Автоматизированный скрининг заболеваний на рентгенограммах грудной клетки с использованием больших языковых моделей и глубоких сверточных нейронных сетей на наборе данных MIMIC-CXR

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1883337

LLM | MIMIC-CXR | диагностика | радиология | сверточные нейронные сети

Введение

Рентгенологи в условиях ограниченных ресурсов часто сталкиваются с высокой рабочей нагрузкой, особенно при интерпретации рентгенограмм грудной клетки. Ручная разметка крупномасштабных наборов медицинских изображений остаётся дорогостоящей и трудоёмкой задачей. Цель данного исследования — изучить возможность использования больших языковых моделей (LLM), в частности GPT-4o, для генерации бинарных меток наличия заболевания на основе текстовых радиологических заключений, а также применения этих меток для обучения моделей глубокого обучения автоматической классификации рентгенограмм грудной клетки.

Методы

Был разработан двухэтапный конвейер обучения с учителем на основе общедоступного набора данных MIMIC-CXR v2.1.0. На первом этапе GPT-4o получала структурированный клинический протокол для классификации каждого радиологического заключения как «заболевание присутствует» или

«заболевание отсутствует». На втором этапе сгенерированные метки использовались для обучения с учителем четырёх свёрточных нейронных сетей: ResNet-18, DenseNet-121, EfficientNet-B1 и ConvNeXt-Tiny. Для предотвращения утечки данных между выборками применялось разбиение на уровне пациентов в пропорции 70/10/20. Каждая модель обучалась с пятью случайными начальными значениями (42–46), а 95% доверительные интервалы вычислялись с использованием t-распределения. Качество меток оценивалось путём сравнения 210 сгенерированных меток с аннотациями сертифицированного рентгенолога.

Результаты

GPT-4o достигла общей точности 92,9% при сравнении с экспертными метками на валидационном наборе из 210 заключений. Для класса «заболевание присутствует» точность составила 97,4%, а полнота — 90,5%; для класса «заболевание отсутствует» точность составила 87,1%, а полнота — 96,4%. Среди моделей CNN, оценённых на отложенном тестовом наборе, ConvNeXt-Tiny достигла наивысшей площади под кривой (AUC=0,832; 95% ДИ [0,801; 0,863]) и сбалансированной точности (0,739), значительно превзойдя EfficientNet-B1 (AUC=0,797; парный t-критерий, p=0,014). ResNet-18 (AUC=0,822) и DenseNet-121 (AUC=0,808) показали промежуточные результаты. Все модели продемонстрировали значения AUC выше 0,79, что подтверждает жизнеспособность слабого контроля, полученного с помощью LLM.

Обсуждение

Данное исследование демонстрирует, что большие языковые модели могут эффективно использоваться для генерации контрольных меток в задачах медицинской визуализации. Предложенный подход предлагает масштабируемое и недорогое решение для предварительного скрининга заболеваний, особенно в условиях здравоохранения с ограниченной доступностью экспертов. Оценка с множественными случайными начальными значениями и доверительными интервалами обеспечивает строгую оценку стабильности моделей. Требуется дальнейшая работа для повышения надёжности меток и расширения подхода до многоклассовой классификации.