

МЕДИЦИНСКИ УНИВЕРСИТЕТ – ПЛЕВЕН
КАТЕДРА „АКУШЕРСТВО И ГИНЕКОЛОГИЯ“

Д-р Георги Даниелов Пранджев

**„Място на изкуствения интелект в съвременната
колпоскопска диагностика на предракови изменения на
маточната шийка“**

АВТОРЕФЕРАТ

на дисертация за присъждане на образователна и научна степен
„Доктор“

Научна специалност „Акушерство и гинекология“

Научни ръководители:

Чл.-кор. проф. д-р Славчо Томов Томов, д.м.н.

Акад. д-р Григор Ангелов Горчев, д.м.н.

Официални рецензенти:

Проф. д-р Явор Димитров Корновски, д.м.н.

Проф. д-р Станислав Христов Славчев, д.м.н.

Плевен, 2026 г.

Дисертационният труд е написан на 182 стандартни печатни страници. Материалът е онагледен с 41 фигури и 20 таблици. Библиографията обхваща 115 литературни източника, от които 2 на кирилица и 113 на латиница.

Авторът е докторант редовна форма на обучение към катедра „Акушерство и гинекология“, Факултет по Медицина към Медицински университет – Плевен.

Дисертационният труд е обсъден и одобрен за публична защита на разширен Катедрен съвет на катедра „Акушерство и гинекология“, Факултет по Медицина, Медицински университет – Плевен, състоял се на 30.07.2026 г.

Забележка: номерацията на фигурите и таблиците не отговаря на тази в дисертационния труд.

Публичната защита на дисертационния труд ще се състои на 07.10.2026 г. от 13:00 часа в зала „Луи Пастьор“, Медицински колеж, факултет "Фармация" на Медицински университет – Плевен.

Научно жури в състав:

Председател: Доц. д-р Добринка Христова Гинчева, д.м.

Членове:

Проф. д-р Явор Димитров Корновски, д.м.н.

Проф. д-р Станислав Христов Славчев, д.м.н.

Проф. д-р Елена Димитрова Димитракова, д.м.

Доц. д-р Илия Илков Карагъзов, д.м.

Доц. д-р Добринка Христова Гинчева, д.м.

Резервни членове:

Доц. д-р Чавдар Атанасов Цветков, д.м.

Доц. д-р Павел Петров Добрев, д.м.

Материалите по защитата са публикувани на интернет страницата на МУ–Плевен (www.mu-pleven.bg) и са на разположение на интересуващите се в Научен отдел на Факултет по Медицина, МУ-Плевен.

СЪДЪРЖАНИЕ

СПИСЪК НА ИЗПОЛЗВАНИТЕ СЪКРАЩЕНИЯ

1. **ВЪВЕДЕНИЕ**
 - 1.1 Епидемиология
 - 1.2 Изкуствен интелект и ограничения в колпоскопията
2. **ЦЕЛ И ЗАДАЧИ**
 - 2.1 Цел
 - 2.2 Задачи
3. **МАТЕРИАЛИ И МЕТОДИ**
 - 4.1 Обща характеристика на клиничния контингент
 - 4.2 Колпоскопска апаратура и методика на изследването
 - 4.3 Подготовка и анотация на колпоскопските изображения
 - 4.4 Архитектура на невронната мрежа
 - 4.5 Хардуерна и софтуерна среда
 - 4.6 Статистически методи и метрики
4. **СОБСТВЕНИ РЕЗУЛТАТИ И ОБСЪЖДАНЕ**
 - 4.1 **ПЪРВА ЗАДАЧА**
 - 4.1.1 Архитектура на модела и експериментален дизайн
 - 4.1.2 **Първа група експерименти** за многокласова сегментация
 - 4.1.2.1 Базов експеримент E000 (оригинална 5-класова сегментация)
 - 4.1.2.2 Експеримент E212 (многокласова + увеличено представяне на редки класове (oversampling))
 - 4.1.3 **Втора група експерименти** за бинарна (foreground) сегментация
 - 4.1.3.1 Експеримент E202 (Foreground + повишено представяне на редки класове (oversampling))
 - 4.1.4 **Трета група експерименти** с алтернативни целеви формулировки
 - 4.1.5 **Обсъждане**
 - 4.1.5.1 Оценка на модела и интерпретация на резултатите
 - 4.1.5.2 Foreground срещу multiclass подход
 - 4.1.5.3 JNEG сигнал и протокол-специфичност
 - 4.1.5.4 Роля на аугментациите и пресемплирането
 - 4.2 **ВТОРА ЗАДАЧА**
 - 4.2.1 Формат на данните и анотационен протокол
 - 4.2.2 Одит на данните и класово разпределение
 - 4.2.3 Структура на набора от данни и подмножества за експериментите
 - 4.2.4 **Обсъждане**
 - 4.3 **ТРЕТА ЗАДАЧА**
 - 4.3.1 **Фактори, свързани с данните**
 - 4.3.1.1 Класов дисбаланс и обем на данните
 - 4.3.1.2 Варибилност между поднаборите и липса на групиране по пациент

- 4.3.1.3 Протокол-специфичен сигнал и липса на метаданни
- 4.3.1.4 Ефект от аугментационните стратегии
- 4.3.2 **Фактори, свързани с анотациите**
- 4.3.3 **Фактори, свързани с модела**
- 4.3.3.1 Целева формулировка – двоична срещу многокласова сегментация
- 4.3.3.2 Калибрация и софтвакс увереност
- 4.3.4 **Обсъждане**
- 4.4 **ЧЕТВЪРТА ЗАДАЧА**
- 4.4.1 Мост между сегментационни и класификационни метрики
- 4.4.2 Сравнение с литературни данни
- 4.4.2.1 Сравнение по ROC-AUC профил с описани в литературата системи
- 4.4.2.2 Сегментационен анализ и сравнение по Dice стойности с описани в литературата системи
- 4.4.3 **Обсъждане**
- 4.4.3.1 Размер и класова фрагментация на набора от данни
- 4.4.3.2 Анотационна субективност и концептуално припокриване на класовете
- 4.4.3.3 Липса на метаданни и протокол-зависим сигнал
- 4.5 **ПЕТА ЗАДАЧА**
- 4.5.1 Асистирана колпоскопска оценка
- 4.5.2 Биопсично насочване
- 4.5.3 Обучение и стандартизация
- 4.5.4 Подобряване на достъпа в слаборесурсни условия
- 4.5.5 **Обсъждане**

5. ПРИНОСИ

6. НАУЧНИ ПУБЛИКАЦИИ И УЧАСТИЯ

СПИСЪК НА ИЗПОЛЗВАНИТЕ СЪКРАЩЕНИЯ

AI – Artificial Intelligence – изкуствен интелект (ИИ)
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 – дълбоко обучение
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 – Глобална база данни към 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. ВЪВЕДЕНИЕ

„Вярвам, че до края на века използваните думи и общоприетите възгледи на образованите хора ще се променят дотолкова, че ще можем да говорим за мислещи машини, без това да предизвиква възмущение.“

Алън Тюринг, 1950 г.

1.1 Епидемиология

Ракът на маточната шийка е четвъртото по честота злокачествено заболяване при жените в световен мащаб. През 2022 г. са регистрирани 662 301 нови случая и 348 874 смъртни случая. [2] При запазване на настоящите тенденции в заболяемостта, до 2050 г. прогнозните модели предвиждат между 948 000 и 973 000 новозабоболели и между 401 000 и 543 000 починали годишно. [3] В България болестта засяга 15.4 на 100 000 жени [4], като държавата попада в най-високата категория по честота на заболяването в Европейския съюз и се нарежда на второ място по смъртност [5]. За 2024 г. Националният статистически институт отчита средно 9.6 починали на 100 000 жени с изразени регионални различия. [6] Човешкият папиломен вирус (HPV) представлява семейство от над 200 типа, от които 40 засягат аногениталния тракт. Инфекцията с високорискови щамове е основният етиологичен фактор за развитието на карцином на маточната шийка. [7]

През 2020 г. Световната здравна организация прие Глобална стратегия за ускорено елиминиране на рака на маточната шийка. [8] Целта е до 2030 г. да се постигне и поддържа праг под 4 нови случая на 100 000 жени годишно, посредством три основни задачи: 90% от момичетата до 15-годишна възраст да бъдат напълно ваксинирани с HPV ваксина; на 70% от жените да се извърши скрининг с високоефективен тест до 35- и отново до 45-годишна възраст; и 90% от жените с установени преканцерозни или инвазивни изменения да получат подходящо лечение.

Покритието с HPV ваксина в България е изключително ниско – през 2019 г. първа доза са получили едва 7.2% от момичетата на 13 години, а втора доза само 11.9 % от 12-годишните. [9] За периода 2020-2024 г. се наблюдава обезпокоително ниско ниво на ваксинация - около 1-2% от всички момичета на възраст 10-14 години. [10] През 2025 г. беше приета нова Национална програма за първична профилактика на раковите заболявания, причинени от HPV (2025-2030), която за първи път включва момчета във ваксинационния план и разширява възрастовите групи. [11]

По отношение на вторичната профилактика България се нарежда сред страните с най-нисък обхват. През 2019 г. едва 45% от българските жени на възраст 20–69 години са съобщили за проведен Рар тест през последните три години, при средно европейско ниво от 59.9%, докато HPV генотипизирането остава нереимбурсирано от НЗОК. [12]

България продължава да бъде сред водещите страни в региона по отношение на негативната статистика на рака на маточната шийка. Първичната профилактика чрез ваксинация не постига достатъчен обхват и явно контрастира с поставената от СЗО цел. Само три държави в Европа са постигнали 90% ваксинационно покритие сред момичетата до 15-годишна възраст, но напредъкът остава неравномерно разпределен в рамките на ЕС. [13] В същото време вторичната профилактика е затруднена от ниско популационно покритие и от липсата на организирана национална скринингова програма. [14] Колпоскопското изследване, от друга страна, остава до голяма степен силно зависимо от субективната интерпретация. Множество проучвания демонстрират значителна междунаблюдателна вариабилност и относително ниска възпроизводимост дори между експертни колпоскописти, със съгласие в около 56–60% от случаите. [15]

HPV-асоциираният карцином на маточната шийка е почти напълно предотвратим чрез ваксинация, но покриването на поставените цели ще даде резултат след около 40 години. Постигането на 90% ваксинация до 2030 г. ще намали смъртността с едва около 0.1%, а двукратният скрининг и лечение би предотвратил 300 000 смъртни случая (34.2%). До 2070 г. ваксинацията самостоятелно би намалила смъртността с приблизително 61.7%, а комбинираната стратегия - с 88.9% при еднократен и с 92.3% при двукратен скрининг през живота, което се равнява на близо 15 милиона спасени женски живота. [16]

До края на века ще наблюдаваме пълното разгръщане на тези целенасочени усилия, но до тогава вниманието следва да бъде насочено към популациите, които нямат възможност да се възползват от ваксинационния потенциал. Развитието на изкуствения интелект (artificial intelligence - AI) в здравеопазването и най-вече в анализа на медицински изображения създава възможности да потърсим решение още днес, очаквайки и подпомагайки Стратегията на СЗО. Внедряването на интелигентни системи за анализ може да подобри диагностичната прецизност в реално време, компенсирайки ограниченията на настоящите профилактични подходи. [17]

1.2 Изкуствен интелект и ограничения в колпоскопията

Изкуственият интелект представлява широка научна и инженерна област, насочена към създаване на системи, способни да изпълняват задачи, които традиционно изискват интелигентно поведение. В класическия учебник *Artificial Intelligence: A Modern Approach* авторите подчертават, че ИИ не се ограничава до разбирането на интелигентността, а цели и изграждането на интелигентни системи. [22] В техния подход централно място заема понятието „рационален агент“ - система, която възприема средата си и избира действия, насочени към постигане на най-добрия възможен или очакван резултат при дадените условия. В най-общ смисъл ИИ може да бъде разглеждан като широка категория, обединяваща различни подходи за възприемане, учене, планиране и адаптация към нова информация.

В медицинската образна диагностика особено значение имат машинното обучение (machine learning, ML) и дълбокото обучение (deep learning, DL). В класическите подходи на машинно обучение изследователят обикновено дефинира предварително характеристиките, които моделът трябва да анализира - например цвят, текстура, форма, интензитет или морфологични параметри. След това тези характеристики се използват за класификация, прогнозиране или оценка на диагностичен резултат. Дълбокото обучение е специализиран подтип на машинното обучение и обучението на представяния (representation learning), основан на модели с множество последователни слоеве. Неговото основно предимство при анализ на изображения е способността автоматично да извлича йерархични визуални представяния директно от входните данни. В ранните слоеве мрежата разпознава базови признаци като ръбове, контури, цветови преходи и текстури, а в по-дълбоките слоеве - по-сложни морфологични и пространствени структури. Това е особено релевантно за колпоскопията, където диагностичната информация се съдържа в комбинация от фини цветови, съдови, повърхностни и пространствени характеристики.

В тази рамка компютърното зрение не се разглежда като подмножество на дълбокото обучение, а като широка област на приложение на изкуствения интелект, насочена към обработка и интерпретация на изображения. Авторите го обсъждат в раздела за приложения на дълбокото обучение, като подчертават, че компютърното зрение традиционно е една от най-активните области за приложението му. Следователно то следва да бъде представяно не като част от самата методологична йерархия AI-ML-DL, а като приложна област, която използва методи на машинно и дълбоко обучение.

Qin и сътр. (2023) извършват систематичен обзор и мета-анализ на 15 проучвания с общо 22 764 участници, използващи IFCPC 2011

терминологията. [24] При праг LSIL+ колпоскопията показва висока чувствителност 92%, но ниска специфичност – 51%. При праг HSIL+ чувствителността намалява до 68%, а специфичността се повишава до 93%.

Подобен компромис е отчетен и в мащабното мултицентрично ESTAMPA проучване, което оценява стандартизирана колпоскопия като триажен метод при HPV-позитивни жени. Valls и сътр. (2023) съобщават чувствителност 91.2% за откриване на CIN3+, но специфичност 50.1% за <CIN2. [25] Това потвърждава, че дори при стандартизиран подход колпоскопията може да бъде високо чувствителна за високостепенни лезии, но остава ограничена по отношение на специфичността, което води до риск от фалшиво положителни оценки и допълнителни биопсии.

Методологичната сложност при оценката на колпоскопията е разгледана от Brown и Tidy (2019), които показват, че публикуваните стойности за точност зависят от начина, по който се дефинира положителния резултат. [26] Когато се оценява колпоскопското впечатление за наличие на CIN2+, претеглената средна чувствителност е 68.5%, специфичността 75.9%, а PPV 68.9%. Когато обаче положителният резултат се дефинира като наличие на видима болест с последваща биопсия, чувствителността нараства до 95.7%, но специфичността спада до 34.2%, а PPV до 54.3%. Това показва, че диагностичната точност на колпоскопията не е фиксирана характеристика на метода, а зависи от зададения диагностичния праг.

Li и сътр. (2021) анализират съгласието между колпоскопското впечатление и хистологичната диагноза при 495 пациентки, изследвани по IFPCPC 2011 терминология. [28] Авторите установяват пълно съвпадение само в 46.9% от случаите като заключават, че колпоскопската оценка по-често подценява, отколкото надценява тежестта на лезиите, особено при случаи на HSIL и карцином.

Ограниченията на колпоскопията не се изчерпват с визуалната интерпретация, а включват и биопсичната стратегия. Underwood и сътр. (2012) показват в системен обзор и мета-анализ, че колпоскопски насочената биопсия има висока обединена чувствителност за CIN2+ - 91.3%, но много ниска специфичност - 24.6%. [29] Wentzensen и сътр. (2015) допълнително демонстрират, че една насочена биопсия е недостатъчна: чувствителността за откриване на HSIL се повишава от 60.6% при една биопсия до 85.6% при две и 95.6% при три биопсии. [30]

Този проблем има и пространствено измерение. Ако биопсията не попадне в зоната с най-тежка дисплазия, лезията може да бъде подценена, независимо че визуално е разпозната като патологична. Следователно диагностичната стойност на колпоскопията зависи не само от това дали лекарят разпознава абнормна находка, но и от това дали успява точно да я локализира върху повърхността на маточната шийка. Това поставя акцент

върху една по-специфична клинична задача - не само да се определи дали изображението е подозрително, а да се посочи къде се намира зоната с най-голяма вероятност за високостепенна лезия.

Едно от основните ограничения на конвенционалната колпоскопия е значимата междуоператорна вариабилност при оценката на колпоскопските находки. Съвременни данни от проучвания за осигуряване на качеството потвърждават, че този проблем остава актуален. Origoni и сътр. (2023) оценяват диагностичното представяне на колпоскописти с различен опит при интерпретация на дигитални случаи и установяват обща чувствителност 61.6% и специфичност 77.1%. [34] При праг CIN2+ специфичността се повишава до 87.7%, но колпоскопските впечатления са свързани с около 20% подценяване на CIN2+ лезии. Това подкрепя необходимостта от систематичен контрол на качеството и от по-обективни инструменти за подпомагане на колпоскопската интерпретация.

В обобщение, литературата потвърждава, че конвенционалната колпоскопия остава основен, но ограничен визуален метод за оценка на предраковите изменения на маточната шийка. Нейната диагностична ефективност не е постоянна величина, а зависи от изборния праг на интерпретация, дефиницията на положителен резултат, биопсичната стратегия, видимостта на трансформационната зона, техническото качество на изображението и опита на изследващия. Колпоскопията не е само задача за разпознаване, а и задача за пространствена локализация. Представянето ѝ зависи както от способността на колпоскописта да установи абнормна находка, така и от уменията му да идентифицира най-сuspектната зона и да насочи биопсията към нея. Ниското съгласие при оценка на основни колпоскопски белези показва, че субективността е структурна характеристика на метода, а не само резултат от индивидуален опит. Тези фактори ограничават възпроизводимостта на колпоскопската оценка и затрудняват създаването на универсални диагностични критерии, приложими в различни клинични условия.

В контекста на колпоскопията компютърното зрение представлява основното приложно направление на изкуствения интелект, чрез което дигиталните изображения на маточната шийка могат да бъдат анализирани автоматизирано. Тази област позволява извличане и интерпретация на визуални характеристики като цвят, текстура, съдови промени, оцетнобял епител, мозайка, пунктуация и йод-негативни участъци. В зависимост от поставената задача AI системите в колпоскопията могат да бъдат насочени към три основни типа анализ. Първият е класификацията, при която цялото изображение се отнася към определена диагностична категория. Вторият е детекцията, при която моделът не само разпознава наличие на патологична находка, но и локализира приблизителната ѝ позиция в изображението чрез ограничителни рамки или региони на интерес. Третият и най-прецизен

подход е пикселната сегментация, при която всеки пиксел от изображението се отнася към конкретен клас, което позволява детайлно очертаване на отделните колпоскопски находки и пространствената им връзка. В този смисъл компютърното зрение осигурява методологичната основа за преминаване от субективна визуална оценка към количествен, възпроизводим и анатомично локализиран анализ на колпоскопските изображения.

2 ЦЕЛ И ЗАДАЧИ

2.1 ЦЕЛ

Като приложим методите на изкуствен интелект за анализ на колпоскопски изображения, да оценим неговото клинично значение за детекция на предракови изменения на маточната шийка и да определим мястото му в съвременната колпоскопска диагностика.

2.2 ЗАДАЧИ

1. Да разработим алгоритъм, базиран на изкуствен интелект, използвайки методите на машинно и/или дълбоко обучение, за автоматизиран анализ на колпоскопски изображения.

2. Да създадем стандартизирана база данни от колпоскопски изображения за интеграция в разработения модел чрез единен протокол за заснемане и анотация, съобразен с медицинските стандарти за колпоскопска номенклатура.

3. Да идентифицираме и анализираме факторите, които влияят върху сегментационната производителност на модела.

4. Да определим диагностичната стойност на разработения модел чрез сравнителен анализ с утвърдени в литературата системи.

5. Да анализираме клиничната приложимост и мястото на изкуствения интелект в съвременната колпоскопска диагностика.

3. МАТЕРИАЛ И МЕТОДИ

3.1. Обща характеристика на клиничния контингент

За целите на дисертационния труд е използвана собствена база данни от колпоскопски изображения, натрупана проспективно при проследяването на 94 жен на възраст между 18-64 години, преминали през УМБАЛ „Света Марина – Плевен“ в периода 2022-2025 г. Към образните данни е събрана информация, включваща възраст, брой раждания, цитологичен статус, HPV статус и/или генотипизиране, тип трансформационна зона (TZ), наличие на абнормни колпоскопски находки, патохистологичен резултат.

3.2. Колпоскопска апаратура и методика на изследването

Колпоскопските изследвания са извършени с дигитален колпоскоп Leisegang Digital Colposcope, оборудван с регулируемо LED осветление с интензитет 45000-52000 lux и цветна температура 5700-6000 K. Устройството осигурява до 30-кратно оптично увеличение и изображение с резолюция 1280 x 960 пиксела.



Фигура 2. Дигитален колпоскоп Leisegang, използван в настоящото проучване.

За отбелязване на колпоскопските находки е използвана терминологията на *International Federation for Cervical Pathology and Colposcopy (IFCPC)*, съгласно критериите за абнормни епителни и съдови изменения. (Фиг. 3) (Bornstein et al., 2012)



International Federation for Cervical Pathology and Colposcopy
Internationale Federation für Zervixpathologie und Kolyposkopie
Fédération Internationale de Pathologie Cervicale y Colposcopie
Federația Internațională de Patologie Cervicală și Colposcopie

2011 IFCPC Nomenclature¹

Accepted in Rio World Congress, July 5, 2011
Nomenclature Committee chairman: Jacob Bornstein MD

2011 IFCPC colposcopic terminology of the cervix ¹			
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 (keratosis, 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

Фигура 3. Колпоскопска терминология спрямо IFCPC 2011

За нуждите на анотацията данните бяха етикетирани със съкратени кодове: WE - бял епител, М - мозайка, Р - пунктуация, L - левкоплакия, JNEG - йод-негативен участък, съгласно абривиатурата на Б. Василев. (Фиг. 4)

РИМСКА НОМЕНКЛАТУРА	АБРЕВИАТУРИ ЗА „КОЛПОСКОПСКА КАРТА НА МАТОЧНАТА ШИЙКА“
Нормален (типичен, обичаен) колпоскопски образ	N
Многослоен плосък невроговяващ епител	N-Sq
Цилиндричен епител	N-Gl
Нормална зона на прехода	N-TZ
Абнормен (атипичен, необичаен) колпоскопски образ	A
Бял епител сака третиране с оцетна киселина *	A-WF **
• с плоска повърхност	A-WF-Pi **
• с микропапилярна или микропапилярна повърхност	A-WF-R **
Пунктирна зона (пунктуация) *	A-P **
Мозайка *	A-M **
Левкоплакия *	A-L **
Йод-негативен участък	A-J (-)
Атипични съдове	A-V **
СЪМНЕНИЕ ЗА ИНВАЗИВЕН РАК	C ?
НЕЗАДОВОЛИТЕЛНА КОЛПОСКОПИЯ	
Границата плосък/цилиндричен епител не се проследява изцяло	UNSAT
Теско възпадение или тежка атрофия	
Не се визуализира маточна шийка	
РАЗНИ	
Микропапилярна повърхност, която не е бял епител	R
Екзофитни кондиломи	Condyl
Външалнис	I
Атрофия	At
Ява	U
Друго (описание)	

Фигура 4. Колпоскопска номенклатура по проф. Божил Василев

3.3 Подготовка и анотация на колпоскопските изображения

След извършване на колпоскопското изследване, получените изображения се архивират в преносим компютър и се отпечатват към пациентски протокол, върху който експерт-колпоскопист извършва първично ръчно ограждане на абнормните участъци. След дигитализация изображенията се импортират в *Supervisely* - специализирана платформа за анотация на медицински изображения и подготовка на бази данни за машинно обучение. Анотацията се извършва повторно чрез инструмент, с помощта на който се очертават границите на лезиите точка по точка при високо увеличение на ниво пиксел. (Фиг. 5)



Фигура 5. Изглед на анотационния процес в *Supervisely*

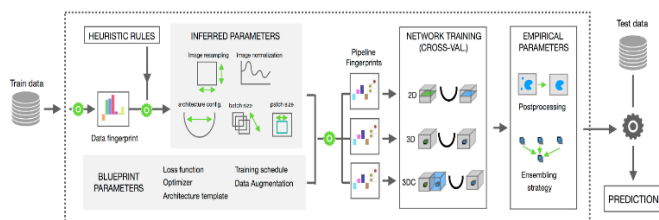
Общият брой изображения е 355, като за всяко от тях е наличен съответен анотационен файл, съдържащ координатите на полигонално оградените находки. В рамките на базата данни са идентифицирани общо 625 полигона. (Табл.1) Полигоните представляват експертно маркирани региони, получени чрез ръчно очертаване на колпоскопски находки върху изображенията. Всеки от тях дефинира пространствените граници на определена зона, която представлява колпоскопски клиничен интерес. Всички оригинални изображения и анотации са използвани без допълнителна ръчна модификация или промяна на суровите данни.

Параметър	Резултат
Брой изображения	355
Брой анотационни файлове	355
Брой полигони	625

Таблица 1. Основни параметри при одита на данните

3.4 Архитектура на невронната мрежа

За анализ на колпоскопски изображения беше избрана deep learning архитектурата nnU-Net v2 с конфигурация Residual Encoder U-Net M (ResEnc-M). (Фиг. 6) Изборът на модела се базира на няколко ключови предимства, които го правят особено подходящ за медицински сегментационни задачи. На първо място, nnU-Net v2 представлява самоконфигурираща се рамка, способна автоматично да адаптира архитектурата, стъпките на предварителна обработка, стратегиите за аугментация и процеса на обучение спрямо характеристиките на конкретния набор от данни. Това позволява минимизиране на ръчната оптимизация и осигурява висока възпроизводимост на резултатите.



Фигура 6. Схема на nnU-Net v2 архитектурата

Архитектурата е разработена специално за сегментация на биомедицински изображения и постига резултати на водещо световно ниво в редица международни състезания върху медицински набори от данни. Основно предимство на nnU-Net е способността му да се адаптира към бази данни с ограничен размер и значителна хетерогенност, което са характерни особености на колпоскопските изображения. Използваният Residual Encoder подобрява извличането на признаци, което улеснява обучението на по-дълбоки невронни мрежи. Това е особено важно при идентифицирането на фино изразени абнормни находки. (Isensee et al., 2021)

3.4 Хардуерна и софтуерна среда

Експериментите са проведени под Ubuntu 24.04.4 LTS на сървър с 4x NVIDIA H100 GPU (80 GB VRAM). Използвана е среда Python (Conda, PyTorch 2.6.0, CUDA 12.4), базирана на архитектурата 2D nnU-Net v2 Residual Encoder U-Net. Обучението е извършено по folds с разделяне на данните на ниво пациент против data leakage. Допълнителната обработка и анализ са реализирани с научни Python библиотеки. Разработени са

персонализирани скриптове за оценка на базовите метрики (Dice, IoU, ROC-AUC, confusion matrices и др.) и анализ на softmax confidence.

3.5 Метрики и статистическа оценка

Качеството на сегментационния модел се оценява според степента на припокриване между предсказаната пикселна маска и експертната анотация. Обикновената точност на ниво изображение не представлява надежден обобщаващ показател, тъй като по-голямата част от колпоскопските снимки се състои от фон или нетаргетна тъкан. По този начин моделът може да демонстрира привидно висока точност чрез предсказване предимно на бекграунда (фона), като същевременно пропуска клинично значими абнормни находки. В настоящото изследване е използвана следната статистическа терминология с медицинско и техническо значение:

Термин	Значение	Интерпретация
Dice	Основен показател за припокриване при сегментация. Сравнява областта, очертана от експерта, с областта, предсказана от модела.	Стойност 1.0 означава перфектно припокриване, а стойност 0.0 - липса на припокриване.
IoU / Jaccard	Друг показател за припокриване, отразяващ пространственото припокриване между експертната анотация и предсказаната от модела област, изчислен като отношение между тяхното сечение и обединение	По-строг показател от Dice, полезен за сравнение между модели, но по-слабо интуитивен за читатели без ML подготовка.
Precision	От всички области, маркирани от модела като лезия/клас, колко действително съответстват на експертната анотация?	Ниска precision стойност означава наличие на много фалшиво-положителни области.
Recall (sensitivity)	От всички области, анотирани от експерта, колко са открити от модела?	Нисък recall означава, че моделът пропуска реални патологични области.

Термин	Значение	Интерпретация
Macro Dice	Средна Dice стойност за всички foreground класове, като всеки клас има еднаква тежест.	Особено важен показател в настоящото изследване, тъй като редките класове не се „скриват“ в средната стойност на честите класове.
Foreground Dice	Dice стойност след обединяване на всички таргетни класове в една foreground област	Полезен за оценка дали моделът разпознава наличие на абнормна/ аотирана зона, но не оценява правилната подтипова класификация.
Pixel accuracy	Процент на всички пиксели, класифицирани правилно.	Не се използва като основен показател, тъй като изображения с преобладаващ background могат да демонстрират висока точност дори при пропускане на лезията.

Таблица 2. Статистическа терминология

За целите на настоящото изследване и приложените методи на изкуствен интелект е използвана следната специализирана терминология:

Термин	Значение за изследването
Семантична сегментация	Класификация на ниво пиксел. Моделът очертава област за всеки таргетен клас, вместо да предоставя заключение единствено на ниво изображение.
Маска	Изображение, при което всеки пиксел съдържа идентификатор на определен клас, например background=0, M=1, WE=2

Полигонална анотация	Експертно очертан контур, използван за създаване на обучаващата маска. Качеството на полигоналната анотация директно влияе върху качеството на модела.
Поднабор (Fold)	Един валидационен поднабор при кръстосана валидация (cross-validation). Резултатите от 5-fold анализ са по-надеждни в сравнение с 2-fold screening резултати.
Групово разделяне на данните	Изображенията със същия префикс в името на файла се разпределят в един и същ поднабор (fold) с цел ограничаване на изтичането на данни между обучаващия и валидационния набор.
Аугментация	Вариации на изображенията, използвани само по време на обучението - промени в яркостта, контраста, отблясъците, изрязването на части от изображението, завъртане. Те имат за цел повишаване устойчивостта на модела без промяна на валидационните изображения.
Повишено представяне на редките класове (Oversampling)	По-често представяне на редките класове по време на обучението. Това подпомага модела да „обръща внимание“ на редките класове, но не може да замести необходимостта от повече висококачествени анотации.
Softmax многокласова класификация	Моделът трябва да избере само един клас за всеки пиксел, т.е. всеки пиксел може да принадлежи само на един клас. Това може да бъде ограничение, когато отделни находки съществуват в рамките на по-обширни трета находка.

Таблица 3. Специализирана техническа терминология

4. СОБСТВЕНИ РЕЗУЛТАТИ И ОБСЪЖДАНЕ

4.1 ПЪРВА ЗАДАЧА: Да разработим алгоритъм, базиран на изкуствен интелект, използвайки методите на машинно и/или дълбоко обучение, за автоматизиран анализ на колпоскопски изображения.

4.1.1 Архитектура на модела и експериментален дизайн

За автоматизиран анализ на колпоскопски изображения беше изграден модел, базиран на архитектурата за дълбоко обучение 2D nnU-Net v2 Residual Encoder U-Net Medium (ResEnc-M). Тази архитектура е специализирана за пикселна сегментация - т.е. всеки пиксел от изображението се класифицира като принадлежащ към определен колпоскопски клас, което позволява прецизно очертаване на абнормните региони на маточната шийка. Работният процес се извършва директно с оригиналните цветни изображения от колпоскопа, без необходимост от ръчно извличане на текстурни или цветови характеристики. Задачата беше реализирана чрез софтуерната платформа nnU-Net v2, която автоматично определя оптималните настройки за обучение (като разделителна способност на входните данни, размер на изображенията и стратегия за предварителна обработка) въз основа на специфичните характеристики на работния набор от данни. Изборът на nnU-Net се основава на неговата доказана ефективност в множество медицински сегментационни задачи, включително при ограничени бази данни, благодарение на вградените механизми за подобряване на обучаващите примери и автоматично адаптиране към особеностите на конкретната задача.

В хода на експериментите бяха приложени няколко техники за подобряване на обучението. Извършените аугментации включваха изкуствени промени в яркостта, цветовете, контраста и остротата на обучаващите изображения, чрез които моделът се подготвя за различни условия на заснемане. Платформата nnU-Net v2 прилага автоматично стандартни геометрични аугментации като завъртания, мащабиране и оразмеряване, които подобряват устойчивостта на модела към различни позиции и ориентации на находките. Върху част от експериментите беше добавена фотометрична обработка - симулиране на различна осветеност, отблясъци (glare), размазване (blur), компресионни артефакти с промяна във формата на файла (JPEG) и потъмняване по краищата на изображението (vignetting). Редките класове бяха представени по-често в хода на обучителния процес (oversampling), за да се компенсира техният малък

брой и да се даде възможност на модела да ги разпознава по-добре. Допълнително беше приложено и изрязване на област около находката (сгор) с размер 768x768 пиксела, което принуждава модела да се фокусира върху малките лезионни региони.

Проведоха се общо 13 експеримента, които могат да бъдат разделени на три групи според изследвания въпрос:

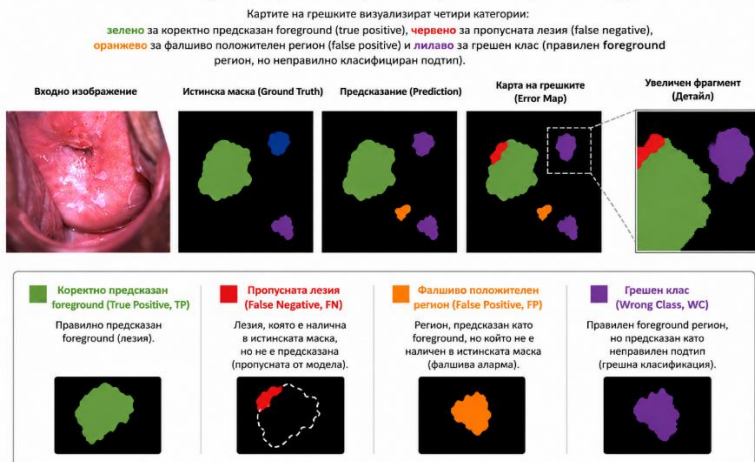
Първата група тества способността на модела да разпознава всички пет колпоскопски класа (мултикласова сегментация) - т.е. директно да класифицира всеки пиксел като мозайка, пунктуация, бял епител, левкоплакия или йод-негативен участък. Това беше най-амбициозната и с най-висока трудност клинична задача поради малкия обем данни и класовия дисбаланс. Тук са базовия експеримент E000 и неговите подобрени версии E102, E211, E212 и E213.

Втората група облекчи задачата, обединявайки всички колпоскопски находки в един общ клас. Така моделът се фокусира единствено върху откриване на абнормна зона, без да прави разлика между нейния подтип - бинарна (foreground) сегментация. Въпреки че този подход не предоставя информация за вида колпоскопска находка, той намалява сложността на задачата, поради което беше приложен в експериментите E101, E201, E202 и E203.

Третата група изследва алтернативни начини за дефиниране на самите целеви класове - например премахване на йод-негативните случаи от базата данни (E103) или работа само с такива изображения (E104), обединяване на визуално сходни признаци като М и Р в общ съдов клас (E105), или промяна на приоритета при припокриване на анотациите (E106). Тези експерименти имаха за цел да разберат доколко точното определяне на класовете влияе върху производителността на модела. Фигура 7 визуализира класирането на проведените експериментите според постигнатата от алгоритъма средната стойност на Dice коефициента, изчислен поотделно за всеки клас (Macro Dice).

За всеки експеримент са визуализирани три нива на припокриване между предсказаната от модела зона и реалната анотация: ниско (Dice = 0.0-0.2), средно (Dice = 0.3-0.6) и високо (Dice = 0.7-0.9). Картата на грешки (error map) визуализира това в четири категории цветове: зелено за коректно предсказан foreground (true positive), червено за пропуснат лезия (false negative), оранжево за фалшиво положителен регион (false positive) и лилаво за грешен клас (правилен foreground регион, но неправилно класифициран подтип). (Фиг. 8)

Визуализация на карти на грешките (Error map)



Фигура 8. Карта на грешките, визуализираща категориите на припокриване

4.1.2 Първа група експерименти за многокласова сегментация (Табл. 4)

ID	Формулировка	Промяна спрямо E000	Fold	Macro Dice	IoU
E000	Оригинална 5-класова (baseline)	Базов експеримент	5	0.154	0.120
E102	5-класова без празни анотации	E000 + премахнати празни анотации	2	0.242	0.193
E211	Многокласова + аугментации	E102 + фотометрични промени/отблясъци/размазване на образа	2	0.245	0.196

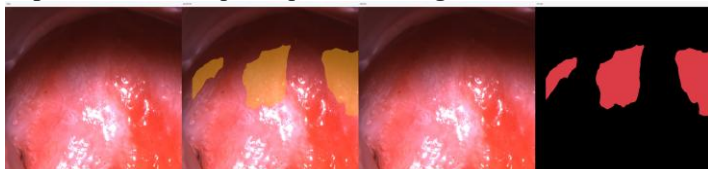
E212	Многокласова + увеличено представяне на редки класове (oversampling)	E102 + увеличено представяне на P/L	5	0.256	0.209
E213	Многокласова + изрязване (crops)	E102 + изрязване около P/L в резолюция 768x768	2	0.227	0.180

Таблица 4. Първа група експерименти

4.1.2.1 Базов експеримент E000 (оригинална 5-класова сегментация)

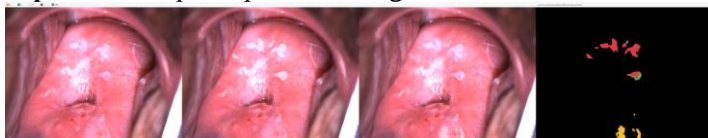
Първият проведен експеримент (E000) служи като отправна точка (*baseline*), спрямо която се измерва ефектът от всяка последваща промяна във формулировката, данните или стратегията за обучение. В него се използва най-основната възможна конфигурация: оригиналната 5-класова softmax целева формулировка (M, P, WE, L, JNEG + background) с всички 355 изображения, включително празните анотационни файлове, без oversampling, без аугментации и допълнителна обработка на данните. Важно е да се отбележи, че E000 е експеримент с 5-кратна крос-валидация, класифициран като потвърден резултат, с macro Dice = 0.154 ± 0.019 и macro IoU = 0.120. Качественият му анализ показва широк диапазон от случаи без никакво припокриване до такива със средна стойност на Dice = 0.675. (Фиг. 9-11)

Пример 1: Липсващо припокриване. Foreground Dice = 0.000.



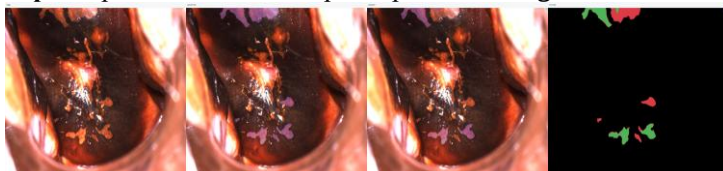
Фигура 9. Качествена визуализация на fold-0 валидация (експеримент E000).

Пример 2: Ниско припокриване. $\text{Foreground Dice} = 0.129$.



Фигура 10. Качествена визуализация на fold-0 валидация (експеримент E000).

Пример 3: Средно към високо припокриване. $\text{Foreground Dice} = 0.675$.



Фигура 11. Качествена визуализация на fold-0 валидация (експеримент E000).

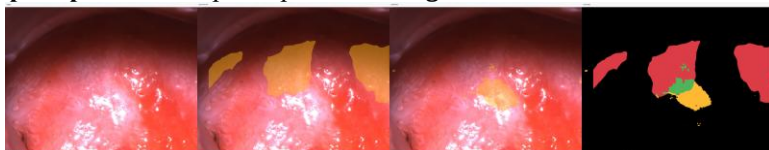
Тези резултати демонстрират, че дори базовият модел успява да разпознае някои абнормни региони, макар и непостоянно. Този широк диапазон е показателен за хетерогенността в данните и подчертава необходимостта от целенасочени подобрения, чийто ефект ще бъде количествено оценен именно спрямо стойностите на E000.

От това основно ниво последващите експерименти тестват различни стратегии за подобряване: премахване на празните анотации (E102), добавяне на аугментации (E211), повишено представяне (oversampling) на редките класове P и L (E212) и изрязване на по-малки региони около тях (E213). Най-добрият многокласов модел в групата беше E212 с macro Dice = 0.256 ± 0.039 и macro IoU = 0.209, като oversampling на P/L повишава Dice от 0.242 (E102) на 0.256.

4.1.2.2 Експеримент E212 (многокласова + увеличено представяне на редки класове (oversampling)) (Фиг. 12-14)

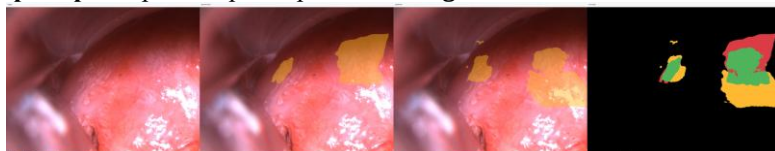
Последващите примери илюстрират трите нива на припокриване при E212 - от предизвикателен случай с нисък Dice = 0.124, през типичен със среден Dice = 0.542 и висококачественото предсказание с Dice = 0.862.

Пример 4: Ниско припокриване. $\text{Foreground Dice} = 0.124$.



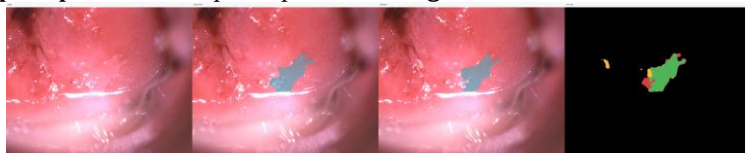
Фигура 12. Качествена визуализация на fold-0 валидация (експеримент E212).

Пример 5: Средно припокриване. $\text{Foreground Dice} = 0.542$.



Фигура 13. Качествена визуализация на fold-0 валидация (експеримент E212).

Пример 6: Високо припокриване. $\text{Foreground Dice} = 0.862$.



Фигура 14. Качествена визуализация на fold-0 валидация (експеримент E212)

4.1.3 Втора група експерименти за бинарна (foreground) сегментация (Табл. 5)

Базовият експеримент E101 постигна foreground Dice от 0.273. Приложените аугментациите (E201) и изрязванията около находките (E203) не доведоха до подобрение спрямо базовото ниво.

ID	Формулировка	Промяна спрямо E101	Folds	Foreground Dice
E101	Бинарна foreground	Базов експеримент	2	0.273
E201	Foreground + аугментации	E101 + фотометрични промени/отблясъци/размазване на образа	2	0.259

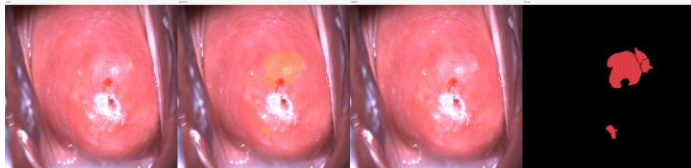
E202	Foreground + повишено представяне на редки класове (oversampling)	E101 + увеличено представяне на P/L	5	0.273
E203	Foreground + изрязване (crops)	E101 + изрязване около P/L в резолюция 768x768	2	0.263

Таблица 5. Втора група експерименти

Чрез допълнително представяне на редките находки P и L (E202) моделът запазва същия Dice от 0.273, но вече потвърден след 5-кратна валидация, което го прави най-надеждният foreground резултат (**Фиг. 15-17**):

4.1.3.1 Експеримент E202 (Foreground + повишено представяне на редки класове (oversampling))

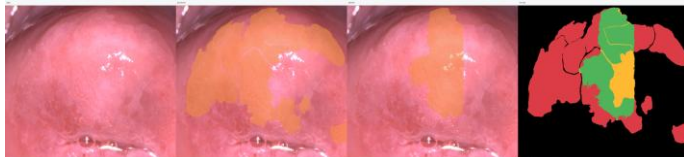
Пример 7: Ниско припокриване. Foreground Dice = 0.001.



Фигура 15. Качествена визуализация на fold-0 валидация (експеримент E202).

Труден, предизвикателен случай на изображение с минимално припокриване, при което моделът не успява да идентифицира абнормния регион. Характерен за снимки с малки, слабо изразени лезии.

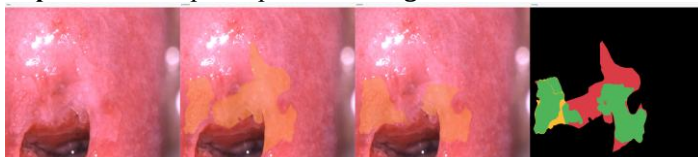
Пример 8: Средно припокриване. Foreground Dice = 0.430.



Фигура 16. Качествена визуализация на fold-0 валидация (експеримент E202).

Типичен случай на изображение с умерено припокриване, при което моделът идентифицира основния абнормен регион, но границите са неточни. Това ниво на производителност е представително за средностатистическия случай.

Пример 9: Високо припокриване. Foreground Dice = 0.732.



Фигура 17. Качествена визуализация на fold-0 валидация (експеримент E202).

Голяма, добре отличаваща се лезия с ясни граници, при която моделът постига много добро припокриване с експертната анотация. Този случай илюстрира горната граница на текущите възможности на модела.

4.1.4 Трета група експерименти с алтернативни целеви формулировки (Табл. 6)

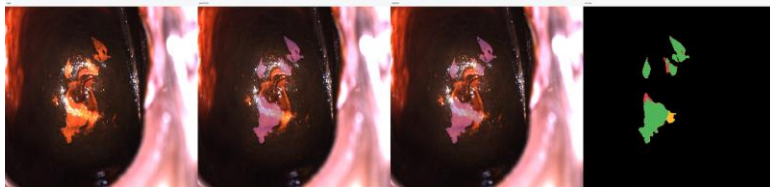
Премахването на йод-негативните случаи (E103, само M/P/WE/L) води до спад на маско Dice до 0.186, докато модел, обучен само върху JNEG (E104), постига Dice от 0.529. Обединяването на M и P в общ съдов клас (E105) дава Dice от 0.230. Промяната на приоритета на припокриване от $L > M > P$ на $M > P > L$ (E106) води до най-ниския резултат от всички 13 експеримента (маско Dice 0.128), дължащ се единствено на различна техническа конфигурация, без клиничен аналог.

ID	Формулировка	Промяна спрямо baseline	Folds	Macro Dice	SD	IoU
E103	4-класова без JNEG	E000 без JNEG случаи	2	0.186	0.007	0.143
E104	Бинарен само JNEG	Еднокласов модел само JNEG	2	0.529	0.137	0.433
E105	Сливане на сходни класове	M/P → общ съдов клас, WE/L → общ клас с промени в бяло	2	0.230	0.020	0.181
E106	Промяна в класовия приоритет (M/P priority)	Приоритет $M > P > L > WE > JNEG$	2	0.128	0.014	0.098

Таблица 6. Трета група експерименти

JNEG-only експериментът E104 постига Macro Dice 0.529, но не е директно сравним, тъй като работи с еднокласова целева дефиниция и обхваща само подмножество от данните. Визуализацията все пак свидетелства за най-високото единични припокриване в цялото изследване, достигащо до $\text{Dice} = 0.913$. (Фиг. 18)

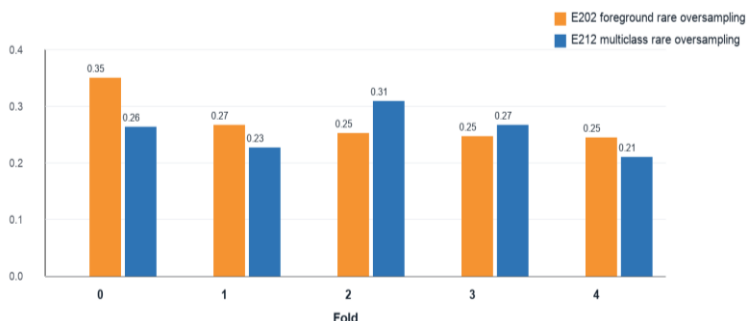
Пример 10: Много високо припокриване. **Foreground Dice = 0.913.**



Фигура 18. Качествена визуализация на fold-0 валидация (експеримент E104).

Обобщено, най-добрият потвърден бинарен модел е E202, а най-добрият потвърден многокласов модел е E212, постигащи macro Dice в диапазон 0.25 - 0.28 и foreground IoU от 0.20-0.21 (Фиг. 19)

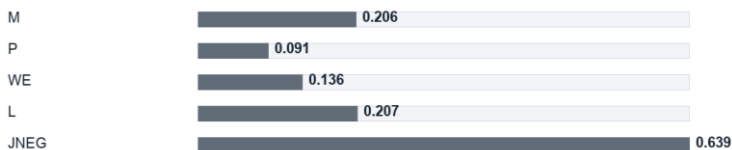
Confirmed Fold-Level Macro Dice



Фигура 19. Представяне по отделни поднабори за всеки от 5-те етапа валидация на двата потвърдени на етап 3 експеримента (E202 и E212)

Детайлният анализ по класове за модела E212 разкрива значителни разлики в разпознаваемостта между отделните колпоскопски класове (Фиг. 20).

E212 Confirmed Per-Class Dice



Фигура 20. Потвърдено представяне по класове от експеримент E212. JNEG доминира мултикласовия сигнал, докато P, WE, M, и L остават слабо изразени.

JNEG доминира многокласовия сигнал с Dice от 0.639 - близо три пъти по-висок от следващия клас. P (пунктуация) е най-слабо представим (Dice 0.091), M, WE и L попадат в сходен диапазон (0.136 - 0.207), което предполага, че моделът ги разграничава трудно един от друг.

4.1.5 Обсъждане

4.1.5.1 Оценка на модела и интерпретация на резултатите

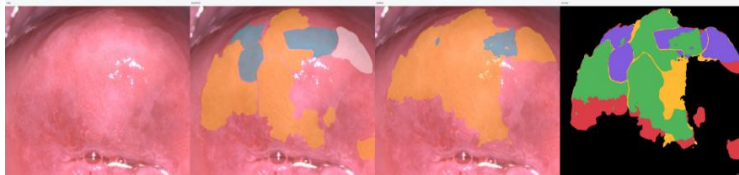
Резултатите от 13-те експеримента демонстрират, че 2D nnU-Net v2 ResEnc-M архитектурата може да се обучи да разпознава и локализира абнормни колпоскопски находки, но степента на припокриване между предсказанията и анотациите е под нивото, необходимо за самостоятелна клинична употреба. Представянето отразява значително по-ранен етап на развитие на модела от публикуваните в литературата системи, които решават предимно класификационни задачи и разполагат с по-големи набори от данни. Важно е да се отбележи, че по-голямата част от тях класифицират на ниво цяло изображение въз основа на хистологичен резултат (напр. HSIL срещу LSIL), докато тук задачата е пикселна сегментация на отделни колпоскопски находки. Този процес е фундаментално различен, по-труден и не е директно сравним с публикуваните класификационни стойности в литературата. Подробен сравнителен анализ е представен в Задача 4.

4.1.5.2 *Foreground срещу multiclass подход*

Първоначалната цел на експериментите беше да се изгради модел, способен да класифицира отделните колпоскопски находки чрез пикселна сегментация. Петкласовият модел (E212) постигна масго Dice от 0.256, което потвърждава, че архитектурата може да улови сигнал за отделните подтипове, но степента на точност остава под клинично приемливото ниво, основно поради малкия обем данни и класовия дисбаланс. Паралено с това, тестваният бинарен модел (E202), обединяващ всички патологични находки в един клас, постигна по-висок резултат (Dice = 0.273). Разликата между двата е само 0.017 пункта, но тя отразява съществено различна сложност на задачата. Бинарният подход трябва само да определи всеки пиксел като "абнормен" или "нормален", докато многокласовият цели допълнително да класифицира коя от петте колпоскопски находки е представена. Разликата в производителността на двата експеримента подкрепя двуетапна стратегия за моделиране: първи етап - детекция на абнормен регион, втори етап - класификация на колпоскопската находка. Този подход е в унисон с клиничното мислене на колпоскописта, който преди да определи типа на находката, първо преценява дали изобщо има абнормност и къде точно се намира тя. Допълнително предимство на двуетапния подход е възможността за настройка на различни нива на сигурност във всяка стъпка. Бинарният модел може да работи с по-нисък праг за разпознаване и да маркира дадена зона като абнормна дори при по-слаб сигнал, което осигурява висока чувствителност и намалява риска от пропуснати лезии. Многокласовият процес, от друга страна, изисква по-висока степен на увереност, за да се избегне неправилно определяне на подтипа, което се изразява във висока специфичност. На фона на тези ограничения, представянето е окуражаващо от гледна точка на осъществимостта на подхода. Резултатите на бинарния модел показват, че алгоритъмът е уловил съществен сегментационен сигнал въпреки малкия обем данни, класовия дисбаланс и липсата на протоколни метаданни. Високите Dice стойности при отделни случаи (0.732-0.913) свидетелстват, че архитектурата е способна да усвои прецизна сегментация, когато входните данни са с достатъчно високо качество и контраст. За това свидетелства качественият анализ на отделни експерименти, който разкрива, че моделът постига високо припокриване дори при не-JNEG класове, когато изображенията са с достатъчен контраст и ясни граници. 4-класовия експеримент (E103), при който JNEG случаите са изключени от обучението, съдържа пример с Dice = 0.813 (Фиг.21). Подобни високи стойности се наблюдават също при 5-класовия експеримент без празни анотации (E102), постигащ 0.811 (Фиг.22) и при многокласовия с добавени

аугментации (E211), където $\text{Dice} = 0.822$. (**Фиг.23**) Тези примери показват, че архитектурата е способна да сегментира М, Р, WE и L, когато данните са благоприятни, но производителността е силно зависима от качеството на конкретното изображение. Тази избирателност обяснява и ниската средна стойност - моделът работи добре в лесните случаи и пропуска в трудните.

Пример 11: Високо припокриване. Foreground Dice = 0.813.



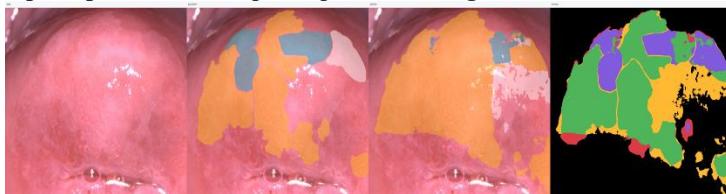
Фигура 21. Качествена визуализация на fold-0 валидация (експеримент E103).

Пример 12: Високо припокриване. Foreground Dice = 0.811.



Фигура 22. Качествена визуализация на fold-0 валидация (експеримент E102).

Пример 13: Високо припокриване. Foreground Dice = 0.822.



Фигура 23. Качествена визуализация на fold-0 валидация (експеримент E211).

4.1.5.3 JNEG сигнал и протокол-специфичност.

JNEG се откроява като най-разпознаваемия клас с Dice от 0.639, което е близо три пъти повече от останалите. Този резултат обаче трябва да се интерпретира внимателно. Йод-негативния участък се оценява само след апликация на разтвор на Лугол, който създава висок контраст между светложълтия лезионен регион и тъмнокафявия здрав епител. За разлика от нея, оцетнобелите класове се проявяват като фини текстурни и цветови вариации върху блед фон, което ги прави значително по-трудни за разпознаване. От клинична гледна точка това означава, че модел, обучен върху смесени изображения (оцетна киселина и Лугол) без протоколни метаданни, може да създаде подвеждащо впечатление за общата си производителност, тъй като високият Dice на JNEG доминира средната стойност. В реални медицински условия е необходимо всяко изображение да бъде етикетирано спрямо типа на протокола, по който е заснето.

4.1.5.4 Роля на аугментациите и пресемплирането

Експериментите с аугментации и промени в честотата на представяне на редките класове дадоха важни практически насоки за бъдещо обучение, постигайки различен ефект в зависимост от целевата формулировка. За да се оцени това, бяха сравнени два базови модела - бинарният E101, който различава само абнормна от нормална тъкан, и многокласовият E102, който разпознава отделните колпоскопски класове M, P, WE, L, JNEG. Върху всеки от тях бяха приложени три стратегии: фотометрични аугментации, увеличено присъствие на редките класове P и L в тренировъчния процес, и изрязване на области с размер 768x768 пиксела около находките M/P.

При многокласовия модел (E212) увеличеното присъствие на случаи с пунктуация (P) и левкоплакия (L) по време на обучението доведе до повишаване на масго Dice от 0.242 (E102) на 0.256. Това подобрение се обяснява с по-голямата сложност на многокласовата задача. Когато моделът трябва едновременно да локализира абнормни региони и да ги класифицира в пет отделни подтипа, допълнителното внимание към редките класове P и L му дава възможност да научи техните специфични характеристики, които иначе биха останали маскирани от по-често срещаните класове.

При бинарния модел (E202), където задачата е само да се разграничи абнормно изображение без подтипова класификация, същата стратегия не доведе до промяна в представянето (Dice = 0.273). Тъй като всички находки са обединени в един общ клас, добавянето на още P и L примери не предоставя на модела нова абнормна информация, защото тези случаи вече са част от базовата версия (E101).

Приложени бяха и различни фотометрични аугментации като симулиране на различна осветеност, отблясъци, размазване и компресионни артефакти. Както в многокласовия, така и в бинарния модел резултатите показаха понижени стойности: за E201 от 0.273 на 0.259, а за E211 се отчете минимално подобрение от 0.242 на 0.245 Dice. Този спад контрастира с очакванията от общата литература за deep learning, но намира обяснение в спецификата на колпоскопската диагностика. Цветът на тъканите носи специфична клинична информация - белия епител, мозаицизма и пунктуациите имат специфични цветови характеристики, които са част от диагностичния алгоритъм на колпоскописта. Фотометричните аугментации могат да нарушат или заличат тези характеристики, което обърква модела. Цветът има семантична стойност и стандартните аугментационни стратегии не трябва да се прилагат автоматично без съобразяване с конкретната диагностична модалност.

Изрязването на области около находката с размер 768x768 пиксела също не доведе до подобрение. В многокласовия модел (E213) се регистрира спад до Dice 0.227 срещу 0.242 при E102. Този резултат предполага, че изолирането на дадена абнормна зона заедно с малък регион от околната цервикална повърхност лишава модела от важен пространствен контекст. За разлика от задачи като детекция на тумори в солидни органи, където лезията е относително самостоятелна, колпоскопските находки съществуват в анатомична рамка с клиничен контекст. Размерът на оцетнобелия участък, позицията му спрямо външния цервикален отвор и връзката му с трансформационната зона са диагностично значими, а оразмеряването елиминира тези ориентири.

В заключение, представените експерименти демонстрират, че 2D nnU-Net v2 може да се обучи да разпознава колпоскопски находки на пикселно ниво. Резултатите разкриват, че производителността на модела се определя в по-голяма степен от формулировката на задачата и качеството на данните, отколкото от архитектурния капацитет. Докато моделът постига високо припокриване при отделни случаи с благоприятни условия, средните стойности остават под клинично приемливото ниво поради малък обем данни, класов дисбаланс и липса на протоколни метаданни. Аугментационните стратегии, пренесени от общата deep learning практика, се оказаха неподходящи за колпоскопската модалност, което подчертава необходимостта от специализиран подход. В този смисъл, задачата поставя методологична основа и идентифицира конкретните точки, които следващите задачи адресират - структурата на данните, влиянието на отделните фактори и пътят към клинична приложимост.

4.2 ВТОРА ЗАДАЧА: Да създадем стандартизирана база данни от колпоскопски изображения за интеграция в разработения модел чрез единен протокол за заснемане и анотация, съобразен с медицинските стандарти за колпоскопска номенклатура.

4.2.1 Формат на данните и анотационен протокол

Създадохме стандартизирана база данни от колпоскопски изображения, обработени съгласно медицинските стандарти на колпоскопската номенклатура, за интеграция в разработения модел. Използвахме собствен набор от 355 изображения, като установените находки в тях бяха структурирани в класове, наименовани съгласно колпоскопската карта на *проф. Божил Василев* и съответстващи на терминологията на *International Federation for Cervical Pathology and Colposcopy (IFCPC)* от 2011-та година. (Фиг.24-26) Допуснати бяха леки съкращения в абривиатурите, премахвайки “А-” пред абнормните находки и описвайки йод-негативния участък като JNEG за улеснение на модела. Въпреки неспецифичния характер на левкоплакия като лезия, тя беше включена в класовото обучение поради цветовото сходство с белия епител. На Таблица 7 е представена класова легенда на приложените колпоскопски анотации в дигиталното хранилище. (Табл. 7)

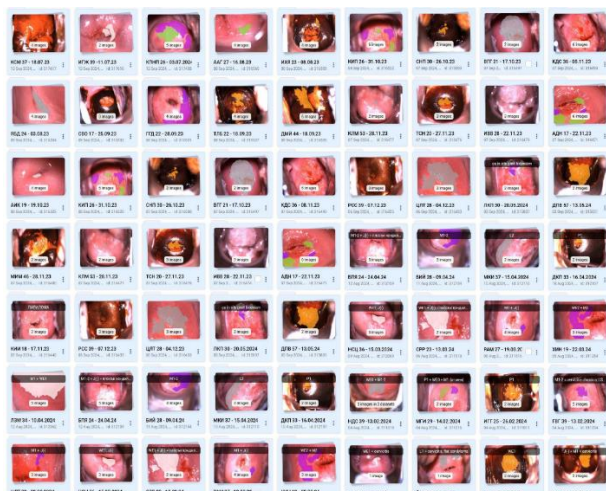
Код	Клас	Клиничен термин	Практическо анотационно описание
0	background	Фон / здрава тъкан	Нетегретен регион
1	M	Мозайка	Паважоподобен мозаичен съдов модел
2	P	Пунктуация	Точковидни съдови образувания върху бял фон
3	WE	Бял епител	Оцетнобяло оцветяване на епитела
4	L	Левкоплакия	Грапава бяла плака; визуално близка до WE
5	JNEG	Йод-негативен участък	Светложълта/оранжева зона на фона на кафяво оцветен здрав епител

Таблица 7. Класова легенда на колпоскопските находки



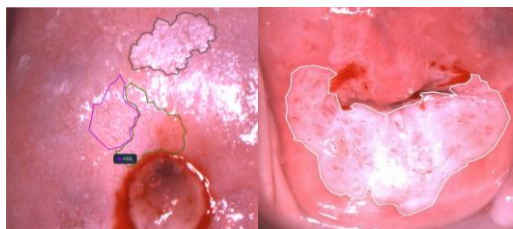
Фигури 24-26. Колпоскопски изглед на класовете WE, M, P

Анотационният процес е осъществен в *Supervisely* - цялостна учебна базирана платформа за изкуствен интелект и компютърно зрение. Тя предоставя унифицирана среда за целия жизнен цикъл на AI проектите - от управлението на данни и етикетването, до обучението, оценката и внедряването на невронни мрежи в реална работна среда. Платформата поддържа полигонална сегментация с професионални инструменти за прецизно очертаване на посочените граници. *Supervisely* позволява дефинирането на потребителски класове, импортиране на изображения в множество формати и експорт на анотациите в различни индустриални стандарти. Изображенията са качени в проекта като 1280x960 px файлове в RGB формат, организирани в структурирани индивидуални папки според техния произход. (Фиг. 27)



Фигура 27. Структура на базата данни – изглед през *Supervisely*

Всяка колпоскопска находка е анотирана чрез ръчно поставяне на точки по контура на лезията. Операторът маркира последователно границата на находката, като всяко отбелязване добавя нова точка, а линиите между тях оформят полигон (многоъгълник). (Фиг. 28) Платформата е специално оптимизирана да поддържа огромни резолюции и изключително детайлна визуализация. Тя позволява приближаване до нивото на отделен пиксел, което дава възможност за увеличаване на изображението, докато отделните пиксели станат ясно видими като големи квадратчета на екрана.



Фигура 28. Анотирани колпоскопски находки – близък план

Петте колпоскопски находки М (мозайка), Р (пунктуация), WE (бял епител), L (левкоплакия) и JNEG (йод-негативен участък) са създадени като отделни обектни класове, като за всеки от тях е зададен уникален цвят за визуална идентификация по време на анотиране. (Фиг. 29)



Фигура 29. Анотирани с характерен цвят колпоскопски класове

Празните анотационни файлове са изображения без патологични находки, които се експортират като празни текстови файлове. Те се третираят като истински отрицателни случаи (true negatives) и са запазени като отделни екземпляри в платформата без прикачени полигони. и служат за истински отрицателни случаи (true negatives). Припокриващите се полигони, като находките М или Р, които се наблюдават в границите на WE, са записани като отделни геометрии върху едно и също изображение. Тяхното сегментиране е извършено на по-късен етап чрез конфигурируема приоритетна йерархия - $L > M > P > WE > JNEG > background$.

След приключване на анотирането, проектът беше експортиран от Supervisely в първичния Supervisely JSON формат, който запазва пълната структура от класове, метаданни и геометрии. Всяко изображение беше

придружено от отделен текстов файл, съдържащ координатите на очертанията на всички колпоскопски находки в него.

4.2.2 Одит на данните и класово разпределение

Обобщено, база данните съдържа 625 колпоскопски находки под формата на полигони, разпределени между 355 изображения с резолюция 1280 x 960 пискела. Всяко изображение носи уникален код, идентифициращ конкретния клиничен преглед, от който произлиза. Например два кадъра с кодове "IMG_001_1" и "IMG_001_2" са заснети по време на един колпоскопски анализ на една и съща пациентка, но от различен ъгъл, увеличение или момент след апликация на разтвор. От общо 355 изображения се формират 123 такива групи, като 108 от тях (87.8%) съдържат по две или повече снимки от един и същи преглед, а останалите 15 са единични кадри. Това групиране има пряко значение за стратегията на разделяне на данните. Ако две снимки от един и същи преглед попаднат едновременно в обучаващия и валидационния набор, моделът ще бъде оценен върху почти идентично съдържание. Това събитие се нарича изтичане на информация (data leakage), което води до изкуствено повишаване на реалната диагностична ефективност и производителност. По тази причина разделянето на данните е извършено на ниво клиничен преглед. Така всички снимки от едно и също колпоскопско изследване остават на заедно или в тренировъчния, или във валидационния сет. След прилагането на тази стратегия, проверката потвърждава нулево изтичане - нито един преглед не попада на две места едновременно.

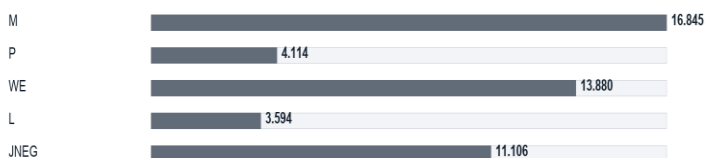
От 355-те изображения, 81 са с празни анотационни файлове и не съдържат колпоскопски находки. Наличието на подобни примери са от съществено значение, тъй като служат като негативни контроли, които учат модела да разпознава липсата на абнормни изменения. По брой полигони клас М (мозайка) е най-често срещаният с 201 полигона, докато L (левкоплакия) е най-редкият с едва 62 полигона. (Табл. 9)

Клас	Полигони	Изображения с находка
М	201	132
WE	142	88
P	117	71
L	62	26
JNEG	101	61

Таблица 9. Разпределението на полигоните спрямо изображенията

Класовият дисбаланс е особено изразен на пикселно ниво, където най-големият клас М заема около 16.8 млн. пикселна площ, близо петкратно (4.69 пъти) повече от най-малкия клас L. (Фиг. 30) Йод-негативните участъци се нареждат на трето място по ангажирана пикселна повърхност, отчитайки обаче най-голяма средна площ на полигон от всички класове - 110 000 пиксела на 1 полигон. С други думи, JNEG участъците са големи, компактни и добре очертани региони с ясна граница спрямо кафявия здрав епител. Това, съчетано с високия контраст на Луголовото оцветяване, прави йод-негативните участъци обективно по-лесни за разпознаване не само заради броя на примерите, а заради самата природа на анатомичния сигнал. Обратно, Р има най-малка средна площ - точковидните съдови образувания са дребни и разпръснати, което ги прави трудни за сегментиране независимо от броя им. (Табл. 10)

Dataset Foreground Pixel Area by Class (Millions)



Фигура 30. Разпределение на foreground пикселната площ по класове в милиони.

Клас	Полигони	Пиксели (foreground)	Дисбаланс спрямо най-големия
М	201	16 844 872	1.00x
Р	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

Таблица 10. Разпределение на foreground пикселната площ по класове и съотношение на дисбаланса. Най-редките класове по пикселна площ са L (левкоплакия) и Р (пунктуация).

4.2.3 Структура на набора от данни и подмножества за експериментите

От 355 колпоскопски изображения, 22.8% са придружени от празни анотационни файлове. След изключването им остават 273 изображения с поне една анотация (E102), като 212 от тях съдържат само класовете М,

WE, P и/или L (E103). В условия на приложен Луголов разтвор разполагаме с 61 JNEG-положителните случаи (E104). Проведените експерименти от Задача 1 използват различни филтрирани подмножества, които разкриват структурната хетерогенност на данните. (Табл. 11)

ID	Подмножества	Общо	Бележка
E000/E101	Всички 355 изображения	355	Включва 81 изображения без анотации
E102	Без празни анотации	273	Само изображения с поне една находка
E103	Само изображения след оцетна киселина	212	JNEG и празните случаи са изключени
E104	Само изображения след приложение на Луголов разтвор	61	JNEG изображения

Таблица 11. Подмножества на данните според филтриращите критерии на експериментите

Това подмножествено филтриране се използва като заместителен подход при липсата на протоколни метаданни. Експеримент E103 симулира набор от изображения, получени след приложение на оцетна киселина, а E104 - набор от изображения след приложение на Луголов разтвор. В двата случая става въпрос за условно разделяне, тъй като за отделните изображения не е налична информация относно реално използвания диагностичен протокол.

В опит да компенсираме класовия дисбаланс, без да разчитаме на събиране на нови клинични данни, приложихме три аугментационни стратегии - фотометрична аугментация на цветовите характеристики, повишено представяне на редките находки (oversampling) и изрязване около P и L анотациите. Резултатът от тези стратегии е видим в **Таблица 12**, представяща пълния списък от 13 експеримента с видовете извършени аугментации и настъпилите промени в базата данни.

ID	Основен брой	След аугментация	Общо	Бележка
E000	355	0	355	Без аугментация, 5-класов модел
E101	355	0	355	Без аугментация, бинарен модел
E102	273	0	273	Без празни анотации

ID	Основен брой	След аугментация	Общо	Бележка
E103	212	0	212	Без празни и JNEG анотации
E104	61	0	61	Само JNEG анотации
E105	355	0	355	М/Р – общ съдов клас WE/L – общ бял клас
E106	355	0	355	Променен приоритет в припокриването
E201	355	710	1065	Две фотометрични копия на всяко изображение
E202	355	629	984	Р/L случаите се повтарят 5 пъти, останалите 1 път
E203	355	267	622	Около всяка Р/L анотация е изрязан 768×768 px регион
E211	273	546	819	Две фотометрични копия на всяко изображение спрямо E102
E212	273	629	902	Повишено представяне на Р/L спрямо E102
E213	273	267	540	Около всяка Р/L анотация е изрязан 768×768 px регион спрямо E102

Таблица 12. Аугментационни процеси в базата данни - качествено и количествено описание

Разликите в броя на случаите между E202 и E212 отразяват различните филтриращи критерии. E202 работи с пълната база от 355 изображения с аугментациите, увеличаващи общия брой до 984 случая. E212 изключва

празните маски и разполага с 273 оригинални изображения, достигащи общо 902 случая след аугментациите.

Повишеното представяне на редки класове (E202, E212) увеличи честотата на поява на Р и L случаите до 629 изображения, чрез повтарянето им 5 пъти повече спрямо останалите единично показани случаи. Това цели да компенсира липсата на равновесие в находките, без да променя визуалното съдържание на образите. Изрязаните изображения около същите редки класове в резолюция 768 x 768 (E203, E213), създаде около 267-273 допълнителни копия за обучение.

Разделянето на данните за кръстосана валидация е извършено на групово ниво, като изображенията с индикатор за произход от един и същи колпоскопски преглед са запазени в един и същи поднабор (fold). Тази стратегия намалява риска от изтичане на информация между тренировъчен и валидационен сет, но не е еквивалентна на истинска пациентска карта поради отсъствието на уникален идентификатор за всяка прегледана жена в групата.

В допълнение, за окончателно потвърдените модели E202 и E212 (5-fold експериментите) е показано разпределението на train и validation случаите за всеки отделен fold, като съотношението варира от близо 90:10 до около 94:6. Валидационният дял варира от 5.8% до 9.7%, което е под стандартните 10-20% и отразява ограничения обем на данните. **(Табл. 13)** Във връзка с груповото разделение на данните моделът разполага с между 750-827 изображения за тренировка срещу 55-81 за валидация за E202 и 688-764 срещу 46-61 съответно за E212.

Експеримент	Фолд	Train случаи	Val случаи	Train дял	Val дял
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%

Таблица 13. Разпределение на тренировъчни и валидационни случаите за E202 и E212

5.2.4 Обсъждане

Създаденият набор от данни покрива всички основни колпоскопски класове според IFCPC 2011 класификацията, което го прави практически приложим за обучение на модел за семантична сегментация. Легендата на класовете съответства на утвърдената колпоскопска номенклатура, а полигоналният формат на анотациите позволява прецизно очертаване на регионите на различните находки без загуба на пространствена разделителна способност. Предимството на натрупаната база данни спрямо описаните в литературата е фокусът върху пикселна сегментация на находките, а не върху класификация на цялото изображение въз основа хистологичен резултат. Това дава възможност за по-детайлен анализ на морфологията на колпоскопските лезии, което е съществено за клиничната интерпретация. Включването на чист JNEG клас позволява оценка на модела и в двата основни колпоскопски протокола - с оцетна киселина и разтвор на Лугол.

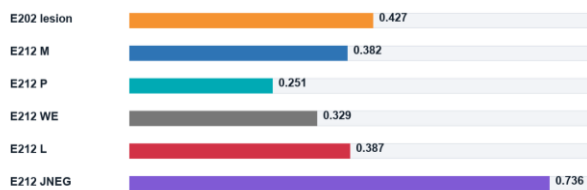
Целта на аугментацията в контекста на обучение за семантична сегментация е изкуствено разширяване на обучаващото множество чрез контролирани трансформации на съществуващите образци. Това намалява риска от overfitting и подобрява генерализацията на модела към непознати примери. За представения ограничен набор от данни, този процес е особено важен, защото позволява на модела да се запознае с вариации в осветлението, цветовия баланс, фокуса и позиционирането на лезиите, без да се налага да събира нови изображения. В настоящото изследване приложихме три различни аугментационни стратегии, като всяка от тях разширява обучаващия набор в различна посока. Фотометричната аугментация генерира по две допълнителни копия на всеки оригинален образ чрез комбинация от промени в яркостта, контраста и насищането на цветовете, заедно със синтетични отблясъци, замъгляване, потъмняване и компресия на формата. Повишеното представяне на редките класове спомогна за увеличаване на честотата на поява на случаите, съдържащи пунктуация (P) и левкоплакия (L), повтаряйки ги 5 пъти спрямо еднократното показване на останалите случаи. Изрязването на изображенията около въпросните редки находки даде на модела подобрен изглед към тези лезии, с цел да балансира класовия дисбаланс, без да променя визуалното съдържание на образите.

Основното предизвикателство при работа с ограничен брой данни е възможността моделът да "запомни" находките от обучаващия набор, вместо да изгради обобщена способност за разпознаване на патологични изменения върху нови, непознати за него изображения. За да се предотврати това, наличните данни се разделят на обучаващ дял (training

set), върху който моделът изгражда своите диагностични критерии, и контролен, валидационен дял (validation set), който остава напълно непознат по време на обучението и служи за независима проверка на неговата диагностична точност. Тъй като единично разделение може да доведе до случайно благоприятна или неблагоприятна разбивка, данните се разделят на n -брой части и обучението се повтаря n -пъти. Така всяка част последователно играе ролята на валидационен дял, а останалите служат за обучение. По този начин всеки образец служи за валидация само веднъж, а крайният резултат е средна стойност от всички fold-ве, което дава значително по-стабилна оценка. Съотношението между тези двата сета е компромисен - повече тренировъчни данни подобряват качеството на обучение, но твърде малък валидационен дял води до нестабилна оценка с висока дисперсия.

Качеството на създадената база данни може да бъде оценено не само чрез количествените показатели на анотациите, но и чрез способността на сегментационния модел да научи последователно всеки клас. Ако дадена находка е представена еднозначно, моделът постига висока разделителна способност, но когато процесът на етикетиране е непостоянен или концептуално различен, метриките остават ниски независимо от архитектурата. Тази зависимост е демонстрирана на **Фигура 30**, където Dice коефициентът за всеки клас е изчислен чрез обединяване на пикселите от петте валидационни групи (pixel-pooled Dice). Този показател се различава от макро Dice, използван при селекцията на модела, тъй като по-големите области получават по-голяма тежест в изчисленията.

Pixel-pooled Dice from full 5-fold confusion matrices



Фигура 30. Dice коефициент, изчислен върху всички пиксели от 5-кратната кръстосана валидация

Получените резултати показват съществена разлика в степента, с която отделните класове могат да бъдат научени от модела. JNEG достига Dice коефициент 0.736, което показва висока консистентност на анотациите и наличие на добре дефинирани визуални характеристики. За разлика от него, класовете P (0.251), WE (0.329) и L (0.387) демонстрират значително

по-ниски стойности. Това вероятно се дължи не само на по-ниската честота на съответните находки, но и на по-висока степен на вариабилност при тяхното очертаване. В условия на приложена оцетна киселина границите между метапластичния епител, белия епител и пунктуацията често са постепенни и припокриващи се, което създава предпоставки за различна интерпретация от отделните експерти. В резултат на това моделът получава по-малко еднозначен обучаващ сигнал и постига по-ниска сегментационна точност. Следователно разликите в Dice коефициентите могат да се разглеждат като индиректен показател за степента на съгласуваност и обективност при аотирането на отделните колпоскопски находки. В заключение, можем приемем, че непостоянството в база данните, е част от причините влияещи на сегментационната производителност на модела, които са разгледани в Задача 3.

4.3 ТРЕТА ЗАДАЧА: Да идентифицираме и анализираме факторите, които влияят върху сегментационната производителност на модела.

За идентификация на факторите, влияещи върху сегментационната производителност на разработения модел, синтезирахме количествените резултати от 13-те експеримента и техническите показатели от одита на базата данни, застъпени в първите две задачи. Проведохме и разширен анализ, включващ допълнителни метрики като ROC-AUC, прецизност, чувствителност и софтвакс калибрация, за да дефинираме основните бариери пред диагностичната точност. Идентифицираните фактори се групират в три категории, свързани с данните, анотации към тях и модела.

4.3.1 Фактори, свързани с данните

4.3.1.1 Класов дисбаланс и обем на данните

Описаният класови дисбаланс се очертава като водещ фактор, ограничаващ многокласовото представяне. Разликата в пикселната площ между най-застъпената колпоскопска находка М (16.8 млн. пиксела) и най-рядката L (3.6 млн. пиксела) е 4.69 пъти. Към това се добавя и съществения брой изображения (22.8%) без анотационно съдържание. Влиянието на тази диспропорция върху производителността е представено чрез пет различни показателя: степен на припокриване Dice, прецизност, чувствителност и ROC-AUC в многокласовия експеримент E212. (Табл. 14)

Клас	Брой полигони	Брой изображения	Прецизност	Чувствителност	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

Таблица 14. Показатели на модела E212 (5-кратна крос-валидация) по класове

Класът JNEG, с 61 изображения и отчетливи визуални признаци, постига ROC-AUC 0.908 и чувствителност 0.762. За разлика от него, Р разполага с 71 изображения, характеризиращи се с фини, разпръснати съдови точки, и постига едва 0.157 чувствителност. М има най-много образци (132), но Dice = 0.206, докато L е с най-малко примери (26), но със

сходен резултат. Това показва, че броят на изображенията не е достатъчен предиктор за успеваемост, а морфологичната сложност и контрастът на находката са еднакво важни. Съществен фактор е и абсолютният обем на данните - 355 изображения за петкласово сегментиране са под препоръчителния минимум. Това количество води до малки валидационни поднабори между 5.8–9.7% от данните, което увеличава разсейването на оценката и подсилва вариабилността в разбиванията.

4.3.1.2 *Вариабилност между поднаборите и липса на групиране по пациент*

Стандартното отклонение на средната стойност на припокриване за всички класове (Macro Dice) между отделните поднабори е 0.044 за E202 и 0.039 за E212. Тази вариабилност отразява чувствителността на модела към конкретното разпределение на редките класове. Разлика от 15 валидационни случая между поднаборите (61 срещу 46 за fold-0 и fold-4 от E212) може значително да промени представителността за рядък клас като P, който е аотиран само в 71 изображения общо. Към това се добавя и липсата на пациентска карта с разделянето на данните на ниво пациент, а не по персонален идентификатор. Така се намалява риска от изтичане на данни, но не се гарантира, че два прегледа на един и същ пациент не попадат в различни поднабори. Това е допълнителен фактор, който увеличава разсейването на оценката, особено при малък набор от данни.

4.3.1.3 *Протокол-специфичен сигнал и липса на метаданни*

Един от основните фактори, определящи производителността на модела, е специфичният цветови контраст на изображенията, получен при различните етапи на колпоскопския протокол. JNEG постига Dice = 0.639 и ROC-AUC = 0.908, стойности, които значително превъзхождат следващият най-добре представим клас M. Тази разлика не е просто резултат от различна честота на класовете, а се дължи на визуалното изпъкване на светложълтия участък върху тъмнокафявия йод-позитивен фон. Експеримент E103 (без JNEG, само M/P/WE/L) показва спад от 0.242 на 0.186 Dice, което означава, че 29% от многокласовия сигнал идва от JNEG. (Табл.6) В текущата база данни липсват метаданни за колпоскопския процес, което принуждава модела да учи смесен протокол и създава риск за асоциация на JNEG с общото понятие "лезия", вместо да се разпознава като специфична находка, зависима от условията.

4.3.1.4 Ефект от аугментационните стратегии

Приложените аугментационни стратегии показаха ограничен или отрицателен ефект върху производителността. Фотометричната аугментация (E201 с Dice = 0.259; E211 с Dice = 0.245) доведе до резултати под изходното ниво (E101 = 0.273). (Табл.4-5) Това предполага, че цветът е критичен диагностичен признак в колпоскопията и неговата произволна промяна обърква модела. Увеличеното повторение на редки примери при E202 и E212 задържа Dice на нивото на изходните експерименти между 0.256 и 0.273, но предизвиква по-високо разсейване. При стандартно отклонение от 0.044 срещу 0.026 за E101 излиза, че oversampling-ът помага на модела да ги вижда по-често, но не подобрява качеството на наученото. Изразването на участъци около лезиите при E203 и E213 редуцира Dice под изходното ниво, вероятно защото изолирането на малки региони лишава модела от пространствения контекст на цялата маточна шийка. Единствената промяна, която значително подобрява Dice, е филтрирането на данните (E104 само с JNEG има Dice = 0.529, което е ръст от 94%), което доказва, че разнородността, а не количеството, е основният проблем. (Табл. 6)

4.3.2 Фактори, свързани с анотациите

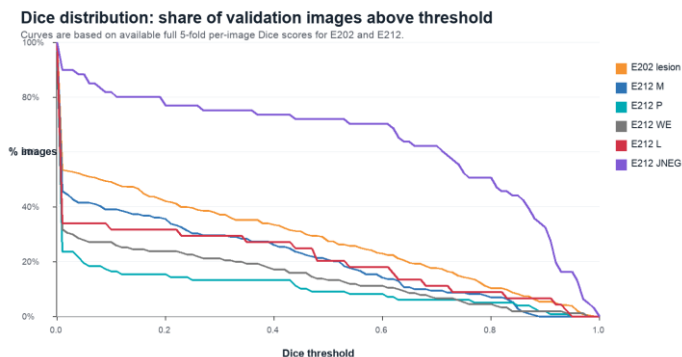
Разпределението на Dice стойностите в E212 предоставя индиректна, но информативна мярка за консистентността на анотациите. От една страна, ако даден клас е анотиран еднозначно, моделът би трябвало да постигне висок Dice. От друга страна, ако анотациите са непоследователни, Dice остава нисък независимо от архитектурата. (Табл. 15)

Клас	Брой	Mean Dice	Медиана	≥ 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%)

Таблица 15. Разпределение на Dice стойностите за E212

Резултатите разкриват значителна разлика в анотационната консистентност между класовете. JNEG се откроява с висок медианен коефициент на припокриване от 0.803, като 72.1% от йод-негативните случаи преминават прага от 0.50. От друга страна, класовете P, WE и L показват нулева медиана, което означава, че в над половината случаи моделът не открива никакво припокриване с анотацията. Едва в 9.3% от P

и 13.3% от WE находките моделът успява да повтори поне половината от лезиите, което не е просто последица от рядкостта на тези класове. Дори когато P и L присъстват в базата данни, анотациите им не са достатъчно постоянни, за да произведат стабилен обучаващ сигнал.



Фигура 31. Дял на изображенията с Dice коефициент над определен праг. Класът JNEG показва най-ясен положителен сигнал, докато P, WE и L остават нестабилни.

Представената визуализация демонстрира по-интуитивно каква част от изображенията достигат практически полезно ниво на припокриване между предсказаната и референтната анотация. (Фиг. 31) При JNEG се наблюдава съществен дял от случаи с високо качество на сегментацията, докато при P, WE и L голяма част от изображенията остават с ниски стойности. Това показва, че средният Dice не прикрива наличието на надежден модел за тези подтипове, а отразява реалната им нестабилност и ограничена възпроизводимост.

Матрицата показва какво точно се случва, когато моделът се опита да класифицира пет клинични категории, които не са взаимно изключващи се. (Фиг. 32)

E212 full 5-fold confusion matrix

	Predicted class					
	background	M	P	WE	L	JNEG
background	98.1% 282,198,945	0.3% 930,723	0.0% 99,621	0.3% 890,009	0.1% 226,855	1.2% 3,427,681
M	69.1% 11,645,939	26.4% 4,440,201	0.4% 69,612	3.7% 629,821	0.4% 59,299	0.0% 0
P	66.8% 2,722,045	6.1% 247,536	15.7% 639,991	11.4% 462,992	0.0% 327	0.0% 0
True						
WE	69.4% 8,374,005	5.8% 703,338	1.3% 162,458	23.4% 2,824,616	0.1% 6,828	0.0% 0
L	63.8% 2,291,196	1.0% 34,248	1.6% 56,127	7.8% 280,544	25.9% 931,412	0.0% 0
JNEG	23.5% 2,610,997	0.2% 27,558	0.0% 0	0.0% 0	0.0% 0	76.2% 8,467,476

Фигура 32. Конфузионна матрица на експеримент E212 при пълна 5-кратна кръстосана валидация.

Почти 70% от пикселите на M, P, WE и L попадат във фона (background), защото алгоритмът просто не разпознава лезията и по подразбиране избира околната повърхност. Разпиляването на останалите 30% не е случаен шум, а следва клиничната логика. Най-честата грешка е P да бъде класифицирана като WE (11.4%), следвана от взаимно объркване между M и WE (3.7–5.8%), тъй като и трите класа споделят една и съща оцветеностна основа. Асиметрията е показателна - P се бърка с WE девет пъти по-често от обратното. Това означава, че при несигурност моделът регресира към WE, така както колпоскопист би дефинирал труден случай просто като „бял епител”. От друга страна, нито един пиксел от M, P, WE или L не бива предсказан като JNEG, което се дължи на двата крайно различни протокола на изследване.

Софтмакс архитектурата на практика налага взаимно изключваща се класификация и експеримент E106 демонстрира значението на този фактор. При промяна на приоритета на припокриване от $L > M > P > WE > JNEG$ на $M > P > L > WE > JNEG$ средният Dice се понижава с 17% при идентични данни и архитектура. Експеримент E105, при който M и P са слети в съдов клас, а WE и L в бял клас, постига 49% подобрене само чрез премахване на нееднозначната граница. Софтмакс функцията изисква всеки пиксел да

принадлежи точно на един клас, но в клиничната практика М и Р са находки, разположени върху WE, а цветовете границите между WE и L са сходни. Общата картина потвърждава, че софтвакс моделът не страда от архитектурен лимит, но се затруднява от изискването да избира категорично между понякога визуално сходни класове.

4.3.3 Фактори, свързани с модела

4.3.3.1 Целева формулировка - двоична срещу многокласова сегментация

Сравнението между двоичната сегментация (E202 с Dice = 0.273) и петкласовата многокласова сегментация (E212 с Dice = 0.256) разкрива, че целевата формулировка сама по себе си е фактор. При бинарния подход (E202) всички лезии се обединяват в един клас и моделът се фокусира върху локализация, докато при многокласовия сценарий моделът трябва допълнително да ги разпредели по подтип. Разликата от около 0.017 Dice може да изглежда малка, но предвид, че само един от петте класа (JNEG) е добре представен, реалната клинична полезност на многокласовия подход е по-ограничена. Това подкрепя двуетапна стратегия - първо откриване на лезия чрез двоична сегментация, а след това последваща класификация.

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

Фигура 33. Конфузионна матрица на експеримент E202 при пълна 5-кратна кръстосана валидация.

При експеримент E202 по-голямата част от пикселите, принадлежащи към фона, са класифицирани правилно. В същото време значителна част от действителните лезионни области погрешно се разпознават като фон. (Фиг. 33) Това обяснява защо моделът постига високи стойности на точност и специфичност, докато Dice коефициентът и чувствителността за

лезионните класове остават ограничени. С други думи, моделът се справя добре с разпознаването на нормалните области, но пропуска значителна част от патологичните изменения.

4.3.3.2 Калибрация и софтвакс увереност

Задълбоченият анализ на softmax вероятностите показва, че моделът генерира висока увереност при правилното разпознаване на ограничена подгрупа от случаи, но систематично пропуска значителна част от лезионните пиксели. Повторното изпълнение на модела върху 273-те валидационни изображения на E212 (**Табл.12**) със запазване на пълните вероятностни карти позволи изчисляването на ROC-AUC и PR-AUC, както и анализ на разпределението на softmax увереността. (**Табл. 16**)

Клас	ROC-AUC	PR-AUC	Медианна вероятност за лезионните пиксели	Увереност при верните предсказания
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

Таблица 16. Вероятностни показатели за E212 (5-кратна крос-валидация)

PR-AUC (Precision-Recall AUC) е по-информативен от ROC-AUC при силно неравномерни данни, тъй като се фокусира само върху лезионните пиксели и отразява частта фалшиво положителните предсказания. За P (PR-AUC 0.180) и WE (0.257) стойностите са значително под ROC-AUC (0.638-0.704), което означава, че повечето предсказания за тези класове са фалшиво положителни.

Най-показателна обаче е третата колона, отчитаща медианната вероятност, която моделът дава на пикселите, маркирани като лезия. За класовете P, L, WE и M тя е между 0.001 и 0.002, т.е. повечето истински лезионни пиксели получават почти нулева вероятност от модела да бъдат такива. Само JNEG има медианна вероятност 0.999. В същото време, когато алгоритъмът познае правилно, той е много уверен и 90% от точните му предсказания са с вероятност над 0.87. Комбинацията между висока увереност в малкото верни случаи и почти нулева вероятност за повечето истински пиксели показва, че моделът научава само малък консистентен брой примери и игнорира останалите като противоречиви.



Фигура 34. Softmax увереност според изхода от класификацията. TP – true positive; FP – false positive; FN – false negative. Висока увереност при FP и FN означава, че моделът често греши с висока увереност.

Моделът демонстрира висока увереност при голяма част от правилно класифицираните пиксели, но същевременно запазва висока увереност и при значителен брой неправилно класифицирани пиксели. (Фиг. 34) Това предполага недостатъчна калибрация на модела и наличие на нееднозначност както в анотациите, така и в самата задача. Поради тази причина стойностите на softmax увереността не следва да се използват като самостоятелен показател за клинична надеждност без допълнителна калибрация и експертна проверка.

4.3.4 Обсъждане

Идентифицираните фактори, влияещи върху сегментационната производителност на модела, са обособени в три категории, свързани с данните, анотациите и модела. Архитектурата nnU-Net разполага с достатъчен капацитет за отлична сегментация, когато диагностичният сигнал е консистентен и висококонтрастен. За това свидетелства високият ROC-AUC за JNEG (0.908). Отслабването на показателите за P, WE и L не отразява архитектурен лимит, а взаимодействието между ограничения обем и дисбаланс на данните, непостоянство на анотациите и несъвместимост на софтвакс формулировката с клиничната реалност.

База данни с 355 колпоскопски изображения за петкласова сегментация е значително под препоръчителния минимум. За сравнение, Yuan и сътр. (2020) използват над 11 000 изображения, а големите медицински сегментационни набори могат да включват от стотици до хиляди изображения, макар размерът им да варира значително в зависимост от задачата и модалността. (Kuş et al., 2024) Към малкия обем

се добавя и класовият дисбаланс, свързан с кратните разлики в заетата пикселна повърхност от различните находки, наред с празните анотационни файлове. Тази комбинация създава ситуация, при която редките класове разполагат с недостатъчен брой изображения и още по-малко смислени лезионни пиксели, от които моделът да се учи. Недостатъчните валидационни поднабори допълнително увеличават разсейването на оценката, като при тази количествена извадка дори минимална разлика между две разбивания може да промени статистическата представителност.

Отвъд количеството, разнородността на данните се оказва по-фундаментален проблем. Протокол-специфичният контраст на йод-негативните участъци е толкова различен от обработка с оцетна киселина, че моделът постига значително по-високи резултати за него дори с малко на брой примери. Експеримент E103 показва, че класът JNEG има съществен принос към крайния резултат, защото след премахването му Dice намалява с една трета. Моделът работи едновременно с два различни типа изображения, но тъй като липсва информация за използвания протокол, те не могат да бъдат разделени и се обработват като единна група, което вероятно затруднява обучението на модела.

Неуспехът на аугментационните стратегии затвърждава този извод. Фотометричните аугментации влошават резултатите, защото произволната промяна в цвета унищожава сигнала. Oversampling-ът задържа Dice на изходното ниво, но не го подобрява, което показва, че повторението на редките примери не добавя нова информация. Единственото драстично подобрение идва от филтрирането на данните, което доказва, че разнородността, а не количеството, е основният ограничител.

Анотационната консистентност се оказва по-сериозно предизвикателство от класовия дисбаланс. Това проличава от две посоки на доказателства, които взаимно се потвърждават. Първо, разпределението на Dice стойностите показва, че проблемът не е просто в рядкостта на класовете. Ако дисбалансът беше единствената причина, бихме очаквали ниски, но равномерно разпределени Dice стойности. Вместо това наблюдаваме нулева медиана за P, WE и L, където в над половината случаи няма никакво припокриване между предсказание и анотация. Това е качествено различна ситуация, тъй като за значителна част от примерите анотацията и визуалният сигнал са в противоречие, което прави невъзможно за модела да научи сигурно находката. Съпоставката с E104 (само JNEG, Dice 0.529) показва, че когато анотациите са консистентни, моделът се справя дори с малко данни.

Второ, конфузионната матрица на E212 визуализира този конфликт на пикселно ниво. (Фиг. 32) Разпиляването на M, P, WE и L не е произволен шум, а следва клиничната логика - при несигурност моделът регресира към

WE като базов клас, където шансът за грешка намалява. JNEG стои напълно отделно, което допълнително затвърждава, че сигналът му произлиза от друг източник, а не от общия набор класове.

Експериментите E105 и E106 предоставят най-силното доказателство, че именно анотационната нееднозначност, а не дизайнът на модела, е основната пречка. Промяна само на приоритета за разрешаване на припокривания води до спад в Dice при идентични данни и архитектура, резултат, който е трудно обясним чрез ограничения в модела. Обратно, премахването на непостоянството чрез сливане на клинично припокриващите се класове подобрява Dice без никаква промяна в модела. Тези два експеримента заедно показват, че софтвакс формулировката е несъвместима с припокриващите се IFCPC категории и трансформира субективността на анотациите в систематично противоречив обучаващ сигнал.

Модел-специфичните фактори са по-слабо изразени, но показателни. Разликата между двоичната и многокласовата формулировка отразява допълнителната тежест върху модела да разпредели лезиите по подтип, когато повечето класове са недобре представени. Това подкрепя двуетапна стратегия, при която първо се открива лезията, след което тя бива класифицирана - подход, който решава проблема чрез два по-лесни за изпълнение етапа.

Анализът на софтвакс калибрацията разкрива още един аспект на същия казус. Разминаването между ROC-AUC и PR-AUC показва, че моделът не просто греша, но и го прави уверено. Високата вероятност за малкото правилни предсказания, съчетана с почти нулева вероятност за повечето истински лезионни пиксели, е класически признак за шум в анотациите. Това прави грубите софтвакс стойности неподходящи като мярка за диагностична увереност без допълнителна калибрация.

В обобщение, йерархията на факторите е следната - данните осигуряват суровия материал, анотациите определят качеството на сигнала, а моделът отразява несъвършенствата и на двете. Бъдещото развитие на системата трябва да премине от количествени опити за аугментация към качествена промяна в анотационния протокол и преосмисляне на софтвакс формулировката за клинично припокриващи се класове.

4.4 ЧЕТВЪРТА ЗАДАЧА: Да определим диагностичната стойност на разработения модел чрез сравнителен анализ с утвърдени в литературата системи

4.4.1 Мост между сегментационни и класификационни метрики

Основното предизвикателство при сравнението на разработената система с публикуваните в литературата алгоритми произтича от различията в поставената задача. Представеният модел функционира на пикселно ниво чрез мултикласова сегментация на колпоскопските находки съгласно IFCPC номенклатурата, докато повечето публикувани решения са насочени към класификация на ниво изображение. При тези условия възможностите за директно съпоставяне са ограничени, като единствената широко използвана метрика, позволяваща сравнително коректен паралел между двата подхода, е площта под ROC кривата (AUC). (Табл. 17)

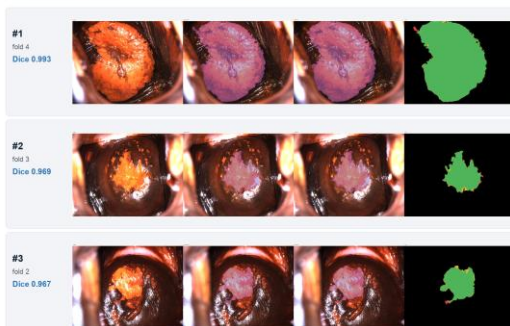
Модел / Клас	ROC-AUC	PR-AUC	Median Dice	Dice > 0.50	Dice > 0.90
E212 - JNEG	0.908	0.654	0.803	72.1%	32.8%
E212 - M	0.737	0.366	0.001	20.8%	0.0%
E212 - WE	0.704	0.257	0.000	13.3%	2.0%
E212 - L	0.686	0.286	0.000	20.5%	6.8%
E212 - P	0.638	0.180	0.000	9.3%	2.1%
E202 (бинарен)	0.767	-	0.085	27.7%	5.5%

Таблица 17. Диагностични показатели на разработените модели (E212 многокласов модел с разделение по находки и E202 бинарен модел).

В мултикласов аспект най-убедителен сегментационен сигнал е отчетен при йод-негативния клас. JNEG се откроява с ROC-AUC 0.908 и медианен Dice от 0.803, като в 72.1% от изображенията се постига над 50% припокриване между анотацията и разпознатото от модела. Най-добрите примери постигат Dice до 0.993. (Фиг. 35) Бинарният модел (E202) заема междинна позиция с ROC-AUC 0.767 и 27.7% изображения над Dice 0.50, като най-високата отчетена стойност е 0.992. (Фиг. 36)

Пример 14: JNEG (E212). Почти перфектно припокриване, изразяващо се с Dice = 0.993. Тези случаи принадлежат към 32.8% от изображенията, за които моделът отчита Dice > 0.90.

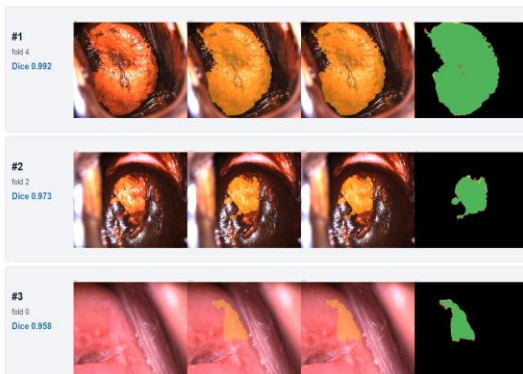
Top available validation examples: E212 JNEG



Фигура 35. Примери от най-доброто представяне на JNEG класа при многокласовия модел E212.

Пример 15: Бинарен модел (E202). Високо ниво на разпознаване на лезия без подлежаща класификация (Dice = 0.958-0.992).

Top available validation examples: E202 lesion



Фигура 36. Примери от най-доброто представяне на бинарния модел E202.

Останалите класове (M, P, WE и L) показват ROC-AUC между 0.638 и 0.737 и медианен Dice от едва 0.000-0.001. Въпреки по-ниските стойности, при средно 15.97% (диапазон 9.3-20.8%) от оцетнобелите находки Dice надвишава 0.50, а най-добрите валидационни примери достигат висока степен на припокриване между 0.81-0.94 Dice (**Фиг. 37-39**).

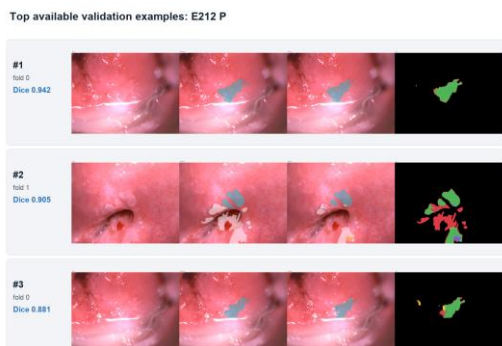
Тези визуализации потвърждават, че моделът може успешно да локализира и сегментира съответните класове, когато образният сигнал е достатъчно изразен.

Пример 16: Клас *М* (E212). Високо припокриване на изразените мозаечни находки (Dice = 0.843-0.875).



Фигура 37. Примери от най-доброто представяне на М класа при многокласовия модел E212.

Пример 17: Клас *Р* (E212). Високо припокриване на визуално отчетлива пунктирана зона (Dice = 0.881-0.942).



Фигура 38. Примери от най-доброто представяне на Р класа при многокласовия модел E212.

Пример 18: *Клас WE (E212).* Най-добро припокриване на WE анотациите се постига при наличие на обширни региони от бял епител (Dice = 0.810-0.838).



Фигура 39. Примери от най-доброто представяне на WE класа при многокласовия модел E212.

Разгледаните примери демонстрират, че моделът не страда от архитектурно ограничение, когато входният сигнал е чист - едри мозаечни полета, изразена точковидна структура, бял епител на голяма площ, JNEG контраст. nnU-Net е способна на висококачествена сегментация, съпоставима с други разработени системи, които са обсъдени в следващия раздел.

4.4.2 Сравнение с литературни данни

Сравнението се базира на изложения вече литературен преглед, систематичен обзор на Takahashi и сътр. (2025), разглеждащ 19 проучвания от периода 2019-2024 г., и мета-анализ на Liu и сътр. (2024), включващ 26 колпоскопски проучвания до 2024 г. Установено е припокриване от 10 проучвания в двете публикации, а общият брой уникални колпоскопски статии е 35.

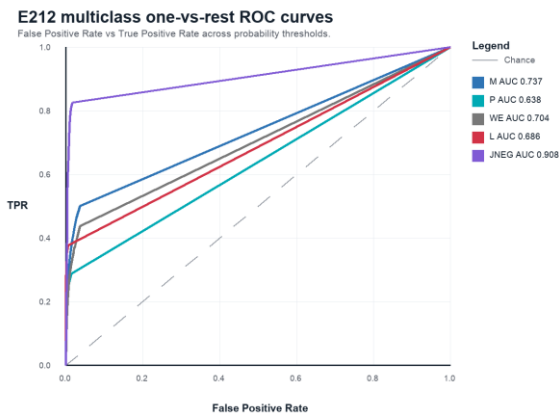
4.4.2.1 Сравнение по ROC-AUC профил с описани в литературата системи

Според обобщените данни от мета-анализа на Liu и сътр. докладваната обединена стойност на AUC е 0.91 (95% CI 0.88-0.93) за обхванатите AI-базирани диагностични системи.

Система	Тип задача	Обем	AUC
E212 многокласов модел	Пикселна сегментация на JNEG	355 изображения	0.908
E202 бинарен модел	Бинарна пикселна сегментация	355 изображения	0.767
Fang и сътр. (2022)	5-класов модел за класификация на ниво изображение	6996 изображения от 1189 пациенти	0.99
Fan и сътр. (2022)	3-класов модел за класификация на ниво изображение	7106 изображения	0.9856
Mascarenhas и сътр. (2024)	HSIL срещу LSIL класификация	22 693 кадъра	0.98
Xue и сътр. (2020)	Степенуване на ниво изображение	101 267 изображения от 19 435 пациенти	0.947
Yue Z и сътр. (2020)	Степенуване според CIN класификация	4753 изображения	≥ 0.94
Yuan и сътр. (2020)	Хибридна система (класификация и сегментация)	22 330 случая	0.93
Fu и сътр. (2022)	Мултимодална система на ниво изображение	2160 пациенти	0.921
Hu и сътр. (2019)	Класификация на ниво CIN2+	9406 изображения	0.91
Cho и сътр. (2020)	Бинарна класификационна система HSIL+/LSIL-	1426 изображения от 791 пациенти	0.781

Таблица 18. ROC-AUC сравнение срещу описани в литературата системи

Постигнатият AUC от 0.908 за JNEG класа попада в докладвания диапазон от Liu и сътр., но на пикселно ниво, което е фундаментално по-трудна задача от класификация на ниво изображение. Бинарният модел (E202) достига AUC от 0.767 резултат, който е сравним със проучването на Cho и сътр. (0.781), което прилага двупосочен подход (HSIL+ срещу LSIL-), но на ниво изображение и в условия на четири пъти повече данни.



Фигура 40. Конвенционални ROC криви по схемата „един срещу всички“ (one-vs-rest) за експеримент E212. Кривата на JNEG е ясно отделена от по-слабите криви на P, WE и L.

ROC анализът показва съществени различия между отделните класове по отношение на тяхната разпознаваемост от модела. (Фиг.40) JNEG демонстрира най-висока дискриминационна способност (ROC-AUC 0.908) с крива, рязко отделена в горния ляв квадрант. M и WE заемат междинна позиция (ROC-AUC 0.737 и 0.704), а P и L остават най-трудни за разграничаване (0.638 и 0.686) със стойности, доближаващи се до диагонала. Тази класова йерархия е стабилна и съвпада с Dice резултатите, което показва, че ограничението не е в избора на метрика, а в самата информация, която всеки клас носи. Основният извод е, че един общ многокласов модел трябва едновременно да усвои класове с коренно различна визуална разпознаваемост, което представлява съществено предизвикателство за обучението.

4.4.2.2 *Сегментационен анализ и сравнение по Dice стойности с описани в литературата системи*

ROC-AUC анализът осигурява всеобхватен поглед, позволяващ съпоставка с хетерогенни по своята задача литературни системи. За оценка на сегментационната точност обаче, следващата стъпка е да се сравни пространственото припокриване между предсказаните и аотираните зони. Моделът е проектиран да оперира в многокласова среда, докато по-голямата част от публикуваните проучвания дефинират сегментацията като бинарна задача - лезия срещу фон. Това разгръща сравнението в две

посоки: мултикласова и бинарна. Първо, E212 се съпоставя с проучвания, докладващи Dice за сегментационния си компонент, като се отчита принципната разлика в броя и дефиницията на класовете. Второ, E202, като методически най-чист паралел, привежда задачата към общия знаменател на литературата и позволява разграничаване на ефекта на многокласовата сложност от ефекта на ограничения обем данни.

От всички литературни източници само три докладват коефициент за пространствено припокриване: Yue Z и сътр. (2021), Yu и сътр. (2022) и Yuan и сътр. (2020). (Табл. 19)

Система	Сегментационна цел	Dice/IoU	AUC	Обем	Бележка
E212	M/P/WE/L/JNEG	0.803 (JNEG) / 0.256 (macro)	0.908 (JNEG)	355	Многокласов модел
E202	Лезия/фон	0.427 (pixel) / 0.273 (macro)	0.767	355	Бинарен модел
Yue Z и сътр. (2021)	Сегментация на оцетнобели региони	0.823	0.973*	3045	Бинарна оцетнобяла сегментация
Yu и сътр. (2022)	LSIL+	0.737	N/A	5455	Бинарна LSIL+ сегментация
Yuan и сътр. (2020)	Лезия	0.616	0.93	22 330	Хибрид сегментация-класификация
Liu J et al. (2022)	Сегментация на оцетнобели региони	0.636 mean IoU/Jaccard	N/A	280	Точност 91.2%, Специфичност 94.9%

Таблица 19. Всички проучвания със сегментационен компонент.

*AUC за класификация

Yue Z и сътр. (2021) постига най-високия Dice в литературата (0.823) при бинарна сегментация на оцетнобели региони (AW зони). Този резултат очертава горната граница на сегментационната точност при достатъчен обем данни и ясна визуална дефиниция на таргета. За целите на сравнението използвахме нашия WE клас като най-близкия аналог на AW зоните. Самостоятелно WE постига среден Dice от едва 0.142, като само 13.3% от случаите превишиха праг 0.50. Тези стойности отразяват не само разликата в обема данни, но и сложността на многокласовата задача. За да

проверим дали класовата фрагментация е основният ограничител, WE и L бяха обединени в общ бял клас (E105), при което Dice достигна до 0.351, което е ръст от близо 2.5 пъти спрямо самостоятелното изследване на WE. Резултатите потвърдиха двата ограничаващи фактора - мултикласовата формулировка намалява Dice с 0.68 спрямо Yue, а малкият обем данни добавя допълнителен спад от около 0.47 дори след премахване на класовата фрагментация.

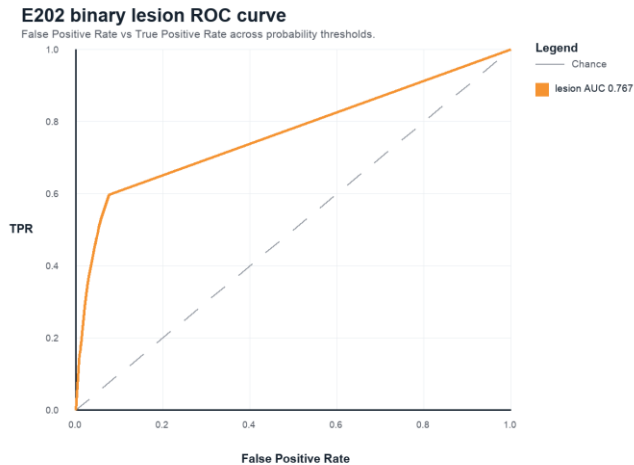
Yu и сътр. (2022) сегментира изображения след приложение на оцетна киселина на фона на поставена бинарна задача LSIL+ срещу нормален резултат. Постигнатият Dice от 0.737 спрямо нашия таско Dice от 0.256 илюстрира цената на разделянето на класове по IFCPC, а именно спад в степента на припокриване с приблизително 0.48 пункта спрямо бинарния подход при съпоставим протокол.

Yuan и сътр. (2020) предлага хибридна система за едновременна сегментация и класификация, постигайки Dice 0.616 за сегментацията и AUC 0.93 за детекцията на ниво изображение. За сравнение, нашият E212 постига макро Dice от 0.256 и макро ROC-AUC от 0.735. Единствената директно съпоставима стойност е JNEG ROC-AUC от 0.908, която се доближава до класификационния AUC на Yuan (0.93), въпреки че произлиза от пикселна мултикласова задача с 63 пъти по-малко данни и йод-контрастен протокол.

Liu и сътр. (2022) извършват оцетнобяла сегментация върху 280 изображения и постигат точност 91.2% и специфичност 94.9%, без да предоставят пространствена метрика. Това проучване е най-близко до нашето по обем на данните, но остава бинарно по същност. Докладваните показатели са доминирани от фоновите пиксели и не отразяват пряко пространственото припокриване на предсказаните маски с референтните анотации, което може да доведе до подвеждащи заключения. Въпреки това, високите стойности при ограничен брой изображения подкрепят извода, че бинарната оцетнобяла сегментация е постижима дори с малък набор, докато мултикласовата IFCPC сегментация изисква значително повече и по-последователно анотирани данни.

Представянето на бинарния модел E202 е значително под литературните стойности (0.616-0.823). (Фиг. 41) Екипът на Yue използва 3045 изображения, на Yu 5455, а на Yuan 22 330, което се равнява на 8 до 60 пъти разлика в количество данни. При толкова големи извадки литературните модели се обучават на значително по-представителна картина на лезионната морфология, което води до по-прецизно маркиране на находки. Въпреки това ROC-AUC на E202 от 0.767 показва, че моделът разпознава лезия от фон и успява да ги отграничи, дори когато пространствената точност, измерена чрез Dice, остава ниска. Това разминаване се дължи на оскъдния брой примери, от които моделът улавя

общия сигнал, но не и прецизните му граници. С повече данни Dice се очаква да нарасне и да съответства на това, което ROC-AUC вече показва.



Фигура 41. Конвенционална ROC крива за експеримент E202

ROC анализът показва, че моделът се представя по-добре от случайното класифициране и успява да разпознава част от лезионните области. Дискриминационната му способност обаче остава ограничена, тъй като не е постигнато едновременно високо ниво на чувствителност и нисък брой фалшиво положителни резултати. Това определя E202 като обещаващ изследователски модел, който се нуждае от допълнителна оптимизация преди евентуално клинично приложение.

4.4.3 Обсъждане

В своя систематичен обзор Takahashi установява висок риск на отклонение при 17 от 19-те проучвания, тъй като е използвана удобна ретроспективна извадка, включваща всички налични случаи, отговарящи на заложените критерии. В мета-анализа на Liu се докладва висока хетерогенност при оценката AI представянето срещу преценката на колпоскопист ($I^2 = 93\%$). Тези ограничения на самата литература трябва да се имат предвид при интерпретацията на всички сравнения по-долу. (Табл. 20)

Параметър	Литературен диапазон	Разработен модел
-----------	----------------------	------------------

Тип задача	Класификация на ниво изображение (CIN2+, HSIL+)	Пикселна сегментация (многокласова)
AUC	0.91 (обединена стойност от мета-анализа на Liu)	0.908 (за JNEG)
AUC (бинарен модел)	0.78–0.91	0.767
Dice	0.823 (Yue Z и сътр.) 0.737 (Yu и сътр.) 0.616 (Yuan и сътр.)	0.803 (JNEG клас) 0.256 (масо Dice всички класове) 0.427 (бинарен модел)
Мултикласова пикселна сегментация по IFCPC	Не е докладвана в литературата	Да (M, P, WE, L, JNEG)
Клиничен обем на данните	134 - 101 267 данни	355 изображения

Таблица 20. Синтезирано сравнение с литературата

Таблицата оформя двете основни линии за дискусия. Първо, JNEG ROC-AUC от 0.908 практически съвпада със средната обединена стойност на AUC от 0.91, докладван от Liu на ниво AI анализ на цяло изображение, но е постигнат на пикселно ниво с 355 изображения. Второ, бинарният модел E202 (AUC 0.767) попада в долната част на литературния диапазон (0.78-0.91), което е очаквано при многократно по-малкото количество данни.

Силното представяне на JNEG класа е съпоставимо с трите сегментационни проучвания с бинарен характер, въпреки че задачата на E212 е 5-класова, която е реализирана с многократно по-малко данни. Това е атестат за възможностите на използваната архитектурата nnU-Net v2 ResEnc-M, но трябва да се анализира в контекста на специфичен цетови протокол. JNEG сегментацията разчита на висок контраст, получен между йод-положителната фонова тъкан в кафяв цвят и йод-негативните участъци в жълто-оранжев цвят, което създава рязка визуална граница. За разлика от нея, споменатите литературни системи сегментират оцетнобели региони, които се проявяват като фини бели градиенти върху розово-червена лигавица, протокол който е крайно различен от Луголовия. Следва да се подчертае, че JNEG трябва да се разглежда като положителна контрола,

потвърждаваща валидността на архитектурата, а не като пряк аналог на AW-базираните системи в литературата. Смисленото сравнение с тези проучвания следва да се основава на WE класа (Dice 0.142, ROC-AUC 0.704), който използва същия протокол и значително изостава от публикуваните стойности.

Резултатите от сравнителния анализ разкриват три фундаментални бариери, които ограничават клиничната стойност на разработената система. Те не са взаимно изолирани, а действат съвкупно, като всяка от тях засилва ефекта на останалите.

4.4.3.1 *Размер и класова фрагментация на набора от данни*

Литературните сегментационни системи, постигащи висока степен на припокриване в колпоскопията, разчитат на 3045 - 5455 изображения, обучени за решаване на една бинарна задача (AW/норма или LSIL+/норма). При 355 изображения, разпределени в пет класа, разработеният модел разполага със средно 70 изображения на клас, като най-редкият клас L е представен от едва 44 случая. Изследванията върху кривите на обучение при медицински изображения показват, че за постигане на стабилна точност над 98% са необходими приблизително 1000 анотирани примера на клас, като при по-малки извадки моделът демонстрира висока вариабилност и нестабилни резултати - ситуация, която пряко кореспондира с представянето на нашия модел. (Cho et al., 2015)

Класовата фрагментация допълнително усилва ефекта на малкия обем. Експериментът E105, при който WE и L са обединени в общ "бял" клас, демонстрира нарастване на Dice от 0.142 до 0.351 (ръст от 2.5 пъти). Това показва, че моделът разполага с достатъчно информация за идентифициране на абнормност, но не и за фината диференциация между подтипозите, която изисква по-голям и по-разнообразен набор от анотирани примери. Класовият дисбаланс между M (168 полигона) и L (44 полигона) означава, че моделът вижда почти четири пъти повече мозаечни примери отколкото левкоплакични, което създава систематична предразположеност към по-често предсказване на M за сметка на по-рядките класове. Двата обзора потвърждават тази оценка - Takahashi изрично посочва, че малките и демографски хомогенни набори ограничават възможността за обобщаване на резултатите, докато Liu препоръчва развитието на AI системи да се насочи и адаптира към среди с ограничен ресурс, където обемите от данни са традиционно по-малки.

4.4.3.2 *Анотационна субективност и концептуално припокриване на класовете*

Клиничната същност на колпоскопските находки е спектрална, резултат от динамичното съревнование между човешкия папиломен вирус и защитните възможности на организма. Мозаицизмът (М) и пунктуацията (Р) представляват съдови промени, които се развиват върху бял епител (WE), а левкоплакията (L) може да включва както WE, така и по-дълбоки стромални промени. Това означава, че в клиничната реалност тези находки често съществуват едновременно върху едно и също изображение и се припокриват пространствено. Количеството обучителни данни, необходими за стабилна сегментация, зависи от визуалната сложност и припокриването на класовете. За хомогенни, пространствено изолирани структури (напр. кости) nnU-Net постига стабилни резултати с 10-30 случая, а за визуално припокриващи се класове със спектрални характеристики изискванията нарастват значително, обусловени от фините дискриминативни различия между подтипове. (Huysentruyt et al., 2024)

Анотациите, изготвени от един експерт, отразяват тази клинична нееднозначност: полигоните за М често са разположени вътре в по-големи WE региони, което създава конфликт при softmax формулировката. Софтуерът разрешава конфликта чрез йерархичен приоритет ($L > M > P > WE$), което принудително елиминира част от анотационната информация при всеки обработен кадър. Експериментите E000 и E106 предоставят директно количествено доказателство за този ефект. Приложената в тях промяна на softmax приоритета от $L > M > P$ на $M > P > L$ води до 17% спад на Dice (от 0.154 на 0.128) без никаква промяна в клиничната картина, анотациите или архитектурата. Тази чувствителност към техническо решение, което няма клиничен аналог, показва, че формулировката не отразява клиничната реалност, при която IFPCS находките се припокриват.

Високата хетерогенност в анотациите се отразява и в разпределението на Dice стойностите. Медианният Dice за Р е 0.000, което означава, че в повече от половината валидационни случаи моделът не открива никакво припокриване с еталонната анотация. Това не е задължително грешка на модела, а отражение на фундаменталната нееднозначност при маркирането на фините пунктуационни структури, които могат да бъдат интерпретирани по различен начин дори от опитни колпоскописти.

4.4.3.3 *Липса на метаданни и протокол-зависим сигнал.*

JNEG доминира мултиклассовия сигнал поради протоколна, а не морфологична причина. Без метаданни от заснемането моделът се обучава върху смесен набор от изображения с коренно различни визуални характеристики, което го насочва към експлоатиране на контрастния сигнал вместо към научаване на морфологичните белези на лезията. По тази причина JNEG не може да бъде оценяван коректно при отсъствие на протоколна информация, тъй като не е ясно на кои изображения е приложим йодният етап от колпоскопското изследване. Това ограничение е в съответствие с оценката на Takahashi, че повечето от включените в неговия обзор проучвания не докладват подробно условията на заснемане, като последователност и вид на приложените разтвори, времето след апликацията и условията на осветление. Неговата препоръка за стандартизация показва, че липсата на метаданни е приета от общността като системен проблем в областта, който ограничава както сравнимостта между проучванията, така и възможността за надеждна оценка на клиничната приложимост на разработените алгоритми. Въвеждането на протоколни метаданни в бъдещите данни би позволило индивидуална оценка, което е предпоставка за клинично значима интерпретация на резултатите.

4.5 ПЕТА ЗАДАЧА: Да анализираме клиничната приложимост и мястото на изкуствения интелект в съвременната колпоскопска диагностика.

Приложението на изкуствен интелект в диагностиката на предраковите изменения на маточната шийка се развива в отговор на няколко трайни ограничения на конвенционалната колпоскопия - субективност на интерпретацията, междубнаблюдателна вариабилност, зависимост от опита на изследващия и неравномерен достъп до експертна колпоскопска оценка. Областта преминава през няколко етапа от ранни опити за количествен текстуриран и цвятанализ, през първите модели за дълбоко обучение, до съвременните мултимодални системи за сегментация, класификация и биопсично насочване. Тази еволюция отразява промяна в самата концепция за компютърно подпомагана колпоскопия и бележи преход от изолирано разпознаване на абнормен образ към подпомагане на конкретни клинични решения.

Включените в систематичният обзор на Takahashi и сътр. модели показват широк диапазон на диагностична ефективност, като докладваната чувствителност варира от 62.1% до 99.7%, а специфичността - от 56.6% до 98.6%. Поради ретроспективния характер на проучванията и подбора на налични и удобни за анализ данни съществува риск от изкривяване на резултатите, свързано с характеристиките на включените пациенти. Обзорът подчертава значителната методологична разнородност между статиите: различни популации, референтни стандарти, дефиниции за положителен резултат, модели и контрол върху качеството на изображенията. Тези резултати демонстрират значителен потенциал, но не могат да бъдат интерпретирани еднозначно като доказателство за готовност за рутинно клинично приложение.

Най-същественият количествен синтез до момента е мета-анализът на Liu и сътр. Данни за точността на изкуствения интелект са извлечени от 26 проучвания, а за опитни колпоскописти от 14 проучвания. За чувствителност, специфичност, положителна и отрицателна прогностична стойност са използвани съответно 16 проучвания за AI и 8 проучвания за опитни колпоскописти. При сравняване с човешката експертна оценка авторите изключват младшите колпоскописти, а когато са участвали няколко специалисти, използват средна стойност на постигнатата от тях диагностична точност. Извършената съпоставка между представянето на опитни колпоскописти и резултатите на AI-асистираната колпоскопия отчита, че изкуственият интелект постига по-висока обобщена точност (81% спрямо 74%) и по-висока специфичност (83% спрямо 67%), при сходна чувствителност (86% спрямо 85%). Това означава, че

публикуваните AI системи имат потенциал да намалят фалшиво положителните оценки и да подпомогнат стандартизирането на колпоскопската преценка. В подгруповия анализ AI превъзхожда клиницистите както за LSIL+ задачата, където достига 84% точност, 86% чувствителност и 85% специфичност, така и при HSIL+ праг - 81% точност, 89% чувствителност и 86% специфичност. Високата хетерогенност между включените проучвания налага резултатите да се разглеждат като изява на клиничен потенциал, а не като основание за автономно функциониране. Голяма част от наличните разработки са ретроспективни, едноцентрови или валидирани върху ограничени и добре контролирани набори от изображения. Реалната стойност на AI може да бъде определена само чрез проспективни, многоцентрови проучвания с външна валидация, които отчитат описаните променливи условия и реалния работен процес.

Колпоскопията остава незаменим диагностичен метод с безспорна клинична стойност, която обаче е силно зависима от опита на изследващия и последващата субективната интерпретация на визуализираните промени. Тази вариабилност има пряко значение за настоящата дисертация, защото показва, че проблемът в колпоскопията не е само диагностичен, а има изразен семиотичен характер. Оцетнобелият епител, мозайката, пунктуацията не са напълно дискретни категории. Те често съществуват едновременно, преливат една в друга и зависят от редица фактори - време след апликация на оцетна киселина, качество на приложения разтвор, технически условия и др. Това обяснява защо чисто образната оценка може да бъде непостоянна дори между експерти. В този контекст, алгоритъм може да прилага една и съща визуална логика върху голям брой изображения, да маркира подозрителни участъци, да измерва площ и вероятностен сигнал, да съхранява възпроизводима информация и да осигурява вид „второ мнение“. Това не елиминира нуждата от колпоскопист, но може да намали непостоянството в онези елементи на метода, които са най-силно свързани с човешката преценка.

Анализът на литературата позволява да се дефинират няколко практически направления, в които AI може да бъде интегриран в клиничния процес - като интелигентен асистент за оценка на суспектни находки и избор на биопсични цели, за обучение и стандартизация и като решение в условия с ограничен достъп и медицински ресурси.

4.5.1 Асистирана колпоскопска оценка.

AI може да подпомогне лекаря чрез маркиране на подозрителни зони, оценка на вероятността за високостепенна лезия и визуализиране на области, които заслужават внимание. В рандомизираното двуцентрово клинично проучване Kim и сътр. показват подобрена локализационната

AUC с 0.12, повишаване на чувствителността от 71.33% на 89.18%, на специфичността от 92.16% на 96.68%, а общата диагностична точност нараства от 75,45% на 86,40%. като най-голяма полза се наблюдава при по-малко опитните колпоскописти. Резултатите подкрепят възможността изкуственият интелект да допълва експертната преценка както при диагностичната класификация, така и при пространственото локализиране на suspectните участъци.

4.5.2 Биопсично насочване

Колпоскопията има клинична стойност не само при визуализация на абнормна находка, а и при насочване към зоната с най-изразени изменения за последващо биопсично изследване. Според ASCCP стандартите между две до четири таргетни биопсии подобряват откриването на налични предракови изменения. (Wentzensen et al., 2017) Това подчертава значението на правилния избор на локализации за хистологично изследване. CAIADS системата на Xue и сътр. демонстрира именно такава възможност, като постига mIoU 0.758 за предсказване на биопсични места. Това показва, че една от най-перспективните клинични функции на AI е пространственото насочване, а не само поставянето на бинарна диагноза. Приносът на подобна система е особено изразен при неопитни колпоскописти или при изображения с нееднозначни находки.

4.5.3 Обучение и стандартизация

AI системите предлагат възможност за стандартизиране на колпоскопската оценка чрез обективен и възпроизводим анализ на изображенията. В образователен контекст AI може да служи като инструмент за обучение на специализанти, като осигурява незабавна обратна връзка спрямо експертни анотации. Сегментационните карти са особено ценни, защото визуализират пространственото разположение на находките. Това ги прави подходящи за обучение, за сравнение между експертна анотация и AI предложение, както и за развитие на визуална компетентност. По този начин може да се намали зависимостта от индивидуалния опит и да се създаде възможност за структурирано обучение между различни центрове. Проучването на Wang и сътр. (2024) сред 284 китайски колпоскописти от 24 провинции установява висока готовност сред участниците за използване на AI-асистирани системи, особено когато технологията се възприема като помощно средство, а не като заместител на лекаря. От тях 82.4% предпочитат колпоскопията да разчита изображението с помощта на CAIADS, 15.49% приемат CAIADS да разчита под лекарски контрол, а само 0.70% биха оставили системата да

разчита самостоятелно. На фона на тези резултати едва една трета от анкетираните познават системата и актуалното ѝ развитие, което показва едновременно възприемчивост към нововъведения и нужда от обучение и организационна подготовка.

4.5.4 *Подобряване на достъпа в слаборесурсни условия*

Глобалната стратегия на СЗО се стреми да понижи честота до 4 нови случая годишно на 100 000 жени. Постигането на това изисква по-голям диагностичен капацитет, който AI има потенциала да осигури, но не като компромис с установените норми, а под формата на стандартизирана подкрепа. В региони с ограничени ресурси AI-базираните колпоскопски системи могат да запълнят критичната липса на обучени специалисти. Shamsunder и сътр. (2024) установяват, че алгоритъмът EVA Visual Check, интегриран в портативен дигитален колпоскоп, постига чувствителност 89.3% за откриване на CIN2+, близка до тази на лекарската колпоскопска оценка - 92.9%. Значително по-ниската специфичност на AI системата спрямо лекарите (26.1% срещу 68.9%) обаче показва, че тя е по-подходяща за първоначален триаж и насочване за експертна оценка, отколкото за самостоятелно поставяне на диагноза. Подобни системи, интегрирани в преносими колпоскопи, смартфон-базирани устройства или облачни платформи, могат да намалят зависимостта от непосредственото присъствие на експерт и да демократизират достъпа до колпоскопска диагностика в отдалечени райони.

4.5.5 *Обсъждане*

Разработеният в настоящия труд модел следва да бъде позициониран като изследователска система за пикселна сегментация на колпоскопски находки, с потенциално клинично приложение в асистираната колпоскопия и обучението.

Основната концептуална разлика с голяма част от публикуваните системи е формулировката на задачата. Повечето литературни модели работят на ниво изображение или пациент и класифицират резултата в двоично направление: нормален-абнормен резултат или нискостепенна-високостепенна лезия. За разлика от тях, нашият опит е за възпроизводство на пространствената задача на колпоскописта, имитирайки мисловните стъпки на „естествения“ човешкия алгоритъм. Доколкото ни е известно,

това е една от първите системи, насочени към многокласова пикселна сегментация, структурирана по IFCPC номенклатура.

Бинарната класификация отговаря на въпроса дали изображението е нормално или патологично, но често не обяснява защо. IFCPC-базираната сегментация, макар и по-трудна и с по-ниски текущи метрики, се стреми да направи решението проследимо чрез локализация на конкретните колпоскопски зони. Тя представлява не просто алтернативен технически подход, а опит за изграждане на по-разбираем, клинично проверим и по-близък до реалната колпоскопска логика изкуствен интелект. За практикуващия лекар е от съществено значение не само какъв краен резултат предлага алгоритъмът, а по какъв начин е достигнал до него, върху кои образни характеристики се основава и дали заключението може да бъде проверено. Именно тук се проявява един от основните проблеми на съвременните AI решения в медицинската образна диагностика - изграждането им като системи тип „черна кутия”. Такъв подход може да бъде статистически ефективен, но остава ограничен от гледна точка на клиничната интерпретация и доверието на лекаря. Тази позиция е в съответствие с Wang и сътр. които подчертават, че клиничното внедряване на AI в цервикалната диагностика изисква не само висока алгоритмична точност, но и обясними интерфейси в т.нар. clinician-in-the-loop условия, при които лекарят остава активен участник в диагностичното решение.

Най-важният извод от собствените резултати не е, че моделът е готов за клинична употреба, а че ясно показва границата между технологичната възможност и клиничната готовност. JNEG класът подкрепя извода, че използваната архитектура може да извлича стабилен визуален сигнал при добре контрастирани и консистентно анотирани находки. Ниските резултати при P, WE и L показват, че ограниченията в общата производителност са свързани не само с модела, а преди всичко с обема, структурата и анотационната консистентност на данните, липсата на протоколни метаданни и концептуалното припокриване между отделните колпоскопски белези. Затова дори при по-ниски стойности на Dice коефициента, многокласовата сегментация запазва важна научна и клинична стойност. Способността за припокриване не трябва да се интерпретира като неуспех на модела, а като отражение на сложността на задачата - неясни граници между находките със сходен сигнал между тях, количествен дисбаланс на класовете, ограничен брой анотирани изображения и субективност дори при експертната човешка оценка. Следователно представената разработка не трябва да се разглежда само през призмата на абсолютната сегментационна ефективност, а като стъпка към по-прозрачен изкуствен интелект в колпоскопията. В реалната клинична практика доверието към една AI система би било по-високо, когато тя не просто маркира вероятност за заболяване, а визуализира

конкретните находки, върху които се основава. Така IFPCP-базираната сегментация може да служи като междинен обясним слой между изображението и диагностичното решение, като запазва ролята на лекаря в интерпретацията на резултатите.

Преминването от изследователски прототип към реална клинична стойност изисква последователно адресиране на няколко приоритета. Първият е мащабирането и балансирането на данните. Настоящата база данни е достатъчна за доказване на концепцията, но е недостатъчна за стабилно обучение на фини и редки класове. Публикуваните високоефективни системи обичайно използват многократно по-големи набори, което показва, че бъдещото развитие трябва да включва повече изображения, балансирани по клас, протокол и клиничен изход.

Вторият приоритет е въвеждането на протоколни метаданни. Липсата на информация дали дадено изображение е след оцетна киселина или след Лугол, какво е времето от апликацията, степента на увеличение и типът на трансформационната зона, лимитира способността на модела да разграничава.

Третият е подобряване на анотационната стратегия и начина, по който представяме на модела клиничната реалност. За класовете е необходим мултианотаторски консенсус, при който степента на междуоператорно съгласие да се измерва количествено с ясни заключения върху дискусабилни региони на интерес. При частично пространствено припокриване на находки една зона може да съдържа елементи от няколко анотации, което технически погледнато, е ситуация, при която един пиксел принадлежи на повече от един клас. В тези случаи йерархичен подход с множество етикети е по-подходящ от класическата софтвакс формулировка, която принуждава модела да избере само един клас и автоматично изключва останалите.

Четвъртият приоритет е провеждането на клинична валидация за оценка на реалната стойност на модела. За целта изображенията трябва да бъдат свързани с пациентска карта, включваща HPV статус, цитология, хистология, вид на биопсията или ексцизионната процедура и проследяване. Само така може да се премине от оценка на пикселното припокриване към оценка на клиничната полза.

Последният приоритет е външна, проспективна мултицентрова валидация, без която не може да се докаже, че моделът работи извън условията, в които е обучен. Това е ключово изискване както от научна, така и от регулаторна гледна точка. В Европейския съюз AI Act определя AI-базирания софтуер, предназначен за медицинска употреба, като високорискова система, изискваща управление, висококачествени данни, ясна информация за потребителя и човешки надзор. Това означава, че

клиничното внедряване е не само технически, но и регулаторен, етичен и организационен процес.

Въз основа на извършения литературен обзор и резултатите от модела, обучен върху собствена база данни, може да се заключи, че изкуственият интелект има своето място в съвременната колпоскопска диагностика на предракови изменения на маточната шийка. На настоящия етап той е подпомагащ инструмент с нарастващо значение, но не може да бъде разглеждан като самостоятелен заместител на колпоскописта.

Мястото на изкуствения интелект в колпоскопията следва да се разглежда не като конкуренция на лекаря, а като средство за повишаване на обективността и стандартизацията на диагностичния процес в рамките на взаимодействие, при което колпоскопистът запазва водещата клинична роля.

5. ПРИНОСИ

Научно-теоретични

- 5.1** Формулирано е мястото на изкуствения интелект в съвременната колпоскопска практика, определени са основните направления за потенциално приложение, предпоставките и предизвикателствата за клиничното му внедряване.

Научно-приложни

- 5.2** За първи път в България е разработена и експериментално изследвана оригинална система, базирана на изкуствен интелект, за автоматизирана сегментация на пикселно ниво на колпоскопски находки на маточната шийка.
- 5.3** Създадена и структурирана е собствена база данни от колпоскопски изображения, подходяща за обучение и сравнително изследване на сегментационни модели чрез методите на изкуствения интелект.
- 5.4** Разработен е унифициран протокол за експертна анотация на колпоскопски изображения, включващ дефиниране на целевите класове, пространствено очертаване на находките, преобразуването им в пикселни маски, одит на структурата и класовото разпределение на данните.

Научно-практични

- 5.5** Извършено е комплексно изследване на приложимостта, техническото и диагностичното представяне на класификационни, сегментационни и мултимодални модели с

изкуствен интелект за автоматизиран анализ на колпоскопски изображения на маточната шийка.

- 5.6** Извършена е контролирана експериментална оценка на влиянието на основните фактори върху сегментационното представяне - многокласова срещу бинарна формулировка, обединяване и изключване на класове, приоритетна йерархия при припокриващи се анотации, увеличено представяне на редките класове, фотометрични аугментации и локално изрязване на областите около находките.
- 5.7** Разработката е заявена за патентна закрила пред United States Patent and Trademark Office № 63/991,571 от 26.02.2026 г.

„Ако очакваме от една машина да бъде безпогрешна, отнемаме от възможността ѝ да бъде интелигентна.“

Алън Тюринг

20 февруари 1947 г.

Декларация за оригиналност

Авторът на настоящия дисертационен труд декларира, че всички описани данни са оригинални и получени в резултат на изследователската му работа в Медицински университет – Плевен и УМБАЛ „Света Марина – Плевен“. Резултатите, обсъжданията и изводите не са заимствани от други източници без цитиране.

Декларатор: д-р Георги Даниелов Праднжев

6. НАУЧНИ ПУБЛИКАЦИИ И УЧАСТИЯ

ПУБЛИКАЦИИ, СВЪРЗАНИ С ДИСЕРТАЦИОННИЯ ТРУД

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>.

НАУЧНИ СЪОБЩЕНИЯ, СВЪРЗАНИ С ДИСЕРТАЦИОННИЯ ТРУД

1. 19-21.06.2025 г. – Втори Национален Конгрес на Българската Асоциация по Минимално Инвазивна и Роботизирана Хирургия (БАМИРХ) с международно участие с постер *„Приложение на изкуствен интелект в съвременната колпоскопия – първи клиничен опит“*, Медицински университет – Плевен Телекомуникационен Ендоскопски Център (ТЕЛЕЦ)„ гр. Плевен
2. **03.11.2024г** - Юбилейната научна конференция "50 години медицинско образование и наука в Плевен", с доклад *„Приложение на изкуствен интелект за детекция на предракови изменения на маточната шийка“*, гр. Плевен.

3. **15.06.2024г** - Единадесетата годишна научно-практическа конференция на Асоциацията на оториноларинголозите от Югоизточна България (АОРЛЮБ) с доклад на тема „*Приложение на изкуствен интелект в онкологията*“, гр. Китев.
4. **10-12.05.2024г** - Първа Национална Конференция Плевенски Минимално Инвазивни Дни на Българската Асоциация по Минимално Инвазивна и Роботизирана Хирургия (БАМИРХ) с международно участие, с доклад на тема „*Изкуствен интелект в минимално инвазивната гинекологична хирургия*“, Телекомуникационен Ендоскопски Център (ТЕЛЕЦ), Медицински университет – Плевен, гр. Плевен.
5. **14.05.2024г** - Гражданска инициатива “Здравен барометър” под патронажа на проф. Даниел Вълчев, декан на Юридическия факултет на Софийския университет „Св. Климент Охридски“, с представяне на презентация “*Приложение на изкуствен интелект в гинекологията*”, х-л „Интерконтинентал“, гр. София.
6. **12.11.2023г** - 14-та годишна Конференция СОНМ – МОРЕ 2023, организирана от Съвместна онкологична национална мрежа (СОНМ), с презентация на тема „*Революция в скрининга на рака на маточната шийка*“ към демо-сесия „Панорама на здравни технологии в българската онкология“, гр. Варна.
7. 04.06.2022 г. - Пленарна лекция на тема „*Изкуственият интелект в медицината*“ на Деветата Национална конференция на БАМИГХ, Медицински университет – Плевен, ТЕЛЕЦ.

УЧАСТИЕ В НАУЧНОИЗСЛЕДОВАТЕЛСКИ ПРОКЕТИ

Проект BG-175467353-2022-04-0257-C01 „Приложение на методите на изкуствения ИНТЕЛект за колпоскопско разпознаване и КАтегоризиране на характеристиките на МАТочната шийка – ИнтелКАМат“, финансиран от Фонд „Научни изследвания“ по процедура конкурс за финансиране на фундаментални научни изследвания – 2022 г.

НАГРАДИ И ОТЛИЧИЯ

1. 31.10.2022г - Победител в **European Institute of Innovation and Technology (EIT) Health i-days**, организирано в Медицински университет – Плевен с награда участие на международното състезание EIT Health Европа. I-days е сътрудничество на Junior Achievement Bulgaria и нейната spin-off компания The Edge: R&BD, в партньорство с EIT Health и различни университети в България.
2. 24-25.11.2022г – Финалист на споменатото международно състезание **European Institute of Innovation and Technology (EIT)** в гр. Виена, Австрия, с участници от 24 държави, където достигнах до финал, трето място и финансова награда.
3. 02.06.2023г - **Първо място** в предприемаческа програма **Beyond pre-accelerator**, организирана от The Edge, Junior Achievement

Bulgaria, за разработен проект за пазарна реализация на научен дисертационен труд – *„Приложение на изкуствен интелект за ранна диагностика на предрак на маточната шийка“*.

4. 14.07.2023г - Спечелена награда **Intel AI Innovators Award 2023**, връчена от **Intel Corporation®** за разработка на *"Изкуствен интелект за ранна диагностика на предрак на маточната шийка"* и представяне на международен форум GEN-E 2023 в Инстанбул, Турция, организиран от Junior Achievement Европа.
5. 09.2023г - Спечелена награда на **Intel® AI Global Impact Festival 2023 - Country Recognition Award** *"Cervical Pre-Cancer Detection"* регионална награда за **България** за принос в разработване на изкуствен интелект за детекция на предракови изменения на маточната шийка.
6. 04-06.09.2023г – Командировка на основание заповед № РД- 19-46 / 28.08.2023г в гр. Саламанка, Кралство Испания със задача: **„Придружаване на министъра на иновациите и растежа**, на най-голямото европейско технологично събитие „Startup Olé 2023“. Трето място на описаното презентационно състезание в графа **Humanitarian Technologies/Social Impact** с презентация, посветена на приложение на изкуствен интелект за ранна диагностика на предрак на маточната шийка.
7. 05.11.2025г – в Брюксел предстои получаване на награда след номинация **във FERD's List**, сектор здравеопазване - признание за вдъхновяващи млади предприемачи и лидери в Европа. Това е

инициатива на Junior Achievement Europe в сътрудничество с FERD, която цели да почете постиженията, предприемаческия път и вдъхновяващото лидерство на завършили програмите на JA.

ПОПУЛЯРИЗИРАНЕ НА РЕЗУЛТАТИ

12.2023г - Списание **FORBES България 30под30** статия-интервю „Изкуствен интелект за диагностика на рак“, брой 12/2023, стр. 62-62.

31.01.2024г - БНТ2, телевизионна интервю за предаване: Знание.БГ, по темата на дисертационния труд, част от репортаж, посветен на 50 години Медицински университет-Плевен (04-06:00 мин.)

08.02.2024г - Bloomberg TV Bulgaria, телевизионно 25-минутно интервю за популяризиране на научната идея, заложена в дисертационния труд, предаване „Светът е бизнес“ с водещ Ивайло Лаков.

16.06.2024г - Bulgaria ON AIR, 10 – минутно телевизионно интервю по темата на дисертационния труд: “Изкуственият интелект помага в ранната диагностика на ракови състояния”.