

Применение BlueCast для классификации клинических кейсов по группам заболеваний на основе симптоматики пациента

Цель: Разработка baseline-модели для прогнозирования группы заболеваний пациента (из 11 классов) на основе клинических описаний и симптомных признаков.



Подготовил: Антон Будняк, группа 633

Made with **GAMMA**

Исходный Датасет и Процесс Выборки

Наш анализ начинается с датасета **PMC-Patients-V2.pkl**, содержащего обширные клинические описания пациентов, метаданные статей и уникальные идентификаторы. Для обеспечения стабильности данных, мы сфокусировались на публикациях (PMID), содержащих не менее 5 клинических кейсов.

Параметр	До Фильтрации	После Фильтрации
Количество записей (df.shape)	15,355	1,535
Уникальные PMID	210,069	431
Уникальные пациенты	250,294	1,200

Разметка Групп Заболеваний: От LLM до "Слабых" Меток

Для создания целевой переменной мы использовали гибридный подход: заголовки статей сопоставлялись с группами заболеваний посредством словарей, созданных с помощью кластеризации LLM. Это позволило получить метки для наших данных.

Заголовок Статьи (Пример)	Присвоенная Группа
Severe pneumonia in immunocompromised patient	Infectious Diseases & Immunology
A case of metastatic colorectal cancer	Oncology
Novel therapy for Parkinson's disease	Neurology & Psychiatry
Congenital heart defect in an infant	Genetic, Congenital & Developmental Disorders
Drug interaction with anticoagulant therapy	Pharmacology, Toxicology & Adverse Effects
Myocardial infarction in young adult	Cardiovascular & Pulmonary Diseases

Всего размечено кейсов: 1535

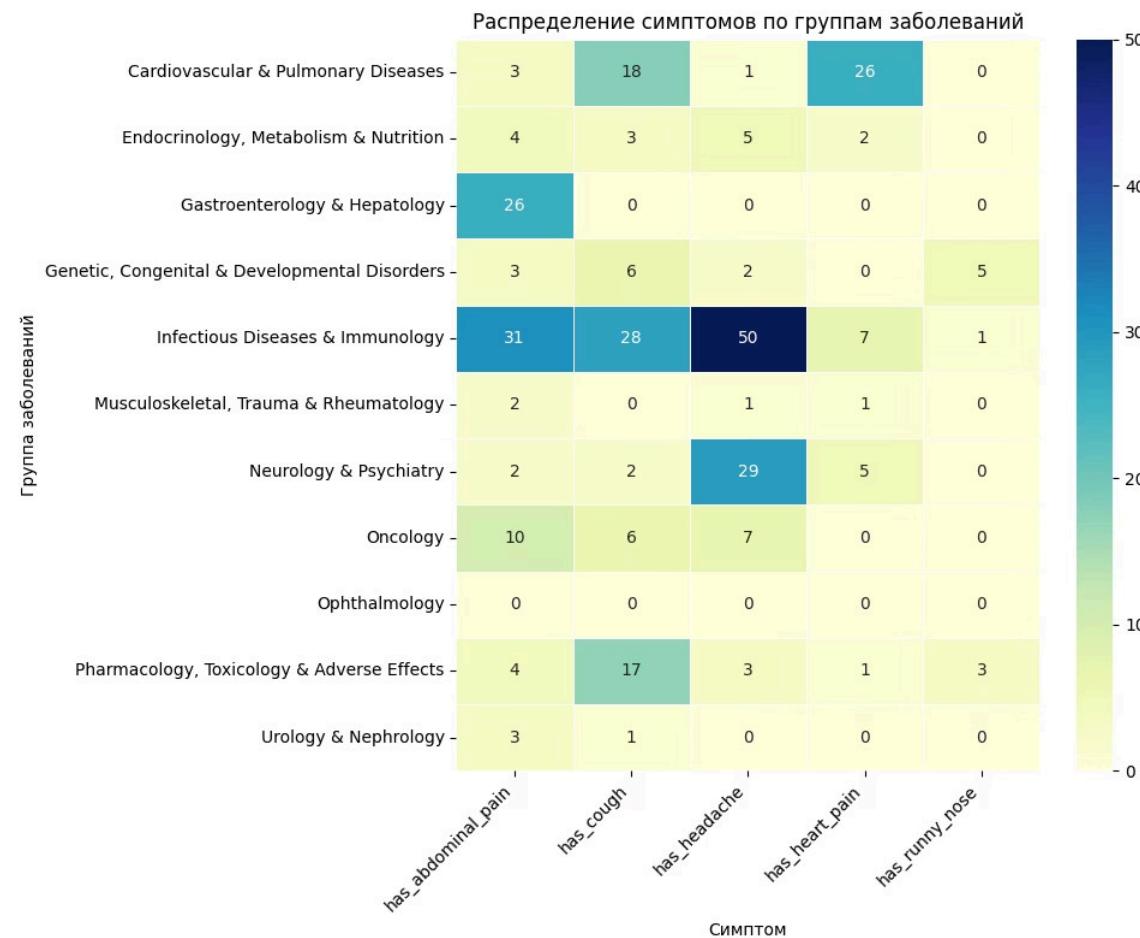
Баланс Классов: Распределение 11 Групп Заболеваний

Group	
Infectious Diseases & Immunology	338
Oncology	231
Neurology & Psychiatry	217
Genetic, Congenital & Developmental Disorders	196
Pharmacology, Toxicology & Adverse Effects	139
Cardiovascular & Pulmonary Diseases	117
Musculoskeletal, Trauma & Rheumatology	98
Ophthalmology	59
Gastroenterology & Hepatology	54
Endocrinology, Metabolism & Nutrition	46
Urology & Nephrology	40

Представленная диаграмма показывает распределение кейсов по 11 основным группам заболеваний, демонстрируя преобладание некоторых категорий, таких как "Инфекционные заболевания и иммунология", что соответствует общей картине клинических исследований.

Симптомные Признаки и Разведочный Анализ Данных (EDA)

Мы извлекли бинарные симптомные признаки из клинических текстов пациентов с использованием ключевых слов, таких как "headache", "abdominal pain", "cough", "runny nose" и "chest/heart pain". Разведочный анализ данных (EDA) включал создание сводной таблицы "Группа × симптом" для выявления корреляций.



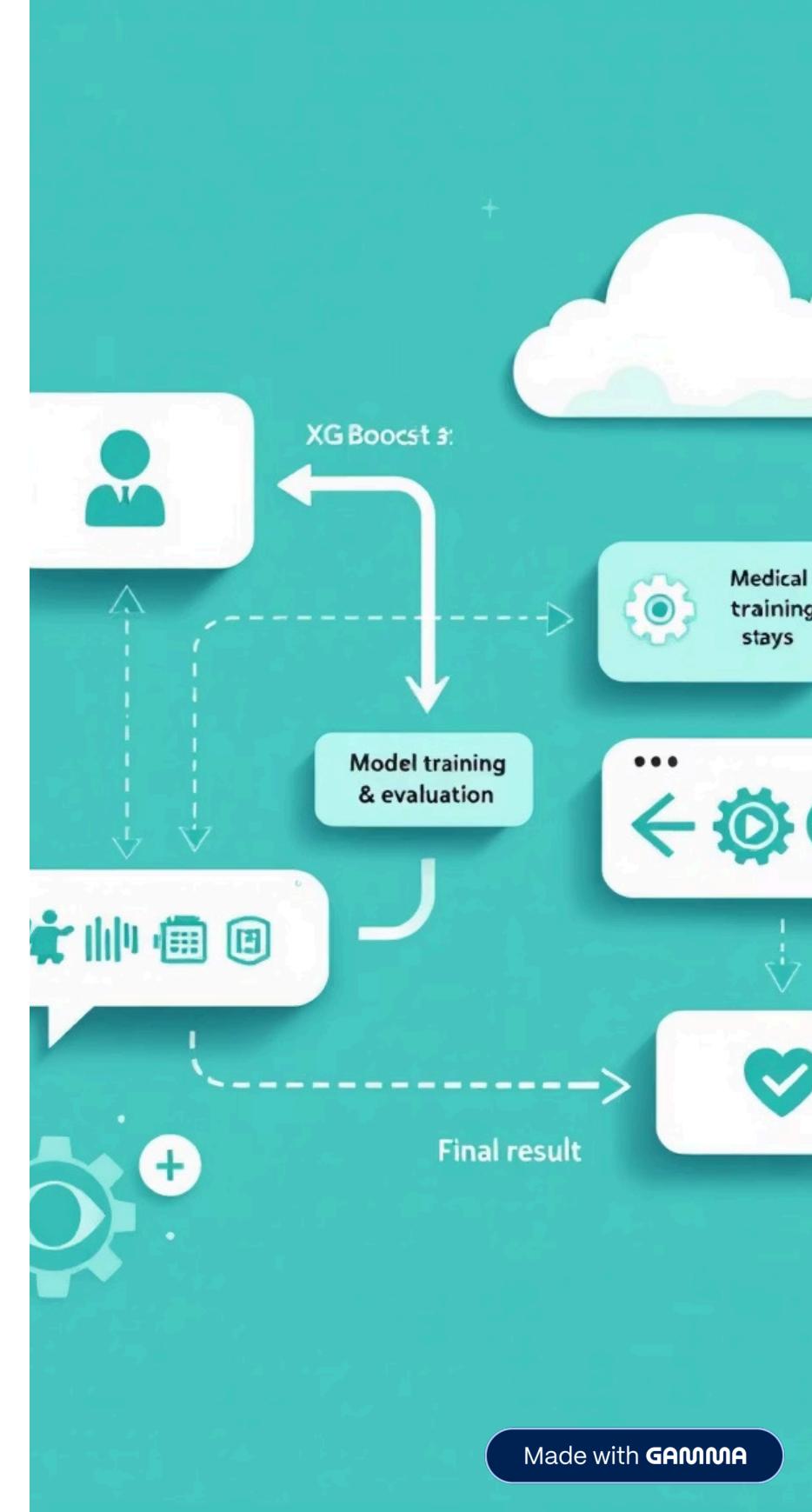
Визуализация в виде тепловой карты позволяет интуитивно оценить связь между наличием определенных симптомов и принадлежностью к конкретным группам заболеваний, подтверждая, что симптомы являются значимыми предикторами.

Постановка Задачи и Baseline-Моделирование с BlueCast

Наша задача — это многоклассовая классификация по 11 группам заболеваний. В качестве входных данных использовались табличные признаки, включающие симптомные флаги, а также подготовленные метаданные и числовые признаки.



Конечная матрица признаков имеет размерность **(1535, 268)**, что указывает на комплексный набор данных для обучения модели.



Конфигурация BlueCast: Автоматизированный Пайплайн

Пайплайн BlueCast автоматизирует ключевые этапы машинного обучения, обеспечивая эффективную подготовку данных и обучение модели.

О1

Определение Типов Признаков

Автоматическое выявление категориальных, числовых и других типов данных.

О2

Обработка Признаков

Удаление константных признаков и приведение схемы данных.

О3

Заполнение Пропусков

Интеллектуальное заполнение отсутствующих значений.

О4

Target Encoding

Преобразование категориальных признаков с учетом целевой переменной.

О5

Обучение XGBoost

Финальное обучение модели для многоклассовой классификации.

```
2025-12-16 10:14:07,840 - root - INFO - Start predicting on new data using Xgboost model.  
2025-12-16 10:14:07,968 - root - INFO - Finished predicting  
2025-12-16 10:14:07,969 - root - INFO - Start reverse-encoding target labels.
```

--- HOLDOUT METRICS ---

Accuracy: 0.6156351791530945

Macro F1 : 0.6113849112348432

Основные Метрики Производительности на Holdout-Выборке

Оценка baseline-модели на отложенной выборке (holdout) показала следующие результаты, подтверждающие наличие устойчивого сигнала для классификации по 11 клиническим группам.

0.6156

Accuracy

Доля правильных предсказаний.

0.6114

Macro F1

Среднее гармоническое взвешенное Precision и Recall.

Значение Macro F1, близкое к Accuracy, свидетельствует о сбалансированной производительности модели по всем классам, несмотря на их дисбаланс.

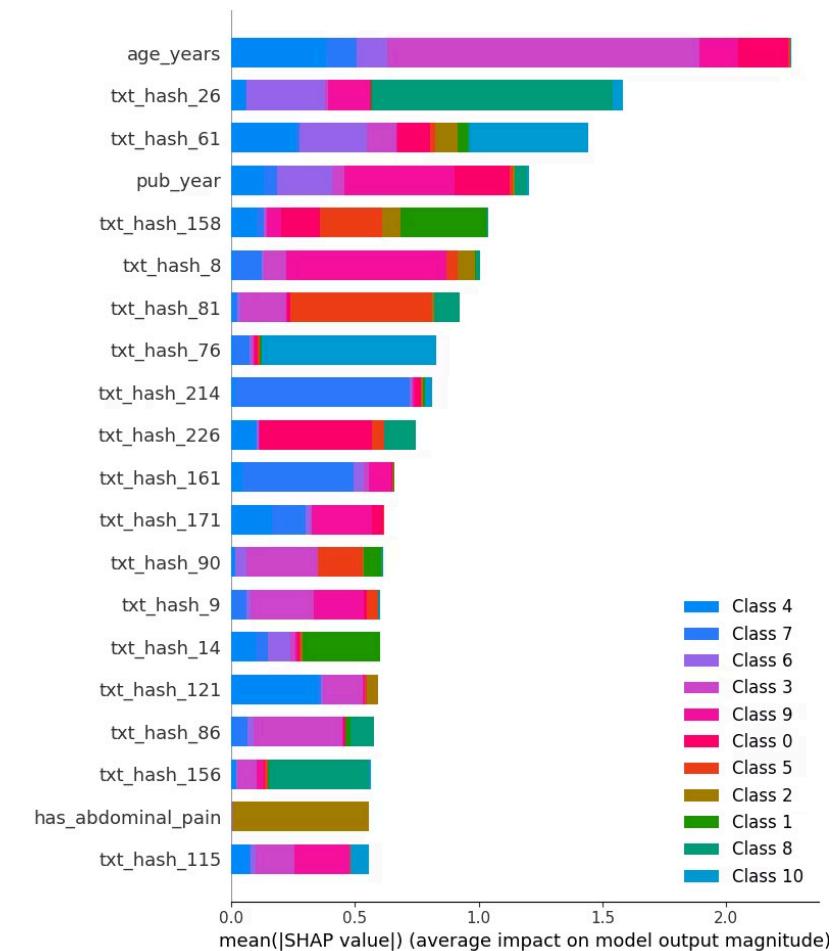
Отчёт по Классам: Детальный Анализ

Precision, Recall и F1-Score

Группа Заболеваний	F1-Score	Precision	Recall
Musculoskeletal, Trauma & Rheumatology	0.778	0.875	0.700
Oncology	0.696	0.696	0.696
Urology & Nephrology	0.667	1.000	0.500
Cardiovascular & Pulmonary Diseases	0.632	0.800	0.522
Infectious Diseases & Immunology	0.614	0.500	0.794
Genetic, Congenital & Developmental Disorders	0.603	0.647	0.564
Neurology & Psychiatry	0.562	0.543	0.581
Ophthalmology	0.560	0.538	0.583
Gastroenterology & Hepatology	0.500	0.800	0.364
Pharmacology, Toxicology & Adverse Effects	0.500	0.688	0.393
Endocrinology, Metabolism & Nutrition	0.615	1.000	0.444

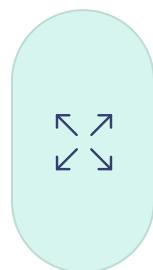
Наилучшие показатели F1-Score наблюдаются в группах **Musculoskeletal, Trauma & Rheumatology (0.778)** и **Oncology (0.696)**, что указывает на высокую предсказательную способность модели в этих областях. Однако в группах **Gastroenterology & Hepatology** и **Pharmacology, Toxicology & Adverse Effects** Recall значительно ниже, что требует дальнейшей оптимизации.

Shap-values по признакам:



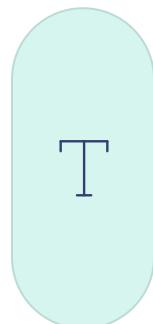
Выводы и Дальнейшие Шаги

Baseline-классификатор на основе BlueCast и XGBoost демонстрирует устойчивую производительность с Macro-F1 около 0.61 для 11 клинических групп - может быть, альтернативой для разметки с помощью LLM в случае жесткой нехватки ресурсов.



Расширение Симптомного Словаря

Улучшение извлечения клинических сущностей из текста для обогащения признаков.



Добавление Текстовых Представлений

Интеграция эмбеддингов текста для учета семантического содержания клинических описаний.



Переход к Сценарию Поиска Схожих Кейсов

Разработка системы retrieval + rerank для поиска наиболее релевантных клинических прецедентов.

