion, and a hemangiopericytoma-like pattern of growth (**Figure 1**). Pericellular collagen deposition is an important diagnostic clue of SS(Sbaraglia et al., 2020). The immunohistochemical profile of the round cell variant of SS is similar to that of more conventional subtypes, and patchy expression of EMA and keratin in scattered cells is typical. Nuclear TLE1 was identified as a highly sensitive and relatively specific marker of SS. CD99 immunoreactivity may be seen in SS and may become a source of diagnostic confusion (especially in small biopsies specimens) because it is also expressed in ES, which is the main differential diagnosis(Sbaraglia et al., 2020). EMA and keratin expression in poorly differentiated synovial sarcoma can be limited, and these markers may be completely negative in small biopsy specimens. In this context, molecular testing is crucial, not only for the proper diagnosis but also for treatment strategies.

Recently antibodies reacting against the fusion product of *SYT* gene have been introduced and represent an optimal surrogate for molecular tests.

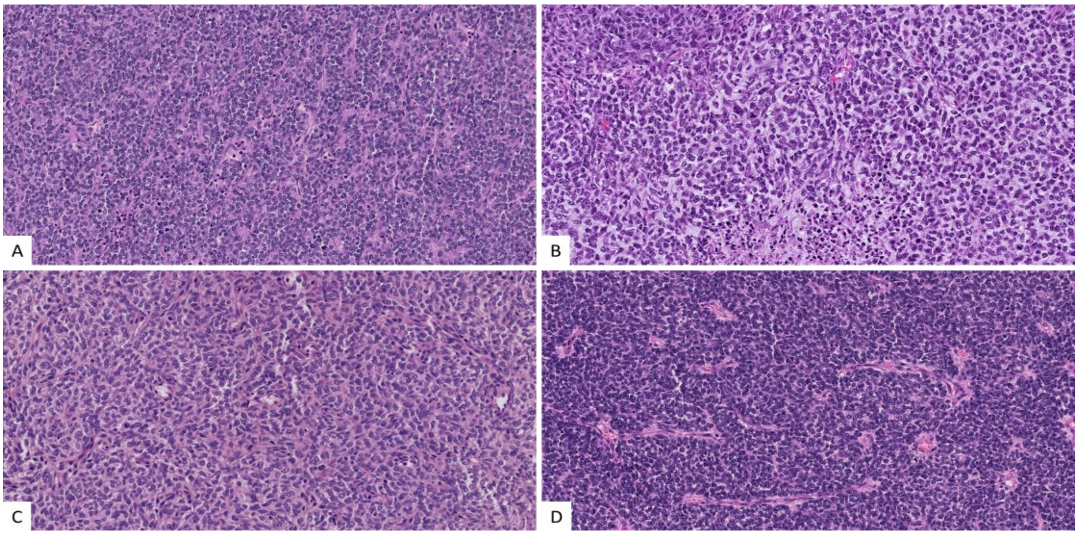


Figure 1. Representative histology of Ewing sarcoma (A); CIC-rearranged sarcoma (B); BCOR-rearranged sarcoma (C) and a monophasic synovial sarcoma with small round cell morphology (D). Despite histological differences among these entities may be appreciated, figure shows as the small round cell morphology may predominate the histological pattern creating confusion and diagnostic challenges. Haematoxylin & eosin stain, original magnification 200x.

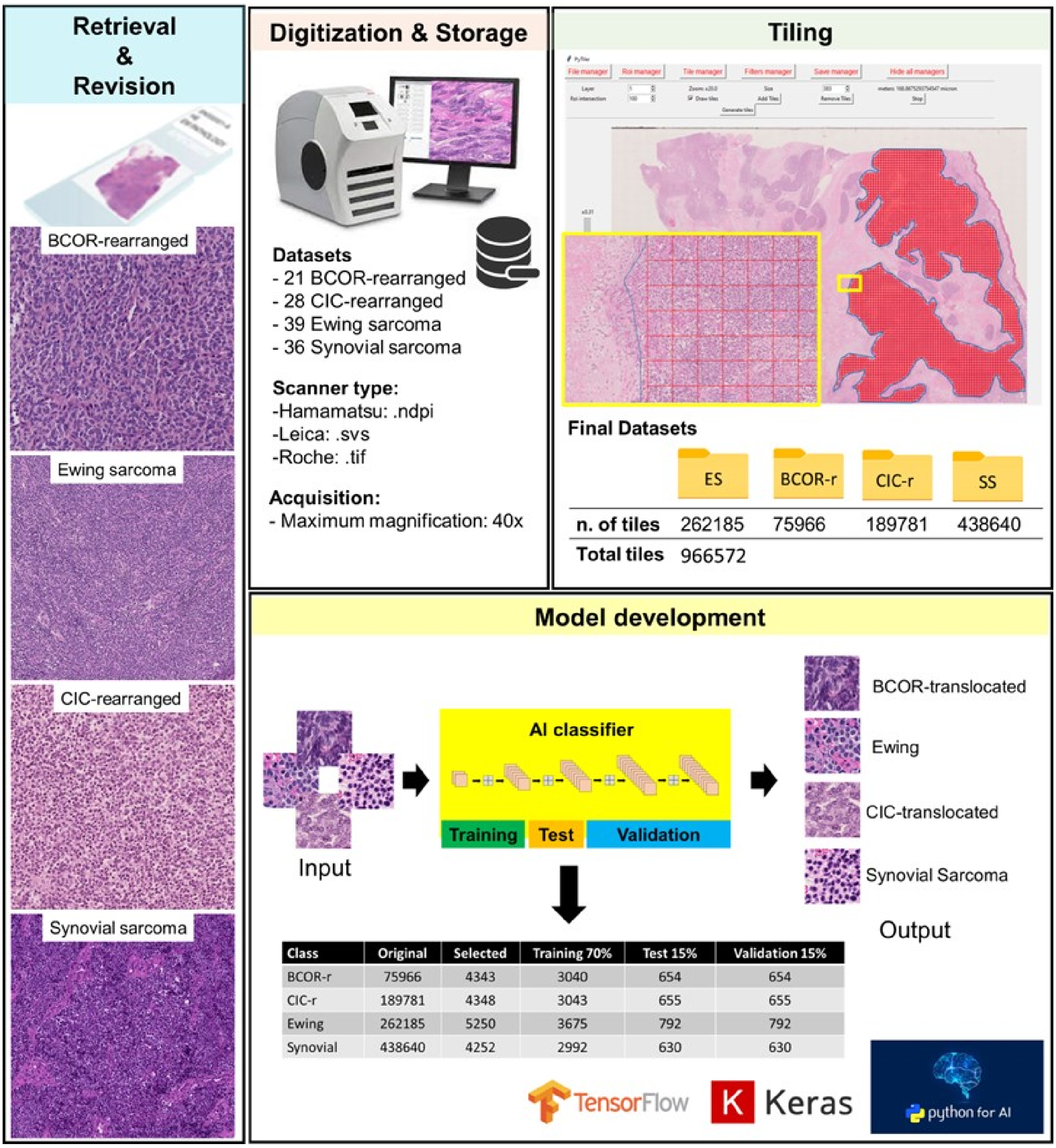
1.2 Artificial Intelligence and Deep Learning in digital pathology

Artificial Intelligence (AI) is a branch of computer science focusing on development of algorithms able to autonomously solve decisional tasks on behalf of the human being(Nativi & Craglia, n.d.). AI systems may be implemented through different computational techniques as Machine Learning (ML) and Deep Learning (DL)(Cohen, 2020). ML is a computational strategy in which the output (e.g. a final decision of a classification task) of the algorithm is automatically defined through statistical and data-driven rules that are not previously defined by humans, and are extracted from a large set of input data. In ML models input data are represented by exemples of the specific event that comprises the relationship between input and output data (e.g. two classes of images that must be classified). However, ML systems are limited by the fact that they require expertise and human engineering to design feature extractors that transform raw input data into suitable mathematical representations from which the algorithm could detect the rules (also called representations) that link the input with the output(Maier et al., 2019). In contrast, DL is a subfield of ML in which the algorithm is designed to autonomously define through the training and the test processes its own mathematical representation from raw data, achieving in this way a final decision without human inferences(Esteva et al., 2019). During the last decades, DL has had a tremendous impact on various fields of science, and in particular in image analysis tasks such as image classification. Convolutional neural networks (CNNs), is a type of DL algorithm inspired by the biological architecture of the neurons that is specifically designed to process data that exhibits natural spatial invariance (e.g. images)(Laak et al., 2021). CNNs play a pivotal role in tasks such as image classification and segmentation, especially in the field of pathology, in which the transition towards digital pathology represents an optimal substrate for the development of these technologies. During the last years, a plethora of tasks comprising tumor detection, tissue classification, biomarkers extractions, prognosis stratification and prediction analysis have been performed by CNN algorithm using H&E stained slides of several types of cancer(Cohen, 2020; Gertych et al., 2019; Hekler et al., 2019; Tran et al., 2021). However, despite these encouraging results many diagnostic challenges still have to be addressed. From a technical point of view, large-scale validation of algorithms is crucial to confirm safety and accuracy for broad applications that represent fundamental requirements for regulatory approval, and actually most of the studies reported results based on internal validation with a limited number of cases, and large public datasets truly representative of the true clinicals scenario are still lacking(Laak et al., 2021). Moreover, AI solutions based on DL models are systems not explicitly programmed, meaning that it is very difficult for humans to understand the exact functioning of the systems that lead to the final decision (the so-called 'black box' concept). To fix this problem, data scientists have focused on strategies of explainable AI, in which techniques are developed to better understand the functioning leading to the final decision of the model(Guidotti et al., 2019). Finally, the use of patient data for algorithms development and the 'black box' problem lead to ethical and legal concerns that have to be addressed properly. The European Commission published guidelines for the development of trustworthy AI, specifying the correct framework to help researchers to achieve AI solutions that are ethical and lawful within a specific clinical context. The main conclusion of the European Commission's amendment is concerned with the need to overcome the concept of 'black box' systems encouraging explainable AI solutions, and also to create a reliable framework for data management and sharing that avoids iniquities with the nature of the data used for the models implementation (such as concerns regarding socioeconomic status, race, ethnic, religious background, gender and disability), and that could affect the generalizability of the decisional pipeline(Cannarsa, 2021). Finally, regarding the scalability of AI solutions in the field of pathology, it is imperative to realize that pathologists formulate diagnosis not only by analyzing a sample of tissue under the microscope, but also integrating clinical information from different sources, their own clinical expertise and the specific circumstances of the patient. It is more reasonable to think that AI solutions in the midterms may help pathologists in tedious and time-consuming activities such as first-line tissue screening or IHC evaluation and control quality, rather than a mere replacement of human activity and judgment.

1.3 Work Hypothesis

Based on the recent improvements of AI applications for image processing, we assume that a DL-based algorithm will be able to recognize the specific histological subtype of SRCS through automatic analysis of histological H&E slides. We also assume that the DL-based classifier can be conceived as part of a data-driven diagnostic pipeline able to improve diagnostic accuracy even with rare cancer histotypes.

1.4 Graphic abstract



	ES	BCOR-r	CIC-r	SS
n. of tiles	262185	75966	189781	438640
Total tiles	966572			

Class	Original	Selected	Training 70%	Test 15%	Validation 15%
BCOR-r	75966	4343	3040	654	654
CIC-r	189781	4348	3043	655	655
Ewing	262185	5250	3675	792	792
Synovial	438640	4252	2992	630	630

Figure 2. Graphic overview representing the four main phases of the work with their key features showed in each box. *Tensorflow, Keras* and *Python* were the software resources used to implement the entire pipeline and to analyse all data.

CHAPTER 2

Experiments implementation

2.1 Datasets implementation (WSI management & quality check)

For this study a dataset for each SRCS entity investigated was created. Cases with a diagnosis of ES, *CIC*-rearranged sarcoma, *BCOR*-rearranged sarcoma and SS (biphasic SS were excluded), were collected from the archives of the Pathology Department of the University of Padova, from the Istituto Oncologico Veneto (IOV, Padova, Italy), and from the Istituto Ortopedico Rizzoli (Bologna, Italy) (**see Table 2 for clinical data**).

Histotype	**Cases**	**Sex (M/F)**	**Age (mean,min-max)**	**Tiles**
Ewing sarcoma	39	20/19	36.4 (15 - 83)	262185
Synovial sarcoma	36	21/15	30.2 (18 - 61)	438640
CIC sarcoma	28	11/17	29.1 (10 - 42)	189781
BCOR sarcoma	21	10/11	17.8 (7 - 25)	75966
Total	**124**	**62/62**	**28.4 (7- 83)**	**966572**

Table 2. Clinical data and tiles generated for each class of tumour investigated

Only cases confirmed by molecular tests were considered. Each case was reviewed by expert pathologists to confirm the diagnosis and then scanned at the maximum resolution available (40X) using three different types of scanner: 1) Hamamatsu Nanozoomer (Hamamatsu Photonics, Hamamatsu, Japan), Aperio (Leica biosystems, Wetzlar, Germany) and Ventana DP200 (Roche, Basel, Switzerland). A total of 124 cases were collected including 39 ES, 36 SS, 28 *CIC*-rearranged sarcomas and 21 *BCOR*-rearranged sarcomas (**Figure 1**). In each slide, a representative tumor area was outlined by a pathologist and then subdivided into non overlapping tiles of size 383x383 pixels, which were then exported to the corresponding dataset. Tiles was generated with TilerPath, a custom tool developed using Python coding language **(see section 2.2 for details about TilerPath**). Finally, 966572 tiles were generated and implemented into the final datasets: 438640 (40%) tiles for SS, 262185 (27%) tiles for ES, 189781 (20%) tiles for *CIC*-rearranged and 75966 (13%) tiles for *BCOR*-rearranged (**Figure 2**).

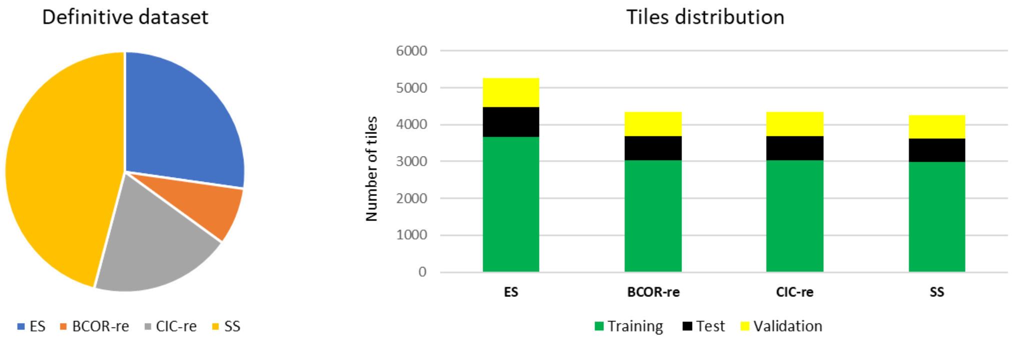
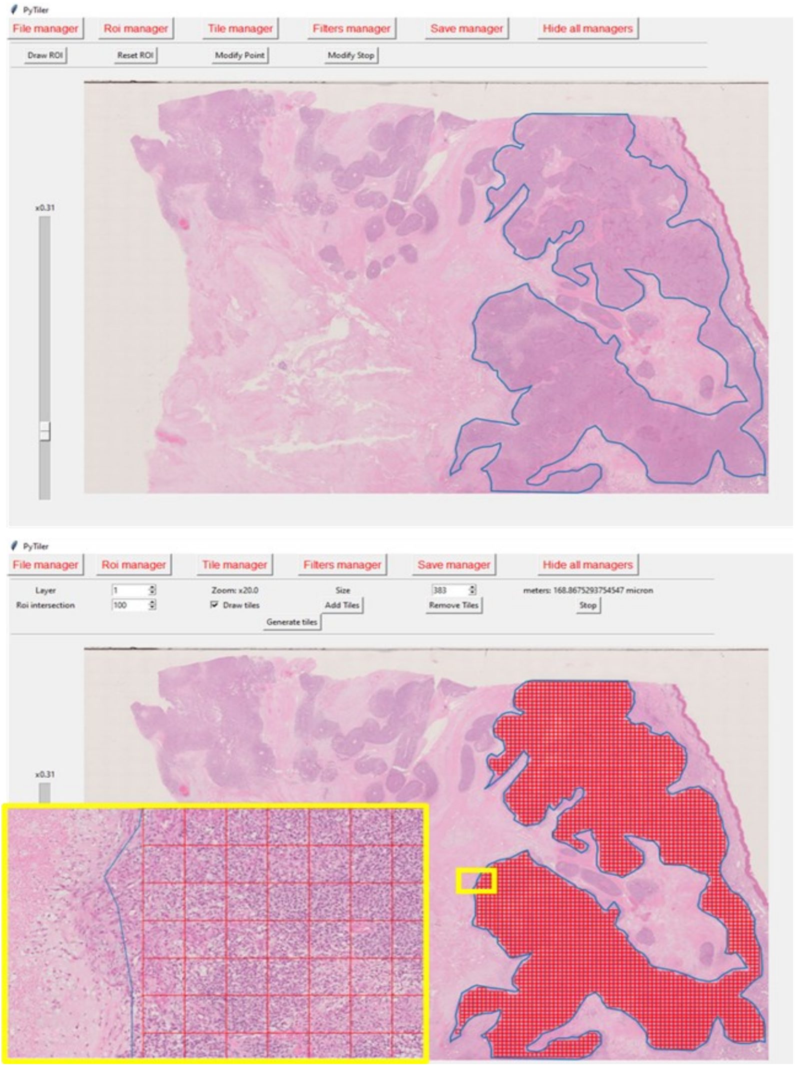


Figure 3. Graphic overview of the final datasets used to develop DeeRasnet. Tiles as reported in the document are balanced across the four different tumors in order to harmonize the training.

2.2 TilerPath

The management of the WSI for AI research purposes remains a challenging task that requires specific actions generally not provided by commercial slides viewers. Such actions comprise direct tissue annotation or labeling, tiles generation and automatic dataset implementation. To fill this gap we developed a graphical user interface called TilerPath. TilerPath was conceived as an open-source tool mainly based on the open-source library Openslide that allow the user to open different WSI format files, to navigate over the tissue at different level of magnification, to create custom region of interest, custom size tiles generation, to perform a quality check on single tiles and to export tiles into specific directory (**Figure 3**). Quality check was based on the ability to analyze each tile by the Stardist algorithm directly implemented into TilerPath to calculate and compare with a threshold set by the user, the complessive percentage of the area of the cells nuclei covering the tile surface. In this way the user can choose the cellularity of each tile and discard unnecessary tiles. Moreover, TilerPath is implemented with a library containing a transformative convolutional neural network that we trained to recognize and discard tiles with common tissue artifacts (tissue folding, dark spots, ink, marker, dusts, electrocution effects and area scanned out of focus) that could affect the classification performance (Manuscript under revision). More technical details about TilerPath are presented in a dedicated paper (Manuscript under revision).



Layer 1
Roi intersection 100
Zoom: x20.0
☑ Draw tiles
Size 383
meters: 168.8675293754547 micron

Figure 4. Representative screenshots of TilerPath during the datasets creation. Each whole slide image may be opened and explored at the original magnification provided by the scanner. A *File manager* tool allows the user to select the path in which are stored the scanned slides, and to select a specific output path in which the tiles generated will be stored. The ROI manager tool allows the user to outline the tissue area that will be tiled. Tiles dimension and capture magnification may be managed within the Tiles manager tool, tils don't overlap each other and remain confined within the selected area (yellow box). Filters manager tool allow the user to run algorithms for the tiles quality control, with this tool cellularity cut-off may be selected and tiles with tissue artifacts automatically excluded from the export. To use Filters manager tool a Graphical Processing Unit (GPU) is recommended, users without GPU can skip Filters manager without interfering with the other functions of the system. Finally, the Save manager tool allows to export the generated tiles to specific directory created by the user. TylerPath was specifically designed to run on Windows 10 or further version.

2.3 The classifier

In this project the Google's inception network (GIN) was applied. GIN is an advanced DL architecture that performs convolution and pooling operations in parallel through the so-called inception block. In particular the InceptionV3 CNN architecture was adopted for the SRCS classification model to classify image tiles from the four different types of SRCS investigated. Input tiles were sized at 383x383 pixels. Model was pretrained on ImageNet and subsequently trained with stochastic gradient descent algorithms in Keras with TensorFlow backend. No data augmentation was adopted for training. The CNN architecture comprised five convolutional layers interconnected by four max-pooling layers and a final fully-connected layer (**Figure 4)**.

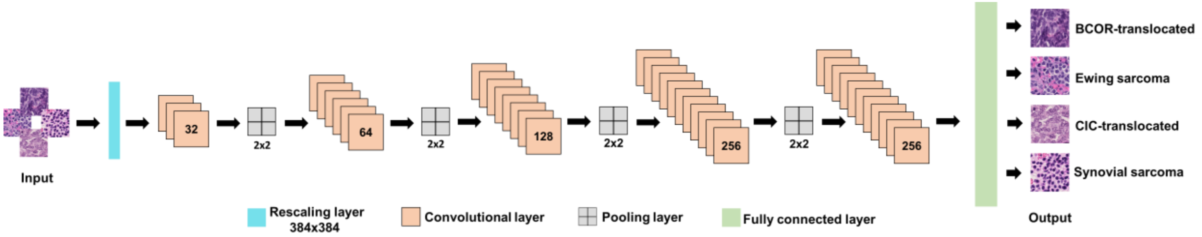
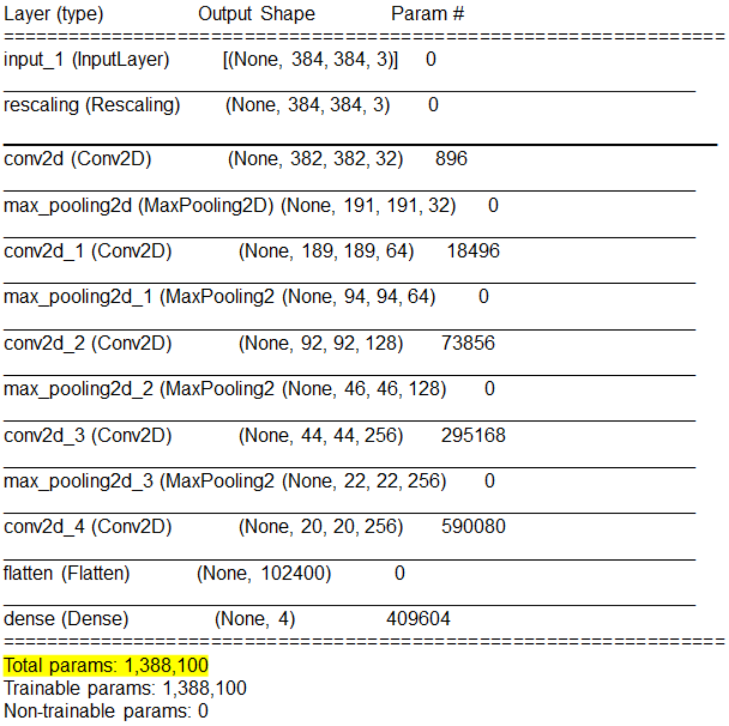


Figure 5. Graphic overview of *DeeRASnet* architecture

2.4 Results: training, test & validation

Morphology of each specific histotype of SRCS is displayed in **Figure 1,** and the basic clinical characteristics of the cases selected for his study are summarized in **Table 2**. ES was the more represented entity (31% of the cases) while the less represented was the *BCOR*-rearranged class (21% of the cases). By automatic quality check with TilerPath a total of 94789 tiles (10%) were removed from the native datasets. Considering that the number of the tiles generated was different among the four classes, the size of the datasets used for the experiments was balanced according to the number of available tiles of the less represented class (*BCOR*-rearranged sarcoma). Finally for the experiments a total of 18193 tiles were randomly selected and used for the experiments. For the training phase the 70% of tiles for each class (3675 tiles for ES, 2995 tiles for SS, 3040 tiles for *BCOR*-rearranged and 3043 tiles for *CIC*-rearranged) was used. The remaining 30% of the tiles were splitted for the test (15%, 792 tiles for ES, 630 tiles for SS, 654 tiles for *BCOR*-rearranged and 655 tiles for *CIC*-rearranged) and the final validation (15%, 792 tiles for ES, 630 tiles for SS, 654 tiles for *BCOR*-rearranged and 630 tiles for *CIC*-rearranged) phase (**Figure 3**). Tiles were assigned to the training, test and validation phases avoiding splitting tiles from the same patient. The model identified a total of 1,388,100 parameters to solve the classification task with the best accuracy (**Figure 5**). The final model performance was shown through a confusion matrix and sensibility, specificity and final accuracy resulted: 0,99, 0,99 and 0,99 for *BCOR*-rearranged sarcoma; 0,99, 0,99 and 0,99 for *CIC*-rearranged sarcoma; 0,98, 0,99 and 0,98 for ES; 0,98, 0,99 and 0,98 for SS (**Figure 6**).



```
Layer (type)                Output Shape            Param #
=================================================================
input_1 (InputLayer)         [(None, 384, 384, 3)]     0

rescaling (Rescaling)        (None, 384, 384, 3)       0

conv2d (Conv2D)              (None, 382, 382, 32)      896

max_pooling2d (MaxPooling2D) (None, 191, 191, 32)      0

conv2d_1 (Conv2D)            (None, 189, 189, 64)      18496

max_pooling2d_1 (MaxPooling2 (None, 94, 94, 64)        0

conv2d_2 (Conv2D)            (None, 92, 92, 128)       73856

max_pooling2d_2 (MaxPooling2 (None, 46, 46, 128)       0

conv2d_3 (Conv2D)            (None, 44, 44, 256)       295168

max_pooling2d_3 (MaxPooling2 (None, 22, 22, 256)       0

conv2d_4 (Conv2D)            (None, 20, 20, 256)       590080

flatten (Flatten)            (None, 102400)            0

dense (Dense)                (None, 4)                 409604
=================================================================
Total params: 1,388,100
Trainable params: 1,388,100
Non-trainable params: 0
```

Figure 6. Details of the parameters used by *DeeRaSnet* to obtain the best classification performance.

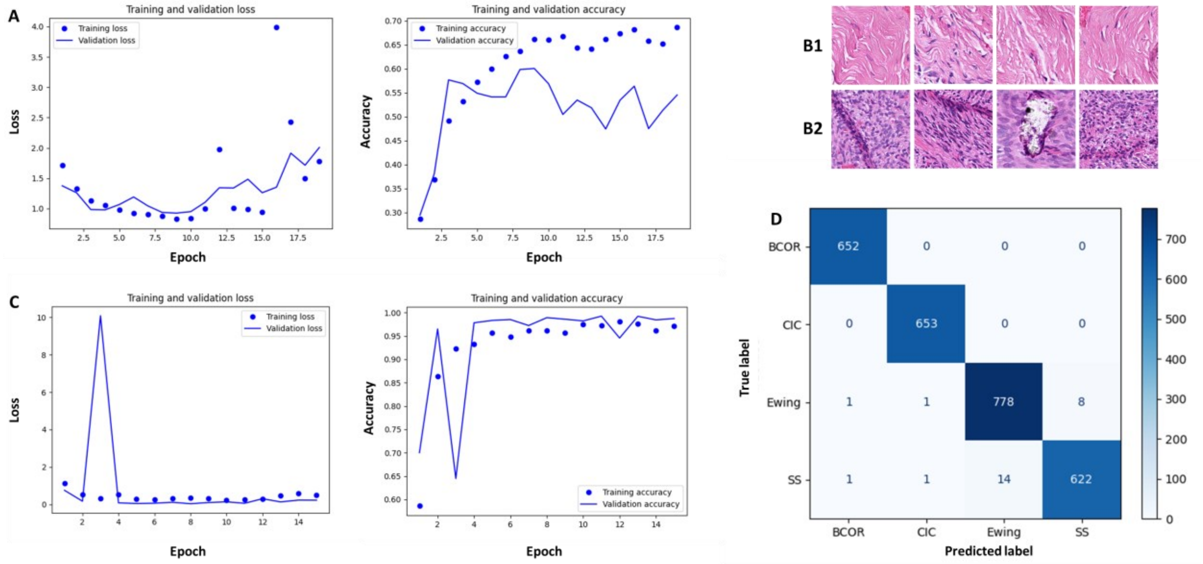


Figure 7. Summary of performance of all the classification tasks before (A) and after (C) pre-processing with *TIlerPath*. Box B shows the significative accuracy increase after quality check, removing tiles with low cellularity (B1) and with tissue artifacts (B2). In D the confusion matrix of all patients evaluated stratified random permutation cross-validation.

CHAPTER 3
Discussion & Conclusion & Perspectives

Soft tissue sarcomas (STSs) refer to a group of neoplasms subdivided in at least 100 different histologic and molecular subtypes, with each subtype displaying variable clinical behavior(W. H. O. Classification WHO Classification of Tumours Editorial Board & Who Classification of Tumours Editorial, 2020). Globally STSs account for nearly 20% of all pediatric solid malignancies and less than 1% of all adult solid malignancies(Burningham et al., 2012). Within STSs, specific entities such as the SRCSs account together for less than 1%, representing ultra rare histotypes affecting principally children and young adults. Despite continuous advances in translational research, rarity, heterogeneous morphology and also overlapping morphological and molecular features make diagnosis of STSs still challenging especially for those with an extremely low incidence(Gamboa et al., 2020). Despite substantial differences in biological behavior, clinical prognosis and response to treatments make a correct histological diagnosis mandatory, expertise in this field of pathology is still limited to few reference centers (Mesko et al., 2014). Moreover, in many peripheral centers, due to bureaucratic quibbles, limitations prevent a tempestive access to referral centers for second opinions and for non routinary diagnostic analysis, leading to a significant delayed diagnosis with a related negative impact on the lives of patients and their caregivers(Gamboa et al., 2020). Digitalization and AI solutions may almost in part play a central role in providing concrete solutions to overcome these challenges. However, despite AI solutions have been largely investigated with encouraging results in most fields such as breast, lung and urological pathology(Cohen, 2020), these techniques have not been investigated into the field of ultra rare cancers as SRCS. In this work, we present a possible solution to approach rare cancer histotypes with undifferentiated small round cell morphology. Our DL-based model, based on a CNN trained to recognize the four main types of SRCS on H&E stained slides, showed a great classification performance with final accuracies ranging from 98% to 99%. A high histological prediction yield in this context means that the pathologists, especially those working in non-reference institutions may make more appropriate decisions focusing on specific available ancillary tests to support the predicted diagnosis avoiding wasting material and preserving it for the confirmatory molecular analysis. Importantly, we stress the fact that DeeRASnet remains a tool that is not designed to substitute experts' opinions, which remain the fundamental steps into the diagnostic workflow of all the rare tumors(Gamboa et al., 2020), but rather as a technological solution to support pathologists within a diagnostic workflow completely under a full human supervision as presented in **Figure 7**. In a research context, scalability and usefulness of DL-based classifiers in the field of STSs have been investigated in a previous study by which authors demonstrated that this supporting technology may contribute to shortening the diagnostic time and saving biological material and costs(Foersch et al., 2021). Despite encouraging results obtained by this study, some critical points need to be pointed out with the perspective to set the basis for further improvements. Firstly, due to their extreme rarity, SRCS entities as *EWSR1*-non_ETS gene rearranged sarcoma, desmoplastic round cell tumor, small round cell osteosarcoma and mesenchymal chondrosarcoma were not included into the training sets and therefore not recognizable by DeeRASnet. Secondly, we are not yet able to explain the mathematical representations (and their corresponding features within the image) that DeeRASnet used to classify the input images, making our approach still at a 'black box' level. Thirdly, a large external validation set to assess the true scalability of DeeRASnet into a larger clinical context is still lacking limiting inference about the true generalizability of the model.

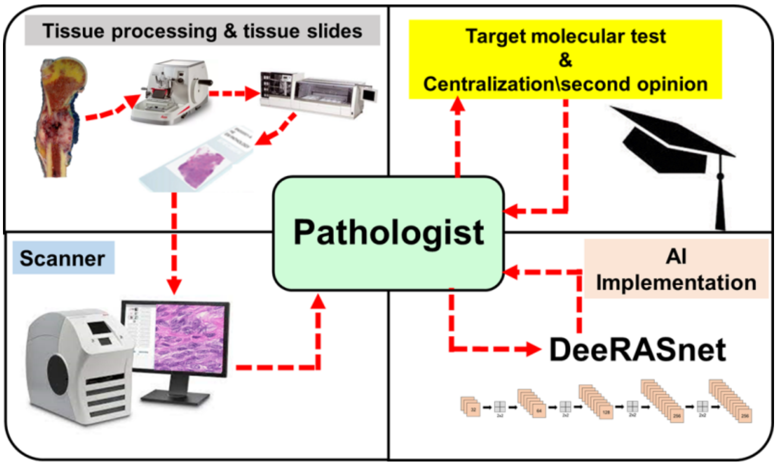


Figure 8. Schematic representation of diagnostic workflow implemented with *DeeRASnet.* Note the pivotal role played by the pathologist who remains responsible for all the steps.

Despite these critical points, results obtained so far are really encouraging; we achieved an high classification performance if compared with most of the recent studies published so far(Coudray et al., 2018; Kather et al., 2020), and the implementation of our training dataset with all the SRCS histotypes remain a priority challenge that we intend to solve in near future extending our collaborative research network. Moreover, this study serves also as a model for showing the feasibility of an AI-based approach to rare and ultra-rare tumours, for us an important point to highlight, since rare tumors represent up to 30% of all tumors meaning that about 1 out of 4 new cancer patients has a diagnosis of rare cancer (Gatta et al., 2011). Finally, to make DeeRASnet a more transparent classifier, through CNN ‘reverse analysis’ and sophisticated visualization techniques we intend to define the morphological parameters that are used by the machine to achieve the final classification, and smart and scalability solutions will be designed to insert and test DeeRASnet within the clinical context.

REFERENCES

Antonescu, C. R., Agaram, N. P., Sung, Y.-S., Zhang, L., & Dickson, B. C. (2020). Undifferentiated round cell sarcomas with novel SS18-POU5F1 fusions. *Genes, Chromosomes & Cancer*, *59*(11), 620–626.

Burningham, Z., Hashibe, M., Spector, L., & Schiffman, J. D. (2012). The epidemiology of sarcoma. *Clinical Sarcoma Research*, *2*(1), 14.

Cannarsa, M. (2021). Ethics Guidelines for Trustworthy AI. In *The Cambridge Handbook of Lawyering in the Digital Age* (pp. 283–297). https://doi.org/10.1017/9781108936040.022

Casali, P. G., Abecassis, N., Aro, H. T., Bauer, S., Biagini, R., Bielack, S., Bonvalot, S., Boukovinas, I., Bovee, J. V. M. G., Brodowicz, T., Broto, J. M., Buonadonna, A., De Álava, E., Dei Tos, A. P., Del Muro, X. G., Dileo, P., Eriksson, M., Fedenko, A., Ferraresi, V., … ESMO Guidelines Committee and EURACAN. (2018). Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, *29*(Suppl 4), iv268–iv269.

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., Huang, F., He, Y., Sun, J., Tabori, U., Kennedy, M., Lieber, D. S., Roels, S., White, J., Otto, G. A., … Frampton, G. M. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, *9*(1), 34.

Charville, G. W., Wang, W.-L., Ingram, D. R., Roy, A., Thomas, D., Patel, R. M., Hornick, J. L., van de Rijn, M., & Lazar, A. J. (2017). EWSR1 fusion proteins mediate PAX7 expression in Ewing sarcoma. In *Modern Pathology* (Vol. 30, Issue 9, pp. 1312–1320).

https://doi.org/10.1038/modpathol.2017.49

Charville, G. W., Wang, W.-L., Ingram, D. R., Roy, A., Thomas, D., Patel, R. M., Hornick, J. L., van de Rijn, M., & Lazar, A. J. (2019). PAX7 expression in sarcomas bearing the EWSR1-NFATC2 translocation. In *Modern Pathology* (Vol. 32, Issue 1, pp. 154–156). https://doi.org/10.1038/s41379-018-0095-6

Cohen, S. (2020). *Artificial Intelligence and Deep Learning in Pathology*. Elsevier Health Sciences.

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. In *Nature Medicine* (Vol. 24, Issue 10, pp. 1559–1567). https://doi.org/10.1038/s41591-018-0177-5

El Beaino, M., Rassy, E., Hadid, B., Araujo, D. M., Pavlidis, N., & Lin, P. P. (2020). Synovial Sarcoma: A Complex Disease with Multifaceted Signaling and Epigenetic Landscapes. *Current Oncology Reports*, *22*(12), 124.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29.

Foersch, S., Eckstein, M., Wagner, D.-C., Gach, F., Woerl, A.-C., Geiger, J., Glasner, C., Schelbert, S., Schulz, S., Porubsky, S., Kreft, A., Hartmann, A., Agaimy, A., & Roth, W. (2021). Deep learning for diagnosis and survival prediction in soft tissue sarcoma. In *Annals of Oncology* (Vol. 32, Issue 9, pp. 1178–1187). https://doi.org/10.1016/j.annonc.2021.06.007

Gamboa, A. C., Gronchi, A., & Cardona, K. (2020). Soft-tissue sarcoma in adults: An update on the current state of histiotype-specific management in an era of personalized medicine. In *CA: A Cancer Journal for Clinicians* (Vol. 70, Issue 3, pp. 200–229). https://doi.org/10.3322/caac.21605

Gatta, G., A.P. Dei Tos, J.M. Van Der Zwan, P.G. Casali, S. Siesling, A. Tavilla, et al., Rare cancers are not so rare: the rare cancer burden in Europe, Eur. J. Cancer 47 (2011) 2493–2511

Gazendam, A. M., Popovic, S., Munir, S., Parasu, N., Wilson, D., & Ghert, M. (2021). Synovial Sarcoma: A Clinical Review. In *Current Oncology* (Vol. 28, Issue 3, pp. 1909–1920). https://doi.org/10.3390/curroncol28030177

Gertych, A., Swiderska-Chadaj, Z., Ma, Z., Ing, N., Markiewicz, T., Cierniak, S., Salemi, H., Guzman, S., Walts, A. E., & Knudsen, B. S. (2019). Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific Reports*, *9*(1), 1483.

Grünewald, T. G. P., Cidre-Aranaz, F., Surdez, D., Tomazou, E. M., de Álava, E., Kovar, H., Sorensen, P. H., Delattre, O., & Dirksen, U. (2018). Ewing sarcoma. *Nature Reviews. Disease Primers*, *4*(1), 5.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. In *ACM Computing Surveys* (Vol. 51, Issue 5, pp. 1–42). https://doi.org/10.1145/3236009

Hekler, A., Utikal, J. S., Enk, A. H., Solass, W., Schmitt, M., Klode, J., Schadendorf, D., Sondermann, W., Franklin, C., Bestvater, F., Flaig, M. J., Krahl, D., von Kalle, C., Fröhling, S., & Brinker, T. J. (2019). Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, *118*, 91–96.

Honoré, C., Mir, O., & Adam, J. (2020). Desmoplastic Small Round Cell Tumors. In *Rare Sarcomas* (pp. 69–81). https://doi.org/10.1007/978-3-030-24697-6_4

Hung, Y. P., Fletcher, C. D. M., & Hornick, J. L. (2016). Evaluation of NKX2-2 expression in round cell sarcomas and other tumors with EWSR1 rearrangement: imperfect specificity for Ewing sarcoma. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *29*(4), 370–380.

Italiano, A., Sung, Y. S., Zhang, L., Singer, S., Maki, R. G., Coindre, J.-M., & Antonescu, C. R. (2012). High prevalence of CIC fusion with double-homeobox (DUX4) transcription factors in EWSR1-negative undifferentiated small blue round cell sarcomas. *Genes, Chromosomes & Cancer*, *51*(3), 207–218.

Kather, J. N., Heij, L. R., Grabsch, H. I., Loeffler, C., Echle, A., Muti, H. S., Krause, J., Niehues, J. M., Sommer, K. A. J., Bankhead, P., Kooreman, L. F. S., Schulte, J. J., Cipriani, N. A., Buelow, R. D., Boor, P., Ortiz-Brüchle, N.-N., Hanby, A. M., Speirs, V., Kochanny, S., … Luedde, T. (2020). Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, *1*(8), 789–799.

Laak, J. van der, van der Laak, J., Litjens, G., & Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. In *Nature Medicine* (Vol. 27, Issue 5, pp. 775–784). https://doi.org/10.1038/s41591-021-01343-4

Lee, Y. (2020). Regulation and function of capicua in mammals. *Experimental & Molecular Medicine*, *52*(4), 531–537.

Le Guellec, S., Velasco, V., Pérot, G., Watson, S., Tirode, F., & Coindre, J.-M. (2016). ETV4 is a useful marker for the diagnosis of CIC-rearranged undifferentiated round-cell sarcomas: a study of 127 cases including mimicking lesions. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *29*(12), 1523–1531.

Le Loarer, F., Pissaloux, D., Watson, S., Godfraind, C., Galmiche-Rolland, L., Silva, K., Mayeur, L., Italiano, A., Michot, A., Pierron, G., Vasiljevic, A., Ranchère-Vince, D., Coindre, J. M., & Tirode, F. (2019). Clinicopathologic Features of CIC-NUTM1 Sarcomas, a New Molecular Variant of the Family of CIC-Fused Sarcomas. *The American Journal of Surgical Pathology*, *43*(2), 268–276.

Lessnick, S. L., Paolo Dei Tos, A., Sorensen, P. H. B., Dileo, P., Baker, L. H., Ferrari, S., & Hall, K. S. (2009). Small Round Cell Sarcomas. In *Seminars in Oncology* (Vol. 36, Issue 4, pp. 338–346). https://doi.org/10.1053/j.seminoncol.2009.06.006

Maier, A., Syben, C., Lasser, T., & Riess, C. (2019). A gentle introduction to deep learning in medical image processing. *Zeitschrift Fur Medizinische Physik*, *29*(2), 86–101.

Mesko, N. W., Mesko, J. L., Gaffney, L. M., Halpern, J. L., Schwartz, H. S., & Holt, G. E. (2014). Medical malpractice and sarcoma care--a thirty-three year review of case resolutions, inciting factors, and at risk physician specialties surrounding a rare diagnosis. *Journal of Surgical Oncology*, *110*(8), 919–929.

Nativi, S., & Craglia, M. (n.d.). *European Commission AI Watch initiative: Artificial Intelligence uptake and the European Geosciences Community*. https://doi.org/10.5194/egusphere-egu2020-5691

Nguyen, T. B., Manova, K., Capodieci, P., Lindon, C., Bottega, S., Wang, X.-Y., Refik-Rogers, J., Pines, J., Wolgemuth, D. J., & Koff, A. (2002). Characterization and expression of mammalian cyclin b3, a prepachytene meiotic cyclin. *The Journal of Biological Chemistry*, *277*(44), 41960–41969.

Pierron, G., Tirode, F., Lucchesi, C., Reynaud, S., Ballet, S., Cohen-Gogo, S., Perrin, V., Coindre, J.-M., & Delattre, O. (2012). A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. In *Nature Genetics* (Vol. 44, Issue 4, pp. 461–466). https://doi.org/10.1038/ng.1107

Puls, F., Niblett, A., Marland, G., Gaston, C. L. L., Douis, H., Chas Mangham, D., Sumathi, V. P., & Kindblom, L.-G. (2014). BCOR-CCNB3 (Ewing-like) Sarcoma. In *The*

American Journal of Surgical Pathology (Vol. 38, Issue 10, pp. 1307–1318). https://doi.org/10.1097/pas.0000000000000223

Puls, F., Niblett, A., Marland, G., Gaston, C. L. L., Douis, H., Mangham, D. C., Sumathi, V. P., & Kindblom, L.-G. (2014). BCOR-CCNB3 (Ewing-like) sarcoma: a clinicopathologic analysis of 10 cases, in comparison with conventional Ewing sarcoma. *The American Journal of Surgical Pathology*, *38*(10), 1307–1318.

Richkind, K. E., Romansky, S. G., & Finklestein, J. Z. (1996). t(4;19)(q35;q13.1): a recurrent change in primitive mesenchymal tumors? *Cancer Genetics and Cytogenetics*, *87*(1), 71–74.

Righi, A., Gambarotti, M., & Tos, A. P. D. (2019). Round Cell Sarcomas. In *Soft Tissue Sarcomas* (pp. 315–362). https://doi.org/10.1017/9781316535097.008

Sbaraglia, M., Righi, A., Gambarotti, M., & Dei Tos, A. P. (2020). Ewing sarcoma and Ewing-like tumors. *Virchows Archiv: An International Journal of Pathology*, *476*(1), 109–119.

Specht, K., Zhang, L., Sung, Y.-S., Nucci, M., Dry, S., Vaiyapuri, S., Richter, G. H. S., Fletcher, C. D. M., & Antonescu, C. R. (2016). Novel BCOR-MAML3 and ZC3H7B-BCOR Gene Fusions in Undifferentiated Small Blue Round Cell Sarcomas. *The American Journal of Surgical Pathology*, *40*(4), 433–442.

Tos, A. P. D. (2018). *Soft Tissue Sarcomas*. Cambridge University Press.

Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, *13*(1), 152.

W. H. O. Classification WHO Classification of Tumours Editorial Board, & Who Classification of Tumours Editorial. (2020). *Soft Tissue and Bone Tumours*.