ПРОЕКТ МЭРА МОСКВЫ

ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ

АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

МОСКОВСКАЯ МЕДИЦИНА

Сквозь турникеты в ML

Задача 8

ИИ-сервис для выявления компьютерных томографий органов грудной клетки с «нормой»

Команда «Сквозь турникеты в ML»

Александр Павлов

- DS ML
- awesome_sp68
- 89027293656

Капитан

Владислав Баланда

- DS ML
- Vlad2ru
- 89145443295

Мзиссана Куртанидзе

- Frontend
- mzissana
- 89296622579

Валерия Никитина

- Backend
- lerin_nikita
- 89036148202

Команда «Сквозь турникеты в ML»

ПРОЕКТ МЭРА МОСКВЫ | ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ | АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

О команде

- Москва, Хабаровск, Тамбов
- 4 участника
- Александр Павлов




Наименование задачи:

ИИ-сервис для выявления компьютерных томографий органов грудной клетки с «нормой»

Описание решения:

ИИ-сервис, который автоматически отсеивает «норму» на КТ ОГК и подсвечивает подозрительные исследования.

Гибрид: SlimCLR + автоэнкодер по «норме» + ResNet-классификатор с TTA.

Быстрее диагностика, меньше пропусков, готово к интеграции.

Как вы планируете дальше использовать или развивать ваше решение:

Улучшение качества: расширяем датасеты, валидируем разметку, настраиваем пороги под сценарии, используем более сложные модели

Функциональный рост: добавляем типы патологий, локализацию и сегментацию, ускоряем сервис

Улучшение пользовательского пути: расширение UX/UI

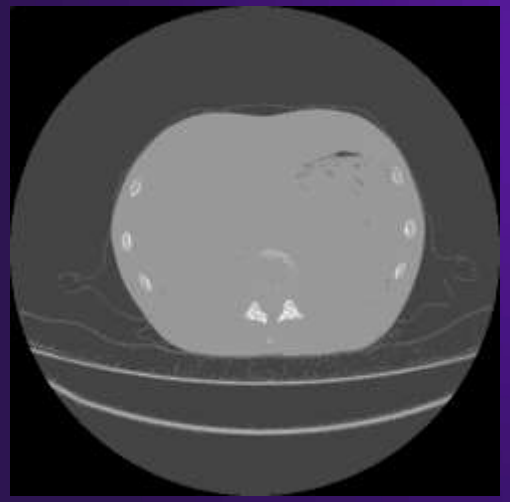
Пробные пилоты у специалистов: собираем клиническую обратную связь

Задача и цель

Задача: разработка автоматизированного решения на основе ИИ, которое автоматически классифицирует КТ ОГК снимки на «норму» и «патологию»

Целевой результат: повышение скорости анализа КТ ОГК снимков, снижение нагрузки на медицинский персонал и минимизация рисков невыявленных патологий

Примеры исследований:



Актуальность

Проблема

- Врачи перегружены рутинной сортировкой исследований, где значительная часть пациентов не имеет патологий
- Работая в контексте одной жалобы легко упустить сопутствующую или скрытую патологию

Альтернативные решения

- Существующие подходы строятся на бинарной классификации для каждой отдельной патологии
- Интеграция множества разрозненных моделей требует значительных ресурсов и времени
- Отсутствует единый сервис, способный решать задачу комплексно

Предложенное решение

Используем гибридный ансамбль моделей, который сочетает четыре подхода:

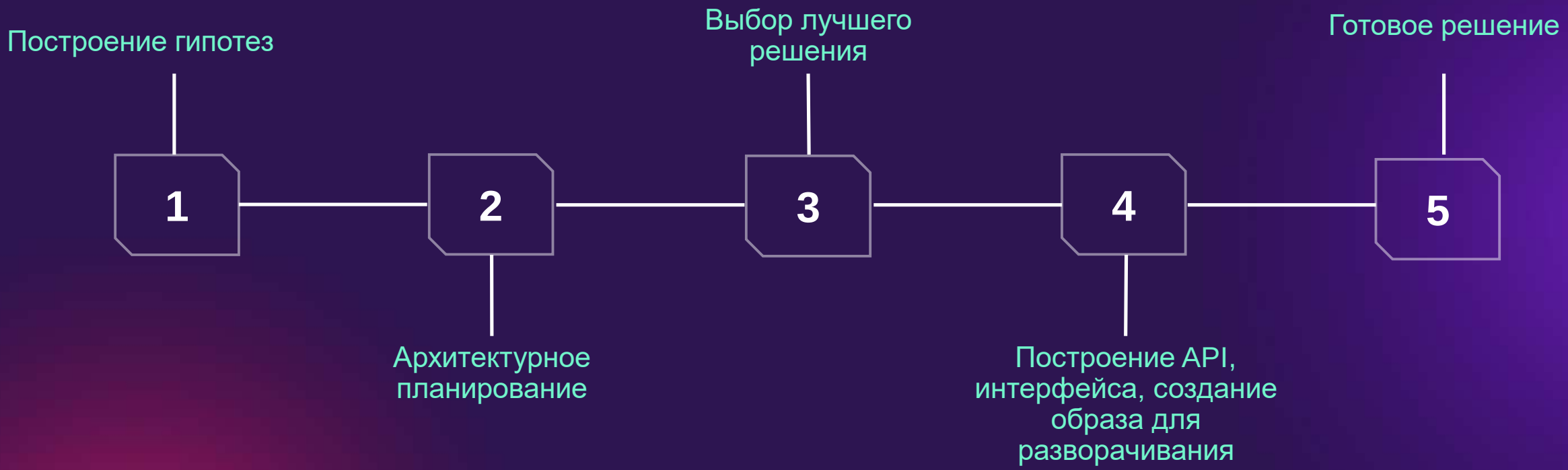
- self-supervised обучение для извлечения признаков

- автоэнкодер для поиска аномалий

- классификатор для финального решения

- классификатор для патологий

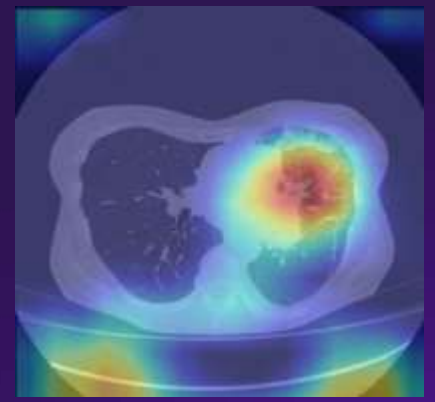
План работы


Построение гипотез
1
Архитектурное планирование
2
Выбор лучшего решения
3
Построение API, интерфейса, создание образа для разворачивания
4
Готовое решение
5

Проверка гипотез и ход работы

1. Были собраны дополнительные данные – КТ снимки для расширения представленного датасета.
2. Написана программа для разделения данных на обучающую (train) и тестовую (test) выборки с учетом пациентов. Изображения одного пациента должны попадать либо в тренировочную. либо в тестовую выборку - это предотвращает утечку данных.
3. Для построения базового решения использовались классификаторы изображений на основе моделей семейства ResNet (ResNet18 и ResNet34).
4. Было проверена гипотеза о 3D признаках – был обучена модель «SlowFast_r50» (модель для классификации видео).
5. Учитывая особенности задачи:
   - разная модальность снимков (полученных с помощью различных аппаратов КТ);
   - большое количество патологий (более 40);
   - необходимость поиска новых патологий, отсутствующих в размеченных данных.

   Мы перешли к решению, основанному на «Self-supervised pretraining» (SSP).



Преимущества данного метода: позволяет достигать высокого качества, используя в разы меньше размеченных данных по сравнению с чисто supervised-подходами, что особенно важно в областях с ограниченной разметкой.

Подготовка датасета

ПРОЕКТ МЭРА МОСКВЫ | ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ | АГЕНТСТВО ИННОВАЦИЙ МОСКВЫ

1. MosMedData НДКТ с признаками рака легкого тип I

- Ссылка: https://mosmed.ai/datasets/datasets/mm/
- Количество классов: 2
- Названия классов: [Без патологии; С патологией]
- Количество по классам: [50; 50]

2. MosMedData КТ с признаками коронавирусной инфекции (COVID-19) тип I

- Ссылка: https://mosmed.ai/datasets/datasets/mosmeddata-kt-s-priznakami-koronavirusnoi-infektsii-covid-19-tip-i/
- Количество классов: 2
- Названия классов: [Без патологии; С патологией]
- Количество по классам: [49; 48]

3. MosMedData: результаты исследований компьютерной томографии органов грудной клетки с признаками COVID-19

- Ссылка: https://mosmed.ai/datasets/datasets/covid191110/
- Количество классов: 2
- Названия классов: [Без патологии; С патологией]
- Количество по классам: [254; 856]

4. MosMedData: КТ с признаками рака легкого тип VIII

- Ссылка: https://mosmed.ai/datasets/datasets/mosmeddata-kt-s-priznakami-raka-legkogo-tip-viii/
- Количество классов: 1
- Названия классов: [С патологией]
- Количество по классам: [536]

Подготовка датасета

5. MosMedData: результаты лучевых исследований пациентов с коронавирусной инфекцией (COVID-19)

- Ссылка: https://mosmed.ai/datasets/datasets/covid19s20/
- Количество классов: 1
- Названия классов: [С патологией]
- Количество по классам: [46]

6. CT-RATE: **Разработка универсальных базовых моделей на основе мультимодального набора данных для 3D-компьютерной томографии**

- Ссылка: https://huggingface.co/datasets/ibrahimhamamci/CT-RATE/tree/main/dataset
- Количество классов: 17
- Названия классов: ['Arterial wall calcification', 'Cardiomegaly', 'Pericardial effusion', 'Coronary artery wall calcification', 'Hiatal hernia', 'Lymphadenopathy', 'Emphysema', 'Atelectasis', 'Lung nodule', 'Lung opacity', 'Pulmonary fibrotic sequela', 'Pleural effusion', 'Mosaic attenuation pattern', 'Peribronchial thickening', 'Consolidation', 'Bronchiectasis', 'Interlobular septal thickening']
- Количество по классам: [Без патологий – 5654, С патологиями - 43745]

Для бинарной классификации были отобраны данные из датасетов:

(Исключены снимки с дефектами: костные ткани, засветы и прочее)

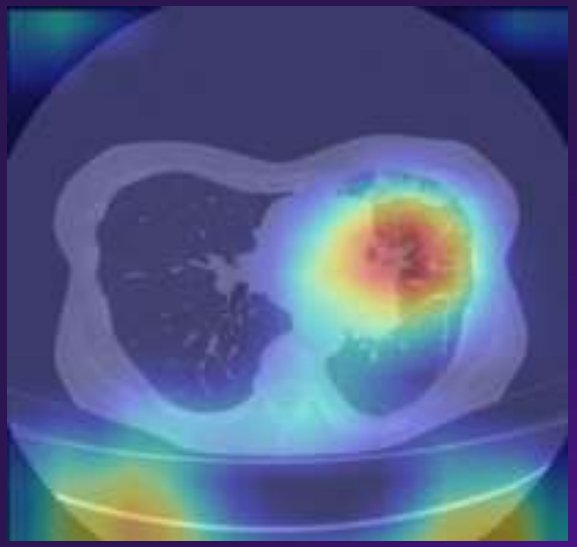
MosMedData (из всех датасетов) – 350 пациентов (Без патологий – 50%, С патологиями – 50%)

CT-Rate – 1500 пациентов (Без патологий – 50%, С патологиями – 50%)

Подготовка датасета

Для классификации патологий отобраны данные из датасета CT-Rate:
В датасет включены только исследования с одной патологией
(в датасете в основном пациенты с двумя и более патологиями)

Патология	Пациентов
Arterial wall calcification	94
Atelectasis	300
Bronchiectasis	114
Cardiomegaly	28
Consolidation	174
Coronary artery wall calcification	119
Emphysema	270
Hiatal hernia	196
Lung nodule	300
Lung opacity	300
Lymphadenopathy	300
Mosaic attenuation pattern	79
Peribronchial thickening	78
Pericardial effusion	61
Pleural effusion	46
Pulmonary fibrotic sequela	300



Принципы решения

Self-Supervised Pretraining (SimCLR)

Метод основан на концепции «self-supervised learning», использующей алгоритм SimCLR.

Особенности:

- Объединение обучающей и валидационной выборок без учета диагнозов (без меток);
- Создание пар изображений через случайные аугментации каждого снимка (случайный кроп/обрезка, отражение по горизонтали, изменение яркости и контраста);
- Контрастивная функция потерь (NTXentLoss) способствует извлечению значимых признаков структуры лёгких.

Реконструкция автоэнкодером

Автоэнкодер обучается только на изображениях класса «норма».

Особенности:

- Эффективное восстановление нормальных паттернов лёгочной ткани.
- Увеличение ошибки реконструкции при появлении патологии.
- Использование маски легких (полученной методом пороговой сегментацией) для исключения влияния некритичных областей (костей, мышц).

Принципы решения (продолжение)

Тонкая настройка бинарного классификатора

Обучение бинарного классификатора осуществляется на основе модели ResNet-18, предварительно обученной на основе принципов «self-supervised pretraining».

Важные моменты:

- Применение аугментаций для увеличения вариативности обучающего набора;
- Учет результатов «Test-Time Augmentation» (TTA) для повышения устойчивости модели к шуму и артефактам.

Принятие решения о результате

Принятие решений о результате («патология» или «норма») производится с использованием ансамбля, объединяющего сигналы двух компонентов:

- Нормализованная ошибка реконструкции (реконструкционный компонент);
- Максимальная вероятность патологии (бинарный классификатор).

Итоговая оценка зависит от ошибки реконструкции (от автоэнкодера) и вероятности патологии (от бинарного классификатора) —по принципу максимума.

Выбор патологии

Принятие решений о классе патологии производится на основе отдельной модели, обученной на датасете, содержащем 16 видов патологий (модель семейства YOLO11).

Преимущества разработанного решения

1. **Выявление «неизвестных» патологий**. Благодаря автоэкондеру, обученному только на здоровом материале, система способна обнаруживать ранее не встречавшиеся для модели типы патологии.
2. **Использование всей базы меток.** Обучение на основе всей доступной базы снимков self-supervised learning без привязки к диагнозам улучшает качество извлекаемых признаков.
3. **Устойчивость к данным.** Применение Test-Time Augmentation существенно повышает надежность результата – снижая зависимость от «шума» изображений и различий между аппаратами КТ.
4. **Баланс точности.** Определение оптимального порога гарантирует баланс между точностью выявления патологии и отсутствием избыточных тревог.

Метрики и доверительные интервалы

После обучения модели, на валидационных данных получены следующие метрики:

ROC-AUC: 0.7782
Патологии (N=176):
Средний скор: 0.7716 ± 0.1146
Диапазон: [0.5227, 1.0000]
Норма (N=175):
Средний скор: 0.6580 ± 0.0867
Диапазон: [0.5245, 0.9421]
Общий диапазон скоров: [0.5227, 1.0000]

Метрики и доверительные интервалы (продолжение)

РЕКОМЕНДУЕМЫЕ ПОРОГИ (ансамбль):

Оптимальный по F1: 0.690 → F1=0.7342

Оптимальный по Youden J: 0.760 → J=0.4538

По чувствительности ≥0.95: 0.530 → Recall=0.994, F1=0.6679, Precision=0.503

Медианный порог: 0.709

Порог для высокой чувствительности: 0.300

95% ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ (бутстреп, n=1000)

Перед bootstrap:

y_true: 351 samples, unique labels: [0 1]

y_scores: 351 samples

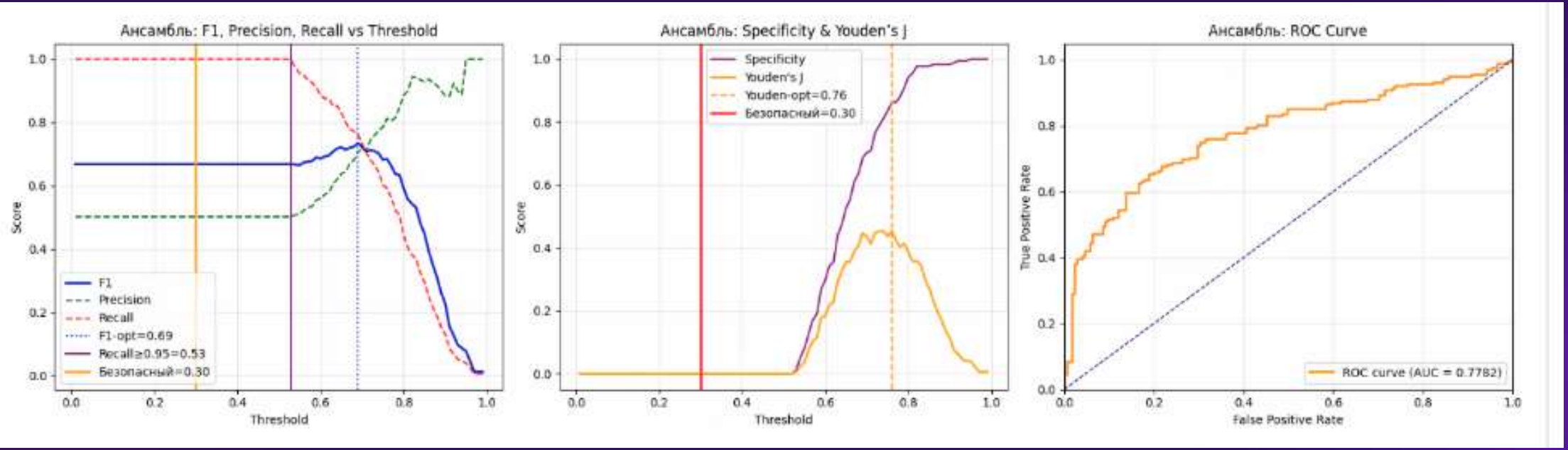
Порог для CI: 0.300

AUC: 0.7790 (95% CI: 0.7272 – 0.8231)

Чувствительность: 1.0000 (95% CI: 1.0000 – 1.0000)

Специфичность: 0.0000 (95% CI: 0.0000 – 0.0000)

Метрики и доверительные интервалы (продолжение)

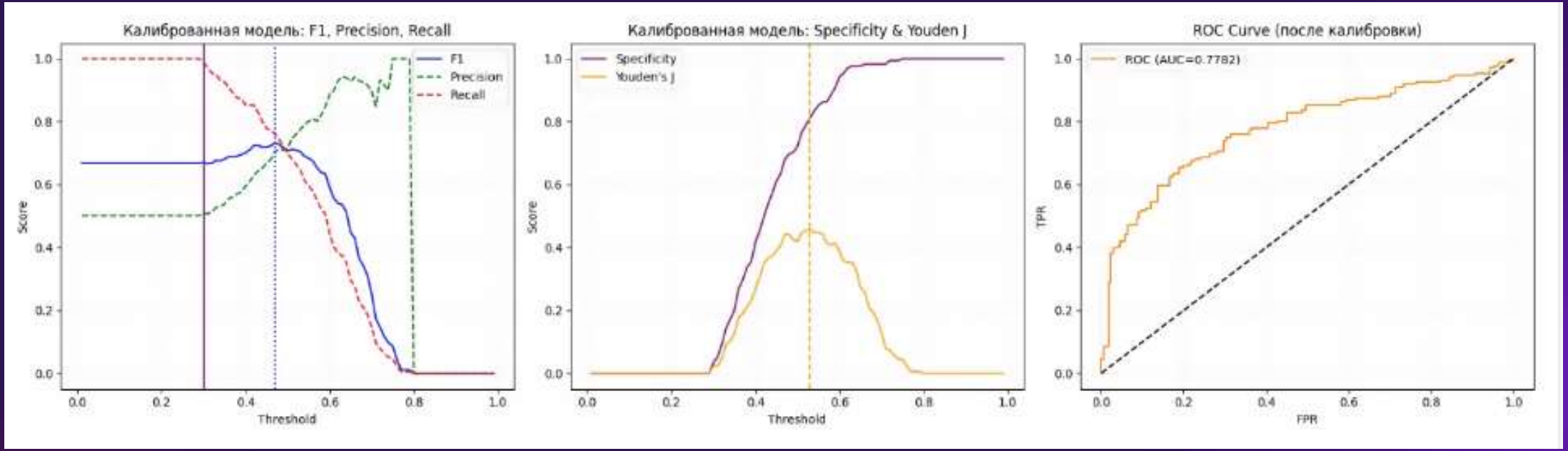


Метрики и доверительные интервалы после калибровки

АНАЛИЗ ПОСЛЕ КАЛИБРОВКИ (patient-level)
ROC-AUC (после калибровки): 0.7782
Патологии (N=176):
Средняя вероятность: 0.5646 ± 0.1265
Диапазон: [0.2914, 0.7924]
Норма (N=175):
Средняя вероятность: 0.4378 ± 0.0974
Диапазон: [0.2932, 0.7445]

РЕКОМЕНДУЕМЫЕ ПОРОГИ (после калибровки):
Оптимальный по F1: 0.470 → F1=0.7322
Оптимальный по Youden J: 0.530 → J=0.4592
По чувствительности ≥0.95: 0.300 → Recall=0.989, F1=0.6705, Precision=0.507
Медианный порог: 0.495

Метрики и доверительные интервалы после калибровки (продолжение)



Метрики классификации патологий

Для классификации патологий обучалась модель «Multiclass classification»

Патология	Пациентов	Accuracy	Recall	F1-Score
Arterial wall calcification	94	0.7024	0.7024	0.6782
Atelectasis	300	0.5076	0.5076	0.4942
Bronchiectasis	114	0.8843	0.8843	0.8251
Cardiomegaly	28	0.0000	0.0000	0.0000
Consolidation	174	0.9906	0.9906	0.9431
Coronary artery wall calcification	119	0.5644	0.5644	0.6773
Emphysema	270	0.7367	0.7367	0.6529
Hiatal hernia	196	0.8640	0.8640	0.7176
Lung nodule	300	0.1807	0.1807	0.2428
Lung opacity	300	0.3433	0.3433	0.3780
Lymphadenopathy	300	0.5778	0.5778	0.5431
Mosaic attenuation pattern	79	0.6667	0.6667	0.6042
Peribronchial thickening	78	0.9933	0.9933	0.7700
Pericardial effusion	61	1.0000	1.0000	0.8621
Pleural effusion	46	1.0000	1.0000	0.9934
Pulmonary fibrotic sequela	300	0.3217	0.3217	0.3437

Усредненные метрики: Accuracy: 0.6458, Recall: 0.6458, F1: 0.6079, AUC: 0.9099

API и интерфейс

API

Предоставляет эндпоинты для работы с моделью и базой

Patients
GET /api/patients - Список записей пациентов
GET /api/patients/{id} - Получение записи пациента
POST /api/patients - Создание записи пациента
PUT /api/patients/{id} - Редактирование записи пациента
DELETE /api/patients/{id} - Удаление записи пациента

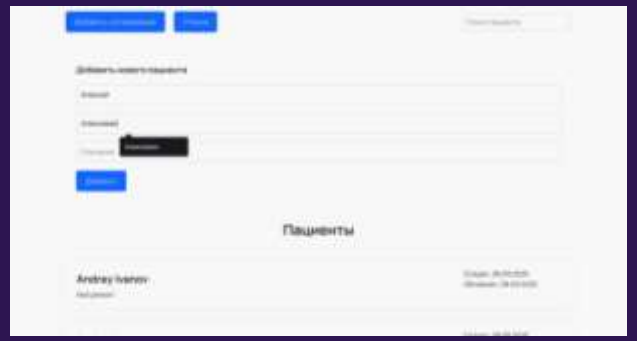
Scans
GET /api/scans — Список исследований
GET /api/scans/?patient_id={patientId} - Получение списка сканов пациента
POST /api/scans — Создание исследования (загрузка файла)
GET /api/scans/{id} — Получение информации об исследовании
GET /api/scans/{id}/file — Скачать исходный бинарник
PUT /api/scans/{id} — Редактирование исследования (description)
POST /api/scans/{id}/analyze — Запустить анализ исследования
GET /api/scans/{id}/report — Получить JSON-отчёт об исследовании
DELETE /api/scans/{id} — Удаление исследования

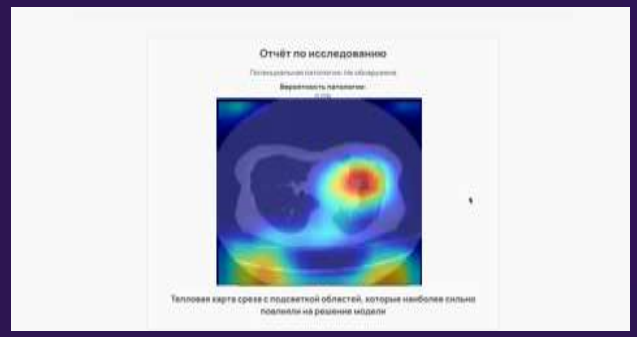
Inference
POST /inference/predict - Получение отчета по исследованию без привязки к пациенту (для массового прогона данных)

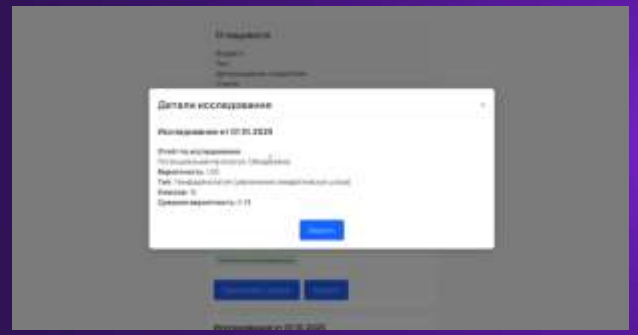
API и интерфейс

Интерфейс

- Frontend для работы врачей и исследователей предоставляет:
  - создание и ведение карточек пациентов,
  - загрузка и просмотр их КТ-исследований,
  - получение результата анализа,
  - просмотр тепловых карт (heatmaps), показывающих наиболее подозрительные области.





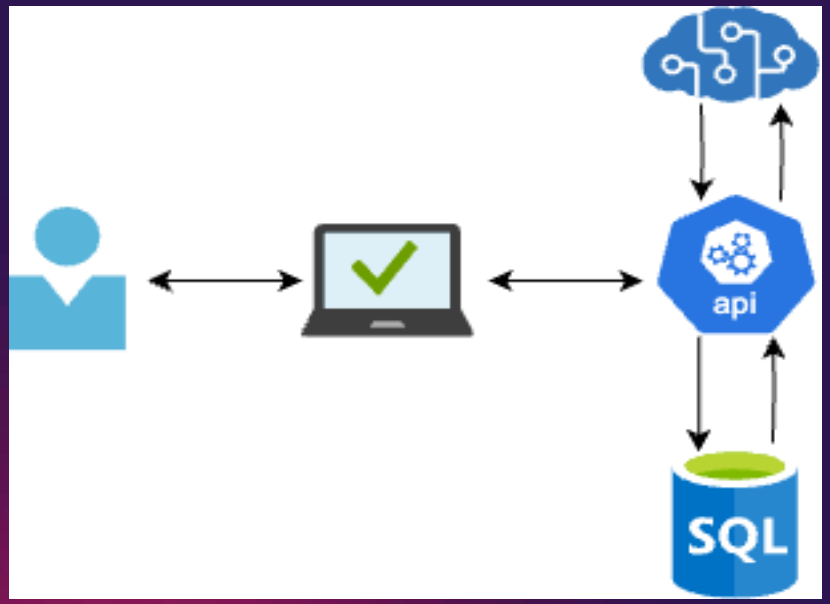


Используемые технологии и ресурсы

01 Python

02 FAST API, Postgres, Docker

03 React



04 Дополнительные источники данных

- Датасет "ibrahimhamamci/CT-RATE" https://huggingface.co/datasets/ibrahimhamamci/CT-RATE
- Датасеты ОГК: https://mosmed.ai/datasets/datasets

05 Обучение модели

В решении обучаются модели:

- SimCLR на объединенных данных (без меток класса, для извлечения общих эмбедингов);
- Автоэнкогдер – только на данных (норма);
- Бинарный классификатор – на тренировочных данных, с валидацией;
- Многоклассовый классификатор - на тренировочных данных, с валидацией.

06 Инференс модели

Технические требования:

ОЗУ 16Г, SSD, при наличии GPU модель автоматически его использует

Развертывание сервиса

Docker

Сервис использует docker-compose.yml

Сборка

docker compose build --no-cache

.env

Файл содержит параметры для поднятия контейнеров

Поднятие

docker compose up -d

Контейнеры

- база данных
- API
- интерфейс

Предложения и развитие проекта

Данные: расширение датасета + повторная валидация разметки

ML-архитектуры: переход к более сложным моделям и ансамблям

Интерфейс: расширение UX/UI и сценариев для врача

Диагностика: определение конкретных типов патологий

Локализация: картирование снимков и визуализация очагов

Производительность: оптимизация инференса

Сквозь турникеты в ML