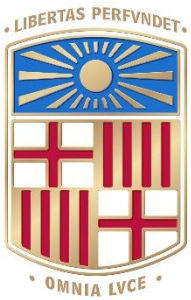

UNIVERSITAT DE BARCELONA

Final Degree Project

Biomedical Engineering Degree

"Automatic Segmentation of Postmenopausal Ovaries in T2-Weighted MRI using Deep Learning"

"Segmentació Automàtica d'Ovaris Postemenopàusics en Imatges per RM ponderades en T2 mitjançant aprenentatge profund"

Barcelona, 10th June 2026

Author: Berta Haro Jovés

Director/s: Aida Niñerola Baizán

Arnau Farré Melero

Tutor: Aida Niñerola Baizán

Abstract

Postmenopausal ovaries are small, low-contrast structures that are difficult to delineate reliably on T2-weighted MRI. Accurate segmentation is a prerequisite for radiomic analysis and AI-based characterisation of adnexal masses, but no published method specifically addresses normal, atrophic postmenopausal ovaries. This work develops and evaluates a deep learning pipeline for automatic segmentation using a clinical dataset from 103 patients at Hospital Clínic de Barcelona.

Several deep learning architectures and training strategies were compared. A patch-based approach with data augmentation, designed to address the severe class imbalance in this target, served as the primary model. Two automatic localisation approaches were also developed. As a downstream application, radiomic features were extracted from the ground-truth masks and used to train a classifier for binary age prediction

The patch-based AttentionU-Net achieved a validation Dice of 0.619 and a test Dice of 0.447, measured with ground-truth centroids. The radiomics classifier achieved 61.9% accuracy on binary age classification (chance level: 50.5%), with shape descriptors as the most discriminative features.

This work delivers the first reported segmentation results for normal atrophic postmenopausal ovaries in T2-weighted MRI, together with a preliminary radiomic characterisation of the segmented structures. Automatic localisation of the ovary remains the main open challenge before a fully automated pipeline becomes feasible.

Key Words

Ovarian segmentation, Deep learning, T2-weighted MRI, AttentionU-Net, Postmenopausal ovary, Radiomics, Medical image segmentation

Resum

Els ovaris postmenopàusics són estructures petites i de baix contrast, difícils de delimitar de forma fiable en RM T2. Una segmentació precisa és un prerequisit per a l'anàlisi radiòmica i la caracterització per IA de les masses anexials, però cap mètode publicat aborda específicament els ovaris postmenopàusics normals i atròfics. Aquest treball desenvolupa i avalua un pipeline de deep learning per a la segmentació automàtica a partir d'un conjunt de dades clínic de 103 pacients de l'Hospital Clínic de Barcelona.

Es van comparar diverses arquitectures i estratègies d'entrenament. El model principal va ser un enfocament basat en pedaços amb augmentació de dades, que resol el desequilibri de classes inherent a aquest problema. També es van desenvolupar dos mètodes de localització automàtica. Com a aplicació secundària, es van extreure característiques radiòmiques de les màscares de referència per entrenar un classificador per a la predicció binària d'edat.

L'AttentionU-Net basada en patches va assolir un Dice de validació de 0,619 i un Dice de test de 0,447. El classificador radiòmic va obtenir un 61,9% d'exactitud en la classificació binària d'edat (nivell d'atzar: 50,5%), amb els descriptors de forma com a característiques més discriminatives.

Aquest treball reporta els primers resultats de segmentació publicats per a ovaris postmenopàusics normals i atròfics en RM T2, amb una caracterització radiòmica preliminar de les estructures segmentades. La localització automàtica de l'ovari és el pas principal pendent de resoldre abans que el pipeline pugui funcionar sense intervenció manual.

Paraules Clau

Segmentació ovàrica, Deep learning, Ressonància magnètica T2, AttentionU-Net, Ovari postmenopàusic, Radiòmica, Segmentació d'imatge mèdica

Acknowledgements

I want to thank my supervisors, Aida Niñerola Baizán and Arnau Farré Melero, for their guidance throughout this project. Their feedback was direct, and they were always available when experimental results raised more questions than answers.

The Radiology Department at Hospital Clínic de Barcelona provided both the clinical dataset and the radiologist's time to produce the ground-truth segmentations, without which this would not have been possible. In particular, I want to thank Olatz for the careful manual segmentation of the entire dataset.

Finally, I would like to thank my family and friends for their support, especially during the months when the models refused to converge.

Glossary

AI: Artificial Intelligence

BCE: Binary Cross-Entropy

CAGR: Compound Annual Growth Rate

CEIC: Clinical Research Ethics Committee

CNN: Convolutional Neural Network

CropCNN: Crop-based Convolutional Neural Network

CT: Computed Tomography

DICOM: Digital Imaging and Communications in Medicine

DL: Deep Learning

DSC: Dice Similarity Coefficient

DWI: Diffusion-Weighted Imaging

ENISA: European Union Agency for Cybersecurity

GDPR: General Data Protection Regulation

GLDM: Grey-Level Dependence Matrix

GLCM: Grey-Level Co-occurrence Matrix

GLRLM: Grey-Level Run-Length Matrix

GLSZM: Grey-Level Size Zone Matrix

GPU: Graphics Processing Unit

IBSI: Image Biomarker Standardisation Initiative

KNN: K-Nearest Neighbours

LCC: Largest Connected Component

LOOCV: Leave-One-Out Cross-Validation

LOPDGDD: Ley Orgánica de Protección de Datos Personales y Garantía de Derechos Digitales

MONAI: Medical Open Network for AI

MRI: Magnetic Resonance Imaging

NGTDM: Neighbouring Grey-Tone Difference Matrix

NIfTI: Neuroimaging Informatics Technology Initiative

nnU-Net: No-New-UNet

O-RADS: Ovarian-Adnexal Reporting and Data System

PERT: Program Evaluation and Review Technique

RF: Random Forest

ROI: Region of Interest

SAM: Segment Anything Model

SSH: Secure Shell

SVM: Support Vector Machine

UB: Universitat de Barcelona

U-Net: U-shaped Convolutional Neural Network

US: Ultrasound

VPN: Virtual Private Network

WBS: Work Breakdown Structure

List of figures

Figure 5.1. Representative T2-weighted MRI axial slices with ground-truth segmentation masks from the clinical dataset.

Figure 6.1. Training and validation loss (left) and Dice (right) over 215 epochs for the best-performing model (AttentionU-Net, 96×96 patches, weight decay 1e-5).

Figure 6.2. Training progression for a representative validation patient. Left: prediction at epoch 10, showing severe over-segmentation (predicted contour in red, ground truth in green). Right: prediction at epoch 210, where the model has converged, and the predicted contour closely matches the ground truth.

Figure 6.3. Validation Dice comparison across four architectures evaluated under identical conditions.

Figure 6.4. Qualitative results on two representative test patients. Each row shows, from left to right: the full axial slice with the centroid marker, the 96×96 input patch, the ground-truth mask, the model prediction, and the overlay of both contours (ground truth in green, prediction in red).

Figure 6.5. Top 15 radiomic features ranked by mean decrease in impurity from the 500-tree Random Forest classifier.

Figure 8.1. Work Breakdown Structure.

Figure 8.2. PERT Diagram.

Figure 8.3. GANTT Chart.

List of tables

Table 2.1. Summary of the main advantages and limitations of the different segmentation approaches.

Table 2.2. Comparative summary of related work. US = ultrasound; DL = deep learning; postmenop. = postmenopausal.

Table 4.1. Overview of segmentation and visualisation software platforms evaluated in this project.

Table 4.2. Overview of image formats considered for the working dataset.

Table 4.3. Comparison of spatial processing approaches for deep learning segmentation.

Table 4.4. Deep learning architectures evaluated for ovarian segmentation.

Table 4.5. Loss functions evaluated for segmentation and localisation training.

Table 4.6. Internal validation strategies considered.

Table 4.7. Overview of evaluation metrics considered for segmentation and localisation.

Table 4.8. Radiomics tools evaluated.

Table 4.9. Classification models considered for binary age prediction.

Table 5.1. Dataset characterisation summary.

Table 6.1. Validation Dice scores for 2D full-image and half-image segmentation strategies.

Table 6.2. Validation Dice scores for 2D patch-based segmentation experiments.

Table 6.3. Validation Dice scores by architecture under the optimal patch-based training configuration.

Table 6.4. Validation Dice scores for 3D segmentation experiments.

Table 6.5. nnU-Net validation Dice scores at 1,000 and 2,000 training epochs.

Table 6.6. Test set performance of the final 2D patch-based model.

Table 6.7. Localisation network performance on the test set.

Table 6.8. Radiomics-based binary age classification results.

Table 8.1. Listing of each task with its identifier, duration in working days, and predecessor tasks.

Table 9.1. SWOT analysis of the project.

Table 10.1. Economic analysis of the project.

Table of contents

1. INTRODUCTION 1
1.1. Motivation of the Project 1
1.2. Objectives 1
1.3. Limitations and Scope 2
1.4. Location and Time Frame 2
2. BACKGROUND 3
2.1. Ovarian Anatomy and MRI Imaging 3
2.2. State of the Art 5
3. MARKET ANALYSIS 7
3.1. Market Sector 7
3.2. Historical Market Evolution 7
3.3. Commercial Solutions 7
3.4. Market Future Perspectives 8
4. CONCEPT ENGINEERING 9
4.1. Study of the solutions 9
4.2. Proposed solution 16
5. DETAILED ENGINEERING 18
5.1. Development environment 18
5.2. Dataset description and preparation 18
5.3. Dataset characterisation 20
5.4. Model training 20
5.5. Localisation pipeline 23
5.6. Inference pipeline 24
5.7. Radiomics 24
6. RESULTS 25
6.1. Segmentation 25
6.2. Test set evaluation 27
6.3. Localisation pipeline 28
6.4. Radiomics: binary age classification 29
7. DISCUSSION AND FUTURE WORK 30
7.1. The challenge of small-structure segmentation 30
7.2. Architecture and dimensionality choices 30
7.3. Pipeline performance and the localisation bottleneck 31
7.4. Radiomics findings and directions for future work 31
8. PROJECT SCHEDULE 33
8.1. Work Breakdown Structure 33

8.2. PERT Diagram 34
8.3. GANTT Chart 35
9. TECHNICAL FEASIBILITY 36
9.1. Strengths 36
9.2. Weaknesses 36
9.3. Opportunities 36
9.4. Threats 37
10. ECONOMIC FEASIBILITY 38
11. REGULATIONS AND LEGAL ASPECTS 39
12. CONCLUSIONS 40
13. BIBLIOGRAPHY 41

1. INTRODUCTION

1.1. Motivation of the Project

Ovarian cancer is the gynaecological malignancy with the highest mortality rate in developed countries, accounting for over 310,000 new diagnoses and 207,000 deaths worldwide in 2020 [1]. This is due to a well-known clinical problem: the disease tends to be asymptomatic in its early stages, so a large proportion of patients are diagnosed at an advanced stage, when five-year survival falls to 20–30% [2]. By contrast, stage I disease carries survival rates around 90%, which makes early and accurate detection an important clinical priority.

Magnetic resonance imaging (MRI) is commonly used as a second-line tool to evaluate adnexal masses that are unclear on ultrasound. Its high soft-tissue contrast allows better assessment of lesion structure and composition. The O-RADS (Ovarian-Adnexal Reporting and Data System) MRI system provides a standardised way to estimate the risk of malignancy and improve consistency in reporting. However, interpreting MRI scans still depends on the radiologist's experience, leading to variability. Therefore, computational tools that help standardise image analysis could be clinically useful, especially in settings without specialised expertise [3].

Both radiomics (the extraction of quantitative features from medical images) and deep learning (a type of AI that learns patterns directly from image data) offer a path towards this standardisation [4,5]. Both approaches, however, require an accurate segmentation of the ovary as a prerequisite: without a reliable mask, neither radiomic features nor lesion-level predictions are meaningful. However, automated segmentation of postmenopausal ovaries is a technically demanding problem. After menopause, ovarian volume decreases substantially, typically falling to only a few cubic centimetres, and the follicular architecture disappears, making the ovary more difficult to identify on imaging [6].

This work is developed within a project at Hospital Clínic de Barcelona, which aims to build an AI-based (Artificial Intelligence) characterisation framework for ovarian masses in postmenopausal women. Automated ovary segmentation is a foundational step in this pipeline, and providing a reliable solution for this problem is the central motivation of the present work.

In simple terms, this work tries to teach a computer to find and outline the ovaries on MRI scans from postmenopausal women. The main challenge is that these ovaries are very small and look similar to the surrounding tissue, making them hard to detect even for trained eyes.

1.2. Objectives

The main objective of this project is to develop and evaluate a deep learning pipeline for the automatic segmentation of postmenopausal ovaries in T2-weighted MRI, using the clinical dataset collected for the project (103 patients, Hospital Clínic de Barcelona).

To reach this goal, seven tasks were defined:

1. Dataset preparation: acquire, convert (DICOM to NIfTI), and preprocess the T2-weighted MRI dataset.

2. Dataset characterisation: describe the dataset in terms of ovary size and acquisition variability.
3. Baseline segmentation experiments: train initial segmentation models on full-resolution slices to establish a performance baseline.
4. Patch-based segmentation model: develop a patch-based segmentation model to address the class imbalance problem and achieve meaningful segmentation performance.
5. Localisation pipeline: develop an automatic ovary localisation approach to enable inference on new patients without manual segmentations.
6. Evaluation: evaluate all models on a held-out test set and analyse failure modes to guide future work.
7. Radiomics and age prediction: extract radiomic features from the segmented masks and train a classifier for binary age prediction as a downstream application.

1.3. Limitations and Scope

This project focuses on segmenting postmenopausal ovaries in T2-weighted MRI. As an exploratory downstream task, radiomic feature extraction and binary age classification are also included, using the ground-truth segmentation masks. While these represent the intended downstream applications for the hospital project, the classification of adnexal masses, the design of new deep learning architectures, and clinical implementation remain out of scope.

The study uses a clinical cohort provided by Hospital Clínic de Barcelona under an ethics committee-approved research protocol. No external datasets were incorporated. The use of a single-institution dataset limits generalisability; this is acknowledged as an important limitation but reflects the typical data constraints of clinical AI research. The dataset includes images from two scanner manufacturers (GE and Siemens), which introduces acquisition heterogeneity. This may also be considered a limitation in radiomics. An additional challenge is the small size of the target structure in this dataset, which makes it difficult to distinguish from adjacent structures.

Two further practical limitations shaped the experimental design. First, model training was conducted on shared GPU infrastructure, which limited the number of experiments that could run in parallel and increased overall training time. Second, the project was time-limited as the TFG must be submitted in June 2026.

1.4. Location and Time Frame

The project was carried out between February and June 2026 at the University of Barcelona, with collaboration from the radiology department at Hospital Clínic of Barcelona, which also provided clinical data and domain expertise. This project was carried out primarily remotely, using the group's computing workstation through a VPN (Virtual Private Network) connection. Additionally, access to collaborative group spaces and ongoing support from biomedical engineers and radiologists were provided.

2. BACKGROUND

2.1. Ovarian Anatomy and MRI Imaging

2.1.1. Ovarian Anatomy and the Postmenopausal Ovary

The ovaries are paired organs located in the lesser pelvis, one on either side of the uterus. In reproductive-age women, each ovary typically measures around 3 x 2 x 2 cm and contains follicles at various stages of development [6]. Following the menopause (defined clinically as the cessation of menstruation for 12 consecutive months), the ovaries undergo progressive atrophy. When folliculogenesis ceases, oestrogen production falls dramatically, and ovarian volume decreases substantially after menopause, typically to a few cubic centimetres [6].

This atrophy makes postmenopausal ovaries significantly more difficult to identify on imaging. Unlike premenopausal ovaries, which are readily identifiable by their follicular architecture, postmenopausal ovaries are small, homogeneous structures that can be easily confused with adjacent tissues [6]. Despite their inconspicuous appearance, these ovaries remain of considerable clinical importance, as the risk of ovarian malignancy is higher in postmenopausal women [1,2]. Any identifiable ovarian mass or morphological change in this population warrants careful evaluation, which is why automated segmentation is worth pursuing despite the difficulty. This imaging challenge is precisely what makes the choice of modality critical.

2.1.2. MRI Imaging of the Pelvis

Magnetic Resonance Imaging (MRI) generates contrast by measuring the relaxation of hydrogen protons following excitation by a radiofrequency pulse within a static magnetic field. This contrast is primarily driven by two concurrent but independent mechanisms: T1 relaxation, which describes the time taken for protons to return to their original energy state by interacting with the surrounding molecular lattice, and T2 relaxation, which reflects local field variations causing dephasing between adjacent spins [7].

T2-weighted MRI is a primary component used for ovarian mass characterisation [3]. It provides high soft-tissue contrast and enables the identification of fluid, mucin and fibrous tissue based on their T2 signal characteristics. Simple fluid appears very bright, fibrous tissue appears dark, and normal ovarian stroma is hypointense relative to any residual follicular fluid. By contrast, T1-weighted images are particularly useful for identifying fat, blood products, and proteinaceous or mucinous content, which often appear hyperintense on T1. The O-RADS MRI scoring system provides a standardised five-point scale for risk stratification of adnexal masses based on these T2 features [3], and the project aims to develop AI tools that may support this scoring system.

The dataset used in this project consists of T2-weighted MRI volumes acquired on GE and Siemens scanners, with varying in-plane resolutions and slice thicknesses. This variability in acquisition is typical of real clinical data and adds an extra challenge for model generalisation.

Reducing this variability requires computational tools that can analyse images consistently, and deep learning has emerged as the dominant approach for this type of task in medical imaging.

2.1.3. Deep Learning for Medical Image Segmentation

Image segmentation is the process of assigning a label to each pixel or voxel in an image to delineate regions of interest. In medical imaging, it is a prerequisite for volumetric measurements, radiomic feature extraction, and treatment planning [5]. Manual segmentation by trained radiologists remains the clinical reference standard but is time-consuming and subject to inter-observer variability [5]. Semi-automated methods offer a compromise between manual delineation and full automation, but they still depend on expert interaction and therefore remain time-intensive in large datasets.

Deep learning methods, particularly convolutional neural networks (CNNs), have achieved state-of-the-art performance on most medical image segmentation benchmarks since 2015 [5]. CNN-based models can reduce inter-observer variability and scale more effectively to large datasets than manual or semi-automatic approaches. However, their performance may deteriorate in highly complex or atypical cases that fall outside the distribution of the training data, and some applications still require human intervention for refinement or quality control. CNNs apply learned filters to the input image, progressively extracting features at increasingly abstract levels. The encoder-decoder architecture is the dominant design for segmentation: the encoder progressively reduces spatial resolution while extracting abstract features, and the decoder reconstructs the segmentation map at full resolution by combining information from the encoder through skip connections [8].

Table 2.1. Summary of the main advantages and limitations of the different segmentation approaches.

Method	Advantages	Limitations
Manual	High accuracy, clinical gold standard	Time-consuming, inter-observer variability
Semi-automated	Faster than manual, expert guidance	Still requires human interaction, limited scalability
Deep Learning	Reduces variability, scales to large datasets	Fails in atypical cases, may need refinement

A key challenge in medical image segmentation is class imbalance: postmenopausal ovaries occupy a very small portion of the field of view in a typical MRI clinical protocol.

Beyond segmentation itself, once a reliable mask is available, quantitative analysis of the segmented region becomes possible. This is the domain of radiomics.

2.1.4. Radiomics

Radiomics is the quantitative extraction of many features from medical images using computational algorithms [4]. These features capture information about shape, intensity distribution, and texture, many of which are imperceptible to the human eye but contain relevant biomedical information. Radiomics is performed after defining a region of interest (ROI) obtained through image segmentation [4].

Radiomic features are generally classified into three categories: shape features (e.g. volume, sphericity, axis lengths, surface area), first-order or intensity-based features (e.g. mean, standard deviation, entropy, skewness), and texture features that describe spatial relationships between voxel intensities, extracted

from texture matrixes such as the GLCM (grey-level co-occurrence matrix), GLRLM (grey-level run-length matrix), and GLSZM (grey-level size-zone matrix) [4].

With this technical context established, the following section reviews what has been done specifically for ovarian segmentation.

2.2. State of the Art

Automated segmentation of ovarian structures has been explored across multiple imaging modalities, including ultrasound, CT, and MRI. The following review covers the most relevant published approaches, with a focus on deep learning methods and their applicability to the specific problem addressed in this work: the segmentation of normal, atrophic postmenopausal ovaries on T2-weighted MRI.

Beyond CT and MRI, automated segmentation of ovarian structures has also been explored in ultrasound, the most widely used modality for routine ovarian assessment. Chen et al. (2022) [9] proposed a U-Net-based network (see Section 4 for an introduction to this architecture) with an integrated edge-detection branch for ovarian follicle segmentation in ultrasound images, reporting improved performance over the standard U-Net. Although the clinical target differs from the postmenopausal ovaries addressed in this study, this work illustrates the broader interest in automating ovarian analysis across imaging modalities. Table 2.2 provides a comparative summary.

Wang et al. (2023) [10] published one of the most comprehensive studies on automated segmentation of ovarian cancer in contrast-enhanced CT. Their dataset included 367 patients with biopsy-confirmed high-grade serous ovarian cancer, and their best-performing deep learning model achieved excellent segmentation accuracy (Dice similarity coefficient of 0.941). They also assessed the stability of radiomic features extracted from automated segmentations, finding that 85% remained highly stable, validating the use of automated masks for downstream radiomics. However, this result must be contextualised: ovarian tumours in CT are typically large, hyperattenuating masses that are much easier to segment than the small, postmenopausal ovaries targeted in this study.

Buddenkotte et al. (2023) [11] addressed the segmentation of multisite ovarian cancer on CT, targeting pelvic/ovarian masses and omental lesions simultaneously. Their model achieved a Dice score of 0.71 for pelvic/ovarian lesions across four international institutions, performing comparably to a trainee radiologist. Their results also demonstrated that careful hyperparameter tuning significantly outperformed default existing configurations; a finding consistent with the difficulties encountered with nnUNet in this project.

Hsu et al. (2025) [12] presented a multicentre study using Meta's Segment Anything Model (SAM) as the segmentation backbone for ovarian lesions on MRI, combined with a DenseNet-121 classifier. The system achieved a Dice of 0.86–0.88 and reduced manual delineation time by approximately 4 minutes per lesion. The integrated model, which classifies ovarian lesions as benign or malignant, achieved an AUC of 0.85, comparable with radiologist performance.

Liang et al, (2025) [13] introduced a multimodal pelvic MRI dataset for endometriosis and proposed RAovSeg, a two-step pipeline that combines a ResNet-based slice classifier with an Attention U-Net segmentation model. The method achieved a Dice score of 0.290 for ovary segmentation, and removing

the slice selection step caused performance to collapse to 0.013, directly confirming the importance of the localisation step. However, the task differs from the present work, as it focuses on the ovaries of patients with endometriosis, which may exhibit structural abnormalities.

Considering the papers mentioned above, we did not identify studies that specifically address the segmentation of healthy, atrophic postmenopausal ovaries on T2-weighted MRI, a fundamentally different and substantially harder task due to the small size, low contrast with surrounding tissue, and the absence of any pathological mass to guide the model. This project directly addresses this gap.

Table 2.2. Comparative summary of related work. US = ultrasound; DL = deep learning; postmenop. = postmenopausal.

Reference	Year	Modality	Target	Architecture	N patients	Dice (test)
Wang et al. [10]	2023	CT	Ovarian tumour	nnUNet 3D cascade	367	0.941
Buddenkotte et al. [11]	2023	CT	Ovarian + omental disease	Custom DL	451	0.71 / 0.61
Hsu et al. [12]	2025	MRI (multi-seq.)	Ovarian lesion	SAM + DenseNet-121	534	0.86–0.88
RAovSeg [13]	2025	MRI (T2, DWI)	Ovary	ResNet18 + AttUNet 2D	132	0.290

3. MARKET ANALYSIS

3.1. Market Sector

This project is primarily designed for the digital health and medical technologies (MedTech) sector, with direct applications in medical radiology, diagnostic imaging, and gynaecological oncology, providing decision-support tools for clinicians. The rapid growth of AI across multiple industries has driven significant advances in medical AI, particularly in imaging.

Automated ovarian segmentation from T2-weighted MRI could improve radiological workflow efficiency and reduce inter-observer variability. The main potential beneficiaries include hospital imaging services, gynaecological oncology units, clinical research teams, and medical AI companies operating in pelvic imaging.

3.2. Historical Market Evolution

AI-assisted radiology has evolved considerably over the past three decades. Early computer-aided detection systems emerged in the 1990s, and the field gained significant momentum in the mid-2010s when deep learning architectures began outperforming traditional image processing methods. This period also marked a substantial increase in regulatory approvals of AI-based medical imaging tools in both the United States and Europe. The number of approved AI/ML-based medical devices increased substantially between 2015 and 2020, reaching 240 CE-marked devices in Europe [14]. More recently, uptake has grown substantially, with nearly half of European radiologists reporting regular use of AI tools in clinical practice as of 2024 [15].

In gynaecological imaging, progress has been somewhat slower. Radiomics applied to ovarian masses began gaining traction in the early 2010s as a means of extracting objective, quantitative descriptors from routine MRI acquisitions [4]. The subsequent integration of deep learning into segmentation workflows has opened new possibilities for automating the delineation of pelvic structures. However, dedicated tools for ovarian analysis remain largely confined to research [16]. This gap between technological potential and clinical availability is what motivates projects like this one.

3.3. Commercial Solutions

Several commercial platforms currently support radiologists in medical image analysis, though none are specifically designed for normal ovarian segmentation in MRI.

Mint Lesion (Mint Medical) is a CE-marked platform integrated into DICOM workflows that combines semi-automatic segmentation, radiomics extraction, and structured reporting in a single environment [17]. Its advanced visualisation tools and export capabilities make it well-suited for clinical segmentation workflows, and it has been adopted in Hospital Clínic de Barcelona for manual delineation tasks. However, it requires manual interaction and does not provide fully automated segmentation.

Syngo.via (Siemens Healthineers) is another CE-marked solution offering automatic and semi-automatic segmentation, multimodal image integration and 3D visualisation [18]. While it is widely deployed in

clinical settings, its performance in anatomically complex pelvic regions is limited and typically requires manual correction.

More recently, AI-driven platforms such as AI-Rad Companion (Siemens Healthineers) have extended AI-based detection and segmentation to multiple organ systems, including the abdomen and pelvis. Still, automated MR-based contouring is currently tailored specifically to the male pelvic anatomy [19]. Dedicated modules for gynaecological structures, such as the ovaries, remain largely absent from these commercial offerings.

Across all these platforms, a common limitation is the lack of specialised models for small, low-contrast structures such as postmenopausal atrophic ovaries, where tissue boundaries are subtle and conventional segmentation approaches often underperform [16]. This gap motivates the development of a dedicated deep learning pipeline, as proposed in this project.

3.4. Market Future Perspectives

The medical AI imaging sector is currently expanding and is forecasted to experience significant growth in the coming years. The radiology AI market is expected to grow at a compound annual growth rate (CAGR) of 24.5% between 2025 and 2030, nearly tripling its 2025 valuation by the end of the decade [20]. This growth will be driven by a paradigm shift in which AI moves from being an experimental tool to an established clinical standard, with regulatory frameworks in both Europe and the United States progressively adapting to accommodate AI-based medical devices.

In gynaecological imaging, dedicated AI tools remain largely absent from clinical practice despite growing interest. The persistent gap between technological capability and clinical availability, identified in Section 3.2, is increasingly recognised by both research institutions and industry. The logical next step is integrating AI-assisted segmentation into structured reporting workflows, enabling faster and more reproducible characterisation of pelvic structures. The combination of automated segmentation and radiomic feature extraction, as explored in this project, represents a key direction for non-invasive ovarian mass characterisation and risk stratification [4].

Combining imaging biomarkers with clinical and molecular data could eventually support more personalised management of patients in gynaecological oncology. As datasets grow and multicentre collaborations become more common, the generalisation of deep learning models, currently one of the main limitations in this field, is likely to improve. This project addresses one part of that problem: automated segmentation of normal postmenopausal ovaries in T2-weighted MRI, a structure that current commercial tools do not handle well.

4. CONCEPT ENGINEERING

4.1. Study of the solutions

4.1.1. Segmentation and visualisation software

Several software platforms were considered for segmentation and visual quality control during the project. Table 4.1 presents a comparative overview.

Table 4.1. Overview of segmentation and visualisation software platforms evaluated in this project.

Software	Description	Decision
Mint Lesion	Commercial clinical platform used at Hospital Clínic de Barcelona. Radiologist segmentations were originally produced and exported from this platform in NIfTI format.	Ground-truth source. Not used during model development.
3D-Slicer	Open-source software widely used for NIfTI visualisation and overlay inspection. Prior experience available in the group.	Selected for visual quality control throughout preprocessing and training.
ITK-SNAP	Open-source tool widely used in medical imaging. No prior experience available.	Not used.
MRIcron	Open-source viewer with less intuitive interface. No prior experience available.	Not used.
syngo.via	CE-marked commercial platform from Siemens Healthineers offering automatic and semi-automatic segmentation, multimodal image integration, and 3D visualisation. Integrated at Hospital Clínic de Barcelona	Not used.

The ground-truth ovarian masks were delineated by a clinical radiologist at Hospital Clinic de Barcelona using Mint Lesion, an advanced clinical platform specifically designed for structured lesion reporting and segmentation. The masks were exported in NIfTI format and delivered together with the DICOM images; because the segmentations were already available before this project, no new ground-truth segmentations needed to be generated. 3D Slicer was selected for all visual quality-control tasks throughout preprocessing and training, owing to its native support for NIfTI overlays and the group's prior experience with it.

4.1.2. Data management

A key preliminary decision was the choice of image format for the working dataset. The images were originally delivered in DICOM format, while the segmentation masks were already in NIfTI format. Table 4.2 summarises the formats considered.

Table 4.2. Overview of image formats considered for the working dataset.

Format	Description	Decision
DICOM	Standard clinical format combining image data and rich metadata (scanner parameters, patient info, acquisition details). All MRI volumes were originally delivered in this format.	Input format only - converted during preprocessing.

NIfTI (.nii /.nii.gz)	Volumetric format storing image data and affine matrix in a single file. Standard in neuroimaging and deep learning workflows. Ground-truth masks were already provided in this format.	Selected as working format for all images and masks.
Analyze (.hdr/.img)	Legacy two-file format. Lacks standardised metadata storage and has been superseded by NIfTI.	Not used.
RTSTRUCT	DICOM radiotherapy format encoding segmentation contours as polygonal outlines per slice. Less practical for voxel-wise deep learning workflows.	Not used - masks already available as NIfTI.

NIfTI (.nii.gz) was selected as the working format for all images and masks.

4.1.3. Preprocessing

Several preprocessing steps were designed and applied to transform the heterogeneous clinical database into a geometrically consistent dataset suitable for deep learning. Unlike in multiple-image datasets, no image registration was required here because all studies correspond to a single T2-weighted acquisition per patient [7]. The preprocessing options evaluated were:

- **Matrix transposition**: Geometric consistency of the converted NIfTI volumes was verified across the dataset. Where the slice dimension was found to be incorrectly ordered, axis permutation correction was applied to restore a uniform spatial orientation. This step was mandatory for geometrically consistent training.
- **Mask alignment verification and correction**: The spatial consistency between each image and its corresponding mask was verified using a verification script. Cases in which affine properties differed beyond numerical tolerance were corrected by reconstructing the mask in image space using DICOM metadata. Visual confirmation on a representative subset confirmed the correct overlay before training.
- **Intensity normalisation**: MRI intensity values are scanner-dependent and lack an absolute scale, making per-volume standardisation essential. Two strategies were evaluated: z-score normalisation (zero mean, unit variance) and min-max scaling (range 0 to 1). Z-score normalisation per volume was selected, as it is more robust to outlier intensity values common in clinical MRI acquisitions.
- **Resampling**: In-plane resampling to a common isotropic spacing was applied for the radiomics pipeline, where IBSI-compliant feature extraction requires a fixed voxel size [21]. For the deep learning models, resampling was not applied. Slice thickness varies across acquisitions (3.0-4.5 mm), but the range is narrow enough that native resolution training is the safer choice; resampling would introduce interpolation artefacts without a clear benefit.
- **Denoising/smoothing**: Not applied. MRI images are already low in noise, and smoothing risks blurring the fine boundaries needed to delineate small ovarian structures.
- **Data augmentation** [22]: Random horizontal flipping, small rotations, zoom, and random contrast shifts were applied on-the-fly during training to increase the effective diversity of the

training set. Elastic deformations were considered but excluded to avoid distorting the small target structure.

Moreover, segmentation models trained on centred patches require a known ovary location at inference time. To remove this dependency and enable fully automated inference on unseen patients, a localisation pipeline was considered as a preprocessing step before segmentation.

Two approaches were evaluated. First, a lightweight convolutional network (CropCNN) that predicts the vertical crop range from a maximum-intensity projection of the volume, reducing the field of view to the pelvic region. Second, a heatmap-based model that predicts a Gaussian centred on the approximate ovary location from the cropped volume.

4.1.4. Deep Learning model

The deep learning model constitutes the core of the project. Several aspects were analysed: the model architecture, the strategy to address class imbalance, the training loss function, the hyperparameter configuration, and the validation approach. All alternatives were evaluated considering the temporal, computational, and data constraints of the project.

Selection of the model

Among the wide variety of deep learning models available, the U-Net [8] and its variants are the most widely used architectures for medical image segmentation, and their design is particularly effective in scenarios with limited training data [8], making them well-suited to this project.

Each of the architectures described below can be applied using different spatial processing strategies. Table 4.3 summarises the three main approaches considered.

Table 4.3. Comparison of spatial processing approaches for deep learning segmentation.

Approach	Description	Advantages	Limitations
2D	Each slice processed independently	Low memory; dataset size scales with number of slices; prior group experience	No inter-slice context
2.5D	Adjacent slices used as additional input channels	Partial volumetric context; moderate memory	Limited depth information; adds input complexity
3D	Full volumes processed at once	Full spatial context across all axes	High memory; small datasets increase overfitting risk

The U-Net architecture (Ronneberger et al., 2015) [8] is the dominant architecture for medical image segmentation. Its name comes from its U-shaped structure, which consists of two main parts: an encoder (that compresses the image) and a decoder (that reconstructs it). These two parts are connected by skip connections that pass information directly from each encoder layer to its corresponding decoder layer. At the centre of the network lies the bottleneck, which can be understood as a compact, low-dimensional representation of the input image that captures its most important features. By combining this high-level information from the bottleneck with fine spatial details preserved through the skip connections, the U-

Net can accurately segment small and complex structures, even when only limited training data is available [8].

The AttentionU-Net (Oktay et al., 2018) [23] extends the standard U-Net by introducing soft attention gates on the skip connections. These gates suppress the contribution of irrelevant or noisy features from the encoder before they are concatenated with the decoder features, focusing the network's attention on the target structure. This is particularly beneficial for small structures embedded in a large background, precisely the scenario encountered in postmenopausal ovary segmentation.

nnUNet (Isensee et al., 2021) [22] is an automated medical image segmentation framework that automatically adapts its architecture, preprocessing, training, and post-processing to any new dataset. Given a dataset, it automatically determines the appropriate resampling target, normalisation strategy, network topology, patch size, and batch size. nnUNet has achieved state-of-the-art performance on several medical image segmentation benchmarks [22].

Other architectures were evaluated before settling on the final selection. Transformer-based models such as **Swin-UNETR** [24] relate distant image regions through global self-attention but require large training sets to generalise, a constraint that makes them unsuitable for this dataset. **UNET-R** [25] pairs a Vision Transformer encoder with a convolutional decoder but shares the same data requirement. **TrUE-Net** [26] applies a separate U-Net to each anatomical plane and merges the outputs, which adds volumetric context with little memory overhead; it was excluded here because it was designed and validated for brain lesion segmentation. U-Net and its attention-gated variant were chosen for their consistent performance in limited-data settings.

Table 4.4 presents a comparative summary of the architectures evaluated.

Table 4.4. Deep learning architectures evaluated for ovarian segmentation.

Architecture	Key properties	Limitations	Role in project
U-Net [8]	Symmetric encoder-decoder with skip connections. Effective in limited-data settings.	No built-in mechanism to focus on small structures; all skip features weighted equally.	Baseline experiments.
AttentionU-Net [23]	Soft attention gates on skip connections suppress background features. Particularly suited to small structures in large FOV.	Slightly higher computational cost than U-Net.	Primary architecture - patches and localisation.
U-Net	Captures inter-slice context and volumetric continuity.	High memory demand; requires large datasets.	Explored but not selected.
nnUNet [22]	Automatically adapts architecture, preprocessing, patch size, and augmentation to any new dataset.	Heuristics designed for large datasets. Computationally demanding.	Evaluated as additional architecture.

U-Net and AttentionU-Net were evaluated in both 2D and 3D configurations. The 2D configuration was selected as the primary approach; 3D variants were also explored and are described in further sections.

The 2D AttentionU-Net [23] was selected as the primary architecture. Processing images slice-by-slice effectively increases the number of training samples, reduces GPU memory requirements, and is well-suited to relative small datasets. The soft attention gates on the skip connections suppress irrelevant background features before concatenation with the decoder, potentially focusing the network on the small

ovarian region, a property particularly valuable for postmenopausal ovary segmentation, where the target structure is embedded in a large pelvic field of view with structurally similar surrounding tissues [6].

The 2D nnUNet [22] was also evaluated. Despite its strong benchmarks on large datasets, its self-configuration heuristics are designed for several hundred volumes.

Class imbalance and loss function

Postmenopausal ovaries occupy a very small fraction of the pelvic field of view, typically well below 1% of voxels. Without mitigation strategies, training on full-resolution images with standard cross-entropy loss is prone to converging to the trivial all-background solution, as background voxels overwhelmingly dominate the gradient signal. Two complementary strategies were considered to address this:

- **Patch-based training** [22]: Small crops centred on the ground-truth ovary centroid are extracted during training, so the target structure occupies a much larger fraction of each sample than it would in a full slice. This directly reduces class imbalance and concentrates the gradient signal on the relevant anatomy. A complementary approach is to train only on 2D slices that contain or are adjacent to the target structure, which has a similar effect. Patch-based training was chosen because it constrains the spatial context directly around the target, without depending on slice selection thresholds.
- **Loss function selection**: Dice Loss was selected as the primary candidate, as it directly optimises the Dice, the primary evaluation metric, regardless of the foreground/background ratio. If Dice Loss proved insufficient, alternative loss functions would be explored as shown in Table 4.5.

Table 4.5. Loss functions evaluated for segmentation and localisation training

Loss function	Description	Role in this project
Dice Loss	Directly optimises the overlap metric (1 - Dice). Robust to class imbalance as it normalises by total positives rather than all voxels.	Primary loss for all segmentation experiments.
Dice + Cross-Entropy (DiceCE)	Combination of Dice and cross-entropy. Cross-entropy stabilises early-training gradients; Dice ensures class-imbalance robustness later.	Used in selected patch-based experiments.
Binary Cross-Entropy (BCE)	Pixel-wise loss. Dominated by background class in highly imbalanced datasets, causing networks to converge to all-background prediction.	Tested in early baselines; discarded.
Mean Squared Error (MSE)	Regression loss over continuous heatmap values. Not used for segmentation; applied exclusively to the heatmap-based localisation sub-task.	Used for heatmap-based localisation models.
Focal Loss	Down-weights easy negatives by (1-p) ^2, focusing learning on hard examples. Designed for extreme class imbalance.	Considered but not adopted.

Hyperparameter tuning

Hyperparameter tuning refers to the selection of training configuration parameters such as learning rate, batch size, or dropout rate to optimise model performance. Exhaustive search methods, such as grid

search or random search, are computationally prohibitive for deep learning models. Given that the primary objective of this project is to explore the feasibility of deep learning for postmenopausal ovary segmentation rather than to produce a fully optimised system, hyperparameters were set based on values reported in the medical image segmentation literature and refined iteratively based on observed validation curves.

Internal validation strategies

Several validation strategies were considered to assess model performance. Table 4.6 presents a comparative overview.

Table 4.6. Internal validation strategies considered.

Strategy	Description	Assessment
Hold-Out (patient-level)	Dataset split into train/validation/test ensuring no patient appears in more than one split. Fixed seed ensures reproducibility across all experiments.	Selected: computationally efficient and sufficient.
K-Fold Cross-Validation	Model trained K times with different validation folds. Improves generalisation estimation at the cost of K times training time.	Rejected due to computational and time constraints.
Leave-One-Out (LOOCV)	Each patient used once for validation. Computationally prohibitive for datasets of this size.	Rejected: due to computational and time constraints.

A patient-level hold-out split was selected as the validation strategy, applied consistently across all tasks (segmentation, localisation, and radiomics) to ensure strict data separation between model development and evaluation. All reported performance metrics are computed on the held-out test set, which was not used during model development.

Evaluation metrics

After training, evaluating model performance is a crucial step. Several commonly used metrics in medical image segmentation were considered. Table 4.7 summarises the metrics evaluated and their role in this project.

Table 4.7. Overview of evaluation metrics considered for segmentation and localisation.

Metric	Description	Use in this project
Dice Similarity Coeff. (Dice)	Measures overlap between predicted and ground-truth masks. Ranges 0–1. Standard metric for medical image segmentation.	Primary metric for segmentation models.
Centroid distance (px)	Euclidean distance between predicted and ground-truth centroid per slice. Used to evaluate localisation models.	Primary metric for localisation models.
IoU (Jaccard)	Ratio of intersection to union of the prediction and ground truth. Used in bounding box localisation experiments to evaluate predicted vs. ground-truth region	Preliminary bounding box localisation experiments.

	overlap. Not used as a segmentation metric..	
(Hausdorff Distance 95th percentile (HD95)(mm)	95th percentile of the minimum distance from each predicted boundary point to the ground truth. Robust to outliers; measures boundary accuracy.	Considered; complementary to Dice in this project.
Detection Rate	Proportion of patients correctly detected, defined as achieving a patient-level Dice ≥ 0.10. Provides a clinically interpretable binary summary of segmentation success per patient.	Used as a complementary metric.

4.1.5. Post-processing

Several post-processing techniques were evaluated to improve the quality of the predicted segmentation masks:

- Largest Connected Component (LCC): Post-processing technique that retains only the largest contiguous predicted region, removing spurious small detections. Applied independently to each anatomical side to obtain one segmentation per ovary. This approach is valid because ovaries are compact, single structures, one per side. It would not be appropriate for pathologies involving multiple separate lesions, such as metastatic disease. Applied in the final inference pipeline.
- Exclusion masking: Generating masks of anatomical regions known not to contain ovaries (e.g., bladder, uterus) and removing corresponding voxels from the prediction. Considered but not implemented. At inference time, masks for adjacent structures are not available without a pre-trained whole-body segmentation model; the annotation gap in this dataset is therefore not the limiting factor. Additionally, the patch-based approach already constrains the prediction to a small region around the ovary, reducing the risk of confusing adjacent structures. This could be addressed with a validated whole-body segmentation model such as TotalSegmentator [27], which provides anatomical structure masks without manual annotation.
- Majority voting: Combining predictions from multiple independently trained models to improve robustness. Considered for ensemble use but not applied, given the focus on a single primary architecture and time limitations.

4.1.6. Radiomics

For radiomic feature extraction, PyRadiomics [28] and SlicerRadiomics were evaluated. Table 4.8 presents the comparison.

Table 4.8. Radiomics tools evaluated.

Tool	Description	Decision
PyRadiomics [28]	Open-source Python library for radiomic feature extraction. Fully compliant with the IBSI [21], ensuring reproducibility. Extracts over 100 features: shape, first-order intensity, texture (GLCM, GLRLM, GLSZM).	Selected: IBSI (Image Biomarker Standardisation Initiative) compliance and Python integration.

SlicerRadiomics	Graphical 3D Slicer extension based on PyRadiomics. Useful for interactive inspection but limited automation in scripted pipelines.	Not used: insufficient automation.

PyRadiomics [28] was selected as the only viable option. It is the de facto standard for reproducible radiomics in Python, is fully compliant with the IBSI [21], and integrates seamlessly with NIfTI-based workflows. Manual selection of a feature subset was considered but not pursued, as PyRadiomics already provides a comprehensive and standardised feature set.

Several classifiers were evaluated for the binary age prediction task. Table 4.9 summarises the options considered.

Table 4.9. Classification models considered for binary age prediction.

Classifier	Description	Decision
Random Forest (RF) [29]	Ensemble of decision trees. Robust to overfitting on small datasets; provides feature importances.	Selected.
Support Vector Machine (SVM) [29]	Finds optimal separating hyperplane in high-dimensional feature spaces.	Considered, not adopted.
K-Nearest Neighbours (KNN) [29]	Classifies based on similarity to the k nearest training samples. Sensitive to dataset size.	Considered, not adopted.

4.2. Proposed solution

The proposed solution is designed to balance segmentation performance, computational feasibility, and development efficiency within the constraints of a 103 patient single-institution clinical dataset acquired at Hospital Clinic de Barcelona.

For visualisation and qualitative quality control, 3D Slicer is used throughout the project. All images are converted from DICOM to NIfTI using dicom2nifti and organised in a standardised dataset structure, with geometric inconsistencies (matrix transposition, mask misalignments) corrected by dedicated preprocessing scripts before any model training is carried out.

The primary segmentation model is a 2D AttentionU-Net [23] implemented in MONAI, trained on patches centred on the ground-truth ovary centroid using Dice Loss and on-the-fly data augmentation. This patch-based strategy is the central design choice for addressing class imbalance [22]. Full-resolution 2D AttentionU-Net experiments are run as baselines to characterise the failure modes of training without patch extraction. The 3D nnUNet framework [22] is evaluated under the same dataset split.

To enable inference on new patients without any manual segmentation, a two-stage localisation pipeline is also explored: a lightweight CropCNN that predicts how much to crop above and below the pelvis from a maximum-intensity projection of the volume, followed by a full-image AttentionU-Net [23] that predicts a Gaussian heatmap centred on the approximate ovary centroid. This localisation step mirrors the design principle validated in the literature [13], in which removing the localisation stage caused performance to collapse from a Dice score of 0.290 to 0.013.

Internal validation uses a patient-level hold-out split (70/15/15%) consistent across all experiments. Segmentation performance is evaluated using the Dice and Detection Rate (the proportion of patients with a patient-level Dice ≥ 0.10), and localisation quality is evaluated using centroid distance in pixels. Post-processing consists of selecting the LCC on each ovarian side.

Finally, radiomic features are extracted from the ground-truth ovarian masks using PyRadiomics [28], with voxel resampling and z-score normalisation as recommended for multi-scanner datasets [21]. A Random Forest classifier is trained for binary age prediction as an exploratory downstream application, illustrating the segmentation pipeline's potential for quantitative ovarian characterisation.

5. DETAILED ENGINEERING

5.1. Development environment

5.1.1. Hardware resources.

Training, inference, and compute-intensive preprocessing all ran on the Alfa server, hosted by the Biomedical Imaging Group at the University of Barcelona. The server has an NVIDIA RTX A5000 GPU with 24 GB of VRAM, shared among group members. Datasets, intermediate outputs, and model weights were stored on its shared file system in a project-specific directory.

The server was accessed via the University's VPN and SSH. Local machines were used only to edit code and inspect outputs, with all execution taking place remotely.

5.1.2. Software environment and libraries

All code was written in Python 3.10 and run on the Alfa server within a dedicated Conda environment, with library versions pinned to ensure all experiments are reproducible. Visual Studio Code was used locally as the development environment for editing code. All execution took place on the Alfa server, accessed exclusively through an SSH terminal connection.

The deep learning pipeline relied on *PyTorch* as the underlying framework and *MONAI* for the segmentation architectures, data transforms, and loss and metric computation. The *nnU-Net v2* framework was installed separately for the self-configuring 2D experiments. Image reading and writing used *nibabel* for NIfTI files and *dicom2nifti* for the original DICOM acquisitions; *scikit-image* handled resizing and connected-component analysis. *NumPy* and *pandas* covered array operations and tabular data; patient splits and experiment settings were stored as YAML files so that every run could be reproduced exactly. *Matplotlib* produced training curves and segmentation figures. *PyRadiomics* extracted radiomic features from the ground-truth masks, and *scikit-learn* provided the Random Forest classifier for the age prediction task.

5.2. Dataset description and preparation

The clinical dataset used in this project consists of 103 postmenopausal patients who underwent T2-weighted pelvic MRI at the Hospital Clínic de Barcelona. The data were provided directly by the hospital's Radiology Department, with all direct patient identifiers removed and each case represented by an anonymous internal identifier. The conversion of raw clinical data into a consistent dataset suitable for deep learning required a series of preprocessing steps, which are explained in the following subsections.

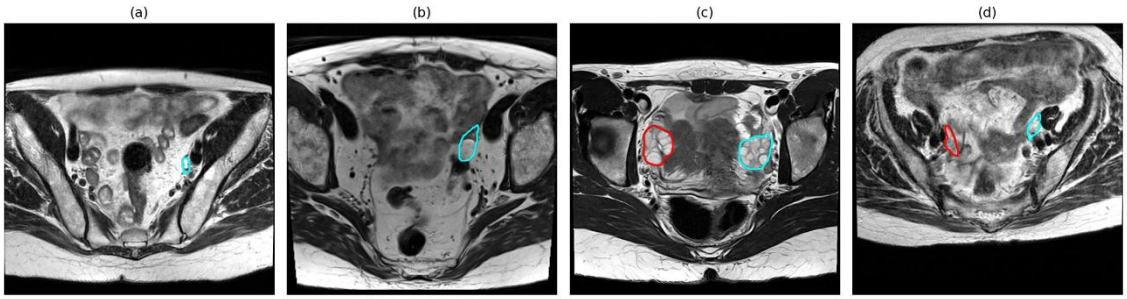

Figure 5.1. Representative T2-weighted MRI axial slices with ground-truth segmentation masks from the clinical dataset. Right ovary contours are shown in red, left ovary contours in blue (or cyan)

5.2.1. Dataset overview and anonymisation

The dataset was delivered as a flat directory containing one folder per patient. For each patient, the original T2-weighted MRI volume was stored as a series of DICOM files inside a baseline subfolder. The segmentation masks of the right and left ovaries, manually delineated by the clinical radiologist using Mint Lesion, were provided as separate NIfTI files along with their corresponding JSON metadata. All acquisitions correspond to axial (or axial-oblique) T2-weighted uterine protocols; the orientation was later verified geometrically, as discussed in Section 5.3.

The hospital performed pseudonymisation before the data left the clinical environment, in accordance with the ethics protocol of the project. All scripts, file names, and intermediate outputs in the project use these pseudonymised identifiers exclusively, and no protected health information is stored alongside the images. The dataset was kept on the Alfa server within a project-specific directory accessible only to authorised members of the Biomedical Imaging Group.

5.2.2. DICOM to NIfTI conversion

As described in Section 4.1.2, NIfTI was selected as the working format since the ground-truth masks were already delivered in this format. All MRI volumes were therefore converted from DICOM to NIfTI using dicom2nifti, with compression and automatic reorientation enabled. A conversion script processed each patient and handled special cases.

The most relevant special case was a patient whose acquisition was split into two independent DICOM series, one per ovary. Both series were separately converted, and the pipeline was made compatible with either a single image or a pair of side-specific images per patient.

After conversion, the dataset was reorganised into a simple layout, with one folder per patient containing the converted T2-weighted image, the right and left ovary masks renamed homogeneously, and their associated JSON files.

5.2.3. Dataset inspection and quality control

Once the dataset had been converted and reorganised, an inspection script was run to verify the integrity of every patient before any model training was attempted. For each patient, the script checked the existence of the expected files, the absence of empty or corrupted volumes, the agreement between mask and image shapes, the presence of segmented voxels in the masks, and the validity of the JSON metadata.

This quality control phase identified a small number of recurrent problems. Twelve patients had only one of the two ovaries segmented; these patients were retained, and their single available mask was used during training. In addition, one patient had an ovary mask that delineated the urinary bladder, with a volume of approximately 88 cm^3, clearly inconsistent with a postmenopausal ovary. This anomaly was flagged in the inspection report and excluded from the localisation pipeline but not from the main segmentation experiments, where it represents less than 1% of the training data. DICOM metadata was also extracted during this phase and used to characterise the dataset in Section 5.3.

The same inspection also identified two geometric issues that required correction. In 13 patients, the volume had been wrongly transposed during conversion, with the slice dimension placed in the second array position instead of the last; this was corrected by reordering the array axes and updating the affine matrix accordingly.

A more subtle issue affected some image–mask pairs: the masks covered only the relevant anatomical slices and had fewer slices than the full image volume. Although this appeared correct in visualisation tools such as 3D Slicer, loading both arrays into a deep learning model would cause shape mismatches. Each mask was therefore reconstructed to match the full image dimensions by placing each slice at its correct position along the Z axis, derived from the original DICOM metadata, and applying a Y-axis flip to reconcile the coordinate conventions between the image and the mask.

5.3. Dataset characterisation

The 103 patients dataset comes from the Hospital Clínic de Barcelona. 64 (62%) were scanned on GE equipment and 39 (38%) on Siemens, with most acquisitions at 1.5 T and the rest at 3 T. The vendor and field-strength mix reflects a routine clinical workflow rather than a purpose-built research cohort.

All acquisitions follow an axial or axial-oblique orientation, and DICOM-to-NIfTI conversion was run with automatic reorientation enabled, so every volume ends up in a consistent orientation with no further reorientation needed at training time.

Most patients fall within the expected range for natural menopause (> 45 years); however, the dataset includes a small number of cases below this threshold. This variability can be regarded as a strength of the dataset, as it reflects the diversity encountered in real clinical practice.

Table 5.1. Dataset characterisation summary.

Characteristic	Value		
Slice thickness (mm)	3.0 - 4.5		
In-plane resolution (mm)	0.23 - 0.94 (mode: 0.47)		
Repetition Time (TR)(ms)	1,800 - 8,830		
Patient age (years)	Median: 65.1	Range: 15.7 - 95.6	Mean: 63.1
Right ovary volume (cm³)	Median: 2.08	IQR: 0.94 - 4.72	Range: 0.12 - 88.54
Left ovary volume (cm³)	Median: 2.12	IQR: 0.95–3.49	

IQR: Interquartile range

5.4. Model training

This section describes the model training experiments, structured as a progression: starting from full-resolution baselines, moving to patch-based training once the failure mode was identified, and finally comparing architectures and exploring 3D approaches.

5.4.1. Data splitting, pre-processing and loading

The 103 patients were split at the patient level into training (70%), validation (15%), and test (15%) sets, giving 72 training patients, 15 validation patients, and 16 test patients. A fixed random seed was used to ensure the split is reproducible across runs.

Each volume was loaded from its NIfTI file and resampled in-plane to 256×256 pixels using bilinear interpolation; the number of axial slices was left unchanged, and the same procedure was applied identically in the 3D patch-based experiments. For the full-volume 3D experiments, the Z dimension was additionally standardised to 32 slices via interpolation-based resampling (chosen to cover most patients, whose slice counts ranged from 16 to 54). Intensity normalisation was then applied per volume using a two-step procedure. First, z-score standardisation:

$$\hat{x} = \frac{x - \mu}{\sigma}$$

where x denotes the raw intensity of each voxel, $\hat{x}$ its z-score standardised value and μ and σ the mean and standard deviation of the volume. This was followed by min-max rescaling to the [0, 1] range:

$$x_{\text{norm}} = \frac{\hat{x} - \hat{x}_{\min}}{\hat{x}_{\max} - \hat{x}_{\min}}$$

This two-step normalisation was applied independently to each volume, avoiding any assumption about consistent absolute intensities across scanners or field strengths.

Training was carried out at the slice level rather than the volume level. Since only slices containing at least one annotated ovarian voxel carried a useful supervision signal, the dataset was divided into positive slices (containing the ovary) and negative slices (background only). To address class imbalance, the training set was balanced by sampling an equal number of negative slices. Validation followed the same positive/negative split but without augmentation.

5.4.2. Baseline experiments: full-slice training

The first step was to establish whether a standard approach, training on full axial slices, could work at all.

The first series of experiments trained the network on full axial slices at the original in-plane resolution of 512×512 pixels, without the downsampling applied in the general pipeline (Section 5.4.1). The model was an AttentionU-Net with five encoder levels and channels (16, 32, 64, 128, 256). Several variants were tested, including different resolutions, dropout rates, and loss functions. Results are reported in Section 6.1.

A parallel approach split each axial slice along the left–right anatomical axis, training the model on each half independently to halve the effective field of view without requiring a known centroid.

Neither the full-slice nor the half-image approach achieved satisfactory segmentation performance, motivating a shift to a patch-based strategy, described in the following section.

5.4.3. Patch-based training

With the failure mode identified, the training strategy was redesigned around a single insight: if the model cannot find the ovary in a full slice, constrain its view to a small region around it.

The key insight from the baseline experiments was that the network lacked any spatial prior about where the ovary was located. The patch-based approach addresses this directly: instead of feeding the full slice

to the network, a 96×96 pixel crop centred on the ovary centroid is extracted from each positive slice, and a random 96×96 crop is extracted from each negative slice. This forces the network to focus exclusively on the region of interest, reducing the effective class imbalance by orders of magnitude.

The architecture and most hyperparameters were kept identical to the baseline: AttentionU-Net with channels (16, 32, 64, 128, 256), Adam optimiser at lr=10^-4, DiceLoss, and ReduceLROnPlateau scheduler (factor 0.5, patience 10), and the same augmentation strategy of horizontal flips, vertical flips, and 90° rotations on positive patches. The only structural change was the patch extraction itself. The Adam optimiser adjusts each parameter's learning rate based on gradient history; ReduceLROnPlateau monitors validation loss and halves the learning rate if no improvement is seen for 10 consecutive epochs.

A final variant introduced random jitter to the patch centre during training: instead of always centring on the exact ovary centroid, a uniform offset of up to ±40 pixels was applied independently along each axis. The motivation was to improve robustness at inference, where the centroid must be estimated automatically rather than read from the ground-truth mask.

Two additional variants were also evaluated. One added L2 regularisation (weight decay 10^-5) to the optimiser, with model selection by minimum validation loss. A second variant introduced more aggressive augmentation (intensity jitter, Gaussian noise, zoom, and patch position jitter) combined with DiceCE loss.

5.4.4. Architecture comparison

With the patch-based setup established, the next question was whether the choice of architecture made a meaningful difference.

To assess whether the choice of architecture contributed meaningfully to the results, three alternatives were evaluated under identical conditions to the best patch-based model: a standard U-Net, UNETR, and SwinUNETR, all implemented via MONAI. All shared the same patch size (96×96), optimiser (Adam, lr=1e-4, weight decay 1e-5), loss function (DiceLoss), and data pipeline.

The AttentionU-Net was selected as the final architecture for all subsequent experiments.

5.4.5. 3D experiments

In parallel with the 2D work, 3D models were explored to assess whether volumetric context could improve segmentation.

Unlike the 2D models, which process individual axial slices, the 3D models operate on volumetric inputs. In parallel with the 2D patch-based work, a series of experiments explored whether modelling the full volumetric context could improve segmentation.Two strategies were tested, both reusing the same 3D AttentionU-Net configuration as the 2D experiments (channels, optimiser, and loss function unchanged).

The first used full volumes resized to 256×256×32 voxels, with the Z dimension standardised by padding or cropping to fit a consistent batch size of 2. .This mirrored the 2D full-slice baseline in three dimensions: the model again collapsed to predicting empty masks and never recovered.

The second strategy applied 3D patch extraction, using a 96×96×16 voxel patch centred on the ovary centroid; the XY dimensions matched the 2D setup, with Z set to 16 to cover approximately the full axial extent of the ovary. The batch size was reduced to 4 due to GPU memory constraints.

5.4.6. nnU-Net

As an independent check, nnU-Net was also tested, a self-configuring framework that had shown strong results on other medical segmentation tasks.

nnU-Net (version 2) [22] was tested as an alternative to the manually designed pipelines. Unlike the other models in this chapter, nnU-Net automatically determines its training configuration from the dataset (patch size, batch size, architecture depth, and preprocessing) without any manual hyperparameter choices. It has a strong track record on medical image segmentation [13,27], so it was a natural candidate to try here.

The dataset was converted to nnU-Net format, and the automatic planner chose the 2D configuration and set the patch size to 512x512 pixels, which covers the full in-plane field of view of each slice. In other words, it processes whole slices rather than the smaller crops used in the patch-based strategy. The batch size was set to 12. The architecture was a UNet with eight encoder stages, feature maps ranging from 32 to 512 channels, Z-score normalisation, and a polynomial learning rate schedule starting at 0.01. Data were divided into five cross-validation folds; the same fold was used for both runs, giving 132 training and 34 validation cases.

Two training runs were attempted. The standard trainer ran for 1,000 epochs. A second run with 2000 epochs was launched on the assumption that training had simply not converged yet.

5.5. Localisation pipeline

The patch-based training strategy in Section 5.4.3 relied on ground-truth masks to compute the ovary centroid per slice, which was used to centre each 96 × 96 crop. At inference time, no masks are available, so the centroid must be estimated automatically before the segmentation model can be applied.

Although the heatmap-based pipeline was the initial design choice, preliminary experiments showed it underperformed relative to direct regression-based alternatives. Two regression-based approaches were therefore developed and trained instead, while the heatmap-based family was kept as a discarded baseline (see below).

Bounding box regression. *LocalitzadorCNN* is a lightweight convolutional network that regresses the bounding boxes of both ovaries from a single resized slice. Five encoder blocks (16 to 256 feature maps) feed into a fully connected regression head with sigmoid output, predicting eight normalised coordinates in a single forward pass. Some patients only have one ovary annotated (the other being absent or not delineated in the ground truth). To avoid penalising the model for "missing" coordinates it cannot learn from, a per-sample mask marks which of the two ovaries are actually present, and the MSE loss is computed only over those. Training used Adam with a learning rate of 10^{-4} and weight decay of 10^{-4}, with early stopping after 50 epochs without improvement. Two runs were conducted. The achieved

localisation precision was not sufficient for reliable 96×96 patch placement over structures as small as postmenopausal ovaries.

Two-stage crop and segmentation. The second approach splits localisation into two steps. A CropCNN takes the maximum-intensity projection of the full volume and predicts two scalar values: the fractions to remove from the top and bottom of the image to isolate the pelvic rows containing the ovaries. In the cropped region, a dedicated AttentionU-Net was trained using 128 × 128 patches and applied during inference via a sliding window (50% overlap, Gaussian weighting). The resulting probability map was thresholded at 0.5 and split at the horizontal midline; the centroid of the largest connected component on each side gave the estimated ovary position.

5.6. Inference pipeline

Model evaluation on the test set followed the same patch extraction procedure used during training, with patches centred on the ground-truth ovary centroid (Section 5.4.3). The model output was then passed through sigmoid activation and thresholded at 0.5 to produce a binary prediction.

Performance was measured on the 16 test patients using two metrics. The first was the mean Dice Similarity Coefficient, aggregated across all patches from all positive slices, with patient-level scores obtained by averaging across all positive slices per patient. The second was the patient-level detection rate: the proportion of patients whose averaged Dice score reached or exceeded 0.10.

5.7. Radiomics

As a downstream application, a radiomics analysis was conducted to explore whether quantitative features extracted from the ovarian masks carry information about patient age. Postmenopausal ovarian tissue undergoes progressive atrophy with age, so shape and texture features derived from the ground-truth segmentations may reflect some of that change.

Feature extraction used PyRadiomics on the ground-truth masks of both ovaries, applied to the original image without resampling to preserve native voxel geometry, with a bin width of 25 for histogram-based features. The extraction produced 107 features per ovary across seven categories: 14 shape descriptors, 18 first-order intensity statistics, 24 Grey-Level Co-occurrence Matrix (GLCM) texture features, 16 Grey-Level Run-Length Matrix (GLRLM) features, 16 Grey-Level Size Zone Matrix (GLSZM) features, 14 Grey-Level Dependence Matrix (GLDM) features, and 5 Neighbouring Grey-Tone Difference Matrix (NGTDM) features. Where both ovaries were annotated, the two feature vectors were averaged into a single per-patient representation.

An initial regression attempt trained a Random Forest to predict age directly from the 107 features. The error was too large to be informative, so the task was reformulated as binary classification.

The threshold was set at the cohort median age of 65 years. Three patients with implausible DICOM ages below 30 years were excluded. 49 were aged 65 or below and 48 above. A 500-tree Random Forest was evaluated by 5-fold stratified cross-validation. Following evaluation, the model was retrained on the full dataset, and feature importances were computed using the mean decrease in impurity (MDI), yielding a ranking of all 107 features by their contribution to age discrimination.

6. RESULTS

6.1. Segmentation

6.1.1. Full volume experiments

The first series of experiments trained an AttentionU-Net on full 2D axial slices. Table 6.1 shows the best validation Dice obtained across all full-slice variants. The model architecture was fixed throughout: five encoder levels with channel sizes (16, 32, 64, 128, 256), an Adam optimiser with lr=1e-4, and a ReduceLROnPlateau scheduler.

Table 6.1. Validation Dice scores for 2D full-image and half-image segmentation strategies.

Experiment	Configuration	Best Validation Dice
Full slice	512x512, DiceCELoss	0.006
Full slice	256x256, DiceCELoss	0.073
Full slice	256x256, DiceLoss, dropout 0.3, augmentation	0.161
Half-image	256x256, left/right split	0.237

At 512×512, the model collapsed to predicting empty masks from the first epoch and never recovered, producing a validation Dice of 0.006 throughout training. Downsampling to 256×256 yielded a first measurable validation Dice of 0.073. Switching the loss from DiceCE to Dice, increasing dropout from 0.2 to 0.3, and introducing augmentation (horizontal and vertical flips, 90° rotations with doubled positive slices) pushed the result to 0.161.

A parallel approach trained the model on left and right half-images independently, aiming to reduce the effective field of view without requiring a centroid estimate. Two configurations were tested: using DiceLoss alone, the model reached a best validation Dice of 0.237; replacing it with a combined DiceCE loss caused convergence to collapse at 0.016. In both cases, test performance was poor (median Dice 0.013, 14/29 cases below 0.01).

6.1.2. Patch-based experiments

Replacing full slices with 96×96 crops centred on the ground-truth ovary centroid immediately resolved the class imbalance. Table 6.2 summarises the patch-based variants.

Table 6.2. Validation Dice scores for 2D patch-based segmentation experiments.

Experiment	Key change	Best val Dice
Patches	96x96, centroid-centred	0.492
Patches + centroid jitter	±40 px position noise during training	0.253
Patches + weight decay	Weight decay 1e-5	0.619
Patches + aggressive augmentation	DiceCE, intensity jitter, Gaussian noise, zoom, position jitter	0.419

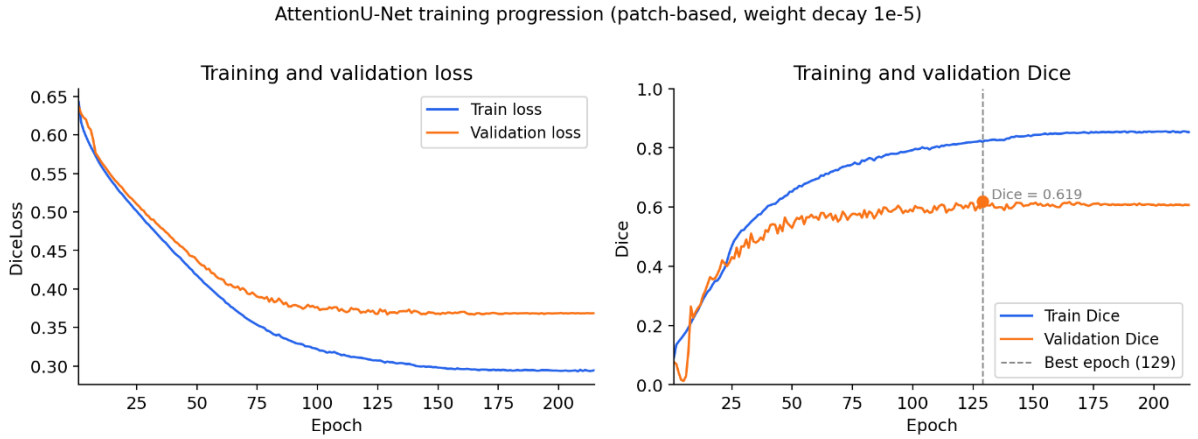

Figure 6.1. Training and validation loss (left) and Dice (right) over 215 epochs for the best-performing model (AttentionU-Net, 96×96 patches, weight decay 1e-5).

The first patch-based experiment reached a validation Dice of 0.492, a threefold improvement over the best full-slice result, with no change to the architecture or optimiser.

Introducing centroid jitter (a uniform offset of up to ±40 pixels applied to the patch centre during training) was intended to improve robustness at inference, where the centroid must be estimated rather than read from the ground truth. The best validation Dice for this variant fell to 0.253.

Adding L2 regularisation (weight decay 1e-5) to the centroid-centred setup yielded a validation Dice of 0.619, the best result across all experiments; the saved model was selected at the epoch of minimum validation loss. Combining a more aggressive augmentation policy, intensity jitter, Gaussian noise, zoom, and patch-position jitter, with a DiceCE loss produced a lower Dice of 0.419.

Figure 6.1 illustrates the training progression for one representative validation patient, showing predictions at epochs 10 and 210.

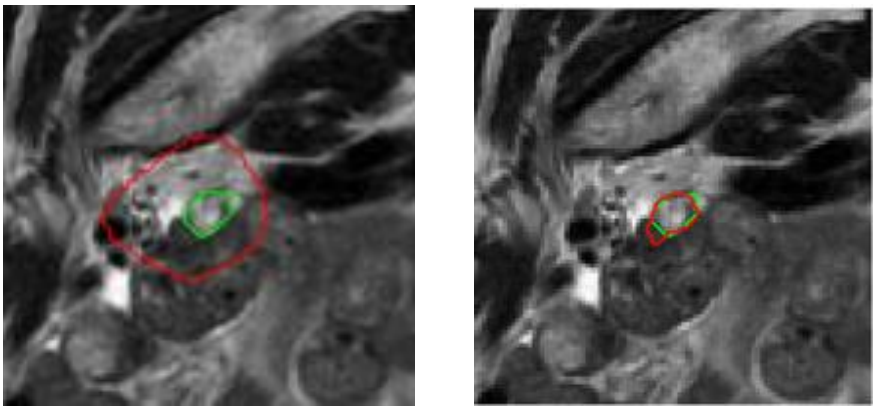

Figure 6.2. Training progression for a representative validation patient. Left: prediction at epoch 10, showing severe over-segmentation (predicted contour in red, ground truth in green). Right: prediction at epoch 210, where the model has converged, and the predicted contour closely matches the ground truth.

6.1.3. Architecture comparison

Table 6.3 compares four architectures evaluated under identical conditions: the patch-based configuration with weight decay 1e-5, 96×96 patches, Adam at lr=1e-4, and DiceLoss.

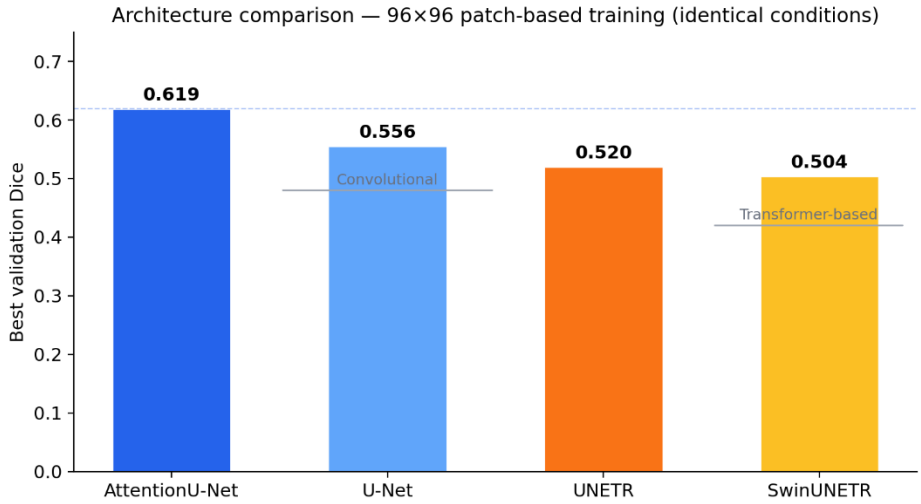

Figure 6.3. Validation Dice comparison across four architectures evaluated under identical conditions.

The AttentionU-Net achieved the highest validation Dice across all architectures, followed by the standard U-Net, UNETR, and SwinUNETR.

6.1.4. 3D experiments

Two 3D strategies were evaluated.

Table 6.4. Validation Dice scores for 3D segmentation experiments.

Configuration	Best val Dice
Full volume, 256x256x32 voxels, batch size 2	0.025
Patches, 96x96x16 voxels, batch size 4	0.175

The full-volume 3D model reproduced the same failure as the 2D full-slice baseline, collapsing to background prediction with a best validation Dice of 0.025. Switching to a patch-based 3D approach brought the score to 0.175; a clear improvement over the 3D baseline, but still well below the 2D result of 0.619.

6.1.5. nnU-Net

nnU-Net was evaluated as a self-configuring reference.

Table 6.5. nnU-Net validation Dice scores at 1,000 and 2,000 training epochs.

Run	Epochs	Final val Dice
nnUNetTrainer	1,000	0.003
nnUNetTrainer_2000epochs	2,000	0.003

Both runs produced a pseudo-Dice below 0.01 throughout training, with final validation Dice scores of 0.003 in both cases. Doubling the training budget had no effect.

6.2. Test set evaluation

The best model (AttentionU-Net, patches, weight decay 1e-5) was evaluated on the 16 held-out test patients: the 96×96 patch was centred on the ground-truth ovary centroid for each positive slice. This

removes the localisation step from the pipeline, so the results represent an upper bound on segmentation performance when localisation is treated as solved.

Table 6.6. Test set performance of the final 2D patch-based model.

Metric	Value
Mean Dice Similarity Coefficient	0.447
Patient-level detection rate (Dice ≥ 0.10)	15/16 (93.75%)

The final model was evaluated on the 16-patient test set using oracle centroids. The mean Dice was 0.447, with a median of 0.431 and a standard deviation of 0.209 (Table 6.6). Two patients achieved Dice scores above 0.70, while the single case below the detection threshold (Dice 0.083) accounted for the only segmentation failure.

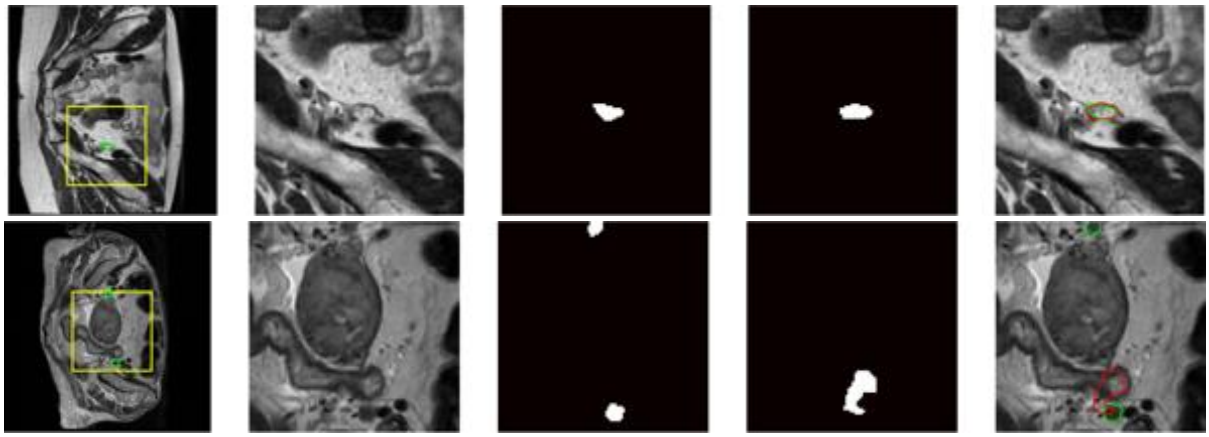

Figure 6.4. Qualitative results on two representative test patients. Each row shows, from left to right: the full axial slice with the centroid marker, the 96×96 input patch, the ground-truth mask, the model prediction, and the overlay of both contours (ground truth in green, prediction in red).

Row 1 illustrates a successful case: the ovary is small and compact, and the predicted contour closely follows the ground truth. Row 2 illustrates a characteristic failure mode: the model segments a large cystic structure that dominates the patch (red contour). At the same time, the ground-truth ovary corresponds to the smaller adjacent structures (green contours). This type of confusion with anatomically prominent neighbouring structures is a known limitation of patch-based approaches trained without explicit anatomical context.

6.3. Localisation pipeline

At inference time, the ground-truth centroid is unavailable. Two automatic localisation approaches were developed and evaluated.

Table 6.7. Localisation network performance on the test set.

Approach	Metric
Bounding box regression (LocalitzadorCNN)	Best val IoU: 0.36
Two-stage: CropCNN + dedicated AttentionU-Net	Crop error: 7.2 px; segmentator val Dice: 0.23

The bounding box regressor (LocalitzadorCNN) reached a best validation IoU of 0.36 across two training runs.

The two-stage approach offered a partial alternative. A dedicated CropCNN, predicting vertical pelvic crop boundaries from a maximum intensity projection, achieved a minimum validation position error of 7.2 pixels, and the AttentionU-Net trained on the resulting crops reached a validation Dice of 0.23 on the cropped pelvic region.

6.4. Radiomics: binary age classification

The classifier achieved 61.9% accuracy, against a majority-class baseline of 50.5%. Sensitivity and specificity were both 65.3%. Shape descriptors dominated the feature importance ranking, consistent with the known reduction in ovarian volume and geometry after menopause.

Table 6.8. Radiomics-based binary age classification results.

Metric	Value
Accuracy	61.9%
Majority-class baseline	50.5%
Sensitivity	65.3%
Specificity	65.3%

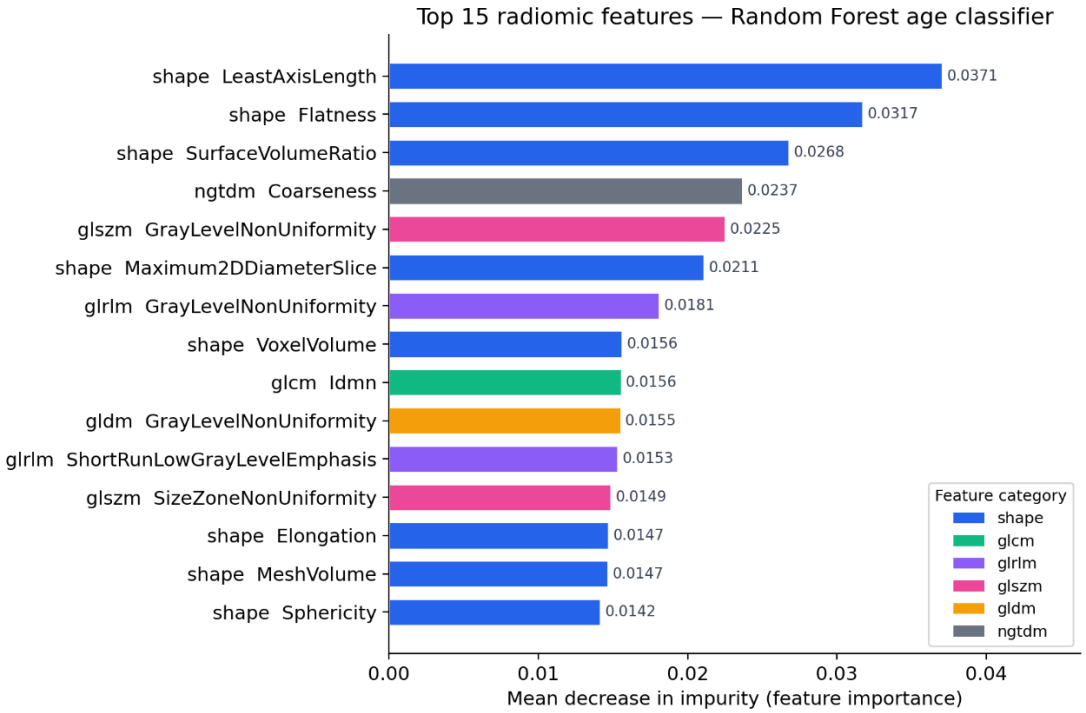

Figure 6.5. Top 15 radiomic features ranked by mean decrease in impurity from the 500-tree Random Forest classifier.

The five highest-ranked features by mean decrease in impurity were Least Axis Length, Flatness, Surface Volume Ratio, Coarseness, and Gray Level Non-Uniformity.

7. DISCUSSION AND FUTURE WORK

7.1. The challenge of small-structure segmentation

A postmenopausal ovary typically measures less than 2 cm in maximum diameter, corresponding to a cross-sectional area of roughly 1–2 cm² against a pelvic field of view of approximately 240 × 240 mm, less than 0.5% of the imaged area. At that scale, the Dice loss gradient during full-resolution training is dominated by background pixels and carries almost no signal from the target. This is why the full-slice baseline collapsed to predicting empty masks from the first epoch and never recovered.

Downsampling to 256×256 reduced the imbalance but did not eliminate it. The network was still required to learn global pelvic anatomy and locate a sub-centimetre structure within it, with no spatial constraint on where the ovary should appear. The half-image experiments, which aimed to shrink the field of view without requiring a centroid estimate, confirmed this: both loss configurations produced poor test generalisation.

Centring 96×96 crops on the ovary centroid addressed the problem at its source. The target now occupied a substantial fraction of each patch rather than a fraction of a percent, and validation Dice increased from 0.161 to 0.492 with no other changes to the architecture or optimiser. This suggests that a substantial part of the network's capacity in full-image training is consumed by learning to ignore irrelevant pelvic structures. Restricting the field of view to the ovarian neighbourhood removes that burden and directs available capacity entirely toward characterising the target.

nnU-Net ran into the same problem. Its automatic configuration chose a 512×512 patch covering the full slice, reproducing the class imbalance that caused the earlier baselines to collapse. Doubling the training budget changed nothing. This is not a comment on the framework generally; nnU-Net performs well on standard organ segmentation tasks. The issue is that its configuration logic derives patch size from image dimensions and available GPU memory, with no term for target structure size. Applied to a sub-centimetre ovary, the resulting foreground-to-background ratio was no different from that of the manual full-slice baseline.

7.2. Architecture and dimensionality choices

Among the four architectures evaluated under identical conditions, AttentionU-Net produced the highest validation Dice. The margin over the standard U-Net was 0.063 points; modest, but attention gates suppress background activations selectively, and for a small, concentrated target, that suppression has a clear opportunity to help. This result is consistent with prior work demonstrating that attention mechanisms improve segmentation of small anatomical structures by suppressing irrelevant background activations.

Transformer-based models (UNETR and SwinUNETR) performed below both convolutional variants. Transformers learn attention patterns from data, and 72 training patients are not enough for them to do so reliably. This is a known pattern in medical imaging: vision transformers tend to outperform convolutional networks at scale, but underperform them when training data is limited.

The 2D patch-based approach outperformed 3D by a wide margin (0.619 versus 0.175). Part of this is structural. The acquisitions are anisotropic: in-plane spacing is approximately 0.47 mm, while slice thickness ranges from 3 to 5.5 mm, so a typical ovary spans only three to five slices in the Z direction. A 3D model that treats all axes symmetrically cannot make effective use of a depth axis with so little spatial detail. Training efficiency also differs: the 2D model extracts one patch per positive slice per patient per epoch, while the 3D model extracts one patch per patient, so the 2D model accumulates far more gradient updates over the same number of epochs.

7.3. Pipeline performance and the localisation bottleneck

Before discussing these results, it is important to note that all segmentation metrics in this work were computed with the patch centred on the ground-truth ovary centroid. Under real inference conditions that centroid is unavailable. The reported Dice of 0.447 therefore represents an upper bound on what a fully automatic pipeline could achieve, not an estimate of end-to-end performance. The gap between this figure and the best localisation result (CropCNN validation Dice 0.23) illustrates how much performance depends on accurate centroid prediction.

The final model achieved a mean Dice of 0.447 on the test set, with a standard deviation of 0.209. The mean sits slightly above the median (0.431), reflecting a modest right skew: most patients fall in a mid-range band, with a small number of well-segmented cases pulling the average up. The spread is wide, and differences in ovary size, cystic content, and ground-truth mask quality across patients are the most plausible sources of that variability.A test DSC of 0.447, while moderate in absolute terms, exceeds the 0.290 reported by RAovSeg [13], the closest published pipeline for normal ovary segmentation in T2-weighted MRI, and represents a 178% improvement over the full-slice baseline (0.161). Higher Dice scores in the literature (0.73–0.94) correspond to macroscopic pathological masses (ovarian tumours or lesions) which are substantially larger and more conspicuous than the atrophied postmenopausal ovary targeted here.

As mentioned before, all of this was measured with the patch centred on the ground-truth centroid. At inference time, that centroid is unavailable. The bounding box regressor reached a best validation IoU of 0.36, which is not enough. At that precision, the predicted centre can sit far enough from the true centroid to exclude the ovary from the patch entirely. The two-stage CropCNN approach reduced the search space but did not solve the problem, reaching a validation Dice of 0.23 on cropped pelvic regions.

Localisation is the main unresolved problem in the pipeline. Small target size, low soft-tissue contrast, and the absence of reliable anatomical landmarks make it harder than segmentation.

7.4. Radiomics findings and directions for future work

A cross-validated accuracy of 61.9% on a balanced binary task reflects a weak classifier. The result is consistent with the expected difficulty: 97 patients and 107 radiomic features is a challenging sample-to-feature ratio, no explicit feature selection was applied before training, and cross-validation estimates at this scale carry wide confidence intervals. The finding is best treated as a preliminary signal, sufficient to suggest that shape features encode age-related morphological information, but not as evidence of a clinically useful predictor.

The most pressing gap is localisation. The segmentation model cannot run without a centroid estimate, and neither approach developed here provided one accurate enough. A dedicated detection network trained on full pelvic volumes would address this directly; foundation models pre-trained on large medical imaging datasets are another option, particularly given the limited data available here.

Dataset size constrains what can be concluded from all parts of this work. With 103 patients from a single institution, generalisation across scanners and acquisition protocols is untested. More data would also allow revisiting transformer architectures and 3D models, both of which were disadvantaged here by data scarcity and acquisition anisotropy rather than by any fundamental unsuitability.

The radiomics features in this study were extracted from ground-truth masks. Once segmentation quality is sufficient to produce reliable predicted masks, the analysis should be repeated on those, since understanding how segmentation error propagates into radiomic features is a prerequisite for any clinical application downstream.

8. PROJECT SCHEDULE

8.1. Work Breakdown Structure

The project runs from February 2 to May 28, 2026, covering 80 working days. Easter (April 2, 3, and 6) and Labour Day (May 1) are excluded. The WBS in Figure 8.1 groups the twenty tasks by function across seven phases: Initialisation, Data Preparation, Deep Learning, Localisation, Radiomics, Evaluation, and Documentation. This organisation reflects the project scope rather than its timeline.

Table 8.1. Listing of each task with its identifier, duration in working days, and predecessor tasks.

ID	Task name	Duration (days)	Predecessors	Phase
Phase 1 – Initialisation				
1.1	Bibliographic research	5	-	Initialisation
1.2	Objectives and scope	4	1.1	Initialisation
1.3	Project planning	2	1.2	Initialisation
Phase 2 – Data preparation				
2.1	Server setup and data access	2	1.3	Data preparation
2.2	Dataset inspection and Quality control	5	2.1	Data preparation
2.3	Image preprocessing	7	2.2	Data preparation
2.4	Dataset characterisation	3	2.3	Data preparation
Phase 3 – Deep Learning				
3.1	Full-slice 2D baseline	6	2.4	Deep Learning
3.2	Patch-based 2D training	8	3.1	Deep Learning
3.3	Architecture comparison	6	3.2	Deep Learning
3.4	3D segmentation experiments	7	3.2	Deep Learning
3.5	nnU-Net experiments	5	3.2	Deep Learning
Phase 4 – Localisation				
4.1	Localitzador CNN	6	3.2	Localisation
4.2	Two-stage CNN pipeline	6	4.1	Localisation
Phase 5 – Radiomics				
5.1	Radiomics feature extraction	4	2.4	Radiomics
5.2	Binary age classification	4	5.1	Radiomics
Phase 6 – Evaluation				
6.1	Test set evaluation	4	3.3, 3.4, 3.5, 4.2, 5.2	Evaluation
6.2	Results analysis	4	6.1	Evaluation
Phase 7 – Documentation				
7.1	Memory writing	12	6.2	Documentation
7.2	Presentation preparation	6	7.1	Documentation

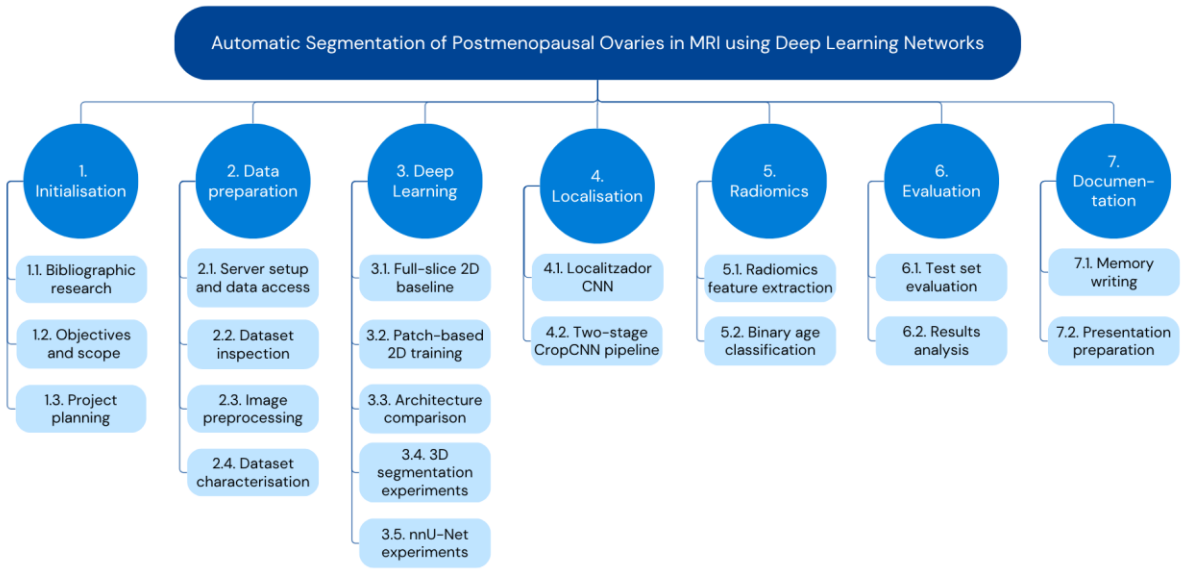

Figure 8.1. Work Breakdown Structure

8.2. PERT Diagram

The PERT diagram in Figure 8.2 follows Activity-on-Arrow notation, where arrows represent tasks and nodes are events; the moments when preceding tasks have all finished, and following ones can begin.

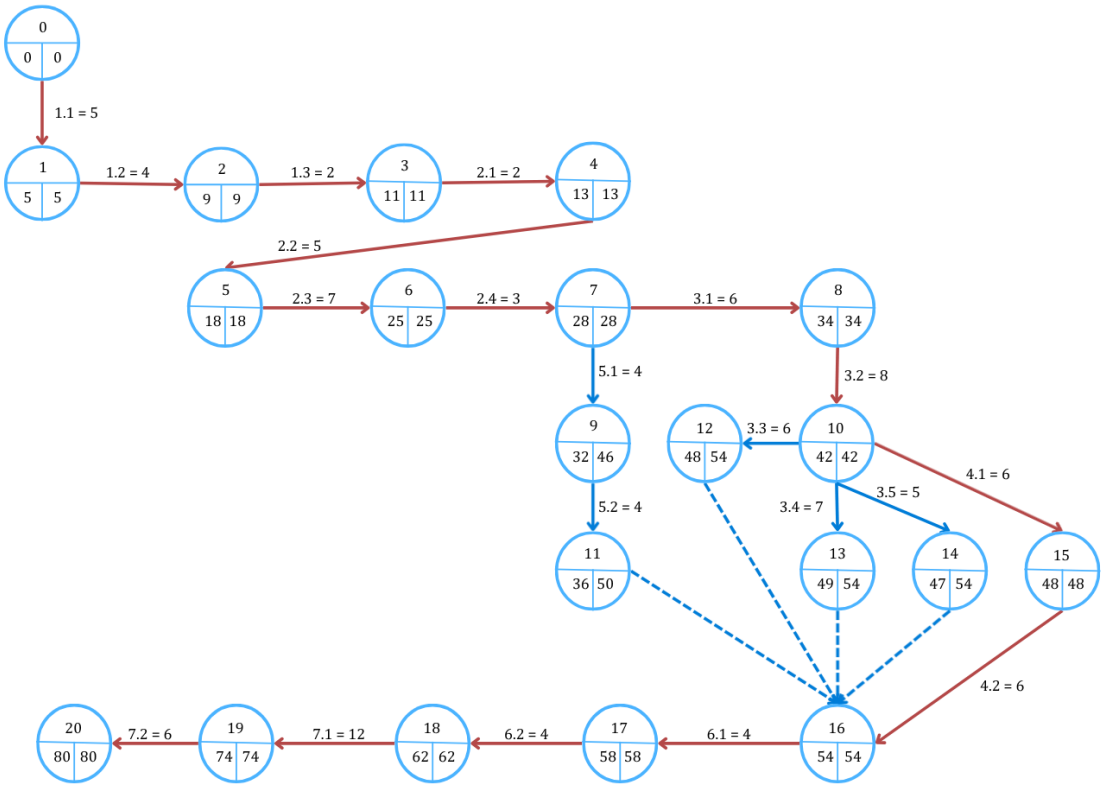

Figure 8.2. PERT Diagram

The critical path: 1.1 → 1.2 → 1.3 → 2.1 → 2.2 → 2.3 → 2.4 → 3.1 → 3.2 → 4.1 → 4.2 → 6.1 → 6.2 → 7.1 → 7.2, runs for 80 working days. Delays on any of these tasks delay the end date by the same amount.

8.3. GANTT Chart

The Gantt chart in Figure 8.3 places each task on the calendar using early-start scheduling. Red bars indicate critical tasks, and blue bars indicate non-critical ones. Tasks 3.3, 3.4, 3.5, 5.1, and 5.2 have between 5 and 18 days of float. The most parallel point in the schedule is day 42, when 3.3, 3.4, 3.5, and 4.1 all start together after Patch-based 2D training (3.2) completes.

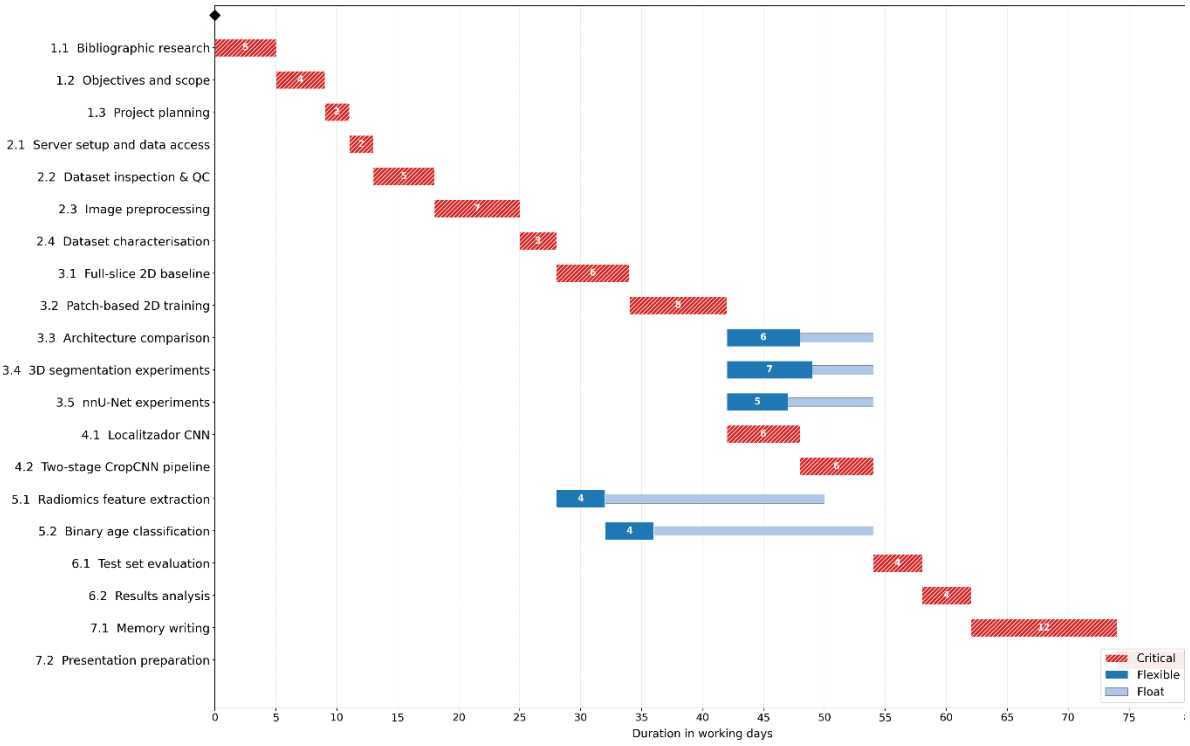

Figure 8.3. GANTT Chart

9. TECHNICAL FEASIBILITY

To assess the project's technical feasibility, a SWOT analysis was conducted to identify the main internal and external factors likely to influence its development. Table 9.1 summarises the four dimensions.

Table 9.1. SWOT analysis of the project.

Strengths	Weaknesses
- Prior experience in Python programming. - GPU access on the group's shared workstation. - Direct collaboration with radiologists and biomedical engineers at Hospital Clínic de Barcelona. - Availability of real patient MRI images.	- Limited execution time. - Limited experience in medical imaging and deep learning. - Small, single-institution dataset. - Small and subtle postmenopausal ovaries. - Shared GPU infrastructure restricted parallel training runs.
Opportunities	**Threats**
- Growing clinical and commercial interest in AI-based radiology tools. - Potential for future multicentre expansion once a baseline model is established. - Expansion of the AI medical applications market.	- Rapid publication pace in medical image segmentation may reduce novelty. - Clinical adoption may meet resistance from radiologists sceptical of AI tools - Need to meet ethical and legal requirements.

9.1. Strengths

Prior experience in Python reduced the setup overhead considerably. GPU access on the group's shared workstation allowed running the training experiments without relying on external cloud resources. Moreover, the collaboration with radiologists and biomedical engineers at Hospital Clínic de Barcelona was an important practical asset, as clinical knowledge was available at critical points. Finally, working directly with a real clinical dataset grounded the work in the actual diagnostic context the pipeline is meant to support.

9.2. Weaknesses

The most binding constraint was time. The TFG submission deadline in June 2026 meant that experimental iterations had to be prioritised carefully, and several directions worth exploring were left out. On the other hand, the dataset size was a related problem: 103 patients from a single institution is a reasonable starting point, but it limits what can be concluded from the work. It also makes it harder to avoid overfitting. The target structure added its own difficulty as postmenopausal ovaries are small.

Finally, the preprocessing steps specific to clinical MRI (handling DICOM metadata inconsistencies, axis permutation corrections, mask alignment checks) required more time than anticipated early in the project.

9.3. Opportunities

Ovarian cancer is diagnosed at an advanced stage in most cases, and early detection tools in routine clinical practice remain limited. Automated segmentation of the postmenopausal ovary is a prerequisite for any downstream mass characterisation analysis, which means this work addresses a real gap in the pipeline Hospital Clínic is building. The radiomics study carried out here, though exploratory, points toward non-invasive tissue characterisation as a concrete downstream application once segmentation

quality improves. On a broader scale, the market for AI-assisted radiology is growing, and a method developed in close collaboration with a clinical radiology department and validated on real patient data would be well placed for future development or integration into a wider diagnostic workflow.

9.4. Threats

The medical image segmentation literature moves fast, and related segmentation methods are published frequently. By the time this system could be extended to clinical use, other groups may have produced approaches with comparable or better performance on similar anatomical targets. Regulatory compliance would also be a significant requirement for any clinical deployment. Beyond regulation, acceptance by clinical staff is not guaranteed. Radiologists are not uniformly receptive to AI tools, particularly when those tools produce outputs they are expected to review and act on, and any future implementation would need to account for that.

10. ECONOMIC FEASIBILITY

An economic analysis was conducted to assess the project's viability. Table 10.1 shows the estimated costs by category.

Table 10.1. Economic analysis of the project.

Category	Description	Cost per unit	Total cost
Data acquisition			
MRI scans	T2-weighted pelvis MRI (103 patients)	200€ / scan	20,600 €
Manual segmentation	Ovary segmentation by a radiologist	52h x 25€ / h	1,300 €
Materials			
Software	Python, 3D Slicer, PyRadiomis	Open source (free)	0 €
MINT Lesion Software License	Segmentation and visualisation tool used by the radiologist	13,000€ + 7,000€ / anual revision	20,000 €
Computing workstation	GPU workstation (UB)	9,000€	9,000 €
Personal computer	Laptop used for the project	1,000€	1,000 €
Human resources			
Student	Biomedical engineering student	300h x 15€ / h	4,500 €
Supervisors	UB supervisors	25h x 2 x 25€ / h	1,250 €
Radiologist	Clinical expert consultations	5 x 25€ / h	125 €
Total cost			57,775 €

Most of the figures are approximate, as the resources were not acquired specifically for this project. The MRI scans were originally obtained for diagnostic purposes at Hospital Clínic de Barcelona, the GPU workstation was already available at the Biomedical Imaging Lab of the Universitat de Barcelona, and the personal computer was pre-owned. The manual segmentation cost is based on an estimate of roughly 30 minutes per case across the 103-patient dataset. For human resources, the student hours are calculated at the standard rate for a junior research role, and the supervisors' contribution covers regular meetings and technical guidance. Other tools used during the project, such as VPN access and SSH infrastructure, were provided by the Universitat de Barcelona at no additional cost. The total estimated cost of the project is approximately 57,775 €.

11. REGULATIONS AND LEGAL ASPECTS

The central regulatory constraint is that the project uses real patient data. At the European level, all personal data processing must comply with the General Data Protection Regulation (GDPR) [30]. Health data falls under a special category of data, which requires an explicit legal basis for processing.

At the Spanish level, the Organic Law 3/2018 on the Protection of Personal Data and Guarantee of Digital Rights (LOPDGDD) also applies [31], adding national-level obligations for sensitive data in a research context.

Since the dataset comprises MRI images from patients at Hospital Clínic de Barcelona, the research protocol was reviewed and approved by the hospital's Clinical Research Ethics Committee (CEIC) [32], which assessed the study methodology and data handling procedures. All images were anonymised before use: identifying DICOM metadata (patient name, date of birth, and study date) was removed before any data left the hospital's infrastructure. The European Cybersecurity Agency (ENISA) recommends pseudonymisation as the appropriate security technique for health research data [33], and the anonymisation pipeline applied here follows that guidance.

All software tools used in this project (Python, PyTorch, 3D Slicer, PyRadiomics, and nnU-Net) are distributed under open-source licences that permit academic and research use without restriction. No proprietary software requiring a specific licence was used.

This project is research-focused and does not include commercialisation plans. Even so, the applicable product regulations are worth keeping in mind. Under European Regulation 2017/745 (MDR) [34], software with a medical purpose is classified as a medical device, and any future clinical deployment of the segmentation pipeline would require CE marking and compliance with the corresponding quality management standards. Obligations under intellectual property law and research-related contractual agreements would also apply at that stage.

12. CONCLUSIONS

This project aimed to develop and evaluate a deep learning pipeline for the automatic segmentation of postmenopausal ovaries in T2-weighted MRI, using a clinical dataset of 103 patients from Hospital Clínic de Barcelona; radiomic feature extraction and binary age classification were included as a downstream application to assess the clinical potential of the resulting masks.

Class imbalance, not architecture choice, proved to be the central obstacle in segmentation. Full-resolution training consistently collapsed to empty-mask predictions; centering 96×96 crops on the ovary centroid resolved the problem directly, raising validation Dice from 0.161 to 0.619. An AttentionU-Net trained under this scheme achieved a test Dice of 0.447, above the 0.290 reported by RAovSeg, the only published system for normal-ovary segmentation in T2-weighted MRI.

Automatic localisation was not achieved. Both the bounding box regressor and the two-stage CropCNN produced centroid estimates too imprecise for reliable inference; the test Dice of 0.447 was measured with ground-truth centroids, and end-to-end performance is lower. Localisation remains the unsolved part of the pipeline.

The radiomics classifier reached 61.9% accuracy on binary age classification against a 50.5% majority-class baseline, with shape features ranking highest, consistent with the volume and sphericity changes that follow menopause. The result is above chance but not clinically useful, and it was obtained on ground-truth masks rather than predicted ones.

There was no prior comparable result for postmenopausal ovary segmentation in MRI; this work provides one. The gap between the current pipeline and a deployable system comes down to localisation: a detection network trained on full pelvic volumes is the most direct path forward, and if that step is solved, the segmentation and radiomics components are already in place.

13. BIBLIOGRAPHY

[1] Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA A Cancer J Clinicians. 2021;71(3):209–249.

[2] Lheureux S, Braunstein M, Oza AM. Epithelial ovarian cancer: Evolution of management in the era of precision medicine. CA A Cancer J Clinicians. 2019;69(4):280–304.

[3] Sadowski EA, Thomassin-Naggara I, Rockall A, et al. O-RADS MRI Risk Stratification System: Guide for Assessing Adnexal Lesions from the ACR O-RADS Committee. Radiology. 2022;303(1):35–47.

[4] Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer. 2012;48(4):441–446.

[5] Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Medical Image Analysis. 2017;42:60–88.

[6] Kurjak A, Kupesic S, Simunic V. Ultrasonic assessment of the peri- and postmenopausal ovary. Maturitas. 2002;41(4):245–254.

[7] Mahesh M. The Essential Physics of Medical Imaging, Third Edition. Medical Physics. 2013;40(7):077301.

[8] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, et al., editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 [Internet]. Cham: Springer International Publishing; 2015 [cited 2026 May 7]. p. 234–241. Available from: http://link.springer.com/10.1007/978-3-319-24574-4_28.

[9] Chen Z, Zhang C, Li Z, et al. Automatic segmentation of ovarian follicles using deep neural network combined with edge information. Front Reprod Health. 2022;4:877216.

[10] Wang Y, Wang M, Cao P, et al. CT-based deep learning segmentation of ovarian cancer and the stability of the extracted radiomics features. Quant Imaging Med Surg. 2023;13(8):5218–5229.

[11] Buddenkotte T, Rundo L, Woitek R, et al. Deep learning-based segmentation of multisite disease in ovarian cancer. Eur Radiol Exp. 2023;7(1):77.

[12] Hsu W-C, Wang Y, Wu Y-F, et al. MRI-based Ovarian Lesion Classification via a Foundation Segmentation Model and Multimodal Analysis: A Multicenter Study. Radiology. 2025;316(2):e243412.

[13] Liang X, Alpuing Radilla LA, Khalaj K, et al. A Multi-Modal Pelvic MRI Dataset for Deep Learning-Based Pelvic Organ Segmentation in Endometriosis. Sci Data. 2025;12(1):1292.

[14] Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health. 2021;3(3):e195–e203.

[15] Zanardo M, Visser JJ, Colarieti A, et al. Impact of AI on radiology: a EuroAIM/EuSoMII 2024 survey among members of the European Society of Radiology. Insights Imaging. 2024;15(1):240.

[16] Xu Y, Quan R, Xu W, et al. Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches. Bioengineering. 2024;11(10):1034.

[17] Mint Medical: developer of the leading radiological software solution - Mint Medical GmbH [Internet]. [cited 2026 May 7]. Available from: https://mint-medical.com/.

[18] syngo.via [Internet]. [cited 2026 May 7]. Available from: https://www.siemens-healthineers.com/es/digital-health-solutions/syngovia.

[19] AI-Rad Companion [Internet]. [cited 2026 May 7]. Available from: https://www.siemens-healthineers.com/es/digital-health-solutions/ai-rad-companion.

[20] MarketsandMarkets. Radiology AI Market worth $2.27 billion by 2030 with 24.5% CAGR | MarketsandMarkets™ [Internet]. [cited 2026 May 7]. Available from: https://www.prnewswire.com/news-releases/radiology-ai-market-worth-2-27-billion-by-2030-with-24-5-cagr--marketsandmarkets-302621509.html.

[21] Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328–338.

[22] Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–211.

[23] Oktay O, Schlemper J, Folgoc LL, et al. Attention U-Net: Learning Where to Look for the Pancreas [Internet]. arXiv; 2018 [cited 2026 May 7]. Available from: http://arxiv.org/abs/1804.03999.

[24] Hatamizadeh A, Nath V, Tang Y, et al. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images [Internet]. arXiv; 2022 [cited 2026 May 24]. Available from: https://arxiv.org/abs/2201.01266.

[25] Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D Medical Image Segmentation [Internet]. arXiv; 2021 [cited 2026 May 24]. Available from: https://arxiv.org/abs/2103.10504.

[26] Sundaresan V, Zamboni G, Rothwell PM, et al. Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images. Medical Image Analysis. 2021;73:102184.

[27] TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images | Radiology: Artificial Intelligence [Internet]. [cited 2026 May 27]. Available from: https://pubs.rsna.org/doi/10.1148/ryai.230024.

[28] Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research. 2017;77(21):e104–e107.

[29] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Machine Learning in Python.

[30] BOE.es - DOUE-L-2016-80807 Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta

al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos). [Internet]. [cited 2026 May 23]. Available from: https://www.boe.es/buscar/doc.php?id=DOUE-L-2016-80807.

[31] BOE-A-2018-16673 Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. [Internet]. [cited 2026 May 23]. Available from: https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673.

[32] Codi Ètic Institucional.

[33] Pseudonymisation techniques and best practices | ENISA [Internet]. [cited 2026 May 23]. Available from: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices.

[34] Regulation - 2017/745 - EN - Medical Device Regulation - EUR-Lex [Internet]. [cited 2026 May 23]. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745.