ISSN 2221-2582

УДК 616-056.7+574.21+616.036.12-039.42-053.1

СИСТЕМА СКРИНИНГА ДЕПЕРСОНАЛИЗИРОВАННЫХ ЭЛЕКТРОННЫХ МЕДИЦИНСКИХ КАРТ (ДЭМК) С ЦЕЛЬЮ ВЫЯВЛЕНИЯ РЕДКИХ ГЕНЕТИЧЕСКИХ ЗАБОЛЕВАНИЙ У ДЕТЕЙ ДО 7 ЛЕТ (IPAVLOV SMART CLINIC PLATFORM — CDSS RD (RARE DISEASES) NEURO-SCANNER)

Акопян Лоран Ваганович

ООО «Айпавлов», НИЦ АО «Швабе» в МФТИ. 141701, Московская обл., Долгопрудный, Научный пр., д. 4
E-mail: ceo@ipavlov.ai

Ключевые слова: *орфанные заболевания; генетические заболевания; NLP; iPavlov; медицинские карты; мукополисахаридоз; Фабри; Помпе; Ниманна–Пика A/B.*

Актуальность. Орфанные (редкие) заболевания, как правило, проявляют себя в раннем возрасте и сопровождают человека на протяжении всей его жизни. О наличии заболевания пациент узнает слишком поздно ввиду отсутствия ярко выраженных симптомов. Платформа анализа ДЭМК акцентирует внимание врачей на пациентах, чьи данные медицинских карт свидетельствуют о предрасположенности к определенным заболеваниям. Первичный этап апробации системы включает вероятность вхождения человека в группу риска по одному из перечисленных орфанных заболеваний:

- мукополисахаридоз типа I (МПС I);
- болезнь Фабри;
- болезнь Помпе;
- болезнь Ниманна–Пика A/B.

Материалы и методы. Порядок проведения исследования (методология):

1. Изучение рассматриваемых болезней, поиск похожих решений (похожие работы в подразделе 3.1 «Существующие методы решения задачи»).
2. Анализ поступающих данных с целью поиска паттернов и особенностей в выражении исследуемой группы людей (подробнее про поступающие и тренировочные данные в подразделе 2.1 «Описание природы входных данных»).
3. Получение и извлечение данных в удобный для обработки формат (подробнее в подразделе 2.4 «Описательная статистика поступающих и тренировочных данных»).
4. Препроцессинг и приведение формата данных в табличный формат, необходимый для работы с моделями (подробнее в подразделе 2.4 «Описательная статистика поступающих и тренировочных данных»).
5. Проверка карт на полноту, качество и деперсонализацию (подробнее в подразделах 2.2 и 2.3 «Критерии и методология проверки деперсонализации данных» и «Критерии и методология проверки полноты и качества данных»).
6. Выбор бейзлайн конфигурации и фиксация бейзлайн метрик качества (подробнее про бейзлайн конфигурацию в подразделе 3.3 «Описание бейзлайн конфигурации»).
7. Включение дополнительных модулей обработки и трансформации данных с контролем изменения качества на каждом из этапов (подробнее про добавляемые модули в подразделе 3.4 «Описание добавляемых модулей»).
8. Выбор наилучшей конфигурации в терминах выбранной конкретной метрики качества и скорости работы (метрики оценки качества описаны в подразделе 1.3 «Метрики оценки качества моделей машинного обучения», методика в подразделе 1.5 «Методика постановки экспериментов»).
9. Выполнение предсказаний на всем корпусе данных путем запуска пайплайна МО и логирование результатов в единый файл (подробнее про инструмент трекинга в подразделе 4.2 «Логирование экспериментов в MLFlow»).
10. Выполнение предсказаний на всем корпусе данных путем запуска алгоритмов диагностики, сформированных на основе правил.
11. Отбор карт, одновременно попавших в предсказания пайплайном, МО и алгоритмами на основе правил. Число отобранных карт (трешолд для пайплайна МО) определяется максимально допустимым значением долевого отличия по номерам карт между двумя подходами путем набора группы в порядке уменьшения вероятности предсказания.
12. Визуальный контроль разработчиками качества работы модели по данным (подробнее в подразделе 4.3 «Контроль предсказанных карт»).
13. Передача контрольных результатов карт потенциальных пациентов на проверку экспертам.

Промежуточные результаты. Наиболее эффективным методом, который лег в основу данной системы, стало выявление конкретных релевантных признаков в карте (переданных экспертами). Здесь реализованы подходы полнотекстового поиска бинарных и численных признаков, которые являются явными биомаркерами-индикаторами наличия редких заболеваний у пациентов.

Валидация модели в процессе обучения также происходит за счет аугментированных (синтетических) позитивных карт.

Данный подход позволил выявить двух пациентов (женщина в возрасте 40 лет и подросток в возрасте 13 лет) с заболеванием Фабри (в период с 01.06.2021 по 01.09.2021).

Разработанная система находится на завершающем этапе установленной методологии «Передача контрольных результатов карт потенциальных пациентов на проверку экспертам». Система призвана внести колоссальный вклад в укрепление здоровья нации и снизить смертность и инвалидизацию в детском возрасте.