



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

**Deep learning systems and methodologies for Spitz
melanocytic tumour analysis**

November, 2025

ELECTRONIC ENGINEERING DEPARTMENT

Author: Laëtítia Launet

Supervisors: Full Prof. Valery Naranjo Ornedo
Prof. Adrián Colomer Granero

Acknowledgements

Esta tesis no es solo el resultado de años de trabajo y aprendizaje, también es el reflejo de todas las personas que me han acompañado a lo largo del camino. Este trabajo es tan mío como de quienes me rodearon, apoyaron, e inspiraron durante este viaje.

Me gustaría empezar por agradecer a Valery, por creer en mí desde el día uno, por acompañarme con confianza, libertad y apoyo constante, y no solo en esta tesis, sino también a nivel personal. Gracias a Adri también, por ser no solo mi codirector, sino también un referente, por su apoyo, implicación y guía a lo largo de estos años.

Al equipo del Ágora, donde compartimos tardes, risas, pausas necesarias y una amistad que se fue tejiendo sin apenas darnos cuenta: Elena, Ro y Cris; y a todas esas personas que fueron llegando al lab y que acabaron siendo amigas también: Nat, María, Alejandro, Ilán, Juan, Pablo(s), Fer(s), y los más peques del grupo. También a quienes estuvieron antes: Gabri, Julio, Claudio. Por las ideas compartidas, los debates (¡con algunos más que otros!), los tardeos, por el buen rollo, y por demostrar que el trabajo también puede ser un espacio de amistad. Gracias también a Silvia y Sandra, por siempre encontrarle una solución a todo.

I would also like to express my gratitude to the CLARIFY team, all supervisors and the incredible fellow PhD students. This project has been a collective journey enriched by collaboration, shared moments, and memorable trips, all of which will stay with me for a long time. It has been a joy getting to know each and every one of you. Special thanks to Zahra and Andrés for their generosity, support, and unconditional friendship.

I am especially grateful to Professor Zhao for the kind welcome and guidance during my stay at the University of Amsterdam, as well as for the ongoing support for months after I left. And to Doctor Carlos Monteagudo, for welcoming me at the INCLIVA during my first secondment and for having the

patience to teach an engineer about pathological anatomy, from sweat gland cells to basal cell carcinoma. I would have truly enjoyed studying pathology!

Y, sobre todo, un agradecimiento muy especial a la mejor familia que me podía tocar. A Pap's y Mam's, mis referentes, por su apoyo incondicional en todas mis locuras e ideas a veces demasiado optimistas, por aceptar cada una de mis facetas y por ser quienes son. Sois luz en mi vida. A mis compañer@s de vida desde siempre, Wadj y Ak, por estar ahí en cada momento, celebrando las alegrías y sosteniendo en las dificultades, por ser mis ejemplos de vida y mi inspiración constante. Haber sido testigo de vuestra evolución y crecimiento ha sido una fuente inmensa de orgullo y motivación para mí. Y, por supuesto, a Papy y Mamie, cuyo amor, cariño y sabiduría siempre iluminan mi camino y son un pilar fundamental en mi vida (tout comme l'est notre rituel du soir !).

Merci aussi à Tiki, ma meilleure amie et boussole, pour être toujours présente quand j'avais besoin de me confier, de rire, de pleurer ou simplement de décompresser, malgré la distance nous séparant. Quelle chance d'avoir une amie aussi exceptionnelle. Merci aussi à Pauline, Eva, Jade, Alice, Max, ainsi qu'à toute la team 3TLR, pour votre présence et amitié.

Y, por supuesto, a Iñigo. Gracias por estar en cada momento, por tu paciencia infinita e incondicional, por ser mi apoyo constante y mi refugio cuando más lo necesitaba. Por compartir risas (y cosquillas), sueños y logros (¡y trofeos!), y por caminar a mi lado en esta aventura de la vida y las que están por venir. No podría haber pedido un mejor compañero de vida.

Finalmente, cómo no mencionar a la que habitó mis primeros recuerdos en la terreta, la Tía, que siempre se ha quedado en mi vida y en mis pensamientos. Tu recuerdo me acompañó durante este proceso y permanecerá conmigo en las etapas por venir.

Abstract

The advent of artificial intelligence has revolutionised many fields, and digital pathology is no exception. The digitisation of biopsies into whole slide images has enabled the application of deep learning that can assist in diagnosis, support pathologists and ultimately reduce errors. However, despite its great potential, the field still faces significant challenges, both in terms of the data itself and the practical implementation of artificial intelligence in clinical workflows. Addressing these challenges requires overcoming the complexities of high-resolution histopathological images, the constraints posed by limited training data, and the integration of such tools into clinical practice.

These limitations can become even more pronounced in the case of rare diseases, where research is clinically challenging and not always extensive. Rare cancers, such as Spitz melanocytic tumours, pose additional difficulties due to their low incidence and clinical ambiguities, making both research in the field and the application of artificial intelligence more complex.

This thesis addresses part of these challenges by exploring and implementing different deep learning strategies tailored to the complexities of digital pathology, with a particular focus on Spitz tumours. Specifically, this research investigates the application of methodologies such as multiple instance learning, active learning, self-training and federated learning to help mitigate the inherent challenges in digital pathology, while maintaining a practical focus on clinical integration. Positioned at the intersection of artificial intelligence and information systems, this work aims to lay the groundwork for future research on Spitz tumours through artificial intelligence-driven solutions.

El avance de la inteligencia artificial ha revolucionado muchos campos, incluyendo la patología digital. La digitalización de biopsias en *whole slide images* de alta resolución ha permitido aplicar metodologías de aprendizaje profundo para ayudar en el diagnóstico, asistir a los patólogos e incluso reducir errores. Sin embargo, a pesar de su potencial, este campo sigue sujeto a retos importantes, tanto en lo que respecta a los datos como a la integración práctica de la inteligencia artificial en entornos clínicos. Superar estos obstáculos requiere abordar la complejidad de las imágenes histopatológicas de alta resolución, la escasez de datos de entrenamiento y la integración efectiva de estas herramientas en la práctica.

Estas limitaciones pueden verse aún más pronunciadas en el caso de enfermedades menos comunes, donde la investigación clínica es compleja y, a menudo, limitada. Cánceres poco comunes, como los tumores melanocíticos de Spitz, suponen dificultades adicionales debido a su baja incidencia y su difícil interpretación clínica, lo que complica tanto la investigación en este ámbito como la aplicación de métodos de inteligencia artificial.

Esta tesis aborda parte de estos desafíos mediante la exploración e implementación de diferentes estrategias de aprendizaje profundo adaptadas al contexto de la patología digital y sus retos, con un enfoque particular en los tumores de Spitz. En concreto, esta tesis investiga la aplicación de metodologías como el *multiple instance learning*, *active learning*, *self-training* y *federated learning* para mitigar los desafíos inherentes al campo de la patología digital, manteniendo un enfoque práctico hacia la integración clínica. Este trabajo, posicionado en la intersección entre la inteligencia artificial y los sistemas de información, busca sentar las bases para futuras investigaciones sobre los tumores de Spitz basadas en sistemas inteligentes.

L'avanç de la intel·ligència artificial ha revolucionat molts camps, incloent-hi la patologia digital. La digitalització de biòpsies en *whole slide images* d'alta resolució ha permés aplicar metodologies d'aprenentatge profund per a ajudar en el diagnòstic, assistir als patòlegs i fins i tot reduir errors. No obstant això, malgrat el seu potencial, este camp segueix subjecte a reptes importants, tant pel que fa a les dades com a la integració pràctica de la intel·ligència artificial en entorns clínics. Superar estos obstacles requerix abordar la complexitat de les imatges histopatològiques d'alta resolució, l'escassetat de dades d'entrenament i la integració efectiva d'estes ferramentes en la pràctica.

Estes limitacions poden veure's encara més pronunciades en el cas de malalties menys comunes, on la investigació clínica és complexa i, sovint, limitada. Càncers poc comuns, com els tumors melanocítics de Spitz, suposen dificultats addicionals a causa de la seua baixa incidència i la seua difícil interpretació clínica, la qual cosa complica tant la investigació en este àmbit com l'aplicació de mètodes d'intel·ligència artificial.

Esta tesi aborda part d'estos desafiaments mitjançant l'exploració i implementació de diferents estratègies d'aprenentatge profund adaptades al context de la patologia digital i els seus reptes, amb un enfocament particular en els tumors de Spitz. En concret, esta tesi investiga l'aplicació de metodologies com el *multiple instance learning*, *active learning*, *self-training* i *federated learning* per a mitigar els desafiaments inherents al camp de la patologia digital, mantenint un enfocament pràctic cap a la integració clínica. Este treball, posicionat en la intersecció entre la intel·ligència artificial i els sistemes d'informació, busca establir les bases per a futures investigacions sobre els tumors de Spitz basades en sistemes intel·ligents.

Contents

List of Figures	xv
List of Tables	xix
1 Introduction	1
1.1 High hopes, bigger challenges: motivation	3
1.2 The puzzling Spitz tumours: clinical context	6
1.3 Objectives	9
1.4 Software infrastructure for digital pathology: Pixnormous	10
1.5 Framework	11
1.6 Outline	11
2 Dealing with image size and complexity: self-training and multiple instance learning	13
2.1 Introduction	14
2.2 Methods	15
2.2.1 Data preprocessing: GigaiHandler	15
2.2.2 Pathologist-like segmentation	19
2.2.3 Weakly-supervised classification	21
2.3 Experimental setting	22
2.3.1 Dataset	22
2.3.2 Training details	23
2.3.3 Pseudo-annotations	23
2.3.4 Classification	24
2.4 Conclusions	25
3 Limited training data: federated learning	27
3.1 Introduction	28
3.2 Related work	30
3.3 Methods: the Notebook Federator	32
3.3.1 System definition and requirements	32

3.3.2	Challenges and assumptions	33
3.3.3	Architecture	34
3.3.4	Technology considerations	36
3.4	Case study: histological image analysis	37
3.4.1	Use case scenarios	37
3.4.2	Federation	38
3.4.3	Federated implementation	39
3.4.4	Experimental results	42
3.5	Discussion and conclusions	44
3.5.1	Achievements	44
3.5.2	Weaknesses and future work	45
4	Barriers to clinical adoption: democratising active learning for pathologists	47
4.1	Introduction	48
4.2	Related work	49
4.2.1	Digital pathology platforms	49
4.2.2	Active learning in pathology	50
4.3	Methods	51
4.3.1	Software components and architecture	51
4.3.2	Active learning pipeline and specific functionalities	56
4.4	Case study: active segmentation in whole slide images of invasive breast cancer	60
4.4.1	Dataset	61
4.4.2	Implementation details	61
4.4.3	AL user workflow	63
4.5	Experimental setting	64
4.6	Discussion	66
4.7	Conclusions	67
5	Scarce annotations: active learning	69
5.1	Introduction	70
5.2	Related work	71
5.3	Material	72
5.3.1	Dataset	72
5.3.2	Experimental setup	73
5.4	Methods	73
5.4.1	Problem formulation	73
5.4.2	Data preprocessing	74
5.4.3	Active learning framework	75
5.4.4	Real-conditions integration	77

5.5	Results	78
5.5.1	Ablation experiments	78
5.5.2	Interobserver study	80
5.6	Discussion	83
5.7	Conclusions	84
6	What happens next? Artificial intelligence in Spitz tumour research	85
6.1	Introduction	86
6.2	Latest clinical advancements: Spitz or spitzoid?	86
6.3	Towards unmasking Spitz tumours with AI?	88
6.3.1	Clinicopathological tabular data	88
6.3.2	Histopathological analysis	92
6.3.3	Genetic analysis	94
6.4	Roadblocks to AI in Spitz tumour analysis	97
6.4.1	Ambiguity in clinical guidelines	97
6.4.2	Limited prognostic information	97
6.5	Future directions	98
6.5.1	Prognostic data: public databases	99
6.5.2	Foundation models	99
6.5.3	Hybrid approaches	100
6.5.4	Cross-speciality collaborations	101
6.5.5	Retrieval systems	101
6.6	Conclusion	102
7	Final conclusions	103
7.1	Global remarks	105
7.2	Specific remarks	105
7.3	Future work	107
	Merits	108
	Bibliography	113

List of Figures

1.1	Histopathological features of Spitz melanocytic tumours extracted from WSIs. (a-c): Spitz naevus (SN) showing symmetry and circumscribed contours (a), large uniform spindle melanocytic cells without mitotic figures (b) and Kamino bodies (c); (d-e): atypical Spitz tumour/melanocytoma (AST) with significant mitotic activity (e); (f-h): Spitz melanoma (SM) with ulceration (f), necrotic areas (g) and atypical mitotic figures (h).	8
1.2	Scheme of the overall objectives.	9
2.1	GigaiHandler: gigapixel preprocessing module, responsible for saving and handling annotations at any magnification level, generating standardised XML files, and extracting tissue fragments and patches for a given input whole slide image. . . .	15
2.2	Class diagram of the GigaiHandler module, illustrating its interactions with the different components responsible for annotation management and image preprocessing.	16
2.3	XML annotation file structure following the proposed standard. It gathers image information, label, annotation coordinates, and corresponding AI-related information.	18
2.4	Overview of the proposed framework. (a) Segmentation model θ^s training using the annotated set S^a . Note that each block follows a residual configuration. (b) Coarse pseudo-masks predictions on the unannotated set S^u . (c) Classification module θ^c trained with the set S^t containing all the images available and their pseudo-annotations m'_i to extract patches in an on-the-fly manner. . . .	20
2.5	Annotations highlighting the tumorous area within tissues of spitzoid melanocytic lesions. (a) Ground truth annotations made by an expert pathologist; (b) coarse pseudo-annotations made by our segmentation model.	21

3.1	An overview of the architecture of Notebook Federator with NaaVRE [54].	35
3.2	Examples of the type of histological patches used in this case study: (a) tumorous patches extracted from the annotated region made by dermatopathologists; (b) non-tumorous patches.	38
3.3	Step-by-step implementation of the federated pipeline, the starting point being a local institution with their corresponding local data, and the result being the federation.	40
3.4	Demonstration of the Notebook Federator method with component containeriser in a Jupyter Notebook: building reusable FL pipeline blocks, e.g., Client Training container.	41
3.5	Implementation of the proposed solution, Notebook Federator, in a federation composed of institutions located in different countries and relying on different hardware resources.	43
4.1	Overall system infrastructure of Pixnormous, enabling the seamless implementation of active learning and associated functionalities, together with the interactions between the different components. Micro services server: contains all submodule services necessary for visualisation, annotation, and preprocessing of gigapixel data. Model server: enables the embedding and interaction with AI models within the platform. Black arrows: specific actions; green arrows: data storage-related actions; blue arrows: actions on the database.	52
4.2	Pixnormous, along with the main functionalities for visualisation and annotation on the left panel.	53
4.3	Interaction flow between the buttons embedded in the interface and the preprocessing and artificial intelligence models' servers.	54
4.4	Active learning pipeline outlining the system's workflow, from the initial user request for prediction and retraining to the final completion notification.	55
4.5	Model (re)training functionality, integrated within Pixnormous proprietary software by means of buttons and pop-up selections.	58
4.6	AL feedback in Pixnormous. Left: the image name is highlighted in red if it is selected for review; once reviewed, users need to fill a checkbox next to the annotation name. Up right: model training and review query notifications for pathologists.	59
4.7	Visual comparison of the models' predictions on a subset of test images. From left to right: GT from pathologists; model trained with GT only; ST model (GT and pseudo-masks); AL model (first iteration); AST model.	65

5.1	Overview of the approach presented in this study. Blue arrows: sample selection strategy; green arrows: model prediction and self-training process.	74
5.2	Data preparation pipeline for the approach proposed in this paper. Top panel: Pixnormous, used for WSI visualisation and annotation (review) by pathologists; lower panel-(left): extraction of the tissue fragments to segment by getting the coordinates of the different regions with tissue within a slide; lower panel-(right): creation of the two main subsets depending on the annotations made by pathologists.	75
5.3	Weighted Dice score of sample selection strategies for AL over the iterations. The lighter curves (blue, purple, orange) represent different seeds for random selection, while the black curve corresponds to the Shannon entropy-based selection.	78
5.4	Comparison of weighted Dice scores for AL strategies with varied seeds and sample selection strategies, with lighter curves corresponding to the ST approach at each step. (A-B-C): AL with ST for three different random seeds for sample selection; (D): AL with ST using the Shannon entropy for informed sample selection.	79
5.5	Comparison of the annotations made by different pathologists vs. our model for a selection of test tissue fragments. (A) Annotations made by pathologist A; (B) annotations made by pathologist B; (C-E) annotation masks predicted on the third iteration by random selection-based models trained only with expert annotations; (F-G) annotation masks predicted by the Shannon-based active learning model on the third round, trained with expert annotations (E) together with predicted pseudo-masks (F). κ : Kappa score between pathologists; κ_A : Kappa score of the model compared with pathologist A's ground truth annotation; κ_B : Kappa score of the model vs. pathologist B's mask.	82
6.1	Histopathological findings of spitzoid melanocytic tumours. (a-b): BRAF mutated atypical spitzoid tumour (ASoidT/BAMS) showing symmetry (a) and large uniform melanocytic cells (b) (c-e); Spitzoid melanoma (SoidM) with an expansible dermal nodular component (c), and large cells with marked nuclear pleomorphism (d).	87

6.2 Simplified overview of typical artificial intelligence approaches
in Spitz tumour analysis depending on the data type. Top:
approaches leveraging tabular data; centre: WSI-based solutions;
bottom: DNA methylation-based approach. 90

List of Tables

1.1	Main histopathological features for the diagnosis of Spitz nevus (SN), atypical Spitz tumour (AST) and Spitz melanoma (SM). .	8
2.1	Description of the Spitz melanocytic lesions database used for the experiments. Number of WSIs, patients, annotated and unannotated fragments with tumour extracted from the WSIs. .	22
2.2	Classification results on the Spitz melanocytic lesions dataset reached during the test step. The last column corresponds to a model trained with data from the annotated set S^a only, while all other models were trained with the complete set S^t	25
3.1	Details of the histological image data used in this use case. Note that non-tumorous patches were then randomly undersampled to balance the dataset before training each model.	38
3.2	Classification results reached on the test set for patch-level ROI selection on the Spitz melanocytic lesions dataset; I_A , I_B , and I_C : centralised setups at each local institution; Federation: decentralised federated setup, leveraging all institutions' data. .	44
4.1	Segmentation metrics on the test set for the different learning experiments proposed. GT: ground truth; ST: self-training; AL: active learning; AST: active self-training. DSC: Dice similarity coefficient. IOU: intersection over union.	64
6.1	Studies included in the review, applying artificial intelligence algorithms to the analysis of Spitz tumours.	89

Introduction

This chapter introduces the motivation and the objectives pursued in this thesis, as well as the main contributions. It also includes the thesis framework and outline.

Contents

1.1	High hopes, bigger challenges: motivation	3
1.2	The puzzling Spitz tumours: clinical context	6
1.3	Objectives	9
1.4	Software infrastructure for digital pathology: Pixnormous	10
1.5	Framework	11
1.6	Outline	11

1.1 High hopes, bigger challenges: motivation

Once a futuristic concept, artificial intelligence (AI) has now become part of our daily lives, sometimes even without knowing it. From voice assistants on smartphones to autonomous vehicles, AI is reshaping industries and offering innovative solutions across fields. And AI goes a long way, tracing back to the early 1900s until pioneers like John McCarthy coined the term in the 1950s, setting the stage for the field's development.

Although early AI systems relied on hand-coded rules, the emergence of machine learning (ML), seen as a subfield of AI, allowed computers to learn from data rather than relying solely on explicit programming. However, traditional ML still relied heavily on manual feature engineering. Before data could be input into a prediction algorithm, such as a classifier or regressor, it needed to be processed and transformed into a set of relevant features that could effectively represent the underlying patterns in the data.

Deep learning (DL) took ML a step further, emerging as a subfield based on artificial neural networks (ANNs). Inspired by the neural networks in the human brain, these networks consist of multiple layers of interconnected nodes that process information and learn patterns from input data, given an objective task. By enabling computers to automatically extract features from raw data, DL eliminated the need for manual feature engineering. DL sparked another revolution, particularly in areas such as image recognition and natural language processing, with architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), and, more recently, transformers.

The potential of DL extended further as it made its way into the medical field, bringing the promise of aiding in diagnosis, reducing errors and, ultimately, improving patient outcomes. With its ability to analyse complex medical data, DL has been increasingly explored for applications in radiology [1], dermatology [2], and pathology [3].

From hype to impact: the promise of digital pathology

For centuries, pathology has been a meticulous process, with pathologists examining tissue samples under a microscope, one slide at a time. This careful, manual analysis has remained the gold standard for diagnosing cancer, relying on the expertise and experience of trained specialists. However, traditional methods can be time-consuming, prone to subjectivity, and suppose an increasing workload for pathologists with the growing volume of biopsies and the rising incidence of cancer [4].

Digital pathology marked a turning point in diagnostic practice. The digitisation of biopsies into high-resolution whole slide images (WSI) enabled the creation of digital pyramid images that preserve all magnification levels, just like those seen through a microscope. With this, digital pathology opens the door to the application of DL and computer vision methods that can analyse these high-resolution gigapixel-sized images and provide relevant insights to pathologists [5, 6].

In particular, CNN-based architectures have been widely used for WSI analysis, due to their ability to effectively process image data and extract meaningful patterns and features. These models have proven effective across a variety of tasks, from tumour localisation and segmentation [7–9] to diagnostic classification by identifying malignant regions [10], grading tumours [11], and distinguishing between tissue subtypes [12]. Beyond these tasks, DL has also been successfully applied to other stages of the WSI analysis pipeline, including mitosis detection [13], artifact removal [14] and stain normalisation [15], among others.

This growing interest in digital pathology has not gone unnoticed by major industry players either. Companies such as Google, Roche, Aiforia, and Aignostics, among others, have made significant investments in AI-powered pathology tools, aiming to bring automation, scalability, and clinical-grade reliability to histopathological diagnosis. For instance, Google has supported research in computational pathology [16], while Roche has advanced its uPath platform to integrate AI solutions into their digital pathology workflow. Meanwhile, Aiforia and Aignostics are dedicated to developing commercial-grade AI tools designed to support both research and clinical needs in digital pathology [17, 18]. The involvement of these companies underscores the strategic value of computational pathology and highlights its potential for real-world impact beyond academic research.

Barriers to deep learning in digital pathology

Yet, despite the promises it brings, applying DL to digital pathology comes with a myriad of challenges [19, 20], mostly related to the nature of digitised histopathological images and the complexities involved in processing them. The vast size of WSIs makes them computationally demanding, as they can often scale up to several gigabytes. Their gigapixel size also makes the annotation process particularly time-consuming, making it difficult to obtain a large and consistent number of annotations. Furthermore, the sensitive nature of patient data and privacy regulations in the medical field often prevent data sharing, limiting access to diverse datasets across different

institutions. However, training robust DL models is data-hungry and requires large, annotated datasets, with fully-supervised learning often being the most straightforward approach.

These challenges become even more pronounced when addressing rare diseases, where research is limited and clinically challenging. The low incidence of such diseases means fewer patient samples, which further amplifies the difficulty of obtaining the large datasets needed. This is particularly the case for Spitz melanocytic tumours, a rare and diagnostically challenging form of skin cancer. With their uncommon nature and the overlap in histological features with other melanocytic lesions such as melanoma, the limited number of cases makes it hard to build large, high-quality datasets for training robust DL models, slowing down progress in a field where DL could be of great support.

To cope with these challenges in applying DL to digital pathology, most of the recent research for WSI analysis has focused on techniques allowing the training of models with less precise annotations and smaller images. One of the most widely used approaches involves weakly-supervised multiple instance learning (MIL) [21], where larger WSIs are cut down into smaller patches to make them more manageable and enable the processing of their gigapixel size [22]. This approach often relies on slide-level labels rather than pixel-level annotations, reducing the need for exhaustive manual annotation. In parallel, other strategies have been explored to deal with the limitations inherent to this field and its data. Among these are active learning, where most informative samples are selected for annotation, and self-training, which uses the model's own predictions to label unlabelled data and expand the training set. These techniques have shown to help alleviate the dependency on large-scale annotated datasets and, in some cases, even improve model performance.

Bringing artificial intelligence closer to clinical practice

Despite the latest advancements in AI-supported pathology, the application of these technologies in the clinical practice remains relatively underexplored, missing out on the potential benefits that AI could offer in real-world settings. Not only are most approaches developed and tested within optimal AI research environments, but pathologists are also usually cut off from the model development process. This disconnect fails to consider the real-world implications of clinical practice, therefore resulting in a lack of trust and integration of AI systems in clinical workflows [23].

Some platforms like QuPath [24] and SlideFlow [25] have brought AI solutions closer to practical usability. While the former is designed with pathologists in

mind, providing a user-friendly environment for WSI visualisation, the latter is more focused on AI research and the simplification of model building. However, IT infrastructures in most clinical settings typically favour user-friendly, plug-and-play solutions [26], and these platforms still require expertise to operate effectively, creating barriers to seamless integration.

Yet to make AI tools truly useful in digital pathology, it is essential to focus not only on overcoming technical limitations in AI model development but also on ensuring these systems are aligned with the workflows and needs of the clinical practice.

This Ph.D. thesis hence positions itself at the intersection of AI and software methodologies in digital pathology, aiming to address both the technical and practical challenges that hinder the effective deployment of AI in clinical practice. In a context where AI solutions are most needed, this work seeks to bridge the gap between DL models and the clinical systems required for their integration into practice, particularly in the case of rare and challenging tumours like Spitz lesions.

1.2 The puzzling Spitz tumours: clinical context

Despite the medical advancements of the last century, certain tumours continue to pose significant challenges, even for expert pathologists, as is the case for Spitz tumours, uncommon skin melanocytic lesions with histological features resembling those of melanoma. First described in 1948 by Sophie Spitz [27], Spitz tumours were initially referred to as "melanomas of childhood" due to their predominance in children and adolescents despite often having favourable clinical outcomes. Spitz neoplasms were later recognised as Spitz naevi, acknowledging their generally benign clinical outcomes. Although they were long believed to be completely benign [28], cases of nodal metastasis subsequently challenged this view [29]. This is how atypical Spitz tumours (AST), also called Spitz tumours of uncertain malignant potential (STUMP) or Spitz melanocytoma [30, 31] were introduced, initially labelled as malignant Spitz naevi [32] or metastasising Spitz tumours [33]. These neoplasms are still considered to date a diagnostic gray area due to their prognostic uncertainty, positioning themselves as an intermediate class between Spitz naevi and Spitz melanoma. However, since histological appearance remains the gold standard for the tumour diagnosis process worldwide, classifying Spitz neoplasms as benign or malignant is particularly challenging.

Over the last decades, research has sought to better understand these tumours, either by analysing their histological features or investigating their genetic background, aiming to improve the characterisation of Spitz tumours. However, to date, these melanocytic tumours remain one of the most challenging and enigmatic lesions to diagnose accurately in histopathology, even for expert pathologists, mostly because of their conflicting morphological features and unpredictable clinical behaviour [34].

Clinical background

Spitz melanocytic lesions in general are characterised by a distinctive morphology with large epithelioid and/or spindle-shaped melanocytes, as depicted in Fig.1.1. Despite their morphological ambiguity often leading to high interobserver variability, representative histological features have been described to facilitate their categorisation [30, 35], summarised in Table 1.1.

Specifically, Spitz naevi are defined as benign melanocytic lesions with a confluence of melanocytes in an organised manner (see Fig.1.1-(b)), typically symmetrical and sharply circumscribed (Fig.1.1-(a)), with minimal mitotic activity and without atypical or deep mitosis (see Fig.1.1-(b)), with the presence of Kamino bodies further reassuring their favourable prognosis (see Fig.1.1-(c)). They can occur at any cutaneous site. In contrast, Spitz melanomas may lead to aggressive clinical outcomes, similar to conventional melanomas. Their usually asymmetrical silhouette with disorganised nests of melanocytes, extensive pagetoid spread, necrosis, high mitotic rate and the presence of deep and atypical mitoses as depicted in Fig.1.1-(f-h) are commonly associated with malignancy.

In the midst of these peculiar lesions, ASTs display overlapping features between Spitz naevi and Spitz melanoma, as shows Fig.1.1-(d-e). ASTs represent tumours with one or more atypical features, showing histological characteristics that differ from conventional Spitz naevi but are not sufficient for a malignancy diagnosis, therefore making their classification difficult.

Why it matters

The clinical and histopathological complexity of Spitz and spitzoid tumours, along with their diagnostic uncertainty and evolving classification criteria, underscores the need for new tools to support their analysis. These challenges make them a relevant and demanding use case for exploring the application of AI in digital pathology. As rare and diagnostically complex tumours, they exemplify how the core limitations already

present in digital pathology can become amplified in more challenging clinical scenarios, making them an ideal case to test the boundaries of current AI methods. This thesis is motivated by the opportunity to develop and evaluate AI-based approaches tailored to these challenges, using the Spitz tumour spectrum as a representative and high-impact case.

Table 1.1: Main histopathological features for the diagnosis of Spitz nevus (SN), atypical Spitz tumour (AST) and Spitz melanoma (SM).

Feature	SN	AST	SM
Size	Usually <5-6 mm	6-10 mm	Usually >1 cm
Symmetry	Yes	Rare	No
Circumscription	Well demarcated	Well or poorly demarcated	Poorly demarcated
Melanocytic nests	No	Well-defined	Ill-defined
Pagetoid spread	If any, central and focal	Greater than in SN	Extensive
Mitotic rate	<2/mm ²	2-6/mm ²	>6/mm ²
Atypical mitosis	No	No	Yes
Ulceration	No	Rare	May be present
Deep/marginal mitosis	No	May be present	Yes
Necrosis	No	Uncommon	May be present
Maturation	Common	Partial or absent	Uncommon
High-grade nuclear atypia	No	May be present	Common
Epidermal consumption	Extremely uncommon	May be present	Common
Kamino bodies	May be present	Rare	Extremely uncommon
Solar elastosis	Uncommon	May be present	More common than AST

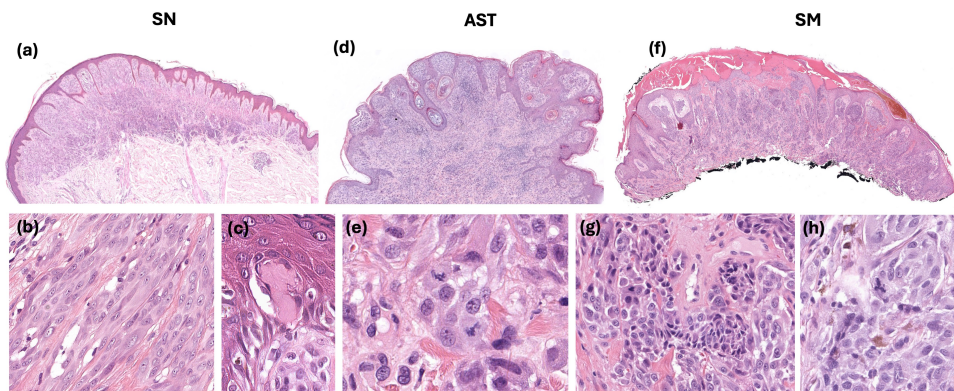


Figure 1.1: Histopathological features of Spitz melanocytic tumours extracted from WSIs. (a-c): Spitz naevus (SN) showing symmetry and circumscribed contours (a), large uniform spindle melanocytic cells without mitotic figures (b) and Kamino bodies (c); (d-e): atypical Spitz tumour/melanocytoma (AST) with significant mitotic activity (e); (f-h): Spitz melanoma (SM) with ulceration (f), necrotic areas (g) and atypical mitotic figures (h).

1.3 Objectives

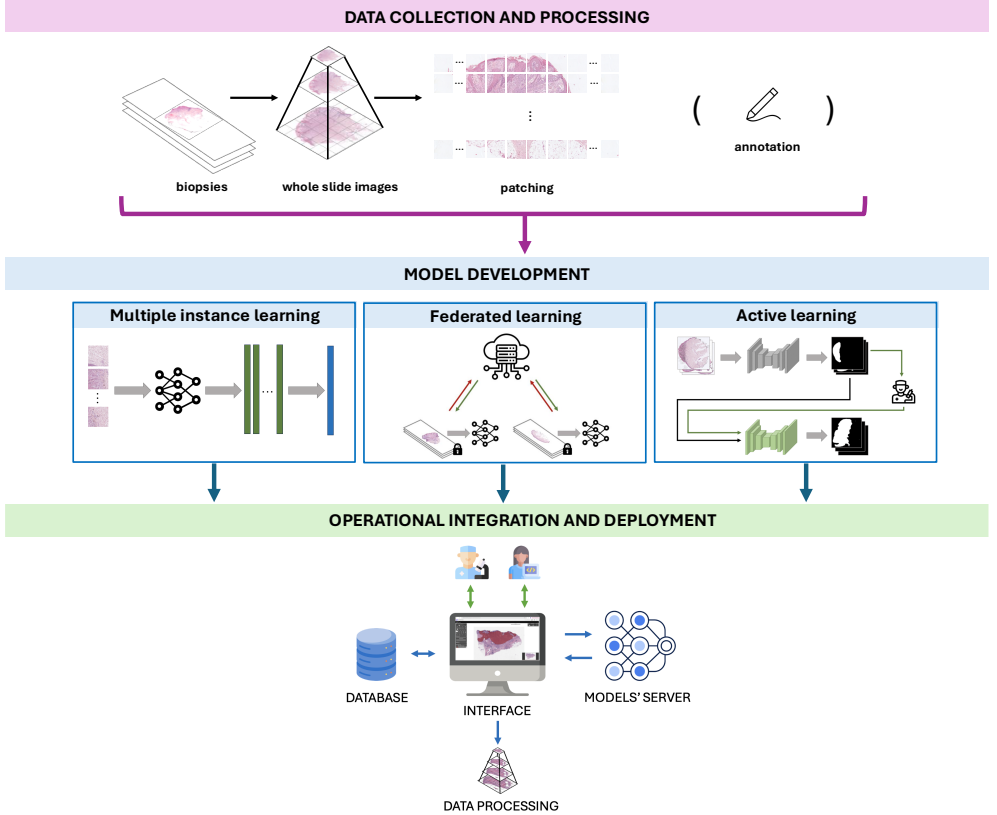


Figure 1.2: Scheme of the overall objectives.

The main objective of this Ph.D. thesis is two-fold, aiming to address both the challenges inherent to DL model development in digital pathology and the systems necessary for their effective integration into clinical practice. Given the complexity of WSI analysis and the particular challenges posed by rare, under-studied tumours, this research operates at the intersection of learning strategies and the systems needed for their integration. Figure 1.2 shows a scheme of the objectives' flow. More specifically, the objectives of this thesis are as follows:

- To collect, process, and condition data from a particularly challenging and under-studied tumour type, ensuring its suitability

for AI-based analysis. Given the gigapixel nature of WSIs and their inherent complexity, this objective’s scope includes handling this data type in a reproducible and optimised manner, under learning paradigms adapted to the data at hand.

- To design and develop AI models under different learning strategies that address the specific challenges posed by DL in digital pathology. Motivated by the inherent limitations of DL in this field, this objective involves the exploration of various learning strategies that include multiple instance learning, federated learning, active learning, and self-training to mitigate issues related to data scarcity and annotation constraints.
- To implement systems that facilitate the integration of DL models into clinical workflows. Given that model development alone is not sufficient for real-world implications, this objective focuses on considering the necessary infrastructure and systems to bridge the gap between AI research and clinical deployment.

1.4 Software infrastructure for digital pathology: Pixnormous

A central pillar of this thesis is Pixnormous, a whole slide image (WSI) visualisation and processing platform designed to handle the gigapixel-sized images typical of digital pathology. Initially developed within the CVB Lab group to support research workflows, Pixnormous gradually evolved into a production-grade tool that supports every stage of this work, from data management and annotation to active learning integration and AI model deployment.

The relationship between this thesis and Pixnormous is reciprocal: a big part of the research relies on the platform’s capabilities, while the work presented here has actively contributed to the ongoing development and refinement of the platform.

As Pixnormous matured beyond a research-only tool, it was transitioned to the commercial sector through the spin-off company ArtiKode

Intelligence, SL. Thus, Pixnornious not only serves as foundational software infrastructure for this thesis but also exemplifies how academic research can translate into practical clinical tools.

1.5 Framework

This thesis was conducted within the framework of the CLARIFY (cloud artificial intelligence for pathology) project, a European initiative aimed at developing AI-driven tools and cloud-based research infrastructure to enhance the interpretation and diagnosis of complex pathological cases using histopathological images. These include triple negative breast cancer, high-risk non-muscle invasive bladder cancer and Spitz melanocytic tumours. Among these, this thesis specifically focuses on Spitz tumours. The CLARIFY project was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 860627.

1.6 Outline

This thesis is divided into six chapters, each addressing different aspects of DL limitations in digital pathology, with a specific focus on Spitz melanocytic tumours.

Chapter 2 focuses on the nature of the analysed images, particularly dealing with the challenges posed by WSIs in digital pathology. Then, Chapter 3 addresses the issue of scarce data, exploring the use of federated learning as a potential solution to this problem, keeping in mind the real-world application of this type of solutions. Chapter 4 focuses on the development of systems that bring DL closer to clinical practice with a focus on active learning methodologies, while Chapter 5 delves deeper into active learning strategies to address the challenge of scarce annotations in a real-world infrastructure. Chapter 6 summarises the current state of research of AI for Spitz tumour analysis as observed upon finalising this thesis, along with the challenges these lesions continue to pose in digital pathology, and outlines future guidelines for improving their interpretation with DL.

To conclude on the work carried out throughout this thesis, Chapter 7 provides final remarks, reflecting on the overall contributions and discussing its broader implications in digital pathology, particularly in the context of Spitz tumours. Finally, the [Merits](#) Section presents the academic achievements resulting from this thesis, including journal publications, presentations at national and international conferences, and related awards.

Dealing with image size and complexity: self-training and multiple instance learning

One of the first challenges when dealing with whole slide images in computational pathology is their enormous size and complexity, which can reach several gigapixels per slide. This chapter addresses this issue by introducing a preprocessing module and a multiple instance learning approach to enable efficient and effective downstream analysis.

Part of this chapter contains work previously published, although adapted and widely complemented for this thesis:

Launet, L., Colomer, A., Mosquera-Zamudio, A., Moscardó, A., Monteagudo, C., & Naranjo, V. A self-training weakly-supervised framework for pathologist-like histopathological image analysis. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 3401-3405). IEEE.

Contents

2.1	Introduction	14
2.2	Methods	15
2.2.1	Data preprocessing: GigaiHandler	15
2.2.2	Pathologist-like segmentation	19
2.2.3	Weakly-supervised classification	21
2.3	Experimental setting	22
2.3.1	Dataset	22
2.3.2	Training details	23
2.3.3	Pseudo-annotations	23
2.3.4	Classification	24
2.4	Conclusions	25

2.1 Introduction

One of the initial challenges in applying DL to histopathological images lies in the nature of WSIs themselves. These gigapixel-scale images can contain millions of pixels, making direct processing by DL models computationally infeasible. To overcome this computational limitation, a widely adopted solution is to divide WSIs into smaller, more manageable patches or tiles [6].

Multiple instance learning (MIL), a weakly-supervised learning paradigm, is generally employed to facilitate this process. MIL enables feature extraction at the patch level before aggregating them to obtain slide-level predictions, treating WSIs as bags and patches as the instances within bags. Unlike traditional supervised learning, which would require annotations for each individual patch, MIL only requires labels at the bag level, thus reducing the need for detailed and time-consuming labelling.

In practice, not all patches within a bag are equally informative, as many patches may contain irrelevant tissue to the diagnosis. Limited by GPU memory size, the number of patches that can be processed at once is often constrained for particularly large WSIs, restricting the model's ability to fully capture the diverse information across the entire WSI. A preliminary tumour detection step is commonly implemented to identify regions of interest (ROIs) [36, 37], thereby reducing the number of patches that need to be processed.

However, existing approaches often perform tumour localisation at the patch level [38], requiring all patches to be individually processed before selecting relevant ones. This per-patch processing remains computationally expensive, especially for large WSIs. In contrast, we propose a different perspective for WSI analysis by mimicking the way pathologists examine biopsies under the microscope. Specifically, we introduce a two-stage pipeline that first applies a rough segmentation to identify the region of interest (ROI) from a bird eye's view at a lower magnification. In the second stage, we extract detailed information from these identified regions by focusing on smaller patches at higher magnification, similar to the focused analysis pathologists perform once they locate regions of interest. To enable this approach and subsequent

developments, we also introduce a generalisable and adaptable data preprocessing module for robust WSI handling. This module manages every step of preprocessing, from tissue fragment extraction to patch extraction with customisable parameters, enabling efficient and flexible adaptation to different histopathology datasets and workflows.

Specifically, our main contributions in this work include:

- A generalisable and customisable data preprocessing module that supports diverse histopathology workflows;
- A novel two-stage WSI analysis pipeline that mimics the diagnostic workflow in pathology, integrating self-training for tumour localisation with MIL for tumour classification.

2.2 Methods

2.2.1 Data preprocessing: GigaiHandler

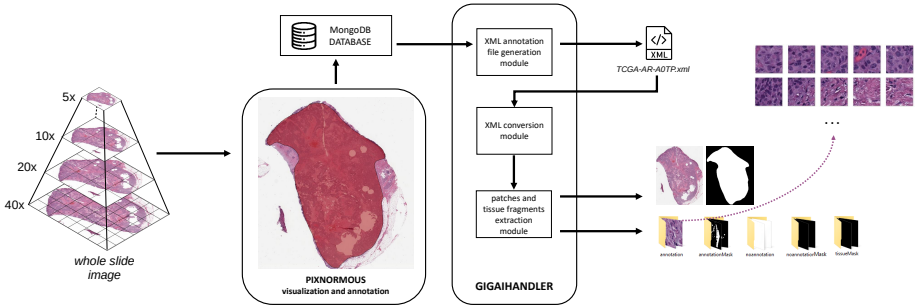


Figure 2.1: GigaiHandler: gigapixel preprocessing module, responsible for saving and handling annotations at any magnification level, generating standardised XML files, and extracting tissue fragments and patches for a given input whole slide image.

Preprocessing WSIs is a critical step in digital pathology pipelines for DL applications, transforming these gigapixel-scale images into a manageable form that can be effectively input to DL models. In this sense, the GigaiHandler module plays a central role in our pipeline by managing and optimising every step of the preprocessing process, from tissue

fragment extraction to patch generation and annotation management. It ensures that input data is properly handled and formatted for both image analysis tasks and model training. GigaiHandler is tightly integrated with Pixnorious (see Figure 2.1).

To efficiently manage the preprocessing workflow, the GigaiHandler module is divided into three different components, each responsible for a specific aspect of the preprocessing pipeline. Specifically, two components are dedicated to annotation management, while the third focuses on proper image processing. Figure 2.2 shows the corresponding class diagram with the relationships between the different Python classes and the functions they encapsulate. All processes are coordinated through a central *launcher*, while the *App* class orchestrates the entire data preprocessing pipeline by handling the appropriate classes and managing task execution. Meanwhile, the *SlideAnnotationToXMLTranslator* class is responsible for handling interactions with the database whenever an annotation is created, ensuring that annotation data is correctly stored and converted into the standardised XML format.

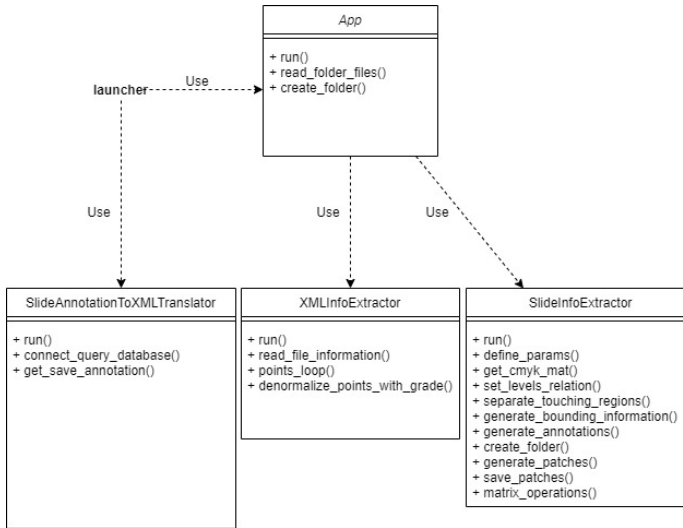


Figure 2.2: Class diagram of the GigaiHandler module, illustrating its interactions with the different components responsible for annotation management and image preprocessing.

- **Annotation management system**

To enable the efficient handling of WSI annotations, we implemented a dedicated annotation management system built around a structured XML standard for annotation files. This XML format serves as the central element for organising and maintaining annotation data while ensuring compatibility across different workflows. It allows to store slide information externally from the database, providing a transferable and easily convertible backup solution.

Whenever a user creates and saves an annotation, the XML generation module automatically retrieves the relevant slide annotations from the database and generates a corresponding normalised XML file for the WSI through the *SlideAnnotationToXMLTranslator* class (see Figure 2.2). As shows Figure 2.1, this file then serves as input for generating annotation masks associated with tissues and patches and is fed into the *XMLInfoExtractor* class, facilitating subsequent processing and conversion to image mask.

This XML standard is designed to store key information such as slide metadata, annotation coordinates, assigned labels, and AI-related parameters, ensuring structured and comprehensive annotation management. The detailed structure of this format is depicted in Figure 2.3 for an example slide.

- **WSI preprocessing**

The other facet of the preprocessing pipeline involves patch and tissue fragment extraction from WSIs. This process is handled by the *SlideInfoExtractor* class (see Figure 2.2), which manages all operations related to WSI data.

Specifically, input WSIs are first converted to CMYK images, as this colour space has shown the most promising results for operations on H&E-stained WSIs. The magenta channel is then extracted at a fixed magnification level of $5\times$, selecting a lower resolution to enable lighter preprocessing operations that will later be converted to the required magnification for subsequent analysis. A 15×15 kernel is then applied to slightly blur the image before using Otsu’s thresholding method to

```
<?xml version="1.0" encoding="UTF-8" ?>
<Images>
  <Image>
    <Name>TCGA-A1-A05M</Name>
    <Height>80501</Height>
    <Width>100794</Width>
    <GlobalLabel>0</GlobalLabel>
    <Observations>0</Observations>
    <Regions>
      <Region1>
        <Grade>IBCA</Grade>
        <x>0.5531 0.55279 0.55279 0.55279 0.55247 0.55216 0.55184
0.55153 ...</x>
        <y>0.26836 0.2688 0.26899 0.26931 0.26962 0.26962 0.26962
0.26962 ...</y>
        <AI>False</AI>
        <modified>True</modified>
        <reviewed>False</reviewed>
        <label_confidence>1</label_confidence>
      </Region1>
    </Regions>
  </Image>
</Images>
```

Figure 2.3: XML annotation file structure following the proposed standard. It gathers image information, label, annotation coordinates, and corresponding AI-related information.

separate tissue from background areas, obtaining a binary image where tissue pixels are assigned a value of 1 and background pixels a value of 0. This binary mask serves as the basis for extracting tissue fragments by identifying the contour coordinates or bounding boxes of each tissue block.

Once tissue fragments are identified, patches are extracted at the required magnification level, ensuring correspondence between the initial $5\times$ matrix generated in the initial preprocessing step and the target magnification. The extraction process is configurable through several parameters, including the desired magnification level, patch size, and the percentage of overlap between patches.

When annotations are available for a slide, a corresponding annotation mask is generated after converting the annotations using the *XMLInfoExtractor* class, with greyscale values assigned to the mask based on the annotation class. Patches are then extracted from both the original tissue fragment and the associated annotation mask at the same coordinates, maintaining spatial alignment between the image matrix and its annotation mask.

2.2.2 Pathologist-like segmentation

Building upon the GigaiHandler preprocessing module to prepare the data necessary for subsequent analysis, we propose a segmentation model θ^s that generates coarse annotations of ROIs. This approach narrows down the areas for subsequent MIL-based classification. Inspired by how pathologists get a first overall picture of a tissue slide, we do not split WSIs yet into smaller patches at this stage. Instead, we leverage the tissue fragments generated by the GigaiHandler module and utilise a low-level magnification, i.e., $5\times$ magnification.

Notation. In this work, we consider the input dataset $D = \{S^a U S^u\}$ composed of both a set of annotated images $S^a = \{x_i, y_i, m_i\}$ and a set of unannotated images $S^u = \{x_i, y_i\}$. x_i represents the RGB image, y_i denotes its label, and m_i designates the corresponding annotation mask, when available. An image corresponds to an extracted tissue fragment from a WSI, therefore a single WSI might sometimes result in multiple tissue fragments. Note that the term *label* is used to refer to slide-level diagnosis (i.e., malignant or benign), while *annotation* corresponds to the specific outline of a ROI made by pathologists to highlight a tumorous area within an image. In this work, although annotations are not always available, each WSI does come with a corresponding label.

Input data. We resize each of the previously extracted tissue fragments and their corresponding ground truth mask to 512×512 squares. This size showed to contain enough details to generate coarse segmentation masks, as demonstrated in Section 2.3.3. We then apply basic augmentation operations to the training data samples to generate a more substantial annotated dataset at this step, i.e., random rotations, shifts, flips, and zoom.

Ground truth images. Expert dermatopathologists carried out most of the ground truth annotations in a freehand manner, thus sometimes including non-tumorous areas within highlighted ROIs. Examples of expert annotations are depicted in Figure 2.5-(a).

Segmentation model architecture. We adopt a U-Net backbone architecture [39] with residual blocks [40], following the same structure

as in [41]. While the encoder part of the network focuses on feature extraction, the decoder allows to reconstruct the images with the probability segmentation maps. Under the residual configuration, skip-additive connections are added to help optimise the model. To train our coarse segmentation network, we use 3 convolutional blocks, each doubling the number of filters from 64 up to 512.

Pseudo-masks generation. As depicted in Figure 2.4-(b), once the segmentation model θ^s trained with the annotated data samples S^a , we use it to predict the annotations of unannotated samples S^u , thus generating pseudo-annotations m'_i that will be used to narrow down ROIs in the classification module.

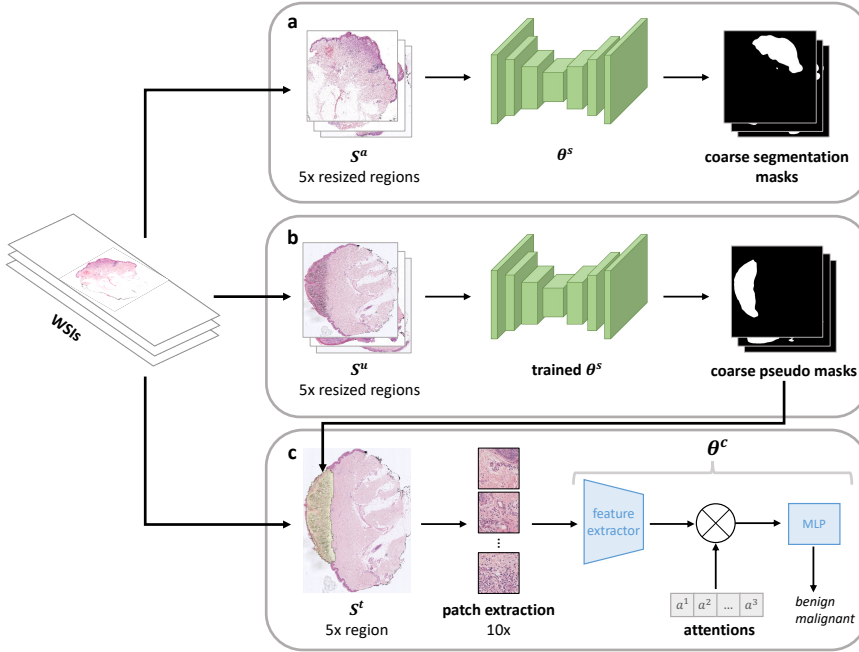


Figure 2.4: Overview of the proposed framework. (a) Segmentation model θ^s training using the annotated set S^a . Note that each block follows a residual configuration. (b) Coarse pseudo-masks predictions on the unannotated set S^u . (c) Classification module θ^c trained with the set S^t containing all the images available and their pseudo-annotations m'_i to extract patches in an on-the-fly manner.

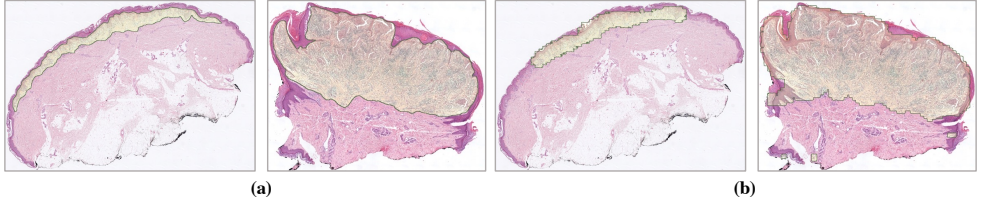


Figure 2.5: Annotations highlighting the tumorous area within tissues of spitzoid melanocytic lesions. (a) Ground truth annotations made by an expert pathologist; (b) coarse pseudo-annotations made by our segmentation model.

2.2.3 Weakly-supervised classification

Once the ROIs are identified, tumorous areas are analysed at a higher magnification level for slide-level classification, as depicted in Figure 2.4-(c).

Input data. Given the resulting set of images $S^t = \{x_i, y_i, m'_i\}$ composed of all the available images x_i along with their corresponding labels y_i and either pathologists' annotations m_i or pseudo-masks m'_i generated with the segmentation model θ^s , we re-adapt the 512×512 pseudo-masks coordinates to a $10\times$ magnification level. During this process, we also apply a dilation operation to the up-sized pseudo-masks to ensure all tumorous areas are included.

On-the-fly patch extraction. WSIs are then cropped during the training process with a sliding window scheme over the pseudo-annotation demarcation at a $10\times$ magnification level and no overlap. The patches where pseudo-annotation accounts for less than 30% are dismissed. Note that the memory cache is emptied at the end of each batch during the training process, one batch corresponding to an entire image.

Classifier θ^c . We train a feature extractor on the extracted patches by fine-tuning different architectures with both light-weight and heavier architectures: ResNet-34, ResNet-50 [40] and VGG16 [42]. Due to the difference in patterns between histological images and the ImageNet dataset, we re-train most of the blocks except for those associated with low-level features (contours, basic shapes) and add two convolutional

layers to each architecture. Following Ilse et al. [43] trainable attention weights, we then use the attention-based MIL pooling to aggregate the obtained patches' features before giving the final diagnosis with a classification layer.

2.3 Experimental setting

2.3.1 Dataset

In this work, we trained the models with a Spitz melanocytic lesions WSI database digitised, labelled, and annotated by pathologists of the Pathology Laboratory from the University Clinic Hospital of Valencia. The Spitz dataset used, described in Table 2.1, is relatively scarce due to the rare occurrence of these lesions among the population [44] and contains 84 biopsies from 79 different patients. Although the ROI was highlighted by pathologists in part of the WSIs, at the time of the development of this framework, a substantial part of the dataset lacked a precise annotation.

Table 2.1: Description of the Spitz melanocytic lesions database used for the experiments. Number of WSIs, patients, annotated and unannotated fragments with tumour extracted from the WSIs.

	Malignant	Benign
# WSIs	34	50
# patients	30	49
# annotated fragments	16	11
# unannotated fragments	29	41

2.3.2 Training details

All models, irrespective of the module, were trained on the NVIDIA DGX A100 system and implemented with the PyTorch 1.9.0 library under Python 3.8.

2.3.3 Pseudo-annotations

We evaluated the coarse segmentation model with the Dice score obtained in the predicted segmentation maps, using the resized masks of pathologists' annotations as ground truth. We trained the U-Net segmentation model using residual connections, ResU-Net [41] for 40 epochs with a Nadam optimiser [45], on resized images of 512×512 and 256×256 using either 3 or 4 blocks. With 256×256 input images to train the model, we achieved an average Dice of 0.9071 and 0.8665 on the test set using 3 and 4 blocks, respectively. With 512×512 input images, we reached a Dice score of 0.9171 using 3 blocks and 0.9204 with 4 blocks.

Figure 2.5 shows an original-size comparison of the pseudo-annotations predicted by our selected model with those made by expert pathologists. As depicted in this figure, the generated pseudo-annotations are coarse and may highlight a bigger ROI than those delineated by pathologists or sometimes smaller ones, neglecting thinner areas. However, the classification results presented in the next sub-section 2.3.4 demonstrate that a coarse segmentation is sufficient.

To further analyse the correlation of the extracted features for each class at this stage, i.e., malignant vs. benign, we clustered the input images using the t-distributed stochastic neighbour embedding (t-SNE) technique [46], applied to the extracted features of the last encoder block. However, the t-SNE representation of the data revealed that the features of the two classes under study were not separated, which is why the encoder trained in this step and the extracted features were not used in the classification module.

2.3.4 Classification

For the slide-level classification training stage, we employed Adam optimisation with a learning rate set to 0.0001 on all the baselines. To evaluate the performances of the models, we selected relevant metrics, including false negative rate (FNR), sensitivity (SN), specificity (SPC), positive predictive value (PPV), negative predictive value (NPV), F1-score (F1S), and accuracy (ACC) predicted on data never seen by the models before, on completion of their training. Note that images corresponding to the same WSI or patient were used in the same set.

We used the pseudo-masks predicted by the 4-block segmentation model to extract patches in an on-the-fly manner, resulting in a total of 2851 patches from the 97 tissue fragments of both annotated and unannotated images. We trained the models for 30 epochs each on baselines that showed promising performance in this field, i.e., ResNet and VGG architectures. Details of the performance results are shown in Table 2.2, and the backbone with the best performance metrics was selected as the baseline for the classification module, that is to say, ResNet-34. Note that, when training heavier architectures such as ResNet-50, some patches had to be randomly discarded at the classification step due to the size of this architecture, thus partly explaining the better performance of ResNet-34.

To demonstrate the efficiency of the proposed framework, we compared the results of the best network, i.e., ResNet-34, to a model trained on the same architecture with the patches extracted directly from the annotations made by pathologists, without using pseudo-masks nor unannotated data samples. As depicted in Table 2.2, models trained under a semi-supervision scenario clearly outperform the one trained with annotated samples only (S^a). Note that the proposed model also reaches a better classification accuracy than that of previous works studying the same type of lesions [38].

Table 2.2: Classification results on the Spitz melanocytic lesions dataset reached during the test step. The last column corresponds to a model trained with data from the annotated set S^a only, while all other models were trained with the complete set S^t .

	VGG16	ResNet-34	ResNet-50	ResNet-34 (S^a)
FNR	0.2222	0.1111	0.2000	0.3750
SN	0.7778	0.8889	0.8000	0.6250
SPC	0.6923	0.8182	0.7500	0.8333
PPV	0.6364	0.8000	0.7273	0.9091
NPV	0.8182	0.9000	0.8182	0.4545
F1S	0.7000	0.8421	0.7619	0.7407
ACC	0.7273	0.8500	0.7727	0.6818

2.4 Conclusions

In this work, we introduced a novel self-training and weakly-supervised learning-based diagnosis tool for whole slide image analysis inspired by how pathologists analyse such images in the daily sign-out of biopsies. The proposed method enables to train models with a limited number of expert annotations, thus fully leveraging the available data by making use of both annotated and unannotated data samples. Despite using a coarse segmentation of the tumorous area to extract patches, we managed to reach a 0.85 classification accuracy on the test set, clearly outperforming models trained on annotated data only, thus demonstrating the efficiency of our approach. Importantly, this framework is supported by our generalisable data preprocessing module GigaiHandler, which ensures consistent, high-quality preprocessing of the WSIs and supports all subsequent analysis steps.

Limited training data: federated learning

While Chapter 2 addresses the challenges of managing the size and complexity of gigapixel whole-slide images in computational pathology, this chapter focuses aims to help deal with the scarcity of training data by exploring federated learning as a means to leverage decentralised medical datasets while preserving patient privacy.

The content of this chapter corresponds to a work previously published, although adapted for this thesis:

Launet, L., Wang, Y., Colomer, A., Igual, J., Pulgarín-Ospina, C., Koulouzis, S., Bianchi, R., Mosquera-Zamudio, A., Monteagudo, C., Naranjo, V., & Zhao, Z. Federating Medical Deep Learning Models from Private Jupyter Notebooks to Distributed Institutions. Applied Sciences, 13(2), 919 (2023).

Contents

3.1	Introduction	28
3.2	Related work	30
3.3	Methods: the Notebook Federator	32
3.3.1	System definition and requirements	32
3.3.2	Challenges and assumptions	33
3.3.3	Architecture	34
3.3.4	Technology considerations	36
3.4	Case study: histological image analysis	37
3.4.1	Use case scenarios	37
3.4.2	Federation	38
3.4.3	Federated implementation	39
3.4.4	Experimental results	42
3.5	Discussion and conclusions	44
3.5.1	Achievements	44
3.5.2	Weaknesses and future work	45

3.1 Introduction

The challenges of applying AI to digital pathology extend beyond gigapixel image sizes and computational constraints. Indeed, developing optimal DL models relies heavily on large, diverse training datasets, which are often difficult to obtain, particularly in the medical field. In many cases, local datasets are too limited to train robust models, leading to performance issues, particularly when dealing with rare and under-represented diseases.

To address data scarcity, AI experts have explored techniques such as synthetic data generation [47] and data augmentation methods [48]. However, synthetic data still raises various concerns about reliability and clinical applicability of models [47]. Meanwhile, vast amounts of medical data, critical for training optimal DL models, do exist across different institutions but remain largely inaccessible due to privacy regulations.

If given access to these diverse medical institutions' local data, AI experts would be able to train optimal models with substantial datasets. However, medical data is decentralised by nature due to the privacy legislation it entails, and therefore usually does not allow for its integration across institutions without raising legal concerns. While anonymisation and pseudonymisation were once considered viable solutions to overcome these limitations and thus eventually allow safe transfers of medical data, studies have shown that they do not fully eliminate the risk of patient re-identification [49].

As a way to guarantee data privacy concerns while gaining access to more substantial datasets, federated learning (FL) [50, 51] was introduced to collaboratively train DL models over decentralised data and already proved to perform similarly to models exposed to single-institutional data during training [52]. However, despite its potential, FL remains in its early stages within the medical field. Global collaboration and data sharing among organisations in the federation settings are important requests but still open issues [53]. Indeed, most of these studies were often limited to simulations because of the complex deployment of FL settings in real-world medical applications.

Meanwhile, although data scientists frequently share prototype code through notebook environments like Jupyter, implementing FL systems in real-world settings requires significant engineering and deployment efforts. The process involves model training at each institution, model aggregation, and establishing communication protocols, all of which demand substantial customisation, infrastructure setup, and debugging. This complexity makes it particularly difficult for domain-specific users, such as hospital pathologists, to implement AI at the edge. As a result, the gap between prototype development in Jupyter and fully operational FL systems remains a major obstacle to real-world adoption.

Prior work [54] introduced how code fragments in Jupyter notebooks could be containerised and reused as workflow building blocks to scale up scientific experiments to cloud infrastructures through an example of processing massive light detection and ranging (LiDAR) data. However, it did not support any AI-based code implementations. In this work, we extend this previous work and present a reproducible step-by-step guideline that enables the creation of a federated setup starting from a centralised Jupyter code for distributed medical applications. Specifically, our main contributions include:

- Reduce the complexity of implementing an FL framework by providing a Jupyter notebook code template tailored for artificial intelligence experts;
- Improve the deployment efficiency of cross-silo FL systems through automatic containerisation of Jupyter components for distribution to multiple users;
- Seamlessly bridge the gap between Jupyter notebooks and FL systems by offering a generic, easy-to-use, and reproducible solution;
- Demonstrate the proposed solution, the Notebook Federator, with a practical example tutorial applied to a real-world cross-country computational pathology classification task.

3.2 Related work

Although FL was initially developed for mobile and edge device use cases under a cross-device paradigm [55], cross-silo FL recently gained traction in various domains, including healthcare [56]. Under this setting, models are trained on siloed data across organisations, and medical data remains local and decentralised in each collaborating institution (client) [55]. Only the resulting weights and parameters are shared and aggregated together before being transferred back to the different clients.

In recent years, these applications have become increasingly popular in the medical image analysis field, with successful implementations in areas such as brain imaging classification and segmentation [52, 57, 58], EEG signal classification [59], and, more recently, histopathology [60, 61]. Li et al. [52] demonstrated the feasibility of using a client-server FL system to perform brain tumour segmentation. Roy et al. [57] introduced BrainTorrent, which is a peer-to-peer serverless FL environment, to perform a whole-brain segmentation. Lu et al. [61] presented a large-scale computational pathology study to demonstrate the feasibility and effectiveness of privacy-preserving FL using thousands of gigapixel whole slide images from multiple institutions. However, due to the complexity of FL implementation in real-world situations and large-scale deployment, most of these medical applications of FL still rely on simulations.

To support the different FL scenarios and simplify their deployment and execution process, some studies proposed easy-to-use open-source FL frameworks, such as PySyft [62], OpenFL [63], FLOWER [64] or Federated Scope [65]. The majority of these solutions focuses on the architecture of the FL framework. Whereas FLOWER [64] provides a core framework architecture with Edge Client Engine and Virtual Client Engine, Federated Scope [65] proposes an event-driven architecture with an asynchronous training protocol in FL for heterogeneity. These open-source frameworks have served as a basis to build federated setups, some of them in the medical field. In [66], Lee et al. leveraged the PySyft framework to implement a thyroid ultrasound image classifier across six medical institutions and compared the performance of federated

training with that of conventional centralised model training. In the same vein, Florescu et al. [67] made use of FLOWER to simulate an FL system for COVID-19 detection on CT images, where the clients were deployed locally on a single machine. In contrast, in a work on OCT imaging [68], although the authors did mention the existence of FL frameworks, they chose to simplify their implementation through an independent development and to simulate the clients' nodes locally on a single supercomputer cluster.

Despite these available FL frameworks, there is no explicit pipeline for the FL training process between the framework and data scientists in these solutions, and most steps have to be done manually without FL automation. It is thus still time-consuming for domain-specific users to deploy their FL applications, not only because of the heterogeneous resources but also because of the complexity of models, such as software dependencies and parameters to deploy.

Additionally, although many AI researchers make use of Jupyter notebooks to develop DL models, there is a clear gap between the Jupyter environment and FL systems. Although some specific frameworks do provide a version of the FL code presented in a Jupyter notebook [65], there is still a need to upload the data to the cloud, and these notebooks are not fully integrated into the FL pipeline. In fact, they are mostly based on simulations rather than real-world scenarios running on different clients' machines.

Despite the many system construction efforts for FL frameworks, there is a clear need to simplify their deployment process and address the gap between Jupyter users and FL systems. Far from aiming to replace the available FL frameworks, in this work, we seek to fill the gap they are not covering by proposing a solution that can be built upon them. Therefore, we propose a novel FL research asset that leverages the Jupyter environment and proposes to reduce the complexity of implementing and deploying a reproducible FL pipeline, the Notebook Federator.

3.3 Methods: the Notebook Federator

The Notebook Federator comes as a handy tool that combines resources rather than aiming to replace them. Therefore, we leverage different previously developed solutions to provide a generic solution that can adapt to the particular needs of the end users.

One of the groups participating in this study previously developed a tool that embeds a virtual research environment (VRE) into the Jupyter environment, Notebook-as-a-VRE (NaaVRE) (<https://github.com/QCDIS/NaaVRE>, accessed on 4 January 2023) [54], which comes with a toolbox aiming to provide researchers with a variety of functional components (services). In this work, we make use of NaaVRE and its component containeriser service to Dockerise the components to share with distributed users automatically.

To implement the proposed solution, we selected a commonly used FL framework as a basis for the Notebook Federator, the FLOWER framework [64]. We chose this particular framework for its broad community and documentation and the possibility to easily modify and integrate it for our specific pipeline. However, as a technology, the Notebook Federator is generic, and could thus be used together with other FL frameworks.

3.3.1 System definition and requirements

For the efficient implementation of a framework that promotes the optimal collaboration of AI scientists, the setup proposed in this work is based on a user-friendly environment, Jupyter Notebook. However, Jupyter relies on the kernel (engine) for interpreting the code and is often limited to the capacity of the machine it is running on, thus making it difficult to support FL. To improve this, we identify a set of requirements. First, users registered within a federation should be able to develop the model architecture and code blocks that will be used to train all local models, and the central node should then automatically be able to update the global model's weights upon completing a communication

round. Finally, the proposed pipeline needs to be reproducible and easily implementable for other federations.

3.3.2 Challenges and assumptions

Implementing an FL ecosystem in a real-world medical environment comes with a set of challenges related to its practical application. First and foremost, not all medical institutions can count on the necessary hardware resources to train AI models. When they do, these might vary across sites, as do operating systems and network conditions. Moreover, different collaborators might store their raw data in varying formats and structures, thus limiting the automation of the data-loading process. Therefore, to enable the correct practical implementation of the FL setup, we proceed on the basis that the following assumptions are valid:

- All the participating institutions are already registered as part of the federation;
- Local data cannot be shipped outside the institution;
- Local data is stored on each client machine following the same – previously defined– nomenclature for labels and metadata across institutions. As some institutions may have their own tools or methods to process raw data, we assume the data has passed the quality check;
- Each medical collaborator either has local machines (e.g., local computers) to run the code or an agreement with a research group that acts as a local infrastructure (e.g., high-performance servers, or computer clusters) and is thus able to train models;
- Each participating entity has a Docker cluster or is able to build Docker images to run the corresponding code.

To reproduce the FL process, it requires configuring the corresponding software dependencies, input and/or output parameters, and services. In a traditional way, a user has to manually install all necessary software, enable services, and feed with suitable inputs, a process that is burdensome and time-consuming.

3.3.3 Architecture

The main goal of this work is to simplify the process of federation setup and make it less burdensome and time-consuming for domain-specific users such as AI experts and pathologists. In this work, we propose our FL setup solution, as depicted in Figure 3.1, illustrating an overview of its architecture. It comprises the following five steps:

1. **Model and aggregator definition:** Suppose a Jupyter Notebook user, e.g., an AI expert, develops a cutting-edge model architecture, aggregation function, and other machine learning-related code fragments for the FL training process.
2. **Create FL pipeline building blocks:** For flexibility purposes when it comes to handling collaborating institutions and updates, as well as reducing the complexity of FL deployment and execution, we propose to adopt one Jupyter Notebook extension named component containeriser on the local experimental environment, e.g., NaaVRE [54], to encapsulate FL pipeline building blocks as reusable services, such as model and aggregator.
3. **Build and push job automation:** Once the model and aggregator are ready, the GitHub workflow and Docker registry enable automatic building and pushing to Docker Hub.
4. **Deliver FL building blocks to distributed resources:** The model can be delivered to distributed client users worldwide, i.e., geographically distributed clients in institutions A, B, and C, by pulling the model to local sites. At the same time, the aggregator can be easily delivered to the cloud infrastructure by pulling the aggregator to the cloud virtual machine.
5. **Federation setup:** Once the aggregator is assigned to the cloud infrastructure, e.g., a cloud virtual machine, it is easy to start the Docker container with specific IP and port number. For geographically distributed users, i.e., clients across wide-area institutions, the local AI model training process mainly contains (A) feeding with local data, (B) assigning suitable computation, and (C)

starting the model training with specific data and computation (e.g., GPU resources) on site. By this time, the FL starts.

Specifically, the proposed solution enables a standard centralised AI environment to a decentralised (e.g., raw data) federated setup, flexibly supporting “using AI at the edge”. In such a setting, only the models’ general characteristics are shared. The Notebook Federator provides a Jupyter notebook template that gives the overall structure of the different functions and classes that will be further integrated into the FL pipeline code (<https://github.com/QCDIS/Notebook-Federator>, accessed on 4 January 2023).

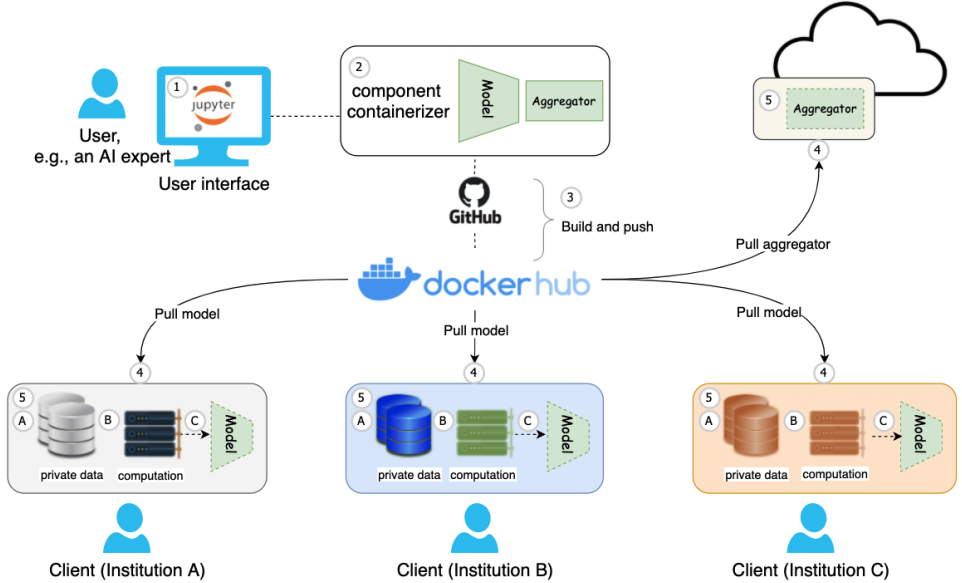


Figure 3.1: An overview of the architecture of Notebook Federator with NaaVRE [54].

3.3.4 Technology considerations

To achieve the automation described in the architecture, we consider the following technologies.

- **Jupyter environment**—We build the FL pipeline with Jupyter Notebook. By default, we suppose that AI experts use a Jupyter environment such as Notebook to design the architecture of the model and aggregation for FL experiments.
- **Component containeriser**—We encapsulate the model and aggregator as reusable FL building blocks based on a Jupyter Notebook. The component containeriser module is one of the Jupyter extensions in NaaVRE.
- **Docker Hub**—We share model and aggregator with distributed users around the world. Docker Hub, i.e., a central repository of containers, is the easiest way to deliver reusable container applications anywhere.
- **Docker Engine**—We enhance the automation of the FL deployment and execution. In this work, we consider the Docker Engine as the critical technology for FL pipeline automation because of its popularity in the community and flourishing software tools such as Docker container, docker-nvidia, docker-compose, or even Docker Swarm. In this paper, we mainly utilise docker-nvidia for local client training with CUDA GPU resources for automating FL pipelines, as we suppose that client users have their demands for local autonomy, e.g., controlling their own data and computation for AI model training, although using Docker Swarm can also achieve large-scale automated deployment for client users.

All in all, the proposed work leverages the available technology assets presented above to bring them together for a specific application, the deployment of FL systems. With all that, the Notebook Federator positions itself as a novel module within the previously developed NaaVRE toolbox [54], providing a Jupyter notebook structured baseline for data scientists to easily reproduce the FL pipeline.

3.4 Case study: histological image analysis

3.4.1 Use case scenarios

We demonstrate the feasibility of the proposed pipeline on a medical image dataset of histological whole slide images (WSIs). More precisely, the WSI dataset used in this study consists of a Spitz melanocytic lesions dataset of 84 biopsies provided, labelled, and annotated by dermatopathologists from the Pathology Laboratory of the University Clinic Hospital of Valencia.

In this case study, we perform a patch-level region of interest (ROI) classification, that is to say, differentiate tumorous patches from non-tumorous ones. To do so, we divide the WSIs into small patches of 224×224 pixels to make the computational needs lighter and allow the use of well-known model architectures with images of the same size. Figure 3.2 depicts a few patch examples of the type of histological data fed into the convolutional neural network in this use case.

Hundreds to thousands of patches can be extracted from a single WSI. Table 3.1 details the number of extracted patches per class, i.e., tumorous vs. non-tumorous, for the WSIs available at each of the three institutions simulated in this case study. To demonstrate the potential of our method, we use twenty of the available biopsies, four of which are used for testing purposes to allow a stable comparison of results across institutions. It is important to highlight the fact that the number of patches is not correlated with the number of WSIs used, as the size of ROIs might vary depending on the WSI under study.

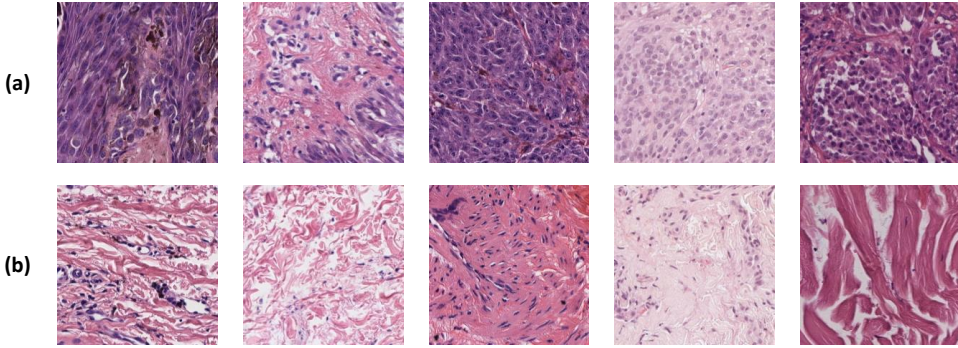


Figure 3.2: Examples of the type of histological patches used in this case study: (a) tumorous patches extracted from the annotated region made by dermatopathologists; (b) non-tumorous patches.

Table 3.1: Details of the histological image data used in this use case. Note that non-tumorous patches were then randomly undersampled to balance the dataset before training each model.

	I_A	I_B	I_C
# WSIs	10	6	8
# tumorous patches	1554	2627	1005
# non-tumorous patches	5609	4694	3979

3.4.2 Federation

The use case presented in this work is part of a concrete medical federation, CLARIFY (<http://www.clarify-project.eu/>, accessed on 4 January 2023). In this multi-sectorial and multidisciplinary consortium, nine institutions across five cities in three different countries are brought together to collaborate on the automatic analysis of specific and challenging cancer types: triple negative breast cancer, high-risk non-muscle invasive bladder cancer, and Spitz melanocytic lesions. In particular, CLARIFY aims to develop an automated diagnostic environment for digital pathology that leverages artificial intelligence methods together with cloud computing, to enable knowledge sharing among institutions and better-informed clinical decisions.

Here, three of the nine collaborators are involved in the different steps of the proposed pipeline, both from the engineering and medical aspects. The cancer type under study in this work, Spitz melanocytic lesions, would strongly benefit from a federation. Indeed, the few incidences of this lesion type in the population is a considerable limitation to gathering sufficient WSIs, both for DL models' training as well as for improving its clinical interpretation.

3.4.3 Federated implementation

To integrate an FL pipeline in the computational pathology use case, we apply the Notebook Federator solution illustrated in Figure 3.1. More precisely, to carry out this implementation on this first use case scenario in realistic conditions, the model was first trained locally on private data at a single collaborating institution before being adapted and deployed for the federation by means of the Notebook Federator.

As one of the requirements of the proposed system is for it to be reproducible and easily implementable for future collaborations, we aim to give a detailed explanation of the steps undertaken to perform that federated implementation, starting from a centralised setup; in other words, the inner workings of the proposed solution. These specific steps are highlighted in Figure 3.3.

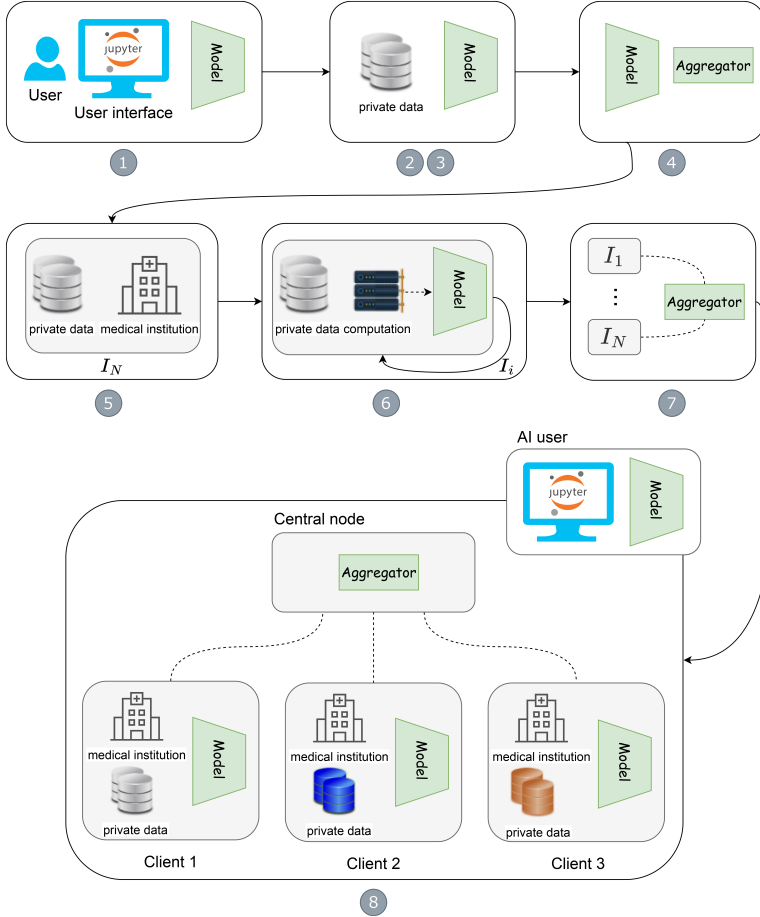


Figure 3.3: Step-by-step implementation of the federated pipeline, the starting point being a local institution with their corresponding local data, and the result being the federation.

First, we fill the Jupyter template provided on NaaVRE (Figure 3.3-(1)) and define a standardised and automated data preparation protocol adapted to the specific type of data under study (Figure 3.3-(2)). As WSI patch extraction was computed beforehand in this use case, the data loading function in the Jupyter template covers loading the patches along with their corresponding label, as well as performing an undersampling of the majority class for each set, as the dataset is clearly unbalanced (see Table 2.1). Then, we define a VGG16-based [42] model architecture

to perform patch-level binary classification with the PyTorch 1.12 library under Python 3.6 and train it locally with the available local data at first to optimize its parameters (Figure 3.3-(3)). To define the aggregation function to update the global model’s weights when mounting the federation, we leverage the federated averaging algorithm FedAvg [69] integrated into FLOWER’s strategies [64] (Figure 3.3-(4)).

Regarding the steps involving the federation, the filled template to share with distributed users is then containerised by means of NaaVRE’s component containeriser, thus allowing their further use by clients by running the corresponding images. This specific step is illustrated in Figure 3.4, which shows how to efficiently containerise reusable model and aggregator components with the Notebook Federator approach. The containers can thus be downloaded and set up at each of the collaborating institutions (Figure 3.3-(5)), and model training can start (Figure 3.3-(6),(7)). The resulting federation (Figure 3.3-(8)) gathers the steps defined previously and communicates across nodes by means of bidirectional gRPC streams.

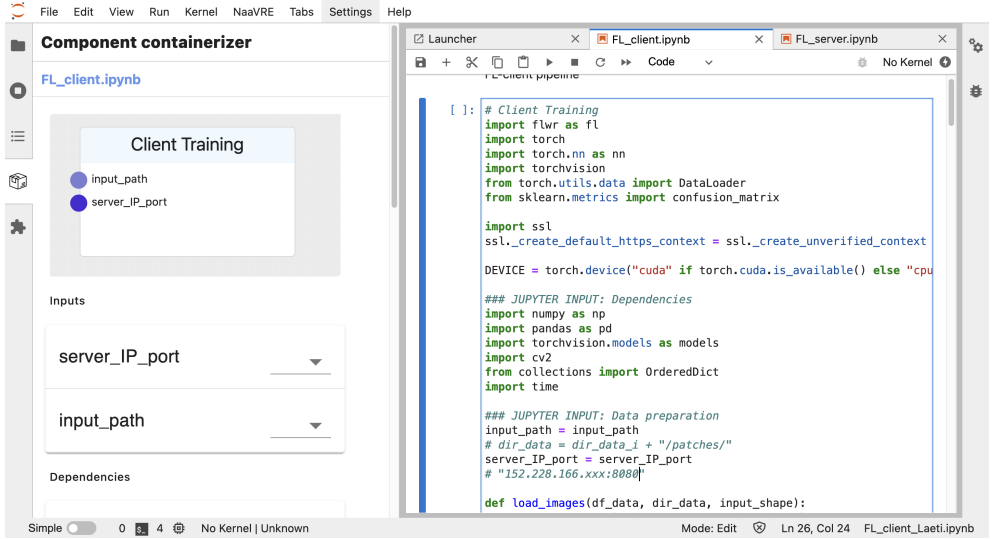


Figure 3.4: Demonstration of the Notebook Federator method with component containeriser in a Jupyter Notebook: building reusable FL pipeline blocks, e.g., Client Training container.

3.4.4 *Experimental results*

3.4.4.1 *FL system setup*

Once the code base is ready on the Jupyter Notebook user side, the component containeriser provided by NaaVRE's toolbox allows users to automate the process of building and pushing a reusable FL-based Docker container to the Docker registry in a matter of minutes. The steps to set up the concrete FL system are quick and simple, as show the times obtained for each final phase of this case study:

- Automated build–push job for the server-aggregation image container (approx. 1.9 GB): 4 m 57 s;
- Building and pushing client-training image container to the Docker hub (approx. 3.16 GB): 10 m 48 s.

Note that the processing time for the build–push jobs depends on the local network environment and image size. The server-aggregation container associated to the host IP address is then deployed on a cloud virtual machine, and distributed users can pull the client-training image and run it locally (e.g., GPU resources in Spain, the Netherlands, and even Norway), as well as mount local medical data volumes to feed the AI model training. For the communication, each client can then join the federation's training process by specifying the server IP address and port number when running the corresponding Docker containers.

3.4.4.2 *AI Training*

As depicted in Figure 3.5, each on-site client used different hardware resources for model training, as is to be expected in real-world scenarios. As a result, whereas the client at institution A took around 60 seconds for 10 epochs on average, clients at institutions B and C trained each communication round for 540 and 534 seconds, respectively. All in all, the whole federation training took 46 minutes for the three communication rounds in this use case, as the central node waits for all three clients to finish training.

To allow for an effective comparison of the prediction performance of models presented in this case study, several metrics of relevance were selected: sensitivity, specificity, positive predictive value, negative predictive value, F1-score, and accuracy. A particularly important metric in pathology is sensitivity, which measures the ability of the model to correctly detect positive instances.

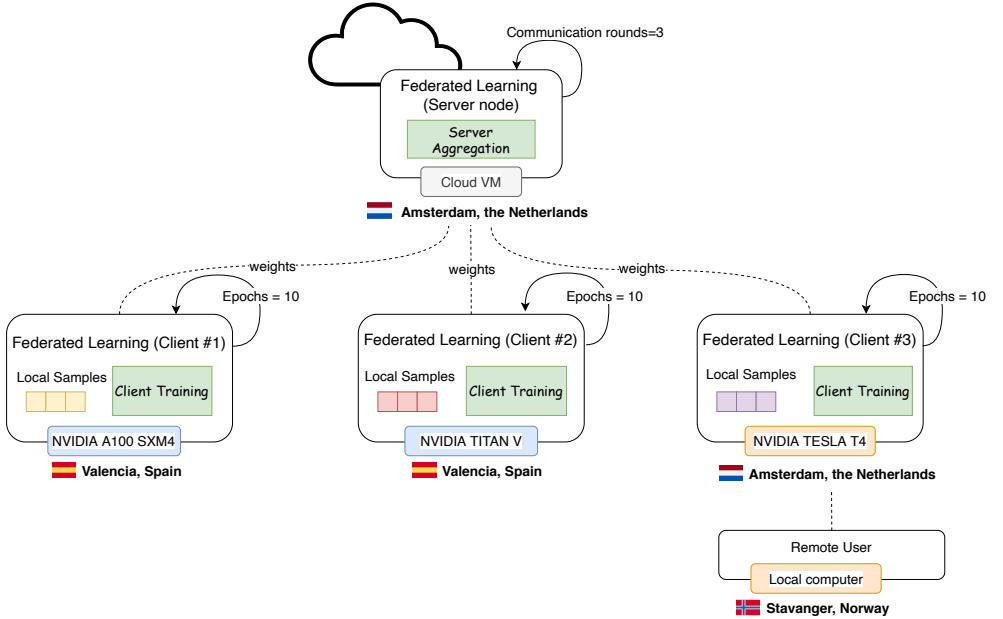


Figure 3.5: Implementation of the proposed solution, Notebook Federator, in a federation composed of institutions located in different countries and relying on different hardware resources.

Table 3.2 shows the metrics on the test set after training for 30 epochs at each institution with its own local data under a centralised configuration, compared to the results when training as a federation with all three clients for three communication rounds of ten epochs each. It is noteworthy to highlight that the scope of this study is not to reach unprecedented classification results but rather to demonstrate the federation’s improvement compared to that of individual institutions.

In that sense, classification metrics on the test set clearly show an overall improvement of the model trained collaboratively (federation) in relation to those trained in a centralised manner. Although I_B reaches considerably high specificity and positive predictive value metrics on the test set when trained under a centralised setup, it is important to highlight that it is mostly due to its tendency to predict a sample as positive (i.e., tumorous) even when negative, as depicted in the negative predictive value results for that model.

Table 3.2: Classification results reached on the test set for patch-level ROI selection on the Spitz melanocytic lesions dataset; I_A , I_B , and I_C : centralised setups at each local institution; Federation: decentralised federated setup, leveraging all institutions’ data.

	I_A	I_B	I_C	Federation
Sensitivity	0.7921	0.5896	0.7887	0.8174
Specificity	0.7842	0.9874	0.8337	0.9247
Positive predictive value	0.7813	0.9961	0.8457	0.9355
Negative predictive value	0.7949	0.3066	0.7734	0.7910
F1-Score	0.7866	0.7407	0.8162	0.8725
Accuracy	0.7881	0.6514	0.8096	0.8633

By and large, the classification results show the potential of the approach, taking into account that, the more institutions participating in model training, the more robust the resulting global model will be when exposed to unseen data.

3.5 Discussion and conclusions

3.5.1 Achievements

In the presented use case, we demonstrated the feasibility of the proposed pipeline to go from a centralised training setup to a federated environment to train collaborative models. Specifically, the Notebook Federator proposes to considerably reduce the complexity of FL systems’ setup for AI Jupyter users by leveraging the latest technologies and available assets.

Starting from a Jupyter environment on the end-user side (i.e., the AI expert or data scientist), we put an emphasis on the reproducibility and scalability of the method for researchers. The Jupyter user only has to complete the indicated lines of the template provided as a code base, and that code is then included in the client code. Building and running the resulting Docker is fast and automatic, thus simplifying the quick implementation of the code on the client machines.

In contrast to other FL solutions and implementations, in this work, we wish to provide a real-world use case, with the server and clients located in different cities and countries, which all have access to different computing resources. Additionally, as our solution aims to simplify the complexity of FL setups and seeks to be easily reproducible, the provided use case leverages a commonly used baseline framework with a strong community and documentation, FLOWER [64]. Although the example we provide is applied to a domain-specific task in computational pathology, it is important to note that the Notebook Federator, as a technology, is generic and can easily be adapted to other tasks and baseline frameworks.

3.5.2 Weaknesses and future work

Some FL-related challenges [51] still need to be tackled in future lines of work. In this study, the GPU and CPU resources used to train the model on the different clients' machines were sufficient, but the computing resources may be heterogeneous across institutions within a federation, and some software or systems may be incompatible if the computing platform cannot handle Docker.

A typical challenge of FL approaches is client reliability, that is to say, whether a client will fail or drop out during a round. In this case study, we consider that the server can only complete a communication round when the three participating institutions finished training, and thus waits for all the clients to have completed data loading and training before performing the aggregation. Even if one of the clients ran the model training on a CPU, the central node waited to receive the weights from all three clients to perform the aggregation. For future implementations

involving a larger number of clients, it could be considered to proceed with aggregation as soon as a certain percentage of clients have completed training, rather than waiting for all.

Another challenge brought by FL ecosystems is security. Although FL helps to tackle the data challenge in DL and the related major privacy limitation, it is still vulnerable to a wide range of challenging security issues. For instance, despite the anonymisation process applied to medical data that was long considered sufficient to protect leaks from patient data when developing AI models, it was shown that, in some cases, sensitive information can actually be inferred [49]. Future lines will thus consider differential privacy preservation.

Currently, some security technologies can be adopted to achieve a privacy-preserving model aggregation for FL [70], such as differential privacy, multi-party computation, or even blockchain-based solutions. Additionally, Docker images may not be optimally secure because of the root access they provide to the system they are running on (<https://researchcomputing.princeton.edu/support/knowledge-base/singularity>, accessed on 4 January 2023). An alternative to the containerisation toolkit would be singularity. This method provides the considerable advantage of limiting users' access and capabilities and includes a compatibility layer that allows Docker images to be run on the platform. The images created via our tool could therefore also be used in a singularity-enabled environment.

Barriers to clinical adoption: democratising active learning for pathologists

While earlier chapters tackle data and technical challenges, this chapter addresses the gap between deep learning model development and clinical use. It presents a user-friendly active learning tool that integrates pathologists directly into the training process, aiming to improve artificial intelligence adoption in practice.

Contents

4.1	Introduction	48
4.2	Related work	49
4.2.1	Digital pathology platforms	49
4.2.2	Active learning in pathology	50
4.3	Methods	51
4.3.1	Software components and architecture	51
4.3.2	Active learning pipeline and specific functionalities	56
4.4	Case study: active segmentation in whole slide images of invasive breast cancer	60
4.4.1	Dataset	61
4.4.2	Implementation details	61
4.4.3	AL user workflow	63
4.5	Experimental setting	64
4.6	Discussion	66
4.7	Conclusions	67

4.1 Introduction

While previous chapters have addressed key challenges in handling WSIs and data scarcity for model development, an equally critical challenge remains: ensuring that AI is truly accessible to pathologists and seamlessly integrated into clinical practice. Despite the rapid progress of AI in digital pathology and the increasing trend of user-AI interactions, its adoption is often hindered by a disconnect between model development and real-world usability. Most AI solutions are designed and fine-tuned by data scientists, leaving pathologists detached from the model development process. This separation arises not only from a lack of technical expertise but also from practical barriers and scepticism about AI reliability and interpretability [23].

This distrust is further amplified for some pathologists by the dependency on specific pathologists' expertise as a reference standard in training AI models, leading to concerns about inter-observer variability and overall trust in the technology [71]. Additionally, many laboratories do not have access to the robust information technology (IT) infrastructure required to support the implementation and deployment of these advanced AI solutions, further complicating their integration [71]. Together, these elements create a significant gap between pathologists and AI model development, considerably affecting technology adoption and AI development in the field.

Meanwhile, training optimal DL models strongly depends on extensive datasets that require manual annotations by expert pathologists, a process that is particularly time-consuming and expensive. Active learning (AL) emerges as a promising solution to reduce the number of annotations needed for model training [72], by selecting the most representative samples to annotate from the data pool [73]. The underlying assumption is that training with the most informative samples will yield insights that allow effective model generalisation, thus reducing the annotation efforts while maximising model performance. And beyond efficiency, AL in digital pathology also has the potential to significantly foster pathologists' trust in AI models by actively involving them in refining model predictions based on their specific expertise. By

allowing pathologists to participate in refining predictions, AL promotes a collaborative approach where human expertise directly influences AI outputs. Since AL allows pathologists to tailor model predictions, it can directly address concerns about inter-observer variability.

However, while AL holds great potential for optimising the annotation process and may help foster AI adoption by pathologists, its effective implementation in real conditions faces considerable complexities. Effective use of AL in real-world settings often requires frequent shifts between platforms and frameworks, posing additional barriers to its integration. To efficiently leverage the benefits of AL in digital pathology, there is a clear need for a seamless solution that can help reduce the barriers to its implementation and simplify its workflow, while giving pathologists the tools to directly interact with the needs of the model for training.

Hence, this work introduces a user-centric AL tool that aims to empower users (i.e., pathologists) by making them an integral part of the training process. Acknowledging the advantages of AL for DL model training, we propose an AL module embedded in a gigapixel visualisation tool, which gets rid of the need to switch between different software tools and libraries to facilitate the integration of AL in digital pathology. By doing so, we aim to democratise AL for pathologists, prioritising their scarce time and ensuring resource-efficient model development under a high-level and user-friendly scheme.

4.2 Related work

4.2.1 Digital pathology platforms

In recent years, several user-friendly platforms have emerged in the digital pathology field. Notably, QuPath [24] stands out as a solution for digital pathology analysis, offering an open-source platform tailored to address specific challenges in WSI analysis. Its user-friendly interface for pathologists and the generic models provided have marked a significant advancement. Following this trend, Dolezal et al. introduced Slideflow

[25], which further contributes by proposing a package to simplify the building and deployment of DL models specifically for digital pathology through a user-friendly interface.

While QuPath [24] primarily serves as a support annotation and visualisation tool for pathologists, Slideflow [25] is more oriented towards simplifying the complexity of model building and deployment, particularly for research purposes. However, these solutions still maintain a clear separation between those who build the models and those who will use them. Additionally, both solutions often require technical expertise for effective use, creating barriers to their integration within established IT infrastructures in many clinical settings [26]. These environments typically favour user-friendly, plug-and-play solutions and often cannot handle high-performance computing requirements.

In contrast, our approach aims to place pathologists at the forefront of model training and optimisation. This goes beyond model deployment, ensuring that pathologists, as end users, take the principal role in the creation of their future clinical algorithms. To do so, we leverage Pixnormous. By choosing to build on a proprietary solution, we can provide tailored and continuous support to pathologists, ensuring commercial-grade performance and seamless integration into existing workflows. This strategic choice directly addresses the scepticism surrounding AI technologies in clinical practice, empowering pathologists to utilise advanced tools without the complications of complex IT requirements or AI expertise. Additionally, a proprietary solution allows us to better accommodate resource limitations often encountered in clinical settings, simplifying the deployment process of both the tool and the models.

4.2.2 Active learning in pathology

Addressing the persistent challenges of annotation in digital pathology, AL brings the promise of mitigating annotation requirements for training models. Its integration typically involves a strategic sample selection process, assessing unannotated samples' informativeness or uncertainty

to prioritise the annotation of the most representative samples, ultimately optimising AL performance [73].

In the context of pathology applications, several works proposed AL frameworks [74, 75], leveraging diverse public datasets. Notably, Li et al. [74] introduced PathAL, an approach aiming to minimise expert annotations by incorporating AL and pseudo masks while excluding noisy samples from the training data. Similarly, Qiu et al. [75] utilised another public dataset to mimic an AL iteration process, proposing a method for dynamically selecting informative annotation regions.

Despite the promise of these AL methodologies in digital pathology, their current implementations remain primarily research-oriented. The practical application of these frameworks in clinical environments still faces significant drawbacks, as the lack of seamless integration and the frequent transitions between platforms and frameworks limit their practical utility. The need for a unified solution that integrates AL seamlessly into the existing digital pathology workflow, without requiring multiple platform switches or additional technical overhead, is therefore crucial for clinical utility.

4.3 Methods

4.3.1 *Software components and architecture*

The proposed AL module was developed as part of the Pixnorous gigapixel visualisation and analysis tool. As such, the seamless integration of a model training and AL module could not be fully described without introducing the overall workflow for WSI handling, together with its components and interactions. Since the complete infrastructure presented in Figure 4.1 is essential for the proper integration of this AL tool, in this Section we present each of its components as well as their relevance in the embedding process of democratising AL for pathologists.

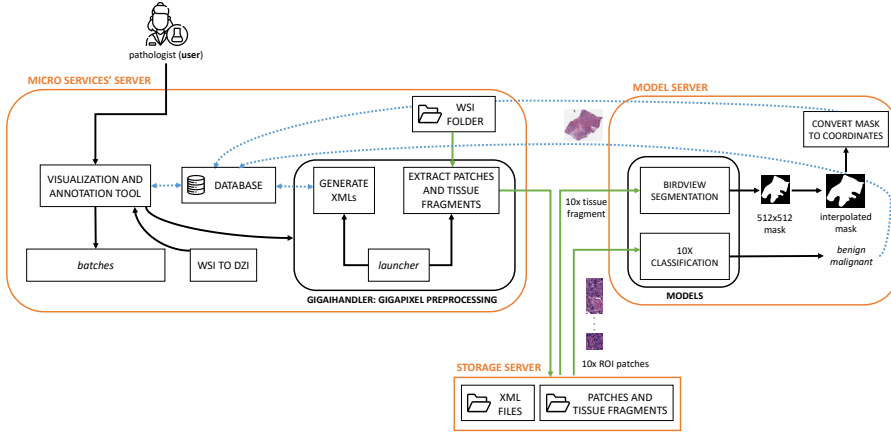


Figure 4.1: Overall system infrastructure of Pixnornious, enabling the seamless implementation of active learning and associated functionalities, together with the interactions between the different components. Micro services server: contains all submodule services necessary for visualisation, annotation, and preprocessing of gigapixel data. Model server: enables the embedding and interaction with AI models within the platform. Black arrows: specific actions; green arrows: data storage-related actions; blue arrows: actions on the database.

4.3.1.1 Visualisation and annotation tool

The central system that enables users and pathologists to visualise and analyse gigapixel images is the frontend of Pixnornious, which serves as a user interface for all functionalities. In a way, it acts as a user-friendly and straightforward communication channel that bridges end users with built-in functionalities and the different services necessary for their proper functioning, represented by the connections with the visualisation and annotation tool block in Figure 4.1-left.

As an all-in-one gigapixel tool, the platform was developed with a set of required functionalities that go from the management of a project, including automatic image preprocessing upon uploading and batch creation, to an extended set of tools for the proper annotation and labelling of WSIs. These can be seen in Figure 4.2, which showcases the visualisation of a breast tissue WSI on Pixnornious along with associated visualisation and annotation tools on the left panel, such as drawing, region growing, rotating the biopsy or zooming in and out, among

most relevant functionalities. This web interface was implemented in JavaScript for both the frontend and backend, and relies on the Paper.js [76] library for drawing and vector graphics interactions.

Since Pixnormous is our main user interface, it acts as a central connector between all components. The different functionalities of the software and preprocessing steps presented below, which are essential for the model training and AL workflow, are thus handled in Pixnormous' background.

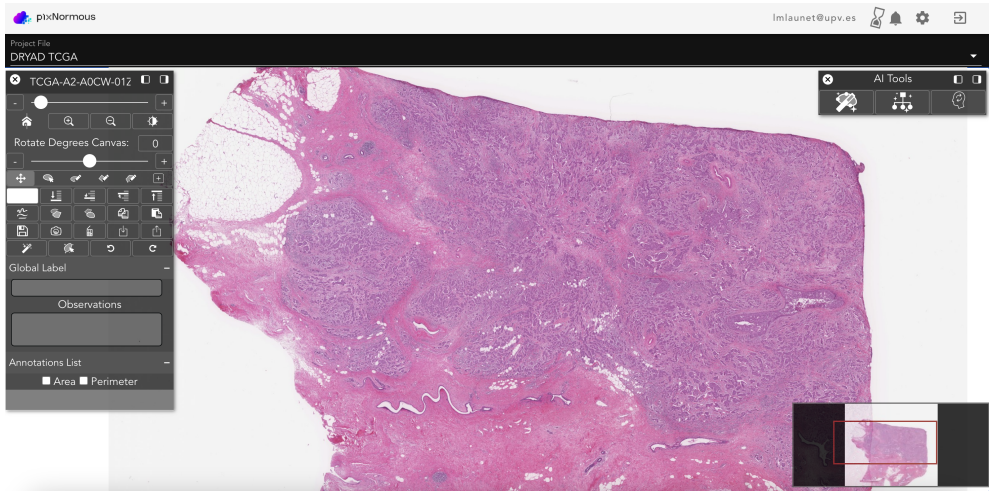


Figure 4.2: Pixnormous, along with the main functionalities for visualisation and annotation on the left panel.

4.3.1.2 Database

To allow the visualisation of annotations on a given WSI, Pixnormous is directly connected to a MongoDB database in its backend to store and access any annotations and labels, together with the corresponding user-related information. These database entries actually represent the key component of the whole system as they are essential in every step of the process, from annotation and label management to data preparation for model training. This is why they are connected to each of the major modules of the system, as show the blue dotted arrows in Figure 4.1.

4.3.1.3 GigaiHandler

The data preprocessing steps are carried out by the GigaiHandler module presented in Chapter 2. GigaiHandler is deployed as a Flask application [77] within a Docker container, allowing access to the different preprocessing functionalities through a REST API. In this way, the preprocessing buttons integrated into the user interface send direct POST requests to the preprocessing server, which are executed in different threads. Once each task is completed, the server communicates with the user interface to transmit the relevant information, as illustrated in the left block of Figure 4.3.

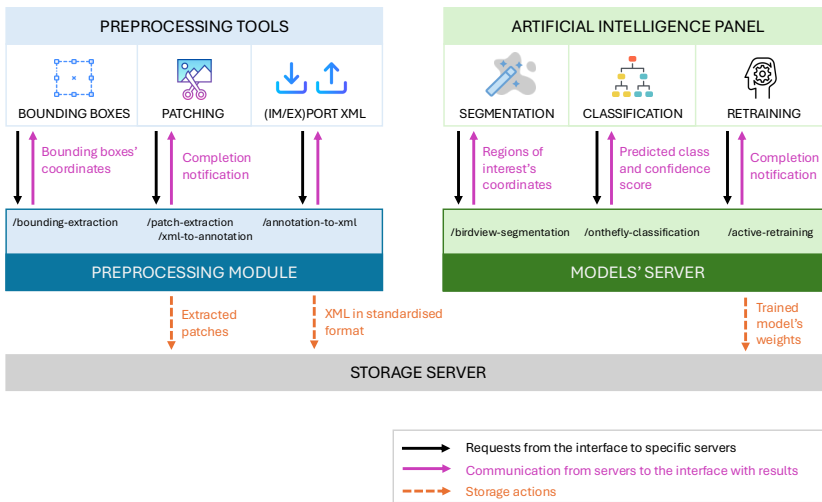


Figure 4.3: Interaction flow between the buttons embedded in the interface and the preprocessing and artificial intelligence models' servers.

4.3.1.4 Model server

While previous modules are the building blocks of the system, the specific module that enables the integration of DL models with AL methods is deployed in a separate server that corresponds to the module in Figure 4.1-right, the model server. This block runs as a Flask application that enables receiving requests from other modules inquiring about a specific segmentation or classification prediction.

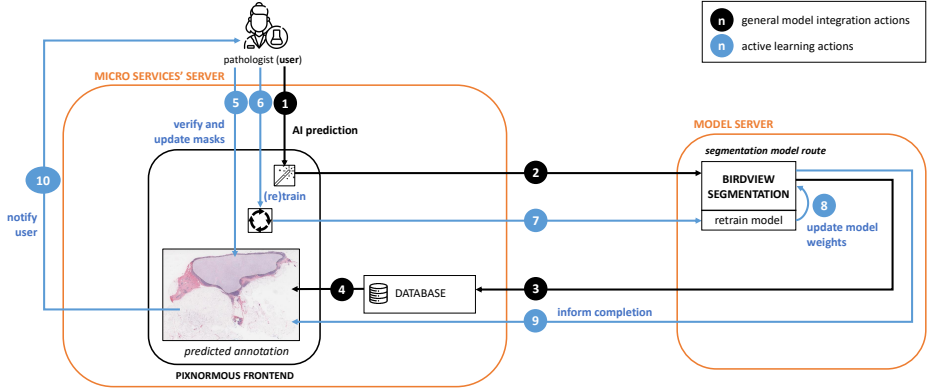


Figure 4.4: Active learning pipeline outlining the system’s workflow, from the initial user request for prediction and retraining to the final completion notification.

- Pipeline.** User-friendly buttons were added to Pixnormous for model inference, as seen in the *AI Tools* box in Figure 4.2-top right. Obtaining a model prediction is as simple as clicking on one of these buttons, which initiates the sequential stages illustrated by the black arrows in Figure 4.4. Upon selecting the segmentation inference button, a request is sent to the corresponding route within the model server to request a new prediction from the model, see right block in Figure 4.3. The predicted mask coordinates are then transmitted to the database, which is directly linked to the visualisations in Pixnormous.
- Data preparation.** As explained in the previous subsection, patches and tissue fragments are automatically generated upon uploading new WSIs to Pixnormous. Therefore, when sending an inference request to the model server, the corresponding preprocessed data is automatically loaded from the storage server illustrated in Figure 4.1-bottom. For segmentation, the loaded image is then resized to a 512×512 square to feed the model, while for classification a patch selection process is carried out by generating the annotation mask from the XML file and only selecting patches located within the annotation coordinates.

- **Inference.** Upon the initiation of the Flask application for the model server, the weights of both segmentation and classification models are loaded. This enables immediate predictions upon receiving requests from Pixnormous. When opting for a different model for predictions, its weights are loaded seamlessly, replacing the previously loaded model.

To address the computational demands of high-resolution pathology datasets, the platform is designed with scalability and performance in mind. First, the platform follows a multithreaded design to manage image loading and prediction tasks, enabling efficient parallel processing and reducing response times under heavy workloads. Additionally, GPU resource utilisation is optimised by dynamically loading only the required model weights during inference, ensuring inactive models do not consume memory or computational resources unnecessarily.

4.3.2 *Active learning pipeline and specific functionalities*

To simplify AL implementation in a user-centric and empowering way for pathologists and users of our platform, we defined and integrated into our ecosystem a set of functionalities specifically dedicated to DL models. By integrating user-friendly buttons and notification-based feedback into Pixnormous, (re)training models under different learning scenarios is only one button (and a few annotations) away for pathologists.

Indeed, by clicking on the (re)training button on the *AI tools* panel (see Figure 4.3-right panel), i.e., the right button, a popup action window appears to enable the user to select the parameters to (re)train the model, as shown in Figure 4.5. These parameters include the learning scenario together with a specific version of the segmentation model for weights' initialisation.

4.3.2.1 Learning scenarios

To make the implementation of AL realistic, we propose to integrate different learning scenarios into our platform, which opens the way to a more sophisticated implementation of AL models. When the (re)training action window pops up, the user can thus select among the following learning schemes:

- **Ground truth.** Annotations directly made by expert pathologists on Pixnormous serve as the sole input for training. Note that all training scenarios mentioned below incorporate images with ground truth for training.
- **Self-training.** Utilise unannotated images by integrating masks predicted by the segmentation model as pseudo ground truths, referred to as pseudo masks.
- **Active learning.** Train the model with the reviewed pseudo masks only, together with ground truth masks. Validation by pathologists is required for considering a mask prediction as reviewed, allowing modifications if needed.
- **Active self-training.** Gather all available masks for training, including ground truth and pseudo masks, whether reviewed or not.

4.3.2.2 Scenarios' management

To facilitate model training under different learning scenarios, specific information regarding annotation types is crucial. Each entry of our MongoDB database thus includes the following properties for every annotation:

- **AI.** Set to False for annotations made by pathologists using Pixnormous, and True for model predictions. All subsequent properties only apply when the AI tag is set to True.
- **Modified.** Indicates if a user modified a pseudo mask after inference.

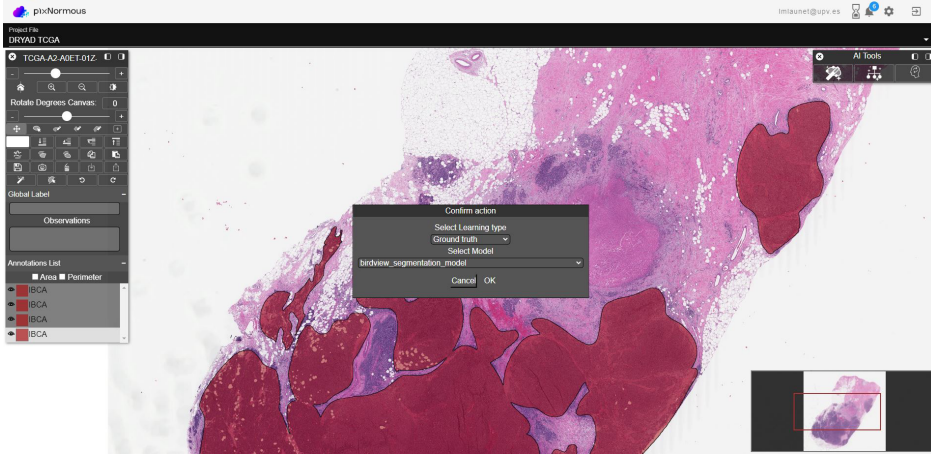


Figure 4.5: Model (re)training functionality, integrated within Pixnormous proprietary software by means of buttons and pop-up selections.

- **Reviewed.** Specifies if the user reviewed and validated the pseudo mask, allowing to further include it in the *active learning* scenario.
- **Label confidence.** Ranges from 0 to 1, denoting the model's confidence in the classification prediction. Set to 1 for ground truth labels.

These properties, stored for each annotation in the database, are also saved as tags in the generated XML files. The associated XML structure standard necessary to prepare data for training, including its specific tags and values, is exemplified in Figure 2.3 in Chapter 2.

4.3.2.3 Training data selection

When selecting a learning scenario on Pixnormous for (re)training, see Figure 4.5, the models' server receives a request with the model name and learning type as input parameters. The latter is essential to prepare the data for training, as it retrieves the annotations of each tissue fragment and filters out annotations that do not match the selected learning scenario. Therefore, the data used for training only contains annotations aligned with the user's chosen learning scenario.

4.3.2.4 Model training and post training

In addition to selecting a learning scenario, users are prompted to choose a model from a predefined set of options, with the initial trained model set as the default choice. This functionality becomes especially useful after multiple training iterations, as it allows users to select an alternative model as the starting point for weight initialisation for the new training round.

Once the training is finalised, the updated model weights are stored in our storage server (see Figure 4.1, lower block) along with information about the training end date, test dice, and test loss. If these computed scores surpass the metrics of the currently loaded model, the model server reloads the weights of the model used for inference, as only one model at a time is active. Users can then change the model selected for inference anytime when inquiring about a new prediction.

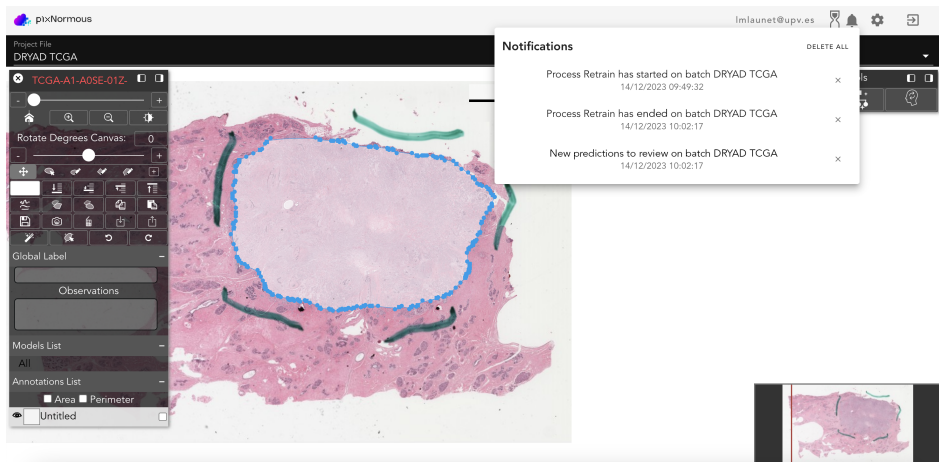


Figure 4.6: AL feedback in Pixnормous. Left: the image name is highlighted in red if it is selected for review; once reviewed, users need to fill a checkbox next to the annotation name. Up right: model training and review query notifications for pathologists.

4.3.2.5 Feedback and annotation review

To enable the seamless integration of AL without the need for constant platforms switches, while empowering pathologists using the platform, an essential component of our solution includes the computation of scores for all pseudo masks upon training finalisation. These pseudo masks correspond to the model’s predictions on unannotated tissue fragments initially lacking ground truth annotations. While various methods are proposed in the literature, the platform currently relies on the Shannon entropy [78] to compute uncertainty scores.

The computed scores serve as the foundation for AL, allowing to select samples to be reviewed or annotated by expert pathologists, who receive notifications and tasks through our platform as illustrated in Figure 4.6. This feedback loop plays an integral role in the AL integration, working hand in hand with pathologists who occupy a central position in our framework, ensuring a seamless synergy between DL models and human expertise.

Upon receiving the review query notification on Pixnornious, pathologists are therefore invited to review the pseudo annotations predicted by the model for the automatically selected samples. Experts then have the option to validate a pseudo annotation, make necessary modifications, or improve the annotations for the samples, understanding that these samples are presumed to offer more pertinent information for the subsequent model training.

4.4 Case study: active segmentation in whole slide images of invasive breast cancer

To validate the efficacy of the developed ecosystem, we propose a comprehensive simulation of a full AL workflow tailored for digital pathology clinical applications, utilising a public dataset. To do so, we leveraged Pixnornious to carry out each step of the workflow seamlessly from a user’s perspective. This user-centric case study ensures that every action, from the initial WSI uploading to executing various model training scenarios, is conveniently accessible within a few clicks.

4.4.1 Dataset

For this case study, we resorted to The Cancer Genome Atlas Breast Cancer (TCGA-BRCA) public database, specifically focusing on invasive breast cancer (IBCA). Recognised as the most aggressive and clinically challenging subtype of breast cancer, IBCA is generally characterised by the uncontrolled spread of cancer cells from the breast tissue to other parts of the body [79].

We collected a total of 195 WSIs from the TCGA database, along with the annotations provided for the test set of the database from Dryad shared in [80]. In this dataset, WSIs were manually annotated by three expert pathologists at a $2\times$ magnification level [80], delineating the invasive regions of IBCA within each WSI. The annotations were provided as binary masks for each image, along with the $2\times$ magnification level of the WSI.

4.4.2 Implementation details

While the primary focus of this work lies on the user-centric AL tool's integration and its benefits for pathologists, it is essential to outline the specific implementation and development details to thoroughly carry out the case study validation.

4.4.2.1 Model architecture

The segmentation model architecture was previously defined based on the U-Net backbone [81] presented in Chapter 2, adapted to suit the specific input and output requirements of this case study. In this specific application, input images for the model were retrieved from the database [80] from Dryad, which provides $5\times$ images together with their corresponding masks. We utilised GigaiHandler to extract tissue fragments from the $5\times$ images and their masks. Subsequently, these were resized to the input size of our model for training, set to 512×512 pixels. A validation set comprising 20% of the training set was used for validation purposes.

The model architecture was then integrated into the model server application along with its parameters, as depicted in Figure 4.4-right, rendering the models and complete platform ready for any prediction or (re)training actions on this specific dataset. All these operations relied on a NVIDIA A100 GPU as a computational backbone for the model server deployment, used for pre-production purposes.

4.4.2.2 Evaluation metrics

While our work primarily focuses on system-level implications rather than being solely model-centric, properly assessing the different approaches was essential to demonstrate the validity and applicability of our proposed approach on an example use case. To do so, we thus resorted to relevant evaluation metrics popularly used in medical image segmentation, the Dice similarity coefficient (DSC) and intersection over union (IOU). These metrics measure the overlap between two sets, with the IOU penalising the discrepancies between the predicted and actual regions more heavily. Specifically, these two metrics are calculated as follows:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}, \quad IOU = \frac{|X \cap Y|}{|X \cup Y|}$$

with X defined as the set of pixels predicted by the model and Y the set of true pixels (ground truth) that should ideally be matched by X .

For the selection of pseudo masks to be reviewed by pathologists, we leveraged a well-established measure of uncertainty, the Shannon entropy [78], mathematically defined as:

$$H(x) = \arg \max_x - \sum_i P(y_i | x) \log P(y_i | x)$$

where x represents the input image and i denotes the number of classes. Using this measure, we identified the 15 images with the highest entropy

scores at each step, indicating greater uncertainty, to query users on the platform.

4.4.3 *AL user workflow*

4.4.3.1 *Data preparation*

To prepare the dataset for training, we downloaded the IBCA WSI directly from the TCGA-BRCA database and uploaded them to a new batch in Pixnormous. Simultaneously, we retrieved the ground truth masks provided by the Dryad database [80] and converted them to XML files following the standard proposed in our study, as exemplified in Figure 2.3.

From the pool of 195 available annotation masks, we randomly selected 29 (15% of the data) for testing, serving as a benchmark for model comparison. For the initial iteration of training for the ground truth-based model, we uploaded one-fourth of the annotations to Pixnormous, totalling 49 XML files.

4.4.3.2 *Use case experiments*

To validate the developed ecosystem, we utilised Pixnormous user interface to carry out the different learning scenarios proposed on our platform as defined in the previous Section. Therefore, the simulated user would execute the following experiments seamlessly on Pixnormous without switching platforms: (1) ground truth training (GT), (2) self-training (ST), (3) active learning (AL), and (4) active self-training (AST).

To carry out these experiments, we exclusively utilised the integrated buttons for model training. These trigger the background steps to initiate the model training specific to each learning paradigm, based on the annotations uploaded or delineated on the platform. In this case study, only one iteration of AL was performed since the primary focus was to demonstrate the platform usability for AL and ST training rather than conducting an exhaustive evaluation of model performance.

The platform’s integrated notifications played a crucial role, serving as indicators for the beginning and conclusion of each model training process. They also highlighted predictions with higher levels of uncertainty, prioritising annotation reviews on specific segmentation masks for AL endeavours. Once training concluded, the newly trained model is ready for utilisation, enabling predictions on the uploaded images of the user’s preference with the trained model of their choice.

4.5 Experimental setting

Table 4.1: Segmentation metrics on the test set for the different learning experiments proposed. GT: ground truth; ST: self-training; AL: active learning; AST: active self-training. DSC: Dice similarity coefficient. IOU: intersection over union.

	DSC	IOU
GT	0.9049	0.7615
ST	0.9186	0.7747
AL	0.9077	0.7526
AST	0.9216	0.7767

The experiments carried out for this use case aimed to evaluate the usability and efficacy of training with different learning paradigms within our user-centric AL platform. The performance metrics used in this set of experiments, i.e. the DSC and IOU, helped assess the segmentation accuracy across varied training experiments. The metrics, depicted in Table 4.1, served as quantitative benchmarks to measure the performance of each experiment for this specific use case.

The GT training, despite using a relatively small dataset, demonstrated robust segmentation metrics with a Dice score of 0.9049, serving as a baseline for subsequent training phases. The ST paradigm, which leveraged the GT dataset together with the pseudo masks predicted by the initial model, exhibited iterative improvement with a Dice score of 0.9186. In contrast, the AL phase, where pathologists were queried to review 15 pseudo masks selected based on the Shannon entropy scores, achieved a better Dice score than the initial GT-based model but fell short compared to any ST-based approach. This difference could

be attributed to the initial model’s robustness in this particular use case, which facilitated the creation of a consistent and reliable pseudo mask dataset for the ST approaches, in comparison to the relatively limited number of images in the AL phase. However, the AST phase, encompassing all available data including GT, pseudo masks reviewed by pathologists during the AL phase, and pseudo masks, demonstrated significant advancement. Leveraging this comprehensive dataset resulted in higher performance metrics, achieving a Dice score of 0.9216, signifying a notable improvement in segmentation accuracy. The AST phase’s success underscores the synergy achieved by joining together user-reviewed data and pseudo masks to considerably augment the dataset, leading to more diverse and refined training.

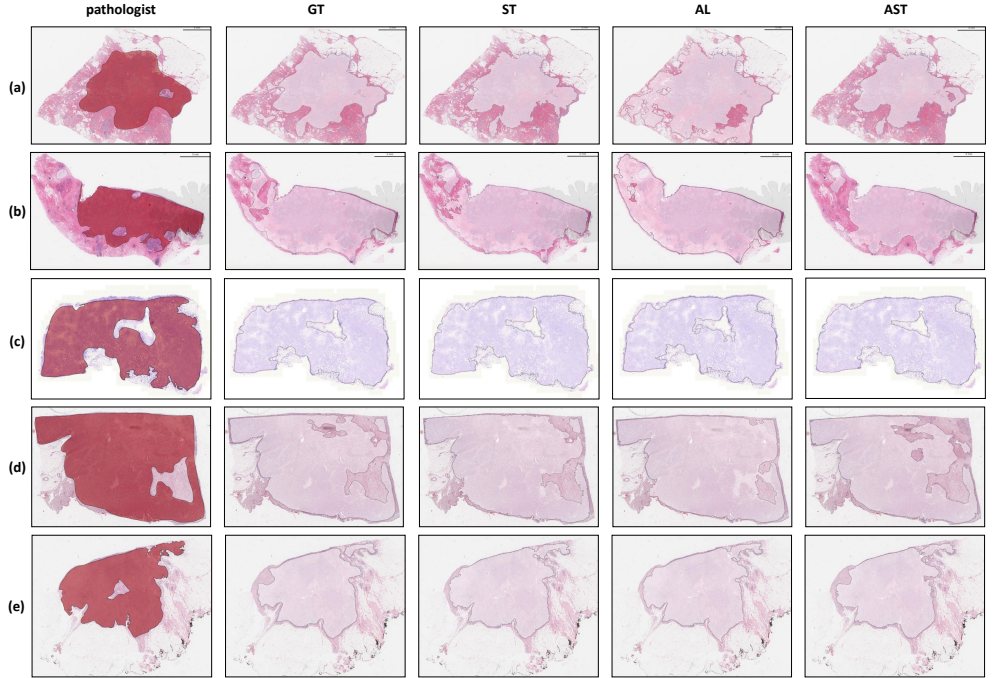


Figure 4.7: Visual comparison of the models’ predictions on a subset of test images. From left to right: GT from pathologists; model trained with GT only; ST model (GT and pseudo-masks); AL model (first iteration); AST model.

These results are visually depicted in Figure 4.7, representing the model predictions generated with the inference buttons on Pixnormous. This figure offers a representative glimpse of the models' performance, thanks to the mask predictions of a subset of test images for each model across all the proposed learning scenarios. The second column highlights the generalisation of the initial GT paradigm with respect to the pathologist's annotation (first column). It sets a baseline performance level, delineating how different learning scenarios refine or deviate from this baseline. Generally, the transition from GT to ST presents a subtle improvement, while the AL scenario tends to perform worse compared to the ST paradigm (Figure 4.7-(a) and (b)), with some exceptions such as Figure 4.7-(c) and (d). The AST scheme, in comparison, generally enhances the segmentation accuracy when compared to other paradigms.

4.6 Discussion

In the presented case study, we demonstrate the efficacy and applicability of the proposed AL system, placing platform users (i.e., pathologists) at the core of the training process, only a few clicks away from models' training results. Through a user-friendly interface that positions AI and AL a few clicks away from end users, pathologists are given the agency to curate images, provide annotations, analyse model predictions, and refine predicted masks. Pathologists actively engage in every step of the process, ensuring their integral involvement at critical stages. Through this approach, our objectives are two-fold: to ensure that the AI models are finely tuned to meet the specific needs of digital pathology, and to empower primary end users to be a part of the development process of the tools intended for their future practices.

By becoming an integral part of the DL and AL training process, end users are given the possibility to connect directly with the freshly obtained results, fostering a deeper understanding of model performance and training iterations. This active participation not only aims to enhance their confidence in utilising AI-based solutions but also aligns these tools with the specific needs and practices of the clinical workflow.

The primary objective of this work was to implement a functional proof of concept working with pathologists to define the platform’s functional and non-functional requirements. Moving forward, the next stages will focus on conducting comprehensive usability studies to collect feedback from pathologists who are not yet familiar with the tool, evaluating their interaction experiences to guide further improvements.

Moreover, testing the platform for a use case of a complete AL pipeline in digital pathology also highlighted some areas for improvement that will be addressed in the next iterations of our work. Among these is the lack of customisation options for model architecture and fine-tuning model parameters, which still entails relying on AI developers to be able to initiate project development. Additionally, models would likely benefit from integrated parameter search mechanisms to ensure generalisability and robustness of models across varied datasets. To better adapt the proposed tool to the specific needs of digital pathology, upcoming iterations will integrate customisation features for model architecture or input data specifications, and automated parameter optimisation techniques. Finally, as pathologists are put at the centre of our approach, special efforts will need to be put into the feedback they receive, notably by implementing explainable activation maps to explain the models’ decisions.

4.7 Conclusions

In this work, we introduce an all-in-one active learning system designed for digital pathology, placing end users at the core of the training process. By bridging the gap between pathologists and the tools destined for integration into their future clinical practices, our approach aims to democratise artificial intelligence and active learning scenarios, bringing artificial intelligence only a few clicks away from end users. The presented case study highlights the platform’s efficacy in engaging end users in the active learning process, creating a collaborative ecosystem that aims to advance healthcare diagnostics and patient care. With all this, this chapter lays the groundwork for subsequent integration of deep learning models and active learning configurations, bringing intelligent systems one step closer to clinical practice.

Scarce annotations: active learning

Building upon the pathologist-centred active learning tool introduced in Chapter 4, this chapter goes further into expert-in-the-loop training by exploring different active learning setups and examining their impact on model performance and annotation efficiency in histopathology.

Contents

5.1	Introduction	70
5.2	Related work	71
5.3	Material	72
	5.3.1 Dataset	72
	5.3.2 Experimental setup	73
5.4	Methods	73
	5.4.1 Problem formulation	73
	5.4.2 Data preprocessing	74
	5.4.3 Active learning framework	75
	5.4.4 Real-conditions integration	77
5.5	Results	78
	5.5.1 Ablation experiments	78
	5.5.2 Interobserver study	80
5.6	Discussion	83
5.7	Conclusions	84

5.1 Introduction

Having established tools to deal with complexity and expert integration, the next step is to dig deeper into how AL can effectively address the challenge of scarce annotations in histopathology. Beyond making DL and AL accessible to pathologists, it is equally important to understand its implications, particularly how different training settings, such as AL configurations and sample selection strategies, can influence model performance in histopathology. Optimising these settings is essential to ensure that DL models are both efficient and reliable.

To ensure the successful integration of AI systems into clinical practice, a practical and effective approach is needed, one that not only provides a robust technical framework but also considers the sensitivity of model performance to various training configurations.

In this work, we propose a histopathology AL framework for segmentation, trained and refined in real conditions. After training the network with annotated images for the first iteration, we combine different training configurations to provide an extensive analysis of the optimal AL setup. To carry out this process, we leverage the system presented in Chapter 4, providing feedback to pathologists in a visual way after each iteration.

In parallel, the vast size of WSIs presents significant challenges when analysing them with AI. Building upon Chapter 2's segmentation model, inspired by how pathologists analyse biopsy slides under the microscope, we propose to validate our method on a pathologist-like segmentation model that locates the regions of interest (ROI) of the different tissue fragments from a bird's eye view.

Specifically, our main contributions include:

- An AL framework designed to function in real clinical environments, integrated into a proprietary platform for WSI navigation, annotation and analysis that enables seamless collaboration with pathologists;

- Empirical comparison of AL configurations to determine the optimal setup for histopathology segmentation on Spitz tumours, considering factors such as sample selection strategies, random seed variations, and training configurations;
- Explore how AL methods can be effectively employed in histopathology workflows, improving segmentation efficiency and reducing annotation burden;
- Actionable insights on interobserver comparisons to validate AI models and improve clinical relevance and applicability.

5.2 Related work

The core idea behind AL is to iteratively select the most informative samples for annotation, enabling the training of DL models using significantly fewer labelled samples [72]. In practice, this involves an initial training phase with a small set of labelled data, followed by the model’s evaluation on unlabelled samples to determine which ones are most uncertain or likely to improve performance if annotated.

Several studies have investigated AL for histopathological image analysis with a strong focus on sample selection methods, particularly in classification tasks [74, 82–84]. Li et al. [74] introduced the PathAL framework, which selects samples for expert annotation based on their informativeness while also incorporating pseudo-labels from confidently predicted samples into training to improve classification performance. Meirelles et al. [82] developed a novel AL selection method leveraging both model prediction uncertainty and the image features of patch samples. In [83], Schmidt et al. proposed a patch-level probabilistic DL approach to ignore artifacts and ambiguous images in the sample selection process, as they do not contain relevant information for tissue classification. In contrast, Kang et al. [84] introduced a method for AL sample selection strategy at the WSI level, defining a novel class distance metric to identify most informative WSIs.

While AL approaches focusing on classification have gained traction, fewer studies have explored AL for segmentation in histopathology. Among these, Qiu et al. [75] proposed an adaptive region selection approach that dynamically selects the region size of annotation regions to improve annotation efficiency. Folmsbee et al. [85] compared random selection strategy with a team of expert pathologists to select samples in each iteration, enabling more efficient ROI-scale segmentation. In contrast, Belharbi et al. [86] found that random sampling could outperform AL-based selection when combined with pseudo-labelling strategies.

Despite these advancements, AL sample selection strategies for segmentation vary significantly across studies, making it difficult to establish a consensus on their effectiveness. In addition, approaches usually rely on pixel- or patch-level selection criteria rather than considering tissue fragments as a whole, as pathologists do when interpreting biopsies in a broader clinical context. Furthermore, integrating AL frameworks into real clinical workflows remains challenging, as experimental settings often fail to capture the complexities of pathology practice.

Another important yet often overlooked issue in AL is interobserver variability, which refers to the inconsistency in annotations among different pathologists. While this variability is well documented across multiple histopathological assessment tasks and can significantly affect model performance [87–89], AL frameworks do not consider expert disagreement during the iterative selection process. Without accounting for interobserver variability, AL models may propagate annotation inconsistencies, ultimately affecting model performance and clinical applicability.

5.3 Material

5.3.1 Dataset

The proposed methodology was evaluated on a private dataset of H&E-stained biopsies that were digitised and provided by the University Clinic

Hospital of Valencia. Specifically, the dataset consists of 203 WSIs of spitzoid melanocytic tumours, an uncommon type of skin lesions that present ambiguous morphological features [90].

5.3.2 Experimental setup

In this study, we adopted the architecture presented in Chapter 2 [91], which utilised a U-Net backbone architecture [81] enhanced with residual blocks [92] to perform a bird’s eye view segmentation of tumour area, mirroring the process used by pathologists, who first identify the tumour area at a lower magnification.

To ensure robustness and generalisability, we trained each model under a 4-fold cross-validation scheme, to help mitigate overfitting and provide a more reliable assessment of model performance by evaluating it on multiple subsets of the data. All models were trained for 40 epochs on an NVIDIA DGX A100 server where the models’ module was deployed and connected to the Pixnormous software for automatic retraining of models.

5.4 Methods

5.4.1 Problem formulation

Following the paradigm of AL, we proceed on the basis that there is a substantial pool of unannotated data S^u available and that we can rely on an expert pathologist to annotate the ROIs of a selected subset of images to iteratively add to the annotated set S^a , as depicted in Figure 5.1.

Data subsets. In this approach, we consider three different subsets: (1) the unannotated set $S_i^u = \{x_j\}$ that only contains the RGB tissue fragment images x_j ; (2) the annotated set $S_i^a = \{x_k, m_k\}$, which also comprises the corresponding annotation mask m_k made by an expert pathologist for each given tissue fragment x_k ; (3) the pseudo-annotated set $S_i^p = \{x_j, m'_j\}$ where the annotation mask m'_j is the mask prediction made by a segmentation model, without revision by experts. Our pipeline

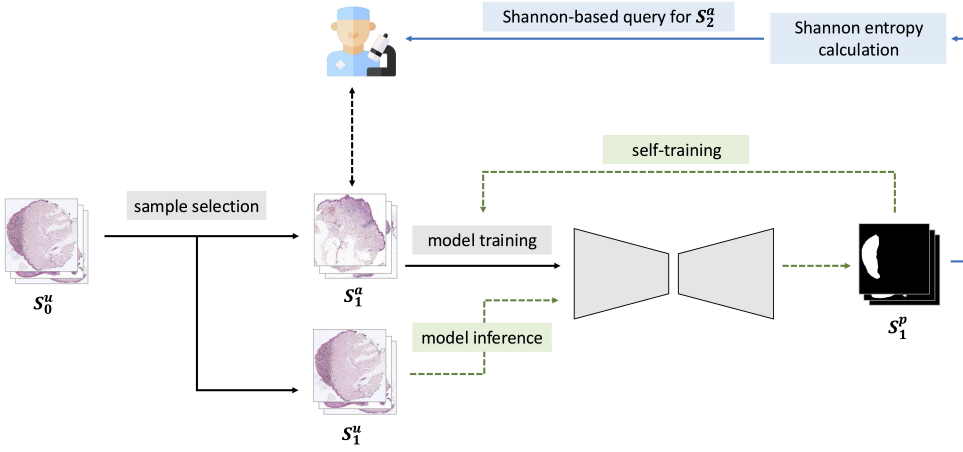


Figure 5.1: Overview of the approach presented in this study. Blue arrows: sample selection strategy; green arrows: model prediction and self-training process.

proposes four setups that emphasise different learning types in real conditions, each of them consisting of combining one or several of these subsets. Note that i indicates the iteration of the learning process, where different data subsets may be used or updated; j indexes unannotated images and their corresponding pseudo-masks, and k corresponds to annotated images with expert-provided masks.

5.4.2 Data preprocessing

As depicted in the lower panel of Figure 5.2, tissue fragments used for segmentation are first automatically extracted from the WSIs, each containing small sections of skin tissue composed of layers of the epidermis, dermis, and sometimes the underlying subcutaneous tissue. To obtain these tissue fragments out of input WSIs, we leverage the GigaiHandler preprocessing module presented in Chapter 2. Note that these steps are automatically embedded in Pixnormous, as described in more detail in Chapter 4. Extracted tissue fragments are then resized to 512×512 squares to feed the models.

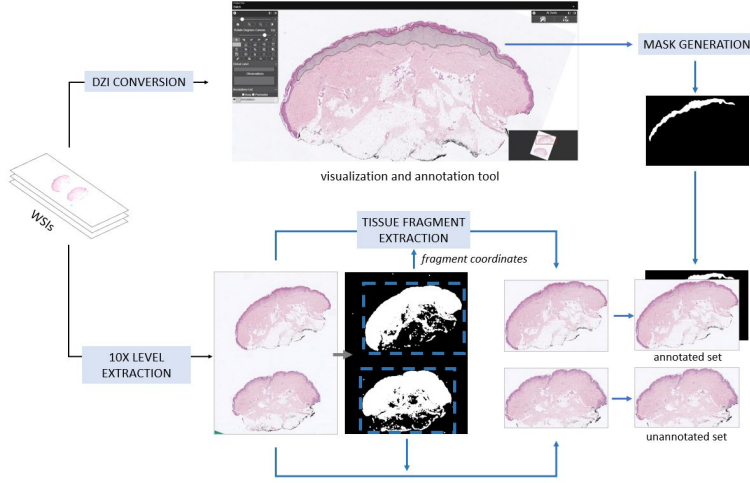


Figure 5.2: Data preparation pipeline for the approach proposed in this paper. Top panel: Pixnornous, used for WSI visualisation and annotation (review) by pathologists; lower panel-(left): extraction of the tissue fragments to segment by getting the coordinates of the different regions with tissue within a slide; lower panel-(right): creation of the two main subsets depending on the annotations made by pathologists.

5.4.3 Active learning framework

Initial setup. Our pipeline proposes different setups that emphasise various learning strategies, each involving one or more of the defined data subsets. The ground truth iteration ($i = 0$) serves as a benchmark and consists of a fully-supervised model trained with a set S_0^a of $k = 76$ ground truth annotations. The corresponding training set is therefore defined as $T_0 = \{S_0^a\}$. Once trained, this initial model is used to generate pseudo-annotations for the samples of the unannotated set S_0^u , resulting in the pseudo-annotated set S_0^p . These pseudo-annotations are then combined with the ground truth annotations to train a self-training model with the training set $T'_0 = \{S_0^a \cup S_0^p\}$.

Active learning iterations. Building upon this initial ground truth model, we conduct a series of 10 AL iterations ($i = 1$ to $i = 10$), where at each step, 10% of the remaining unannotated samples are queried and added to the annotated set. For each iteration, certainty scores are computed for predicted samples in the prior iteration. For the first

iteration ($i = 1$), we thus leverage S_0^p to select the subset $Q_1 \subseteq S_0^p \setminus S_0^a$ for review, based on a confidence threshold described below. This subset will serve to train the first AL iteration with $S_1^a = \{S_0^a \cup Q_1^r\}$ where Q_1^r denotes the reviewed pseudo-annotations. The training set for the first AL iteration is therefore defined as $T_1 = S_1^a$, and the self-training set $T_1' = S_1^a \cup S_1^p$ includes both expert annotations and new pseudo-annotations generated by the current model.

This iterative process repeats for $i = 2$ to $i = 10$, with the models continuously refined by incorporating newly reviewed and predicted annotations.

Querying strategy. To enable the optimal selection of the most informative samples, a confidence-based sample selection approach is employed. This method prioritises the annotation of challenging samples that are more likely to enhance the model’s learning and overall performance. Specifically, a pool-based sampling strategy is used, where at each iteration the current model generates the pseudo-masks of all images from the unannotated pool. The top n samples (10% of the remaining unannotated dataset) with the highest uncertainty are then selected for annotation. Although performing inference on all unannotated samples might be computationally expensive, it has been shown to be the most promising approach when combined with deep learning methods [73].

In this context, Shannon entropy [93] is the most commonly used uncertainty metric due to its effectiveness in quantifying the unpredictability of a model’s predictions, with higher entropy indicating greater uncertainty. Specifically, Shannon entropy is defined as follows:

$$H(x) = \arg \max_x - \sum_i P(y_i | x) \log P(y_i | x)$$

with x being the input image and i the number of classes.

5.4.4 Real-conditions integration

Integrated platform. The digitised WSIs were uploaded to the Pixnormous high-resolution zoomable images visualisation tool, building upon the models' integration presented in the previous chapter (see Chapter 4).

Pathologist annotations. Expert dermatopathologists annotated part of the available WSIs in a freehand manner directly on Pixnormous. The overall slides annotation masks were then generated following the methodology presented in the previous subsection.

Real-conditions AL pipeline. To put these experiments into practice, we also relied on Pixnormous. Specifically, the model-related functions were previously embedded into the software to be directly linked to the front-end buttons on the visualisation tool for segmentation mask prediction, review validation, and model retraining.

The overall pipeline, developed in real-world conditions, is depicted in Figure 5.1. Specifically, all WSIs were first uploaded to the software in order to be visualised, and in some cases, annotated, while their pre-processing was run in the background at the same time. Once uploaded, an expert pathologist was invited to annotate a random set (S_1^2) of 76 initial images for training and run the first model training directly through the Pixnormous platform. Once the first model θ_1 trained with the ground truth annotations subset S_1^a , the sample selection process was performed in the back-end before querying the pathologist with the selected images for expert review. The user could then use the front-end tool to review masks and run the model retraining, selecting the specific AL configuration a drop-down menu: ground truth, active, or active self-training.

Note that, in addition to the 76 initial images for training in the first iteration, 15 additional images were also randomly selected from the 203 available tissue fragments for testing purposes.

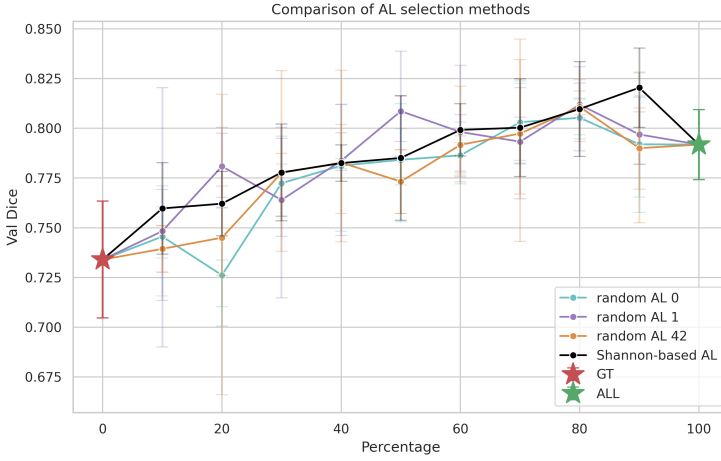


Figure 5.3: Weighted Dice score of sample selection strategies for AL over the iterations. The lighter curves (blue, purple, orange) represent different seeds for random selection, while the black curve corresponds to the Shannon entropy-based selection.

5.5 Results

5.5.1 Ablation experiments

In this Section, we conducted extensive ablation experiments to evaluate different AL training configurations and identify the optimal conditions for reducing the annotation burden for pathologists.

(1) Ground truth baseline. As a comparison baseline, we trained a reference model using only the 76 WSIs annotated by expert pathologists, serving as the benchmark for comparing the effectiveness of the subsequent AL strategies. This ground truth-based model reached a weighted Dice coefficient of 0.7340 across all folds, represented by the green star in Figure 5.3 and Figure 5.4.

(2) Random selection with different seeds. As a first step in comparing the AL training configurations, we implemented the AL scenario under a random selection strategy. In this setup, 10% of the 112 remaining WSIs was selected iteratively until reaching 100% of the dataset. To assess the impact of different random seeds on AL training, we used three distinct seeds, maintaining the same seed across all iterations for a given experiment. All iterations were trained from

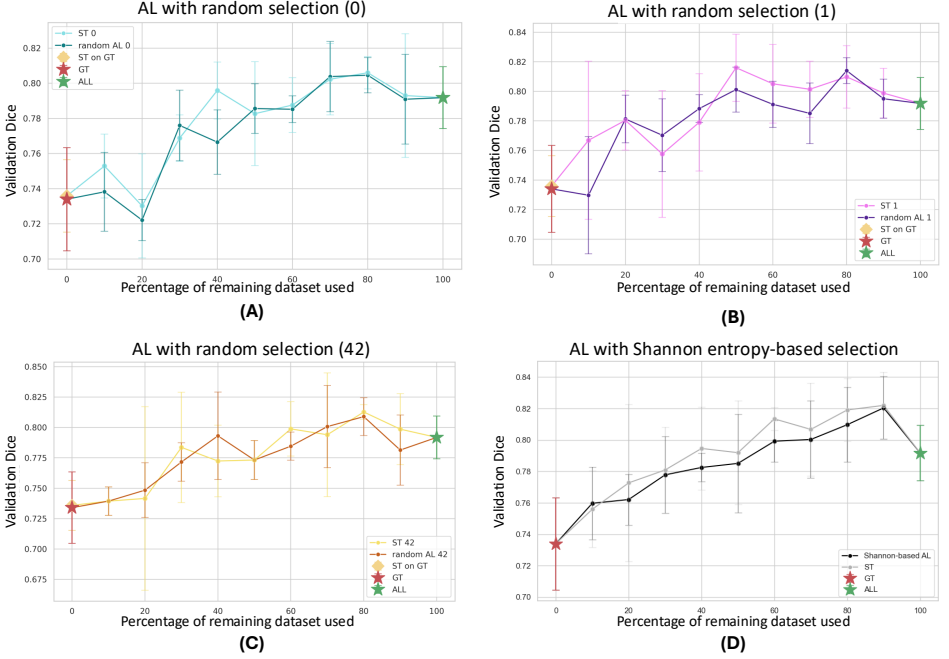


Figure 5.4: Comparison of weighted Dice scores for AL strategies with varied seeds and sample selection strategies, with lighter curves corresponding to the ST approach at each step. (A-B-C): AL with ST for three different random seeds for sample selection; (D): AL with ST using the Shannon entropy for informed sample selection.

scratch to allow for a more representative comparison of the model’s performance at each point.

Results are illustrated in Figure 5.3, with the coloured curves (blue, purple, orange) representing the different random seeds. The coloured curves demonstrate overall improvement in performance across iterations, with noticeable fluctuations in performance across iterations.

(3) Informed selection. To compare the different AL configurations, we also employed an informed selection strategy based on Shannon entropy. Specifically, at each iteration, the resulting model was used to generate the mask predictions (pseudo-masks) for the remaining unannotated samples. The Shannon entropy was then calculated on

these samples to identify the most uncertain predictions, which were included in the next iteration as expert annotations after being reviewed and modified by an expert pathologist.

The results, depicted by the black curve in Figure 5.3, demonstrate a more consistent and steady rise in the Dice coefficient compared to the random selection methods overall. As from the first iteration, the informed selection showed better performance with the representative selected data in comparison with other methods. Although random selection occasionally outperformed the informed approach in certain iterations, the informed method generally maintained a steadier performance improvement overall.

(4) Self-training comparison. At each iteration, a self-training model was trained combining the annotated samples with the pseudo-annotations generated by the current model, considering them as pseudo ground truth to feed the model with more data. The results are depicted in Figure 5.4 for each random seed and the Shannon-based approach. While the self-training on ground truth data alone did not enhance much the model performance, achieving a Dice coefficient of 0.7359 compared to the baseline of 0.7340, Shannon entropy-based models exhibited small but consistent gains when combined with self-training across most iterations. Conversely, the benefit of self-training for random selection-based models was notably less steady. This highlights that the most stable baseline models, such as those based on Shannon entropy, exhibited the most steadier improvement with self-training complements, as shown in Figure 5.4-(D).

5.5.2 Interobserver study

Interobserver study

To further explore the role of AL in digital pathology and its practical implications, we conducted an interobserver study comparing annotations on the test set from multiple expert pathologists with those produced by our models. Specifically, we asked an additional expert pathologist to annotate some of the test set images, hence resulting in a

subset of images annotated by two different pathologists—pathologists A and B. To quantify the level of agreement between different annotations, we resorted to Cohen’s Kappa (κ) score, a widely used metric to measure inter-rater reliability. Specifically, Cohen’s Kappa is defined as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o represents the observed agreement between two raters, and p_e corresponds to the expected agreement by chance.

For this interobserver study, we evaluated models at an early stage of the AL process to effectively compare the impact of different selection methods. Specifically, we analysed model performance at the third iteration, corresponding to 30% of the remaining training data. We compared models trained with the three different random seeds (0, 1, and 42), with the model trained using Shannon-based selection at the same iteration. The predicted segmentation masks and their corresponding κ scores are depicted in Figure 5.5.

The results first highlight the inherent variability in histopathology annotations: the agreement between the two expert pathologists varies considerably across images, with κ values ranging from 0.50 to 0.92. This establishes an important baseline, as model-pathologist agreement should be interpreted in the context of inter-pathologist variability.

Notably, the choice of random seed had a substantial impact on model performance. Random seed 1 (column D) consistently achieved lower agreement scores compared to random seeds 0 and 42 (columns C and E, respectively), demonstrating the instability of random selection strategies. This variability is particularly evident in rows 1, 3 and 4, where seed 1 shows markedly poorer agreement, and even negative agreement in row 1.

In contrast, the Shannon entropy-based selection method (columns F and G) demonstrates more stable and consistent performance across all images. Although some model predictions reach substantially lower agreement for challenging cases, as observed in row 6, in other cases,

the models achieve a level of agreement with the pathologists that is comparable to the agreement observed between the pathologists themselves. For instance, in rows 2 and 3, most models achieve κ values within the range of inter-pathologist agreement. Notably, in row 3, the Shannon-based models and random seeds 0 and 42 even surpass the inter-pathologist agreement ($\kappa = 0.90$), achieving $\kappa > 0.91$, suggesting that in some cases, well-trained model predictions can be as consistent as (or even more consistent than) expert annotations.

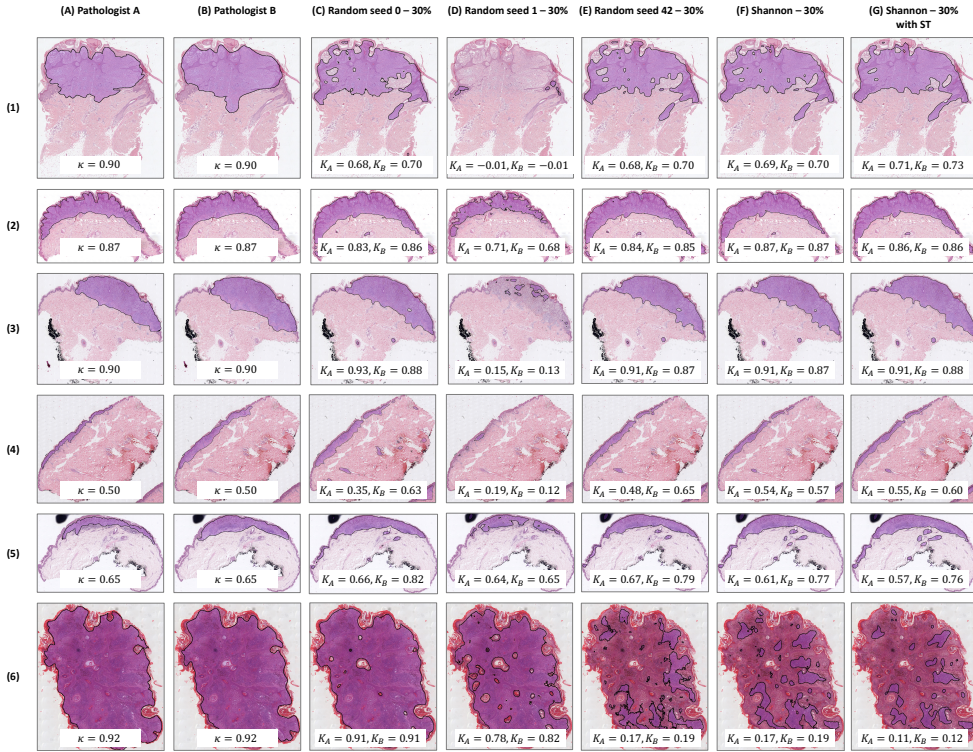


Figure 5.5: Comparison of the annotations made by different pathologists vs. our model for a selection of test tissue fragments. (A) Annotations made by pathologist A; (B) annotations made by pathologist B; (C-E) annotation masks predicted on the third iteration by random selection-based models trained only with expert annotations; (F-G) annotation masks predicted by the Shannon-based active learning model on the third round, trained with expert annotations (E) together with predicted pseudo-masks (F). κ : Kappa score between pathologists; κ_A : Kappa score of the model compared with pathologist A’s ground truth annotation; κ_B : Kappa score of the model vs. pathologist B’s mask.

5.6 Discussion

The obtained results highlight the ability of AL to perform well starting from early training stages, with models using Shannon entropy-based selection strategies achieving more consistent performance than random selection methods. Despite the relatively small dataset used in this study, the informed sample selection method proved effective in prioritising the most informative data for training.

One important consideration in this study is the impact of random seeds. The variability in model performance observed across different random seeds underscores the sensitivity of AL models to the specific composition of the training set. This effect is further evidenced by the variation observed in the 4-fold cross-validation, which highlights that performance can vary significantly between different training subsets. While random selection strategies showed fluctuating performance, models based on Shannon entropy selection demonstrated a more stable and consistent improvement, suggesting that informed sample selection helps mitigate some of the variability in random seed initialisation. In addition, while the self-training strategy did not always enhance model performance, it showed more stable improvement over iterations when combined with Shannon entropy-based AL.

Beyond model variability, a key challenge in histopathological annotation is the inherent variability in interobserver agreement, even among expert pathologists. Although pathologists generally exhibited high agreement, some samples showed lower levels of concordance, highlighting the difficulty in achieving consistency across complex or ambiguous cases. In contrast, models trained with specific random seeds or Shannon entropy-based selection achieved agreement levels comparable to, and in some cases exceeding, inter-pathologist agreement. This interobserver variability underscores the importance of validating AI models against multiple annotators rather than a single reference standard, ensuring consistency and reliability in real-world applications. Therefore, calculating and accounting for interobserver variability becomes crucial for ensuring that AI models generalise well across different clinical environments and minimise potential biases.

Although this study serves as a baseline for applying AL in histopathology segmentation, several aspects must still be considered. First, the dataset used in this study is relatively small and should be extended to allow for more reliable results across a wider range of cases. Another consideration for future work is exploring ways of correlating the uncertainty in model predictions with interobserver variability, providing insights into the reliability of expert annotations and that of models before their clinical deployment. Additionally, future work will also extend this framework to histopathological classification, investigating both AL's performance in classification tasks and the variability across annotators in these tasks.

5.7 Conclusions

This study demonstrates the potential of AL in supporting histopathology segmentation from early steps of the training process while reducing annotation burden, offering a promising framework for real-world clinical environments. By comparing various AL configurations and sample selection methods, we highlighted how AL can effectively prioritise the most informative data, improving model performance from early stages and ensuring more consistent results compared to random selection methods.

The results underscore the importance of validating AI models across multiple annotators, as variability in interobserver agreement remains a challenge in histopathology. The ability of the AL framework to match, and at times surpass, the agreement between expert pathologists highlights the potential of AL to support pathologists in making more reliable and consistent diagnoses. This finding calls for further research into the integration of multi-annotator consensus to address the inherent uncertainty in ground truth annotations.

What happens next? Artificial intelligence in Spitz tumour research

While previous chapters focused on artificial model development and systems, this chapter situates those efforts within the broader research context of artificial intelligence for diagnostically ambiguous Spitz tumours, synthesising current knowledge and outlining future directions.

*The content of this chapter is based on previously published work, which has been modified and adapted to fit the scope of this thesis:
Launet, L., Colomer, Mosquera-Zamudio, A., Monteagudo, C., Naranjo, V. The puzzling Spitz tumours: is artificial intelligence the key to their understanding? Histopathology (2025).*

Contents

6.1	Introduction	86
6.2	Latest clinical advancements: Spitz or spitzoid?	86
6.3	Towards unmasking Spitz tumours with AI?	88
6.3.1	Clinicopathological tabular data	88
6.3.2	Histopathological analysis	92
6.3.3	Genetic analysis	94
6.4	Roadblocks to AI in Spitz tumour analysis	97
6.4.1	Ambiguity in clinical guidelines	97
6.4.2	Limited prognostic information	97
6.5	Future directions	98
6.5.1	Prognostic data: public databases	99
6.5.2	Foundation models	99
6.5.3	Hybrid approaches	100
6.5.4	Cross-speciality collaborations	101
6.5.5	Retrieval systems	101
6.6	Conclusion	102

6.1 Introduction

While the previous chapters of this thesis have focused on specific technical and methodological challenges related to the development of AI models for WSI analysis in diagnostically ambiguous Spitz tumours, this chapter shifts the focus to a broader research and clinical context. In parallel with the work presented here, an increasing number of studies have begun to explore the application of AI to Spitz tumours from different perspectives in recent years.

Despite growing interest in applying AI to ambiguous melanocytic lesions, the literature remains fragmented, and often overlooks the specific complexities posed by borderline lesions such as Spitz tumours. But, in this new era of AI-powered medicine, what hope is there for a better understanding of the mysterious Spitz tumours?

In light of these limitations, this chapter situates this work within the broader landscape of AI research related to Spitz tumours, identifying existing challenges, opportunities, and areas requiring further investigation. Based on this analysis, it proposes a set of future directions and guidelines to support the development of clinically robust and interpretable AI systems suited to the diagnostic ambiguity of these lesions.

6.2 Latest clinical advancements: Spitz or spitzoid?

Since 2018, the World Health Organisation (WHO) classification of skin tumours [90, 94] integrated molecular alterations into the characterisation of Spitz neoplasms, introducing a novel distinction between Spitz melanoma and spitzoid melanoma. This classification defines Spitz tumours as melanocytic tumours with typical spitzoid morphology that possess specific genetic alterations, namely receptor kinase fusions or HRAS mutations. In contrast, this updated categorisation excludes neoplasms with BRAF or NRAS mutations from the Spitz category as they are not genetically related [95], thus redefining them as spitzoid tumours (Spitz-like), depicted in Figure 6.1.

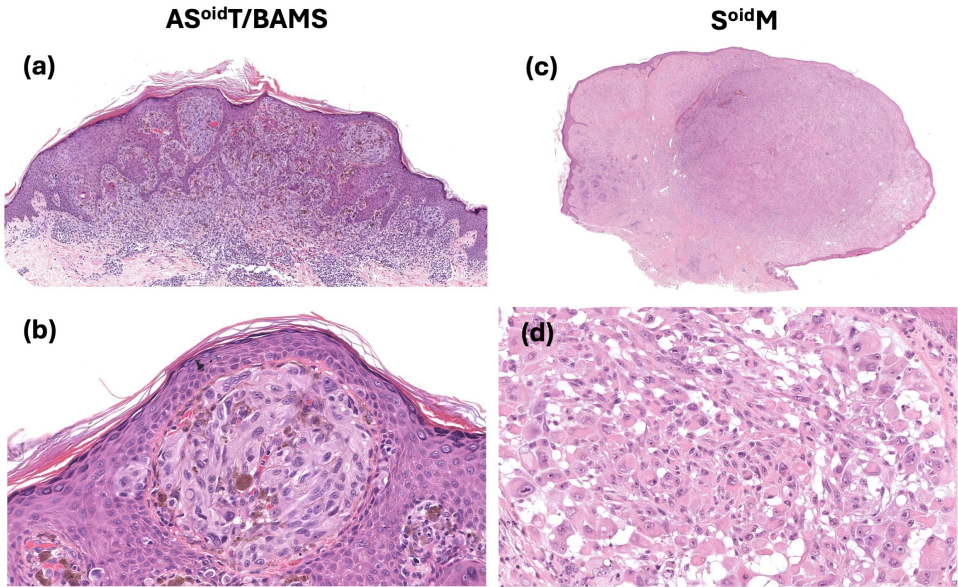


Figure 6.1: Histopathological findings of spitzoid melanocytic tumours. (a-b): BRAF mutated atypical spitzoid tumour (ASoidT/BAMS) showing symmetry (a) and large uniform melanocytic cells (b) (c-e); Spitzoid melanoma (SoidM) with an expansile dermal nodular component (c), and large cells with marked nuclear pleomorphism (d).

With the integration of this genetic information, the categorisation of these challenging melanocytic lesions thus divides them into two global categories based on the mutations they harbour: Spitz tumours and spitzoid tumours. Each category can then be further subclassified into the three initial classes based on their histopathological features, e.g., Spitz naevus (SN), Spitz melanoma (SM), and atypical Spitz tumour (AST), a term that has recently gained popularity as Spitz melanocytoma [31]. The genomic subtypes of Spitz melanocytoma are considered similar to those of Spitz naevus as they carry initiating gene fusions or HRAS mutations found in Spitz naevi, yet secondary alterations to help distinguish them from Spitz tumour and SM remain to be characterised [94]. More recently, atypical spitzoid tumours with BRAF mutations that do not meet melanoma criteria have been categorised as BRAF-mutated

and morphologically spitzoid (BAMS) [96]. An example is depicted in Figure 6.1-(a-b).

Despite this updated categorisation of Spitz tumours, a certain risk of progression is acknowledged for each category. Even Spitz naevus, which is considered to have a minimal risk, still carries some potential for progression [97].

6.3 Towards unmasking Spitz tumours with AI?

With the rise of AI technologies in digital pathology, some studies have explored various approaches to analysing Spitz tumours across different data modalities. These approaches range from simulating real clinical conditions with tabulated features extracted from pathologists' analysis, to leveraging the digitised WSIs or incorporating genetic information. This Section provides a detailed overview of AI-based studies involving Spitz tumour data, with a summary of their characteristics presented in Table 2, and their typical characteristics simplified in Figure 6.2.

6.3.1 *Clinicopathological tabular data*

As depicted in the upper part of Figure 6.2, examining clinical and histopathological variables directly extracted by pathologists can provide essential insights that complement histopathological assessments, while reproducing settings closely aligned with clinical practice, particularly in laboratories where access to advanced scanners may be limited.

In [98], Belaala et al. applied AI methods to analyse STs for the first time. Specifically, they proposed to integrate clinical, histological and immunohistochemical variables for distinguishing between SN and AST, using a dataset with 29 attributes for each sample: 3 clinical (gender, localisation, age), 20 histological, and 6 immunohistochemical (P16, Ki-67, BRAF, ALK IH, ALK FISH, melanin pigmentation), each represented numerically. The authors employed a genetic algorithm to manage high-dimensional data and select the most relevant features for the classification stage. The genetic algorithm consistently selected features such as localisation, cytonuclear atypia, atypical mitosis,

Table 6.1: Studies included in the review, applying artificial intelligence algorithms to the analysis of Spitz tumours.

Study	Year	Studied classes	Dataset size	Data modality	Magnification	AI method	Advantages	Disadvantages
Belaala et al. [98]	2021	SN and AST	54 (7 AST, 47 SN)	Tabular (clinical, histological, immunohistochemical)	-	ML algorithms and MLP	First ML-based approach for ST analysis	Not considering SM in the study nor considering mutation-based categorisation
Mosquera-Zamudio et al. [35]	2024	SN, SM and AST integrating the updated WHO [94] categorisation	122 (39 SM, 39 SN, 44 AST)	Tabular (clinical, histopathological)	-	ML algorithms	Incorporated mutation-based categorisation, analysed histological features' relevance	Manual feature analysis
Del Amor et al. [99]	2021	SN vs SM	51 (21 SM, 30 SN)	WSI	10x	MIL classification with attention	Explainability through attention maps	Did not consider mutation-based categorisation
Launet et al. [91]	2022	SN vs SM	84 (34 SM, 50 SN)	WSI	10x	U-Net segmentation and MIL classification with attention	Leverages both annotated and unannotated samples	mutation-based categorisation, reduced validation data
Del Amor et al. [100]	2022	SN vs SM	43 (16 SM, 27 SN)	WSI	5x, 10x, 20x	Multi-resolution fusion	Integrates multiple resolutions	Patch-level only
Tabatabaei et al. [101]	2024	SN vs SM, AST	126 (34 SM, 50 SN, 42 AST)	WSI	10x	Siamese networks for content-based image retrieval	Similarity-based approach	Did not consider mutation-based categorisation, no analysis of AST prognosis predictions
Del Amor et al. [102]	2021	SN vs SM, AST	39 (8 SM, 12 SN, 19 AST) with 491 CpG sites	DNA methylation level	-	Autoencoder dimensionality reduction and deep clustering	Leverages DNAm insights	Very little data
Lang et al. [103]	2020	SN, AST and SM	68 (21 SN, 21 AST, 26 SM) and independent validation cohort of 56 additional cases	Tabular (clinical and histological)	-	ML algorithms	Reduced feature set, AST prognosis	Did not consider mutation-based categorisation

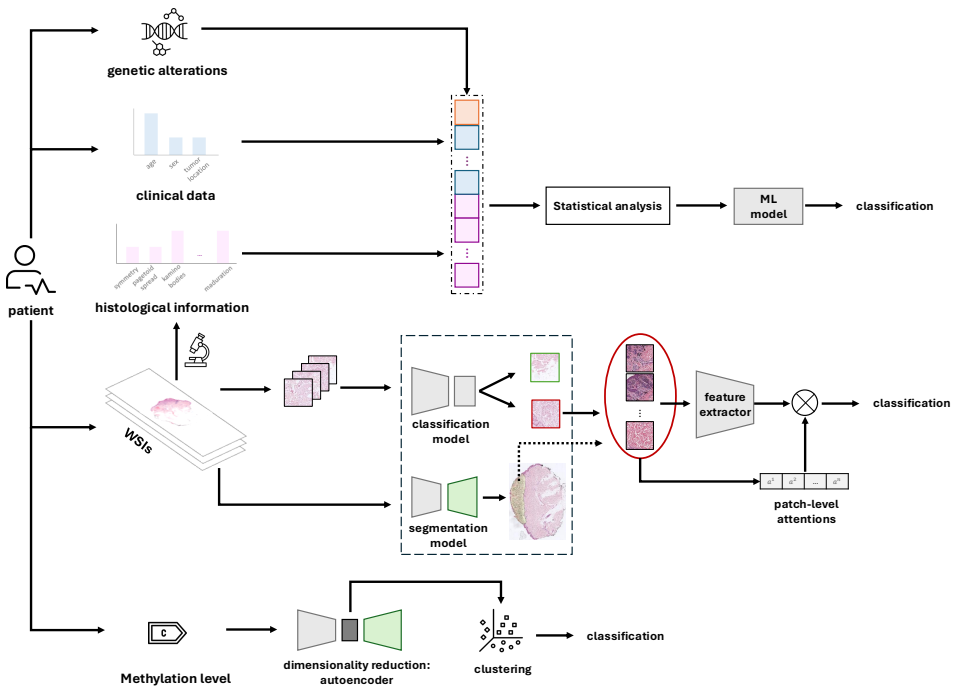


Figure 6.2: Simplified overview of typical artificial intelligence approaches in Spitz tumour analysis depending on the data type. Top: approaches leveraging tabular data; centre: WSI-based solutions; bottom: DNA methylation-based approach.

asymmetry, and Ki-67. Different ML algorithms were then evaluated for the classification step, with the Multilayer Perceptron (MLP) achieving the best performance. Using 14 features selected by the genetic algorithm, this approach reached an accuracy of 0.98 and an F1-score of 0.99 for SN vs AST classification.

In contrast, the study presented by Mosquera-Zamudio et al. in [35] compared various ML algorithms to assess their effectiveness and generalisability in distinguishing between different ST categories simultaneously (SN, AS^{oid}T and S^{oid}M), as well as specific subcategories such as benign and malignant spitzoid tumours. Trained on the more than 20 representative histological variables retrieved from pathologists' slide analyses, tabulated alongside clinical variables (age at diagnosis, sex, and tumour location), the results demonstrated that ML models are capable of effectively distinguishing between certain categories of Spitz tumours. However, the morphological complexity of these lesions was challenging for other classes, such as in differentiating between all three categories due to the overlapping morphological characteristics, thereby highlighting the resemblance between AST and SM.

In addition, the analysis of variance (ANOVA) performed enabled to rank the clinicohistopathological variables used for ST diagnosis according to their contribution to the final diagnosis, revealing consistent trends in the importance of specific variables across experiments. Specifically, variables such as atypical mitosis, atypia/pleomorphism, and mitosis/mm² were consistently found to be relevant for the diagnostic decision, whereas variables like patient sex and pigmentation showed less relevance.

Besides highlighting the inter-observer variability observed in ambiguous tumours, the findings from this study underscore the ongoing need for further exploration of STs. This suggests that not all clinicopathological features initially considered for ST diagnosis may be universally relevant across all Spitz tumours, and their importance may likely vary depending on the specific category under study.

6.3.2 Histopathological analysis

Histopathological image analysis remains the gold standard for diagnosing STs, especially using Hematoxylin and Eosin (H&E) stained slides. In light of the rise of AI methodologies in the study of WSIs within digital pathology, various studies have applied techniques tailored to this type of data, as depicted in Figure 6.2-centre.

- **Multiple instance learning**

Due to the vast size of WSIs and the time-consuming annotation process for pathologists, the general approach for classifying STs with H&E WSI data usually involves a MIL scenario, extracting features at patch level before combining them using an aggregation mechanism.

In [99], the authors proposed an attention-based MIL framework for ST diagnosis. Building on a VGG16 [104] architecture for the feature extractor, they trained a first model at a patch level to identify tumour patches and select regions of interest (ROIs), and refined it with Squeeze-and-Excitation blocks [105] to enhance contextual information for tumour patch identification. For WSI-level classification, the model was trained under a MIL configuration, combining patch embeddings with an attention-based trainable aggregation function based on Ilse et al.'s attention mechanism [106]. The results obtained in this approach showed an accuracy and F1-score of 0.92 and 0.89, respectively, for the source model, and 0.80 and 0.73 for the classification model. In addition, class activation and attention maps were generated to provide additional insights to the inner model workings, aiming to enhance explainability for pathologists.

In [91], a two-step methodology was also proposed to try to mimic how pathologists examine biopsies, first localising the tumour area from a low magnification view before focusing on its features at a higher magnification level. To do so, the authors first trained a fully-supervised U-Net-based segmentation model [81] with residual blocks [92], on resized tissue fragments which they used to generate the pseudo-annotations of tumorous areas for unannotated samples. Leveraging both the ground truth ROIs and pseudo-ROIs, the authors extracted patches from tumorous areas only at a 10× magnification level, to feed a MIL

classifier with attention-based MIL pooling. With this approach, the segmentation model reached a Dice score of 0.92 and the top-performing classification model obtained an accuracy of 0.85 and F1-score of 0.84.

- **Multi-resolution**

WSIs enable to store multiple levels of magnification similar to those available in microscopes, thus facilitating access to different levels of details. While lower resolutions provide contextual information about the lesion under study, higher resolutions offer detailed morphological features at the cellular level. Typically, approaches select specific magnification levels tailored to task requirements, but combining multiple resolutions could integrate information across varying magnification levels, thus enriching the analysis with a comprehensive range of detail.

Following the reviewing strategy of pathologists, Del Amor et al. [100] introduced a multi-resolution framework that integrates the perspectives from $5\times$, $10\times$ and $20\times$ magnification levels. Initially, the authors separately pre-trained the network at each resolution to determine the optimal backbone architecture and optimize hyperparameters using single resolutions. Subsequently, they established a multi-resolution framework with three CNN feature extractors, each corresponding to one resolution and initialised with the top-performing single-resolution model weights through transductive learning. The feature maps extracted from each CNN were then combined using a late fusion approach, where they were concatenated to gather information across multiple magnification levels, before employing a projection head for the final prediction.

In comparison, the study found that the multi-resolution framework significantly outperformed the top-performing single-resolution model ($5\times$), achieving an accuracy of 0.81 compared to 0.56 and an AUC of 0.83 compared to 0.55.

- **Content-based image retrieval**

In particularly challenging tumours like Spitz tumours, CBIR methodologies bring the opportunity of supporting pathologists in the decision-making process by providing them images of similar Spitz tumour cases for reference.

For the first time, a CBIR system designed to retrieve the most similar cases to provide a potential prognosis for AST samples was introduced in [101], utilising Siamese networks on patches solely extracted from the ROIs. In a sense, Siamese networks could be considered as comparator networks since they aim to obtain similarity scores between two input images. They consist of two identical CNNs running in parallel, sharing the same parameters and weights, each working as feature extractors to generate representative feature embeddings from their respective input images.

In this study, the authors trained the Siamese networks on benign and malignant Spitz tumour samples, aiming to minimise the distance between feature embeddings from the same class while maximising the distance between embeddings from different classes (benign vs malignant) during training. When computing the similarity between input pairs, the proposed approach achieved an F1-score of 0.64 and 0.94 for the SN and SM classes in the top-3 scenario, respectively, and an overall accuracy of 0.89. The authors then proposed to apply this framework to AST samples to assess their similarity to benign vs malignant ST samples in the feature database, and provided gradient-weighted class activation maps (Grad-CAMs) to provide interpretable details for these complex cases.

6.3.3 Genetic analysis

Given the challenging morphological appearance of Spitz tumours and the discrepancy with their clinical evolution, some studies proposed to consider the molecular, biomarker and genetic information to improve the algorithms based on histological features only. Gene fusions and mutations, copy number variation evaluated by comparative genomic hybridisation (CGH) or fluorescence in situ hybridisation (FISH), DNA methylation (DNAm) analysis, and immunohistochemical studies targeting biomarkers like p16, HMB45, and Ki-67, provide valuable insights that may complement histopathological data, particularly for these challenging tumours.

However, the availability of data for these approaches can vary significantly due to factors such as accessibility, cost, technical expertise and the rarity of Spitz tumours. Therefore, the development of AI algorithms that integrate this valuable information remains limited, despite its potential.

- **Mutations**

Considering latest advancements in the categorisation of Spitz neoplasms, which incorporate genetic alterations to distinguish between Spitz and spitzoid tumours [94], integrating these molecular insights into AI-based studies might become necessary. Acknowledging the significance of these biomarkers, Mosquera-Zamudio et al. [35] gathered information on BRAF and NRAS mutation status to differentiate between Spitz and spitzoid tumours. However, despite efforts to obtain immunohistochemical features such as HMB-45, Ki67 and p16 when available, they were unable to include this information due to the limited availability of samples containing this information.

- **Molecular markers**

Primary cilia (PC) are sensory organelles crucial for cellular processes like proliferation, differentiation, and development. The ciliation index (CI), which measures the percentage of ciliated melanocytes [107], has emerged as a significant molecular marker for distinguishing melanoma (lacking primary cilia) from melanocytic naevi (retaining primary cilia). Building on this distinction, Lang et al. [103] investigated the diagnostic utility of CI for ambiguous lesions, particularly STs. They analysed 14 clinicohistopathologic features tabulated alongside the CI for each case using a random forest classifier to evaluate the significance of each feature. Additionally, they utilised a logistic regression model to determine the value of CI when combined with other histopathologic and molecular features. The results consistently highlighted a low CI as a top predictive feature for malignancy, and identified other significant features such as asymmetry, hyperchromatism, and cytologic atypia.

While the logistic regression model trained on CI alone reached an area under the curve (AUC) of 0.84, it achieved an AUC of 0.92

when incorporating the top four predictive features (asymmetry, CI, hyperchromatism, and cytologic atypia).

• DNA methylation

As DNA methylation (DNAm) frequently manifests in melanocytic tumors and is directly related to their development and metastatic progression, it is considered a valuable biomarker for diagnostic, prognostic, and therapeutic purposes. Specifically, DNAm involves the methylation of CpG sites, which are regions of DNA where cytosine nucleotides are followed by guanine nucleotides, and play a crucial role in regulating gene expression and tumour behaviour.

Building on this understanding, Del Amor et al. [102] introduced a novel approach to analyse Spitz tumours using DNAm data, aiming to differentiate between naevus and melanoma and, for the first time, utilised this data to propose a potential prognosis for ASTs. Since DNAm analysis of CpG sites generates a large amount of data relative to the typically limited samples available, the authors were able to reduce the 491 DNAm features for each sample to 7 through initial statistical analysis and dimensionality reduction.

The experiments conducted in this study demonstrated better performance for the unsupervised deep clustering approach, which simultaneously refines the latent space and cluster centroids during training. This method achieved promising results, with an accuracy of 0.90 and an F1-score of 0.88. Additionally, the authors utilised the top-performing encoder output, trained with malignant and benign samples, to compute the embedded space for patients initially diagnosed with AST. The Euclidean distance was then calculated and fed into a softmax function to calculate the class probabilities for each AST sample. The results suggested a predominantly benign prognosis for AST samples (63.13%), as their features closely aligned with those of patients labelled as benign.

6.4 Roadblocks to AI in Spitz tumour analysis

Despite the promising potential of AI in advancing Spitz tumour analysis, its application in this area remains limited. The complexity in categorising and understanding Spitz tumours, combined with their low incidence, significantly contributes to the challenges faced in this field.

6.4.1 *Ambiguity in clinical guidelines*

Since their first description, the morphological and clinical understanding of Spitz tumours has been subject to continuous refinement and re-evaluation. However, these changes have not always been uniformly interpreted and adopted in clinical settings, resulting in variations in how Spitz tumours are diagnosed across laboratories. In fact, only one study applying AI to Spitz tumours [98] integrated immunohistochemical data supporting their categorisation into their approach (see Table 6.1).

Additionally, the complexity of these tumours and ambiguity in the clinical guidelines contribute to substantial interobserver variability among pathologists and even intraobserver variability, where a single pathologist's diagnosis may change over time [108]. These inconsistencies might lead to datasets with ambiguous and noisy labels, posing a challenge for training accurate and robust AI models.

6.4.2 *Limited prognostic information*

When considering the training requirements of AI models, they are widely recognised for their need for large datasets to achieve better generalisation and more robust outcomes. However, as reflected in Table 6.1, studies focusing on ST analysis have utilised relatively small datasets, ranging from only 39 to 122 samples, as shown in Table 6.1. This scarcity can be attributed in part to the rare incidence of such tumours.

While resources like The Cancer Genome Atlas (TCGA) and open challenges such as PANDA [109] and BreakHis [110] provide thousands of WSIs publicly for other tumour types, only one WSI dataset has

been specifically released for Spitz tumours, the SOPHIE dataset [111]. This dataset, collected over more than 30 years, contains a total of 61 H&E WSIs from 58 patients, with 30 patients diagnosed as SN, 18 as SM, and 10 as AST. In addition to the limited dataset sizes, the lack of comprehensive annotation data for the vast WSIs presents further challenges for this modality in general.

Furthermore, the absence of prognostic data detailing patient evolution over the years for AST samples limits the ability to draw conclusions about the malignant potential of these ambiguous tumours. Studies such as [101] and [102] have proposed possible outcomes for these tumours, or to divide their categorisation into low-grade vs high-grade ASTs [35], but the lack of long-term follow-up data makes it difficult to confirm the actual outcomes and malignancy risk.

As per the latest WHO classification [94], integrating specific mutation information is crucial for distinguishing between Spitz tumours and spitzoid tumours based on their genetic alterations. However, this information is still generally unavailable, except for two studies [35, 98], as highlighted in Table 6.1. In clinical practice, the high costs and logistical challenges associated with genetic and immunohistochemical studies often result in these tests not being conducted.

6.5 Future directions

Amidst the considerable diagnostic challenges that Spitz tumours have brought along since their first description, initial AI applications for Spitz tumour analysis have demonstrated their potential, although currently limited by some challenges. This Section aims to propose strategies to address current limitations and advance the landscape of Spitz tumour analysis.

6.5.1 Prognostic data: public databases

The first step in advancing Spitz tumour analysis with AI involves the acquisition of comprehensive datasets. Recently, the release of the first public WSI dataset dedicated to these tumours, SOPHIE [111], marked a significant milestone. The field requires more extensive public repositories, including WSIs or IHC data but also incorporate molecular information to refine their categorisation and integrate additional data types. Access to hospital archives with data on AST cases could also provide valuable prognostic information, aiding in the identification of patterns in long-term clinical evolution.

In the context of collecting extensive datasets, privacy legislations in the medical field often pose barriers to data sharing across institutions. A promising solution lies in the implementation of broad federated learning (FL) frameworks [112], which ensure data privacy by keeping sensitive data local while enabling collaborative model training by sharing model weights across decentralised datasets. In the digital pathology field, the FL paradigm remains in its infancy, necessitating efforts to democratise FL frameworks for its adoption outside research settings [113].

6.5.2 Foundation models

Recently, the field of AI has witnessed a significant evolution from specialised systems confined to narrow tasks to the emergence of foundation models (FMs), which were recently successfully applied to digital pathology [114]. Trained on extensive and diverse data, these models offer powerful data representations that enhance adaptability in limited data scenarios, such as zero or few-shot learning. By leveraging this broad training data, FMs like PLIP [115], CONCH [116], TITAN [117] and KEEP [118] have proven to be effective feature extractors. With promising initial results on rare diseases using zero-shot approaches, FM-based feature extractors hold potential for enhancing applications in Spitz tumours.

6.5.3 *Hybrid approaches*

While single-mode approaches have demonstrated promising results, combining multiple data modalities could provide deeper insights into the biological mechanisms and clinical behaviour of these tumours. Most studies presented so far often overlooked this aspect, focusing on either histopathological features [91, 99, 100] or genetic data [102] independently.

Integrating morphological patterns extracted from WSIs with molecular data has proven beneficial in identifying new biomarkers associated with various diseases, such as melanoma [119] or breast cancer [120]. This strategy is referred to as multimodal learning, which involves integrating multiple data modalities or sources to achieve a more comprehensive understanding of disease mechanisms and clinical outcomes. In parallel, hybrid models refer to the combination of several AI architectures, which can involve one or more data modalities.

These different approaches can be implemented through various strategies. One common strategy involves analysing each modality separately, using AI models tailored to extract relevant features from each modality independently, before combining them using a late fusion approach in a later stage. Another approach is to consider modalities simultaneously from the beginning, allowing for joint analysis of their combined features, under an early fusion mechanism.

Combining WSI features with biomarker information and genetic data using multimodal (hybrid) models could facilitate the development of comprehensive analyses to explore potential correlations between histopathological features and molecular characteristics. Such approaches could provide valuable insights into how biomarker expressions may influence the morphological presentation of Spitz tumours.

6.5.4 *Cross-speciality collaborations*

Advancing AI-based ST analysis and digital pathology in general requires more than technical innovation. In fact, future directions should focus on fostering collaborative ecosystems, building AI systems tailored to the needs of medical experts.

Such collaborations should start with integrating pathologists into the AI model pipeline [121], providing them with feedback and soliciting their input on specific aspects. This feedback goes with ensuring that the developed models are interpretable for end-users, particularly in the clinical field where trusting black-box models is risky. For image data, explainable techniques such as class activation maps (CAMs) or attention maps can be used to highlight which regions of an image were most influential in the model's prediction, providing pathologists with visual cues on where the model focused its attention [99]. For models handling numerical features, statistical analysis [35] or SHapley Additive exPlanations (SHAP) [122] can be employed to show which features were most critical in the decision-making process.

Cross-speciality collaborations could also involve crowdsourcing approaches, which involve distributing data labelling tasks among a diverse group of contributors. While most studies have utilised noisy non-experts WSI labels in their approach [123], future directions for Spitz tumour analysis could leverage expert labels. Indeed, since expert labels tend to be noisy by nature in Spitz tumour analysis due to their clinical ambiguity, incorporating their collaborative input could help reduce interobserver variability and contribute to training more robust models.

6.5.5 *Retrieval systems*

In addition to supporting pathologists in a complementary and informative manner, CBIR systems can help find similarities across Spitz tumour samples. While most CBIR approaches have concentrated on single modalities, such as WSI embeddings in histopathology [101], future CBIR systems should adopt multimodal strategies to gain deeper insights into Spitz tumours.

For instance, integrating WSI embeddings with genetic data could help uncover relevant similarities between samples. One such example of multimodal retrieval system is the PLIP model [115], which enables retrieval of similar image cases through both text or image input queries. Additionally, CBIR systems for Spitz tumour analysis would greatly benefit from incorporating distinctions between spitzoid and Spitz tumours, as well as their subclasses. This integration would enable the models to evaluate whether morphological patterns vary according to specific mutations in Spitz tumours.

6.6 Conclusion

Although the application of artificial intelligence to the analysis of spitz(oid) tumours remains relatively under-explored, it has shown significant potential in refining diagnostic accuracy by analysing key diagnostic features. However, to fully leverage these advancements, future research approaches need to integrate the latest updates in spitz(oid) categorisation and incorporate mutation information. In addition, given the complexity and biological significance of these tumours, adopting multimodal strategies will be essential to gain deeper insights and identify meaningful correlations to enhance their understanding. Achieving these goals will ultimately depend on fostering collaborative efforts among researchers, clinicians, and technologists to maximise the potential and impact of artificial intelligence in this field.

Final conclusions

This chapter relates the conclusions to the final aim of this Ph.D. thesis, summarising concluding remarks and suggesting future research lines.

7.1 Global remarks

Bringing artificial intelligence (AI) closer to clinical reality in pathology requires more than accurate models. It also involves practical considerations that ensure their applicability and relevance in real-world settings. To address both these technical and practical challenges, in this thesis we have designed, developed and validated deep learning (DL) systems and methodologies tailored to the complexities of digital pathology, with a particular focus on Spitz tumours. Specifically, we investigated multiple instance learning (MIL) to address the complex nature of whole slide images (WSIs), federated learning (FL) to tackle the challenges of limited training data and associated privacy regulations, and active learning (AL) alongside self-training to optimise model performance when working with limited annotations. Throughout the exploration of each of these methodologies, we maintained a practical focus on their integration in real-world settings.

7.2 Specific remarks

Each chapter of this thesis focused on a specific methodological contribution addressing the different challenges, aligned with both technical and practical aspects.

In Chapter 2, we first addressed the nature of the data itself, introducing a preprocessing module for WSI analysis that served as a foundation for all subsequent work. Mimicking the way pathologists analyse biopsies to reach a diagnosis, we then proposed a two-stage WSI analysis methodology designed to efficiently process the large WSIs and extract meaningful diagnostic features. With a coarse U-Net-based segmentation of the tumorous area from which to extract patches reaching a Dice score of 0.91, we achieved a 0.85 classification accuracy on the test set, demonstrating the effectiveness of our approach.

However, in digital pathology, challenges go beyond the nature of the data, with one of the most significant ones being the limited availability of training data. This is particularly true in the medical field, where privacy regulations often prevent data sharing. To address this,

Chapter 3 introduced a practical and reproducible FL framework, the Notebook Federator, which bridges the gap between AI developers and notebook environments while simplifying the deployment of FL systems. Once the code base is prepared on the provided Jupyter Notebook template on the user side, the component containeriser provided by the group-owned virtual research environment automates the process, simplifying deployment and significantly reducing the time required for each step. The proposed solution was successfully deployed under real-world conditions, running on four different GPUs with varying hardware resources located across three countries, demonstrating its scalability in distributed environments. The Notebook Federator successfully automated the patch-based segmentation of histology images, allowing users to build and push reusable FL-based Docker containers to a registry in a matter of minutes.

Once data could be efficiently processed, and made more accessible through frameworks like FL, the next step was to optimise the training process itself by reducing the need for exhaustive manual annotation, which are often scarce and difficult to obtain due to the time and expertise required. Chapter 4 explored this through the lens of AL, introducing a flexible and lightweight pipeline that enables iterative model improvement by selectively querying the most informative samples. This approach integrated AL with efficient expert-machine interaction, enabling pathologists to play an active role in the training loop by reviewing and annotating only the most informative samples. By allowing experts to iteratively guide model improvement, this method brings AI one step closer to clinical practice, a necessary step for deploying trustworthy and adaptable tools in real-world diagnostic settings.

Building on this tool, Chapter 5 applied this pipeline to the segmentation of Spitz tumours, optimising AL configurations and comparing different model setups. Experiments showed that using Shannon entropy as a scoring strategy led to more effective sample selection at each iteration compared to random sampling, an improvement that was consistent across multiple random seeds. In addition to the core AL approach, the study also explored the integration of self-training. The combined

approach highlighted an overall trend of steady performance gains, with self-training typically improving results across AL iterations. We also analysed interobserver variability between two different expert pathologists, and found inconsistencies in annotations, a reminder of the inherent subjectivity in histopathological annotation and the importance of developing robust, uncertainty-aware models.

Finally, as a final step in this thesis, Chapter 6 situated this work within the broader research context of AI applied to diagnostically ambiguous Spitz tumours, acknowledging the growing body of complementary studies that have emerged in recent years in parallel to this thesis. It identified key challenges and opportunities in the field and proposed future directions to support the development of clinically robust and interpretable AI systems tailored to the complexity of these lesions.

7.3 Future work

The methodologies developed in this thesis have laid the groundwork for advancing DL approaches tailored to the analysis of ambiguous cancers such as Spitz tumours. While this work has demonstrated both the feasibility and potential of these methods, it also highlights important gaps and opportunities for future research. Addressing these areas will be essential to translate technical progress into real-world clinical impact.

A key priority moving forward is improving access to broader and more representative datasets, especially for rare tumour types. In this regard, expanding FL frameworks offers a promising direction, as it can facilitate secure, large-scale collaborations across institutions while preserving patient privacy. Looking ahead, this includes moving beyond diagnostic tasks to incorporate prognostic modelling. This is especially important for tumours of uncertain malignant potential, where outcome prediction remains an unmet clinical need.

In parallel, there is an exciting opportunity in the exploration of foundation models as feature extractors for DL approaches targeting Spitz tumours. These large pretrained models, developed using vast

and diverse datasets, could offer a valuable starting point to improve performance, especially in settings with limited data.

As model performance improves, so too must the emphasis on clinical usability and interpretability. In particular, explainability will play a central role in the safe and meaningful integration of AI into pathology workflows. Future research should continue to explore and refine explainability methods that provide clear, case-specific insights into model predictions. These tools must be designed not only to meet technical standards but also to align with the diagnostic reasoning of expert pathologists, ultimately helping to build trust in AI-assisted decision-making. Furthermore, the need for multimodal approaches becomes particularly evident when dealing with highly ambiguous tumours such as Spitz lesions, where single-modality inputs may fall short in capturing diagnostic complexity.

Finally, we emphasise the importance of sustained, cross-speciality collaboration. Meaningful progress in this domain will most certainly depend on strong partnerships between AI researchers, pathologists, clinicians, and software developers. This includes involving clinical stakeholders early in the design process to ensure that tools are aligned with real-world workflows and constraints. Equally important is the development of user-centred interfaces that allow for interactive exploration, feedback, and expert oversight. By working across disciplines, future efforts can help ensure that AI tools are not only accurate, but also practical, interpretable, and, last but not least, ethically responsible.

Journal papers

- Launet, L.**, Colomer, A., Mosquera-Zamudio, A., Monteagudo, C., & Naranjo, V. The puzzling Spitz tumours: is artificial intelligence the key to their understanding? *Histopathology* (2025).
- Mosquera-Zamudio, A., **Launet, L.**, Colomer, A., ..., Naranjo, V., & Monteagudo, C. Histological interpretation of spitzoid tumours: an extensive machine learning-based concordance analysis for improving decision making. *Histopathology*, 85(1):155-170 (2024).
- Mosquera-Zamudio, A., **Launet, L.**, del Amor, R., Moscardó, A., Colomer, A., Naranjo, V., & Monteagudo, C. A Spitzoid Tumor dataset with clinical metadata and Whole Slide Images for Deep Learning models. *Scientific Data*, 10, 704 (2023).
- Launet, L.**, Wang, Y., Colomer, A., Igual, J., Pulgarín-Ospina, C., Koulouzis, S., Bianchi, R., Mosquera-Zamudio, A., Monteagudo, C., Naranjo, V., & Zhao, Z. Federating Medical Deep Learning Models from Private Jupyter Notebooks to Distributed Institutions. *Applied Sciences*, 13(2), 919 (2023).
- Mosquera-Zamudio, A., **Launet, L.**, Tabatabaei, Z., Parra-Medina, R., Colomer, A., Oliver Moll, J., Monteagudo, C., Janssen, E., & Naranjo, V. Deep Learning for Skin Melanocytic Tumors in Whole-Slide Images: A Systematic Review. *Cancers*, 15(1), 42 (2022).

del Amor, R., **Launet, L.**, Colomer, A., Moscardó, A., Mosquera-Zamudio, A., Monteagudo, C., Naranjo, V. An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. *Artificial Intelligence in Medicine*, 121 (2021).

International conferences

P.García-de-la-Puente, N., López-Pérez, M., **Launet, L.**, Naranjo, V. *Domain Adaptation for Unsupervised Cancer Detection: An application for skin Whole Slides Images from an interhospital dataset* in *2024 Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2024) 58-68.

Pulgarín-Ospina, C., **Launet, L.**, Colomer, A., Naranjo, V. *Optimizing Deep Learning Models for Edge Computing in Histopathology: Bridging the Gap to Clinical Practice* in *28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES)* (2024) 2549-2557.

Launet, L., Colomer, A., Mosquera-Zamudio, A., Moscardó, A., Monteagudo, C., Naranjo, V. *A Self-Training Weakly-Supervised Framework for Pathologist-Like Histopathological Image Analysis* in *2022 IEEE International Conference on Image Processing (ICIP)* (2022) 2291-2295.

Launet, L., del Amor, R., Colomer, A., Mosquera-Zamudio, A., Moscardó, A., Monteagudo, C., Zhao, Z., Naranjo, V. *Federating Unlabeled Samples: A Semi-supervised Collaborative Framework for Whole Slide Image Analysis* in *2022 Intelligent Data Engineering and Automated Learning (IDEAL)* (2022) 64-72.

del Amor, R., Curieses, F. J., **Launet, L.**, Colomer, A., Moscardó, A., Mosquera-Zamudio, A., Monteagudo, C., Naranjo, V. *Multi-Resolution Framework For Spitzoid Neoplasm Classification Using Histological Data* in *2022 IEEE Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)* (2022).

National conferences

Launet, L., Colomer, A., Peris, M., Pulgarín-Ospina, C.C., del Amor, R., Naranjo, V. *Fomentando las colaboraciones transversales en patología digital con Pixnormous* in *XLII Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB)* (2024).

Mosquera-Zamudio, A., **Launet, L.**, del Amor, R., Naranjo, V., Monteagudo, C. *La inteligencia artificial como guía en el laberinto de los tumores melanocíticos spitzoides* in *48^a Reunión del Grupo Español de Dermatopatología* (2024).

Research awards

Launet, L. (2023). Deep learning methodologies for spitzoid melanocytic tumor characterization. *First Place, 3 Minute Thesis Award* at the 32nd European Signal Processing Conference (EU-SIPCO), Helsinki, Finland.

Technology transfer activities

D4HEALTH: Digital diaries for diagnosis of psychological disorders for mental health tracking (S-097-2021). Naranjo Ornedo, Valeriana; Colomer Granero, Adrián; López Mir, Fernando; del Amor del Amor, Rocío; Payá Bosch, Elena; Pulgarín Ospina, Cristian Camilo; Launet, Laëtitia. 13/07/2021.

Bibliography

1. Atmakuru, A. *et al.* Deep learning in radiology for lung cancer diagnostics: A systematic review of classification, segmentation, and predictive modeling techniques. *Expert Systems with Applications*, 124665 (2024).
2. Naseri, H. & Safaei, A. A. Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review. *BMC cancer* **25**, 75 (2025).
3. Mosquera-Zamudio, A. *et al.* Deep learning for skin melanocytic tumors in whole-slide images: A systematic review. *Cancers* **15**, 42 (2022).
4. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **71**, 209–249 (2021).
5. Acs, B., Rantalainen, M. & Hartman, J. Artificial intelligence as the next step towards precision pathology. *Journal of internal medicine* **288**, 62–81 (2020).
6. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25**, 1301–1309 (2019).
7. Feng, R. *et al.* A deep learning approach for colonoscopy pathology WSI analysis: accurate segmentation and classification. *IEEE Journal of Biomedical and Health Informatics* **25**, 3700–3708 (2020).
8. Tayebi, R. M. *et al.* Automated bone marrow cytology using deep learning to generate a histogram of cell types. *Communications medicine* **2**, 45 (2022).
9. Han, C. *et al.* Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. *Medical Image Analysis* **80**, 102487 (2022).

10. Del Amor, R. *et al.* A self-contrastive learning framework for skin cancer detection using histological images in 2022 IEEE International Conference on Image Processing (ICIP) (2022), 2291–2295.
11. Silva-Rodriguez, J., Colomer, A., Dolz, J. & Naranjo, V. Self-learning for weakly supervised gleason grading of local patterns. *IEEE journal of biomedical and health informatics* **25**, 3094–3104 (2021).
12. Tafavvoghi, M. *et al.* Deep learning-based classification of breast cancer molecular subtypes from H&E whole-slide images. *Journal of Pathology Informatics* **16**, 100410 (2025).
13. Fernandez-Martín, C. *et al.* Uninformed Teacher-Student for hard-samples distillation in weakly supervised mitosis localization. *Computerized Medical Imaging and Graphics* **112**, 102328 (2024).
14. Kanwal, N. *et al.* Are you sure it's an artifact? Artifact detection and uncertainty quantification in histological images. *Computerized Medical Imaging and Graphics* **112**, 102321 (2024).
15. Boyd, J. *et al.* Region-guided cyclegans for stain transfer in whole slide images in International Conference on Medical Image Computing and Computer-Assisted Intervention (2022), 356–365.
16. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine* **25**, 954–961 (2019).
17. Dippel, J. *et al.* Ai-based anomaly detection for clinical-grade histopathological diagnostics. *Nejm Ai* **1**, AIoa2400468 (2024).
18. Denic, A. *et al.* Chronic changes on kidney histology by a multiclass AI model. *Kidney International Reports* (2025).
19. Reza Tizhoosh, H. & Pantanowitz, L. Artificial intelligence and digital pathology: Challenges and opportunities. *Journal of Pathology Informatics*. ISSN: 21533539 (2018).
20. Morales, S., Engan, K. & Naranjo, V. Artificial intelligence in computational pathology—challenges and future directions. *Digital Signal Processing* **119**, 103196 (2021).
21. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*. ISSN: 1546170X (2019).
22. Coudray, N. *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*. ISSN: 1546170X (2018).

-
23. Cheng, J. Y., Abel, J. T., Balis, U. G., McClintock, D. S. & Pantanowitz, L. Challenges in the development, deployment, and regulation of artificial intelligence in anatomic pathology. *The American Journal of Pathology* **191**, 1684–1692 (2021).
 24. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Scientific reports* **7**, 1–7 (2017).
 25. Dolezal, J. M. *et al.* Slideflow: Deep learning for digital histopathology with real-time whole-slide visualization. *arXiv preprint arXiv:2304.04142* (2023).
 26. Faa, G. *et al.* The Quest for the Application of Artificial Intelligence to Whole Slide Imaging: Unique Prospective from New Advanced Tools. *Algorithms* **17**, 254 (2024).
 27. Spitz, S. Melanomas of childhood. *The American journal of pathology* **24**, 591 (1948).
 28. Lallas, A. *et al.* Atypical Spitz tumours and sentinel lymph node biopsy: a systematic review. *The lancet oncology* **15**, e178–e183 (2014).
 29. Murali, R. *et al.* Sentinel lymph node biopsy in histologically ambiguous melanocytic tumors with spitzoid features (so-called atypical spitzoid tumors). *Annals of Surgical Oncology* **15**, 302–309 (2008).
 30. Massi, D., De Giorgi, V. & Mandalà, M. The complex management of atypical Spitz tumours. *Pathology* **48**, 132–141 (2016).
 31. Yeh, I. & Busam, K. J. Spitz melanocytic tumours—a review. *Histopathology* **80**, 122–134 (2022).
 32. Smith, K. J., Barrett, T. L., Skelton III, H. G., Lupton, G. P. & Graham, J. H. Spindle cell and epithelioid cell nevi with atypia and metastasis (malignant Spitz nevus). *The American journal of surgical pathology* **13**, 931–939 (1989).
 33. Barnhill, R. L., Flotte, T. J., Fleischli, M. & Perez-Atayde, A. Cutaneous melanoma and atypical Spitz tumors in childhood. *Cancer* **76**, 1833–1845 (1995).
 34. Wiesner, T. *et al.* Genomic aberrations in spitzoid melanocytic tumours and their implications for diagnosis, prognosis and therapy. *Pathology* **48**, 113–131 (2016).
 35. Mosquera-Zamudio, A. *et al.* Histological interpretation of spitzoid tumours: an extensive machine learning-based concordance analysis for improving decision making. *Histopathology* (2024).

36. Li, J. *et al.* A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in biology and medicine* **131**, 104253 (2021).
37. Yan, R. *et al.* Histopathological bladder cancer gene mutation prediction with hierarchical deep multiple-instance learning. *Medical Image Analysis* **87**, 102824 (2023).
38. Del Amor, R. *et al.* An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. *Artificial Intelligence in Medicine*. ISSN: 18732860 (2021).
39. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2015). ISBN: 9783319245737.
40. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2016). ISBN: 9781467388504. arXiv: 1512.03385.
41. Kalapahar, A., Silva-Rodríguez, J., Colomer, A., *et al.* *Gleason grading of histology prostate images through semantic segmentation via residual U-Net* in *2020 IEEE International Conference on Image Processing (ICIP)* (2020), 2501–2505.
42. Simonyan, K. & Zisserman, A. *Very deep convolutional networks for large-scale image recognition* in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015). arXiv: 1409.1556.
43. Ilse, M., Tomczak, J. M. & Welling, M. *Attention-based deep multiple instance learning* in *35th International Conference on Machine Learning, ICML 2018* (2018). ISBN: 9781510867963. arXiv: 1802.04712.
44. Barnhill, R. L. *The Spitzoid lesion: Rethinking Spitz tumors, atypical variants, 'Spitzoid melanoma' and risk assessment* in *Modern Pathology* (2006).
45. Dozat, T. Incorporating Nesterov Momentum into Adam. *ICLR Workshop* (2016).
46. Van Der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*. ISSN: 1532-4435 (2008).
47. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. & Mahmood, F. *Synthetic data in machine learning for medicine and healthcare* 2021.

48. Oza, P., Sharma, P., Patel, S., Adedoyin, F. & Bruno, A. Image Augmentation Techniques for Mammogram Analysis. *Journal of Imaging* **8**, 141 (2022).
49. Rocher, L., Hendrickx, J. M. & de Montjoye, Y. A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*. ISSN: 20411723 (2019).
50. Konecný, J., McMahan, H. B., Ramage, D. & Richtárik, P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *ArXiv* **abs/1610.02527** (2016).
51. Li, T., Sahu, A. K., Talwalkar, A. & Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*. ISSN: 15580792. arXiv: 1908.07873 (2020).
52. Li, W., Milletari, F., Xu, D., Rieke, N., *et al.* Privacy-preserving federated brain tumour segmentation in *International workshop on machine learning in medical imaging* (2019), 133–141.
53. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., *et al.* Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports* **10**, 1–12 (2020).
54. Zhao, Z., Koulouzis, S., Bianchi, R., Farshidi, S., *et al.* Notebook-as-a-VRE (NaaVRE): From private notebooks to a collaborative cloud virtual research environment. *Software: Practice and Experience*. ISSN: 0038-0644 (May 2022).
55. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., *et al.* Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* **14**, 1–210 (2021).
56. Rieke, N., Hancox, J., Li, W., Milletari, F., *et al.* The future of digital health with federated learning. *NPJ digital medicine* **3**, 1–7 (2020).
57. Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N. & Wachinger, C. BrainTorrent: A Peer-to-Peer Environment for Decentralized Federated Learning. *ArXiv* **abs/1905.06731** (2019).
58. Li, X. *et al.* Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Medical Image Analysis*. ISSN: 13618423. arXiv: 2001.05647 (2020).
59. Ju, C. *et al.* Federated Transfer Learning for EEG Signal Classification in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (2020). ISBN: 9781728119908. arXiv: 2004.12321.

60. Andreux, M., du Terrail, J. O., Beguier, C. & Tramel, E. W. *Siloed Federated Learning for Multi-centric Histopathology Datasets in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2020). ISBN: 9783030605476. arXiv: 2008.07424.
61. Lu, M. Y., Chen, R. J., Kong, D., Lipkova, J., *et al.* Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*. ISSN: 13618415. arXiv: 2009.10190 (2022).
62. Ziller, A., Trask, A., Lopardo, A., Szymkow, B., *et al.* in *Federated Learning Systems: Towards Next-Generation AI* (eds ur Rehman, M. H. & Gaber, M. M.) 111–139 (Springer International Publishing, Cham, 2021). ISBN: 978-3-030-70604-3.
63. Reina, G. A., Gruzdev, A., Foley, P., Perepelkina, O., *et al.* *OpenFL: An open-source framework for Federated Learning* 2021.
64. Beutel, D. J. *et al.* Flower: A Friendly Federated Learning Research Framework. *arXiv preprint arXiv:2007.14390* (2020).
65. Xie, Y., Wang, Z., Chen, D., Gao, D., *et al.* FederatedScope: A Flexible Federated Learning Platform for Heterogeneity. *arXiv preprint arXiv:2204.05011* (2022).
66. Lee, H. *et al.* Federated learning for thyroid ultrasound image analysis to protect personal information: Validation study in a real health care environment. *JMIR medical informatics* **9**, e25869 (2021).
67. Florescu, L. M. *et al.* Federated Learning Approach with Pre-Trained Deep Learning Models for COVID-19 Detection from Unsegmented CT images. *Life* **12**, 958 (2022).
68. Lo, J. *et al.* Federated learning for microvasculature segmentation and diabetic retinopathy classification of OCT data. *Ophthalmology Science* **1**, 100069 (2021).
69. Brendan McMahan, H., Moore, E., Ramage, D., Hampson, S. & Agüera y Arcas, B. *Communication-efficient learning of deep networks from decentralized data in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017* (2017). arXiv: 1602.05629.
70. Kanagavelu, R., Li, Z., Samsudin, J., Yang, Y., *et al.* Two-Phase Multi-Party Computation Enabled Privacy-Preserving Federated Learning. *Proceedings - 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGRID 2020*, 410–419. arXiv: 2005.11901 (2020).

-
71. Swillens, J. E. *et al.* Pathologists' first opinions on barriers and facilitators of computational pathology adoption in oncological pathology: an international study. *Oncogene* **42**, 2816–2827 (2023).
 72. Settles, B. *Active learning literature survey* University of Wisconsin-Madison Department of Computer Sciences. 2009.
 73. Budd, S., Robinson, E. C. & Kainz, B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis* **71**, 102062 (2021).
 74. Li, W. *et al.* Pathal: An active learning framework for histopathology image analysis. *IEEE Transactions on Medical Imaging* **41**, 1176–1187 (2021).
 75. Qiu, J. *et al.* Adaptive Region Selection for Active Learning in Whole Slide Image Semantic Segmentation in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023), 90–100.
 76. Lehni, J. & Puckey, J. *Paper.js — The Swiss Army Knife of Vector Graphics Scripting* <http://paperjs.org/>. Version 0.12.11. Open source vector graphics scripting framework. Distributed under the MIT License. 2023.
 77. Grinberg, M. *Flask web development* (" O'Reilly Media, Inc.", 2018).
 78. Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal* **27**, 379–423 (1948).
 79. Dillon, D. *et al.* *Diseases of the Breast* (Wolters Kluwer Health, Philadelphia, PA, 2014).
 80. Cruz-Roa, A. *et al.* High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. *PloS one* **13**, e0196828 (2018).
 81. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* **18** (2015), 234–241.
 82. Meirelles, A. L. *et al.* Effective and efficient active learning for deep learning-based tissue image analysis. *Bioinformatics* **39**, btad138 (2023).
 83. Schmidt, A. *et al.* Focused active learning for histopathological image classification. *Medical Image Analysis* **95**, 103162 (2024).

84. Kang, H., Kim, M., Ko, Y. S., Cho, Y. & Yi, M. Y. WISE: Efficient WSI selection for active learning in histopathology. *Computerized Medical Imaging and Graphics* **118**, 102455 (2024).
85. Folmsbee, J. *et al.* Histology segmentation using active learning on regions of interest in oral cavity squamous cell carcinoma. *Journal of Pathology Informatics* **13**, 100146 (2022).
86. Belharbi, S., Ben Ayed, I., McCaffrey, L. & Granger, E. *Deep active learning for joint classification & segmentation with weak annotator* in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021), 3338–3347.
87. Van Bockstal, M. R. *et al.* Interobserver variability in the assessment of stromal tumor-infiltrating lymphocytes (sTILs) in triple-negative invasive breast carcinoma influences the association with pathological complete response: the IVITA study. *Modern Pathology* **34**, 2130–2140 (2021).
88. Smits, L. J. *et al.* Diagnostic variability in the histopathological assessment of advanced colorectal adenomas and early colorectal cancer in a screening population. *Histopathology* **80**, 790–798 (2022).
89. Sanz-Motilva, V. *et al.* Interobserver Variability in the Histopathological Evaluation of Melanoma: Analysis of 60 Cases. *Actas Dermo-Sifiliográficas* (2024).
90. Xavier-Junior, J. C. C. & Ocanha-Xavier, J. P. *WHO (2018) classification of skin tumors* 2019.
91. Launet, L. *et al.* *A self-training weakly-supervised framework for pathologist-like histopathological image analysis* in *2022 IEEE International Conference on Image Processing (ICIP)* (2022), 3401–3405.
92. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.
93. Shannon, C. E. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* **5**, 3–55 (2001).
94. Gerami, P. *et al.* in *WHO classification of tumours* (eds Elder, D. & Barnhill, R.) 5th ed. [Internet; beta version ahead of print] (International Agency for Research on Cancer, Lyon, France, 2023).
95. Hagstrom, M., Fumero-Velázquez, M., Dhillon, S., Olivares, S. & Gerami, P. An update on genomic aberrations in Spitz naevi and tumours. *Pathology* **55**, 196–205 (2023).

96. Gerami, P. *et al.* BRAF Mutated and Morphologically Spitzoid Tumors, a Subgroup of Melanocytic Neoplasms Difficult to Distinguish From True Spitz Neoplasms. *The American Journal of Surgical Pathology* **48**, 538–545 (2024).
97. Barnhill, R. *et al.* Malignant spitz tumour (spitz melanoma). *World Health Organization*, 108–10 (2018).
98. Belaala, A., Terrissa, L. S., Zerhouni, N. & Devalland, C. Computer-aided diagnosis for spitzoid lesions classification using artificial intelligence techniques. *International Journal of Healthcare Information Systems and Informatics (IJHISI)* **16**, 16–37 (2021).
99. Del Amor, R. *et al.* An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. *Artificial intelligence in medicine* **121**, 102197 (2021).
100. Del Amor, R. *et al.* Multi-resolution framework for spitzoid neoplasm classification using histological data in 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP) (2022), 1–5.
101. Tabatabaei, Z., Colomer, A., Moll, J. O. & Naranjo, V. Siamese Content-based Search Engine for a More Transparent Skin and Breast Cancer Diagnosis through Histological Imaging. *arXiv preprint arXiv:2401.08272* (2024).
102. Del Amor, R. *et al.* A Deep Embedded Framework for Spitzoid Neoplasm Classification Using DNA Methylation Data in 2021 29th European Signal Processing Conference (EUSIPCO) (2021), 1271–1275.
103. Lang, U. E. *et al.* Ciliation index is a useful diagnostic tool in challenging spitzoid melanocytic neoplasms. *Journal of Investigative Dermatology* **140**, 1401–1409 (2020).
104. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
105. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 7132–7141.
106. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning in *International conference on machine learning* (2018), 2127–2136.
107. Kim, J., Dabiri, S. & Seeley, E. S. Primary cilium depletion typifies cutaneous melanoma in situ and malignant melanoma. *PloS one* **6**, e27410 (2011).

108. Elmore, J. G. *et al.* Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *bmj* **357** (2017).
109. Bulten, W. *et al.* Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature medicine* **28**, 154–163 (2022).
110. Spanhol, F. A., Oliveira, L. S., Petitjean, C. & Heutte, L. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering* **63**, 1455–1462 (2015).
111. Mosquera-Zamudio, A. *et al.* A Spitzoid Tumor dataset with clinical metadata and Whole Slide Images for Deep Learning models. *Scientific Data* **10**, 704 (2023).
112. Lu, M. Y. *et al.* Federated learning for computational pathology on gigapixel whole slide images. *Medical image analysis* **76**, 102298 (2022).
113. Launet, L. *et al.* Federating medical deep learning models from private Jupyter notebooks to distributed institutions. *Applied Sciences* **13**, 919 (2023).
114. Dippel, J. *et al.* RudolfV: a foundation model by pathologists for pathologists. *arXiv preprint arXiv:2401.04079* (2024).
115. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical twitter. *Nature medicine* **29**, 2307–2316 (2023).
116. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874 (2024).
117. Ding, T. *et al.* Multimodal whole slide foundation model for pathology. *arXiv preprint arXiv:2411.19666* (2024).
118. Zhou, X. *et al.* A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis. *arXiv preprint arXiv:2412.13126* (2024).
119. March, J., Hand, M., Truong, A. & Grossman, D. Practical application of new technologies for melanoma diagnosis: Part II. Molecular approaches. *Journal of the American Academy of Dermatology* **72**, 943–958 (2015).
120. Bonilla, J. M., Tabanera, M. T. & Mendoza, L. R. El cáncer de mama en el siglo XXI: de la detección precoz a los nuevos tratamientos. *Radiología* **59**, 368–379 (2017).

121. Pulgarín-Ospina, C. C., del Amor, R., Colomera, A., Silva-Rodríguez, J. & Naranjo, V. Histocolai: an open-source web platform for collaborative digital histology image annotation with ai-driven predictive integration. *arXiv preprint arXiv:2307.07525* (2023).
122. Lundberg, S. M. & Lee, S.-I. *A Unified Approach to Interpreting Model Predictions* in *Advances in Neural Information Processing Systems* (eds Guyon, I. *et al.*) **30** (Curran Associates, Inc., 2017).
123. Del Amor, R. *et al.* Annotation protocol and crowdsourcing multiple instance learning classification of skin histological images: The CR-AI4SkIN dataset. *Artificial Intelligence in Medicine* **145**, 102686 (2023).