

MEDICAL UNIVERSITY – PLEVEN
DEPARTMENT OF OBSTETRICS AND GYNECOLOGY

Dr. Georgi Danielov Prandzhev

**“Role of Artificial Intelligence in Contemporary Colposcopic
Diagnosis of Precancerous Cervical Lesions”**

DISSERTATION ABSTRACT

of a dissertation submitted for the award of the educational and
scientific degree “Doctor” (PhD)
Scientific specialty: “Obstetrics and Gynecology”

Dissertation supervisors:

Corresponding Member Prof. Slavcho Tomov Tomov, MD, DSc
Academician Grigor Angelov Gorchev, MD, DSc

Official reviewers:

Prof. Yavor Dimitrov Kornovski, MD, DSc
Prof. Stanislav Hristov Slavchev, MD, DSc

Pleven, 2026

The dissertation comprises 182 standard printed pages. The material is illustrated with 41 figures and 20 tables. The bibliography includes 115 references, of which 2 are in Cyrillic and 113 in Latin script.

The author is a full-time doctoral student at the Department of Obstetrics and Gynecology, Faculty of Medicine, Medical University – Pleven.

The dissertation was discussed and approved for public defense by an Expanded Departmental Council of the Department of Obstetrics and Gynecology, Faculty of Medicine, Medical University – Pleven, held on 30 July 2026.

Note: The numbering of the figures and tables does not correspond to that in the dissertation.

The public defense of the dissertation will take place on 7 October 2026 at 13:00 in the Louis Pasteur Hall, Medical College, Faculty of Pharmacy, Medical University – Pleven.

Scientific Jury:

Chair: Assoc. Prof. Dobrinka Hristova Gincheva, MD, PhD

Members:

Prof. Yavor Dimitrov Kornovski, MD, DSc

Prof. Stanislav Hristov Slavchev, MD, DSc

Prof. Elena Dimitrova Dimitrakova, MD, PhD

Assoc. Prof. Iliya Ilkov Karagyozev, MD, PhD

Assoc. Prof. Dobrinka Hristova Gincheva, MD, PhD

Alternate members:

Assoc. Prof. Chavdar Atanasov Tsvetkov, MD, PhD

Assoc. Prof. Pavel Petrov Dobrev, MD, PhD

The defense materials are published on the website of Medical University – Pleven (www.mu-pleven.bg) and are available to interested parties at the Research Department of the Faculty of Medicine, Medical University – Pleven.

CONTENTS

LIST OF ABBREVIATIONS

1. INTRODUCTION

- 1.1 Epidemiology
- 1.2 Artificial intelligence and limitations of colposcopy

2. AIM AND OBJECTIVES

- 2.1 Aim
- 2.2 Objectives

3. MATERIALS AND METHODS

- 4.1 General characteristics of the clinical cohort
- 4.2 Colposcopic equipment and examination methodology
- 4.3 Preparation and annotation of colposcopic images
- 4.4 Neural network architecture
- 4.5 Hardware and software environment
- 4.6 Statistical methods and metrics

4. ORIGINAL RESULTS AND DISCUSSION

4.1 OBJECTIVE 1

- 4.1.1 Model architecture and experimental design
- 4.1.2 **First group of multiclass segmentation experiments**
 - 4.1.2.1 Baseline experiment E000 (original 5-class segmentation)
 - 4.1.2.2 Experiment E212 (multiclass + oversampling of rare classes)
- 4.1.3 **Second group of binary (foreground) segmentation experiments**
 - 4.1.3.1 Experiment E202 (foreground + oversampling of rare classes)
- 4.1.4 **Third group of experiments with alternative target formulations**
- 4.1.5 **Discussion**
 - 4.1.5.1 Model evaluation and interpretation of the results
 - 4.1.5.2 Foreground versus multiclass approach
 - 4.1.5.3 JNEG signal and protocol specificity
 - 4.1.5.4 Role of augmentation and resampling

4.2 OBJECTIVE 2

- 4.2.1 Data format and annotation protocol
- 4.2.2 Data audit and class distribution
- 4.2.3 Dataset structure and experimental subsets
- 4.2.4 **Discussion**

4.3 OBJECTIVE 3

4.3.1 Data-related factors

- 4.3.1.1 Class imbalance and data volume
- 4.3.1.2 Variability between subsets and lack of patient-level grouping
- 4.3.1.3 Protocol-specific signal and lack of metadata
- 4.3.1.4 Effect of augmentation strategies

4.3.2 Annotation-related factors

4.3.3 Model-related factors

- 4.3.3.1 Target formulation – binary versus multiclass segmentation

4.3.3.2 Calibration and softmax confidence

4.3.4 Discussion

4.4 OBJECTIVE 4

4.4.1 Bridging segmentation and classification metrics

4.4.2 Comparison with published data

4.4.2.1 ROC-AUC profile comparison with systems reported in the literature

4.4.2.2 Segmentation analysis and comparison of Dice scores with systems reported in the literature

4.4.3 Discussion

4.4.3.1 Dataset size and class fragmentation

4.4.3.2 Annotation subjectivity and conceptual overlap between classes

4.4.3.3 Lack of metadata and protocol-dependent signal

4.5 OBJECTIVE 5

4.5.1 AI-assisted colposcopic assessment

4.5.2 Biopsy targeting

4.5.3 Training and standardization

4.5.4 Improving access in low-resource settings

4.5.5 Discussion

5. CONTRIBUTIONS

6. SCIENTIFIC PUBLICATIONS AND PARTICIPATION

LIST OF ABBREVIATIONS

AI – Artificial Intelligence
ALTS – ASC-US/LSIL Triage Study
ANN – Artificial Neural Network
ASCCP – American Society for Colposcopy and Cervical Pathology
AVE – Automated Visual Evaluation
AW – Acetowhite
AwRF – Acetowhite Response Function
Ca – Carcinoma
CAD – Computer-Aided Diagnosis
CAIADS – Colposcopic Artificial Intelligence Auxiliary Diagnostic System
CIN – Cervical Intraepithelial Neoplasia
CMF-CNN – Colposcopic Multimodal Fusion Convolutional Neural Network
CNN – Convolutional Neural Network
CRF – Conditional Random Fields
DL – Deep Learning
ECDC – European Centre for Disease Prevention and Control
ECIS – European Cancer Information System
EVA – Enhanced Visual Assessment
Faster R-CNN – Faster Region-based Convolutional Neural Network
GCN – Graph Convolutional Network
GLOBOCAN – global database of the Global Cancer Observatory
HPV – Human Papillomavirus
HSIL – High-grade Squamous Intraepithelial Lesion
IFCPC – International Federation for Cervical Pathology and Colposcopy
ILSVRC – ImageNet Large Scale Visual Recognition Challenge
LAST – Lower Anogenital Squamous Terminology
LLETZ – Large Loop Excision of the Transformation Zone
LSIL – Low-grade Squamous Intraepithelial Lesion
Mask R-CNN – Mask Region-based Convolutional Neural Network
ML – Machine Learning
MLP – Multilayer Perceptron
OECD – Organisation for Economic Co-operation and Development
R-CNN – Region-based Convolutional Neural Network
ResEnc-M – Residual Encoder U-Net Medium
ResNet – Residual Network
RGB – Red, Green, Blue
ROI – Region of Interest
SIL – Squamous Intraepithelial Lesion
SOM – Self-Organizing Map
TZ – Transformation Zone
YOLO – You Only Look Once

1. INTRODUCTION

“I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.”

Alan Turing, 1950 [1]

1.1 Epidemiology

Cervical cancer is the fourth most common malignant disease among women worldwide. In 2022, 662,301 new cases and 348,874 deaths were recorded. [2] If current incidence trends persist, predictive models estimate between 948,000 and 973,000 new cases and between 401,000 and 543,000 deaths annually by 2050. [3] In Bulgaria, the disease affects 15.4 per 100,000 women [4], placing the country in the highest incidence category within the European Union and second in mortality [5]. For 2024, the National Statistical Institute reported an average mortality rate of 9.6 per 100,000 women, with marked regional variation. [6] Human papillomavirus (HPV) comprises a family of more than 200 types, 40 of which affect the anogenital tract. Infection with high-risk HPV types is the principal etiological factor in the development of cervical cancer. [7]

In 2020, the World Health Organization adopted the Global Strategy to Accelerate the Elimination of Cervical Cancer. [8] The goal is to achieve and maintain an incidence below 4 new cases per 100,000 women per year by 2030 through three principal targets: 90% of girls fully vaccinated against HPV by age 15; 70% of women screened with a high-performance test by age 35 and again by age 45; and 90% of women with identified precancerous or invasive lesions receiving appropriate treatment.

HPV vaccination coverage in Bulgaria is exceptionally low. In 2019, only 7.2% of 13-year-old girls received a first dose, while only 11.9% of 12-year-old girls received a second dose. [9] During 2020–2024, vaccination remained alarmingly low, at approximately 1–2% of all girls aged 10–14 years. [10] In 2025, a new National Programme for the Primary Prevention of HPV-Related Cancers (2025–2030) was adopted, for the first time including boys in the vaccination programme and expanding the eligible age groups. [11]

With regard to secondary prevention, Bulgaria ranks among the countries with the lowest coverage. In 2019, only 45% of Bulgarian women aged 20–69 years reported having undergone a Pap test within the previous three years, compared with a European average of 59.9%, while HPV genotyping remains unreimbursed by the National Health Insurance Fund. [12]

Bulgaria continues to rank among the countries in the region with the most unfavorable cervical cancer statistics. Primary prevention through vaccination has not achieved sufficient coverage and stands in clear contrast to the WHO target. Only three European countries have achieved 90% vaccination coverage among girls by age 15, while progress remains uneven across the EU. [13] At the same time, secondary prevention is hindered by low population coverage and the absence of an organized national screening programme. [14] Colposcopy, on the other hand, remains highly dependent on subjective interpretation. Numerous studies demonstrate substantial interobserver variability and relatively low reproducibility even among expert colposcopists, with agreement in approximately 56–60% of cases. [15]

HPV-associated cervical cancer is almost entirely preventable through vaccination, yet achievement of the current targets will produce its full effect only after approximately 40 years. Reaching 90% vaccination coverage by 2030 would reduce mortality by only about 0.1%, whereas twice-lifetime screening and treatment could prevent 300,000 deaths (34.2%). By 2070, vaccination alone could reduce mortality by approximately 61.7%, while the combined strategy could achieve reductions of 88.9% with once-lifetime screening and 92.3% with twice-lifetime screening, corresponding to nearly 15 million women's lives saved. [16]

By the end of the century, the full impact of these targeted efforts should become apparent; until then, attention must also be directed toward populations unable to benefit from the potential of vaccination. The development of artificial intelligence (AI) in healthcare, particularly in medical image analysis, creates opportunities to seek solutions today while anticipating and supporting the WHO strategy. The implementation of intelligent analytical systems may improve real-time diagnostic precision and help compensate for the limitations of current preventive approaches. [17]

1.2 Artificial intelligence and limitations of colposcopy

Artificial intelligence is a broad scientific and engineering field concerned with developing systems capable of performing tasks that traditionally require intelligent behavior. In the classic textbook *Artificial Intelligence: A Modern Approach*, the authors emphasize that AI is not limited to understanding intelligence but also aims to build intelligent systems. [22] Central to their approach is the concept of the “rational agent” – a system that perceives its environment and selects actions intended to achieve the best possible or expected

outcome under the given conditions. In the broadest sense, AI may be viewed as an umbrella field encompassing diverse approaches to perception, learning, planning, and adaptation to new information.

In medical imaging, machine learning (ML) and deep learning (DL) are of particular importance. In classical machine-learning approaches, the researcher typically defines in advance the features that the model should analyze, such as color, texture, shape, intensity, or morphological parameters. These features are then used for classification, prediction, or estimation of a diagnostic outcome. Deep learning is a specialized subfield of machine learning and representation learning based on models with multiple sequential layers. Its principal advantage in image analysis is the ability to automatically extract hierarchical visual representations directly from the input data. In the early layers, the network identifies basic features such as edges, contours, color transitions, and textures, whereas deeper layers capture more complex morphological and spatial structures. This is particularly relevant to colposcopy, where diagnostic information is contained in combinations of subtle color, vascular, surface, and spatial characteristics.

Within this framework, computer vision is not regarded as a subset of deep learning, but rather as a broad application area of artificial intelligence focused on image processing and interpretation. The authors discuss it in the context of deep-learning applications, emphasizing that computer vision has traditionally been one of the most active areas of application. It should therefore be presented not as part of the AI–ML–DL methodological hierarchy itself, but as an applied field that uses machine-learning and deep-learning methods.

Qin et al. (2023) conducted a systematic review and meta-analysis of 15 studies including a total of 22,764 participants and using the IFCPC 2011 terminology. [24] At an LSIL+ threshold, colposcopy showed high sensitivity of 92% but low specificity of 51%. At an HSIL+ threshold, sensitivity decreased to 68%, while specificity increased to 93%.

A similar trade-off was reported in the large multicenter ESTAMPA study, which evaluated standardized colposcopy as a triage method in HPV-positive women. Valls et al. (2023) reported a sensitivity of 91.2% for detection of CIN3+ but a specificity of 50.1% for <CIN2. [25] This confirms that, even when standardized, colposcopy may be highly sensitive for high-grade lesions but remains limited in specificity, creating a risk of false-positive assessments and additional biopsies.

The methodological complexity of evaluating colposcopy was examined by Brown and Tidy (2019), who showed that published accuracy estimates depend on how a positive result is defined. [26] When colposcopic impression of CIN2+ is assessed, the weighted mean sensitivity is 68.5%, specificity 75.9%, and PPV 68.9%. When a positive result is instead defined as visible disease followed by biopsy, sensitivity rises to 95.7%, while specificity falls to 34.2% and PPV to

54.3%. This demonstrates that the diagnostic accuracy of colposcopy is not a fixed property of the method, but depends on the diagnostic threshold applied.

Li et al. (2021) analyzed agreement between colposcopic impression and histological diagnosis in 495 patients assessed using the IFCPC 2011 terminology. [28] The authors found complete agreement in only 46.9% of cases and concluded that colposcopic assessment more often underestimates than overestimates lesion severity, particularly in cases of HSIL and carcinoma.

The limitations of colposcopy extend beyond visual interpretation and also involve the biopsy strategy. In a systematic review and meta-analysis, Underwood et al. (2012) showed that colposcopically directed biopsy has a high pooled sensitivity for CIN2+ of 91.3%, but very low specificity of 24.6%. [29] Wentzensen et al. (2015) further demonstrated that a single directed biopsy is insufficient: sensitivity for detecting HSIL increases from 60.6% with one biopsy to 85.6% with two and 95.6% with three biopsies. [30]

This problem also has a spatial dimension. If the biopsy does not sample the area containing the most severe dysplasia, the lesion may be underestimated even when it has been visually recognized as abnormal. Thus, the diagnostic value of colposcopy depends not only on whether the clinician recognizes an abnormal finding, but also on whether it can be accurately localized on the cervical surface. This emphasizes a more specific clinical task: not merely determining whether an image is suspicious, but identifying where the area with the highest probability of a high-grade lesion is located.

One of the major limitations of conventional colposcopy is the substantial interobserver variability in the assessment of colposcopic findings. Contemporary quality-assurance studies confirm that this problem remains relevant. Origoni et al. (2023) evaluated the diagnostic performance of colposcopists with varying levels of experience in interpreting digital cases and found an overall sensitivity of 61.6% and specificity of 77.1%. [34] At a CIN2+ threshold, specificity increased to 87.7%, but colposcopic impressions were associated with approximately 20% underestimation of CIN2+ lesions. These findings support the need for systematic quality control and more objective tools to assist colposcopic interpretation.

In summary, the literature confirms that conventional colposcopy remains a fundamental yet limited visual method for assessing precancerous cervical lesions. Its diagnostic performance is not constant; rather, it depends on the selected interpretation threshold, the definition of a positive result, the biopsy strategy, visibility of the transformation zone, technical image quality, and the examiner's experience. Colposcopy is not only a recognition task, but also a task of spatial localization. Its performance depends both on the colposcopist's ability to identify an abnormal finding and on the ability to determine the most suspicious area and direct the biopsy toward it. The low agreement in the assessment of key colposcopic features indicates that subjectivity is a structural

characteristic of the method, rather than merely a consequence of individual experience. These factors limit the reproducibility of colposcopic assessment and complicate the development of universal diagnostic criteria applicable across different clinical settings.

In the context of colposcopy, computer vision is the principal applied area of artificial intelligence through which digital images of the cervix can be analyzed automatically. It enables the extraction and interpretation of visual features such as color, texture, vascular changes, acetowhite epithelium, mosaicism, punctuation, and iodine-negative areas. Depending on the task, AI systems in colposcopy may be directed toward three main types of analysis. The first is classification, in which the entire image is assigned to a diagnostic category. The second is detection, in which the model not only recognizes the presence of a pathological finding but also approximately localizes its position in the image using bounding boxes or regions of interest. The third, and most precise, approach is pixel-level segmentation, in which each image pixel is assigned to a specific class, allowing detailed delineation of individual colposcopic findings and their spatial relationships. In this sense, computer vision provides the methodological basis for moving from subjective visual assessment toward quantitative, reproducible, and anatomically localized analysis of colposcopic images.

2 AIM AND OBJECTIVES

2.1 AIM

By applying artificial-intelligence methods to the analysis of colposcopic images, to assess their clinical significance for detecting precancerous cervical lesions and to determine the role of AI in contemporary colposcopic diagnosis.

2.2 OBJECTIVES

1. To develop an artificial-intelligence-based algorithm using machine-learning and/or deep-learning methods for automated analysis of colposcopic images.
2. To create a standardized database of colposcopic images for integration into the developed model using a unified image-acquisition and annotation protocol consistent with medical standards for colposcopic nomenclature.
3. To identify and analyze the factors that influence the model's segmentation performance.
4. To determine the diagnostic value of the developed model through comparative analysis with established systems reported in the literature.
5. To analyze the clinical applicability and role of artificial intelligence in contemporary colposcopic diagnosis.

3. MATERIALS AND METHODS

3.1. General characteristics of the clinical cohort

For the purposes of the dissertation, an in-house database of colposcopic images was used, prospectively accumulated during the follow-up of 94 women aged 18–64 years who were examined at St. Marina University Hospital – Pleven between 2022 and 2025. The imaging data were accompanied by information on age, number of births, cytological status, HPV status and/or genotyping, transformation zone (TZ) type, presence of abnormal colposcopic findings, and histopathological result.

3.2. Colposcopic equipment and examination methodology

Colposcopic examinations were performed using a Leisegang Digital Colposcope equipped with adjustable LED illumination at an intensity of 45,000–52,000 lux and a color temperature of 5,700–6,000 K. The device provides up to 30× optical magnification and an image resolution of 1280 × 960 pixels.



Figure 2. Leisegang digital colposcope used in the present study.

Colposcopic findings were described using the terminology of the International Federation for Cervical Pathology and Colposcopy (IFCPC), according to the criteria for abnormal epithelial and vascular changes. (Fig. 3) (Bornstein et al., 2012)

 International Federation for Cervical Pathology and Colposcopy Internationale Federation für Zervixpathologie und Kolposkopie Fédération Internationale de Pathologie Cervicale et Colposcopie		
2011 IFCPC Nomenclature¹ Accepted in Rio World Congress, July 5, 2011 Nomenclature committee chairman: Jacob Bornstein MD		
2011 IFCPC colposcopic terminology of the cervix¹		
General assessment	- Adequate/inadequate for the reason ... (i.e.: cervix obscured by inflammation, bleeding, scar) - Squamo-columnar junction visibility: completely visible, partially visible, not visible - Transformation zone types 1,2,3	
Normal colposcopic findings	Original squamous epithelium: - Mature - Atrophic Columnar epithelium - Ectopy Metaplastic squamous epithelium - Nabothian cysts - Crypt (gland) openings Deciduous in pregnancy	
Abnormal colposcopic findings	General principles	Location of the lesion: inside or outside the T-zone, Location of the lesion by clock position Size of the lesion: Number of cervical quadrants the lesion covers, Size of the lesion in percentage of cervix,
	Grade 1 (Minor)	Thin aceto-white epithelium Irregular, geographic border Fine mosaic, Fine punctuation
	Grade 2 (Major)	Dense aceto-white epithelium, Rapid appearance of aceto-whitening Cuffed crypt (gland) openings Coarse mosaic, Coarse punctuation, Sharp border, Inner border sign, Ridge sign
	Non specific	Leukoplakia (leukotosis, hyperkeratosis), Erosion Lugol's staining (Schiller's test): stained/non-stained
	Suspicious for invasion	Atypical vessels Additional signs: Fragile vessels, irregular surface, Exophytic lesion, Necrosis, Ulceration (necrotic), tumor/gross neoplasm
	Miscellaneous finding	Congenital transformation zone, Condyloma, Polyp (ectocervical/endocervical) Inflammation, Stenosis, Congenital anomaly, Post treatment consequence, Endometriosis

Figure 3. Colposcopic terminology according to IFCPC 2011.

For annotation purposes, the data were labeled using abbreviated codes: WE – white epithelium, M – mosaicism, P – punctuation, L – leukoplakia, and JNEG – iodine-negative area, following the abbreviations proposed by B. Vasilev. (Fig. 4)

РИМСКА НОМЕНКЛАТУРА	АБРЕВИЯТУРИ ЗА „КОЛПОСКОПСКА КАРТА НА МАТОЧНАТА ШИЙКА“
Нормален (типичен, обичаен) колпоскопски образ	N
Многослоен плосък непроговяващ епител	N-Sq
Цилиндричен епител	N-Gl
Нормална зона на преход	N-TZ
Абнормен (атипичен, необичаен) колпоскопски образ	A
Бил епител с адхезионен с оксидна киселина *	A-WF **
• с плоска повърхност	A-WF-PJ **
• с микропапилярна или микропапилярна повърхност	A-WF-R **
Пунктирна зона (пунктуация) *	A-P **
Мозайка *	A-M **
Левкоплакия *	A-L **
Иод-негативен участък	A-I(-)
Атипични съдове	A-V **
СЪМНЕНИЕ ЗА ИНВАЗИВЕН РАК	C ?
НЕЗАДОВОЛИТЕЛНА КОЛПОСКОПИЯ	
Границата плосък/цилиндричен епител не се проследява изцяло	UNSAT
Тежко възпадение или тежка атрофия	
Не се визуализира маточна шийка	
РАЗНИ	
Микропапилярна повърхност, която не е бил епител	R
Екзофитни кондиломи	Condyl
Външалесия	I
Атрофия	Atf
Язва	U
Друго	(описание)

Figure 4. Colposcopic nomenclature according to Prof. Bozhil Vasilev.

3.3 Preparation and annotation of colposcopic images

Following the colposcopic examination, the acquired images were archived on a laptop computer and printed as part of the patient record, on which an expert colposcopist performed the initial manual delineation of abnormal areas. After digitization, the images were imported into Supervisely, a specialized platform for medical-image annotation and preparation of machine-learning datasets. Annotation was then repeated using a tool that allowed lesion boundaries to be delineated point by point at high magnification and pixel-level resolution. (Fig. 5)



Figure 5. View of the annotation process in Supervisely.

The total number of images was 355, each accompanied by a corresponding annotation file containing the coordinates of polygonally delineated findings. A total of 625 polygons were identified within the dataset. (Table 1) The polygons represent expert-labeled regions obtained through manual delineation of colposcopic findings on the images. Each polygon defines the spatial boundaries of a specific area of clinical colposcopic interest. All original images and

annotations were used without additional manual modification or alteration of the raw data.

Parameter	Result
Number of images	355
Number of annotation files	355
Number of polygons	625

Table 1. Key parameters of the data audit.

3.4 Neural network architecture

For the analysis of colposcopic images, the nnU-Net v2 deep-learning architecture with the Residual Encoder U-Net M (ResEnc-M) configuration was selected. (Fig. 6) The choice of model was based on several key advantages that make it particularly suitable for medical segmentation tasks. First, nnU-Net v2 is a self-configuring framework capable of automatically adapting the architecture, preprocessing steps, augmentation strategies, and training procedure to the characteristics of the specific dataset. This minimizes the need for manual optimization and provides high reproducibility of the results.

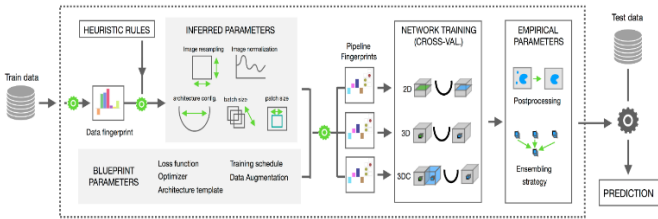


Figure 6. Schematic representation of the nnU-Net v2 architecture.

The architecture was developed specifically for biomedical image segmentation and has achieved state-of-the-art results in a number of international competitions using medical datasets. A major advantage of nnU-Net is its ability to adapt to datasets of limited size and substantial heterogeneity,

both characteristic features of colposcopic images. The Residual Encoder used in the present model improves feature extraction and facilitates the training of deeper neural networks. This is particularly important for identifying subtle abnormal findings. (Isensee et al., 2021)

3.4 Hardware and software environment

The experiments were conducted under Ubuntu 24.04.4 LTS on a server equipped with $4 \times$ NVIDIA H100 GPUs (80 GB VRAM each). A Python environment (Conda, PyTorch 2.6.0, CUDA 12.4) based on the 2D nnU-Net v2 Residual Encoder U-Net architecture was used. Training was performed across folds with patient-level data splitting to prevent data leakage. Additional processing and analysis were implemented using scientific Python libraries. Custom scripts were developed to evaluate the core metrics (Dice, IoU, ROC-AUC, confusion matrices, etc.) and to analyze softmax confidence.

3.5 Metrics and statistical evaluation

The quality of the segmentation model was evaluated according to the degree of overlap between the predicted pixel mask and the expert annotation. Conventional image-level accuracy is not a reliable summary measure because most of a colposcopic image consists of background or non-target tissue. A model may therefore appear to achieve high accuracy by predominantly predicting background while simultaneously missing clinically significant abnormal findings. The following statistical terminology, with both medical and technical relevance, was used in the present study:

Term	Meaning	Interpretation
Dice	Primary overlap metric for segmentation. Compares the region delineated by the expert with the region predicted by the model.	A value of 1.0 indicates perfect overlap, whereas 0.0 indicates no overlap.
IoU / Jaccard	Another overlap metric reflecting the spatial overlap between the expert annotation and the region predicted by the	A stricter metric than Dice, useful for model comparison, but less intuitive for readers without an ML background.

Term	Meaning	Interpretation
	model, calculated as the ratio of their intersection to their union.	
Precision	Of all regions labeled by the model as a lesion/class, how many actually correspond to the expert annotation?	A low precision value indicates many false-positive regions.
Recall (sensitivity)	Of all regions annotated by the expert, how many are detected by the model?	Low recall means that the model misses true pathological regions.
Macro Dice	Mean Dice value across all foreground classes, with each class weighted equally.	A particularly important metric in the present study because rare classes are not “hidden” within the average value of more frequent classes.
Foreground Dice	Dice value after merging all target classes into a single foreground region.	Useful for assessing whether the model recognizes the presence of an abnormal/annotated region, but does not evaluate correct subtype classification.
Pixel accuracy	Percentage of all pixels classified correctly.	Not used as a primary metric because images dominated by background can show high accuracy even when the lesion is missed.

Table 2. Statistical terminology.

For the purposes of the present study and the artificial-intelligence methods applied, the following specialized terminology was used:

Term	Meaning in the present study
Semantic segmentation	Pixel-level classification. The model delineates a region for each target class rather than providing a conclusion only at the image level.
Mask	An image in which each pixel contains an identifier for a specific class, e.g., background=0, M=1, WE=2.
Polygon annotation	An expert-delineated contour used to create the training mask. The quality of the polygon annotation directly affects model quality.
Fold	A validation subset used in cross-validation. Results from 5-fold analysis are more reliable than 2-fold screening results.
Grouped data splitting	Images with the same prefix in the filename are assigned to the same fold in order to limit data leakage between the training and validation sets.
Augmentation	Image variations used only during training, including changes in brightness, contrast, glare, cropping, and rotation. Their purpose is to improve model robustness without altering the validation images.
Oversampling of rare classes	More frequent presentation of rare classes during training. This helps the model “pay attention” to rare classes, but cannot replace the need for more high-quality annotations.
Softmax multiclass classification	The model must select only one class for each pixel; that is, each pixel can belong to only one class. This may be a limitation when individual findings are contained within a larger overlapping finding.

Table 3. Specialized technical terminology.

4. ORIGINAL RESULTS AND DISCUSSION

4.1 OBJECTIVE 1: To develop an artificial-intelligence-based algorithm using machine-learning and/or deep-learning methods for automated analysis of colposcopic images.

4.1.1 *Model architecture and experimental design*

For automated analysis of colposcopic images, a model based on the 2D nnU-Net v2 Residual Encoder U-Net Medium (ResEnc-M) deep-learning architecture was developed. This architecture is specialized for pixel-level segmentation; that is, each image pixel is classified as belonging to a specific colposcopic class, allowing precise delineation of abnormal cervical regions. The workflow operates directly on the original color images acquired by the colposcope, without the need for manual extraction of texture or color features. The task was implemented using the nnU-Net v2 software platform, which automatically determines optimal training settings, such as input resolution, image size, and preprocessing strategy, based on the specific characteristics of the working dataset. The choice of nnU-Net was based on its demonstrated effectiveness across numerous medical segmentation tasks, including limited datasets, owing to its built-in mechanisms for improving training samples and automatically adapting to the characteristics of the specific task.

Several techniques were applied during the experiments to improve training. The augmentations included artificial changes in brightness, color, contrast, and sharpness of the training images, thereby exposing the model to different acquisition conditions. The nnU-Net v2 platform automatically applies standard geometric augmentations such as rotation, scaling, and resizing, which improve model robustness to different positions and orientations of the findings. In some experiments, additional photometric processing was introduced, including simulation of varying illumination, glare, blur, JPEG compression artifacts, and edge darkening (vignetting). Rare classes were sampled more frequently during training (oversampling) to compensate for their small numbers and give the model greater opportunity to learn their appearance. In addition, a 768×768 -pixel crop around the finding was used in selected experiments, forcing the model to focus on small lesional regions.

A total of 13 experiments were conducted and can be divided into three groups according to the research question addressed:

The first group tested the model's ability to recognize all five colposcopic classes (multiclass segmentation), i.e., to directly classify each pixel as mosaicism, punctuation, white epithelium, leukoplakia, or an iodine-negative

area. This was the most ambitious and clinically challenging task because of the limited amount of data and the class imbalance. This group included the baseline experiment E000 and its improved variants E102, E211, E212, and E213.

The second group simplified the task by merging all colposcopic findings into a single class. The model thus focused solely on detecting an abnormal area without distinguishing between subtypes – binary (foreground) segmentation. Although this approach does not provide information on the type of colposcopic finding, it reduces task complexity and was therefore applied in experiments E101, E201, E202, and E203.

The third group investigated alternative ways of defining the target classes themselves, for example by removing iodine-negative cases from the dataset (E103), training only on such images (E104), merging visually similar features such as M and P into a common vascular class (E105), or changing the priority used when annotations overlapped (E106). These experiments were designed to determine the extent to which class definition influences model performance. Figure 7 visualizes the ranking of the experiments according to the mean Dice coefficient achieved by the algorithm, calculated separately for each class (Macro Dice).

For each experiment, three levels of overlap between the model-predicted region and the ground-truth annotation are visualized: low (Dice = 0.0–0.2), moderate (Dice = 0.3–0.6), and high (Dice = 0.7–0.9). The error map displays four color-coded categories: green for correctly predicted foreground (true positive), red for a missed lesion (false negative), orange for a false-positive region, and purple for an incorrect class (correct foreground region but incorrectly classified subtype). (Fig. 8)

Визуализация на карти на грешките (Error map)

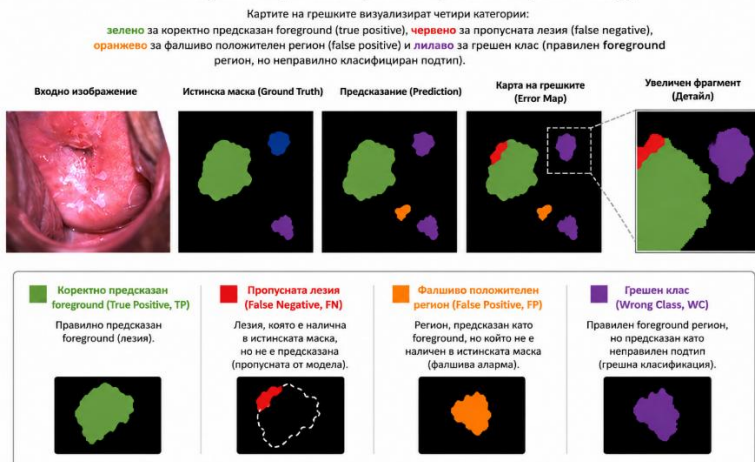


Figure 8. Error map visualizing the overlap categories.

4.1.2 First group of multiclass segmentation experiments (Table 4)

ID	Formulation	Change relative to E000	Fold	Macro Dice	IoU
E000	Original 5-class (baseline)	Baseline experiment	5	0.154	0.120
E102	5-class without empty annotations	E000 + empty annotations removed	2	0.242	0.193
E211	Multiclass + augmentations	E102 + photometric changes/glare/image blur	2	0.245	0.196

E212	Multiclass + oversampling of rare classes	E102 + increased representation of P/L	5	0.256	0.209
E213	Multiclass + cropping	E102 + 768x768 crops around P/L	2	0.227	0.180

Table 4. First group of experiments.

4.1.2.1 Baseline experiment E000 (original 5-class segmentation)

The first experiment conducted (E000) served as the baseline against which the effect of each subsequent change in formulation, data, or training strategy was measured. It used the most basic possible configuration: the original 5-class softmax target formulation (M, P, WE, L, JNEG + background) with all 355 images, including empty annotation files, without oversampling, augmentation, or additional data processing. Importantly, E000 was a 5-fold cross-validation experiment classified as a confirmed result, with macro Dice = 0.154 ± 0.019 and macro IoU = 0.120. Qualitative analysis showed a broad range of cases, from no overlap at all to cases with a Dice value of 0.675. (Figs. 9–11)

Example 1: No overlap. Foreground Dice = 0.000.

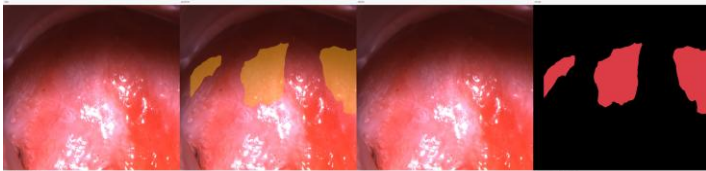


Figure 9. Qualitative visualization of fold-0 validation (experiment E000).

Example 2: Low overlap. Foreground Dice = 0.129.

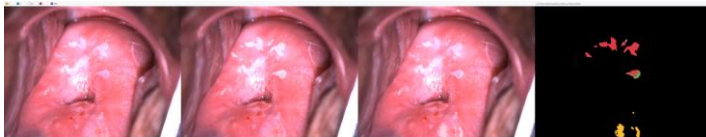


Figure 10. Qualitative visualization of fold-0 validation (experiment E000).

Example 3: Moderate-to-high overlap. Foreground Dice = 0.675.

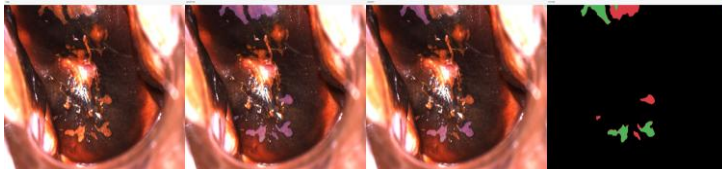


Figure 11. Qualitative visualization of fold-0 validation (experiment E000).

These results demonstrate that even the baseline model is able to recognize some abnormal regions, albeit inconsistently. The wide performance range reflects the heterogeneity of the data and highlights the need for targeted improvements, the effects of which are quantified relative to the E000 baseline.

Starting from this baseline, subsequent experiments tested different improvement strategies: removal of empty annotations (E102), addition of augmentations (E211), oversampling of the rare P and L classes (E212), and cropping smaller regions around them (E213). The best multiclass model in this group was E212, with macro Dice = 0.256 ± 0.039 and macro IoU = 0.209; oversampling of P/L increased Dice from 0.242 (E102) to 0.256.

4.1.2.2 Experiment E212 (multiclass + oversampling of rare classes) (Figs. 12–14)

The following examples illustrate the three levels of overlap in E212, ranging from a challenging case with low Dice = 0.124, through a typical case with moderate Dice = 0.542, to a high-quality prediction with Dice = 0.862.

Example 4: Low overlap. Foreground Dice = 0.124.

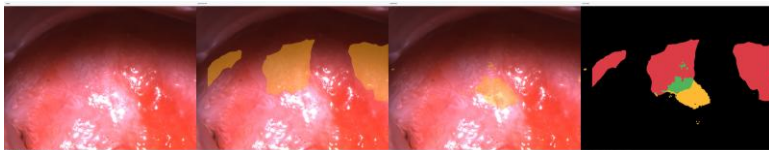


Figure 12. Qualitative visualization of fold-0 validation (experiment E212).

Example 5: Moderate overlap. Foreground Dice = 0.542.

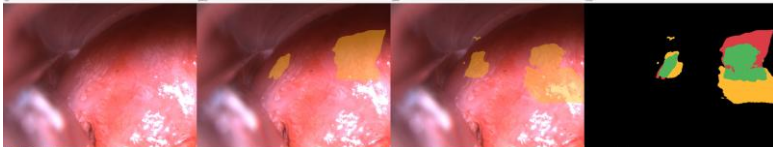


Figure 13. Qualitative visualization of fold-0 validation (experiment E212).

Example 6: High overlap. Foreground Dice = 0.862.

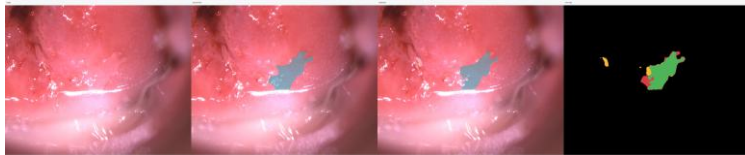


Figure 14. Qualitative visualization of fold-0 validation (experiment E212).

4.1.3 Second group of binary (foreground) segmentation experiments (Table 5)

The baseline experiment E101 achieved a foreground Dice of 0.273. The applied augmentations (E201) and crops around the findings (E203) did not improve performance relative to the baseline.

ID	Formulation	Change relative to E101	Folds	Foreground Dice
E101	Binary foreground	Baseline experiment	2	0.273
E201	Foreground + augmentations	E101 + photometric changes/glare/image blur	2	0.259
E202	Foreground + oversampling of rare classes	E101 + increased representation of P/L	5	0.273

E203	Foreground + cropping	E101 + 768x768 crops around P/L	2	0.263
-------------	-----------------------	------------------------------------	---	-------

Table 5. Second group of experiments.

With additional sampling of the rare P and L findings (E202), the model retained the same Dice of 0.273, now confirmed after 5-fold validation, making it the most reliable foreground result (Figs. 15–17):

4.1.3.1 **Experiment E202 (foreground + oversampling of rare classes)**

Example 7: Low overlap. Foreground Dice = 0.001.



Figure 15. Qualitative visualization of fold-0 validation (experiment E202).

A difficult, challenging image with minimal overlap, in which the model fails to identify the abnormal region. This is characteristic of images containing small, subtle lesions.

Example 8: Moderate overlap. Foreground Dice = 0.430.

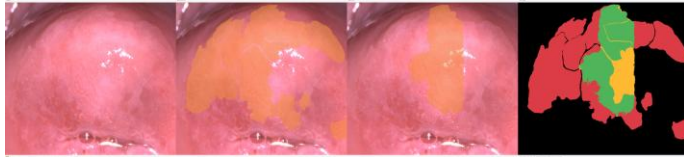


Figure 16. Qualitative visualization of fold-0 validation (experiment E202).

A typical image with moderate overlap, in which the model identifies the main abnormal region but delineates its boundaries imprecisely. This level of performance is representative of an average case.

Example 9: High overlap. Foreground Dice = 0.732.

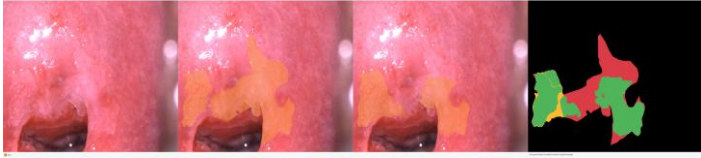


Figure 17. Qualitative visualization of fold-0 validation (experiment E202).

A large, clearly distinguishable lesion with well-defined boundaries, for which the model achieves very good overlap with the expert annotation. This case illustrates the upper range of the model’s current capabilities.

4.1.4 Third group of experiments with alternative target formulations (Table 6)

Removing the iodine-negative cases (E103, M/P/WE/L only) reduced macro Dice to 0.186, whereas a model trained exclusively on JNEG (E104) achieved a Dice of 0.529. Merging M and P into a common vascular class (E105) yielded a Dice of 0.230. Changing the overlap priority from $L > M > P$ to $M > P > L$ (E106) resulted in the lowest performance among all 13 experiments (macro Dice 0.128), attributable solely to a different technical configuration with no direct clinical analogue.

ID	Formulation	Change relative to baseline	Folds	Macro Dice	SD	IoU
E103	4-class without JNEG	E000 without JNEG cases	2	0.186	0.007	0.143
E104	Binary JNEG only	Single-class JNEG-only model	2	0.529	0.137	0.433
E105	Merging of similar classes	M/P → common vascular class, WE/L → common white-change class	2	0.230	0.020	0.181
E106	Change in class priority (M/P priority)	Priority $M > P > L > WE > JNEG$	2	0.128	0.014	0.098

Table 6. Third group of experiments.

The JNEG-only experiment E104 achieved a Macro Dice of 0.529, but it is not directly comparable because it uses a single-class target definition and

includes only a subset of the data. Nevertheless, the visualization demonstrates the highest single-case overlap in the entire study, reaching Dice = 0.913. (Fig. 18)

Example 10: Very high overlap. Foreground Dice = 0.913.

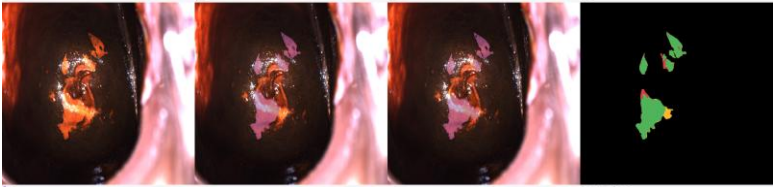


Figure 18. Qualitative visualization of fold-0 validation (experiment E104).

In summary, the best confirmed binary model was E202, while the best confirmed multiclass model was E212, achieving macro Dice values in the range of 0.25–0.28 and foreground IoU of 0.20–0.21 (Fig. 19).

Confirmed Fold-Level Macro Dice

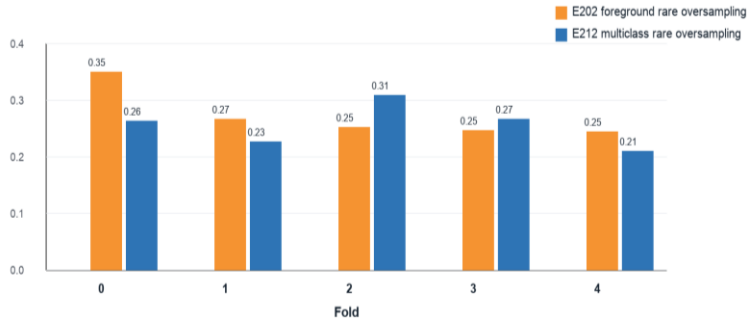


Figure 19. Performance across individual subsets for each of the five validation folds of the two experiments confirmed at stage 3 (E202 and E212).

Detailed class-wise analysis of model E212 revealed substantial differences in recognizability among the individual colposcopic classes (Fig. 20).

E212 Confirmed Per-Class Dice

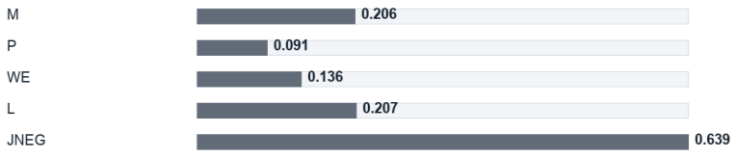


Figure 20. Confirmed class-wise performance of experiment E212. JNEG dominates the multiclass signal, whereas P, WE, M, and L remain weakly represented.

JNEG dominates the multiclass signal with a Dice of 0.639, nearly three times higher than the next-best class. P (punctuation) shows the poorest performance (Dice 0.091), while M, WE, and L fall within a similar range (0.136–0.207), suggesting that the model has difficulty distinguishing them from one another.

4.1.5 Discussion

4.1.5.1 Model evaluation and interpretation of the results

The results of the 13 experiments demonstrate that the 2D nnU-Net v2 ResEnc-M architecture can be trained to recognize and localize abnormal colposcopic findings, but the degree of overlap between predictions and annotations remains below the level required for stand-alone clinical use. The observed performance reflects a substantially earlier stage of model development than that of systems published in the literature, which predominantly address classification tasks and are trained on larger datasets. Importantly, most of those systems perform whole-image classification based on histological outcome (e.g., HSIL versus LSIL), whereas the present task is pixel-level segmentation of individual colposcopic findings. This is a fundamentally different and more difficult task and is not directly comparable with published image-level classification results. A detailed comparative analysis is presented under Objective 4.

4.1.5.2 *Foreground versus multiclass approach*

The initial aim of the experiments was to develop a model capable of classifying individual colposcopic findings through pixel-level segmentation. The five-class model (E212) achieved a macro Dice of 0.256, confirming that the architecture can capture signal from individual subtypes, although accuracy remains below a clinically acceptable level, primarily because of the limited dataset size and class imbalance. In parallel, the binary model (E202), which merges all pathological findings into a single class, achieved a higher Dice of 0.273. The numerical difference is only 0.017, but the two tasks differ substantially in complexity. The binary approach must only classify each pixel as “abnormal” or “normal,” whereas the multiclass approach must additionally determine which of the five colposcopic findings is represented. The performance difference between the two experiments supports a two-stage modeling strategy: first, detection of an abnormal region; second, classification of the colposcopic finding. This approach is consistent with the clinical reasoning of the colposcopist, who first determines whether an abnormality is present and where it is located before characterizing its type. An additional advantage of the two-stage approach is the possibility of using different confidence thresholds at each step. The binary model may operate at a lower detection threshold and mark an area as abnormal even when the signal is weaker, thereby favoring high sensitivity and reducing the risk of missed lesions. The multiclass stage, by contrast, requires a higher degree of confidence to avoid incorrect subtype assignment, favoring greater specificity. Against the background of these limitations, the observed performance is encouraging in terms of feasibility. The binary-model results indicate that the algorithm captured a meaningful segmentation signal despite the limited data volume, class imbalance, and absence of protocol metadata. High Dice values in individual cases (0.732–0.913) demonstrate that the architecture is capable of precise segmentation when the input data have sufficiently high quality and contrast. This is further supported by qualitative analysis of individual experiments, showing high overlap even for non-JNEG classes when images provide sufficient contrast and clearly defined boundaries. The four-class experiment E103, in which JNEG cases were excluded from training, included an example with Dice = 0.813 (Fig. 21). Similarly high values were observed in the five-class experiment without empty annotations (E102), which reached 0.811 (Fig. 22), and in the augmented multiclass experiment E211, which reached Dice = 0.822 (Fig. 23). These examples show that the architecture can segment M, P, WE, and L under favorable conditions, but performance remains highly dependent on the quality of the individual image. This selectivity also explains the low mean value: the model performs well in easier cases but fails in more challenging ones.

Example 11: High overlap. Foreground Dice = 0.813.

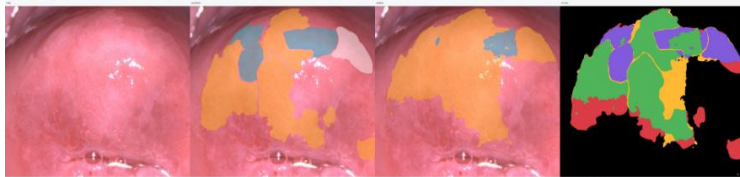


Figure 21. Qualitative visualization of fold-0 validation (experiment E103).

Example 12: High overlap. Foreground Dice = 0.811.



Figure 22. Qualitative visualization of fold-0 validation (experiment E102).

Example 13: High overlap. Foreground Dice = 0.822.

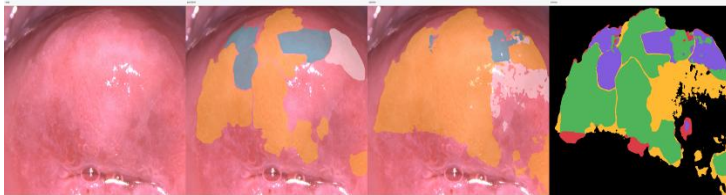


Figure 23. Qualitative visualization of fold-0 validation (experiment E211).

4.1.5.3 JNEG signal and protocol specificity

JNEG stands out as the most readily recognizable class, with a Dice of 0.639, nearly three times higher than the other classes. This result, however, must be interpreted cautiously. Iodine-negative areas are assessed only after application of Lugol's iodine, which creates strong contrast between the light-

yellow lesional region and the dark-brown normal epithelium. In contrast, acetowhite classes appear as subtle textural and color variations against a pale background, making them considerably more difficult to recognize. Clinically, this means that a model trained on mixed image types (acetic acid and Lugol's iodine) without protocol metadata may create a misleading impression of overall performance because the high JNEG Dice dominates the mean. In real-world medical use, each image must therefore be labeled according to the acquisition protocol under which it was obtained.

4.1.5.4 *Role of augmentation and resampling*

Experiments involving augmentation and altered sampling frequency of rare classes provided important practical guidance for future training, with effects differing according to the target formulation. To evaluate this, two baseline models were compared: the binary E101 model, which distinguishes only abnormal from normal tissue, and the multiclass E102 model, which recognizes the individual colposcopic classes M, P, WE, L, and JNEG. Three strategies were applied to each: photometric augmentations, increased representation of the rare P and L classes during training, and cropping of 768×768 -pixel regions around M/P findings.

In the multiclass model (E212), increased representation of punctuation (P) and leukoplakia (L) cases during training increased macro Dice from 0.242 (E102) to 0.256. This improvement can be explained by the greater complexity of the multiclass task. When the model must simultaneously localize abnormal regions and classify them into five separate subtypes, additional exposure to the rare P and L classes allows it to learn their specific characteristics, which would otherwise be overshadowed by the more common classes.

In the binary model (E202), where the task is only to distinguish abnormal tissue without subtype classification, the same strategy produced no change in performance (Dice = 0.273). Because all findings are merged into a single class, adding more P and L examples does not provide fundamentally new abnormal information to the model, as these cases are already represented in the baseline version (E101).

Various photometric augmentations were also applied, including simulated variations in illumination, glare, blur, and compression artifacts. In both the multiclass and binary models, performance decreased or changed minimally: E201 fell from 0.273 to 0.259, whereas E211 showed only a marginal increase in Dice from 0.242 to 0.245. This finding contrasts with expectations from the broader deep-learning literature but can be explained by the specific nature of colposcopic diagnosis. Tissue color carries clinically meaningful information: white epithelium, mosaicism, and punctuation each have characteristic color features that form part of the colposcopist's diagnostic assessment. Photometric

augmentation may distort or erase these characteristics and thereby confuse the model. Color therefore has semantic value, and standard augmentation strategies should not be applied automatically without consideration of the specific diagnostic modality.

Cropping 768×768 -pixel regions around the findings likewise failed to improve performance. In the multiclass model (E213), Dice decreased to 0.227 compared with 0.242 in E102. This result suggests that isolating an abnormal area together with only a small region of surrounding cervical surface deprives the model of important spatial context. Unlike tasks such as tumor detection in solid organs, where the lesion is relatively self-contained, colposcopic findings exist within an anatomical framework that carries clinical context. The size of the acetowhite area, its position relative to the external cervical os, and its relationship to the transformation zone are diagnostically meaningful, and cropping removes these landmarks.

In conclusion, the experiments demonstrate that 2D nnU-Net v2 can be trained to recognize colposcopic findings at the pixel level. The results indicate that model performance is determined to a greater extent by task formulation and data quality than by architectural capacity. Although the model achieves high overlap in selected favorable cases, mean performance remains below a clinically acceptable level because of limited data volume, class imbalance, and the absence of protocol metadata. Augmentation strategies transferred from general deep-learning practice proved unsuitable for the colposcopic modality, underscoring the need for a domain-specific approach. In this respect, Objective 1 establishes a methodological foundation and identifies the specific issues addressed by the subsequent objectives: data structure, the influence of individual factors, and the path toward clinical applicability.

4.2 OBJECTIVE 2: To create a standardized database of colposcopic images for integration into the developed model using a unified image-acquisition and annotation protocol consistent with medical standards for colposcopic nomenclature.

4.2.1 Data format and annotation protocol

We created a standardized database of colposcopic images processed in accordance with established standards of colposcopic nomenclature for integration into the developed model. Our in-house dataset comprised 355 images, and the findings were organized into classes named according to the colposcopic chart of Prof. Bozhil Vasilev and corresponding to the 2011 terminology of the International Federation for Cervical Pathology and Colposcopy (IFCPC). (Figs. 24–26) Minor simplifications were introduced in the abbreviations by omitting the prefix “A-” before abnormal findings and labeling iodine-negative areas as JNEG to facilitate model processing. Although leukoplakia is a nonspecific lesion, it was included as a training class because of its color similarity to white epithelium. Table 7 presents the class legend for the colposcopic annotations used in the digital repository. (Table 7)

Code	Class	Clinical term	Practical annotation description
0	background	Background / normal tissue	Non-target region
1	M	Mosaicism	Paving-stone-like mosaic vascular pattern
2	P	Punctuation	Punctate vascular structures on a white background
3	WE	Acetowhite epithelium	Acetowhite staining of the epithelium
4	L	Leukoplakia	Rough white plaque; visually similar to WE
5	JNEG	Iodine-negative area	Light yellow/orange area against brown-stained normal epithelium

Table 7. Class legend for the colposcopic findings.



Figures 24–26. Colposcopic appearance of the WE, M, and P classes.

The annotation process was carried out in Supervisely, a comprehensive web-based platform for artificial intelligence and computer vision. It provides a unified environment for the entire AI project life cycle, from data management and labeling to neural-network training, evaluation, and deployment in real-world workflows. The platform supports polygonal segmentation with professional tools for precise delineation of specified boundaries. Supervisely enables the definition of custom classes, import of images in multiple formats, and export of annotations in various industry-standard formats. The images were uploaded to the project as 1280 × 960-pixel RGB files, organized in structured individual folders according to their source. (Fig. 27)

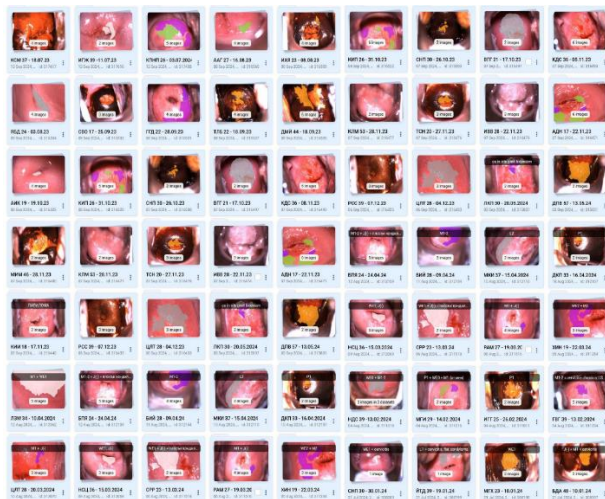


Figure 27. Dataset structure – view in Supervisely.

Each colposcopic finding was annotated by manually placing points along the lesion contour. The operator sequentially marked the boundary of the finding, with each click adding a new point and the lines between them forming a polygon. (Fig. 28) The platform is specifically optimized to support very high resolutions and highly detailed visualization. It allows zooming to the level of individual pixels, enabling the image to be enlarged until individual pixels are clearly visible as large squares on the screen.

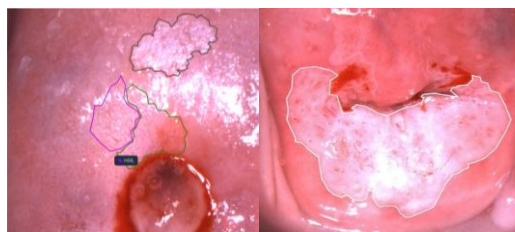


Figure 28. Annotated colposcopic findings – close-up view.

The five colposcopic findings M (mosaicism), P (punctuation), WE (white epithelium), L (leukoplakia), and JNEG (iodine-negative area) were created as separate object classes, each assigned a unique color for visual identification during annotation. (Fig. 29)

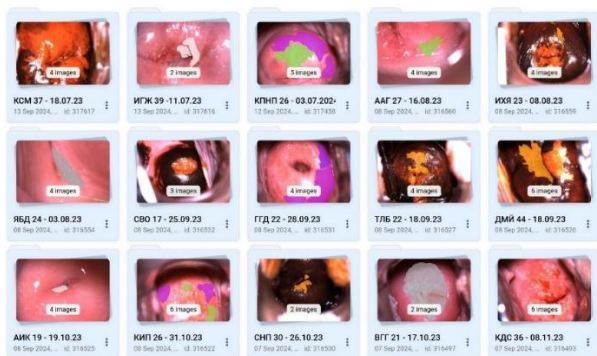


Figure 29. Colposcopic classes annotated with distinctive colors.

Empty annotation files correspond to images without pathological findings and are exported as empty text files. They were treated as true-negative cases and retained as separate instances in the platform without attached polygons. Overlapping polygons, such as M or P findings located within the boundaries of WE, were stored as separate geometries on the same image. Their segmentation was resolved at a later stage using a configurable priority hierarchy: L > M > P > WE > JNEG > background.

After annotation was completed, the project was exported from Supervisely in its native Supervisely JSON format, preserving the complete class, metadata, and geometry structure. Each image was accompanied by a separate text file containing the coordinates of the outlines of all colposcopic findings present in that image.

4.2.2 Data audit and class distribution

Overall, the database contains 625 colposcopic findings represented as polygons across 355 images with a resolution of 1280×960 pixels. Each image carries a unique code identifying the specific clinical examination from which it originated. For example, two frames coded “IMG_001_1” and “IMG_001_2” were acquired during the same colposcopic examination of the same patient, but from a different angle, at a different magnification, or at a different time point after solution application. The 355 images form 123 such groups, 108 of which (87.8%) contain two or more images from the same examination, while the remaining 15 consist of a single frame. This grouping has direct implications for the data-splitting strategy. If two images from the same examination are placed in both the training and validation sets, the model would be evaluated on almost identical content. This constitutes data leakage and leads to an artificial inflation of apparent diagnostic performance. For this reason, the data were split at the level of the clinical examination. Thus, all images from a given colposcopic examination remain together in either the training or validation set. After implementation of this strategy, verification confirmed zero leakage: no examination appeared in both sets.

Of the 355 images, 81 have empty annotation files and contain no colposcopic findings. Such examples are essential because they serve as negative controls, teaching the model to recognize the absence of abnormal changes. By number of polygons, class M (mosaicism) is the most frequent, with 201 polygons, whereas L (leukoplakia) is the rarest, with only 62 polygons. (Table 9)

Class	Polygons	Images with finding
M	201	132
WE	142	88
P	117	71
L	62	26
JNEG	101	61

Table 9. Distribution of polygons across images.

Class imbalance is particularly pronounced at the pixel level, where the largest class, M, occupies approximately 16.8 million pixels, nearly five times ($4.69\times$) the area of the smallest class, L. (Fig. 30) Iodine-negative areas rank third in total pixel area but have the largest mean polygon area of all classes – 110,000 pixels per polygon. In other words, JNEG regions are large, compact, and well-delineated areas with a clear boundary against the brown normal epithelium. Combined with the high contrast produced by Lugol’s iodine staining, this makes iodine-negative areas objectively easier to recognize not merely because of the number of examples, but because of the intrinsic nature of the anatomical signal.

Conversely, P has the smallest mean area: punctate vascular formations are small and dispersed, making them difficult to segment regardless of their frequency. (Table 10)

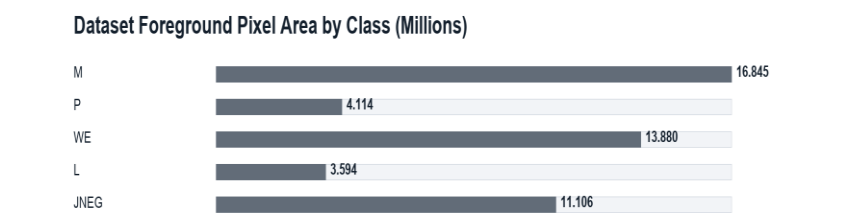


Figure 30. Distribution of foreground pixel area by class, in millions of pixels.

Class	Polygons	Pixels (foreground)	Imbalance relative to largest class
M	201	16 844 872	1.00x
P	117	4 113 934	4.09x
WE	142	13 879 830	1.21x
L	62	3 593 527	4.69x
JNEG	101	11 106 031	1.52x

Table 10. Distribution of foreground pixel area by class and imbalance ratio. The rarest classes by pixel area are L (leukoplakia) and P (punctuation).

4.2.3 Dataset structure and experimental subsets

Of the 355 colposcopic images, 22.8% are accompanied by empty annotation files. After excluding these, 273 images with at least one annotation remain (E102), of which 212 contain only the M, WE, P, and/or L classes (E103). Under conditions following application of Lugol’s iodine, 61 JNEG-positive cases are available (E104). The experiments conducted under Objective 1 use different filtered subsets, revealing the structural heterogeneity of the data. (Table 11)

ID	Subsets	Total	Note
E000/E101	All 355 images	355	Includes 81 images without annotations
E102	Without empty annotations	273	Only images with at least one finding
E103	Only images after acetic acid application	212	JNEG and empty cases excluded

E104	Only images after Lugol's iodine application	61	JNEG images
------	--	----	-------------

Table 11. Data subsets according to the filtering criteria used in the experiments.

This subset filtering was used as a surrogate approach in the absence of protocol metadata. Experiment E103 simulates a set of images acquired after application of acetic acid, whereas E104 simulates a set acquired after application of Lugol's iodine. In both cases, the separation is approximate because information on the actual diagnostic protocol used for each individual image is unavailable.

In an attempt to compensate for class imbalance without relying on collection of new clinical data, we applied three augmentation strategies: photometric augmentation of color characteristics, oversampling of rare findings, and cropping around P and L annotations. The effect of these strategies is summarized in Table 12, which presents the complete list of 13 experiments, the types of augmentation applied, and the resulting changes in the dataset.

ID	Baseline count	After augmentation	Total	Note
E000	355	0	355	No augmentation, 5-class model
E101	355	0	355	No augmentation, binary model
E102	273	0	273	Without empty annotations
E103	212	0	212	Without empty and JNEG annotations
E104	61	0	61	JNEG annotations only
E105	355	0	355	M/P – common vascular class WE/L – common white class
E106	355	0	355	Changed overlap priority
E201	355	710	1065	Two photometric copies of each image

ID	Baseline count	After augmentation	Total	Note
E202	355	629	984	P/L cases repeated 5 times; all others once
E203	355	267	622	A 768×768 px region cropped around each P/L annotation
E211	273	546	819	Two photometric copies of each image relative to E102
E212	273	629	902	Increased representation of P/L relative to E102
E213	273	267	540	A 768×768 px region cropped around each P/L annotation relative to E102

Table 12. Augmentation procedures applied to the dataset – qualitative and quantitative description.

The difference in case numbers between E202 and E212 reflects their different filtering criteria. E202 uses the complete dataset of 355 images, with augmentation increasing the total to 984 cases. E212 excludes empty masks and uses 273 original images, reaching a total of 902 cases after augmentation.

Oversampling of rare classes (E202, E212) increased the frequency of P- and L-containing cases to 629 images by presenting them five times more often than the remaining cases, which were shown once. This strategy was intended to compensate for class imbalance without altering the visual content of the images. Cropping around the same rare classes at a resolution of 768 × 768 pixels (E203, E213) generated approximately 267–273 additional training copies.

The data split for cross-validation was performed at group level, with images indicating origin from the same colposcopic examination retained within the same fold. This strategy reduces the risk of data leakage between the training and validation sets, but it is not equivalent to true patient-level splitting because a unique identifier for each examined woman is unavailable.

In addition, for the final confirmed models E202 and E212 (5-fold experiments), the distribution of training and validation cases is shown for each individual fold, with the ratio ranging from approximately 90:10 to 94:6. The validation proportion ranges from 5.8% to 9.7%, below the standard 10–20%, reflecting the limited data volume. (Table 13) Because of group-level splitting,

the model has between 750 and 827 training images versus 55–81 validation images for E202, and 688–764 versus 46–61, respectively, for E212.

Experiment	Fold	Training cases	Validation cases	Training share	Validation share
E202	0	750	81	90.3%	9.7%
E202	1	777	79	90.8%	9.2%
E202	2	763	74	91.2%	8.8%
E202	3	827	66	92.6%	7.4%
E202	4	819	55	93.7%	6.3%
E212	0	688	61	91.9%	8.1%
E212	1	714	60	92.2%	7.8%
E212	2	696	59	92.2%	7.8%
E212	3	764	47	94.2%	5.8%
E212	4	746	46	94.2%	5.8%

Table 13. *Distribution of training and validation cases for E202 and E212.*

5.2.4 Discussion

The created dataset covers all major colposcopic classes according to the IFCPC 2011 classification, making it practically suitable for training a semantic-segmentation model. The class legend corresponds to established colposcopic nomenclature, while the polygonal annotation format permits precise delineation of regions representing different findings without loss of spatial resolution. Compared with datasets described in the literature, the principal advantage of the present database is its focus on pixel-level segmentation of findings rather than whole-image classification based on histological outcome. This enables more detailed analysis of colposcopic lesion morphology, which is important for clinical interpretation. Inclusion of a dedicated JNEG class also allows evaluation of the model under both principal colposcopic protocols – acetic acid and Lugol’s iodine.

In semantic-segmentation training, augmentation aims to artificially expand the training set through controlled transformations of existing samples. This reduces the risk of overfitting and improves model generalization to unseen examples. For the limited dataset used in the present study, this process is particularly important because it exposes the model to variation in illumination, color balance, focus, and lesion positioning without requiring acquisition of new images. Three different augmentation strategies were applied, each expanding

the training set in a different manner. Photometric augmentation generated two additional copies of each original image through combinations of changes in brightness, contrast, and color saturation, together with synthetic glare, blur, darkening, and compression artifacts. Oversampling of rare classes increased the frequency of cases containing punctuation (P) and leukoplakia (L), presenting them five times versus a single presentation of the remaining cases. Cropping images around these rare findings gave the model an enlarged view of the lesions in an attempt to reduce class imbalance without changing the visual content of the images themselves.

The principal challenge when working with a limited amount of data is that the model may “memorize” findings from the training set rather than develop a generalized ability to recognize pathological changes in new, previously unseen images. To reduce this risk, the available data are divided into a training set, on which the model develops its diagnostic criteria, and a control validation set, which remains entirely unseen during training and is used for independent assessment of diagnostic performance. Because a single split may by chance be unusually favorable or unfavorable, the data are divided into n parts and training is repeated n times. Each part serves once as the validation subset while the remaining parts are used for training. Thus, every sample is used for validation exactly once, and the final result is the average across all folds, providing a substantially more stable estimate. The ratio between the two sets represents a compromise: more training data improve learning, but an excessively small validation set produces an unstable estimate with high variance.

The quality of the created dataset can be evaluated not only through quantitative annotation characteristics, but also through the segmentation model’s ability to learn each class consistently. If a finding is represented unambiguously, the model should achieve high discriminative performance; when labeling is inconsistent or conceptually heterogeneous, the metrics remain low regardless of architecture. This relationship is demonstrated in Figure 30, where the Dice coefficient for each class is calculated by pooling pixels from the five validation folds (pixel-pooled Dice). This metric differs from the macro Dice used for model selection because larger regions receive greater weight in the calculation.

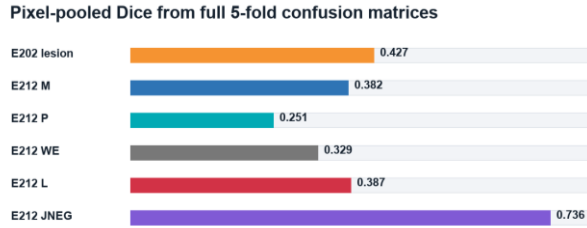


Figure 30. Dice coefficient calculated across all pixels from 5-fold cross-validation.

The results demonstrate substantial differences in the extent to which individual classes can be learned by the model. JNEG reaches a Dice coefficient of 0.736, indicating high annotation consistency and the presence of well-defined visual characteristics. In contrast, P (0.251), WE (0.329), and L (0.387) show markedly lower values. This likely reflects not only the lower frequency of these findings, but also greater variability in their delineation. After application of acetic acid, the boundaries between metaplastic epithelium, white epithelium, and punctuation are often gradual and overlapping, creating opportunities for different interpretations by individual experts. As a result, the model receives a less unambiguous training signal and achieves lower segmentation accuracy. Differences in Dice coefficients may therefore be regarded as an indirect indicator of the degree of consistency and objectivity in the annotation of individual colposcopic findings. In conclusion, inconsistency within the dataset can be considered one of the factors affecting segmentation performance, as further examined under Objective 3.

4.3 OBJECTIVE 3: To identify and analyze the factors that influence the model's segmentation performance.

To identify the factors affecting the segmentation performance of the developed model, we synthesized the quantitative results of the 13 experiments and the technical indicators from the dataset audit described under the first two objectives. We also performed an expanded analysis incorporating additional metrics, including ROC-AUC, precision, sensitivity, and softmax calibration, in order to define the principal barriers to diagnostic accuracy. The identified factors were grouped into three categories related to the data, the annotations, and the model.

4.3.1 Data-related factors

4.3.1.1 Class imbalance and data volume

The class imbalance described above emerges as a major factor limiting multiclass performance. The difference in pixel area between the most prevalent colposcopic finding, M (16.8 million pixels), and the rarest, L (3.6 million pixels), is 4.69-fold. This is compounded by the substantial proportion of images (22.8%) with no annotation content. The effect of this imbalance on performance is represented using five measures in the multiclass experiment E212, including Dice overlap, precision, sensitivity, and ROC-AUC. (Table 14)

Class	Number of polygons	Number of images	Precision	Sensitivity	Dice	ROC-AUC
JNEG	101	61	0.712	0.762	0.639	0.908
M	201	132	0.696	0.264	0.206	0.629
WE	142	88	0.555	0.234	0.136	0.704
L	62	26	0.761	0.259	0.207	0.686
P	117	71	0.623	0.157	0.091	0.638

Table 14. Class-wise performance metrics for model E212 (5-fold cross-validation).

The JNEG class, represented in 61 images and characterized by distinctive visual features, achieves ROC-AUC = 0.908 and sensitivity = 0.762. By contrast, P is represented in 71 images containing subtle, dispersed vascular dots and reaches a sensitivity of only 0.157. M has the largest number of samples (132) but Dice = 0.206, whereas L has the fewest examples (26) yet a similar result. This indicates that the number of images alone is not a sufficient predictor of performance; morphological complexity and lesion contrast are equally important. Absolute data volume is also a major factor: 355 images for five-class segmentation are below the recommended minimum. This quantity results in small validation subsets representing 5.8–9.7% of the data, increasing the variance of the estimate and amplifying variability between splits.

4.3.1.2 Variability between subsets and lack of patient-level grouping

The standard deviation of the mean overlap across all classes (Macro Dice) between individual folds is 0.044 for E202 and 0.039 for E212. This variability reflects the model’s sensitivity to the specific distribution of rare classes. A difference of 15 validation cases between folds (61 versus 46 for fold 0 and fold 4 of E212) can substantially alter representativeness for a rare class such as P, which is annotated in only 71 images overall. This is compounded by the lack of a patient-level mapping based on a unique personal identifier. Grouping by examination reduces the risk of data leakage but does not guarantee that two

separate examinations of the same patient cannot fall into different folds. This is an additional source of variability, particularly in a small dataset.

4.3.1.3 *Protocol-specific signal and lack of metadata*

One of the principal determinants of model performance is the protocol-specific color contrast generated at different stages of the colposcopic examination. JNEG achieves Dice = 0.639 and ROC-AUC = 0.908, values that substantially exceed those of the next-best class, M. This difference is not merely a consequence of class frequency, but reflects the visual prominence of a light-yellow region against a dark-brown iodine-positive background. Experiment E103 (without JNEG; M/P/WE/L only) shows a decrease in Dice from 0.242 to 0.186, indicating that approximately 29% of the multiclass signal is contributed by JNEG. (Table 6) The current dataset lacks metadata describing the colposcopic protocol, forcing the model to learn from mixed-protocol data and creating a risk that it associates JNEG with the general concept of “lesion” rather than recognizing it as a specific finding dependent on examination conditions.

4.3.1.4 *Effect of augmentation strategies*

The applied augmentation strategies had limited or negative effects on performance. Photometric augmentation (E201 with Dice = 0.259; E211 with Dice = 0.245) produced results below the respective baseline levels (E101 = 0.273). (Tables 4–5) This suggests that color is a critical diagnostic feature in colposcopy and that arbitrary color modification confuses the model. Increased repetition of rare examples in E202 and E212 kept Dice near the baseline range of 0.256–0.273 but increased variability. With a standard deviation of 0.044 compared with 0.026 for E101, oversampling appears to help the model encounter rare classes more frequently without improving the quality of the representation learned. Cropping regions around lesions in E203 and E213 reduced Dice below baseline, probably because isolating small regions deprives the model of the spatial context of the entire cervix. The only modification that substantially improved Dice was data filtering: E104, containing only JNEG, achieved Dice = 0.529, a 94% increase. This demonstrates that heterogeneity, rather than quantity alone, is the dominant problem. (Table 6)

4.3.2 Annotation-related factors

The distribution of Dice values in E212 provides an indirect but informative measure of annotation consistency. If a given class is annotated consistently, the model should be able to achieve a high Dice score. Conversely, if annotations are inconsistent, Dice remains low regardless of architecture. (Table 15)

Class	Count	Mean Dice	Median	≥ 0.50	≥ 0.75	≥ 0.90
M	168	0.204	0.001	35 (20.8%)	15 (8.9%)	0 (0.0%)
P	97	0.100	0.000	9 (9.3%)	6 (6.2%)	2 (2.1%)
WE	150	0.142	0.000	20 (13.3%)	8 (5.3%)	3 (2.0%)
L	44	0.201	0.000	9 (20.5%)	4 (9.1%)	3 (6.8%)
JNEG	61	0.638	0.803	44 (72.1%)	32 (52.5%)	20 (32.8%)

Table 15. Distribution of Dice values for E212.

The results reveal substantial differences in annotation consistency between classes. JNEG stands out with a high median overlap coefficient of 0.803, with 72.1% of iodine-negative cases exceeding the 0.50 threshold. By contrast, P, WE, and L have a median Dice of zero, meaning that in more than half of cases the model achieves no overlap with the annotation. In only 9.3% of P findings and 13.3% of WE findings does the model reproduce at least half of the lesion area, which cannot be explained solely by the rarity of these classes. Even when P and L are present in the dataset, their annotations are not sufficiently consistent to generate a stable training signal.

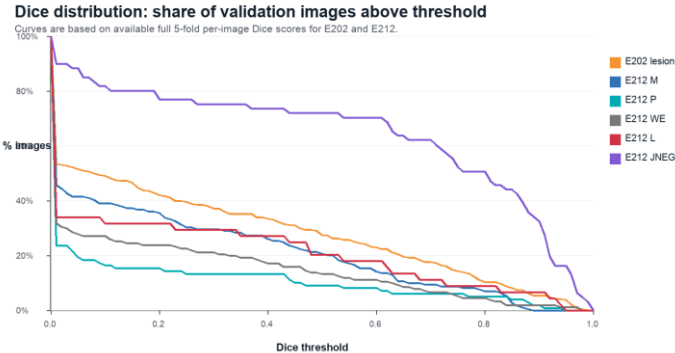


Figure 31. Proportion of images with a Dice coefficient above a specified threshold. JNEG shows the clearest positive signal, whereas P, WE, and L remain unstable.

This visualization more intuitively demonstrates the proportion of images that reach a practically useful level of overlap between the predicted and reference annotations. (Fig. 31) A substantial proportion of JNEG cases show high-quality segmentation, whereas for P, WE, and L a large share of images remain at low values. This indicates that the mean Dice is not concealing an otherwise reliable model for these subtypes, but rather reflects their genuine instability and limited reproducibility.

The matrix shows what occurs when the model attempts to classify five clinical categories that are not mutually exclusive. (Fig. 32)

E212 full 5-fold confusion matrix

		Predicted class					
		background	M	P	WE	L	JNEG
background		98.1%	0.3%	0.0%	0.3%	0.1%	1.2%
		282,138,945	930,723	99,621	890,009	226,855	3,427,681
M		69.1%	26.4%	0.4%	3.7%	0.4%	0.0%
		11,645,939	4,440,201	69,612	629,821	59,299	0
P		66.8%	6.1%	15.7%	11.4%	0.0%	0.0%
		2,722,045	247,536	639,991	462,992	327	0
True							
WE		69.4%	5.8%	1.3%	23.4%	0.1%	0.0%
		8,374,005	703,338	162,458	2,824,616	6,828	0
L		63.8%	1.0%	1.6%	7.8%	25.9%	0.0%
		2,291,196	34,248	56,127	280,544	931,412	0
JNEG		23.5%	0.2%	0.0%	0.0%	0.0%	76.2%
		2,610,997	27,558	0	0	0	8,467,476

Figure 32. Confusion matrix for experiment E212 using complete 5-fold cross-validation.

Nearly 70% of pixels belonging to M, P, WE, and L are assigned to background because the algorithm simply fails to recognize the lesion and defaults to the surrounding surface. The distribution of the remaining 30% is not random noise but follows clinical logic. The most common error is classification of P as WE (11.4%), followed by mutual confusion between M and WE (3.7–5.8%), because all three classes share the same acetowhite background. The asymmetry is informative: P is confused with WE nine times more often than the reverse. This means that under uncertainty the model regresses toward WE, much as a colposcopist might describe a difficult case simply as “white epithelium.” By contrast, not a single pixel from M, P, WE, or L is predicted as JNEG, reflecting the two fundamentally different examination protocols.

In practice, the softmax architecture imposes mutually exclusive classification, and experiment E106 demonstrates the importance of this factor. Changing the overlap priority from L > M > P > WE > JNEG to M > P > L > WE > JNEG reduces mean Dice by 17% with identical data and architecture. Experiment E105, in which M and P are merged into a vascular class and WE

and L into a white-change class, achieves a 49% improvement simply by removing an ambiguous boundary. The softmax function requires each pixel to belong to exactly one class, whereas in clinical practice M and P are findings superimposed on WE, and the visual boundaries between WE and L may be similar. Overall, the findings indicate that the softmax model is not primarily limited by architecture, but by the requirement to make categorical choices between classes that may be visually and conceptually overlapping.

4.3.3 Model-related factors

4.3.3.1 Target formulation – binary versus multiclass segmentation

Comparison between binary segmentation (E202, Dice = 0.273) and five-class multiclass segmentation (E212, Dice = 0.256) shows that target formulation is itself an influential factor. In the binary approach (E202), all lesions are merged into a single class and the model focuses on localization, whereas in the multiclass setting it must additionally assign each lesion to a subtype. The difference of approximately 0.017 Dice may appear small, but given that only one of the five classes (JNEG) is well represented, the actual clinical utility of the multiclass approach is more limited. This supports a two-stage strategy: first lesion detection using binary segmentation, followed by subsequent classification.

E202 full 5-fold confusion matrix

		Predicted class	
		background	lesion
True	background	97.5% 378,824,845	2.5% 9,710,589
	lesion	67.3% 32,096,360	32.7% 15,592,206

Figure 33. Confusion matrix for experiment E202 using complete 5-fold cross-validation.

In experiment E202, most pixels belonging to the background are classified correctly. At the same time, a substantial proportion of true lesional areas are incorrectly classified as background. (Fig. 33) This explains why the model achieves high accuracy and specificity while Dice and sensitivity for lesion

classes remain limited. In other words, the model performs well at recognizing normal areas but misses a substantial proportion of pathological changes.

4.3.3.2 Calibration and softmax confidence

Detailed analysis of softmax probabilities shows that the model generates high confidence when correctly recognizing a limited subset of cases, but systematically misses a substantial proportion of lesional pixels. Re-running the model on the 273 validation images from E212 (Table 12) while retaining the full probability maps enabled calculation of ROC-AUC and PR-AUC, as well as analysis of the softmax-confidence distribution. (Table 16)

Class	ROC-AUC	PR-AUC	Median probability for lesion pixels	Confidence in correct predictions
M	0.737	0.366	0.002	0.936
P	0.638	0.180	0.001	0.947
WE	0.704	0.257	0.001	0.929
L	0.686	0.286	0.001	0.953
JNEG	0.908	0.654	0.999	0.983

Table 16. Probability-based metrics for E212 (5-fold cross-validation).

PR-AUC (Precision–Recall AUC) is more informative than ROC-AUC for highly imbalanced data because it focuses on lesion pixels and better reflects the burden of false-positive predictions. For P (PR-AUC 0.180) and WE (0.257), the values are substantially lower than the corresponding ROC-AUC values (0.638–0.704), indicating that many predictions for these classes are false positives.

The third column, however, is particularly informative: it reports the median probability assigned by the model to pixels marked as lesion. For P, L, WE, and M, this value is only 0.001–0.002, meaning that most true lesional pixels are assigned almost zero probability of belonging to their correct class. Only JNEG has a median probability of 0.999. At the same time, when the algorithm is correct, it is highly confident: 90% of its correct predictions have a probability above 0.87. The combination of high confidence in a small number of correct cases and near-zero probability for most true lesion pixels indicates that the model learns only a small, internally consistent subset of examples and effectively ignores the remainder as contradictory.

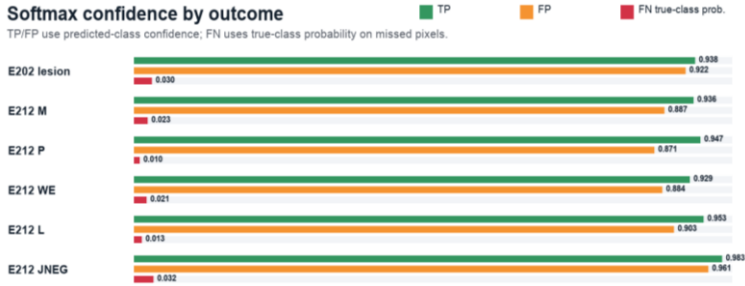


Figure 34. Softmax confidence according to classification outcome. TP – true positive; FP – false positive; FN – false negative. High confidence in FP and FN predictions indicates that the model often makes errors with high confidence.

The model demonstrates high confidence for a large proportion of correctly classified pixels, but it also retains high confidence for a substantial number of incorrectly classified pixels. (Fig. 34) This suggests inadequate calibration of the model and ambiguity both in the annotations and in the task itself. Softmax-confidence values should therefore not be used as a stand-alone indicator of clinical reliability without additional calibration and expert review.

4.3.4 Discussion

The factors identified as affecting segmentation performance fall into three categories related to the data, annotations, and model. The nnU-Net architecture has sufficient capacity to achieve excellent segmentation when the diagnostic signal is consistent and high-contrast, as demonstrated by the high ROC-AUC for JNEG (0.908). The weaker performance for P, WE, and L does not primarily reflect an architectural limitation, but rather the interaction between limited and imbalanced data, annotation inconsistency, and incompatibility of the softmax formulation with the clinical reality.

A dataset of 355 colposcopic images for five-class segmentation is substantially below the recommended minimum. By comparison, Yuan et al. (2020) used more than 11,000 images, and large medical-segmentation datasets may include hundreds to thousands of images, although their size varies considerably according to task and modality. (Kuş et al., 2024) The small dataset is further complicated by class imbalance, reflected in several-fold differences in the pixel area occupied by the various findings, as well as by empty annotation files. This combination leaves the rare classes with too few images and even fewer meaningful lesional pixels from which the model can learn. Small

validation subsets further increase the variance of the estimate; with a sample of this size, even a minimal difference between two splits may alter statistical representativeness.

Beyond quantity, data heterogeneity emerges as a more fundamental problem. The protocol-specific contrast of iodine-negative areas differs so markedly from acetic-acid images that the model achieves substantially higher performance for JNEG despite the relatively small number of examples. Experiment E103 shows that JNEG contributes materially to the overall result, as removal of this class reduces Dice by approximately one third. The model is trained simultaneously on two different image types, but because information on the protocol used is unavailable, they cannot be separated and are processed as a single group, which likely complicates model training.

The failure of the augmentation strategies reinforces this conclusion. Photometric augmentations worsen performance because arbitrary changes in color destroy diagnostically meaningful signal. Oversampling maintains Dice near baseline but does not improve it, indicating that repetition of rare examples adds no new information. The only marked improvement results from filtering the data, further demonstrating that heterogeneity, rather than quantity alone, is the principal limiting factor.

Annotation consistency proved to be an even greater challenge than class imbalance. This is supported by two mutually reinforcing lines of evidence. First, the distribution of Dice values shows that the problem is not simply class rarity. If imbalance were the sole cause, low but more evenly distributed Dice values would be expected. Instead, P, WE, and L have a median Dice of zero, meaning that in more than half of cases there is no overlap at all between prediction and annotation. This is qualitatively different: for a substantial proportion of examples, the annotation and visual signal are effectively inconsistent, making it impossible for the model to learn the finding reliably. Comparison with E104 (JNEG only, Dice 0.529) shows that when annotations are consistent, the model performs well even with relatively little data.

Second, the E212 confusion matrix visualizes this conflict at the pixel level. (Fig. 32) The dispersion of M, P, WE, and L is not random noise but follows clinical logic: under uncertainty, the model regresses toward WE as a baseline class, where the probability of error is reduced. JNEG remains completely separate, further supporting the conclusion that its signal arises from a different source rather than from the same shared class space.

Experiments E105 and E106 provide the strongest evidence that annotation ambiguity, rather than model design, is the principal obstacle. Changing only the priority used to resolve overlaps causes a decrease in Dice with identical data and architecture, a result that is difficult to explain by architectural limitations. Conversely, reducing inconsistency by merging clinically overlapping classes improves Dice without any change to the model itself. Together, these two

experiments show that the softmax formulation is poorly suited to overlapping IFPC categories and transforms annotation subjectivity into a systematically contradictory training signal.

Model-specific factors are less pronounced but remain informative. The difference between binary and multiclass formulations reflects the additional burden placed on the model when it must assign lesions to subtypes while most classes are poorly represented. This supports a two-stage strategy in which the lesion is first detected and then classified, thereby decomposing the problem into two simpler tasks.

Analysis of softmax calibration reveals another aspect of the same problem. The discrepancy between ROC-AUC and PR-AUC shows that the model not only makes errors, but often does so confidently. High probability for the relatively few correct predictions, combined with near-zero probability for most true lesional pixels, is a classic pattern associated with noisy or inconsistent supervision. Raw softmax values are therefore unsuitable as a measure of diagnostic confidence without additional calibration.

In summary, the hierarchy of factors is as follows: the data provide the raw material, the annotations determine the quality of the signal, and the model reflects imperfections in both. Future development of the system should move away from primarily quantitative attempts at augmentation toward qualitative improvement of the annotation protocol and reconsideration of the softmax formulation for clinically overlapping classes.

4.4 OBJECTIVE 4: To determine the diagnostic value of the developed model through comparative analysis with established systems reported in the literature.

4.4.1 Bridging segmentation and classification metrics

The principal challenge in comparing the developed system with algorithms published in the literature arises from differences in the task being addressed. The present model operates at the pixel level through multiclass segmentation of colposcopic findings according to IFCPC nomenclature, whereas most published solutions perform image-level classification. Under these conditions, direct comparison is limited, and the only widely used metric that permits a reasonably valid parallel between the two approaches is the area under the ROC curve (AUC). (Table 17)

Model / Class	ROC-AUC	PR-AUC	Median Dice	Dice > 0.50	Dice > 0.90
E212 - JNEG	0.908	0.654	0.803	72.1%	32.8%
E212 - M	0.737	0.366	0.001	20.8%	0.0%
E212 - WE	0.704	0.257	0.000	13.3%	2.0%
E212 - L	0.686	0.286	0.000	20.5%	6.8%
E212 - P	0.638	0.180	0.000	9.3%	2.1%
E202 (binary)	0.767	-	0.085	27.7%	5.5%

Table 17. Diagnostic performance of the developed models (E212 multiclass model with finding-specific outputs and E202 binary model).

Within the multiclass setting, the most convincing segmentation signal was observed for the iodine-negative class. JNEG achieved ROC-AUC = 0.908 and a median Dice of 0.803, with more than 50% overlap between annotation and model prediction in 72.1% of images. The best examples reached Dice values as high as 0.993. (Fig. 35) The binary model (E202) showed intermediate performance, with ROC-AUC = 0.767 and 27.7% of images exceeding Dice 0.50; its highest recorded Dice was 0.992. (Fig. 36)

Example 14: JNEG (E212). Near-perfect overlap, with Dice = 0.993. These cases belong to the 32.8% of images for which the model achieved Dice > 0.90.

Top available validation examples: E212 JNEG

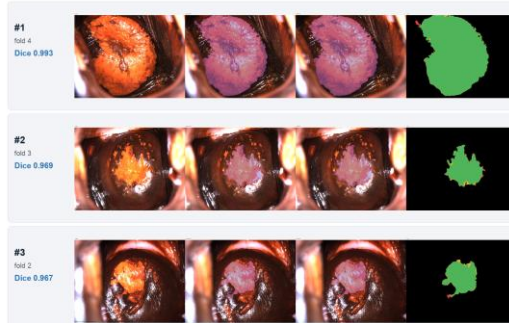


Figure 35. Examples of the best performance of the JNEG class in the multiclass model E212.

Example 15: Binary model (E202). High-quality lesion localization without underlying subtype classification (Dice = 0.958–0.992).

Top available validation examples: E202 lesion

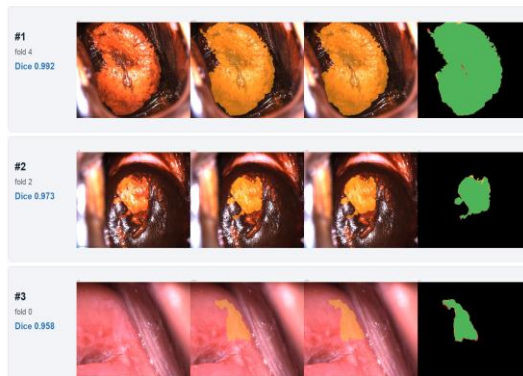


Figure 36. Examples of the best performance of the binary model E202.

The remaining classes (M, P, WE, and L) show ROC-AUC values between 0.638 and 0.737 and median Dice values of only 0.000–0.001. Despite these lower overall values, an average of 15.97% (range 9.3–20.8%) of acetowhite findings exceed Dice 0.50, while the best validation examples achieve high overlap of 0.81–0.94 Dice (Figs. 37–39).

These visualizations confirm that the model can successfully localize and segment the corresponding classes when the imaging signal is sufficiently pronounced.

Example 16: Class M (E212). High overlap in prominent mosaic findings (Dice = 0.843–0.875).

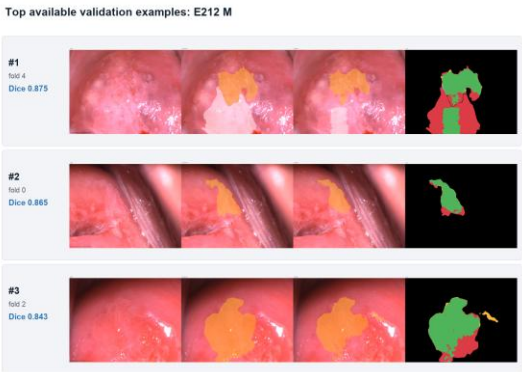


Figure 37. Examples of the best performance of class M in the multiclass model E212.

Example 17: Class P (E212). High overlap in a visually distinct punctate region (Dice = 0.881–0.942).

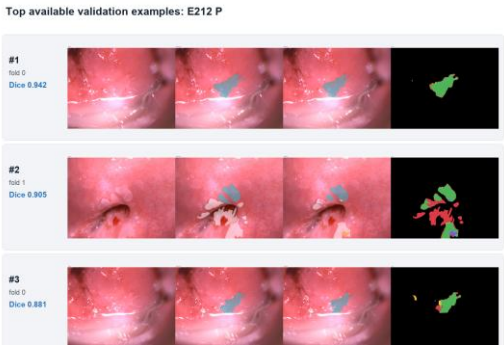


Figure 38. Examples of the best performance of class P in the multiclass model E212.

Example 18: Class WE (E212). The best overlap for WE annotations is achieved in the presence of extensive regions of white epithelium (Dice = 0.810–0.838).

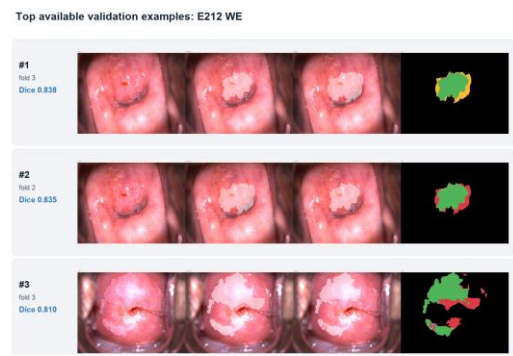


Figure 39. Examples of the best performance of class WE in the multiclass model E212.

The examples demonstrate that the model is not fundamentally limited by its architecture when the input signal is clear: large mosaic fields, prominent punctate structures, extensive white epithelium, or high-contrast JNEG regions. Under such conditions, nnU-Net is capable of high-quality segmentation comparable to other developed systems discussed in the following section.

4.4.2 Comparison with published data

The comparison is based on the literature review presented above, the systematic review by Takahashi et al. (2025), which examined 19 studies published between 2019 and 2024, and the meta-analysis by Liu et al. (2024), which included 26 colposcopy studies published through 2024. Ten studies overlapped between the two publications, yielding a total of 35 unique colposcopy articles.

4.4.2.1 ROC-AUC profile comparison with systems reported in the literature

According to the pooled data from the meta-analysis by Liu et al., the reported summary AUC was 0.91 (95% CI 0.88–0.93) for the included AI-based diagnostic systems.

System	Task type	Dataset size	AUC
--------	-----------	--------------	-----

E212 multiclass model	Pixel-level segmentation of JNEG	355 images	0.908
E202 binary model	Binary pixel-level segmentation	355 images	0.767
Fang et al. (2022)	5-class image-level classification model	6,996 images from 1,189 patients	0.99
Fan et al. (2022)	3-class image-level classification model	7,106 images	0.9856
Mascarenhas et al. (2024)	HSIL versus LSIL classification	22,693 frames	0.98
Xue et al. (2020)	Image-level grading	101,267 images from 19,435 patients	0.947
<u>Yue Z et al. (2020)</u>	Grading according to CIN classification	4,753 images	≥ 0.94
Yuan et al. (2020)	Hybrid system (classification and segmentation)	22,330 cases	0.93
Fu et al. (2022)	Multimodal image-level system	2,160 patients	0.921
Hu et al. (2019)	CIN2+ level classification	9,406 images	0.91
Cho et al. (2020)	Binary HSIL+/LSIL- classification system	1,426 images from 791 patients	0.781

Table 18. ROC-AUC comparison with systems reported in the literature.

The AUC of 0.908 achieved for the JNEG class falls within the range reported by Liu et al., but was obtained at the pixel level, a fundamentally more difficult task than image-level classification. The binary model (E202) reaches AUC = 0.767, a result comparable with the study by Cho et al. (0.781), which used a binary approach (HSIL+ versus LSIL-) at image level and with approximately four times more data.

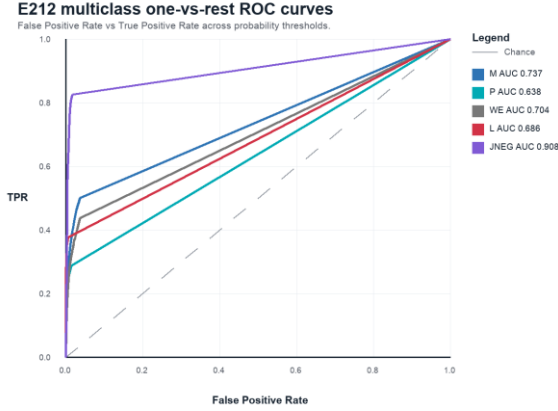


Figure 40. Conventional one-vs-rest ROC curves for experiment E212. The JNEG curve is clearly separated from the weaker P, WE, and L curves.

ROC analysis demonstrates substantial class-specific differences in recognizability. (Fig. 40) JNEG shows the highest discriminatory ability (ROC-AUC 0.908), with a curve sharply displaced toward the upper-left quadrant. M and WE occupy an intermediate position (ROC-AUC 0.737 and 0.704), while P and L remain the most difficult to distinguish (0.638 and 0.686), with curves closer to the diagonal. This class hierarchy is stable and mirrors the Dice results, indicating that the limitation does not arise from the choice of metric but from the information carried by each class itself. The key implication is that a single multiclass model must simultaneously learn classes with fundamentally different levels of visual recognizability, creating a substantial training challenge.

4.4.2.2 Segmentation analysis and comparison of Dice scores with systems reported in the literature

ROC-AUC analysis provides a broad perspective that permits comparison with published systems addressing heterogeneous tasks. To assess segmentation accuracy, however, the next step is to compare the spatial overlap between predicted and annotated regions. The present model was designed to operate in a multiclass setting, whereas most published studies define segmentation as a binary task – lesion versus background. This leads to two complementary comparisons: multiclass and binary. First, E212 is compared with studies reporting Dice for their segmentation component, while accounting for the fundamental differences in the number and definition of classes. Second, E202 provides the methodologically cleanest parallel by reducing the task to the

common binary denominator used in the literature, allowing the effect of multiclass complexity to be distinguished from the effect of limited data volume.

Among all reviewed sources, only three report a spatial-overlap coefficient: Yue Z et al. (2021), Yu et al. (2022), and Yuan et al. (2020). (Table 19)

System	Segmentation target	Dice/IoU	AUC	N	Note
E212	M/P/WE/L/JNEG	0.803 (JNEG) / 0.256 (macro)	0.908 (JNEG)	355	Multiclass model
E202	Lesion/background	0.427 (pixel) / 0.273 (macro)	0.767	355	Binary model
Yue Z et al. (2021)	Segmentation of acetowhite regions	0.823	0.973*	3045	Binary acetowhite segmentation
Yu et al. (2022)	LSIL+	0.737	N/A	5455	Binary LSIL+ segmentation
Yuan et al. (2020)	Lesion	0.616	0.93	22 330	Hybrid segmentation-classification
Liu J et al. (2022)	Segmentation of acetowhite regions	0.636 mean IoU/Jaccard	N/A	280	Accuracy 91.2%, Specificity 94.9%

Table 19. Studies with a segmentation component.

**AUC for classification.*

Yue Z et al. (2021) achieved the highest Dice reported in the literature (0.823) for binary segmentation of acetowhite regions (AW areas). This result indicates the upper range of segmentation accuracy when sufficient data and a clearly defined visual target are available. For comparison, we used our WE class as the closest analogue to AW regions. WE alone achieved a mean Dice of only 0.142, with just 13.3% of cases exceeding the 0.50 threshold. These values reflect not only the difference in data volume but also the complexity of the multiclass task. To test whether class fragmentation was a principal limitation, WE and L were merged into a common white-change class (E105), increasing Dice to 0.351, approximately 2.5 times higher than for WE alone. The results confirm two limiting factors: the multiclass formulation accounts for a Dice difference of approximately 0.68 relative to Yue, while limited data contribute a further difference of approximately 0.47 even after class fragmentation is reduced.

Yu et al. (2022) segmented images acquired after acetic-acid application within a binary LSIL+ versus normal task. Their Dice of 0.737 compared with

our macro Dice of 0.256 illustrates the cost of dividing findings into separate IFPC classes, corresponding to a decrease in overlap of approximately 0.48 points relative to a binary approach using a comparable protocol.

Yuan et al. (2020) proposed a hybrid system for simultaneous segmentation and classification, achieving Dice = 0.616 for segmentation and AUC = 0.93 for image-level detection. By comparison, our E212 model achieves macro Dice = 0.256 and macro ROC-AUC = 0.735. The only directly comparable value is the JNEG ROC-AUC of 0.908, which approaches Yuan's classification AUC of 0.93 despite being derived from a pixel-level multiclass task using approximately 63 times less data and an iodine-contrast protocol.

Liu et al. (2022) performed acetowhite segmentation on 280 images and achieved accuracy of 91.2% and specificity of 94.9%, without reporting a spatial-overlap metric. This study is the closest to ours in data volume but remains binary in nature. The reported metrics are dominated by background pixels and do not directly reflect spatial overlap between predicted masks and reference annotations, potentially leading to misleading conclusions. Nevertheless, the high values achieved with a limited number of images support the conclusion that binary acetowhite segmentation is feasible even with a small dataset, whereas multiclass IFPC segmentation requires substantially more data with more consistent annotation.

The performance of the binary model E202 remains substantially below the values reported in the literature (Dice 0.616–0.823). (Fig. 41) Yue's team used 3,045 images, Yu's 5,455, and Yuan's 22,330, representing approximately 8- to 60-fold more data. With such large samples, published models are trained on a substantially more representative range of lesion morphology, enabling more precise delineation. Nevertheless, E202's ROC-AUC of 0.767 shows that the model can distinguish lesion from background even when spatial accuracy measured by Dice remains low. This discrepancy likely reflects the limited number of examples: the model captures the overall signal but not its precise boundaries. With more data, Dice would be expected to improve and more closely reflect the discrimination already indicated by ROC-AUC.

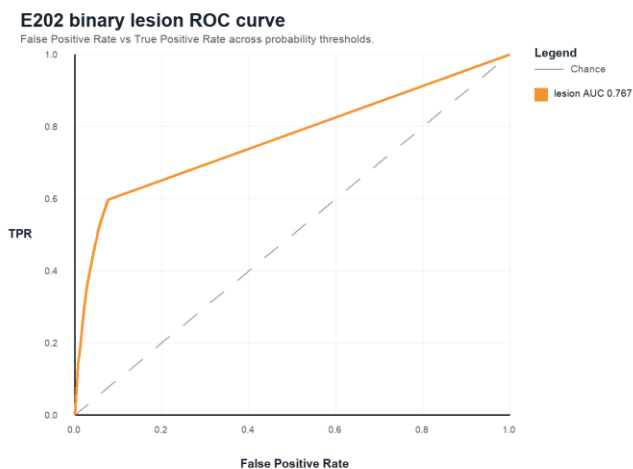


Figure 41. Conventional ROC curve for experiment E202.

ROC analysis shows that the model performs better than random classification and can identify a proportion of lesional areas. Its discriminatory ability remains limited, however, because high sensitivity and a low false-positive rate have not been achieved simultaneously. E202 should therefore be regarded as a promising research model requiring further optimization before potential clinical application.

4.4.3 Discussion

In his systematic review, Takahashi identified a high risk of bias in 17 of the 19 studies because convenience retrospective samples were used, including all available cases that met the predefined criteria. The meta-analysis by Liu reported high heterogeneity when comparing AI performance with colposcopist assessment ($I^2 = 93\%$). These limitations of the underlying literature must be considered when interpreting the comparisons below. (Table 20)

Parameter	Range reported in the literature	Developed model
Task type	Image-level classification (CIN2+, HSIL+)	Pixel-level segmentation (multiclass)
AUC	0.91 (pooled value from the Liu meta-analysis)	0.908 (for JNEG)

AUC (binary model)	0.78–0.91	0.767
Dice	0.823 (Yue Z et al.) 0.737 (Yu et al.) 0.616 (Yuan et al.)	0.803 (JNEG class) 0.256 (macro Dice, all classes) 0.427 (binary model)
IFCPC-based multiclass pixel-level segmentation	Not reported in the literature	Yes (M, P, WE, L, JNEG)
Clinical dataset size	134–101,267 data points	355 images

Table 20. Synthesized comparison with the literature.

The table highlights two principal lines of discussion. First, the JNEG ROC-AUC of 0.908 is nearly identical to the pooled mean AUC of 0.91 reported by Liu for AI analysis at whole-image level, yet it was achieved at pixel level using 355 images. Second, the binary model E202 (AUC 0.767) falls at the lower end of the published range (0.78–0.91), which is expected given the substantially smaller amount of data.

The strong performance of the JNEG class is comparable with the three binary segmentation studies, even though E212 addresses a five-class task using many times less data. This supports the capacity of the nnU-Net v2 ResEnc-M architecture, but the result must be interpreted in the context of a protocol-specific color signal. JNEG segmentation relies on the high contrast between brown iodine-positive background tissue and yellow-orange iodine-negative areas, producing a sharp visual boundary. By contrast, the cited published systems segment acetowhite regions, which appear as subtle white gradients on a pink-red mucosal background, a protocol fundamentally different from Lugol’s iodine. JNEG should therefore be regarded as a positive control confirming the validity of the architecture rather than as a direct analogue of AW-based systems in the literature. A more meaningful comparison with those studies should be based on the WE class (Dice 0.142, ROC-AUC 0.704), which uses the same acetic-acid protocol and remains substantially below published values.

The comparative analysis identifies three fundamental barriers that limit the clinical value of the developed system. These barriers are not independent; rather, they act in combination, with each amplifying the effects of the others.

4.4.3.1 Dataset size and class fragmentation

Predicted segmentation systems achieving high overlap in colposcopy rely on 3,045–5,455 images trained to solve a single binary task (AW/normal or

LSIL+/normal). With 355 images distributed across five classes, the developed model has an average of approximately 70 images per class, with the rarest class, L, represented by only 44 cases. Studies of learning curves in medical imaging indicate that approximately 1,000 annotated examples per class may be required to achieve stable accuracy above 98%; with smaller samples, models show high variability and unstable results, a pattern directly corresponding to the performance observed here. (Cho et al., 2015)

Class fragmentation further magnifies the effect of limited data volume. In experiment E105, merging WE and L into a common “white” class increased Dice from 0.142 to 0.351, a 2.5-fold increase. This suggests that the model contains sufficient information to identify abnormality, but not to perform the finer differentiation between subtypes, which requires a larger and more diverse set of annotated examples. The class imbalance between M (168 polygons) and L (44 polygons) means that the model encounters nearly four times more mosaic examples than leukoplakia examples, creating a systematic tendency to predict M more often at the expense of rarer classes. Both reviews support this assessment: Takahashi explicitly notes that small and demographically homogeneous datasets limit generalizability, while Liu recommends that AI systems also be developed and adapted for resource-limited settings, where data volumes are traditionally smaller.

4.4.3.2 *Annotation subjectivity and conceptual overlap between classes*

The clinical nature of colposcopic findings is inherently continuous and spectral, reflecting the dynamic interaction between human papillomavirus infection and host defense mechanisms. Mosaicism (M) and punctuation (P) are vascular changes that develop on a background of white epithelium (WE), while leukoplakia (L) may involve both WE and deeper stromal changes. In clinical reality, these findings therefore often coexist within the same image and overlap spatially. The amount of training data required for stable segmentation depends on the visual complexity and degree of class overlap. For homogeneous, spatially isolated structures (e.g., bone), nnU-Net can achieve stable results with 10–30 cases, whereas visually overlapping classes with continuous features require substantially more data because of the subtle discriminative differences between subtypes. (Huysentruyt et al., 2024)

The annotations created by a single expert reflect this clinical ambiguity: polygons for M are often located within larger WE regions, creating a conflict

under the softmax formulation. The software resolves this conflict through a hierarchical priority ($L > M > P > WE$), which forcibly eliminates part of the annotation information in each processed image. Experiments E000 and E106 provide direct quantitative evidence of this effect. Changing the softmax priority from $L > M > P$ to $M > P > L$ resulted in a 17% decrease in Dice (from 0.154 to 0.128), with no change in the clinical image, annotations, or architecture. This sensitivity to a technical decision with no clinical analogue demonstrates that the formulation does not fully reflect the clinical reality, in which IFCPC findings may overlap.

The high heterogeneity of the annotations is also reflected in the distribution of Dice values. The median Dice for P is 0.000, meaning that in more than half of the validation cases the model shows no overlap with the reference annotation. This is not necessarily a model error, but rather a reflection of the fundamental ambiguity involved in marking fine punctuation patterns, which may be interpreted differently even by experienced colposcopists.

4.4.3.3 *Lack of metadata and protocol-dependent signal*

JNEG dominates the multiclass signal for protocol-related rather than morphological reasons. In the absence of acquisition metadata, the model is trained on a mixed set of images with fundamentally different visual characteristics, encouraging it to exploit the contrast signal rather than learn the morphological features of the lesion. For this reason, JNEG cannot be evaluated correctly without protocol information, because it is unclear to which images the iodine stage of the colposcopic examination applies. This limitation is consistent with Takahashi's assessment that most studies included in the review do not report acquisition conditions in sufficient detail, such as the sequence and type of solutions applied, the time elapsed after application, and the lighting conditions. The recommendation for standardization indicates that the lack of metadata is recognized by the field as a systemic problem that limits both comparability across studies and reliable assessment of the clinical applicability of the developed algorithms. Incorporating protocol metadata into future datasets would allow individualized evaluation, which is a prerequisite for clinically meaningful interpretation of the results.

4.5 FIFTH OBJECTIVE: To analyze the clinical applicability and role of artificial intelligence in contemporary colposcopic diagnosis.

The application of artificial intelligence in the diagnosis of precancerous cervical lesions is developing in response to several persistent limitations of conventional colposcopy: subjectivity of interpretation, interobserver variability, dependence on examiner experience, and unequal access to expert colposcopic assessment. The field has progressed through several stages, from early attempts at quantitative texture and color analysis, through the first deep-learning models, to contemporary multimodal systems for segmentation, classification, and biopsy guidance. This evolution reflects a change in the very concept of computer-assisted colposcopy and marks a transition from isolated recognition of an abnormal image to support for specific clinical decisions.

The models included in the systematic review by Takahashi et al. show a wide range of diagnostic performance, with reported sensitivity ranging from 62.1% to 99.7% and specificity from 56.6% to 98.6%. Because of the retrospective design of the studies and the selection of readily available and convenient-to-analyze data, there is a risk of bias related to the characteristics of the included patients. The review emphasizes substantial methodological heterogeneity across publications, including differences in populations, reference standards, definitions of a positive result, model architectures, and image-quality control. These findings demonstrate considerable potential, but they cannot be interpreted unambiguously as evidence of readiness for routine clinical implementation.

The most substantial quantitative synthesis to date is the meta-analysis by Liu et al. Data on the accuracy of artificial intelligence were extracted from 26 studies, while data for experienced colposcopists were obtained from 14 studies. For sensitivity, specificity, positive predictive value, and negative predictive value, 16 studies of AI and 8 studies of experienced colposcopists, respectively, were included. When comparing AI with human expert assessment, the authors excluded junior colposcopists, and when several specialists participated, the mean diagnostic accuracy achieved by them was used. The comparison between experienced colposcopists and AI-assisted colposcopy showed that artificial intelligence achieved higher pooled accuracy (81% vs 74%) and higher specificity (83% vs 67%), with comparable sensitivity (86% vs 85%). This suggests that published AI systems have the potential to reduce false-positive assessments and support standardization of colposcopic judgment. In subgroup analysis, AI outperformed clinicians both for the LSIL+ task, achieving 84% accuracy, 86% sensitivity, and 85% specificity, and at the HSIL+ threshold, with 81% accuracy, 89% sensitivity, and 86% specificity. The high heterogeneity

among the included studies requires these results to be interpreted as evidence of clinical potential rather than as justification for autonomous use. A large proportion of the available systems have been evaluated retrospectively, in single centers, or on limited and highly controlled image datasets. The true value of AI can be established only through prospective, multicenter studies with external validation that account for the variable conditions described above and the realities of clinical workflow.

Colposcopy remains an indispensable diagnostic method with unquestionable clinical value; however, its performance is highly dependent on examiner experience and the subsequent subjective interpretation of visualized changes. This variability is directly relevant to the present dissertation because it shows that the challenge in colposcopy is not purely diagnostic but also distinctly semiotic. Acetowhite epithelium, mosaicism, and punctation are not completely discrete categories. They often coexist, merge into one another, and depend on multiple factors, including the time after acetic acid application, the quality of the applied solution, technical conditions, and others. This helps explain why image-based assessment alone may be inconsistent even among experts. In this context, an algorithm can apply the same visual logic to a large number of images, highlight suspicious regions, quantify their area and probability signal, preserve reproducible information, and provide a form of “second opinion.” This does not eliminate the need for a colposcopist, but it may reduce inconsistency in those elements of the method that are most strongly influenced by human judgment.

The literature analysis allows several practical directions to be defined in which AI may be integrated into the clinical workflow: as an intelligent assistant for assessing suspicious findings and selecting biopsy targets; as a tool for training and standardization; and as a solution in settings with limited access to medical expertise and resources.

4.5.1 *Assisted colposcopic assessment*

AI can support the physician by highlighting suspicious regions, estimating the probability of a high-grade lesion, and visualizing areas that warrant particular attention. In the randomized two-center clinical trial by Kim et al., localization AUC improved by 0.12, sensitivity increased from 71.33% to 89.18%, specificity from 92.16% to 96.68%, and overall diagnostic accuracy from 75.45% to 86.40%, with the greatest benefit observed among less experienced colposcopists. These results support the potential of artificial intelligence to complement expert judgment both in diagnostic classification and in the spatial localization of suspicious areas.

4.5.2 *Biopsy guidance*

The clinical value of colposcopy lies not only in visualizing abnormal findings but also in directing biopsy toward the area with the most pronounced changes. According to ASCCP standards, obtaining two to four targeted biopsies improves detection of existing precancerous lesions. (Wentzensen et al., 2017) This underscores the importance of selecting the correct sites for histological examination. The CAIADS system developed by Xue et al. demonstrates precisely this capability, achieving an mIoU of 0.758 for prediction of biopsy sites. This suggests that one of the most promising clinical functions of AI is spatial guidance rather than merely providing a binary diagnosis. The value of such a system may be particularly pronounced for inexperienced colposcopists or in images with ambiguous findings.

4.5.3 *Training and standardization*

AI systems offer an opportunity to standardize colposcopic assessment through objective and reproducible image analysis. In an educational setting, AI can serve as a training tool for residents by providing immediate feedback against expert annotations. Segmentation maps are particularly valuable because they visualize the spatial distribution of findings. This makes them suitable for teaching, for comparison between expert annotations and AI-generated suggestions, and for developing visual diagnostic competence. In this way, dependence on individual experience may be reduced and structured training across different centers may become possible. The study by Wang et al. (2024), involving 284 Chinese colposcopists from 24 provinces, found a high level of willingness to use AI-assisted systems, particularly when the technology was perceived as an aid rather than a replacement for the physician. Of the respondents, 82.4% preferred the colposcopist to interpret the image with CAIADS assistance, 15.49% accepted CAIADS interpretation under physician supervision, and only 0.70% would allow the system to interpret images independently. Against this background, only one third of respondents were familiar with the system and its current development, indicating both openness to innovation and a need for training and organizational preparation.

4.5.4 *Improving access in resource-limited settings*

The WHO global strategy aims to reduce cervical cancer incidence to 4 new cases per 100,000 women per year. Achieving this target requires greater diagnostic capacity, which AI has the potential to provide—not as a compromise with established standards, but as a form of standardized support. In resource-limited regions, AI-based colposcopy systems may help address the critical

shortage of trained specialists. Shamsunder et al. (2024) found that the EVA Visual Check algorithm, integrated into a portable digital colposcope, achieved a sensitivity of 89.3% for detecting CIN2+, close to that of physician-performed colposcopic assessment (92.9%). However, the substantially lower specificity of the AI system compared with physicians (26.1% vs 68.9%) indicates that it is better suited to initial triage and referral for expert assessment than to independent diagnosis. Similar systems integrated into portable colposcopes, smartphone-based devices, or cloud platforms could reduce dependence on the immediate availability of an expert and democratize access to colposcopic diagnosis in remote areas.

4.5.5 Discussion

The model developed in the present work should be positioned as a research system for pixel-level segmentation of colposcopic findings, with potential clinical applications in assisted colposcopy and training.

The principal conceptual difference from many published systems lies in the formulation of the task. Most models described in the literature operate at the image or patient level and classify the outcome into binary categories, such as normal versus abnormal or low-grade versus high-grade lesion. By contrast, our approach attempts to reproduce the spatial task performed by the colposcopist, mimicking the reasoning steps of the “natural” human algorithm. To the best of our knowledge, this is one of the first systems aimed at multiclass pixel-level segmentation structured according to the IFPCP nomenclature.

Binary classification answers the question of whether an image is normal or pathological, but it often does not explain why. Although IFPCP-based segmentation is more difficult and currently yields lower performance metrics, it seeks to make the decision traceable by localizing specific colposcopic regions. It is therefore not merely an alternative technical approach, but an attempt to build artificial intelligence that is more interpretable, clinically verifiable, and closer to real colposcopic reasoning. For the practicing physician, it is essential to know not only what final result an algorithm provides, but also how it arrived at that result, which imaging features it relied upon, and whether its conclusion can be verified. This is precisely where one of the major limitations of contemporary AI solutions in medical imaging becomes apparent: their development as “black-box” systems. Such an approach may be statistically effective but remains limited in terms of clinical interpretability and physician trust. This position is consistent with Wang et al., who emphasize that clinical implementation of AI in cervical diagnostics requires not only high algorithmic

accuracy but also explainable interfaces in so-called clinician-in-the-loop settings, in which the physician remains an active participant in the diagnostic decision.

The most important conclusion from our own results is not that the model is ready for clinical use, but that it clearly demonstrates the boundary between technological feasibility and clinical readiness. The JNEG class supports the conclusion that the architecture used can extract a stable visual signal from well-contrasted and consistently annotated findings. The low results for P, WE, and L indicate that the limitations in overall performance are related not only to the model itself but, above all, to the volume, structure, and annotation consistency of the data, the lack of protocol metadata, and the conceptual overlap among individual colposcopic features. Thus, even with lower Dice values, multiclass segmentation retains important scientific and clinical value. Limited overlap should not be interpreted simply as model failure, but as a reflection of the complexity of the task: indistinct boundaries between findings with similar visual signals, quantitative class imbalance, a limited number of annotated images, and subjectivity even in expert human assessment. The presented system should therefore not be viewed solely through the lens of absolute segmentation performance, but as a step toward more transparent artificial intelligence in colposcopy. In real clinical practice, trust in an AI system would likely be higher if it did not merely assign a probability of disease but also visualized the specific findings on which its assessment was based. IFCPC-based segmentation may therefore serve as an intermediate explainable layer between the image and the diagnostic decision while preserving the physician's role in interpreting the results.

Transitioning from a research prototype to genuine clinical value requires several priorities to be addressed systematically. The first is scaling and balancing the data. The current dataset is sufficient to demonstrate proof of concept, but insufficient for stable learning of subtle and rare classes. Published high-performing systems generally use datasets many times larger, indicating that future development should include more images balanced by class, protocol, and clinical outcome.

The second priority is the inclusion of protocol metadata. The absence of information on whether an image was acquired after acetic acid or Lugol's iodine, the time elapsed since application, the magnification level, and the type of transformation zone limits the model's ability to differentiate findings appropriately.

The third priority is to improve the annotation strategy and the way in which clinical reality is represented to the model. For individual classes, a multi-annotator consensus is needed, with interobserver agreement quantified and clear adjudication of debatable regions of interest. When findings partially overlap spatially, a single region may contain elements from several annotations, which

technically means that one pixel may belong to more than one class. In such cases, a hierarchical multilabel approach is more appropriate than the conventional softmax formulation, which forces the model to choose a single class and automatically excludes the others.

The fourth priority is clinical validation to assess the model's actual value. For this purpose, images must be linked to patient-level records including HPV status, cytology, histology, type of biopsy or excisional procedure, and follow-up. Only then can evaluation move beyond pixel overlap to assessment of clinical benefit.

The final priority is external, prospective, multicenter validation, without which it cannot be demonstrated that the model performs beyond the conditions in which it was trained. This is a key requirement from both scientific and regulatory perspectives. In the European Union, the AI Act classifies AI-based software intended for medical use as a high-risk system requiring appropriate governance, high-quality data, clear information for users, and human oversight. Clinical implementation is therefore not merely a technical process, but also a regulatory, ethical, and organizational one.

Based on the literature review and the results of the model trained on our own dataset, it can be concluded that artificial intelligence has a role in contemporary colposcopic diagnosis of precancerous cervical lesions. At present, it should be regarded as an assistive tool of increasing importance, but not as an independent substitute for the colposcopist.

The role of artificial intelligence in colposcopy should not be viewed as competition with the physician, but as a means of increasing the objectivity and standardization of the diagnostic process within a collaborative framework in which the colposcopist retains the leading clinical role.

5. CONTRIBUTIONS

Scientific and theoretical contributions

- 5.1** The role of artificial intelligence in contemporary colposcopic practice has been defined, together with the principal areas of potential application, the prerequisites for implementation, and the challenges associated with clinical integration.

Scientific and applied contributions

- 5.2** For the first time in Bulgaria, an original artificial intelligence-based system for automated pixel-level segmentation of cervical colposcopic findings has been developed and experimentally evaluated.

- 5.3 A dedicated colposcopic image dataset suitable for training and comparative evaluation of segmentation models using artificial intelligence methods has been created and systematically structured.
- 5.4 A unified protocol for expert annotation of colposcopic images has been developed, including definition of target classes, spatial delineation of findings, conversion into pixel masks, and auditing of the dataset structure and class distribution.

Scientific and practical contributions

- 5.5 A comprehensive evaluation was performed of the applicability, technical performance, and diagnostic performance of classification, segmentation, and multimodal artificial intelligence models for automated analysis of cervical colposcopic images.
- 5.6 A controlled experimental assessment was performed of the effects of key factors on segmentation performance, including multiclass versus binary formulation, class merging and exclusion, priority hierarchy for overlapping annotations, oversampling of rare classes, photometric augmentation, and local cropping around findings.
- 5.7 The development has been submitted for patent protection to the United States Patent and Trademark Office, application No. 63/991,571, dated February 26, 2026.

“If a machine is expected to be infallible, it cannot also be intelligent.”

Alan Turing

February 20, 1947

Declaration of Originality

The author of this dissertation declares that all data described herein are original and were obtained as a result of his research work at Medical University – Pleven and University Hospital “St. Marina – Pleven.” The results, discussions, and conclusions have not been taken from other sources without appropriate citation.

Declarant: Georgi Danielov Prandzhev, MD

6. SCIENTIFIC PUBLICATIONS AND PARTICIPATION

PUBLICATIONS RELATED TO THE DISSERTATION

1. **Prandzhev GD**, Gortchev GA, Dimitrov DD, Miltchev RI, Tomov ST. Development of an Artificial Intelligence Algorithm for Colposcopic Recognition of Cervical Features - Introduction of an Annotation Protocol. *Balkan Med J.* 2025 Jul 1;42(4):373-375. doi: 10.4274/balkanmedj.galenos.2025.2024-12-35.
2. **Prandzhev, G. D.**, Gortchev, G. A., Dimitrov, D. D., Miltchev, R. I., Rangelov, D., Tomov, S. T. (2025). Artificial intelligence analysis of the transformation zone of the uterine cervix. *Journal of Biomedical and Clinical Research*, 18, 47-53. doi:10.3897/jbcr.e144006.
3. Vasileva, M., **Prandzhev, G.**, Dimitrov, D., Gorcheva, Z. (2024). Current state of application of Artificial Intelligence in preoperative MRI assessment of endometrial cancer – a mini review. *Series on Biomechanics*, 38(4), 127–132. <https://doi.org/10.7546/SB.17.04.2024>.

SCIENTIFIC PRESENTATIONS RELATED TO THE DISSERTATION

1. June 19–21, 2025 – Second National Congress of the Bulgarian Association of Minimally Invasive and Robotic Surgery (BAMIRS), with international participation; poster presentation: “Application of Artificial Intelligence in Contemporary Colposcopy – First Clinical Experience,” Medical University – Pleven, Telecommunication Endoscopic Center (TELEC), Pleven, Bulgaria.
2. November 3, 2024 – Jubilee Scientific Conference “50 Years of Medical Education and Science in Pleven”; oral presentation: “Application of Artificial Intelligence for the Detection of Precancerous Cervical Lesions,” Pleven, Bulgaria.
3. June 15, 2024 – Eleventh Annual Scientific and Practical Conference of the Association of Otorhinolaryngologists of Southeastern Bulgaria; presentation: “Application of Artificial Intelligence in Oncology,” Kiten, Bulgaria.
4. May 10–12, 2024 – First National Conference “Pleven Minimally Invasive Days” of the Bulgarian Association of Minimally Invasive and Robotic Surgery (BAMIRS), with international participation; presentation: “Artificial Intelligence in Minimally Invasive Gynecologic Surgery,” Telecommunication Endoscopic Center (TELEC), Medical University – Pleven, Pleven, Bulgaria.
5. May 14, 2024 – “Health Barometer” civic initiative under the patronage of Prof. Daniel Valchev, Dean of the Faculty of Law at Sofia University “St. Kliment Ohridski”; presentation: “Application of Artificial Intelligence in Gynecology,” InterContinental Hotel, Sofia, Bulgaria.
6. November 12, 2023 – 14th Annual SONM – MORE 2023 Conference, organized by the Joint Oncology National Network (SONM);

presentation: “A Revolution in Cervical Cancer Screening” as part of the demo session “Panorama of Health Technologies in Bulgarian Oncology,” Varna, Bulgaria.

7. June 4, 2022 – Plenary lecture: “Artificial Intelligence in Medicine” at the Ninth National Conference of BAMIGS, Medical University – Pleven, TELEC.

Project BG-175467353-2022-04-0257-C01, “Application of ARTIFICIAL INTELLIGENCE Methods for Colposcopic Recognition and Categorization of Cervical Characteristics – IntelCAMat,” funded

1. October 31, 2022 – Winner of the European Institute of Innovation and Technology (EIT) Health i-Days, organized at Medical University – Pleven, receiving the opportunity to participate in the international EIT Health Europe competition. i-Days is a collaboration between Junior Achievement Bulgaria and its spin-off company The Edge: R&BD, in partnership with EIT Health and universities across Bulgaria.
2. November 24–25, 2022 – Finalist in the above-mentioned international European Institute of Innovation and Technology (EIT) competition in

Vienna, Austria, with participants from 24 countries; reached the final, placed third, and received a financial award.

3. June 2, 2023 – First place in the Beyond Pre-Accelerator entrepreneurship program, organized by The Edge and Junior Achievement Bulgaria, for a project aimed at the market implementation of the dissertation research: “Application of Artificial Intelligence for Early Detection of Cervical Precancer.”
4. July 14, 2023 – Recipient of the Intel AI Innovators Award 2023, presented by Intel Corporation® for the development “Artificial Intelligence for Early Detection of Cervical Precancer” and its presentation at the international GEN-E 2023 forum in Istanbul, Türkiye, organized by Junior Achievement Europe.
5. September 2023 – Recipient of the Intel® AI Global Impact Festival 2023 Country Recognition Award, “Cervical Pre-Cancer Detection,” a regional recognition for Bulgaria for contributions to the development of artificial intelligence for the detection of precancerous cervical lesions.
6. September 4–6, 2023 – Official travel pursuant to Order No. RD-19-46 / August 28, 2023, to Salamanca, Kingdom of Spain, with the assignment: “Accompanying the Minister of Innovation and Growth to Europe’s largest technology event, Startup Olé 2023.” Third place in the Humanitarian Technologies/Social Impact category of the presentation competition, with a presentation on the application of artificial intelligence for early detection of cervical precancer.

7. November 5, 2025 – Award ceremony in Brussels following nomination to FERD’s List, Healthcare sector, recognizing inspiring young entrepreneurs and leaders in Europe. FERD’s List is an initiative of Junior Achievement Europe in collaboration with FERD, designed to honor the achievements, entrepreneurial journeys, and inspiring leadership of JA alumni.

1. December 2023 – FORBES Bulgaria “30 Under 30” interview article, “Artificial Intelligence for Cancer Diagnosis,” issue 12/2023, pp. 62–62.

2. January 31, 2024 – BNT 2 television interview for the program Znanie.BG on the topic of the dissertation, included in a feature marking 50 years of Medical University – Pleven (04:00–06:00 min.).

3. February 8, 2024 – Bloomberg TV Bulgaria, 25-minute television interview presenting the scientific concept developed in the dissertation, on the program “The World Is Business,” hosted by Ivaylo Lakov.

4. June 16, 2024 – Bulgaria ON AIR, 10-minute television interview on the topic of the dissertation: “Artificial Intelligence Helps in the Early Diagnosis of Cancerous Conditions.”