

ISEN

ALL IS DIGITAL!

OUEST

Institut Supérieur de l'Électronique et du
Numérique
ISEN Yncréa Ouest, campus de Nantes
33 avenue du Champ de Manœuvres
44470 Carquefou, FRANCE
Tél. : +33 (0)2 30 13 05 60



Stage d'application M1

Année scolaire 2025/2026

Automated analysis of capsule endoscopy videos: lesion detection and anatomical localisation

Effectué du 01/06/26 au 27/09/26 à Coimbra (Portugal)

Institute of Systems and Robotics, University of Coimbra
Rua Sílvio Lima, Pólo II, 3030-290 Coimbra, Portugal

Tuteur entreprise

ARAÚJO Hélder

Full Professor, DEEC,
helder@isr.uc.pt

Correspondant ISEN

DEPUYDT Pauline

ISEN Yncréa Ouest, Nantes

Étudiant

MÉNAHEZE Ewan

option : Intelligence Artificielle

Acknowledgements

The work presented in this report was carried out at the Institute of Systems and Robotics of the University of Coimbra, within the Department of Electrical and Computer Engineering.

Sincere thanks are addressed to Professor Hélder Araújo, supervisor of this internship, for hosting this project in his research group, for defining its subject and for the scientific freedom granted throughout its execution.

Thanks are also addressed to Marina Oliveira, doctoral researcher at the institute, whose published work on capsule endoscopy imaging was studied at the beginning of the internship. All data used in this project come from public datasets.

Coimbra, September 2026

Contents

Acknowledgements	1
Glossary	5
List of figures	6
1 Introduction	7
2 Presentation of the host organisation	8
2.1 The Institute of Systems and Robotics	8
2.2 Organisation and research areas	8
2.3 Position of the internship within the institute	8
3 Context	9
3.1 Clinical context: capsule endoscopy	9
3.2 Current reading practice	9
3.3 Available data	10
3.4 State of the art	11
3.5 Existing work in the group	11
3.6 Societal responsibility and sustainable development	12
3.6.1 Strategy of the host institution	12
3.6.2 Observations in the laboratory	12
3.6.3 Contribution of the internship	12
3.6.4 Critical analysis and recommendations	13
3.6.5 Position of the engineer	13
4 Specifications and technical objectives	14
4.1 Initial request	14
4.2 Analysis of the initial objectives	14
4.3 Redefined objectives	15
4.4 Constraints	15
4.5 Measurable requirements	15
5 Technical development	17
5.1 First approaches to localisation	17
5.2 Problem modelling and architecture	17
5.2.1 Formalisation	17
5.2.2 Architecture	18
5.3 Evaluation protocol	18

5.3.1 What is measured	18
5.3.2 How the data are split	19
5.3.3 Replay and consistency	20
5.4 First pipeline and its tuning	20
5.4.1 Components	20
5.4.2 Tuning with the replay tool	20
5.5 Overlap between public datasets: an error found and corrected	21
5.6 A clean strategy: pilot study and phase 1	22
5.6.1 Pilot study	22
5.6.2 Phase 1: three pillars at the scale of Kvasir-Capsule	22
5.7 Phase 2: cross-validated fine-tuning	23
5.7.1 Training design	23
5.7.2 Results	24
5.8 Anatomical localisation	25
5.9 Bleeding and external validation	27
5.9.1 Diagnosis	27
5.9.2 Phase 3: external bleeding data, audited	28
5.9.3 External test	29
5.9.4 Effect on the Kvasir videos	29
5.10 Integration and end-to-end verification	30
5.10.1 Integration	30
5.10.2 Verification on the 16 videos	30
5.10.3 Final results	32
5.10.4 Limitations	32
6 Project management	33
6.1 Method: decisions driven by measurement	33
6.2 Timeline	33
6.3 Evolution of the technical choices and impact on the project	33
6.4 Decision log	35
6.5 Risk management	35
6.6 Tools and resources	36
6.6.1 Software stack	36
6.6.2 Models tested	37
6.6.3 Code and hardware	37
7 Conclusion	38

Bibliography	40
Appendices	42
Appendix A: Example of event produced by the pipeline	42
Appendix B: Running the pipeline	42
Appendix C: Results per video	42
Appendix D: Organisation of the code	44
Abstract	45
Résumé	45

Glossary

Table 1: Acronyms and technical terms used in this report

Term	Definition
AI	Artificial Intelligence.
AUROC	Area Under the Receiver Operating Characteristic curve: probability that a randomly chosen positive image receives a higher score than a randomly chosen negative one (0.5 = chance, 1 = perfect).
CNN	Convolutional Neural Network.
CPU / GPU	Central / Graphics Processing Unit.
DDRS / RSE	Développement Durable et Responsabilité Sociétale / Responsabilité Sociétale des Entreprises (sustainable development and corporate social responsibility).
DEEC	Department of Electrical and Computer Engineering, University of Coimbra.
dHash	Difference hash: 64-bit perceptual fingerprint of an image, robust to resizing and compression.
Episode	Continuous annotated occurrence of one lesion in a video (a gap longer than 1 s starts a new episode).
False place	Zone of the video flagged by the system while containing no annotated lesion.
Fold	Subset of patients held out for testing in cross-validation.
FPR	False Positive Rate.
Grad-CAM	Gradient-weighted Class Activation Mapping: visual explanation locating the image region that drove a decision.
Homography	Projective transformation between two images of the same plane, used here to estimate the motion between two frames.
ISR	Institute of Systems and Robotics, University of Coimbra.
RAFT	Recurrent All-Pairs Field Transforms: neural network estimating the optical flow between two images.
RANSAC	Random Sample Consensus: robust estimation method discarding outlier correspondences.
JSON	JavaScript Object Notation: structured text format used for the event reports.
k-NN	k-Nearest Neighbours.
OOD	Out-Of-Distribution: input unlike the training data.
Pylorus	Valve between stomach and small bowel: first anatomical landmark.
Ileocecal valve	Valve between small bowel and colon: second anatomical landmark.
VCE / WCE	Video / Wireless Capsule Endoscopy.
Zone	Group of contiguous alerts (gap of at most 2 s): one place for the physician to check.

List of figures

Figure 1: Kvasir-Capsule labelled frames per class (left) and number of patients in which each class appears (right)	10
Figure 2: Architecture of the video analysis pipeline	18
Figure 3: Cross-validation by patient used for all results from Section 5.6 onwards	19
Figure 4: Left: images of each public dataset found to duplicate the data of this work (logarithmic scale; red = duplicates, removed). Right: held-out lesions found with the correct label, before and after removing the duplicates	21
Figure 5: Forty "false places" drawn at random: gastric folds, bubbles and debris, and unannotated red spots	23
Figure 6: Lesions found versus false places per minute on complete videos: thresholds chosen on tuning videos (left), then applied unchanged to held-out videos (right)	24
Figure 7: Section probabilities over one video judged by a model that never saw it, with estimated (red) and annotated (dashed) crossings	25
Figure 8: Error on the crossing times, same 8 videos: bar = median, dot = one video (logarithmic scale); dashed line = published mean error	25
Figure 9: Position of three lesions found by the pipeline, in percent of small-bowel transit, on an 18-minute video	26
Figure 10: Annotated fresh-blood frames of the held-out video: lesion probability (top) and class predicted by the model that never saw the video (bottom)	28
Figure 11: External test on WCEbleedGen: AUROC of each fold model and of the final model (1 = perfect, 0.5 = chance); red = bleeding output, grey = generic lesion score	29
Figure 12: Examples of correct detections of the final pipeline, each on a video never seen by its model: predicted category and confidence (top), Grad-CAM box, annotated class and position given by the pipeline (bottom)	31
Figure 13: Events raised by the pipeline on a held-out video, coloured by clinical category, against the annotated lesion episodes: 63 events over 29 minutes for 29 annotated episodes, of which 25 are found	31
Figure 14: Timeline of the internship	33

1 Introduction

Wireless capsule endoscopy (WCE) is the reference examination of the small bowel. The patient swallows a capsule the size of a large pill that films the digestive tract for eight to twelve hours while it is carried by peristalsis. The recording, typically between 50,000 and 100,000 useful frames, is then reviewed by a gastroenterologist. Reading takes between thirty minutes and two hours per examination, and lesions visible on a handful of frames can be missed through fatigue. Assisting this reading is therefore a well identified need, and one on which public data now exist.

The internship subject, defined with the Institute of Systems and Robotics (ISR) of the University of Coimbra, concerns the localisation of the capsule within the gastrointestinal tract, with the purpose of locating and characterising pathologies. Its final objective was refined during the internship into a concrete deliverable: a pipeline that takes a complete capsule video as input and outputs, for each suspected anomaly, its timecode, an annotated screenshot, its type, and its anatomical location: stomach, small bowel or colon, and position within the small bowel expressed as a percentage of transit time. Distances in centimetres are deliberately excluded, since capsule progression is not uniform and no public data allow them to be measured.

The main difficulty of the subject is not building such a pipeline, but **proving how well it works**. Public capsule datasets are small in number of patients, several of them re-distribute each other's images, and a model evaluated on images of a patient it was trained on appears far better than it is. A significant part of the work reported here is therefore the design of an evaluation in which every video is judged by a model that has never seen it.

This report is organised as follows. Chapter 2 presents the host institute. Chapter 3 describes the clinical context, the current reading practice, the available data and the state of the art, together with the societal and environmental analysis requested by the school. Chapter 4 states the initial objectives, the constraints, and the technical specifications as redefined during the internship. Chapter 5, the core of the report, details the technical development: system architecture, evaluation protocol, the correction of a data overlap between public datasets, the three successive versions of the lesion detector, the anatomical localisation, the bleeding problem and its external validation, and the end-to-end verification of the final system. Chapter 6 analyses project management and the evolution of technical choices. Chapter 7 concludes and gives the perspectives.

2 Presentation of the host organisation

2.1 The Institute of Systems and Robotics

The Institute of Systems and Robotics of the University of Coimbra (ISR-UC) is a private, non-profit research institution founded in 1992, hosted by the Department of Electrical and Computer Engineering (DEEC) of the Faculty of Sciences and Technology, on the Pólo II campus of the University of Coimbra, Portugal [1]. Its mission is to carry out multidisciplinary research of international level in systems and robotics, and to transfer its results to industry and society.

The institute is integrated into the Associate Laboratory ARISE (Advanced Production and Intelligent Systems) and obtained the rating "Excellent" at the 2018 evaluation of research units by the Portuguese Foundation for Science and Technology (FCT) [2].

2.2 Organisation and research areas

The research of ISR-UC is organised into seven activity areas with strong scientific synergies [2]:

- Human-Centred Mobile Robotics;
- Field Robotics;
- **Medical Systems and Robotics;**
- Computational Intelligence and Control;
- Artificial Perception and Cognition;
- **Visual Perception;**
- Intelligent Energy Systems.

The institute collaborates with the Faculty of Medicine of the University of Coimbra, with health institutions and with biomedical research centres, which gives its medical activities direct access to clinical questions and data [2].

2.3 Position of the internship within the institute

The internship took place in the group of Professor Hélder Araújo, Full Professor at the DEEC, whose work covers computer vision and its medical applications, at the intersection of the *Visual Perception* and *Medical Systems and Robotics* areas. Capsule endoscopy, and in particular the localisation of the capsule from images and from electromagnetic signals, is one of the topics studied in this group, as the subject of this internship reflects.

The internship was conducted as an individual research and development project. All data come from public datasets. Computing relied on a laptop computer without a dedicated graphics card for all video processing, and on the free GPU quota of the Kaggle platform (NVIDIA Tesla T4) for model training.

3 Context

3.1 Clinical context: capsule endoscopy

The capsule used in the data of this project is the Olympus EndoCapsule EC-S10, the second most sold capsule system after Medtronic PillCam; together they account for most of the market. It acquires images of 336×336 pixels, exported at 30 frames per second in the videos available for this work.

A capsule examination crosses three regions in a fixed order: the **stomach**, the **small bowel** and the **colon**, separated by two landmarks, the **pylorus** and the **ileocecal valve** (Figure 7). The small bowel, several metres long and out of reach of conventional endoscopes, is the main target of the examination. The lesions of interest fall into a few clinical families:

- **vascular**: angiectasia (dilated vessels), fresh or digested blood;
- **inflammatory**: erythema, erosion, ulcer;
- **lymphangiectasia** (dilated lymphatic vessels);
- **protruding**: polyps;
- **foreign bodies** (for instance retained food or a previous capsule).

Knowing *where* a lesion is matters as much as knowing *what* it is: the choice between a push enteroscopy from the mouth or a retrograde enteroscopy from the colon depends on whether the lesion lies in the proximal or distal part of the small bowel.

3.2 Current reading practice

Without automatic assistance, the physician reads the whole recording on the vendor's workstation and marks suspicious frames by hand; more than an hour is usually needed for the 50,000 to 100,000 frames of an examination [3]. The reading software offers intermediate aids rather than a complete analysis:

- **fast-reading modes**, such as QuickView in the Medtronic software, which score every frame and display only the most significant ones: reading time drops sharply, but with a lesion miss rate reported between 6.5 % and 12 %, mostly isolated lesions such as small vascular lesions and polyps [3];
- **red-colour detectors**, such as the Suspected Blood Indicator, which flag frames containing red pixels: even in patients with active bleeding their sensitivity remains below 60 % [3];
- **deep-learning reading assistants**, available on recent systems or under clinical evaluation, which pre-select suspicious frames; they are proprietary, tied to one capsule brand, and their output is still reviewed by the physician.

In practice, therefore, the reading remains largely manual: the aids shorten it but cannot replace a complete review, and none of them states where in the small bowel a finding lies. This is the gap addressed by the present work: an open

pipeline that proposes a short list of places to check, each with its category and its position.

3.3 Available data

Three public datasets were used, and one was studied without being downloaded (Table 2).

Table 2: Datasets used in this work

Dataset	Content	Role in this work
Kvasir-Capsule [4]	117 videos (Olympus EndoCapsule, Norway); 47,238 labelled frames from 43 videos, 14 classes	Training and evaluation; 16 complete videos available locally
Capsule Vision 2024 [5]	about 54,000 labelled images, 10 classes, several capsule brands	First training set; later only as a source of bleeding and polyp images, after audit
WCEbleedGen [6]	2,618 images, bleeding / non-bleeding, several cameras	External test of bleeding detection
Galar [7]	80 complete examinations, 22 of them with the same camera	Studied for external validation; not downloaded (over 600 GB)

Kvasir-Capsule is the only public dataset combining complete videos, the same camera throughout, and labelled anatomical landmarks. Its main limitation appears clearly in Figure 1: lesions are concentrated in few patients. Fresh blood appears in only 2 of the 43 labelled videos, polyps in one. A class present in a single patient cannot be validated, whatever the model: this constraint shapes several decisions of this report.

Kvasir-Capsule: 47,238 labelled images from 43 videos (lesions in red)

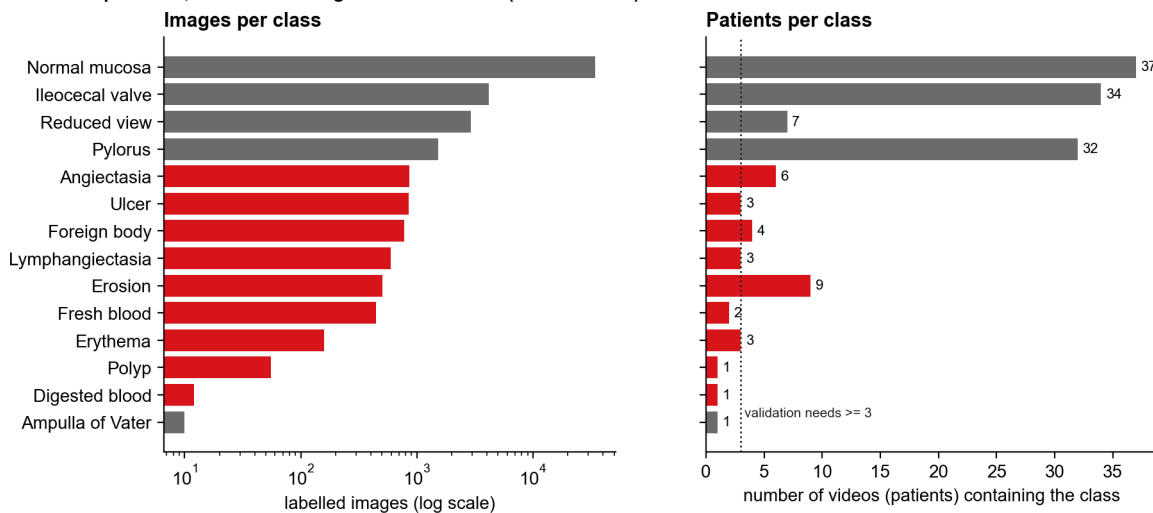


Figure 1: Kvasir-Capsule labelled frames per class (left) and number of patients in which each class appears (right)

3.4 State of the art

Two families of published results must be distinguished. Studies reporting accuracies above 95 % on capsule images generally split the data **by image**: frames of the same patient, often nearly identical consecutive frames, appear both in training and in test. With a split **by patient**, results collapse. On the official video-level split of Kvasir-Capsule, the original article reports a macro-averaged F1 score of 0.26 over the lesion classes [4]; a 2026 benchmark of eleven modern architectures obtains at best 0.35 [8]; and the best teams of the 2026 RARE-VISION challenge, built on the Galar dataset, reach a temporal mean average precision of about 0.35–0.37 [9] [10].

Three lessons were drawn from this review and guided the work:

- naming ten fine-grained lesion classes reliably from public single-camera data is not achieved by anyone today; clinical studies reaching 90 % sensitivity rely on thousands of patients and on **grouped clinical categories**;
- for anatomical localisation, the largest gains in the 2026 challenge came from **enforcing the anatomical order** of the regions rather than from the image classifier itself; a published small-bowel detector with temporal filtering reports mean errors of 38.8 s at the pylorus and 32.0 s at the ileocecal valve [11];
- generic foundation models (DINOv2 [12]) and endoscopy-specific ones (GastroNet-5M [13]) exist; their benefit had to be measured rather than assumed.

3.5 Existing work in the group

Capsule endoscopy imaging is an established line of work in the group, which was studied at the beginning of the internship.

Local panoramas from a multi-camera capsule. This work builds local forward-motion panoramas from the two cameras of the PillCam Colon2 capsule [14], after earlier work on the registration of consecutive capsule frames for 3D motion estimation [15]. Frames are projected onto a cylinder (pipe representation [16]), converted to polar coordinates around the focus of expansion, registered with the optical flow of the RAFT network [17] and a global homography estimated by singular value decomposition and RANSAC, then stitched. The method is validated on a synthetic colon-like tube rendered with Blender, which provides the true motion: registration is measured by the sum of squared differences, the intensity variance and the correlation, and panorama quality by no-reference image quality metrics.

Deep learning applied to capsule images. The literature review of a doctoral thesis of the group [18] covers binary and multi-class anomaly classification, and capsule localisation by organ classification, temporal filtering of the predictions and detection of the entrance and the exit of the small bowel.

Homography estimation. A reference essay on the subject [19] covers the estimation of a homography from the basic linear algorithm to statistical optimisation, together with estimation robust to outlier correspondences.

Two conclusions were drawn from this review. The motion between two frames can only be validated where the true motion is known, which in this line of work is a synthetic model; no public capsule dataset provides it. The anatomical landmarks, in contrast, are labelled in Kvasir-Capsule, so a localisation built on them can be measured against annotations.

3.6 Societal responsibility and sustainable development

3.6.1 Strategy of the host institution

The University of Coimbra, to which ISR is attached, has adopted a formal **Plan for Equality, Equity and Diversity (PIED@UC)**, developed with the European project SUPERA and aligned with the Sustainable Development Goals 5 (gender equality) and 10 (reduced inequalities) [20]. The plan covers work–life balance, the prevention of harassment, and programmes such as Equal.STEAM, which aims to attract and retain women in engineering and technology courses; the university was distinguished in 2025 as a pioneering institution in the promotion of gender equality [20]. ISR, as a research unit of the university, falls under this plan; no separate formal policy specific to the institute was identified during the internship.

3.6.2 Observations in the laboratory

The research teams observed during the internship were in majority male, but men and women held comparable positions and responsibilities, without visible difference in treatment. The laboratory is strongly international: its members come from many countries and backgrounds, and the working language is English, Portuguese being rarely used in the offices. This diversity made the integration of a foreign intern straightforward from the point of view of language.

Integration of newcomers, on the other hand, is little formalised. Arrival consisted of a visit of the laboratory with a member of the team; there is no welcome event or induction programme. Offices are organised by research domain, so that an intern shares an office with researchers working on neighbouring topics (computer vision in this case) but on different projects. Social contact relies mainly on the activities and events of the university campus, with which the laboratory is closely linked.

3.6.3 Contribution of the internship

The project follows the orientation of ISR towards research of public benefit in health. Its contributions to sustainable development are the following:

- **ethics of the data:** all models were trained and evaluated only on public, anonymised and documented datasets, whose licences allow their reuse for research; no patient data were collected or left the computers used;

- **health and access to care:** the system aims at reducing the reading time of the physician (about 26 places to check per video in median, instead of reviewing the whole recording) while leaving the final decision to the physician, and its limits are stated explicitly rather than hidden;
- **frugality:** the complete pipeline runs on a laptop processor without a graphics card (a 48-minute video is processed in about 21 minutes), and the final models required about 3 hours of shared cloud GPU time in total;
- **continuity:** the work builds on earlier research carried out at ISR and is documented so that it can be taken over by the next person working on the subject.

3.6.4 Critical analysis and recommendations

The main opportunity observed is the international and open environment of the laboratory. The main weakness is the absence of a structured integration of newcomers: a short introduction to the people sharing the same office and to their projects would have created exchanges between researchers who work side by side every day without collaborating. For the project itself, the priority for a responsible use is a validation on patients from other hospitals before any clinical use, since a detection tool that works on one population only can create inequalities of care.

3.6.5 Position of the engineer

This internship confirmed that computer vision can serve applications of direct social value, in medicine in particular. In such projects, the first responsibility of the engineer is to measure the system on patients it has never seen and to report its limits as clearly as its results. This rule will be kept in future projects.

4 Specifications and technical objectives

4.1 Initial request

The internship agreement describes the subject as the *localisation of the capsule within the gastrointestinal system*, with the activity: *track and localise the capsule throughout the gastrointestinal system to accurately localise and classify pathologies*. The expected low-level results were: (1) estimate the capsule trajectory from 2D images; (2) refine and correct this trajectory using the electromagnetic radiation of the capsule.

No written specification was provided beyond this description. The specification below was therefore established during the internship, from the agreement, from the review of available data, and from the needs expressed around the analysis of capsule videos.

4.2 Analysis of the initial objectives

Objective (2) requires electromagnetic signals recorded by an external antenna array during the examination. **No public dataset provides such signals together with videos**, and none were available within the group for this internship. Objective (1), a metric trajectory from images alone, is ill-posed for a monocular capsule moving in a deformable tube without any ground truth; no public dataset provides a reference trajectory either.

The part of the subject that could be both realised and **measured** with the available data was retained: localising the capsule *anatomically* (region and position within the small bowel, as a fraction of transit time), and detecting and characterising the pathologies along the video. This redefinition keeps the purpose stated in the agreement (localising and classifying pathologies) while replacing a metric localisation, which could not be validated, by an anatomical one, which can.

Expected precision. The precision that a localisation from images alone can reach was estimated before any development. A position in centimetres is out of reach: the capsule is moved by peristalsis, it advances, stops and sometimes moves backwards; with a single camera the scale of a displacement is unknown; and no public video provides the true motion that would be needed to calibrate or validate it. The useful precision is anatomical: the region of a lesion, and its position within the small bowel as a fraction of transit time. The published reference places the entrance and the exit of the small bowel with mean errors of 38.8 s and 32.0 s [11]; since the small-bowel transit lasts 966 s in median in the annotated Kvasir-Capsule videos, an error of 30 s at a crossing represents about 3 % of the transit. The targets of Table 3 follow from this analysis.

4.3 Redefined objectives

The deliverable is a pipeline processing a complete capsule video and producing, for each suspected anomaly, one event containing:

- the **timecode** and frame number;
- an **annotated screenshot** showing the region that triggered the alert;
- the **type** of the anomaly: a clinical category, with the most likely fine class given as a suggestion;
- the **anatomical location**: region (stomach, small bowel, colon), and within the small bowel, the third (proximal, middle, distal) and the percentage of transit time.

The events are written to a JSON file, grouped into **zones** (places to check), and summarised on a timeline.

4.4 Constraints

- **Single camera**: the Olympus EndoCapsule EC-S10 only. Generalisation to other capsules is explicitly out of scope.
- **No metric distance**: anatomical positions are expressed as fractions of transit time, never in centimetres.
- **Computing resources**: video processing on a laptop CPU; training on a free cloud GPU quota (about 30 hours per week, sessions of at most 12 hours).
- **Evaluation on unseen patients**: every figure must come from videos of patients never used to train or tune the measured model.

4.5 Measurable requirements

Table 3: Requirements and measurable acceptance criteria

Function	Indicator	Target
Detect lesions	Annotated lesion episodes found within ± 5 s	≥ 75 % (stop criterion set before phase 1)
Limit the reading burden	False places to check per minute of video	about 1 per minute (tunable)
Name lesions	Found episodes given the right clinical category	as high as achievable; measured per category
Locate the landmarks	Median error on pylorus and ileocecal valve crossings	better than the initial method and than published values (38.8 s / 32.0 s)
Locate within the small bowel	Median error on the position of lesions (percentage points)	a few points
Reproducibility	Pipeline decisions identical to the measured ones	exact identity, verified on every video

The 75 % detection threshold was fixed as a *stop criterion* before the first large-scale measurement: had the detection fallen below it, the approach would have been abandoned and analysed rather than tuned further.

5 Technical development

5.1 First approaches to localisation

Two ways of placing the capsule were explored before the video system was specified.

Progression by homography. Following the panorama work presented in Section 3.5, the forward progression of the capsule was estimated from the displacement of the intestinal walls between consecutive frames, through a homography estimated on matched points. The approach gives the relative motion between two frames, but it cannot produce a validated position on real videos: the scale is unknown with a single camera, the motion mixes translation and rotation, frames with bubbles, debris or little texture break the matches, and above all no public capsule video provides the true motion against which the result could be checked. The approach was therefore left as a perspective, to be combined with electromagnetic recordings when such data exist.

Anatomical landmarks. The pylorus and the ileocecal valve are labelled in Kvasir-Capsule, so a localisation built on them can be compared with annotations. Recognition models for the two landmarks were trained and retained for the pipeline; the video system was then specified around them.

5.2 Problem modelling and architecture

5.2.1 Formalisation

A capsule video is sampled at 2 frames per second, giving a sequence of N frames $x(1), \dots, x(N)$ with timecodes $t(1), \dots, t(N)$ (a 40-minute video yields about 4,800 frames). Three quantities are estimated for each frame $x(t)$:

- a **lesion probability** $p(x(t))$;
- a **distribution over clinical categories**, used only for flagged frames;
- a **distribution over anatomical sections** $P(s \mid x(t))$, where s is one of stomach, small bowel or colon.

The two problems are then solved at the level of the whole video, not frame by frame.

Detection. A frame is flagged when $p(x(t))$ exceeds a threshold. Flagged frames closer than 2 s form one **episode**, which becomes one event; its category is the vote of its frames, that is the category maximising the sum of their log-probabilities. Contiguous events form a **zone**, the unit a physician reviews.

Localisation. Since the capsule crosses the three sections in a fixed order, the two landmark crossings are the indices $i \leq j$ that best cut the video into three contiguous blocks:

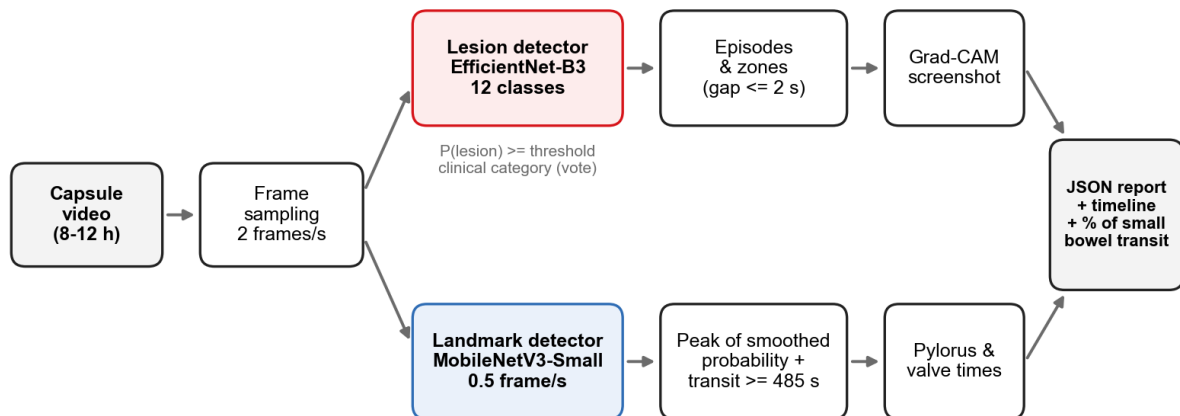
$$(i^*, j^*) = \underset{i \leq j}{\operatorname{argmax}} \sum_{t < i} \log P(\text{stomach} \mid x(t)) + \sum_{i \leq t < j} \log P(\text{small bowel} \mid x(t)) + \sum_{t \geq j} \log P(\text{colon} \mid x(t))$$

This optimum is computed exactly in linear time with cumulative sums and a running maximum; it was validated against exhaustive search on 300 random cases. The position of an event at time t within the small bowel is then $100 \cdot (t - t_{\text{pylorus}}) / (t_{\text{valve}} - t_{\text{pylorus}})$, in percent of transit time, and its third (proximal, middle, distal) follows.

5.2.2 Architecture

Figure 2 shows the resulting system. It has two branches sharing the same sampled frames: the lesion branch detects and names, the localisation branch situates the capsule. Their outputs are merged into one JSON report per video. In the delivered pipeline, the localisation branch uses the landmark detector described below; the monotone segmentation was measured on cached features and is not yet wired into it (Section 5.8).

Lesion branch: detect and name



Localisation branch: situate the capsule

Measured separately, next to integrate: section probe + monotone segmentation (anatomical order)

Figure 2: Architecture of the video analysis pipeline

The software is organised as independent Python modules: video sampling (OpenCV), lesion detector, episode grouping, zone grouping, anatomical localisation, Grad-CAM screenshots [21], report writing, and a command-line entry point. Every model decision on every sampled frame is saved alongside the report. This choice allows any decision rule to be **replayed offline** in seconds on cached model outputs, instead of re-running hours of inference for each experiment.

5.3 Evaluation protocol

5.3.1 What is measured

Ground truth comes from the Kvasir-Capsule annotations, which give the class of individual frames. Consecutive annotated frames of one class form a lesion **episode**; a gap longer than 1 s starts a new one. Three indicators are reported (Table 4).

Table 4: Evaluation indicators

Indicator	Definition	Why
Lesions found	Share of annotated episodes touched by an event within ± 5 s	A physician looks around an alert; ± 5 s is 10 frames
Lesions found and named	Same, with the event carrying the correct clinical category	Measures naming only where it matters
False places per minute	Zones containing no annotated lesion, divided by video duration	The reading burden; one zone = one place to check

A **chance level** is computed for every configuration by shifting the events circularly in time 20 times: a system that alerts often finds lesions by chance, and this level shows how much of a score is due to it.

5.3.2 How the data are split

A model must never be evaluated on a patient it has seen. Two successive splits were used.

Initial split. The 16 local videos were divided into 6 *tuning* videos, on which settings were chosen, 5 *test* videos used to confirm them, and 5 *held-out* videos, measured once at the end.

Cross-validation by patient. After the correction described in Section 5.5, a stricter protocol was adopted (Figure 3). The 21 Kvasir videos never seen by the feature extractor (the 16 local videos plus 5 validation videos) were divided into 4 folds, balanced by hand so that each lesion type tested in a fold remains represented in the training data of that fold. Four models are trained, each without the videos of its fold; every video is judged only by the model that has never seen it. The 22 other Kvasir videos are always used for training. Thresholds are chosen on the 11 tuning videos and applied **unchanged** to the 5 held-out videos.

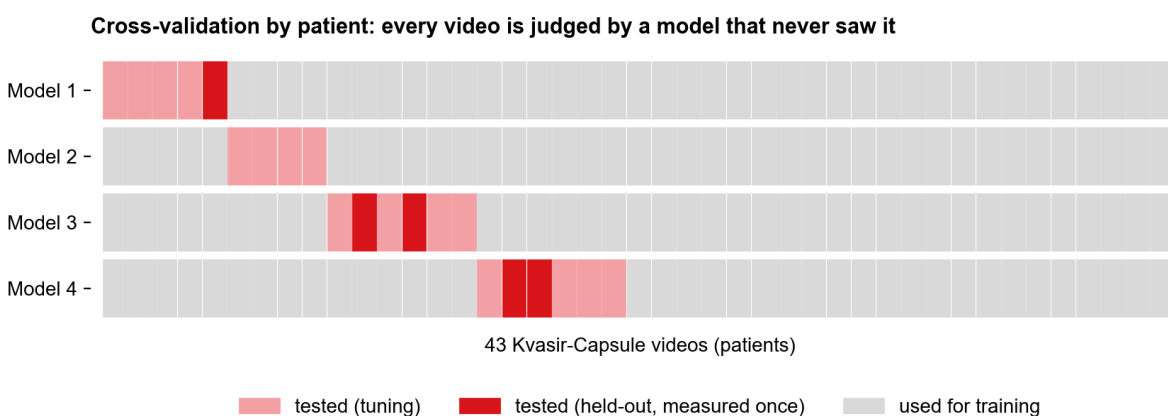


Figure 3: Cross-validation by patient used for all results from Section 5.6 onwards

5.3.3 Replay and consistency

Model outputs of the 16 videos are cached once (about 47,000 frames per model). All decision rules are then evaluated offline. A consistency check compares the replayed decisions with the decisions logged by the real pipeline: 27,790 frames out of 27,790 were identical on the first check, and the same check was repeated on every video for the final system (Section 5.10).

5.4 First pipeline and its tuning

5.4.1 Components

The first version of the pipeline was built on components developed during the first months of the internship:

- an **EfficientNet-B3** image classifier [22] trained on Capsule Vision 2024, 10 classes, with square-root class weighting (an earlier unweighted version over-detected massively, with a precision of about 0.39);
- an **energy score** on the classifier outputs to flag *unknown* anomalies without training a separate model [23]: a frame is "unknown" when its energy exceeds the 99th percentile of annotated non-lesion frames (threshold -2.862) and it is not confidently classified as normal;
- **Grad-CAM** to derive a bounding box and an annotated screenshot for each event;
- a **landmark detector** (MobileNetV3-Small [24], three classes: pylorus, ileocecal valve, other) applied at 0.5 frame per second, taking the peak of the smoothed probability, with an anatomical constraint: the small-bowel transit is never shorter than 485 s in the 24 annotated videos that contain both landmarks (median 966 s).

5.4.2 Tuning with the replay tool

Settings were chosen on the 6 tuning videos with a rule fixed in advance: a change is kept only if it removes **at least 15 % of false places without losing a single lesion**. Table 5 summarises the decisions.

Table 5: Tuning decisions on the first pipeline

Change	Measured effect	Decision
Triage gate: a small 3-class network (unusable / normal / suspect) discards alerts judged normal or unusable (threshold 0.90)	-20 % false places on tuning AND test videos, no lesion lost	Adopted
Margin: ignore frames where the lesion class barely beats "normal"	False places increased (2.75 $\rightarrow$ 2.93 per minute): events were split, not removed	Rejected
Temporal smoothing of probabilities	No gain at equal detection	Rejected
Bubble detector, Hough circle counting	Lesions lost	Rejected
Bubble detector, network trained on 300 hand-labelled frames	Even at threshold 0.95: 35 $\rightarrow$ 28 of 36 lesions found	Rejected

On the 5 held-out videos, this system found 70 of 79 lesion episodes with the correct label, a score far above the published state of the art; the next section explains why.

5.5 Overlap between public datasets: an error found and corrected

The 70/79 score of the first pipeline, together with a fine-tuning run reaching an F1 score of exactly 1.000, was inconsistent with published results on the same dataset (0.26–0.35 in macro-F1 with patient-level splits). The cause was an error in the choice of the training data: Capsule Vision 2024, used to train the first classifier, re-distributes Kvasir-Capsule images, including images of the evaluation videos, and this had not been checked before training.

The check. Since the copied images may have been resized, re-encoded or cropped, the check compares **perceptual fingerprints**: each image is reduced to 9×8 grey pixels and summarised by a 64-bit *difference hash*; each Kvasir image is also indexed with 8 cropped versions (0 to 14 %); candidate pairs differing by at most 3 bits are retrieved without comparing every image with every other, then confirmed on a 32×32 thumbnail. On 300 synthetic cases, the method recovers 92 to 98 % of copies up to 15 % cropping, with no false match.

What it found. 38,592 Capsule Vision images come from Kvasir-Capsule, **16,315 of them from the 16 evaluation videos**: the first results were partly memorisation. A clean classifier was retrained from generic ImageNet weights without these images. With the same decision rule, the held-out score fell from 70/79 to **11/79** lesions with the correct label (Figure 4); the system still located lesions (52 of 79 within ± 5 s with any label) but could no longer name them.

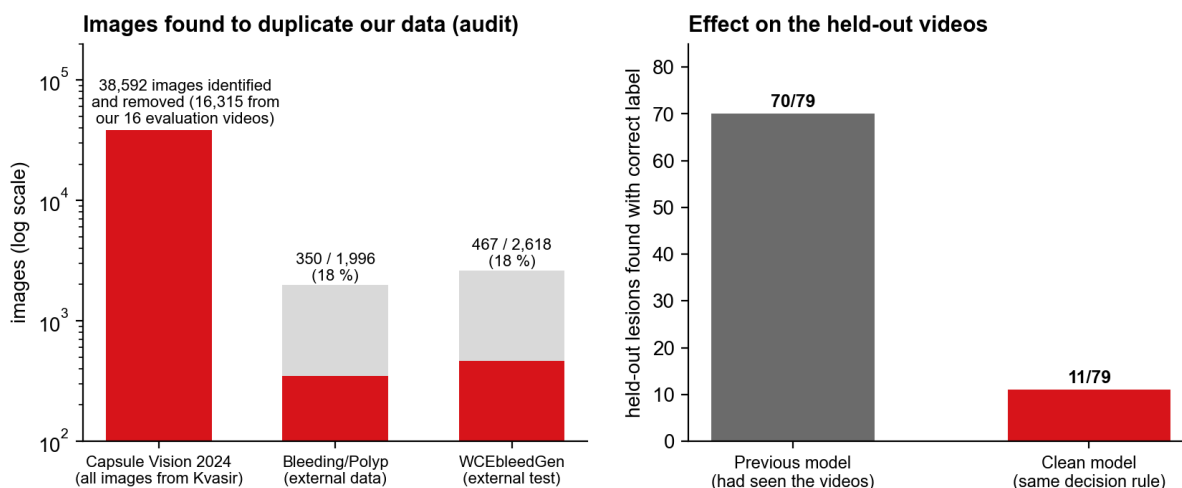


Figure 4: Left: images of each public dataset found to duplicate the data of this work (logarithmic scale; red = duplicates, removed). Right: held-out lesions found with the correct label, before and after removing the duplicates

Consequence. The absolute value of all previous measurements was invalidated, which led to the strategy review below and to the cross-validation by patient of

Section 5.3.2. The same check was applied before every later use of an external dataset (Section 5.9), where it removed 18 % of the external bleeding test set. The cost of this error for the project is analysed in Section 6.3.

5.6 A clean strategy: pilot study and phase 1

5.6.1 Pilot study

Before any new training, three hypotheses were measured on the 16 local videos, each video being judged by models built on the 15 others (Table 6).

Table 6: Pilot study: frozen features, leave-one-video-out on 16 videos

Question	Generic foundation model (DINOv2)	Clean classifier features
Anomaly as distance to normal mucosa of other patients (AUROC)	0.704	0.920
Lesion frames flagged when 10 % of normal frames alert	39.2 %	81.5 %
Lesion episodes given the right clinical category	29 %	60 %

Three conclusions followed. The generic foundation model was clearly inferior to the clean classifier for lesions. Measuring a **deviation from normal mucosa** requires no lesion example at all, and also covers lesions outside the known classes. A third test suggested that enforcing the anatomical order would place the landmark crossings more precisely than the initial method; it is measured in Section 5.8.

5.6.2 Phase 1: three pillars at the scale of Kvasir-Capsule

The strategy was organised in three pillars, **detect** (deviation from normal), **name** (clinical categories) and **situate** (monotone segmentation), and measured on all 43 labelled Kvasir videos. The features of the 47,238 labelled frames were extracted once on the cloud GPU; all analyses then ran locally.

On annotated frames, detection met the objective: 86 % of lesion episodes found with 1–2 % of normal frames alerting, on tuning and held-out videos alike. On **complete videos**, however, the picture changed (Figure 6): at about one false place per minute, phase 1 found 67 % of held-out lesions at 1.30 false places per minute, against 66 % at 1.14 for the initial rule, that is no real gain at equal reading burden.

The false places were inspected visually (Figure 5). They fell into three families: normal gastric mucosa (the Kvasir bank of normal frames is mostly small bowel), bubbles, bile and debris, and small red spots **without annotation**. In one video, only 2.4 of 42 minutes carry annotations: part of the "false" alarms are probably real findings, and the false-alarm figures of this report are therefore pessimistic.

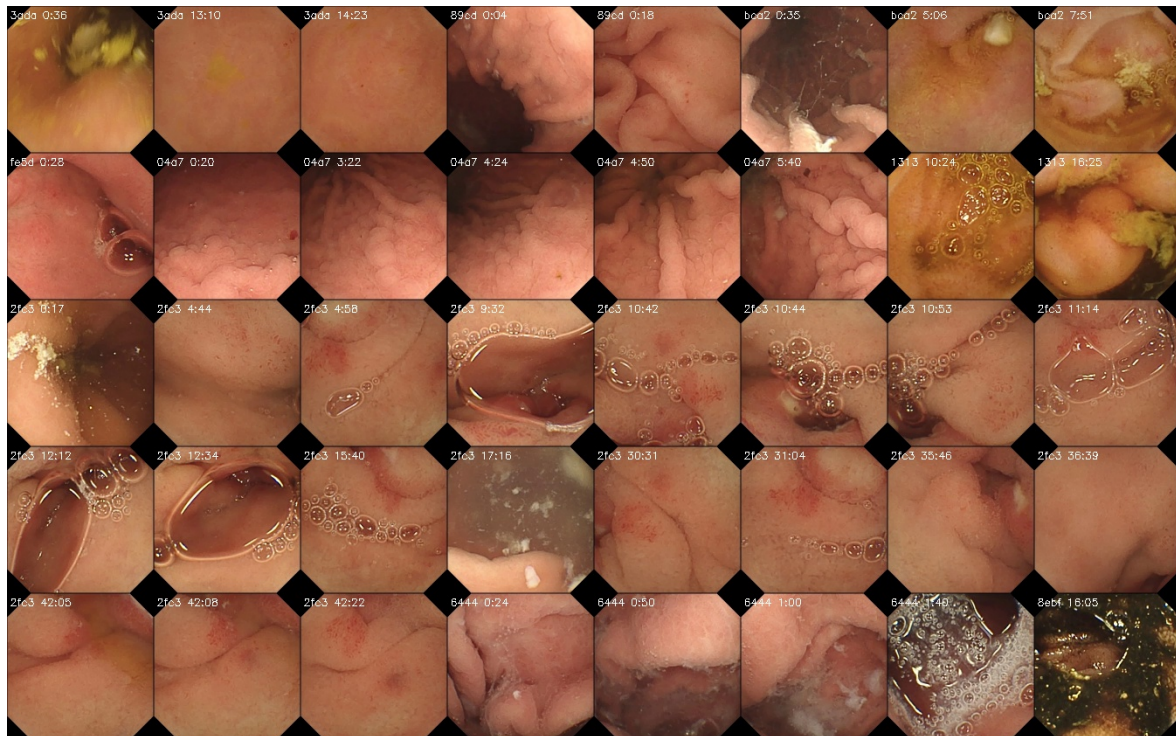


Figure 5: Forty "false places" drawn at random: gastric folds, bubbles and debris, and unannotated red spots

Enriching the bank of normal frames did not help: adding the other non-lesion Kvasir classes changed nothing significant, and adding frames of the local videos away from annotated lesions made detection **collapse** to 45 % at one false place per minute, as unannotated lesions polluted the "normal" bank. Naming reached a plateau of 56 % on tuning videos. The diagnosis was clear: the feature extractor had never seen a single Kvasir lesion, and it was the limiting factor.

5.7 Phase 2: cross-validated fine-tuning

5.7.1 Training design

The clean classifier was fine-tuned on Kvasir-Capsule images, four times, once per fold of Figure 3. The design choices, all fixed **before** any measurement, were:

- **12 classes:** the 8 lesion classes (fresh and digested blood merged, 12 images of the latter only), normal mucosa, reduced mucosal view (bubbles, debris), pylorus, ileocecal valve; the lesion probability is the sum of the 8 lesion classes;
- **balanced sampling:** each class receives a sampling mass (normal 3, reduced view 1.5, landmarks 0.5, each lesion 1), **split equally between the patients who have it**, so that a video rich in one lesion does not dominate that class;
- augmentations respecting capsule images (rotations by multiples of 90°, flips, mild colour changes);

- AdamW optimiser, one-cycle learning-rate schedule, 8 epochs of 16,000 sampled images, label smoothing 0.1, mixed precision; about 13 minutes per fold on a Tesla T4.

The notebook was first executed end to end on a synthetic 43-video dataset on CPU, with checks for leakage between folds, determinism, and consistency between saved features and scores, before consuming any GPU time.

5.7.2 Results

At about one false place per minute, the fine-tuned models found **73 % of held-out lesions at 1.16 false places per minute**, against 66 % at 1.14 for the initial rule and 67 % at 1.30 for phase 1 (Figure 6). At a wider operating point, 86 % were found at 2.06 per minute, against 82 % at 2.69 for phase 1: more lesions **and** fewer false places. Naming improved as well: 72 % of lesion episodes on tuning videos (56 % in phase 1), 82 % on held-out videos (77 %), with the largest gains on foreign bodies (12 → 21 of 27) and vascular lesions (8 → 15 of 24).

Figure 6 is read as follows: each point is one threshold; moving right, the system raises more false places per minute and finds more lesions, so a curve higher and further left is better. The shaded band marks the default setting, about one false place per minute.

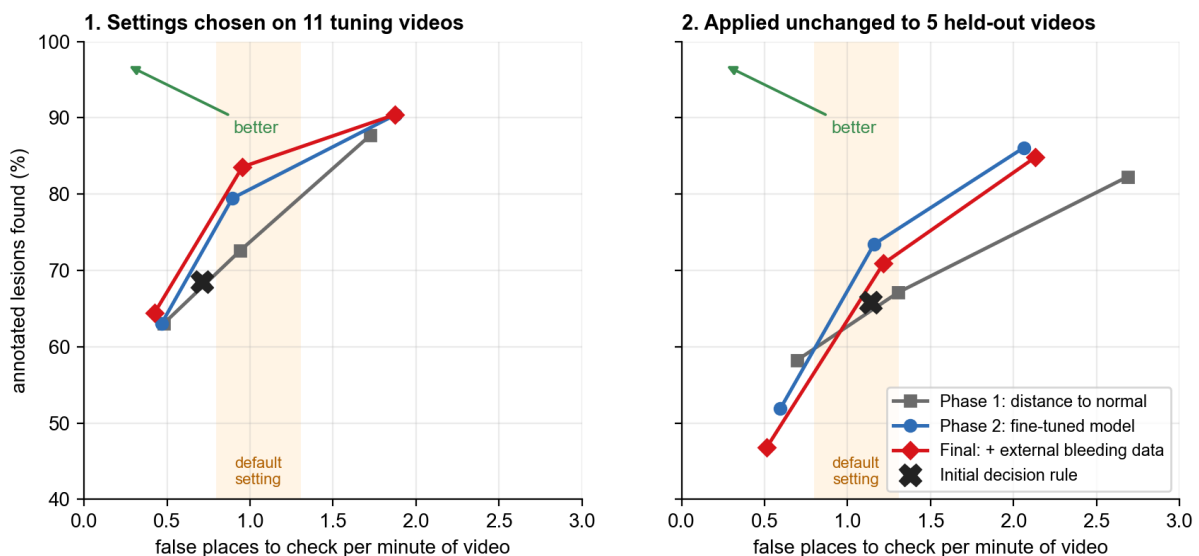


Figure 6: Lesions found versus false places per minute on complete videos: thresholds chosen on tuning videos (left), then applied unchanged to held-out videos (right)

One more lesson came out of this phase: the best detection score is now the **model probability itself**. The distance-to-normal score, central in phase 1, drifts between patients after fine-tuning: a threshold chosen on tuning videos produces too many alerts on held-out ones.

5.8 Anatomical localisation

The section classifier is a regularised logistic regression on the clean classifier features, trained on the dense sequences of the other local videos and on the labelled Kvasir frames whose section is known from the first annotated pylorus and valve frames. It is kept on the frozen features: the four fine-tuned models live in four different feature spaces, and a section model must be trained in the same space as the video it judges.

All crossing errors in this report are measured on the same **8 local videos**, the only ones where both methods can be compared. The median error falls from 13 s to **5 s at the pylorus** and from 55 s to **35 s at the ileocecal valve** (Figure 8); for reference, a published CNN with temporal filtering reports mean errors of 38.8 s and 32.0 s [11]. Figure 7 shows how the cut is obtained on one video: the three curves are the probabilities of each section, the red lines the estimated crossings and the dashed lines the annotated ones.

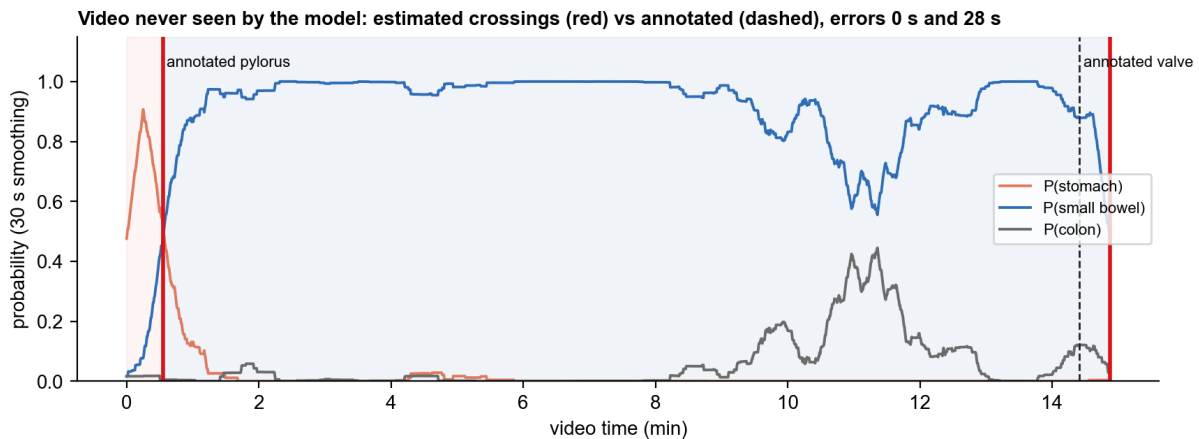


Figure 7: Section probabilities over one video judged by a model that never saw it, with estimated (red) and annotated (dashed) crossings

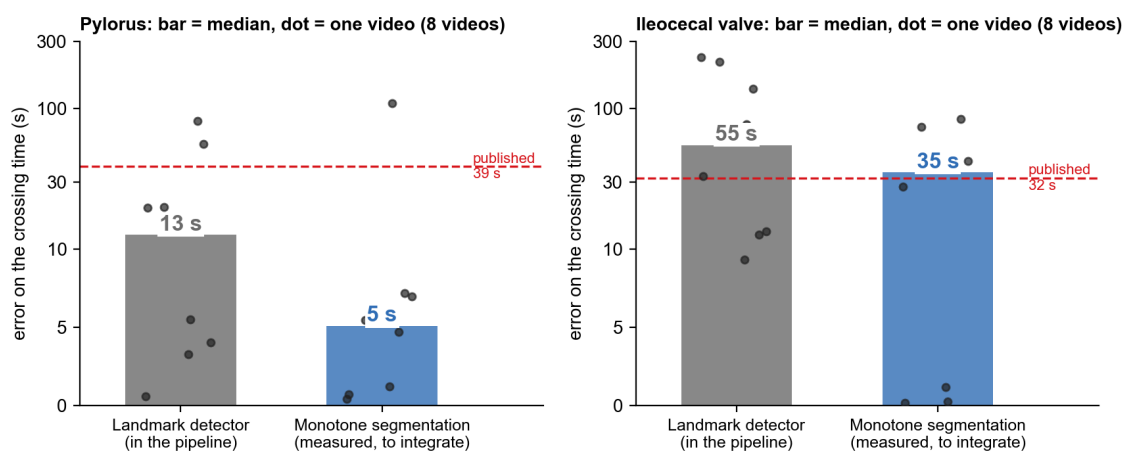


Figure 8: Error on the crossing times, same 8 videos: bar = median, dot = one video (logarithmic scale); dashed line = published mean error

For the clinical question of where the lesion lies within the small bowel, Figure 9 shows how a position is expressed: the time since the pylorus divided by the total small-bowel transit time. On the 8 videos, the median error on this position is 0.8 percentage points with the segmentation and 0.8 with the landmark detector (Table 7).

One failure was observed: in one video the pylorus is placed 108 s too early, and 18 foreign-body episodes lying in the stomach are then attributed to the small bowel.

Position in the small bowel = time since the pylorus / total small-bowel transit time

Crossings estimated by the pipeline on an 18-minute video; annotated crossings: 2.0 and 17.9 min

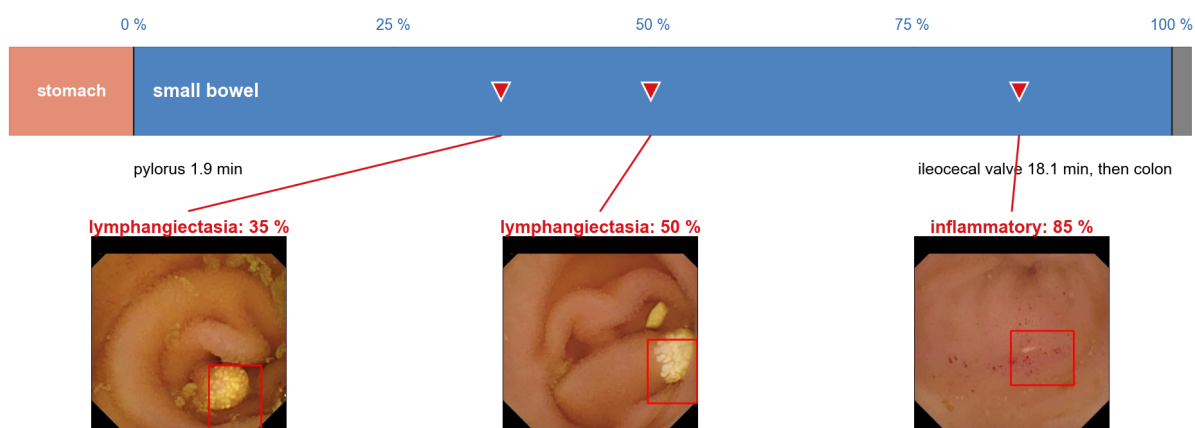


Figure 9: Position of three lesions found by the pipeline, in percent of small-bowel transit, on an 18-minute video

Status in the pipeline. The delivered pipeline still localises with the landmark detector. Table 7 compares the two methods on the same 8 videos: the segmentation places the crossings more precisely, but its single pylorus failure moves 18 lesions to the wrong region, and the landmark detector assigns more lesions to the correct region, with a similar position error within the small bowel. The landmark detector was, however, probably trained on frames of these videos, which favours it. The segmentation was therefore not substituted; combining the two (the segmentation for the cut, the landmark detector as a check on the pylorus) and validating the result with a landmark detector retrained without these videos is the first perspective of this work.

Table 7: Anatomical localisation: the two methods on the same 8 videos

Indicator	Landmark detector (used by the pipeline)	Monotone segmentation (measured)
Pylorus crossing, median error	13 s	5 s
Ileocecal valve crossing, median error	55 s	35 s
Lesions placed in the correct region	64 of 67	46 of 67
Position within the small bowel, median error	0.8 points	0.8 points

5.9 Bleeding and external validation

5.9.1 Diagnosis

One held-out video concentrated most of the missed lesions: 4 of its 14 episodes found. Its lesions are fresh blood. The cause is in the data: **424 of the 446 fresh-blood frames of Kvasir-Capsule belong to this very video**. When it is held out, its model has seen only 22 blood frames, and it answers "ileocecal valve" or "normal" on obvious blood (Figure 10). The median lesion probability is 0.13 on these frames, against 0.07 on normal mucosa: not a threshold problem, but an absence of learning.

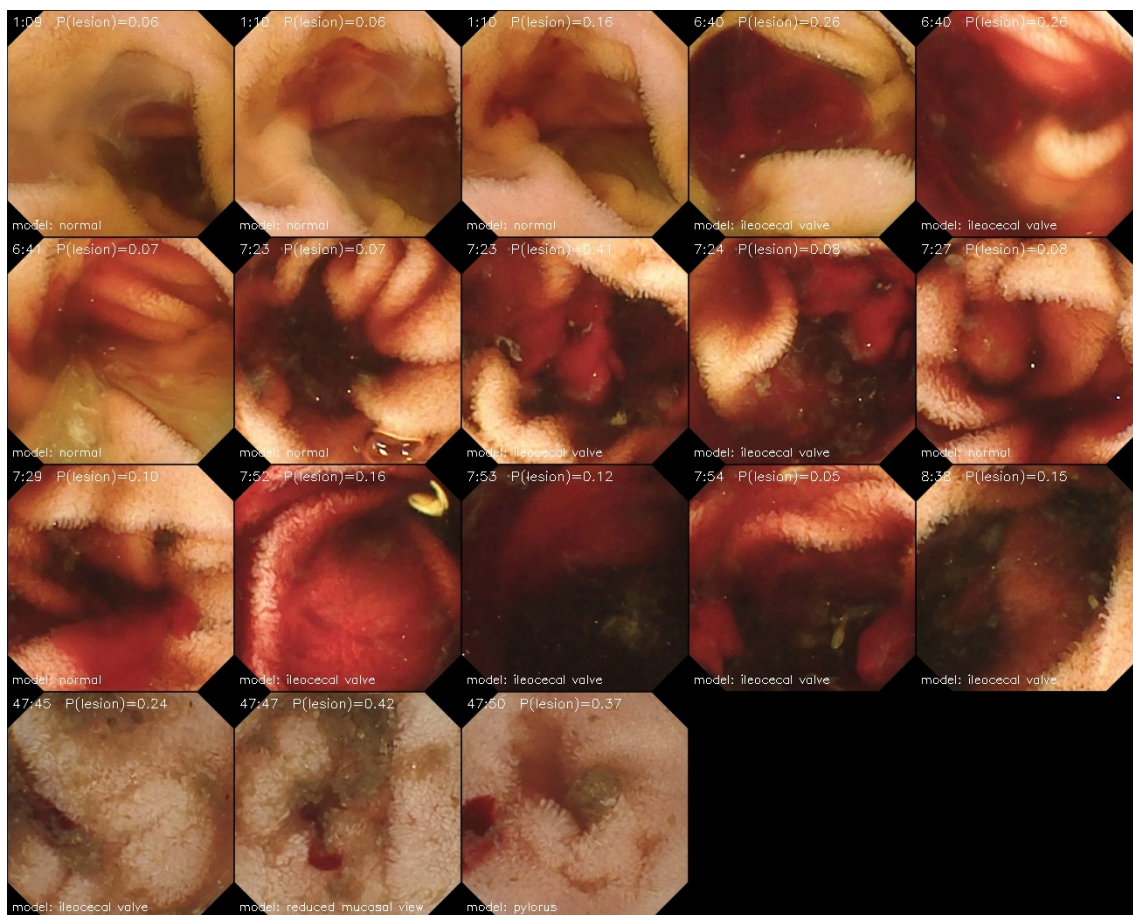


Figure 10: Annotated fresh-blood frames of the held-out video: lesion probability (top) and class predicted by the model that never saw the video (bottom)

The signal exists elsewhere: the clean classifier, trained on other capsule brands, ranks 91 % of these blood frames above the threshold giving 5 % of false alarms on normal frames. Two quick fixes were measured. Adding the distance-to-normal score to the model score was rejected (66 % of held-out lesions instead of 73 %). Adding the bleeding output of the clean classifier found 78 %, but this gain was measured on the very videos used to diagnose the problem, and therefore **cannot be claimed as an independent result**. It was not retained.

5.9.2 Phase 3: external bleeding data, audited

The proper fix is to put bleeding examples back into training. Capsule Vision images of the classes that Kvasir cannot teach (bleeding and polyp) were audited with the method of Section 5.5: **350 of 1996 were in fact Kvasir images** (312 of the 834 bleeding ones) and were removed. The remaining 522 bleeding and 1,124 polyp images, from other cameras, were added, each source being treated as one pseudo-patient by the sampler. All other settings were unchanged, and the four fold models and the final model were retrained.

5.9.3 External test

Independent evidence came from **WCEbleedGen** [6], a public set of bleeding and non-bleeding capsule images from other patients, cameras and annotators. The audit found that **467 of its 2618 images were already present in our training data** (18 %); they were removed, leaving 2151 images never seen by any model. On them, the bleeding output of the fold models reaches an AUROC of 0.81 to 0.89, and **0.90 for the final model**, which detects 69 % of bleeding images at 5 % of false alarms (Figure 11).

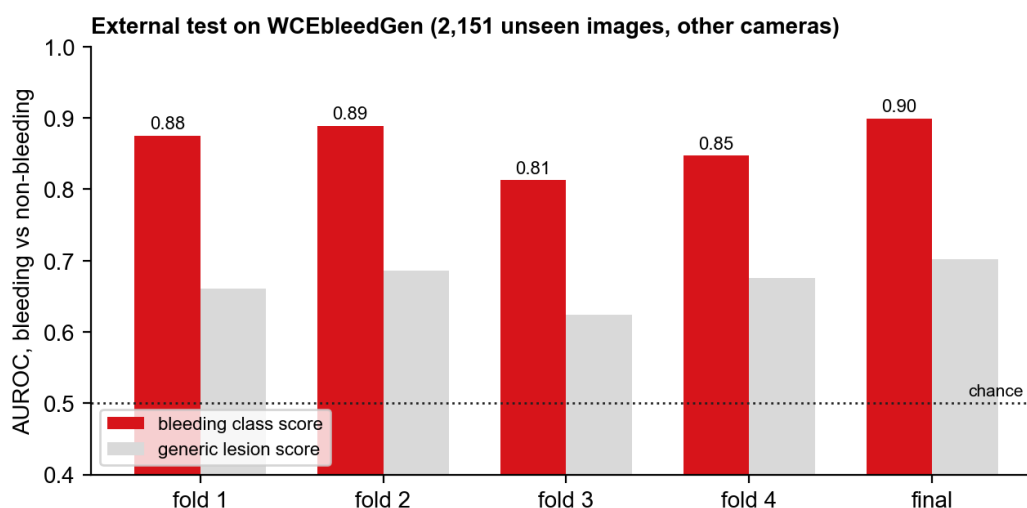


Figure 11: External test on WCEbleedGen: AUROC of each fold model and of the final model (1 = perfect, 0.5 = chance); red = bleeding output, grey = generic lesion score

5.9.4 Effect on the Kvasir videos

On annotated Kvasir frames, phase 3 improved detection on both groups: 92 % of tuning episodes (89 % in phase 2) and 86 % of held-out episodes at 0.9 % of normal frames alerting (82 % at 1.8 % in phase 2). Naming improved on held-out videos (87 % against 82 %; vascular lesions 10 of 14 against 6 of 14) and decreased slightly on tuning videos (68 % against 72 %).

On complete videos, at about one false place per minute, phase 3 finds 84 % of tuning lesions at 0.95 per minute (phase 2: 79 % at 0.89) and 71 % of held-out lesions at 1.22 per minute (phase 2: 73 % at 1.16). At the wider operating point, it finds 85 % of held-out lesions at 2.13 per minute.

The rule fixed before training (keep the new models only if detection does not decrease on tuning videos at equal reading burden) is satisfied: **the phase 3 models are retained**. On the held-out videos, the two versions are within two episodes of each other (56/79 against 58/79 at the default point, 67/79 against 68/79 at the wider one), while more of the lesions found are correctly named at the default point (53/79 against 51/79). On the bleeding video itself, the gain appears only at the wider operating point: 12 of its 14 lesions are found instead of 10; at the default point, bleeding frames are now scored as lesions but mostly below the

threshold (3 of 14 found, 4 before). The decision was taken by the rule, not by the held-out figures, which would otherwise have been used for tuning.

5.10 Integration and end-to-end verification

5.10.1 Integration

The detector (in production, the phase 3 model trained on all videos) is integrated into the pipeline behind a command-line option, without changing the behaviour of the previous version when the option is absent. In this mode, an event is one place: the flagged frames closer than 2 s form one episode, whose category is the vote of its frames, which is exactly the rule that was measured. The fine class most likely *within* that category is attached as a suggestion with its probability, never as a diagnosis. The triage gate is disabled, since the detector has its own "reduced mucosal view" class; it remains available on request.

5.10.2 Verification on the 16 videos

Figures produced by replaying cached scores are worth nothing if the real pipeline behaves differently. The complete pipeline (sampling, detection, grouping, Grad-CAM screenshots, localisation, JSON report) was therefore run on the 16 videos, each with the fold model that never saw it, and its decisions were compared with the replay on every video, first with the phase 2 models, then again with the final phase 3 models. In both runs, the lesion probabilities differ by at most 0.00000 and **the number and boundaries of events are identical on all 16 videos**. The first run revealed one bug before any figure was published: grouping by label, designed for the 10-class classifier, split one place into several events when the category changed between frames (39 events instead of 32 on one video). It was corrected so that the pipeline matches the measured rule.

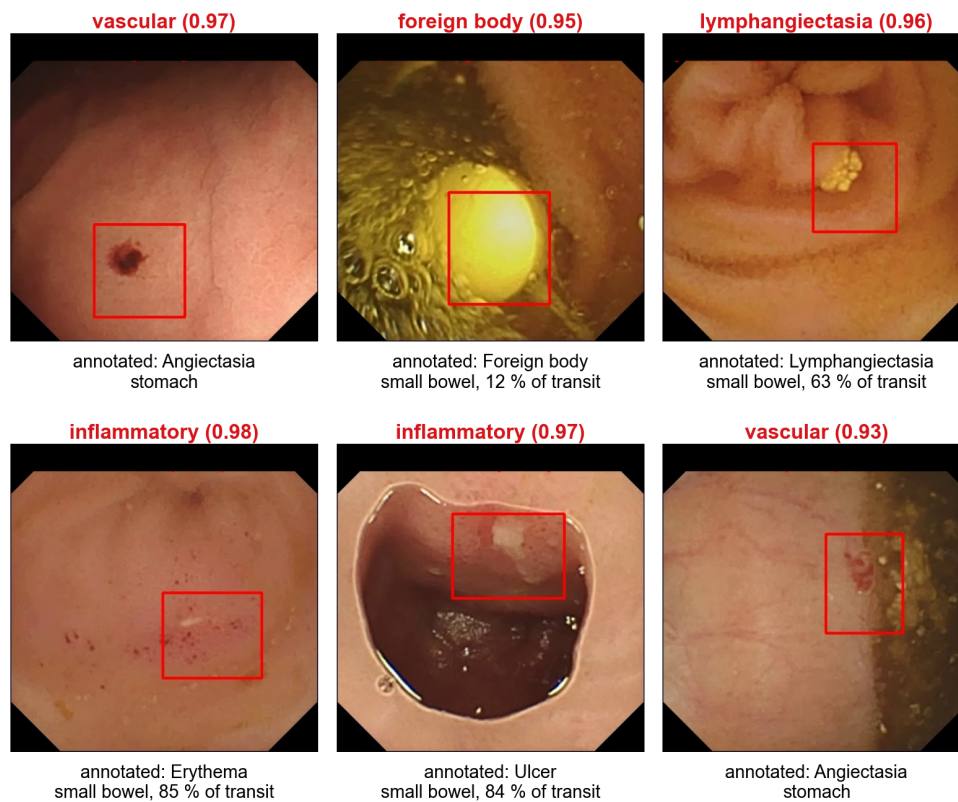


Figure 12: Examples of correct detections of the final pipeline, each on a video never seen by its model: predicted category and confidence (top), Grad-CAM box, annotated class and position given by the pipeline (bottom)

Figure 13 shows the limits of these results: on a whole video, many events fall outside the annotated lesions. Part of them are real findings that were not annotated (Section 5.6.2), but most are normal folds, bubbles or debris; this is the main limitation discussed below.

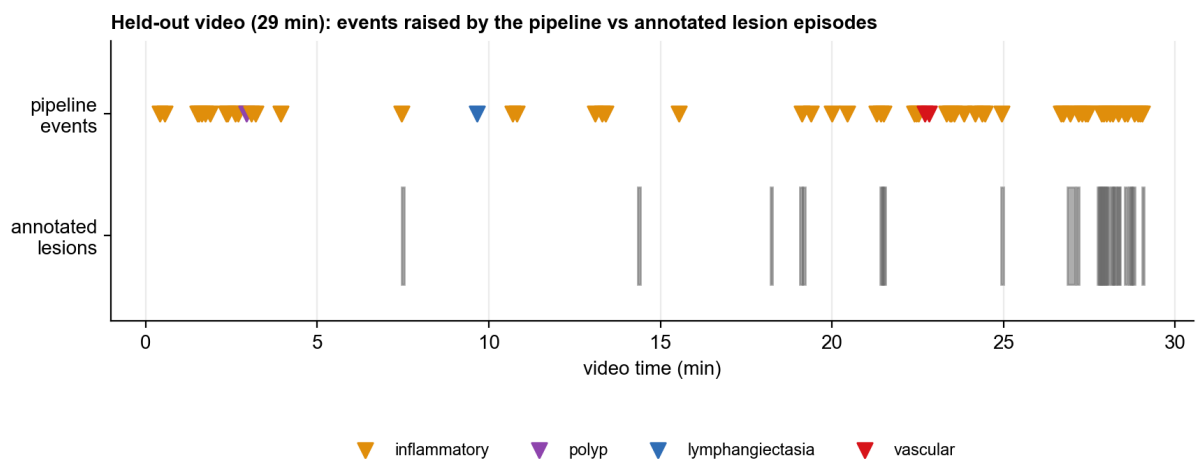


Figure 13: Events raised by the pipeline on a held-out video, coloured by clinical category, against the annotated lesion episodes: 63 events over 29 minutes for 29 annotated episodes, of which 25 are found

5.10.3 Final results

Table 8 gathers the final figures, all measured on videos never seen by the measured model.

Table 8: Final results on the 5 held-out videos (79 annotated lesion episodes)

Indicator	Initial rule	Phase 2 system	Final system (real pipeline outputs)
Lesions found, about 1 false place per minute	66 % at 1.14	73 % at 1.16	71 % at 1.22
Lesions found and correctly named	n/a	51/79	53/79
Lesions found, wider operating point	n/a	86 % at 2.06	85 % at 2.13
Patients with lesions who are flagged	n/a	n/a	13 of 13 (all 16 videos)
Lesions placed in the correct region (8 videos)	n/a	n/a	64 of 67 (landmark detector)
Bleeding, external test (AUROC)	n/a	n/a	0.90

On the 16 videos (391 minutes), the pipeline raises a median of 26 places to check per video, and every patient with annotated lesions is flagged. Processing runs on a laptop CPU at about 0.4 times real time (4.7 min for an 11-minute video, 21 min for a 48-minute one).

5.10.4 Limitations

- **About three lesions in ten are missed** at the default operating point; clinical reading software announces more than 90 %. The wider point reaches 85 % at about twice the reading burden; the choice between them is a clinical one.
- **Validation is limited to one hospital and one camera:** 16 local videos, and 21 Kvasir videos in cross-validation. Classes present in a single patient (polyp, digested blood) cannot be validated.
- **False-alarm figures are pessimistic**, since annotations are incomplete; their true value can only be measured with complete annotations.
- **The alert rate remains too high for clinical use.** At about one false place per minute, a complete examination of several hours would produce hundreds of places to check. A usable tool would need either far fewer false alerts, or a ranking showing only the most suspicious places first.
- **Bleeding** is validated on external images, not on external videos.
- **Localisation** in the pipeline relies on the landmark detector, whose crossing errors are probably optimistic; the more precise monotone segmentation is measured but not yet integrated. The detector can also place a valve crossing in a video that does not contain one, which then labels the following lesions as colon.

6 Project management

6.1 Method: decisions driven by measurement

The project was run as a sequence of short iterations, each one ending with a measurement. Three rules governed the decisions:

- **criteria fixed before measuring:** a change is adopted only if it meets a threshold decided in advance (for instance, at least 15 % fewer false places and no lesion lost), and a stop criterion (75 % of lesions detected) was set before the first large-scale experiment;
- **settings chosen on tuning videos only:** the held-out videos are measured once, with thresholds frozen beforehand;
- **every result is reproducible:** model outputs are cached, decision rules are replayed from the cache, and the real pipeline is checked against the replay.

These rules have a cost (several attractive ideas were abandoned when they failed their criterion) but they are what makes the final figures credible.

6.2 Timeline

Figure 14 shows the organisation of the internship. June was devoted to documentation: the public datasets, and the existing work of the group on capsule imaging. A first lesion classification model was built at the end of June and refined until its final version in August. From July, the localisation was studied: the precision that could be expected, and first trials, mainly by homography. In August, the landmark models were trained and the video pipeline was specified. At the beginning of September the models were closed and the pipeline was implemented; the rest of September was devoted to its evaluation, to the correction of the training data, to three retraining phases and to the final verification.

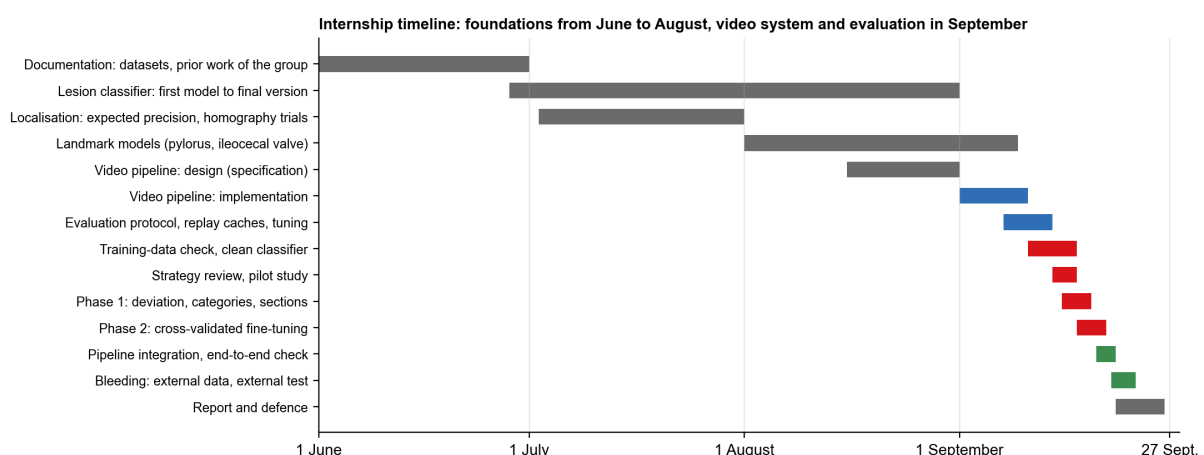


Figure 14: Timeline of the internship

6.3 Evolution of the technical choices and impact on the project

Four decisions reshaped the plan.

From metric trajectory to anatomical localisation. The trajectory objectives of the agreement could neither be realised nor validated without electromagnetic data or ground-truth trajectories (Section 4.2). Redefining the localisation as anatomical (region and fraction of small-bowel transit) kept the purpose of the subject while making it measurable. The electromagnetic refinement was moved to the perspectives.

An error that cost time. Training the first classifier on a public dataset without checking its overlap with the evaluation data (Section 5.5) invalidated about five days of tuning results. The response was a short stop (review of the state of the art and pilot study) followed by a new plan with explicit phases, a stop criterion and cross-validation by patient. The schedule was reorganised around three GPU runs of one to two hours, each tested locally on a synthetic dataset beforehand, so that no GPU session was lost to a programming error. Since then, every external dataset has been checked against the evaluation data before use.

Abandoned directions. Two directions were studied and deliberately dropped. The endoscopy foundation model GastroNet-5M [13] was no longer freely downloadable; since the pilot study had shown a generic foundation model to be clearly inferior to the clean classifier, the effort was redirected to fine-tuning. The Galar dataset [7], which would provide an external validation with the same camera, is distributed in archives of ten examinations of 23 to 203 GB (over 600 GB in total), incompatible with the remaining time; it is the first perspective of this work.

The bleeding problem. The end-to-end verification revealed a video where most lesions were missed. Its diagnosis (a class present in a single patient) led to a third training phase with external bleeding data, and to the search for an independent test set, WCEbleedGen. The phase was bounded by a rule decided before training: keep the new models only if detection on the tuning videos does not decrease.

6.4 Decision log

Table 9: Main technical decisions, with the measurement that motivated them

Decision	Measurement	Outcome
Triage gate on alerts (threshold 0.90)	−20 % false places, no lesion lost	Adopted, later replaced by the "reduced view" class
Margin versus "normal", temporal smoothing, bubble detectors	Failed the adoption criterion	Rejected
Retrain the classifier without duplicated images	16,315 evaluation images found in the training set	Adopted; previous results invalidated
Generic foundation model (DINOv2) as feature extractor	AUROC 0.70 against 0.92	Rejected
Capsule progression by homography between frames	No public video provides the true motion: no validation possible	Not retained; kept as a perspective with electromagnetic data
Crossings by monotone segmentation (anatomical order)	Pylorus 13 → 5 s, valve 55 → 35 s; but lesions in the correct region 64 → 46 of 67	Kept in the measurement code; not yet substituted in the pipeline
Enriched bank of normal frames	Detection 45 % at 1 false place/min	Rejected
Fine-tuning with cross-validation by patient	67 → 73 % of held-out lesions at equal burden	Adopted
One event per place, category by vote	Pipeline identical to the measured rule	Adopted
Score unions for bleeding	Rejected or not independently measurable	Not retained
External bleeding data after audit	External AUROC 0.90; no loss on tuning videos	Adopted (rule met)

6.5 Risk management

Table 10: Risks identified and how they were handled

Risk	Consequence	Mitigation
Leakage between training and test data	Overestimated results	Perceptual audit of every external dataset; cross-validation by patient
Tuning on the test videos	Overestimated results	Thresholds chosen on tuning videos; held-out videos measured once
Loss of a cloud GPU session	Hours of computation lost	Resumable notebooks; atomic writes of every result
Silent programming error	False results	Notebooks run end to end on synthetic data before each GPU run; consistency checks between saved features and scores
Replay different from real pipeline	Figures not valid for the system	Frame-by-frame comparison of the real pipeline and the replay on all 16 videos

6.6 Tools and resources

6.6.1 Software stack

The whole project is written in Python. Table 11 lists the libraries it relies on, with the versions of the local environment; training ran in the standard Kaggle GPU image, with the same libraries.

Table 11: Software stack of the project

Component	Version	Role in the project
Python	3.13	Language of the pipeline and of the measurements
PyTorch / torchvision	2.9 / 0.24	Networks (EfficientNet-B3, MobileNetV3-Small), training, inference, pre-trained ImageNet weights
OpenCV	4.13	Video decoding and frame sampling, drawing of the annotated screenshots
scikit-learn	1.6	Section probe (logistic regression), distance to normal (k-NN), AUROC and other metrics
NumPy / pandas	2.1 / 2.2	Caches of model outputs, replay of decision rules, result tables
Pillow	11.0	Image reading, perceptual fingerprints of the overlap check
matplotlib	3.9	Timelines of the event reports, result figures
git	n/a	Versioning: over 60 commits during the development of the video system

6.6.2 Models tested

Table 12: Models and methods tested, and what was kept

Model or method	Use	Outcome
EfficientNet-B3, ImageNet weights	Lesion classifier, then fine-tuned 12-class detector	Kept: final detector
MobileNetV3-Small	Landmark detector (pylorus, ileocecal valve)	Kept in the pipeline
Energy score on the classifier outputs	Unknown anomalies without extra training	Used in the first pipeline, replaced by the fine-tuned detector
Small 3-class triage network	Discard alerts on normal or unusable frames	Adopted, then made unnecessary by the "reduced view" class
Bubble detectors (Hough circles; EfficientNet-B0 on 300 labelled frames)	Remove alerts on bubbles	Rejected: lesions lost
DINOv2 ViT-S/14	Generic foundation model as feature extractor	Rejected: AUROC 0.70 against 0.92
GastroNet-5M	Endoscopy foundation model	Not tested: no longer freely downloadable
Homography between frames (OpenCV)	Capsule progression	Not retained: no ground truth to validate it
Logistic regression + monotone segmentation	Section of each frame, landmark crossings	Measured, next to integrate

6.6.3 Code and hardware

The code is organised in 41 Python modules (about 6,200 lines) and 10 notebooks for feature extraction and training. The pipeline modules are listed in Appendix D; the measurement modules replay every decision rule on cached outputs. All video processing, measurements and figures ran on a laptop processor without a graphics card; training used the free Kaggle quota (NVIDIA Tesla T4, about 30 hours per week, sessions of at most 12 hours), of which about 3 hours of GPU time for the final models.

7 Conclusion

The objective of this internship was to locate the capsule within the gastrointestinal tract in order to locate and characterise pathologies in capsule endoscopy videos. Since the metric trajectory of the initial subject could not be measured with the available data, the objective was redefined as an anatomical localisation, and a complete pipeline was built: from a capsule video, it produces for each suspected lesion a timecode, an annotated screenshot, a clinical category with a fine-class suggestion, and the position of the lesion within the digestive tract.

The results were obtained with a protocol in which every video is judged by a model that has never seen it, and the final figures were measured once on held-out videos. With this protocol, the final system finds **71 % of held-out lesion episodes at 1.22 false places per minute**, against 66 % for the initial decision rule at the same burden, and 85 % at a wider operating point. Every patient with annotated lesions is flagged, and most lesions found receive the correct clinical category. The pipeline places 64 of 67 lesions in the correct anatomical region on the videos where this can be compared, with a median error of 0.8 percentage points on their position within the small bowel; a monotone segmentation, measured separately, places the pylorus and the ileocecal valve more precisely on the same videos (5 s and 35 s in median, against 13 s and 55 s) and is the next component to integrate. Bleeding detection, first unlearnable from a dataset where bleeding is present in a single patient, reaches an AUROC of 0.90 on an external set of images never seen by the model.

An error in the choice of the first training data (a public dataset containing images of the evaluation videos) was found and corrected during the internship (Section 5.5); every figure above comes from the corrected protocol.

The objectives are therefore reached for anatomical localisation at the level of regions, with a more precise placement of the crossings measured and ready to be integrated; partially reached for detection, where about three lesions in ten are still missed at the default operating point, against more than 90 % announced for clinical software; and not reached for the metric trajectory, which requires data not available during the internship.

The priorities for future work are, in order:

- **integration of the monotone segmentation** into the pipeline, with the landmark detector as a check on the pylorus, validated with a landmark detector retrained without the evaluation videos (cost: a few days, all on existing data);
- **external validation with the same camera** on the 22 Olympus examinations of the Galar dataset, without any retraining (cost: about two weeks, most of it for downloading and converting 200 to 600 GB of data);

- **complete annotation** of a few local videos by a physician, to measure the true false-alarm rate, currently overestimated (cost: a few hours of clinical time per video);
- **more patients for rare lesions** (polyps, bleeding), the main limit of detection (possible only through a clinical collaboration);
- a **temporal model** for the section boundaries, suggested by the 2026 challenge results (cost: a few days of GPU training);
- the **electromagnetic refinement** of the capsule position, as in the initial subject, when synchronised electromagnetic and video recordings become available.

Bibliography

- [1] Institute of Systems and Robotics, University of Coimbra, "ISR – Coimbra." [Online]. Available: <https://www.isr.uc.pt/> (accessed Sep. 2026).
- [2] ARISE Associate Laboratory, "ISR-UC — Institute of Systems and Robotics, University of Coimbra." [Online]. Available: <https://arise-la.pt/isr-uc/> (accessed Sep. 2026).
- [3] S. H. Kim, D.-H. Yang and J. S. Kim, "Current status of interpretation of small bowel capsule endoscopy," *Clinical Endoscopy*, vol. 51, no. 4, pp. 329–333, 2018.
- [4] P. H. Smedsrud et al., "Kvasir-Capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, art. 142, 2021.
- [5] P. Handa et al., "Capsule Vision 2024 Challenge: Multi-class abnormality classification for video capsule endoscopy," arXiv:2408.04940, 2024.
- [6] P. Handa et al., "WCEbleedGen: A wireless capsule endoscopy dataset and its benchmarking for automatic bleeding classification, detection, and segmentation," arXiv:2408.12466, 2024. Dataset: Zenodo, record 10156571.
- [7] M. Le Floch, F. Wolf, L. McIntyre et al., "Galar — a large multi-label video capsule endoscopy dataset," *Scientific Data*, vol. 12, art. 828, 2025.
- [8] D. Hanson and D. Jha, "Benchmarking the domain gap: Model selection instability under domain shift in video capsule endoscopy," arXiv:2607.22736, 2026.
- [9] B.-C. Qiu, F.-Y. Lin, M.-H. Sun, Y.-F. Lin, C.-M. Lee and C.-C. Hsu, "VISTA: Validation-guided integration of spatial and temporal foundation models with anatomical decoding for rare-pathology VCE event detection," arXiv:2605.22096, 2026.
- [10] J. Won, S. Lee and S. K. Jung, "GALAR-TemporalNet v2: Anatomy-guided dual-branch temporal classification with bidirectional Mamba and dual-graph GCN for video capsule endoscopy," arXiv:2605.22209, 2026.
- [11] G. Son, T. Eo, J. An, D. J. Oh, Y. Shin, H. Rha, Y. J. Kim, Y. J. Lim and D. Hwang, "Small bowel detection for wireless capsule endoscopy using convolutional neural networks with temporal filtering," *Diagnostics*, vol. 12, art. 1858, 2022.
- [12] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.
- [13] "GastroNet-5M: A multicenter dataset for developing foundation models in gastrointestinal endoscopy," *Gastroenterology*, 2025.
- [14] M. Oliveira and H. Araújo, "Local forward-motion panoramic views for localization and lesion detection for multi-camera wireless capsule endoscopy videos," in *Proc. 16th Int. Joint Conf. on Biomedical Engineering Systems and Technologies (BIOSTEC)*, SCITEPRESS, 2023.
- [15] M. Oliveira, H. Araújo, I. N. Figueiredo, L. Pinto, E. Curto and L. Perdigoto, "Registration of consecutive frames from wireless capsule endoscopy for 3D motion estimation," *IEEE Access*, vol. 9, pp. 119533–119545, 2021.
- [16] B. Rousso, S. Peleg, I. Finci and A. Rav-Acha, "Universal mosaicing using pipe projection," in *Proc. IEEE ICCV*, 1998, pp. 945–952.
- [17] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. ECCV*, 2020, pp. 402–419.
- [18] E. Macedo, "Thesis – Chapter 3 (version 1): Literature review," PhD thesis in preparation, 25 Feb. 2024, Institute of Systems and Robotics, University of Coimbra, unpublished.
- [19] E. Dubrofsky, "Homography estimation," Master's essay, University of British Columbia, Vancouver, 2009.
- [20] University of Coimbra, "Plano para a Igualdade, Equidade e Diversidade envolve toda a UC" and "Universidade de Coimbra distinguida como instituição pioneira na promoção da igualdade de género," UC news. [Online]. Available: <https://www.uc.pt/sustentabilidade/> and <https://noticias.uc.pt/> (accessed Sep. 2026).
- [21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

- [22] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. ICML*, 2019, pp. 6105–6114.
- [23] W. Liu, X. Wang, J. Owens and Y. Li, “Energy-based out-of-distribution detection,” in *Proc. NeurIPS*, 2020.
- [24] A. Howard et al., “Searching for MobileNetV3,” in *Proc. IEEE ICCV*, 2019, pp. 1314–1324.

Appendices

Appendix A: Example of event produced by the pipeline

The first event of the report of a held-out video (29 minutes), as written by the pipeline. The anatomical fields give the region and, within the small bowel, the third and the position in percent of transit time.

```
{
  "video_id": "d626f4f4a5ac4785",
  "timecode_s": 447.5,
  "frame_index": 13425,
  "screenshot_path": "screens/d626f4f4a5ac4785/evt_0016.jpg",
  "detection": {
    "type": "known",
    "label": "inflammatoire",
    "confidence": 0.9273,
    "ood_score": null
  },
  "image_localization": {
    "bbox_xyxy": [105, 159, 172, 233]
  },
  "anatomical_localization": {
    "region": "small_bowel",
    "subpart": "proximal",
    "progress_in_subpart_pct": 85.9,
    "progress_in_bowel_pct": 28.6
  },
  "episode": {
    "debut_s": 447.5,
    "fin_s": 447.5,
    "n_frames": 1
  },
  "suggestion_classe_fine": {
    "classe": "erosion",
    "probabilite": 0.8792
  },
  "zone_id": 16
}
```

Appendix B: Running the pipeline

One command, shown on three lines:

```
python run_pipeline.py --video data/sample_videos/<video>.mp4
--detecteur-ckpt models/phase3/phase3_final.pt
--landmark-ckpt models/landmark_model.pt --out-dir outputs
```

Default operating point: lesion threshold 0.8532 (about 1 false place per minute).
 Wider operating point: add `--seuil-lesion 0.2277 --min-images-evenement 2`.
 Without `--detecteur-ckpt`, the first version of the pipeline is used unchanged.

Appendix C: Results per video

Real outputs of the pipeline on the 16 local videos, each processed by the final-phase fold model that never saw it.

Table 13: Pipeline outputs per video (held-out videos last)

Video	Group	Duration (min)	Annotated lesions	Found	Named	False places	Alerts
04a78ef0...	tuning	28.3	9	9	9	88	99
131368cc...	tuning	17.6	1	1	0	6	7
2fc3db47...	tuning	42.7	4	3	2	49	52
3ada4222...	tuning	22.5	27	18	18	7	18
64440803...	tuning	18.4	16	16	7	21	32
89cdd412...	tuning	16.6	0	0	0	2	2
8ebf0e48...	tuning	16.6	7	6	5	17	27
b2134f4a...	tuning	16.4	2	2	2	2	4
bca26705...	tuning	26.9	7	6	4	22	26
df6b47ba...	tuning	10.7	0	0	0	0	0
fe5d372e...	tuning	14.9	0	0	0	7	7
5e59c7fd...	held-out	26.7	16	15	15	43	60
7a47e8ea...	held-out	36.6	7	6	6	26	31
7ad22d50...	held-out	18.8	13	7	4	14	19
d369e4f1...	held-out	48.5	14	3	3	66	67
d626f4f4...	held-out	29.1	29	25	25	45	63

Appendix D: Organisation of the code

Table 14: Main files of the project

File	Role
run_pipeline.py	Command-line entry point
src/pipeline.py	Orchestration: sampling, detection, grouping, screenshots, localisation, JSON and timeline
src/detecteur_lesions.py	Fine-tuned lesion detector: lesion probability, category vote, fine-class suggestion
src/episodes.py, src/zones.py	Grouping of flagged frames into events and places to check
src/landmarks.py	Anatomical localisation used by the pipeline (landmark detector, peak and transit constraint)
src/_phase1_sections.py	Monotone segmentation of section probabilities (measurement code)
src/_evaluate_final.py, src/_rejouer.py	Evaluation indicators and replay of decision rules on cached outputs
src/_phase1_*.py, src/_phase2_*.py	Measurements of phases 1 to 3
src/_verifier_pipeline_phase2.py	Frame-by-frame check of the real pipeline against the replay
train_classifier_propre.ipynb	Overlap check between datasets and clean classifier
train_phase2_plis.ipynb, train_phase3_sang.ipynb	Cross-validated fine-tuning; external data and external test

Abstract

This internship, carried out at the Institute of Systems and Robotics of the University of Coimbra, addresses the automatic analysis of wireless capsule endoscopy videos. A pipeline was developed that processes a complete video and reports, for each suspected lesion, its timecode, an annotated screenshot, a clinical category, and its anatomical position: stomach, small bowel or colon, and position within the small bowel as a fraction of transit time. All models were trained and evaluated with cross-validation by patient, every video being judged by a model that never saw it, after an overlap between two public datasets had been found and removed. On held-out videos, the final system finds 71 % of lesion episodes at about one false place to check per minute, against 66 % for the initial method, flags every patient with lesions, places 64 of 67 comparable lesions in the correct anatomical region, and detects bleeding on an external dataset with an AUROC of 0.90. Its limits are stated explicitly: about three lesions in ten missed at the default setting, validation restricted to one camera and one hospital.

Keywords: capsule endoscopy, deep learning, lesion detection, anatomical localisation, cross-validation by patient.

Résumé

Ce stage, réalisé à l'Institut des systèmes et de la robotique de l'Université de Coimbra, porte sur l'analyse automatique des vidéos de vidéocapsule endoscopique. À partir d'une vidéo complète, la chaîne de traitement développée fournit, pour chaque lésion suspecte, son instant, une capture annotée, une catégorie clinique et sa position anatomique : estomac, intestin grêle ou côlon, puis position dans le grêle en fraction du temps de transit. Tous les modèles ont été entraînés et évalués en validation croisée par patient, chaque vidéo étant jugée par un modèle qui ne l'a jamais vue, après la suppression d'images communes à deux jeux de données publics. Sur les vidéos tenues à l'écart, le système retrouve 71 % des épisodes de lésion pour environ un faux endroit par minute, contre 66 % pour la méthode initiale, signale tous les patients porteurs de lésions et détecte le saignement sur un jeu externe avec une AUROC de 0.90. Ses limites sont explicitées : trois lésions sur dix manquées, un taux d'alertes trop élevé pour un usage clinique, une validation limitée à une caméra et à un hôpital.

Mots-clés : vidéocapsule endoscopique, apprentissage profond, détection de lésions, localisation anatomique, validation croisée par patient.