
Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра системного программирования

Направление подготовки / специальность: 03.03.01 Прикладные математика и физика

Направленность (профиль) подготовки: Математическая физика, компьютерные технологии и
математическое моделирование в экономике

ПРИМЕНЕНИЕ НЕЙРОСЕТЕВОЙ МОДЕЛИ ДЕТЕКЦИИ К ГИСТОЛОГИЧЕСКИМ ИЗОБРАЖЕНИЯМ ДЛЯ ВЫЯВЛЕНИЯ ХРОНИЧЕСКОГО ЭНДОМЕТРИТА

(бакалаврская работа)

Студент:
Макарчук Алексей

(подпись студента)

Научный руководитель:
Мутилин Вадим Сергеевич,
канд. физ.-мат. наук

(подпись научного руководителя)

Консультант (при наличии):

(подпись консультанта)

Москва 2023

Аннотация

Применение нейросетевой модели детекции к гистологическим изображениям для выявления хронического эндометрита

Макарчук Алексей

В работе исследуется применение нейросетевой модели детекции к гистологическим изображениям с целью обнаружения плазматических клеток для выявления хронического эндометрита. Был разработан двухэтапный алгоритм для обнаружения плазматических клеток. На первом этапе была использована нейросетевая модель архитектуры CenterNet для детекции стромальных и эпителиальных клеток. Нейросеть была обучена на открытом наборе данных с гистологическими изображениями и дообучена с использованием дополнительного размеченного набора данных. Был использован протокол разметки, а так же подсчитан коэффициент согласованности двух экспертов, который оказался равен 0.81. На втором этапе с помощью разработанного алгоритма, основанного на методах компьютерного зрения, были определены плазматические клетки, а так же подсчитаны их границы HSV цвета. Для двухэтапного алгоритма были получены следующие метрики качества $\text{precision}=0.70$, $\text{recall}=0.43$, $\text{f1-score}=0.53$. Затем модель была модифицирована для детекции только плазматических клеток и обучена на наборе данных с изображениями, содержащими размеченные плазматические клетки. Метрики качества модифицированной модели детекции $\text{precision}=0.73$, $\text{recall}=0.89$, $\text{f1-score}=0.8$. В результате сравнения, подход с модифицированной моделью детекции показал лучшие метрики качества. Автоматизация работы по подсчету плазматических клеток позволит врачу тратить меньше времени на рутинные действия.

Abstract

Application of a neural network detection model to histological images
to identify chronic endometritis

Makarchuk Aleksey

Study explores the use of a neural network detection model to identify plasma cells on histological images for chronic endometritis diagnosing. There was developed a two-stage algorithm for plasma cells detection. In the first stage, a neural network detection model architecture CenterNet was used to separate stromal cells from glandular cells. The neural network was trained on an open dataset of histological images and fine-tuned using an additional labeled dataset. Annotation rules were agreed upon, also the agreement between two experts-annotators was assessed using Cohen's kappa that turned out to be 0.81. At the second stage, using the developed algorithm, based on computer vision methods, the stained plasma cells were identified and their HSV color bounds were counted. The accuracy of two-stage algorithm turned out to be precision=0.70, recall=0.43, f1-score=0.53. Then the neural network detection model was modified for specifically plasma cells detection and trained on the dataset with images containing labeled plasma cells. The accuracy of modified detection model turned out to be precision=0.73, recall=0.89, f1-score=0.8. As a result of the comparison, the approach with the modified detection model showed the best quality metrics. Automating the work of counting plasma cells will allow the doctor to spend less time on routine activities.

Содержание

1 Глоссарий	6
2 Введение	9
2.1 Актуальность	9
2.2 Нейросетевая модель детекции <i>EndoNet</i>	10
2.3 Оценки Качества	12
2.3.1 Качество классификации	12
2.3.2 Качество детекции и сегментации	13
2.3.3 Качество согласованности	14
3 Постановка задачи	16
4 Обзор существующих решений	17
4.1 Применение нейросетевых моделей детекции для решения медицинских задач	17
4.2 Встроенные методы <i>QuPath</i>	18
4.3 Решение задачи детекции клеток на гистологических изображениях (ИСП РАН)	19
4.4 Окрашивание гистологических изображений двумя маркерами <i>CD138</i> и <i>MUM1</i>	19
5 Исследование и построение решения задачи	23
5.1 Данные	23
5.2 Дообучение нейросетевой модели	23
5.3 Разработка алгоритма поиска кандидатов на плазматические клетки, используя методы компьютерного зрения	24
5.4 Алгоритм поиска плазматических клеток	26
5.5 Модификация <i>EndoNet</i> на детекцию плазматических клеток	28
5.6 Модифицированный алгоритм поиска плазматических клеток	30
6 Описание практической части	32
6.1 Инструменты	32

6.2	Применение нейросетевой модели EndoNet , обученной на наборе данных EndoNuke [1] к изображениям с гистологических слайдов, которые представили врачи-эксперты	33
6.3	Дообучение EndoNet на гистологических изображениях, которые разметили врачи-эксперты (набор данных endometrium)	34
6.4	Разработка алгоритма поиска кандидатов на плазматические клетки, используя методы компьютерного зрения	36
6.5	Сбор и разметка набора данных гистологических изображений с плазматическими клетками	39
6.6	Валидация и подсчет точности алгоритма поиска плазматических клеток	40
6.7	Модификация модели EndoNet на детекцию плазматических клеток . . .	41
6.8	Обучение модифицированного EndoNet на детекцию плазматических клеток	43
6.9	Валидация и подсчет точности модифицированного алгоритма поиска плазматических клеток	44
7	Заключение	46
	Список литературы	47

1 Глоссарий

- Детекция объектов (англ. object detection) - задача машинного обучения, которая заключается в определении и классификации нескольких объектов на изображении путем нахождения координат ограничивающих рамок и определения классов этих объектов из набора заранее известных классов.
- Сегментация (англ. semantic segmentation) - задача машинного обучения, при которой изображение разделяется на отдельные области, каждой из которых присваивается метка класса, указывающая на тип объекта в данной области.
- Голова сегментации (англ. segmentation head) - это часть нейросетевой модели, которая отвечает за сегментацию изображения путем присвоения каждому пикслю определенной метки класса или категории (например, объект или фон). Она обычно является последним слоем нейросетевой модели.
- Веса модели - набор внутренних параметров модели.
- Предобучение (англ. pretraining) - процесс обучения нейронной сети или ее компонентов на данных, которые имеют схожую природу с целевыми данными, с целью выявления общих закономерностей, характерных для предметной области задачи.
- Дообучение (англ. fine-tuning) - процесс настройки или обучения нейронной сети на основе предварительно обученных весов или модели. Это позволяет адаптировать модель к конкретной задаче или набору данных. Вместо обучения модели с нуля, дообучение использует предыдущие параметры модели, чтобы быстрее достичь хороших результатов на новой задаче или данных.
- Аугментация данных (англ. data augmentation) - это метод, который позволяет создавать искусственные данные на основе имеющихся, с целью расширения объема данных, используемых в процессе обучения.
- Эндометрий (лат. endometrium) - это слизистый слой, который покрывает внутреннюю поверхность матки.
- Эпителий – это тип ткани, который выстилает железы. Он характеризуется плотно расположенными ядрами и имеет кольцевидную или вытянутую структуру.

- Строма — это тип ткани, который выступает в качестве опоры для эпителия, покрывающего поверхность эндометрия. При наличии хронического эндометрита, строма часто подвергается воспалительным изменениям.
- Хронический эндометрит - это гинекологическое заболевание, которое может привести к бесплодию и другим тяжелым осложнениям. Это заболевание приводит к изменению тканей эндометрия. Раннее обнаружение хронического эндометрита – это один из наиболее эффективных способов его лечения.
- Гистологические изображения - изображения, которые представляют собой фрагменты ткани, полученные методом гистологического исследования. Гистологические изображения получаются путем удаления тонких слоев ткани, последующей обработки и окрашивания, и фиксации на специальных стеклах. Затем, полученные стекла сканируются и полученные слайды используются для дальнейшего анализа.
- Маркер CD138 - маркер, который окрашивает плазматические клетки в оттенки коричневого цвета на гистологических изображениях.
- Маркер MUM1 - маркер, который окрашивает плазматические клетки в оттенки красного цвета на гистологических изображениях.
- Плазматические клетки - клетки иммунной системы, отвечающие за создание и выработку антител в организме. Детекция плазматических клеток на гистологических изображениях является основным методом в выявлении хронического эндометрита. Важной особенностью плазматических клеток является наличие в них гранул, в которых находятся антитела. Эти гранулы могут быть определены на гистологических изображениях как яркие округлые или овальные объекты вокруг стромальной клетки. И в зависимости от маркера окрашивания могут иметь разные цвета. Например, если окрашивать маркером CD138, плазматические клетки будут иметь коричневый оттенок, а если MUM1, то красный.
- Тайл (англ. Tile) - гистологическое изображение размером 200 x 200 мкм и 790 x 790 пкс.

- HSV цвет - Цветовое пространство, в котором каждый пиксель представлен комбинацией тона (Hue), насыщенности (Saturation) и яркости (Value) цвета.
- EndoNet - нейросетевая модель, разработанная для детекции клеток стромы и эпителия на гистологических изображениях. Эта модель основана на архитектуре CenterNet. Разработка ИСП РАН.
- EndoNuke - большой открытый набор данных с гистологическими изображениями, размеченными профессиональными врачами-гистологами.
- Набор данных plasmatic - набор данных, который содержит 373 гистологических изображения (тайла) с размеченными плазматическими клетками. Изображения получены с гистологических слайдов, окрашенных маркером CD138, предоставленных врачами. Этот датасет используется для обучения, валидации и оценки качества модифицированной на поиск плазматических клеток модели EndoNet.
- Набор данных endometrium - датасет, который содержит 54 гистологических изображения (тайла) с размеченными клетками стромы и эпителия. Изображения получены с гистологических слайдов, окрашенных маркером CD138, предоставленных врачами. Этот датасет используется для дообучения модели EndoNet для более точной детекции клеток стромы и эпителия на предоставленных гистологических изображениях.

2 Введение

2.1 Актуальность

В настоящее время применение нейросетевых моделей стало крайне важным и актуальным во многих сферах деятельности. Нейросетевые модели - это модели, способные обучаться на больших объемах данных и принимать решения на основе полученного опыта. Они имеют огромный потенциал и преимущества в решении различных задач. Во-первых, нейросети позволяют эффективно обрабатывать и анализировать большие объемы данных. Они способны автоматически извлекать полезную информацию, выявлять скрытые закономерности и предсказывать тренды, что помогает в принятии более точных и автоматизированных решений. Во-вторых, нейросети нашли применение во многих отраслях. В последние годы нейронные сети и компьютерное зрение играют все более значимую роль в области медицины. Они предоставляют новые возможности для автоматического и точного анализа медицинских изображений. Одна из важных областей применения нейросетевых моделей в медицине - это область в диагностике различных заболеваний путем анализа изображений, таких как рентгеновские снимки, компьютерная томография (КТ), магнитно-резонансная томография (МРТ) и гистологические срезы. Они могут определять наличие патологий, таких как опухоли, инфекции или другие критические маркеры. В частности, хронический эндометрит, распространенное заболевание женской репродуктивной системы, требует точной диагностики для его эффективного лечения. Он характеризуется хроническим воспалением эндометрия, что может привести к нарушениям репродуктивной функции, бесплодию, а также повышает риск развития рака эндометрия. Раннее выявление хронического эндометрита имеет большое значение для своевременного начала лечения и предотвращения развития осложнений. Выявление плазматических клеток является одним из главных критериев определения наличия хронического эндометрита, поэтому автоматизация процесса детекции плазматических клеток на гистологических изображениях может значительно облегчить процесс диагностики данного заболевания.

В целом, применение нейросетевых моделей в медицине позволяет автоматизировать сложные задачи, повышать эффективность работы врачей и улучшать качество принимаемых ими решений. Они могут значительно ускорить процессы, снизить затраты и сделать работу врачей более эффективной.

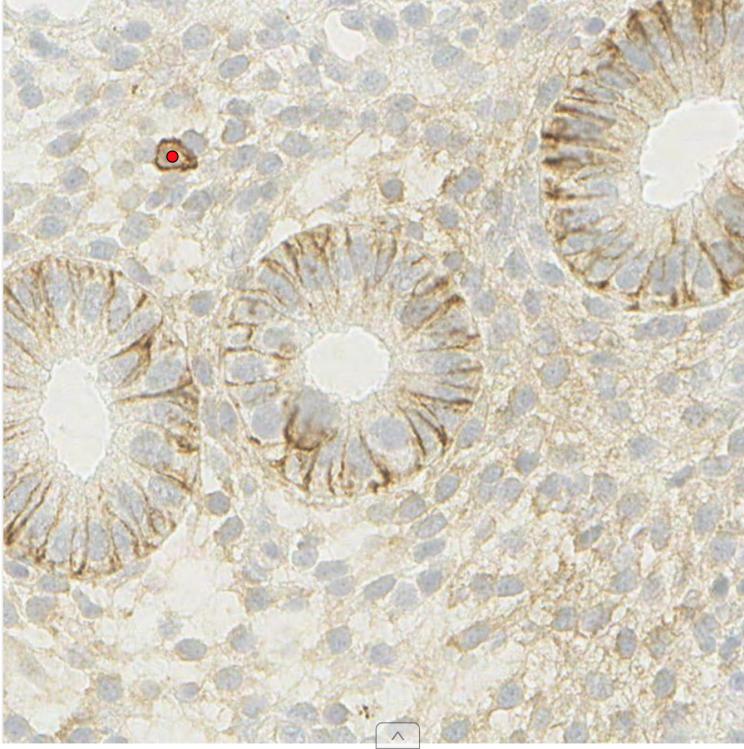


Рис. 1: Пример плазматической клетки на гистологическом изображении (тайле)

2.2 Нейросетевая модель детекции EndoNet

Для решения задачи детекции стромальных и эпителиальных клеток на гистологических изображениях применяется нейросетевая модель EndoNet архитектура CenterNet[2]. Модель принимает RGB изображения в качестве входных данных, и результатом ее работы является множество координат классифицированных ключевых точек - кей-поинтов. Архитектура CenterNet[2] состоит из двух основных компонентов: основы (backbone), которая строит тепловую карту (heatmap[3]) и извлекателя ключевых точек (keypoint extractor), который извлекает набор ключевых точек кей-поинтов из тепловой карты.

Определим понятия, используемые при описании архитектуры CenterNet[2]:

- Основа (англ. Backbone): Пусть X - входное изображение размером $H \times W \times C$, где H , W и C соответствуют высоте, ширине и числу каналов изображения соответственно. Backbone представляет собой сверточную сеть, которая принимает на вход изображение X и генерирует карты признаков F с помощью серии сверточных операций. Тогда $F \in \mathbb{R}^{H' \times W' \times D}$ - это набор карт признаков, где H' , W' и D

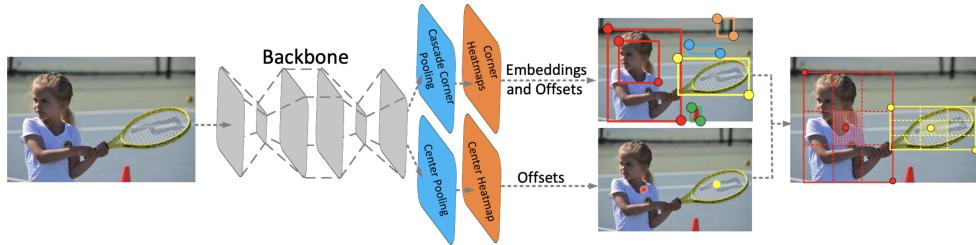


Рис. 2: Архитектура CenterNet

соответствуют новой высоте, новой ширине и числу признаков соответственно.

- Каскадный пулинг углов (англ. Cascade Corner Pooling[4]): Каскадный пулинг углов применяется к картам признаков F для генерации двух карт углов. Обозначим эти карты как $H_c \in \mathbb{R}^{H' \times W'}$ и $W_c \in \mathbb{R}^{H' \times W'}$, которые представляют вертикальные и горизонтальные углы соответственно для каждой позиции на карте признаков.
- Пулинг центра (англ. Center Pooling[4]): Пулинг центра применяется к картам признаков F для генерации карты центров ключевых точек. Обозначим эту карту как $C \in \mathbb{R}^{H' \times W'}$, которая представляет собой карту значений центров для каждой ключевой точки.
- Определение потенциальных ограничивающих рамок: С использованием обнаруженных углов определяются потенциальные ограничивающие рамки. Алгоритмы для определения рамок могут отличаться в зависимости от конкретной реализации CenterNet[2].
- Определение окончательных ограничивающих рамок: Обнаруженные центры ключевых точек, а так же обнаруженные потенциальные ограничивающие рамки используются для определения окончательных ограничивающих рамок.

CenterNet[2] — это одноэтапный детектор объектов, который определяет каждый объект как тройку ключевых точек. Сначала входное изображение обрабатывается **backbone** частью. **Backbone** часть генерирует карты признаков (англ. feature maps) с помощью серии сверточных операций. Эти карты признаков отражают высокоуровневые представления входного изображения. Затем происходит каскадный пулинг углов для обнаружения углов потенциальных ограничивающих рамок. С использованием обнаруженных углов определяются потенциальные ограничивающие рамки. Затем происходит

пулинг центра для обнаружения центров ключевых точек у потенциальных ограничивающих рамок. Обнаруженные центры ключевых точек используются для определения окончательных ограничивающих рамок.

2.3 Оценки Качества

2.3.1 Качество классификации

Задача классификации состоит в предсказании принадлежности каждого объекта выборки к определенному классу или нескольким классам. Если классов всего два (негативный и позитивный класс), то такая классификация называется бинарной. Если классов больше двух, то это мультиклассовая классификация. Если одному изображению может соответствовать несколько, включая ноль, меток, то такая классификация называется многометковой.

Класс	Описание
TP (True Positive)	Верно положительный результат
FP (False Positive)	Ложноположительный результат
TN (True Negative)	Верно отрицательный результат
FN (False Negative)	Ложноотрицательный результат

Таблица 1: Определения классов TP, FP, TN, FN

Для оценки качества бинарной классификации используются следующие метрики.

Точность (Precision):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Полнота (Recall):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F_1 -мера (F_1 -score):

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Точность (Accuracy):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

В данной работе, когда измеряется качество классификации плазматических клеток

на `test`, для всех изображений (тайлов) на `test` считаются общие значения TP, FP, TN, FN и далее итоговые значения Precision, Recall, F_1 -score вычисляются согласно формулам, описанным выше.

2.3.2 Качество детекции и сегментации

Задача детекции заключается в обнаружении и классификации объектов на изображении. При детекции кейпойнтов определяется только местоположение объектов с помощью координат x и y их центров без указания размеров. В случае детекции объектов, используются ограничительные рамки (bounding boxes), которые задают координаты и размеры объектов. Задача instance сегментации похожа на задачу детекции, но вместо рамок используются сегментационные маски, точно отображающие форму каждого объекта.

Для измерения степени совпадения между кейпойнтами используется метрика Keypoint Similarity[5]. Для оценки совпадения ограничительных рамок или сегментационных масок одного объекта применяется IoU (Intersection over Union), которая вычисляется как отношение площади пересечения к площади объединения рамок или масок:

$$IoU(A, B) = \frac{A \cap B}{A \cup B}$$

Оценка качества детекции и instance сегментации выполняется с использованием метрики mAP (mean Average Precision). mAP является средним значением AP (Average Precision), которая представляет собой площадь под кривой Точности-Полноты (Precision-Recall). Для каждого класса и каждого заданного порога точности локализации вычисляется AP. Значения Точности (Precision) и Полноты (Recall) всегда лежат в интервале [0,1]. Следовательно, и AP лежит в том же интервале. Перед тем как определять AP, часто производится сглаживание. Графически, для каждого значения Полноты (Recall) замещается значение Точности (Precision) его максимальным значением до момента поворота тренда. Благодаря этому кривая начинает убывать монотонно вместо зигзагообразного тренда. Рассчитанная AP (Average Precision) будет менее подвержена небольшим вариациям.

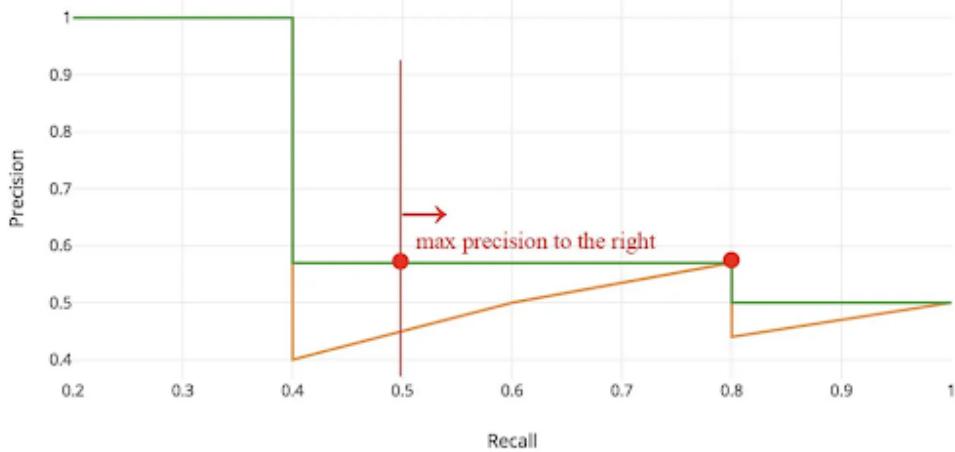


Рис. 3: График кривой Точности-Полноты (Precision-Recall) со сглаживанием

Средняя точность AP (Average Precision) вычисляется как интеграл по кривой Точности-Полноты (Precision-Recall):

$$AP = \int_0^1 p_{\text{interp}}(r)dr,$$

где $p_{\text{interp}}(r) = \max_{\hat{r} > r} p(\hat{r})$

Для оценки качества детекции и instance сегментации, используется метрика mAP[0.5,0.95], где AP вычисляется для порогов IoU от 0.5 до 0.95 с шагом 0.05 для каждого класса, после чего происходит усреднение по всем классам. Средняя Точность по всем классам mAP (mean Average Precision) вычисляется как среднее значение AP для каждого класса:

$$mAP = \frac{1}{n} \sum_{k=1}^n AP_k$$

где AP_k - это AP (Average Precision) класса k , а n - количество классов.

2.3.3 Качество согласованности

Качество согласованности врачей-экспертов в задаче с разметкой гистологических изображений может быть оценено при помощи коэффициента согласия Коэна (Cohen's kappa [6]). Коэффициент согласия Коэна измеряет степень согласованности между дву-

мя врачами-экспертами при разметке изображений и детекции нескольких классов клеток: клетки стромы, клетки эпителия, плазматические клетки.

Формула для вычисления коэффициента согласия Коэна следующая:

$$\text{Карра} = \frac{p_o - p_e}{1 - p_e}$$

где: p_o - наблюдаемая согласованность (процент согласованных меток). Наблюдаемая согласованность p_o определяется как отношение суммы согласованных предсказаний n_{kk} к общему количеству наблюдений N .

$$p_o = \frac{1}{N} \sum_k n_{kk}$$

Здесь k обозначает категории, N — общее количество наблюдений, а n_{kk} представляет количество раз, когда оба оценщика согласовались и предсказали одну и ту же категорию k . Наблюдаемая согласованность p_o представляет собой фактическую согласованность между оценщиками. Ожидаемая согласованность p_e представляет собой процент согласованных меток, которые можно было бы получить случайным образом при классификации объектов на несколько классов p_e - ожидаемая согласованность (процент согласованных меток, которые могли бы быть достигнуты случайным образом).

$$p_e = \frac{1}{2} \sum_k \frac{n_{k1} \cdot n_{k2}}{N^2}$$

Здесь k обозначает категории, N — общее количество наблюдений, а n_{k1} и n_{k2} представляют количество раз, когда оценщик i предсказал категорию k . Значение коэффициента согласия Коэна может варьироваться от -1 до 1. Высокое значение, близкое к 1, указывает на хорошую согласованность между экспертами, в то время как низкое значение, близкое к -1, указывает на плохую согласованность. Коэффициент согласия Коэна является полезным инструментом для измерения согласованности в случае, когда эксперты классифицируют объекты на несколько классов.

3 Постановка задачи

Необходимо подсчитать положительно окрашенные плазматические клетки на гистологических изображениях, окрашенных маркером CD138 (для автоматического определения хронического эндометрита).

Для этого необходимо решить следующие подзадачи:

1. Применить нейросетевую модель EndoNet, обученную на наборе данных EndoNuke[1] к изображениям с гистологических слайдов, которые предоставили врачи-эксперты.
2. Дообучить EndoNet на гистологических изображениях, которые разметили врачи-эксперты.
3. Разработать алгоритм поиска кандидатов на плазматические клетки, используя методы компьютерного зрения.
4. При помощи врачей-экспертов, собрать и разметить набор данных гистологических изображений с плазматическими клетками.
5. Провалидировать и подсчитать точность алгоритма поиска плазматических клеток.
6. Модифицировать модель EndoNet на детекцию плазматических клеток.
7. Обучить модифицированный EndoNet на детекцию плазматических клеток.
8. Провалидировать и подсчитать точность модифицированного алгоритма поиска плазматических клеток.

Требования к решению:

1. Использовать нейросетевую модель EndoNet, которая решает задачу детекции клеток на гистологических изображениях.

4 Обзор существующих решений

4.1 Применение нейросетевых моделей детекции для решения медицинских задач

Для многих задач нейросетевого профиля: классификации, детекции, сегментации, в различных областях медицины существуют примеры успешного применения нейросетевых моделей. Эти нейросетевые модели могут обнаруживать области интереса (region of interest, ROI) и различные маркеры патологий на медицинских изображениях с высокой точностью, а так же позволяют автоматизировать процессы диагностики и помогают в исследованиях, анализируя большие объемы медицинских данных.

В области исследования рака молочной железы, нейросетевые модели детекции, такие как "DeepBreast"^[7] и "YOLOv3"^[8], показывают высокую точность в обнаружении и классификации подозрительных изменений на маммографических изображениях. Это позволяет врачам рано выявлять и диагностировать рак молочной железы, увеличивая шансы на успешное лечение.

В области дерматологии, нейросетевые модели детекции, например "Derm-NN"^[9], обнаруживают и классифицируют различные типы кожных заболеваний на основе фотографий родинок и образцов кожи. Это помогает в ранней диагностике и управлении лечения рака кожи, уменьшая риски и повышая эффективность лечения.

В задачах обнаружения патологий на изображениях глаза, таких как диабетическая ретинопатия или отслойка сетчатки, модели детекции, например "RetinaNet"^[10], демонстрируют высокую точность в определении аномалий и помогают в ранней диагностике и предотвращении возможных осложнений.

Большинство современных нейросетевых моделей в медицине основаны на классических сверточных сетях, используемых в задачах компьютерного зрения таких как "ResNet"^[11], "Unet"^[12], "VGG"^[13], "Faster R-CNN"^[14], "YOLO"^[8], "EfficientNet"^[15]. В свое время данные нейросетевые модели показали лучшие результаты при решении тех или иных задач, и они легли в основу большинства современных нейросетевых моделей, используемых в задачах компьютерного зрения в медицине.

4.2 Встроенные методы QuPath

QuPath - это программное обеспечение с открытым исходным кодом для анализа всего гистологического среза, который представляет собой слайд WSI (whole slide images). QuPath широко используются в гистологии для работы с гистологическими изображениями, тк он предоставляет широкий набор инструментов для визуализации, нарезки, разметки и анализа изображений большого размера. В QuPath есть встроенный инструмент поиска областей интереса (англ. region of interest), а так же базовые инструменты обработки изображений (англ. image processing). К гистологическим изображениям был применен один из таких инструментов, который построил контуры вокруг клеток и выделил красным те клетки, которые он посчитал плазматическими.

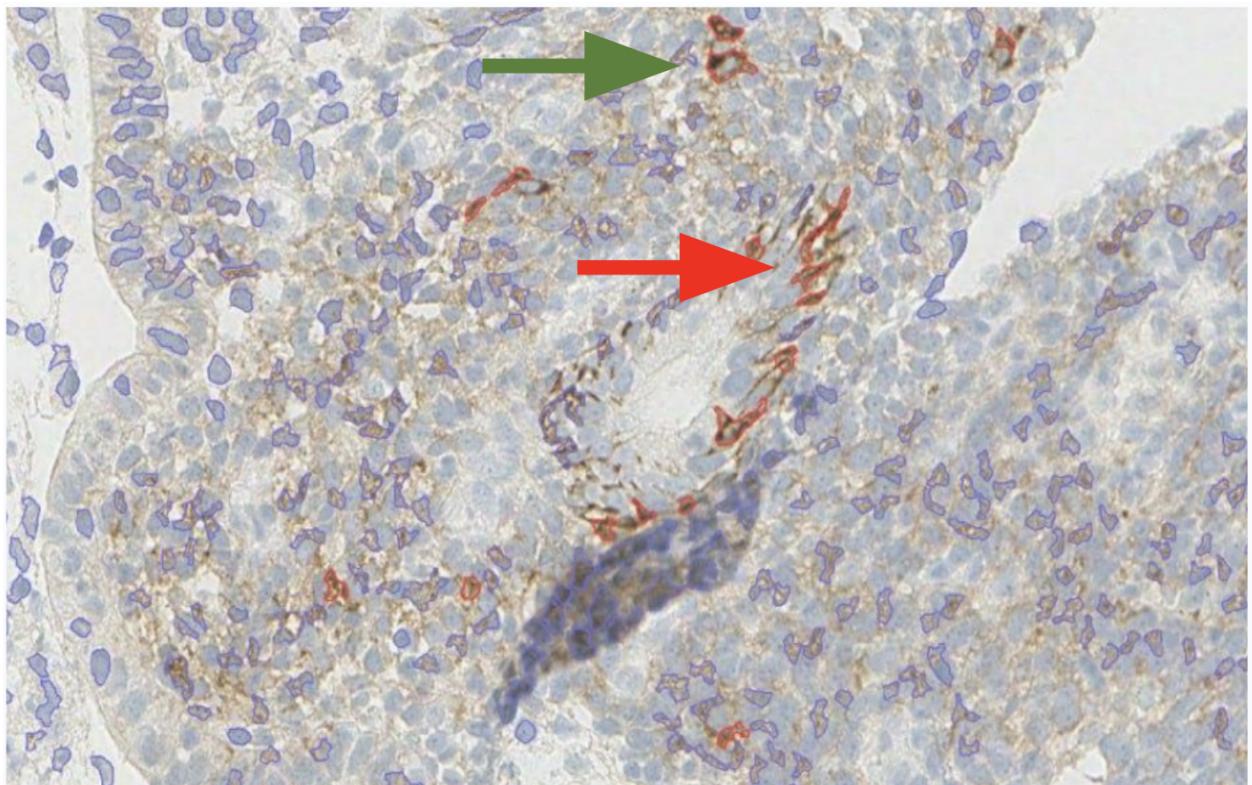


Рис. 4: Гистологическое изображение после применения встроенных методов QuPath. Зеленая стрелка указывает на плазматическую клетку. Красная стрелка указывает на клетку эпителия.

Как можно заметить, встроенный инструмент QuPath не может отличить окрашенные маркером CD138 плазматические клетки от окрашенных маркером CD138 клеток эпителия, поэтому данный инструмент не подходит для решения поставленной задачи

и нужно использовать другие методы.

4.3 Решение задачи детекции клеток на гистологических изображениях (ИСП РАН)

Для решения задачи детекции клеток стром и эпителия на открытом датасете с гистологическими изображениями EndoNuke[1] существуют решения с архитектурой CenterNet[2], которые показывают высокое качество. В ИСП РАН проводилось исследование таких моделей, в ходе которого сравнивались различные варианты моделей и модель с backbone UNet++[16] ResNet50[11] продемонстрировала самые лучшие результаты. В ходе исследования было сформировано несколько тестовых выборок, был проведён подбор оптимального набора аугментаций и других гиперпараметров, а так же были подобраны оптимальные параметры оптимизатора и функции потерь. Данная архитектура с backbone UNet++[16] ResNet50[11] и подобранными оптимальными параметрами была названа EndoNet и легла в основу разработки решения подсчета плазматических клеток на гистологических изображениях.

4.4 Окрашивание гистологических изображений двумя маркерами CD138 и MUM1

Гистологические изображения, использованные при решении поставленной задачи, окрашены одним маркером CD138. Логика решения пытается избавиться от существующей проблемы: CD138 может быть выражен не только в плазматических клетках, но также в структурах тканей, которые могут быть похожи на плазматические клетки, таких как клетки эпителия. Существуют исследования, которые решают эту проблему при помощи окрашивания дополнительным маркером MUM1. Данный маркер окрашивает плазматические клетки в оттенки красного, однако он так же окрашивает активированные В-клетки и Т-клетки, что затрудняет определение плазматических клеток. Двойное окрашивание маркерами CD138 и MUM1 решает недостатки каждого из них по отдельности.

Рассмотрим статью, опубликованную 21.12.2022[17], где решают задачу поиска плазматических клеток при помощи двойного окрашивания. Анализ используемого датасета :

- 298 гистологических слайдов : 100 слайдов train / 198 слайдов test
- 4 вида окрашивания для каждого гистологического слайда : без окрашивания / CD138 / MUM1 / CD138 & MUM1
- было размечено 1308 плазматических клеток на 2000 изображениях 640x640 для обучения Resnet18 и YOLOX-s

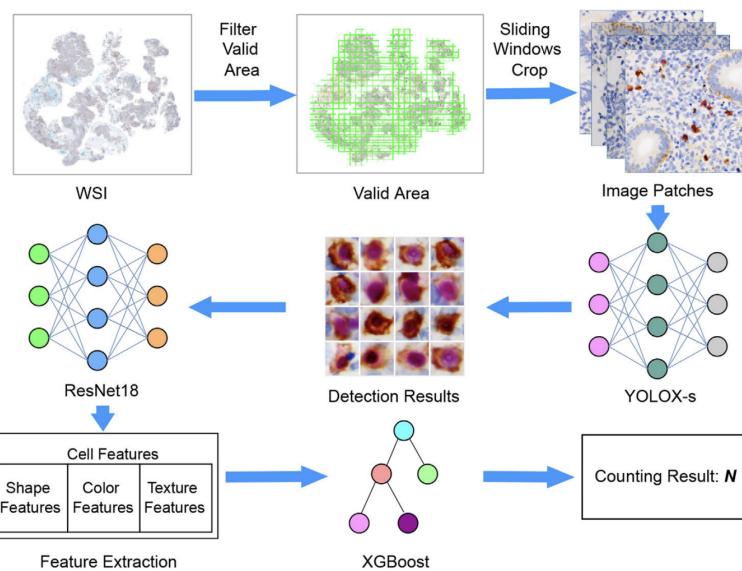


Рис. 5: Схема решения задачи детекции плазматических клеток в статье с двойными окрашиванием маркерами CD138 & MUM1

Описание схемы решения:

1. Был использован алгоритм Оцу[18] (метод автоматического выбора порога для бинаризации изображения) для рассмотрения областей интереса на гистологическом слайде.
2. Был обрезан слайд на изображения 640x640 пикселей при 40x увеличении с помощью скользящего окна внутри области интереса (для обучения моделей было аннотировано 1308 плазматических клеток на 2000 изображениях 640x640).
3. Был обучен YOLOX-s [8] на аннотированных изображениях.

4. Каждое новое изображение для анализа потом входило в обученную модель YOLOX-s [8] для обнаружения объектов.
5. В результате работы YOLOX-s [8] получались кандидаты на клетки, которые включали начальные координаты, конечные координаты и уверенность.
6. Был обучен ResNet18 [11] на размеченных изображениях.
7. Все изображения, включающие кандидатов на клетки, подавались в обученный ResNet18 [11]
8. После работы ResNet18 [11], все кандидаты на клетки были обрезаны и отнесены к клеткам или фоновым объектам.
9. Если кандидаты на клетки были классифицированы как клетки, извлекались дополнительные признаки клеток, включая форму, цвет и текстуру.
10. На основе этого обучался классификатор XGBoost [19] для принятия окончательного решения.

Точность данного подхода, постулированная в статье:

Маркер Окраски	Чувствительность	Специфичность	Точность
CD138	0.92	0.73	0.80
MUM1	1	0.62	0.76
CD138 и MUM1	1	1	1

Таблица 2: Точность решения задачи детекции плазматических клеток в статье с двойными окрашиванием маркерами CD138 & MUM1

Профессиональные патологи проанализировали 198/298 тестовых слайлов, окрашенных двумя маркерами CD138 и MUM1 под микроскопом и диагностировали 96 случаев ХЭ (хронического эндометрита) и 102 случая не-ХЭ. Разработанная диагностическая система на основе искусственного интеллекта диагностировала все 96 случаев ХЭ. Из 102 случаев не-ХЭ, 85 случаев были также интерпретированы как не-ХЭ, а 17 случаев были неправильно определены как ХЭ. Данные результаты свидетельствуют о высокой диагностической точности разработанной системы для гистологических изображений, окрашенных сразу двумя маркерами CD138 и MUM1. Была предпринята попытка создать

систему на основе искусственного интеллекта на гистологических изображениях, окрашенных маркером CD138, результаты который бы превысили результаты, полученные в рассмотренной статье. Забегая вперед можно сказать, что получилось достичь сравнимых результатов с использованием совсем иного подхода.

5 Исследование и построение решения задачи

Работа по исследованию и построению решения поставленной задачи проходила в несколько этапов, соответствующих подзадачам.

5.1 Данные

Используемые при решении задачи данные, были предоставлены врачами из НМИЦ акушерства и гинекологии им. Кулакова. Гистологические образцы представляют собой набор из 14 гистологических слайдов (whole slide images, WSI) в формате .svs размера 100 000x100 000 пикселей и 25 000x25 000 мкм. Для работы со слайдами использовался сервис QuPath. Для уменьшения количества внутренних параметров нейросетевой модели, данные слайды были разбиты на гистологические изображения (тайлы) размера 790x790 пикселей и 200x200 мкм.

5.2 Дообучение нейросетевой модели

Для решения задачи детекции ядер применяется нейросетевая модель EndoNet архитектура CenterNet[2]. Модель принимает RGB изображения в качестве входных данных, и результатом ее работы является множество координат классифицированных ключевых точек - кейпойнтов. Согласно проведенному исследованию в ИСП РАН, модель EndoNet архитектура CenterNet[2] с backbone UNet++[16] ResNet50[11] продемонстрировала самые лучшие результаты при решении задачи детекции ядер на открытом датасете с гистологическими изображениями EndoNuke[1]. Было произведено обучение данного варианта модели EndoNet на открытом наборе данных с гистологическими изображениями EndoNuke[1].

При помощи врачей, была произведена разметка изображений с гистологических слайдов в системе разметки CVAT. На изображениях врачи-эксперты размечали клетки стром и эпителия. Собранный и размеченный набор данных был обозначен как *endometrium*. Была произведена оценка качества разметки двух экспертов при помощи подсчета коэффициента согласованности Cohen's kappa для тестового тайла. Из-за того, что значение коэффициента согласованности оказалось низкое, был составлен протокол разметки. Далее был выдан основной набор тайлов экспертам для разметки. По совпадающим тайлам был подсчитан коэффициент согласованности Cohen's kappa.

Далее было произведено дообучение (fine-tuning) модели EndoNet на размеченном экспертами наборе данных `endometrium` с различными предобученными весами, а также произведено сравнение результатов. Размеченный датасет `endometrium` был разделен на `train/val` часть в соотношении 70% / 30%, что соответствует 37/17 размеченным изображениям (тайлам). Был произведен анализ результатов дообучения и сделан вывод о его необходимости.

5.3 Разработка алгоритма поиска кандидатов на плазматические клетки, используя методы компьютерного зрения

К гистологическим изображениям (тайлам) был применен медианный фильтр и фильтр Лапласса, благодаря чему были построены контуры кандидатов на плазматические клетки:

- К гистологическим изображениям (тайлам) был применен медианный фильтр. Этот фильтр помог уменьшить шум на изображении и сгладить контуры[20]. Для применения медианного фильтра, фильтрующее окно (размер окна - гиперпараметр в работе обозначался как `median_filter_size`) центрируется на каждом пикселе изображения, и значения пикселей внутри окна сортируются по возрастанию. Медианное значение, которое является средним элементом в отсортированном списке, затем выбирается в качестве нового значения для рассматриваемого пикселя. Этот процесс повторяется для каждого пикселя в изображении, что приводит к сглаженной версии исходного изображения.
- Далее гистологические изображения (тайлы) были преобразованы в оттенки серого.
- После преобразования в оттенки серого, гистологические изображения (тайлы) были подвергнуты бинаризации по пороговому значению. Бинаризация преобразует изображение в черно-белый формат, где пиксели с яркостью выше определенного порогового значения становятся белыми, а остальные становятся черными. Это помогло выделить объекты и контуры на изображении.
- Далее к гистологическим изображениям (тайлам) был применен фильтр Лапласса. Он помог выделить границы и текстуры объектов, сделав контуры более четкими.

кими и различимыми. В данной работе используется ядро размером 3x3, которое представлено следующей матрицей:

$$\text{kernel} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & \text{center_kernel_value} & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

где center_kernel_value - это гиперпараметр, который подбирался экспериментально. Для каждого пикселя изображения применяется свертка с ядром Лапласа. Результатирующее значение пикселя рассчитывается путем перемножения значений пикселей и соответствующих элементов ядра, а затем суммирования полученных произведений. Результат свертки присваивается центральному пикслю. Отрицательное значение центрального элемента ядра (`center_kernel_value`) необходимо для подчеркивания краев и областей с высокими изменениями яркости.

- Вокруг белых сегментов на бинаризованных гистологических изображениях (тайлах) были построены контуры. Это позволило выделить кандидатов на патологические клетки.

Из полученных на предыдущем шаге контуров, были сделаны выпуклые объекты, которые можно далее анализировать:

- Для каждого контура, найденного на предыдущем шаге, была построена его выпуклая оболочка.
- Благодаря этим действиям, контуры становятся выпуклыми объектами, охватывающими форму и границы каждого выделенного объекта на бинаризованных гистологических изображениях (тайлах).

Далее была произведена предобработка выпуклых оболочек, полученных на предыдущем этапе, которая заключалась в отсеве оболочек с шумовой длиной и отсеве дублирующих оболочек:

- Произведено удаление оболочек, длина которых меньше порогового значения, считая их шумом.

- На этапе создания контуров, были созданы контуры двух ориентаций: против часовой стрелки и по часовой стрелке, поэтому было произведено удаление дублирующих оболочек.

Выпуклая оболочка представляет собой наименьший выпуклый многоугольник, содержащий все точки контура. Предобработанные выпуклые оболочки были сгруппированы во вложенные выпуклые оболочки в виде графа.

- Каждая выпуклая оболочка рассматривается в качестве вершины графа. Для каждой выпуклой оболочки определяются связанные с ней выпуклые оболочки на основе их вложенности. Если одна выпуклая оболочка полностью содержится внутри другой выпуклой оболочки, то между ними устанавливается ребро.
- Граф может содержать вложенные уровни, где выпуклые оболочки могут быть вложены друг в друга на разных уровнях иерархии. Например, вершина графа может представлять собой выпуклую оболочку, охватывающую группу контуров, в то время как другая вершина графа может представлять собой более крупную выпуклую оболочку, охватывающую несколько таких групп.

Таким образом, при помощи методов компьютерного зрения и обработки изображений, были выделены объекты, которые больше всего похожи на плазматические клетки. Данные объекты были названы кандидатами на плазматические клетки.

5.4 Алгоритм поиска плазматических клеток

Был собран и размечен врачами-экспертами набор данных с изображениями, на которых есть плазматические клетки. Было подготовлено по 219 гистологических изображений (тайлов) для каждого эксперта (154 уникальных и 65 совпадающих). Итого уникальных 373 изображения. Был составлен протокол разметки. Для осуществления разметки использовался сервис CVAT. Совпадающие тайлы были нужны, чтобы по ним подсчитать коэффициент согласованности Cohen's kappa. Размеченный набор данных с гистологическими изображениями, содержащими плазматические клетки, был назван *plasmatic*.

Гистологические изображения окрашены с помощью маркера CD138, поэтому алгоритм поиска плазматических клеток был определен таким образом, что он пытается

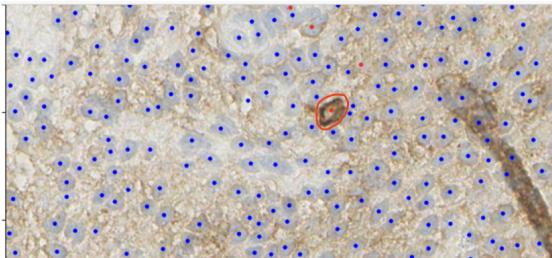
избавиться от существующей проблемы: CD138 может быть выражен не только в плазматических клетках, но также в структурах тканей, которые могут быть похожи на плазматические клетки, таких как эпителиальные клетки. Поэтому были установлены условия, при которых клетка считалась плазматической:

- Цветовое ограничение: hsv-цвет внутри выпуклой оболочки кандидата на плазматическую клетку должен находиться в пределах определенного порогового значения. Это означает, что цвет плазматической клетки должен соответствовать определенному диапазону значений в модели цветового пространства hsv.
- Ограничение по классу: Внутри выпуклой оболочки кандидата на плазматическую клетку должен находиться ровно один кейпоинт, который определила дообученная нейросеть EndoNet и этот кейпоинт имеет класс стромы.

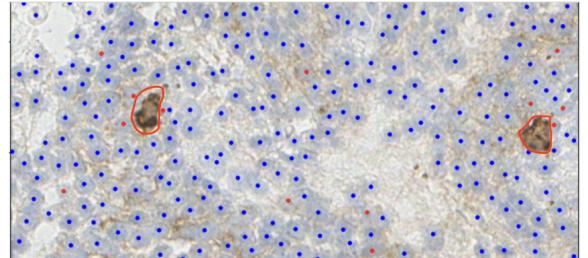
Таким образом, если выполняются оба условия - цветовое ограничение и ограничение по классу, то клетка будет считаться плазматической.

По цвету плазматических клеток были определены пороги hsv цвета, когда клетка считается плазматической. Для валидации гиперпараметров и подсчета точности алгоритма поиска плазматических клеток была взята val часть набора данных plasmatic, состоящая из 56 изображений (тайлов). Далее был провалидирован алгоритм и экспериментально получены его гиперпараметры. А также подсчитана стандартная метрика Точность (Precision), Полнота (Recall) и F_1 -мера (F_1 -score) на test, состоящем из 345 изображений (173 непустые / 172 пустые) для задачи бинарной классификации: плазматическая клетка/фон.

Был произведен анализ результатов и низкая Полнота (Recall) получилась из-за того, что дообученная нейросеть EndoNet не научилась классифицировать с хорошей точностью плазматические клетки как клетки стромы:



Пример: плазматическую клетку
классифицировало как клетку эпителия



Пример: плазматическую клетку не
классифицировало вообще

Рис. 6: Пример плазматических клеток, которые EndoNet не классифицировал как клетки стромы

В результате было принято решение модифицировать модель EndoNet, которая решает задачу детекции клеток эндометрия и их классификации на строму и эпителий на задачу детекции только плазматических клеток.

5.5 Модификация EndoNet на детекцию плазматических клеток

Для того, чтобы модифицировать модель EndoNet на детекцию только одного класса плазматических клеток, пришлось изменить некоторые ее параметры. В результате модификации, архитектура и основа (`backbone`) модели остались без изменений. Количество классов было сокращено до одного класса - класса, который соответствует плазматическим клеткам. В функции потерь (Huber[21]) был убран вес одного из классов. Изменились гиперпараметры тепловой карты (heatmap[3]) и извлечателя ключевых точек (keypoint extractor), а так же некоторые параметры аугментаций, тк плазматические клетки на краях, при прошлых параметрах аугментации, выходили за рассматриваемые рамки и некорректно обрабатывались.

Далее был определен новый класс `ModifiedModel` для модификации оригинальной модели, заменяя последний слой головы сегментации (`segmentation_head`), чтобы модель могла работать с новым количеством классов. В нашем случае - это один класс плазматических клеток. Данные действия были проделаны для использования весов у предобученной модели с другим количеством классов (EndoNet, обученный на датасете EndoNuke[1] и `endometrium`). Набор данных `plasmatic` был разделен на `train/val/test`. Для более корректного и обобщающего определения точности, в `test` было добавлено еще 160 изображений (тайлов) с гистологических слайдов, на которых нет ни одной

плазматической клетки. Такие изображения были названы "пустые". Изображения, на которых есть хотя бы одна плазматическая клетка были названы "непустые". Итого в `test` : 345 изображений (173 непустые / 172 пустые). Был обучен модифицированный на детекцию плазматических клеток `EndoNet` на `train/val` части набора данных `plasmatic` с использованием различных предобученных весов. Было произведено сравнение результатов.

Далее была взята модель `EndoNet`, решающая задачу детекции плазматических клеток и обученная на `train/val` части набора данных `plasmatic` с предобученными весами с набора данных `endometrium` и подсчитана стандартная метрика Точность (Precision), Полнота (Recall) и F_1 -мера (F_1 -score) для задачи бинарной классификации: плазматическая клетка/фон на `test`, состоящем из 345 изображений (173 непустые / 172 пустые).

В результате анализа поставленных нейросетью ключевых точек (кейпоинтов), стало понятно, что `EndoNet` ставит ключевые точки (кейпоинты) на некоторые клетки, которые не окрашены маркером CD-138 и соответственно не являются плазматическими.

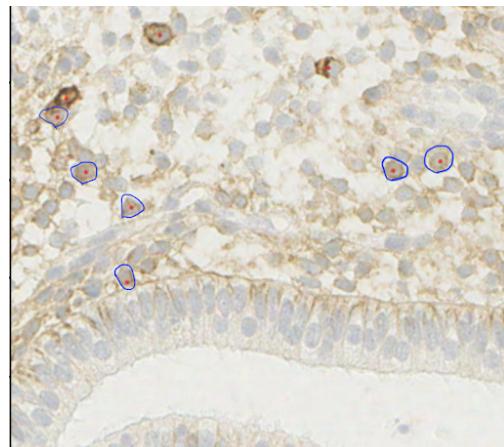


Рис. 7: Пример клеток, которые не окрашены маркером CD-138, однако `EndoNet` определил их как плазматические обведены синим

Чтобы устранить эту проблему, было решено сделать отсечение по цвету `hsv`, который имеют плазматические клетки аналогичное тому, которое было сделано в не модифицированном первом подходе.

5.6 Модифицированный алгоритм поиска плазматических клеток

После анализа работы модифицированной модели EndoNet, которая была обучена на `train/val` части набора данных `plasmatic` для задачи детекции плазматических клеток, было обнаружено, что некоторые ключевые точки (кейпоинты), которые ставит модель, находятся на клетках, которые не окрашены маркером CD-138 и, следовательно, не являются плазматическими. Поэтому были установлены условия, при которых клетка считалась плазматической:

- Цветовое ограничение: `hsv`-цвет внутри выпуклой оболочки кандидата на плазматическую клетку должен находиться в пределах определенного порогового значения. Это означает, что цвет плазматической клетки должен соответствовать определенному диапазону значений в модели цветового пространства `hsv` (аналогично цветовому ограничению в главе 5.4).
- Внутри выпуклой оболочки кандидата на плазматическую клетку должен находиться ровно один кейпоинт, который модифицированная на поиск плазматических клеток нейросеть EndoNet определила как плазматический.

Таким образом, если выполняются оба условия - цветовое ограничение и ограничение по классу, то клетка будет считаться плазматической.

Для выполнения условия ограничения по цвету рассматривались те же пороги `hsv`-цвета, который имеют плазматические клетки, что и подсчитанные ранее. Для валидации гиперпараметров модифицированного алгоритма поиска плазматических клеток была взята `val` часть набора данных `plasmatic`, состоящая из 56 изображений (тайлов). Далее был провалиден алгоритм и экспериментально получены его гиперпараметры. А также подсчитана стандартная метрика Точность (Precision), Полнота (Recall) и F_1 -мера (F_1 -score) на `test`, состоящем из 345 изображений (173 непустые / 172 пустые) для задачи бинарной классификации: плазматическая клетка/фон.

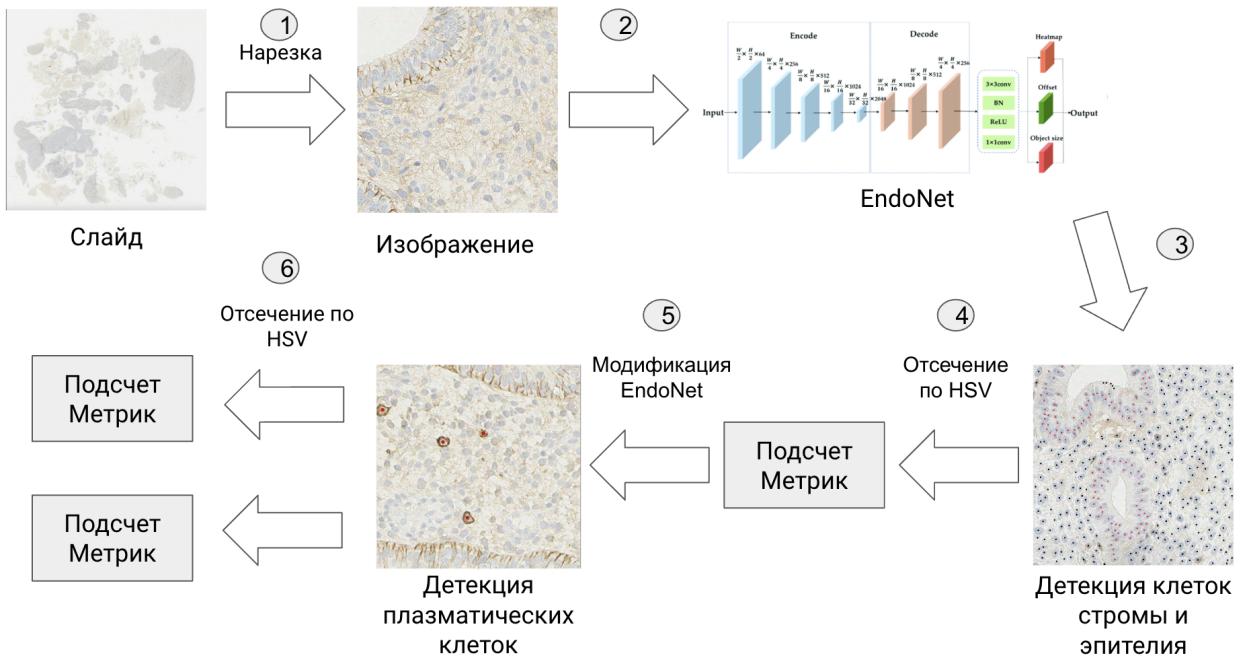


Рис. 8: Схема с этапами решения задачи

Далее было проведено сравнение и анализ результатов. Таким образом, были выполнены все подзадачи, сформулированные в постановке задачи и построено итоговое решение.

6 Описание практической части

6.1 Инструменты

Для решения задачи детекции положительно окрашенных плазматических клеток на гистологических изображениях (тайлах), окрашенных маркером CD138 был использован язык **Python 3.9**, из-за его удобства, гибкости и большого набора библиотек.

Основным инструментом при разработке решения стал **Jupyter Notebook** - интерактивная среда разработки, которая позволяет поэтапно запускать ячейки кода, просматривать результат выполнения, визуализировать, а также по мере необходимости изменять исходный код для достижения желаемого результата.

Для обучения нейронных сетей и измерения качества использовался фреймворк **PyTorch**[22]. Для обучения использовались видеокарты Nvidia Tesla T4 с 16 ГБ видеопамяти и Tesla A100 с 40 ГБ видеопамяти.

Кроме этого, в процессе разработки были использованы следующие инструменты и библиотеки:

- **cv2** - библиотека компьютерного зрения OpenCV, которая позволяет выполнять множество операций с изображениями.
- **PIL** - библиотека Python Imaging Library, которая предоставляет широкий набор инструментов для обработки изображений.
- **openslide** - библиотека для работы с изображениями WSI (Whole Slide Images), которая позволяет читать и обрабатывать гистологические изображения большого размера.
- **albumentations** - библиотека, используемая для аугментации данных.
- **QuPath** - инструмент для просмотра и анализа гистологических слайдов (whole slide images, WSI) и изображений большого размера.
- **CUDA**- параллельный вычислительный фреймворк, предоставляющий доступ к процессору графической системы NVIDIA.
- **CVAT** - бесплатная и открытая платформа для аннотирования данных, которые потом используются при обучении моделей.

6.2 Применение нейросетевой модели EndoNet, обученной на наборе данных EndoNuke[1] к изображениям с гистологических слайдов, которые предоставили врачи-эксперты

Отобразим в таблице некоторые параметры используемой нейросетевой модели EndoNet:

backbone	Параметры оптимизатора			
	оптимизатор	learning_rate	weight_decay	amsgrad
UNet++ ResNet50	Adam	0.0001	0	True

Таблица 3: Параметры нейросети EndoNet

В качестве функции потерь использовался взвешенный Huber Loss[21]. Из-за несбалансированности классов, для клеток типа строма значение веса было 1, для клеток типа эпителий 4.

Далее был обучен EndoNet с заданными параметрами на открытом наборе данных EndoNuke[1]. В качестве контроля точности при обучении использовалось значение метрики mAP (mean Average Precision).

backbone	EndoNuke	
	train	valid
UNet++ ResNet50	0.83	0.76

Таблица 4: Значения метрики mAP при обучении EndoNet на EndoNuke для задачи детекции клеток стромы и клеток эпителия на гистологических изображениях

Далее с помощью встроенных методов QuPath, рассматриваемая область на гистологическом слайде была разбита на изображения (тайлы). Было подготовлено по 30 гистологических изображений (тайлов) размера 200x200 мкм и 790x90 пикселей для каждого врача-эксперта. К ним был добавлен контекст (это все, что снаружи зеленой рамки на рис.9). Контекст нужен, чтобы корректно классифицировать клетки на краях на строму и эпителий. Для осуществления разметки использовался сервис CVAT.

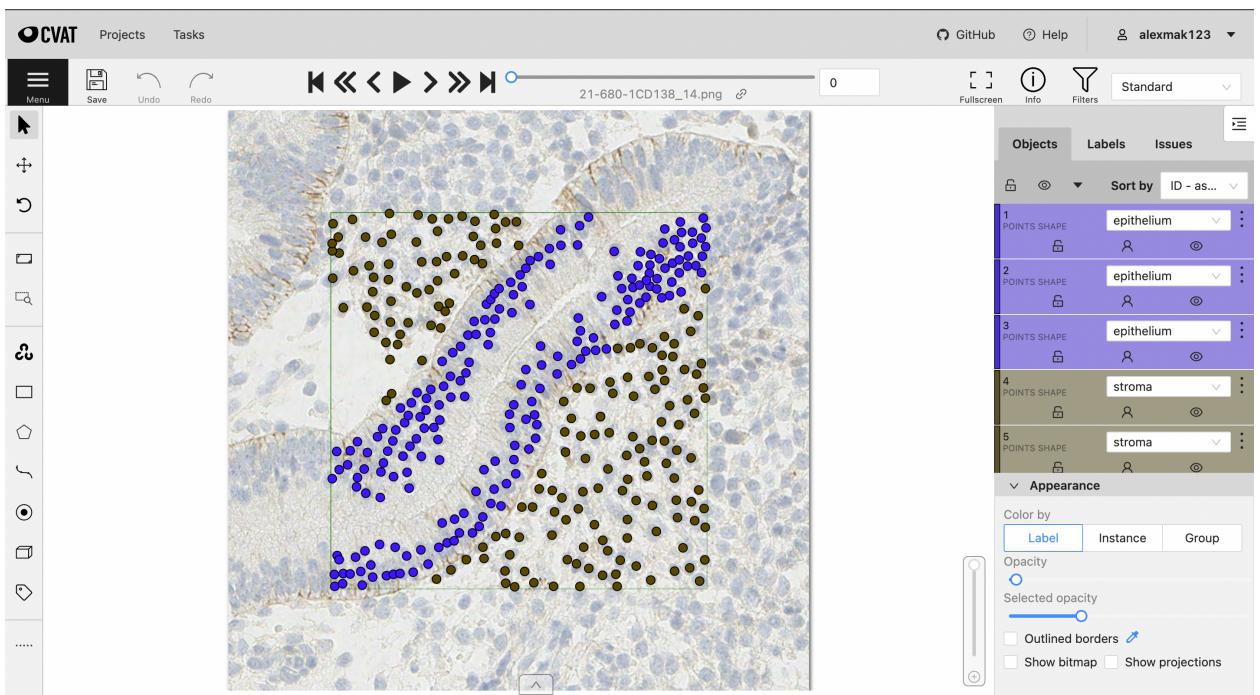


Рис. 9: Разметка врачами-экспертами клеток типа строма и эпителий на тайле в CVAT

У каждого эксперта было по 24 уникальных и 6 совпадающих изображений (тайлов). Итого уникальных 54. Было взято по 4 изображения (тайла) с каждого гистологического слайда и на каждом присутствовала и строма и эпителий. Совпадающие тайлы были нужны, чтобы по ним подсчитать коэффициент согласованности Cohen's kappa. При разметке тестового тайла, коэффициент согласованности Cohen's kappa двух экспертов оказался равен 0.68. Был составлен протокол разметки. После обсуждения протокола разметки с врачами-экспертами, им был выдан основной набор гистологических изображений (тайлов), состоящий из 54 уникальных изображений. При подсчете коэффициента согласованности Cohen's kappa по совпадающим изображениям на основном наборе тайлов, он оказался равен 0.81. И это можно считать отличным результатом согласия в разметке. Данный размеченный набор изображений был назван **endometrium**.

6.3 Дообучение EndoNet на гистологических изображениях, которые разметили врачи-эксперты (набор данных **endometrium**)

Далее было произведено дообучение (fine-tuning) модели EndoNet на размеченном экспертами наборе данных **endometrium** с различными предобученными весами. В качестве

контроля точности при обучении использовалось значение метрики mAP (mean Average Precision). Таблица с результатами при дообучении:

предобученные веса	первая эпоха		лучшая эпоха	
	train	valid	train	valid
с набор данных EndoNuke	0.46	0.49	0.84	0.72
imagenet	0.05	0.09	0.75	0.57

Таблица 5: Значения метрики mAP при дообучении EndoNet с различными предобученными весами на размеченном наборе данных `endometrium` для задачи детекции клеток стромы и клеток эпителия.

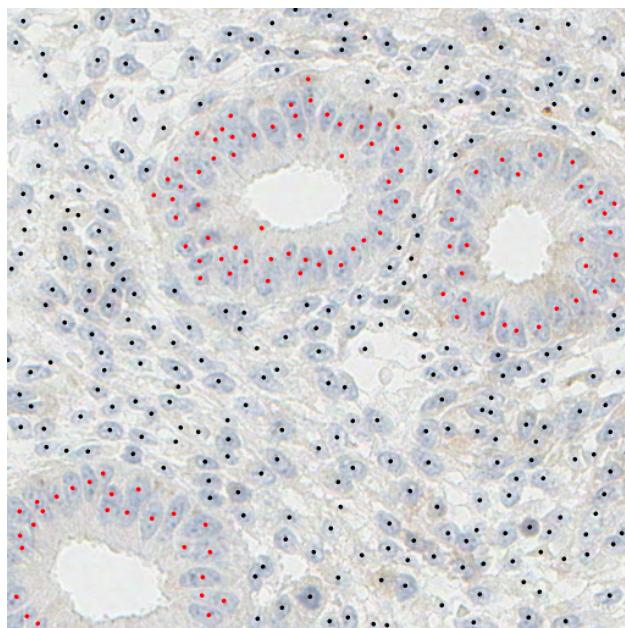


Рис. 10: Пример работы дообученной на наборе данных `endometrium` нейросети EndoNet, примененной к гистологическому тайлу. Красным помечены клетки эпителия. Синим помечены клетки стромы.

Анализируя результаты, можно сделать вывод, что дообучение и разметка гистологических изображений, при помощи врачей-экспертов проводились не зря. Если посмотреть на первую эпоху с использованием предобученных весов с набора данных EndoNuke[1], можно увидеть, что если бы дообучения не проводилось, точность детекции ядер на гистологических изображениях была бы $mAP = 0.49$, что является низким результатом. Благодаря дообучению, можно увидеть, что теперь нейросеть EndoNet имеет точность детекции ядер на гистологических изображениях $mAP = 0.72$. Так же можно

заметить, что если просто брать предобученные веса `imagenet`[23] и не обучать EndoNet на наборе данных `EndoNuke`[1] точность детекции ядер на гистологических изображениях получается равной $mAP = 0.57$, что является низким результатом. Таким образом, разметка набора данных `endometrium` и последующее дообучение модели на этом наборе данных в значительной степени улучшило точность детекции клеток стромы и клеток эпителия на гистологических изображениях, которые предоставили врачи.

6.4 Разработка алгоритма поиска кандидатов на плазматические клетки, используя методы компьютерного зрения

К гистологическим изображениям (тайлам) был применен медианный фильтр и фильтр Лапласа, благодаря чему были построены контуры кандидатов на плазматические клетки. Процесс построения контуров кандидатов на плазматические клетки можно описать следующими этапами:

- Применяется медианный фильтр с размером окна фильтра `median_filter_size` (это гиперпараметр) к текущему гистологическому изображению. Используется функция `cv2.medianBlur` из библиотеки CV2.
- После применения медианного фильтра, полученное изображение преобразуется в оттенки серого. Используется функция `cv2.cvtColor` из библиотеки CV2.
- Далее, к полученному в оттенках серого изображению, применяется бинаризация по пороговому значению. Пиксели с яркостью выше заданного порогового значения `global_threshold` (это гиперпараметр) становятся белыми, а остальные — черными. Используется функция `cv2.threshold` из библиотеки CV2.
- Фильтр Лапласа применяется к бинаризованному изображению с ядром `center_kernel_val` (это гиперпараметр). Используется функция `cv2.filter2D` из библиотеки CV2.
- На полученном бинарном изображении после применения фильтра Лапласа строятся контуры вокруг белых сегментов. Используется функция `cv2.findContours` из библиотеки CV2.

Таким образом, после выполнения данных действий были построены контуры вокруг кандидатов на плазматические клетки.

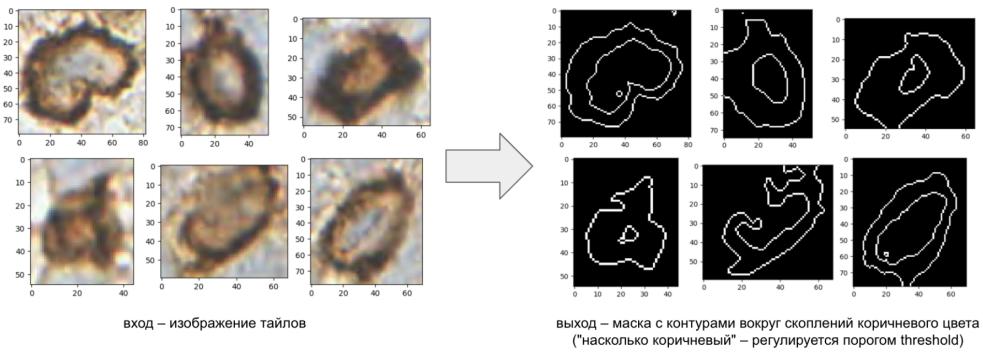


Рис. 11: Построение контуров вокруг кандидатов на плазматические клетки

Далее, из полученных на предыдущем шаге контуров, делаются выпуклые объекты, которые можно далее анализировать. Используется функция `cv2.ConvexHull` из библиотеки CV2.

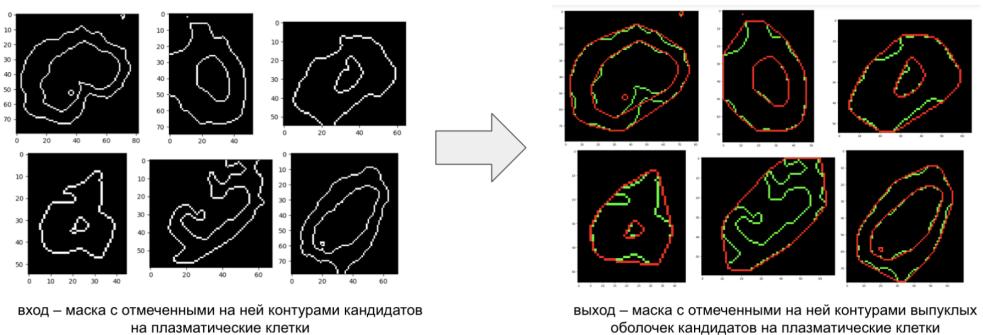


Рис. 12: Создание выпуклых оболочек из контуров, которые были построены вокруг кандидатов на плазматические клетки

Далее производится предобработка выпуклых оболочек, полученных на предыдущем этапе. Выполняются следующие действия:

- Отсев оболочек с шумовой длиной. Если длина полученной выпуклой оболочки меньше пороговой `global_noise_threshold` (это гиперпараметр), то данная оболочка далее не рассматривается.
- Отсев дублирующих оболочек. Так как были созданы контуры двух ориентаций: против часовой стрелки и по часовой стрелке, поэтому было произведено удаление дублирующих оболочек.

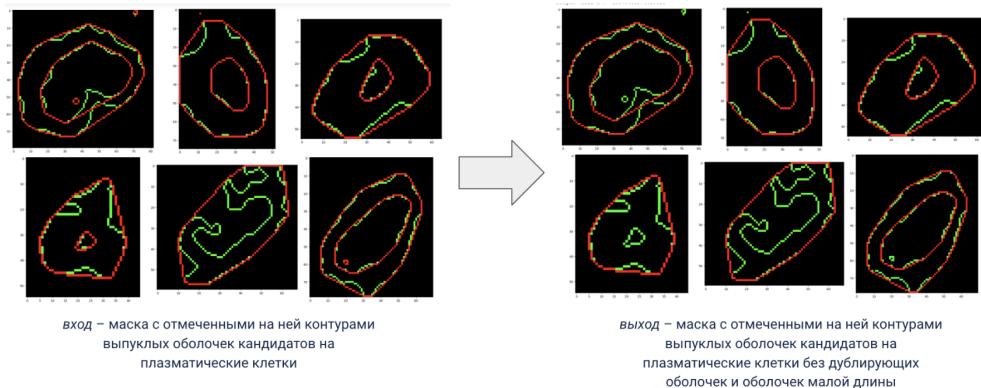


Рис. 13: Предобработка выпуклых оболочек: отсев дублирующих оболочек и удаление оболочек с шумовой длиной

Предобработанные выпуклые оболочки были сгруппированы во вложенные выпуклые оболочки в виде графа. Каждая выпуклая оболочка рассматривается в качестве вершины графа и если одна выпуклая оболочка полностью содержится внутри другой выпуклой оболочки, то между ними устанавливается ребро.

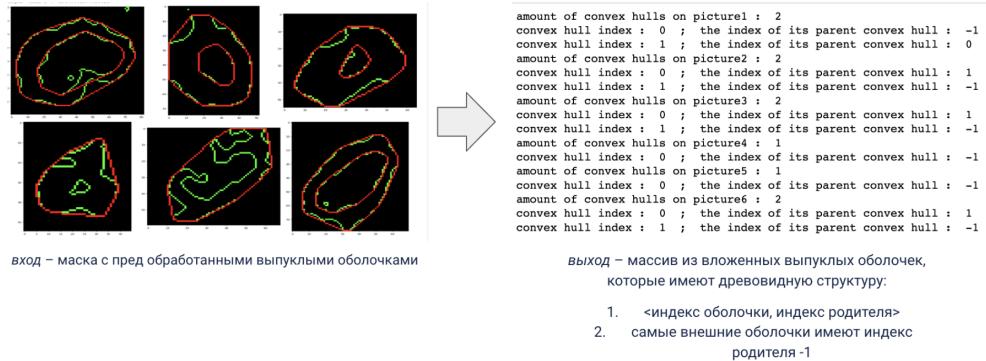


Рис. 14: Предобработанные выпуклые оболочки, представленные в виде графа

Структура в виде графа необходима для корректного подсчета границ hsv цвета плазматической клетки, чтобы учитывать цвет только тех пикселей, которые окрашены маркером CD138 (область на рис.15 отмечена красным) и не учитывать цвет тех пикселей, которые этим маркером не окрашены (область на рис.15 отмечена зеленым).

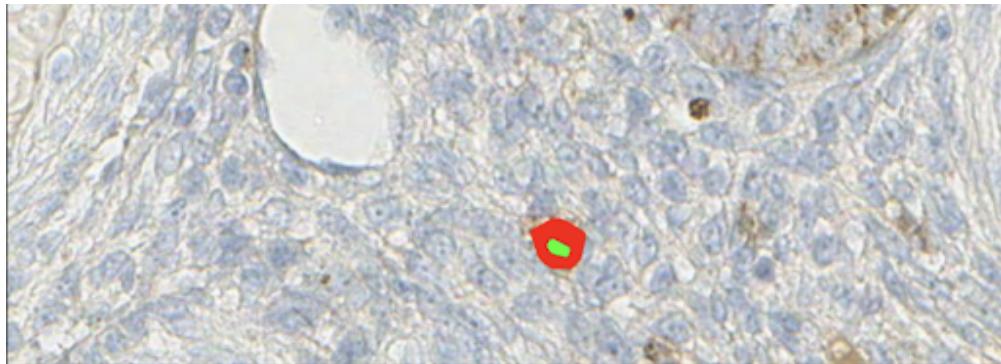


Рис. 15: Окрашенная маркером CD138 плазматическая клетка для иллюстрации того, по каким пикселям (отмечены красным) считается `hsv` цвет клетки

6.5 Сбор и разметка набора данных гистологических изображений с плазматическими клетками

Был собран и размечен врачами-экспертами набор гистологических изображений, на которых есть плазматические клетки. Было подготовлено по 219 тайлов для каждого эксперта (154 уникальных и 65 совпадающих). Итого уникальных 373 изображения. Совпадающие тайлы были нужны, чтобы по ним подсчитать коэффициент согласованности. При подсчете коэффициента согласованности Cohen's kappa, он оказался равен 0.89. И это можно считать отличным результатом согласия в разметке. Собранный и размеченный врачами-экспертами набор данных был назван `plasmatic`.

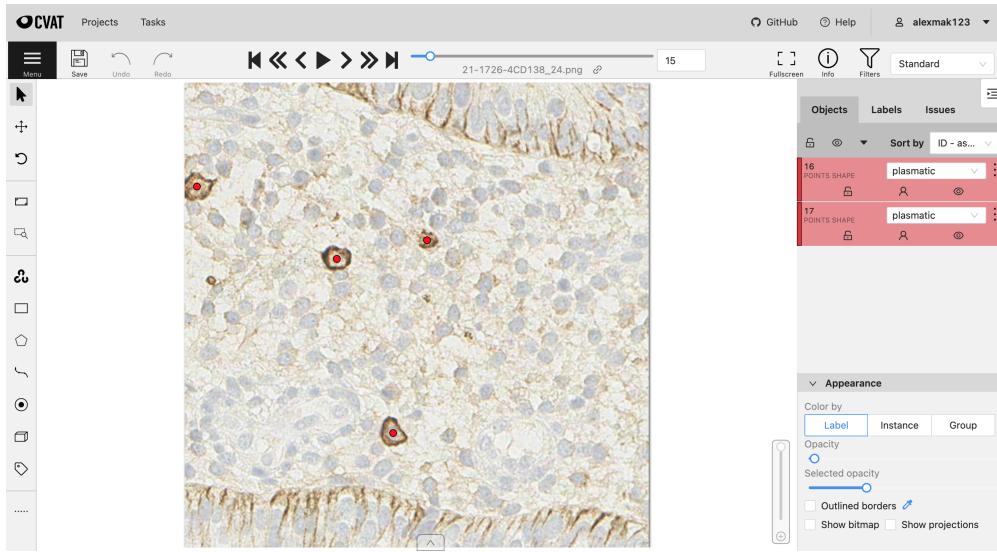


Рис. 16: Разметка плазматических клеток на гистологических изображениях в CVAT

Анализ набора данных `plasmatic`:

- 373 уникальных тайла
- 826 размеченных плазматических клетки на этих тайлах
- в среднем ~ 2.21 плазматические клетки на тайл
- 23/373 тайла в наборе данных, на которых эксперты не нашли плазматические клетки

6.6 Валидация и подсчет точности алгоритма поиска плазматических клеток

По цвету плазматических клеток были определены пороги `hsv` цвета, когда клетка считается плазматической: $5 \leq hue \leq 20$. Для валидации гиперпараметров алгоритма поиска плазматических клеток была взята `val` часть набора данных `plasmatic`, состоящая из 56 изображений (тайлов). На 52/56 изображениях была размечена хотя бы одна плазматическая клетка, на 4/56 изображениях не было размечено ни одной плазматической клетки. Экспериментально были получены гиперпараметры алгоритма:

Гиперпараметр	Значение
Значение Медианного Фильтра	17
Порог Бинаризации Изображения	147
Значение Ядра Для Фильтра Лапласа	-20
Порог Длины Оболочки, Которую Считаем Шумовой	23

Таблица 6: Значения гиперпараметров алгоритма поиска плазматических клеток

А также была произведена оценка точности работы алгоритма поиска плазматических клеток на 345 тайлах (173 непустые / 172 пустые) в стандартных метриках Точность (Precision), Полнота (Recall) и F_1 -мера (F_1 -score) для задачи бинарной классификации плазматическая клетка/фон:

Метрика	Значение
Precision	0.70
Recall	0.43
F_1 -score	0.53

Таблица 7: Точность работы алгоритма поиска плазматических клеток на `test`, состоящем из 345 изображений

Был произведен анализ результатов и низкая полнота получилась из-за того, что дообученная нейросеть `EndoNet` не научилась классифицировать с хорошей точностью плазматические клетки как клетки стромы, поэтому было принято решение модифицировать `EndoNet` на детекцию только плазматических клеток.

6.7 Модификация модели EndoNet на детекцию плазматических клеток

Для того, чтобы модифицировать модель `EndoNet` на детекцию только одного класса плазматических клеток, пришлось изменить некоторые ее параметры. Изменились гиперпараметры тепловой карты (heatmap) и извлечателя ключевых точек (keypoint extractor): Минимальное Значение Пика (min_peak_value), Масштабирование Пулинга (pooling_scale) и Диапазон Подавления (supression_range). Параметры аугментации были изменены следующим образом: вероятность горизонтального отражения (p_flip_hor) и вероятность вертикального отражения (p_flip_vert) были установлены в 0.0. Веро-

ятность поворота (`p_rotate`), масштабирования (`p_scale`) и сдвига (`p_shift`) были установлены в 0.0. Угол поворота (`rotate_angle`) был установлен в 0. Это потребовалось сделать, тк плазматические клетки на краях, при прошлых параметрах аугментации, выходили за рассматриваемые рамки и некорректно обрабатывались.

Параметр	До модификации	После модификации
Архитектура	CenterNet	CenterNet
Backbone	Unet++ resnet50	Unet++ resnet50
Количество Классов	2	1
Функция Потерь	type: Huber class weights: [1.0, 4.0]	type: Huber class weights: [1.0]
Минимальное Значение Пика	0.181	0.144
Масштабирование Пулинга	13	23
Диапазон Подавления	9.222	14.197
Размер Входного Изображения (pxl)	1024x1024	1024x1024
Параметры Аугментации	mosiac_bbox_size: 0.1 noise_var: 0.1 p_flip_hor: 0.8 p_flip_vert: 0.8 p_hsv: 0.5 p_mixup: 0 p_mosaic: 0 p_noise: 0.7 p_perspective: 0.2 p_rotate: 0.3 p_scale: 0.3 p_shift: 0.3 perspective_factor: 0.01 rotate_angle: 1 scale_factor: 0.05 shift_factor: 0.02	mosiac_bbox_size: 0.1 noise_var: 0.1 p_flip_hor: 0.0 p_flip_vert: 0.0 p_hsv: 0.5 p_mixup: 0 p_mosaic: 0 p_noise: 0.7 p_perspective: 0.2 p_rotate: 0.0 p_scale: 0.0 p_shift: 0.0 perspective_factor: 0.00 rotate_angle: 0 scale_factor: 0.05 shift_factor: 0.02

Таблица 8: Разница в параметрах модели `EndoNet` до и после модификации

Далее был определен класс `ModifiedModel`. Он наследуется от `nn.Module` и принимает входные параметры: оригинальную модель (`original_model`), которая в нашем случае `EndoNet` с новыми параметрами и количество выходных классов (`num_classes`).

```

import torch
import torch.nn as nn

class ModifiedModel(nn.Module):
    def __init__(self, original_model, num_classes):
        super(ModifiedModel, self).__init__()
        self.original_model = original_model
        self.original_model.segmentation_head[0] = nn.Conv2d(
            16, num_classes, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1)
)

    def forward(self, x):
        original_output = self.original_model(x)
        return original_output

```

Метод `__init__` класса `ModifiedModel` выполняет инициализацию объекта. В нем происходит модификация оригинальной модели: изменяется последний слой головы сегментации (`segmentation_head`). Голова сегментации (`segmentation_head`) в архитектуре нейронной сети отвечает за генерацию сегментационной карты, которая представляет собой маску, указывающую на принадлежность каждого пикселя к определенному классу или области на изображении. Последний слой головы сегментации заменяется на новый, аналогичный слой, но с другим количеством выходных классов, значение которого мы передаем при инициализации объекта `ModifiedModel`. В соответствии с проведенными изменениями, так же переопределен метод `forward`.

6.8 Обучение модифицированного EndoNet на детекцию плазматических клеток

Полученный набор данных `plasmatic` будет разделен для обучения на детекцию плазматических клеток на `train/val/test` в соотношениях:

- 35%/15%/50%
- в `train` 132 изображений (тайлов), 7 пустых

- в `val` 56 изображений (тайлов), 4 пустых
- в `test` 185 изображений (тайлов), 12 пустых

Далее модифицированный на детекцию плазматических клеток `EndoNet` будет обучен на `train/val` части набора данных `plasmatic`. Будут использованы предобученные веса с набора данных `EndoNuke`[1], с набора данных `endometrium` и предобученные веса `imagenet`[23]. Для контроля точности будет использоваться метрика `mAP` (mean Average Precision). А так же будет проведено сравнение и анализ результатов.

Предобученные Веса	Лучшая Эпоха	
	Train	Valid
<code>imagenet</code>	0.73	0.54
С набора данных <code>EndoNuke</code>	0.97	0.70
С набора данных <code>endometrium</code>	0.98	0.75

Таблица 9: Значения метрики `mAP` для задачи детекции плазматических клеток при обучении `EndoNet` на `train/val` части набора данных `plasmatic`

По результатам проведенного эксперимента можно сделать вывод, что лучшее качество детекции плазматических клеток достигается, если использовать предобученные веса с набора данных `endometrium`. А также была произведена оценка точности работы обученного на детекцию плазматических клеток `EndoNet` на 345 тайлах (173 непустые / 172 пустые) в стандартных метриках Точность (Precision), Полнота (Recall) и F_1 -мера (F_1 -score) для задачи бинарной классификации плазматическая клетка/фон:

Метрика	Значение
Precision	0.73
Recall	0.89
F_1 -score	0.80

Таблица 10: Точность обученного на детекцию плазматических клеток `EndoNet` на `test`, состоящем из 345 изображений

6.9 Валидация и подсчет точности модифицированного алгоритма поиска плазматических клеток

Был определен модифицированный алгоритм поиска плазматических клеток. Для валидации гиперпараметров модифицированного алгоритма было взято 56 тайлов из `val`

части набора данных `plasmatic`. На 52/56 изображениях была размечена хотя бы одна плазматическая клетка, на 4/56 изображениях не было размечено ни одной плазматической клетки. Экспериментально были получены гиперпараметры алгоритма:

Гиперпараметр	Значение
Значение Медианного Фильтра	17
Порог Бинаризации Изображения	165
Значение Ядра Для Фильтра Лапласа	-8
Порог Длины Оболочки, Которую Считаем Шумовой	20

Таблица 11: Значения гиперпараметров модифицированного алгоритма поиска плазматических клеток

А также была произведена оценка точности работы модифицированного алгоритма поиска плазматических клеток на 345 тайлах (173 непустые / 172 пустые) в стандартных метриках Точность (Precision), Полнота (Recall) и F_1 -мера (F_1 -score) для задачи бинарной классификации плазматическая клетка/фон:

Метрика	Значение
Precision	0.84
Recall	0.76
F_1 -score	0.80

Таблица 12: Точность работы модифицированного алгоритма поиска плазматических клеток на `test`, состоящем из 345 изображений

7 Заключение

В ходе выпускной квалификационной работы было исследовано применение нейросетевой модели детекции к гистологическим изображениям с целью обнаружения плазматических клеток для выявления хронического эндометрита. Был разработан двухэтапный алгоритм, подсчитывающий положительно окрашенные плазматические клетки на гистологических изображениях, окрашенных маркером CD138. На первом этапе была использована нейросетевая модель для детекции стромальных и эпителиальных клеток на гистологических изображениях. Модель была дообучена с использованием дополнительного набора данных, размеченного двумя врачами-экспертами. Благодаря дообучению, точность детекции стромальных и эпителиальных клеток увеличилась с $mAP = 0.49$ до $mAP = 0.72$. На втором этапе с помощью разработанного алгоритма, основанного на методах компьютерного зрения и результата работы дообученной нейросети, были определены плазматические клетки. Метрики качества разработанного алгоритма поиска плазматических клеток оказались следующие $Precision = 0.70$, $Recall = 0.43$, $F_1 - score = 0.53$. Был проведен анализ работы алгоритма и сделан вывод о необходимости модификации исходной модели на детекцию только плазматических клеток. Обучение модифицированной модели проводилось с использованием различных предобученных весов. Сравнение результатов показало, что лучшая точность детекции плазматических клеток достигается с использованием предобученных весов с набора данных **endometrium**. Метрики качества обнаружения плазматических клеток при использовании модифицированной модели увеличились до значений $Precision = 0.73$, $Recall = 0.89$, $F_1 - score = 0.8$. Был определен модифицированный алгоритм поиска плазматических клеток, основанный на методах компьютерного зрения и результата работы модифицированной на детекцию плазматических клеток нейросетевой модели. Для модифицированного алгоритма поиска плазматических клеток были получены следующие метрики качества $Precision = 0.84$, $Recall = 0.76$, $F_1 - score = 0.8$. Таким образом, были выполнены все подзадачи, сформулированные в постановке и подход с модификацией нейросетевой модели на детекцию плазматических клеток показал лучшие результаты.

Список литературы

- [1] *Naumov Anton Egor Ushakov, Andrey Ivanov Konstantin Midiber Tatyana Khovanskaya Alexandra Konyukova Polina Vishnyakova Sergei Nora Liudmila Mikhaleva Timur Fatkhudinov Evgeny Karpulevich.* EndoNuke: Nuclei Detection Dataset for Estrogen and Progesterone Stained IHC Endometrium Scans / Andrey Ivanov Konstantin Midiber Tatyana Khovanskaya Alexandra Konyukova Polina Vishnyakova Sergei Nora Liudmila Mikhaleva Timur Fatkhudinov Evgeny Karpulevich Naumov Anton, Egor Ushakov // *Data.* — 2022. — This article belongs to the Section Computational Biology, Bioinformatics, and Biomedical Data Science. <https://doi.org/10.3390/data7060075>.
- [2] *Kaiwen Duan Song Bai, Lingxi Xie Honggang Qi Qingming Huang Qi Tian.* CenterNet: Keypoint Triplets for Object Detection / Lingxi Xie Honggang Qi Qingming Huang Qi Tian Kaiwen Duan, Song Bai. — 2019. <https://arxiv.org/abs/1904.08189>.
- [3] *Fabian Amherd, Elias Rodriguez.* Heatmap-based Object Detection and Tracking with a Fully Convolutional Neural Network / Elias Rodriguez Fabian Amherd. — 2021. <https://arxiv.org/pdf/2101.03541.pdf>.
- [4] Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study / Rajendran Nirthika, Siyamalan Manivannan, Amirthalingam Ramanan, Ruixuan Wang // *Neural Computing and Applications.* — 2022.
- [5] *Matteo Ruggero Ronchi, Pietro Perona.* Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation / Pietro Perona Matteo Ruggero Ronchi // *arXiv preprint arXiv:1707.05388.* — 2017. <https://arxiv.org/abs/1707.05388>.
- [6] *McHugh, Mary L.* Interrater reliability: the kappa statistic / Mary L McHugh // *Biochem Med (Zagreb).* — 2012. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/>.
- [7] *Abo-Youssef, Magdy Abd-Elghany Zeid; Khaled El-Bahnasy; S. E.* DeepBreast: Building Optimized Framework for Prognosis of Breast Cancer Classification Based on

Computational Intelligence / Magdy Abd-Elghany Zeid; Khaled El-Bahnasy; S. E. Abo-Youssef. — 2022. — 05. — Pp. 438–445.

- [8] *Joseph Redmon, Ali Farhadi*. YOLOv3: An Incremental Improvement / Ali Farhadi Joseph Redmon // arXiv preprint arXiv:1804.02767. — 2018. — Apr.
- [9] *Tanzina Afroz Rimi Nishat Sultana, Md. Ferdouse Ahmed Foysal*. Derm-NN: Skin Diseases Detection Using Convolutional Neural Network / Md. Ferdouse Ahmed Foysal Tanzina Afroz Rimi, Nishat Sultana // 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). — 2020. — May.
- [10] *Tsung-Yi Lin Priya Goyal, Ross Girshick Kaiming He Piotr Dollár*. Focal Loss for Dense Object Detection / Ross Girshick Kaiming He Piotr Dollár Tsung-Yi Lin, Priya Goyal // arXiv preprint arXiv:1708.02002. — 2018. — Submitted on 7 Aug 2017 (v1), last revised 7 Feb 2018 (v2). <https://arxiv.org/abs/1708.02002>.
- [11] *Kaiming He Xiangyu Zhang, Shaoqing Ren Jian Sun*. Deep Residual Learning for Image Recognition / Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang // arXiv preprint arXiv:1512.03385. — 2015. — Submitted on 10 Dec 2015. <https://arxiv.org/abs/1512.03385>.
- [12] *Olaf Ronneberger Philipp Fischer, Thomas Brox*. U-Net: Convolutional Networks for Biomedical Image Segmentation / Thomas Brox Olaf Ronneberger, Philipp Fischer // arXiv preprint arXiv:1505.04597. — 2015. — Conditionally accepted at MICCAI 2015. <https://arxiv.org/abs/1505.04597>.
- [13] *Karen Simonyan, Andrew Zisserman*. Very Deep Convolutional Networks for Large-Scale Image Recognition / Andrew Zisserman Karen Simonyan // arXiv preprint arXiv:1409.1556. — 2014. — Submitted on 4 Sep 2014 (v1), last revised 10 Apr 2015 (v6). <https://arxiv.org/abs/1409.1556>.
- [14] *Shaoqing Ren Kaiming He, Ross Girshick Jian Sun*. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks / Ross Girshick Jian Sun Shaoqing Ren, Kaiming He // arXiv preprint arXiv:1506.01497. — 2015.

- Submitted on 4 Jun 2015 (v1), last revised 6 Jan 2016 (v3). <https://arxiv.org/abs/1506.01497>.
- [15] *Mingxing Tan, Quoc V. Le.* EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks / Quoc V. Le Mingxing Tan // arXiv preprint arXiv:1905.11946. — 2019. — ICML 2019. <https://arxiv.org/abs/1905.11946>.
- [16] *Zongwei Zhou Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh Jianming Liang.* UNet++: A Nested U-Net Architecture for Medical Image Segmentation / Nima Tajbakhsh Jianming Liang Zongwei Zhou, Md Mahfuzur Rahman Siddiquee. — 2018.
- [17] *Zhongtang Xiong Wei Zhang, Shaoyan Liu Kai Liu Jin Wang Ping Qin Yuping Liu Qingping Jiang.* The combination of CD138/MUM1 dual-staining and artificial intelligence for plasma cell counting in the diagnosis of chronic endometritis / Shaoyan Liu Kai Liu Jin Wang Ping Qin Yuping Liu Qingping Jiang Zhongtang Xiong, Wei Zhang // American Journal of Reproductive Immunology. — 2022. <https://doi.org/10.1111/aji.13671>.
- [18] *Xiaolu Yang Xuanjing Shen, Jianwu Long Haipeng Chen.* An Improved Median-based Otsu Image Thresholding Algorithm / Jianwu Long Haipeng Chen Xiaolu Yang, Xuanjing Shen // Advanced Science, Engineering and Medicine. — 2013. <https://doi.org/10.1016/j.aasri.2012.11.074>.
- [19] *Chen Tianqi, Guestrin Carlos.* XGBoost: A Scalable Tree Boosting System / Guestrin Carlos Chen Tianqi // arXiv preprint arXiv:1603.02754. — 2016. <https://arxiv.org/abs/1603.02754>.
- [20] *Anwar Shah Javed Iqbal Bangash, Abdul Waheed Khan Imran Ahmed Abdullah Khan Asfandyar Khan Arshad Khan.* Comparative analysis of median filter and its variants for removal of impulse noise from grayscale images / Abdul Waheed Khan Imran Ahmed Abdullah Khan Asfandyar Khan Arshad Khan Anwar Shah, Javed Iqbal Bangash // Journal of King Saud University-Computer and Information Sciences. — 2020. <https://www.sciencedirect.com/science/article/pii/S1319157820300749>.

- [21] *Kaan Gokcesu, Hakan Gokcesu.* Generalized Huber Loss for Robust Learning and its Efficient Minimization for a Robust Statistics / Hakan Gokcesu Kaan Gokcesu. — 2021. — Submitted on 28 Aug 2021 (v1). <https://arxiv.org/abs/2108.12627>.
- [22] *Adam Paszke Sam Gross, Francisco Massa Adam Lerer James Bradbury Gregory Chanan Trevor Killeen Zeming Lin Natalia Gimelshein Luca Antiga Alban Desmaison Andreas Köpf Edward Yang Zach DeVito Martin Raison Alykhan Tejani Sasank Chilamkurthy Benoit Steiner Lu Fang Junjie Bai Soumith Chintala.* PyTorch: An Imperative Style, High-Performance Deep Learning Library / Francisco Massa Adam Lerer James Bradbury Gregory Chanan Trevor Killeen Zeming Lin Natalia Gimelshein Luca Antiga Alban Desmaison Andreas Köpf Edward Yang Zach DeVito Martin Raison Alykhan Tejani Sasank Chilamkurthy Benoit Steiner Lu Fang Junjie Bai Soumith Chintala Adam Paszke, Sam Gross. — 2019. <https://arxiv.org/abs/1912.01703>.
- [23] *Olga Russakovsky Jia Deng, Hao Su Jonathan Krause Sanjeev Satheesh Sean Ma Zhiheng Huang Andrej Karpathy Aditya Khosla Michael Bernstein Alexander C. Berg Li Fei-Fei.* ImageNet Large Scale Visual Recognition Challenge / Hao Su Jonathan Krause Sanjeev Satheesh Sean Ma Zhiheng Huang Andrej Karpathy Aditya Khosla Michael Bernstein Alexander C. Berg Li Fei-Fei Olga Russakovsky, Jia Deng. — 2014. <https://arxiv.org/abs/1409.0575>.