Секция «2.2 Биоинформатика»

Сравнительный анализ алгоритмов машинного обучения в цитологической диагностике новообразований молочной железы: фокус на клиническую безопасность

Научный руководитель – Саввина Екатерина Андреевна

Шалгинская Божена Игоревна

Студент (специалист)

Воронежский государственный университет инженерных технологий, Воронежская область, Россия

E-mail: bozhena.shalginskaia@mail.ru

Рак молочной железы остается одним из самых распространенных онкологических заболеваний в мире. Эффективность терапии и выживаемость пациентов напрямую зависят от скорости и точности первичной диагностики. Внедрение систем поддержки принятия врачебных решений (СППВР) на базе искусственного интеллекта позволяет автоматизировать анализ цитологических исследований и снизить влияние человеческого фактора [1]. Однако большинство современных IT-исследований в области медицины фокусируется на максимизации общей математической точности моделей (Accuracy). В реальной клинической практике такой подход недопустим, так как цена ложноотрицательного результата (пропуска злокачественной опухоли) — несвоевременное начало лечения и угроза жизни пациента. Целью данной работы является сравнительная оценка алгоритмов машинного обучения с приоритетом на метрику полноты (Recall) для обоснования выбора наиболее клинически безопасной модели скрининга [2].

Исследование проводилось на базе открытого набора данных Breast Cancer Wisconsin (Diagnostic) Dataset [3], содержащего 569 образцов и 30 числовых морфологических признаков клеточных ядер (радиус, текстура, периметр, площадь, гладкость и др.). Выборка была случайным образом разделена на обучающую (80%) и тестовую (20%).

В качестве основного классификатора была выбрана логистическая регрессия. Данный алгоритм строит линейную математическую модель, где вероятность принадлежности новообразования к доброкачественному или злокачественному классу вычисляется с помощью логистической функции (сигмоиды) от взвешенной суммы 30 морфологических признаков клеточных ядер. Поскольку морфологические характеристики клеток имеют различный порядок значений (от сотых долей до тысяч), для корректной оптимизации весов модели была проведена предварительная обработка (препроцессинг) данных — стандартизация всех признаков с использованием алгоритма StandardScaler. Это позволило обеспечить стабильную сходимость алгоритма. Для сравнительного анализа в качестве второй модели применялся алгоритм ансамблевого обучения случайный лес (Random Forest). Разработка велась на языке Python с использованием библиотеки scikit-learn, исходный код исследования открыт и доступен для воспроизведения [4].

На тестовой выборке (n=114) обе модели продемонстрировали высокую общую точность (Accuracy = 97%). Однако детальный анализ матриц ошибок выявил критическую разницу в клинической безопасности алгоритмов.

Модель случайного леса ошибочно классифицировала 3 злокачественных новообразования как доброкачественные (ложноотрицательный результат, см. рис. 2). В свою очередь, алгоритм логистической регрессии после стандартизации признаков показал более высокий уровень безопасности, пропустив лишь 2 случая патологии (Recall для злокачественного класса составил 95%, полнота выявления доброкачественных опухолей — 99%,

см. рис. 1). Таким образом, математически более простой алгоритм логистической регрессии оказался предпочтительнее для задач медицинского скрининга, так как он эффективнее минимизирует риск жизнеугрожающих ошибок второго рода.

Продемонстрировано, что при разработке и интеграции медицинских систем искусственного интеллекта в онкологии приоритетной метрикой оценки должна выступать полнота (Recall), а не общая точность (Accuracy). Исследование доказало, что базовые, интерпретируемые и вычислительно легкие математические модели, такие как логистическая регрессия (при условии грамотного препроцессинга данных), способны обеспечивать более высокий уровень клинической безопасности, чем сложные ансамблевые методы. Это делает их оптимальными кандидатами для внедрения в лабораторные информационные системы поликлиник в качестве надежного «второго мнения» для врачей-цитологов.

Источники и литература

1) Радченко В.Н., Сидоров А.А. Применение методов машинного обучения в задачах медицинской диагностики // Врач и информационные технологии. 2022.
2) McKinney S. M. et al. International evaluation of an AI system for breast cancer screening // Nature. 2020. Vol. 577. P. 89-94.
3) UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Dataset: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
4) Google Colab. Исходный код исследования: https://colab.research.google.com/drive/1bp6X6qrUkxNTCeHuT2ZX5cGVJaZl-GjM?usp=sharing

Иллюстрации

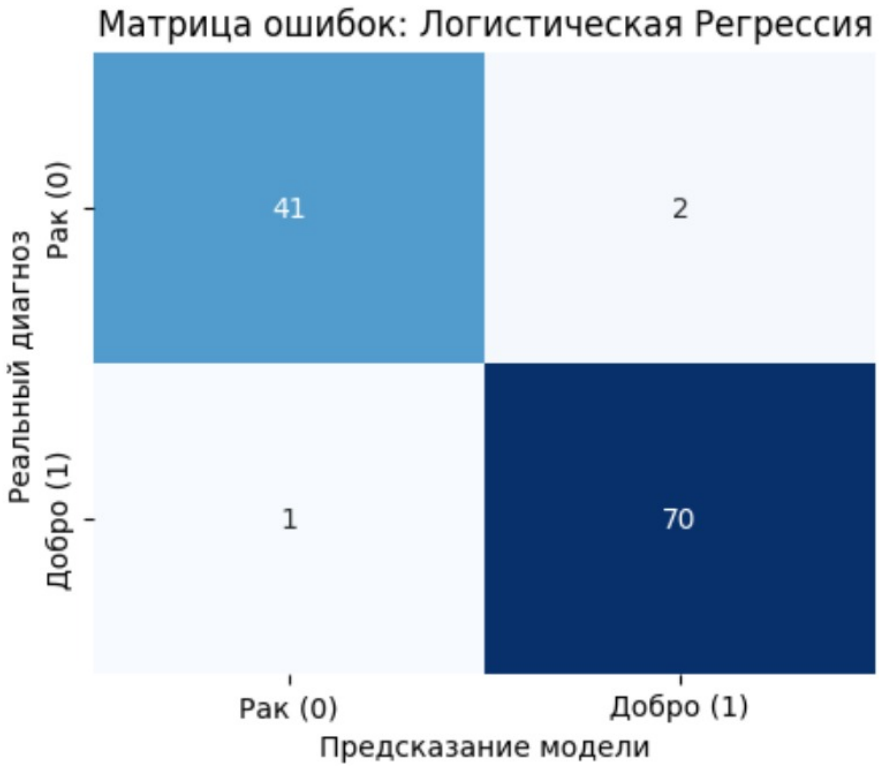

Рис. : Матрица Ошибок: Логистическая Регрессия

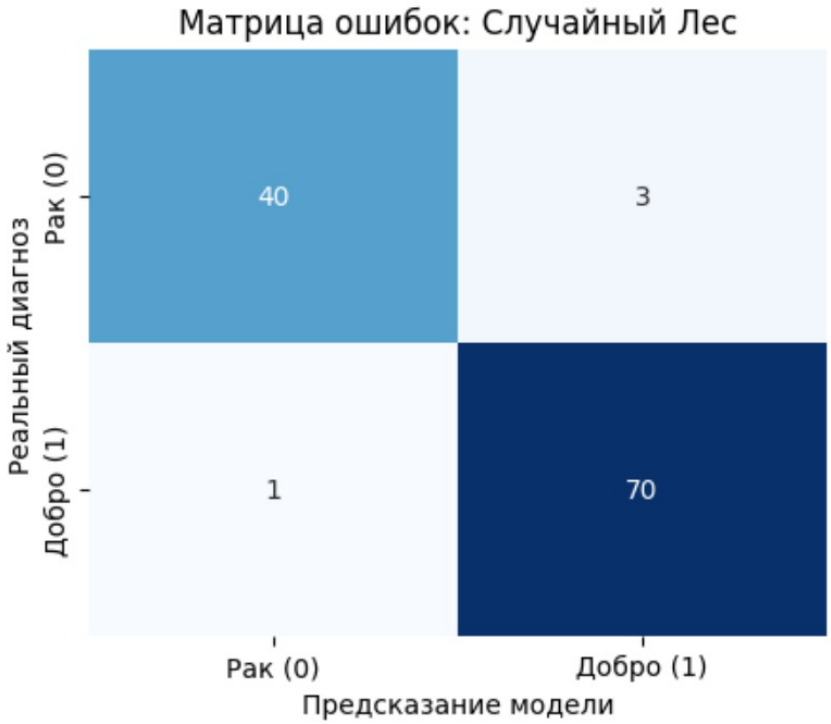

Рис. : Матрица Ошибок: Случайный Лес