Diagnostic accuracy of DeepClass dysplasia classification compared to pathologists' classification from colon histology of polypectomised subjects who underwent endoscopy: a cross-sectional study

Eliane MAALOUF

(FICTIONAL) STUDY PROTOCOL

CAS en Recherche Clinique 2024-2025

11.08.2025

Table of contents

1. Introduction

1.1 Context and rational

Colon cancer (CC) is the most frequent sub-type of colorectal cancer. It had the fifth highest incidence and mortality rates in 2022 in western countries (Europe and North America) according to WHO Global Cancer Observatory data (IARC, 2025). Screening programs are a central element of CC prevention. Screening methods include colonoscopy, flexible sigmoidoscopy, Fecal Occult Blood Testing (European Commission, Karsa, Segnan, & Patnick, 2010). Colonoscopy is the gold standard method to detect colon polyps and adenoma (Jover, Bretthauer, & Dekker, 2021). When polyps are observed during colonoscopy, they are removed. A histology assessment is made on the extracted tissue. Pathologists analyse Hematoxylin and Eosin (H&E) stained tissues to assess cell's dysplasia/neoplasia grade as none (normal cell), low-grade, high-grade or carcinoma (European Commission, Karsa, Segnan, & Patnick, 2010). This classification is crucial in informing about the likelihood of a lesion's progression into invasive cancer, to guide surveillance intervals, follow-up treatment and diagnosis decisions (Truninger, Lugli, & Koeberle, 2022). Switzerland initiated large scale screening programs in low-risk groups once every 10 years from 50 years of age, and earlier and more frequently in higher risk groups (e.g., family history of colon cancer, previously observed polyps, Inflammatory Bowel Disease). Such programs inevitably lead to more polypectomies and consequent histopathology examinations. Furthermore, modern endoscopes are equipped with Computer-aided detection (CADe) tools that are known to increase the number of detected polyps, polypectomies and histopathological assessments (Jin, Ma, Shi, & Cai, 2024) (Soleymanjahi, Huebner, & Elmansy, 2024) (Mori, et al., 2023). Increased detection rates and subsequent examinations are expected to increase the short-term health costs despite a potential reduction of those costs in the long term (Mori, et al., 2023). With shortages of pathologists and increasing workloads, assistive technologies in the histology assessment workflow become crucial in delivering reliable and consistent results quickly and at scale. Artificial intelligence is being developed in multiple areas of digital pathology and diagnosis showing competitive results (McGenity, Clarke, & Jennings, 2024) (Bera, Schalper, & Rimm, 2019). Their rigorous validation remains a crucial phase in establishing a new diagnostic tool in clinical practice following the development phase (Adams & Leveson, 2012).

DeepSpot (Nonchev, et al., 2025) is a deep-learning algorithm that predicts spatial transcriptomics from H&E images and showed integrability in multiple downstream clinically relevant tasks. It was developed on curated and standardized datasets for a variety of cancers and cells types. It won the "Autoimmune disease machine learning challenge" (Crunch Foundation and The Eric and Wendy Schmidt Center, 2024), an online competition to develop spatial transcriptomics prediction tools. Despite its versatility and adaptability, the performance of DeepSpot-integrated workflows remains to be proven in specific clinical contexts.

DeepClass is a new histopathology diagnostic workflow that integrates DeepSpot for colon cells' dysplasia classification. The workflow takes as input Whole Slide Images (WSI) (Kumar, Gupta, & Gupta, 2020) of H&E stained colon tissues and outputs a segmentation of cells clusters in the images with a dysplasia grade as "negative for dysplasia", "low-grade", "high-grade", "adenocarcinoma" and "undetermined" attached to each cluster. The full image is classified with the highest grade observed in the clusters.

1.2 Research question and hypothesis

This study aims to answer the following: "*What is the diagnostic accuracy (sensitivity/specificity) of DeepClass dysplasia classification compared to pathologists' classification in colon's polypectomised subjects who underwent endoscopy?*". Histology assessment is systematically performed after colon polypectomy (Truninger, Lugli, & Koeberle, 2022) and is a decisive step for risk classification. Furthermore, the pathologists' decisions is the current reference in multiple diagnostic quality assessments (McGenity, Clarke, & Jennings, 2024). A high degree of concordance between DeepClass classification and pathologists' classification would be a good indicator of clinical relevance of DeepClass. We further aim to assess the hypothesis that DeepClass has a significantly higher sensitivity than the pathologists for dysplasia classification. Despite the clinical relevance of the resulting evidence from our study, we are not concerned with proving clinical benefits of DeepClass (e.g., cost reduction, quality of care improvement, etc.).

1.3 Main and secondary objectives

Our main objective is to assess the sensitivity and specificity of DeepClass, with a primary focus on sensitivity. A secondary objective is to explore the dependence of sensitivity/specificity on polyps' characteristics (such as type and size) and subject related variables (such as sex, age and comorbidities).

2. Methods

2.1 Study type

Our study methodology follows a retrospective cross-sectional design.

More specifically, we collect existing colon polyps' tissues data from existing pathology biobanks. Our format of interest is the previously fixed and H&E stained tissue. If a WSI already exists for the stained tissue we only collect the image along the metadata including the previous pathologists' dysplasia classification. If such image doesn't exist, we first request or proceed to its generation. Since all tissues materials, variables measurement and pathologists' assessment are already present at the time of the start of the study, our design is retrospective. However, in order to compare the pathologists' diagnostic accuracy to DeepClass accuracy, we construct an experts panel to assess the same images in the sample and to deliver the "Gold standard" classification.

All WSI are provided as input to DeepClass to classify dysplasia level in each image. The pathologists' classification was collected for all the tissues pertaining to the same WSI. The experts' panel will also classify the same WSI. Figure 1 illustrates the study flow from sample selection to analysis. In terms of temporality, we can assume that both classifiers, the pathologists and DeepClass, as well as the "Gold standard", are being applied simultaneously. First, we can safely assume that the WSI used represent the same tissue that the pathologists assessed in the past. Second, if WSI are generated during the study, we can also safely assume that the fixed and stained tissue slides are correctly preserved to guarantee that the digitalized material at the time of the study is the same biological material at the time of collection. Furthermore, we consider that the tissue collection and the outcomes (i.e., for both classifiers and for the gold standard) are measured at the same time. In that sense the study is a paired-sample cross-sectional design.

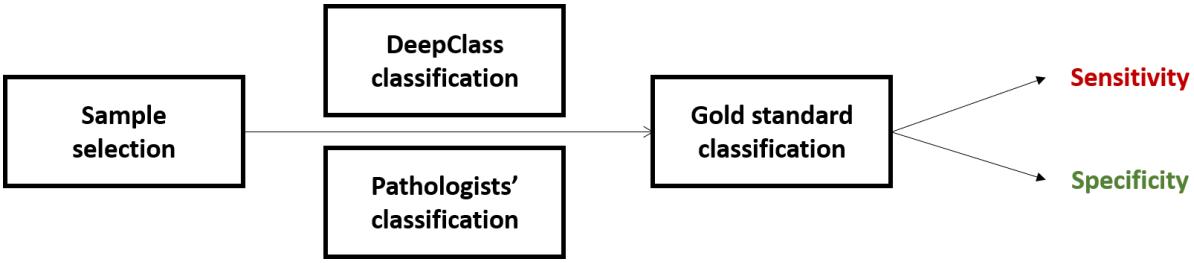

FIGURE 1 - STUDY FLOW DIAGRAM

2.2 Population (eligibility criteria)

Target population contains adult (≥ 18) subjects at risk of developing colon cancer. We base our classification comparisons and clinically relevant cut-offs on recommendations by Truninger et al. (Truninger, Lugli, & Koeberle, 2022). Those guidelines do not apply to subjects with a family history or hereditary colon cancer or with a Serrated Polyposis Syndrome. Therefore, our target population does not contain those subjects either and we exclude them from our analysis.

Source population is the population of adult subjects undergoing a colonoscopy during which one or more polyps were biopsied. Specifically, our subjects correspond to those with tissues stored in three large Swiss university hospitals biobanks (detailed in section 2.3 & 4.5). Subjects included in these biobanks might have participated in cantonal CC screening programs, might have participated in colon related research projects that involved polypectomies (e.g., cohort studies) or might have been prescribed a colonoscopy by their doctor (e.g., general practitioner, gastroenterologist) as part of their routine checkups or as monitoring of a comorbid condition (e.g., Inflammatory Bowel Disease, diabetes, etc.).

Eligibility criteria:

- Subjects who consented to the secondary use (outside of the medical procedure where the biopsy was required) of their tissue samples and relative metadata for the purpose of scientific research. These subjects could have participated in cantonal CC screening programs, in colon related clinical research projects or were prescribed a colonoscopy by their doctor.
- Aged 18 or higher at the time of the biopsy
- Colon polyp tissue not including rectum tissue. The polyp types of interest are: adenomatous and serrated (hyperplastic, sessile serrated lesions, traditional serrated adenomas)

Exclusion criteria:

- Subjects suspected of a hereditary colon cancer, Serrated Polyposis Syndrome, or having a family antecedent of colon cancer

Comparing target and source populations:

Some of the main differences between the target and source population might concern the following aspects, which should be accounted for during analysis (see section 2.6 – secondary analysis):

- the proportion of older adults (i.e., ≥ 50 years) might be higher in the source population compared to its proportion in the target, mostly issued from large scale screening programs.
- the proportion of males in the source population might be higher than their proportion in the target population, mostly because women are usually underrepresented in health research in general.
- the source population might exhibit more frequent occurrences of the comorbidities compared to their frequency in the target population. This would happen because subjects suffering from these health conditions are prescribed colonoscopies more often.

2.3 Participants selection

A random sample of colon polyp data from three pathology tissue biobanks (Centre Hospitalier Universitaire Vaudois - Biobanque de Pathologie; Hôpitaux Universitaires de Genève - Geneva Pathology Tissue Biobank; Biobank Pathology at University of Basel) from eligible subjects. Tissues should have been collected between January 2022 and December 2024. The 2022-2024 timeframe is selected to fit with the latest modifications of classifications and recommendations of naming of polyps and dysplasia grading (Truninger, Lugli, & Koeberle, 2022) (Swiss Society of Pathology, 2011). Furthermore, it is known that the staining degrades with time. If a WSI was not taken at the time of the fixed tissue storage, we would be required to make such image at the risk of producing a low-quality image (risk of being discarded from analysis).

2.4 Compared tests description

Pathologists' diagnosis (current reference test): our reference is the previously applied pathologist dysplasia classification for each of the retained tissue slides. The pathologist assessment is not repeated in this study. We expect the classification to be one of, or equivalent to, "negative for dysplasia", "low-grade", "high-grade", "adenocarcinoma".

DeepClass diagnosis (index test): DeepClass is the index test being evaluated. All retained WSI will be presented as input to DeepClass which is expected to output one of the following classes for each: "negative for dysplasia", "low-grade", "high-grade", "adenocarcinoma" and "undetermined". The "undetermined" WSI will be dropped from the final assessment. Previous reference test results and other variables (see Section 2.5) are not provided as input to the DeepClass workflow. No fine-tuning of the underlying model is performed in this study.

Postprocessing of the classifications: the recommendations by Truninger et al. (Truninger, Lugli, & Koeberle, 2022) distinguish two situations and their corresponding dysplasia grading cutoffs: (a) with serrated polyps the absence or presence of dysplasia (independent of the grade) is the most important; (b) with adenomatous polyps it is important to distinguish low grade from high grade when dysplasia is present. Therefore, we will assess DeepClass sensitivity/specificity in two classification situations:

1) Absence vs presence of dysplasia: regardless of the polyp type, we group the three dysplasia grades classes ("low-grade", "high-grade", "adenocarcinoma") as the "positive" class and "negative for dysplasia" as the "negative" class. We perform this postprocessing for both the reference and the index tests classifications as well as the "Gold standard".
2) Low-grade vs high-grade dysplasia: regardless of the poly type and specifically for the polyps where dysplasia was marked as "positive", we group the classes "high-grade", "adenocarcinoma" as the "positive" class and the "low-grade" as the "negative". We perform this postprocessing for both the reference and the index tests classifications as well as the Gold standard.

Gold standard: a panel of 5 experts will assess the images in the sample and provide an assessment for: absence vs presence of dysplasia as "positive" for presence or "negative" for absence; low-grade vs high-grade dysplasia as "positive" for high-grade or "negative" for low-grade; or "undetermined" in case the image is of low quality. In a large sample size like ours, we avoid running all WSI in the sample by each expert. Instead, we divide the sample into subsets, each subset is assessed by 3 experts (see section 2.6 – Sample size estimation). Each WSI receives its "Gold standard" classification, for each classification task, by majority voting among the 3 experts assigned classes. The experts on the panel will receive the WSIs and the patient metadata attached to it without the previous pathologists' classifications.

2.5 Measures

- ### Outcomes

Our main outcomes are sensitivity and specificity assessed in both classification tasks presented above.

To assess these outcomes, we require for each polyp tissue in the sample:

1) Pathologists assigned dysplasia grade (postprocessed as described in Section 2.4);
2) DeepClass assigned dysplasia grade (postprocessed as described in Section 2.4)
3) "Gold standard" assigned dyplasia presence/absence and low grade/high grade classifications.

- ### Other variables

Auxiliary variables required in the secondary analysis, for each polyp tissue in the sample:

1) Polyp type: adenomatous or serrated (hyperplastic, sessile serrated lesions, traditional serrated adenomas)
2) Polyp size: to be categorized as small (6-9 mm), medium (≥10 mm - <20mm), large (≥20 mm). These are common size classifications (Dornblaser, Young, & Shaukat, 2024).
3) Subject age: to be categorized in 3 groups, G1 (18 to 40 years), G2 (41 to 60 years), G3 (≥60 years).
4) Subject sex
5) Subject comorbidities: Inflammatory Bowel Disease, diabetes, hypertension, smoking status, other

2.6 Analysis

- ### Sample size estimation

Assumptions:

In both classification situations, correctly detecting the positive class is more critical than correctly detecting the negative class. We will base our sample size estimation on sensitivity.

Since distinguishing low-grade from high-grade dysplasia is performed on a subsample (i.e., where dysplasia was previously identified as present), we compute the sample size to guarantee that the sensitivity assessment in low-grade vs high-grade classification is correctly powered. The prevalence of high-grade dysplasia in colon polyps increases with the polyp size (Pickhardt, Hain, & Kim, 2010). This prevalence also varies in the average-risk population with different factors such as age, sex and dietary

and smoking habits (National Colorectal Cancer Roundtable Advanced Adenoma Working Group, 2018). Different types of polyps also have largely varying prevalences (National Colorectal Cancer Roundtable Advanced Adenoma Working Group, 2018). Sessile Serrated polyps' prevalence is between 2% to 9% (National Colorectal Cancer Roundtable Advanced Adenoma Working Group, 2018) in the average-risk group. Adenomatous polyps are generally more frequent with prevalences > 60% (Bulur & Çakır, 2021). In terms of dysplasia grades, low-grade is approximately 59% prevalent in polyps while the high-grade is approximately 3% prevalent in polyps (Bulur & Çakır, 2021). To account for the rarity of the high-grade dysplasia, we will inflate the sample size by the prevalence of 3%. This simultaneously accounts for the rarity of Sessile Serrated polyps considering an average prevalence of 4%. We will consider the general prevalence of dysplasia in polyps to be around 62%.

The hypothesis test of interest here is the two-sided McNemar test (Sundjaja, Shrestha, & Krishan, 2023) (paired-samples two proportions comparison) to detect a statistically significant difference between the sensitivity of DeepClass compared to the sensitivity of the pathologists. Pathologists' concordance is commonly set to 70% (Perlo, Tartaglione, Bertero, Cassoni, & Grangetto, 2022). We assume a similar level of concordance between the pathologists' classification and DeepClass classification. We aim for a power of 80% and risk alpha at 5%.

We aim to detect at least 15 points of percentage increase in sensitivity with DeepClass compared to the reference sensitivity, in both classification assessments. Literature on the assessment of pathologists' sensitivity in the detection of high-grade dysplasia is sporadic. We roughly estimated it to be around 45% from data provided in (Denis, Peters, Chapelain, Kleinclaus, & Fricker, 2009) for community pathologists. We expect expert pathologists to have much higher sensitivity in the 80% range. For our sample size estimation, we will use a 75% sensitivity for the pathologist's diagnosis (denoted as Reference below), for both classification tasks. Therefore, we set the expected DeepClass (denoted as Index below) sensitivity to be 90%.

Estimation:

Estimation of discordant pairs based on 70% concordance rate, 75% sensitivity of the reference test, 90% sensitivity of the index test gives:

	Index +	Index -	Total
Reference +	**67.5 (concordance)**	**7.5 (7.5% disc.)**	**75 (Ref. sens.)**
Reference -	**22.5 (22.5 % disc.)**	2.5	25
Total	**90 (Index sens.)**	10	100

Using the sample size estimation for a two-sided McNemar test comparing two proportions on a paired sample we require n_1 = 103 cases with high-grade dysplasia. Inflating for the high-grade dysplasia prevalence of 3%, we get n_2 = 103/0.03 = 3'433 polyp tissues to assess.

To further account for dysplasia prevalence, n_2 should account for the 62% of the whole sample leading to n_3 = 3433/0.62 = 5'537 polyp tissues to assess.

Furthermore, we will consider that in the final sample we will drop some cases because they might be undetermined (e.g., due to low WSI quality). We assume this drop rate to be a further 10%. We inflate

the sample such that n_3 would coincide with 90% of the final sample size, this leads to $n_4 = n_3/0.9 = 6'152$ polyp tissues to assess.

We aim for a sample size of **N = 6'155** polyp tissues to assess, guaranteeing adequate power to significantly detect an increase in sensitivity from the reference 75% to 90% with DeepClass in both classification tasks.

Note about the application of the "Gold standard" assessment:
We divide the total sample into subsets of 3'693 images (and subgroups of 1'231 images each). Each expert will assess one subset. With 5 experts in total, each subset will be assessed by 3 experts.

	G1	G2	G3	G4	G5
size	1231	1231	1231	1231	1231
Expert 1	X	X	X		
Expert 2		X	X	X	
Expert 3			X	X	X
Expert 4	X			X	X
Expert 5	X	X			X

- Planned analysis

Primary analysis:

Our primary analysis concerns the estimation of sensitivity and specificity of DeepClass in two classification tasks (see tasks' description in section 2.4). Those measures are proportions; we will report them along their 95% confidence intervals based on the Wilson method.

We also perform the McNemar hypothesis test to assess whether the estimated sensitivities differ significantly from the pathologists expected 75% sensitivity. The pathologist's sensitivity will be reassessed during this study by comparison to the "Gold standard" experts panel assessments.

Secondary analysis:

Dysplasia grade severity changes with polyp type, size, subject age, sex and comorbid behaviours or conditions (National Colorectal Cancer Roundtable Advanced Adenoma Working Group, 2018). Our secondary analysis aims to assess whether sensitivity and specificity significantly change between stratum based on those variables. For this assessment, we will stratify the polyp dysplasia classification based on the variable's classes and re-estimate sensitivity and specificity along their confidence intervals in each stratum.

Furthermore, and as discussed in section 2.2, we expect that our source population might overrepresent older adults, males and comorbid conditions, compared to their prevalence in the target population. We propose to assess the effect of this bias by also reporting weighted sensitivity and specificity estimates.

To illustrate with the comorbidities, each comorbid condition is assigned a weight that is equal to the ratio:

$$W_{c1} = \frac{Proportion\ of\ comorbid\ condition\ \#1\ in\ our\ sample}{Proportion\ of\ comorbid\ condition\ \#1\ in\ target\ population}$$

Based on the sensitivity estimates in the comorbidities' strata (assuming there are 3 of them), the weighted sensitivity is equivalent to:

$$Sensitivity_{weighted} = \frac{W_{c1}Sens._{c1} + W_{c2}Sens._{c2} + W_{c3}Sens._{c3}}{W_{c1} + W_{c2} + W_{c3}}$$

A similar weighting can be calculated for the specificity estimates as well as for the age groups and the sex groups.

3. Study limits

3.1 Risk of bias

We based our risk of bias assessments on QUADAS-2 (the QUADAS-2 Group, 2011), a standardized tool for quality assessments of diagnostic accuracy studies. This tool includes 4 domains: (1) patient selection, (2) index test, (3) reference standard, (4) flow and timing. In the following we answer the probing question from the tool's checklist.

- ### Selection bias

  Patient selection:

  - Was a consecutive or random sample of patients enrolled? Yes (a random sample from biobanks where all included subjects underwent a standardized inclusion process, specifically, having underwent a polypectomy during colonoscopy).
  - Was a case-control design avoided? Yes (avoids spectrum bias).
  - Did the study avoid inappropriate exclusions? Yes (limited exclusion criteria to conditions not covered by the current guidelines on polypectomy follow-up).

  Could the selection of patients have introduced bias? RISK: LOW

  Is there concern that the included patients do not match the review question? CONCERN: LOW

- ### Measurement biases

  Index test:

  - Were the index test results interpreted without knowledge of the results of the reference standard? Yes (the DeepClass classification task runs without knowledge of the reference class).
  - If a threshold was used, was it pre-specified? Yes (we followed current clinical guidelines for classification and we predefine them in the protocol before running the study. Furthermore, all DeepClass classifications are automated guaranteeing the consistency of the application of the index test).

  Could the conduct or interpretation of the index test have introduced bias? RISK: LOW

  Is there concern that the index test, its conduct, or interpretation differ from the review question? CONCERN: LOW

  Reference test:

  - Is the reference standard likely to correctly classify the target condition? Yes (despite its imperfection, the pathologists' classification is currently accepted as basis for treatment decisions. The quality of the reference improves with years of experience and targeted continuous training).
  - Were the reference standard results interpreted without knowledge of the results of the index test? Yes (the reference results were predefined before the index test application).

  Could the reference standard, its conduct, or its interpretation have introduced bias? RISK: LOW

  Is there concern that the target condition as defined by the reference standard does not match the review question? CONCERN: LOW

  Flow and timing:

  - Was there an appropriate interval between index test(s) and reference standard? Yes
  - Did all patients receive a reference standard? Yes
  - Did patients receive the same reference standard? Unclear (as a retrospective study, we cannot guarantee consistency in the guidelines application for dysplasia classification by all pathologists).
  - Were all patients included in the analysis? Unclear (the study aims to include all eligible polyp tissues; however, we account for a possible loss due to low quality imaging or algorithmic errors, we sized the sample to account for such losses).

Could the patient flow have introduced bias? RISK: Unclear

3.2 Internal and external validity

Our study design, sample selection, collected measurements and planned analysis are expected to guarantee the internal validity of the study.

However, this study's external validity could be limited. We opted for collecting data from Swiss university hospitals biobanks. We made this selection to guarantee high quality of conserved biological materials and their attached data and highly experienced pathologists' assessments. We believe that the results of our study would be generalizable to similar settings in other Swiss university hospitals biobanks or other community biobanks following similar quality practices.

4. Study conduct

4.1 Planning

Start date: 01.10.2025
Duration: 12 months

Preliminary work packages planning:

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Study management	X	X	X	X	X	X	X	X	X	X	X	X
Data collection	X	X	X									
Applying Gold Standard		X	X	X	X	X	X	X	X	X	X	
Running index test					X	X						
Data analysis										X	X	
Reporting results												X

4.2 Collaborators

- 1 colon histopathology expert (Study manager, Scientific expert)
- 1 biostatistician/data scientist
- 1 lab technician
- Independent panel of 5 histopathology experts

4.3 Materials

- Computational resources to run DeepClass: compute power equivalent to 930 TFLOPS (100 GFLOPS per image classification); disk storage of at least 13 TB (1.3 Gb average storage requirements for compressed WSI).
- Statistical analysis: STATA 19 BE (license)

4.4 Budget

Main costs sources:

- Making the digital slides from fixed tissue slides:
  We expect to digitalize 80% of the sample tissues while we expect to receive the rest already as digital slides. This incurs costs related to requesting and manipulating tissue slides, transportation, and digitalization. We account for 30 CHF per tissue slide[1] and a 20% overhead, for a total of **177'264 CHF**.
- Computational resources:
  930 TFLOPS leads to an average 12 CHF hourly rate. Assuming that DeepClass will run for 490 hours and an 20% overhead, gives **7'056 CHF**. We do not account for software licenses since we assume that they are already available at the hosting institution.
- Salaries:
  - We assume that the histopathology expert will be involved at a 20% rate on the project for 12 months for a total of 384 hours at a 200 CHF hourly rate, leading to **76'800 CHF**.
  - Experts panel composed of 5 experts, each assessing a total of 3'693 cases at 20 cases per day (3 hours per day, for 10 months) for a total of 555 hours each at a 200 CHF hourly rate, leading to **555'000 CHF**.
  - We assume that the biostatistician/data scientist will be involved at a 100% rate on the project for 4 months for a total of 640 hours at a 75 CHF hourly rate, leading to **48'000 CHF**.
  - We assume that the lab technician will be involved at a 100% rate on the project for 3 months for a total of 480 hours at a 45 CHF hourly rate, leading to **21'600 CHF**.

  Total salaries, including a 20% overhead, is equivalent to **841'680 CHF**.

We expect that the major costs of the study will be of approximately **1'026'000 CHF**.

4.5 Data (collection, databases, conservation)

As detailed previously, our main data requirements are the digital tissue slides and their corresponding auxiliary variables, namely the polyp size and type and the related subject age, sex and comorbidities. Our analysis can be run with full anonymization.

Such data will be collected from three biobanks: Centre Hospitalier Universitaire Vaudois - Biobanque de Pathologie; Hôpitaux Universitaires de Genève - Geneva Pathology Tissue Biobank; Biobank Pathology at University of Basel. We will provide our eligibility criteria and exclusion criteria and request in return a list of identifiers fitting our requirements in order to form a random selection sequence for the eligible tissues.

Whenever digital slides are available, we will only request those. However, multiple fixed and stained tissue slides might not already be digitalized. To guarantee enough polyp tissues for our sample size, we will request access to fixed tissues slides for the purpose of digitalization only or request digitalization by the biobank directly (if such service is provided). We limit our conservation needs to the duration of the generation of the images. The slides are restored to the biobank immediately after digitalization.

We expect that the auxiliary variables are stored in the biobank along the polyp tissue and can be collected simultaneously.

[1] Based on public prices listed:
https://www.vetpathology.uzh.ch/de/Andere-Services/Specific-technical-services.html

4.6 Feasibility

We provide below our preliminary risk analysis for this study:

			Risk assessment			
Risk	**Cause**	**Effect**	**Likelihood**	**Impact**	**Risk level**	**Measures**
Regulatory delays	Study approval delays	Study cannot be performed	Possible	Catastrophic	**High**	study can be run with full anonymization; consent will be verified and eventually requested for data use
Insufficient effective sample size	Not enough samples found eligible	Reduced power of the statistical tests and resulting conclusions	Possible	Significant	**Medium**	Including three large biobanks in the study; plan for other biobanks/databases access
Data collection delay	delays in eligible samples identification, in biobank processing time, in transportation and digitalization, gold standard assessment	delays in the study execution	Possible	Moderate	**Medium**	planning three months for data collection; clear communication of requirements to the biobanks and a pre-identification of valid cases from biobanks' administrative records; plan for reuse consent requests time; plan 10 months for experts panel assessments; distribute load among experts
Data quality issues	Digital slides might not be of enough high quality for an accurate classification by DeepClass	Reduces the effective sample size impacting statistical power	Unlikely	Significant	**Medium**	account for possible drops in the sample size estimation; clear define and communicate about data quality requirements and continuously assess them during data collection; request corrections whenever possible.
Unexpected costs	Not accounting for all costs sources or unexpected events	Study might be stopped	Possible	Moderate	**Medium**	study budget accounts for the main costs with a 20% inflation for each to account for possible errors in assessment; Budget is controlled and verified by the institute's accounting services as well as IT administrators.
Scope creep	Expansion of project requirements or protocol amendments during the study	Delays in the study execution	Unlikely	Low	**Low**	protocol clearly delimits the main objective of the study and all expected analysis
Infrastructure issues	electricity blackout, internet outage, datacenter outages	Study might be stopped	Very unlikely	Catastrophic	**Low**	follow recommendations and operations continuity planning of the hosting organization
Staff turnover	team members leaving the project or being assigned to other projects simultaneously	Delays in the study execution	Low	Moderate	**Low**	study planned to run for maximum of 12 months

Based on this risk analysis and the planned measures to accommodate for the identified risks, we believe that the study has a high chance to run with success, after ethical approval is guaranteed and funds are assigned.

5. Ethical and reglementary aspects

Our study requires an ethical approval since it manipulates human related data and potentially biological material.

Reglementary main aspects of concern:

- Consent for reuse: consent is tacitly given during the original endoscopy medical procedure, but consent for reuse might have not been provided (Swiss Society of Pathology, 2011). We will validate this consent and plan actions for requesting it if the eligible polyp tissues end up not enough to constitute our sample. To avoid delays due to consent requests, we will first explore the extension of the study to other biobanks and clinics databases where previous consent might have already been provided.
- AI algorithm: during this study the AI algorithm will be run locally (on the internal institution cloud infrastructure). No data will be transferred externally. The AI algorithm at the basis of the classification will not be retrained on this study data.

References

Abdeljawad, K., Vemulapalli, K., Kahi, C., Cummings, O., Snover, D., & Rex, D. (2015). Sessile serrated polyp prevalence determined by a colonoscopist with a high lesion detection rate and an experienced pathologist. *Gastrointestinal Endoscopy*, 517-524.

Adams, S., & Leveson, S. (2012). Clinical prediction rules. *BMJ*, d8312-d8312.

Bera, K., Schalper, K., & Rimm, D. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 703-715.

Bulur, A., & Çakır, A. (2021). An Evaluation of Colorectal Polyps by Size and Advanced Histological Features: In a Secondary Referral Centre. *Turkish Journal of Colorectal Disease*.

Crunch Foundation and The Eric and Wendy Schmidt Center. (2024). *Autoimmune disease machine learning challenge*. Retrieved from crunchdao.com: https://www.crunchdao.com/live/ewsc-challenge

Denis, B., Peters, C., Chapelain, C., Kleinclaus, I., & Fricker, A. (2009). Diagnostic accuracy of community pathologists in the interpretation of colorectal polyps. *European Journal of Gastroenterology & Hepatology*, 1153.

Dornblaser, D., Young, S., & Shaukat, A. (2024). Colon polyps: updates in classification and management. *Current Opinion in Gastroenterology*, 14-20.

European Commission, Karsa, L. v., Segnan, N., & Patnick, J. (2010). *European guidelines for quality assurance in colorectal cancer screening and diagnosis.* Publications Office. Retrieved from https://op.europa.eu/s/z6NI

IARC, T. I. (2025, June 9). *Global Cancer Observatory*. Retrieved from https://gco.iarc.fr/

Jin, X.-F., Ma, H.-Y., Shi, J.-W., & Cai, J.-T. (2024). Efficacy of artificial intelligence in reducing miss rates of GI adenomas, polyps, and sessile serrated lesions: a meta-analysis of randomized controlled trials. *Gastrointestinal Endoscopy*, 667-675.e1.

Jover, R., Bretthauer, M., & Dekker, E. (2021, 04 29). *EPoS the European Polyp Surveillance trial group - Protocol version 14.* Retrieved 2025, from eposstudy.org: https://www.eposstudy.org/uploads/1/2/1/4/121492320/epos_protocol_v14_april_29th_2021.pdf

Kumar, N., Gupta, R., & Gupta, S. (2020). Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions. *Journal of Digital Imaging*, 1034-1040.

McGenity, C., Clarke, E., & Jennings, C. (2024). Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. *npj Digital Medicine*, 1-19.

Mori, Y., East, J., Hassan, C., Halvorsen, N., Berzin, T., Byrne, M., . . . Rex, D. (2023). Benefits and challenges in implementation of artificial intelligence in colonoscopy: World Endoscopy Organization position statement. *Digestive Endoscopy: Official Journal of the Japan Gastroenterological Endoscopy Society*, 422-429.

National Colorectal Cancer Roundtable Advanced Adenoma Working Group. (2018). *Advanced Colorectal Polyp | GI brief.* Retrieved from National Colorectal Cancer Roundtable: https://nccrt.org/wp-content/uploads/2024/03/Advanced_Colorectal_Polyp_Brief_Final_Ver1.pdf

Nonchev, K., Dawo, S., Silina, K., Moch, H., Andani, S., Koelzer, V., & Rätsch, G. (2025). *DeepSpot: Leveraging Spatial Context for Enhanced Spatial Transcriptomics Prediction from H&E Images.* medRxiv.

Perlo, D., Tartaglione, E., Bertero, L., Cassoni, P., & Grangetto, M. (2022). Dysplasia Grading of Colorectal Polyps Through Convolutional Neural Network Analysis of Whole Slide Images. *Proceedings of 2021 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2021)* (pp. 325-334). Springer.

Pickhardt, P., Hain, K., & Kim, D. (2010). Low rates of cancer or high-grade dysplasia in colorectal polyps collected from computed tomography colonography screening. *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association*, 610-615.

Soleymanjahi, S., Huebner, J., & Elmansy, L. (2024). Artificial Intelligence–Assisted Colonoscopy for Polyp Detection: A Systematic Review and Meta-analysis. *Annals of Internal Medicine*, 1652-1663.

Sundjaja, J., Shrestha, R., & Krishan, K. (2023, July 17). *McNemar And Mann-Whitney U Tests*. Retrieved from StatPearls: https://www.ncbi.nlm.nih.gov/books/NBK560699/

Swiss Society of Pathology. (2011). *Assurance qualité en pathologie.* Retrieved from https://sgpath.ch/: https://sgpath.ch/customer/files/231/Recommandations-pour-la-qualite-SSPath-2011_FR.pdf

the QUADAS-2 Group. (2011). QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Annals of Internal Medicine*, 529-536.

Truninger, K., Lugli, A., & Koeberle, D. (2022, 05 24). *Suivi après polypectomie coloscopique et traitement du cancer colorectal.* Retrieved from sggssg.ch: https://sggssg.ch/fileadmin/user_upload/Empfehlungen/smf_09114_fr.pdf