Анализ метаданных рентгеновских снимков при COVID-19

Дмитрий Брагин

НИЯУ МИФИ

Декабрь 2025

Архитектура решения

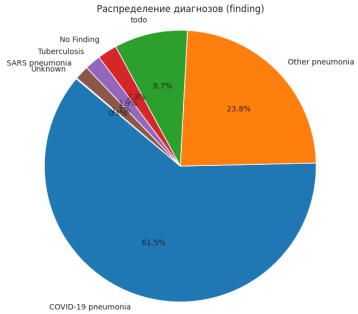
- **Источник данных**:
  - Открытый набор `covid-chestxray-dataset` (GitHub)
  - 950 записей, включая случаи из *The Lancet* (Ухань), *NEJM* (Вьетнам, Тайвань)
- **Обработка**: PySpark — заполнение пропусков, унификация диагнозов, удаление дубликатов.
- **Аналитика**:
  - SQL-запросы (Spark SQL): статистика, топ-3 по возрасту, временные тренды
  - UDF для категоризации возраста: `child`, `young_adult`, `middle_aged`, `senior`
- **Визуализация**: Matplotlib + Seaborn (через Pandas)
- **Выход**: Parquet-файл, партиционированный по `finding`

Технологии: PySpark, Pandas, Matplotlib, Seaborn, Google Colab

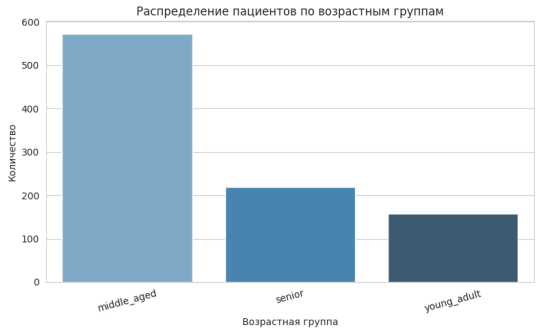
Ключевые статистики

Показатель	Значение
Всего записей	950
Средний возраст	53.7 года
Мужчины	68%
Женщины	33%
Неизвестный пол	8%
COVID-19 pneumonia	61.5%
Other pneumonia	23.8%
SARS pneumonia	1.7%
Bilateral pneumonia	75%
Лимфопения	35%
ICU-госпитализация	23%

Визуализации и выводы

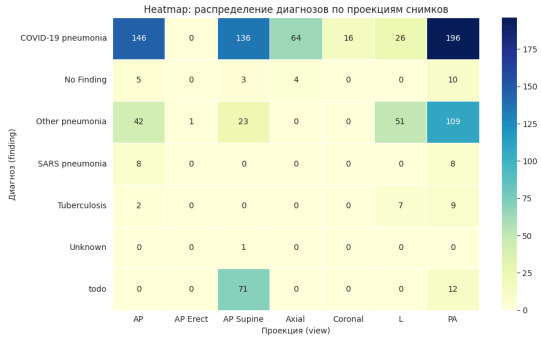


Распределение диагнозов



Возрастные группы

Временной тренд (2003–2020)



Heatmap: диагнозы × проекции

Выводы:

- Доминирование COVID-19 (61.5%) — отражает вспышку 2020 г.

Общий вывод: Метаданные отражают клиническую и эпидемиологическую картину пандемии. PA проекция наиболее подходит для выявления COVID-19. Наибольшее число заболевших зафиксировано у мужчин старше 40 лет.