Секция «2.2 Биоинформатика»

Ранняя диагностика заболеваний крови: моделирование с использованием лабораторных и амбулаторных данных

Научный руководитель – Шишканов Дмитрий Валерьевич

Иштуганова В.В.[1], *Епифанцев С.А.*[2], *Пискунова О.С.*[3], *Кондратюк Н.А.*[4], *Андреева А.В.*[5], *Бессмертный Д.К.*[6]

1 - Санкт-Петербургский государственный университет, Биологический факультет, Санкт-Петербург, Россия, *E-mail: valeriishtuganova@gmail.com*; 2 - Сколковский институт науки и технологий, Информационные технологии, Москва, Россия, *E-mail: sa.epifancev@gmail.com*; 3 - Московский физико-технический институт, Москва, Россия, *E-mail: piskunova.os@phystech.edu*; 4 - Московский физико-технический институт, Москва, Россия, *E-mail: natalyakon43571@gmail.com*; 5 - Московский физико-технический институт, Москва, Россия, *E-mail: andreeva.av@phystech.edu*; 6 - Российский национальный исследовательский медицинский университет имени Н.И. Пирогова, Москва, Россия, *E-mail: dmitry_ bessmertnyy@mail.ru*

Введение и цель. Гематологические заболевания представляют серьезную медицинскую и социальную проблему. Их своевременная диагностика критически важна, однако нередко осложняется сходством ранних клинических и лабораторных проявлений. Методы машинного обучения демонстрируют высокую эффективность при анализе лабораторных данных, помогая выявлять скрытые закономерности.

Цель работы – разработка модели, способной прогнозировать гематологические заболевания на ранних стадиях по результатам анализа крови пациентов.

Материалы и методы. Данные были получены от 38 041 пациента ФГБУ “НМИЦ Гематологии” Минздрава России (г. Москва) и включали результаты общего и биохимического анализа крови с подтвержденным диагнозом. Для моделирования отбирался первый визит пациента, выполнялась очистка, нормализация и унификация данных. Из 307 диагнозов после фильтрации по значимости и представленности оставлено 32 нозологии, объединённые в 18 классов и 3 группы по МКБ-10. Итоговая выборка составила 11 452 наблюдения.

Среди протестированных алгоритмов машинного обучения лучшие результаты показали модели градиентного бустинга: CatBoost (v.1.2.8) и XGBoost (v.3.1.2). Оптимизация гиперпараметров проводилась с использованием Optuna (v.5.0), применялась стратифицированная кросс-валидация. Ключевым решением для повышения качества стало применение иерархического подхода: на первом уровне строится модель предсказания широкой группы заболеваний, на втором — модель классификации внутри группы по нозологическим классам и на третьем – диагностика до конкретного заболевания. Интерпретируемость моделей проверялась методом SHAP (0.48.0), был учтен дисбаланс классов и проведена калибровка вероятностей предсказания. Репозиторий проекта: https://github.com/Valeriisht/AltynCode

Результаты. По результатам кросс-валидации выбрана модель CatBoost. Значение метрики macro F1-score на тестовой выборке составило 0.68 при предсказании классов заболеваний; анализ SHAP выявил информативные показатели для каждого класса, соответствующие клиническим представлениям о патогенезе. Применение двухуровневого иерархического подхода продемонстрировало улучшение качества модели (macro F1-score 0.71, Accuracy 0.8), что подтверждает перспективность данного метода. Далее планируется внедрение полноценной трехуровневой классификации.

Заключение. Совершенствование модели и ее внедрение в медицинские учреждения будет способствовать повышению клинической осведомленности о гематологических за-

болеваниях, своевременному направлению пациентов на диагностику и более раннему выявлению патологий.

Источники и литература

1) Ning W, Wang Z, Gu Y, et al. Machine learning models based on routine blood and biochemical test data for diagnosis of neurological diseases // Sci Rep. – 2025. – Vol. 15. – DOI: 10.1038/s41598-025-09439-4.
2) Palak, Singla I, Malhotra D, Kumar K. Machine Learning-Driven Insights into Autoimmune Disease Prediction and Patient Outcomes // Proc Int Conf Cybernation Computation. – 2024. – P. 466–470.
3) https://mkb-10.com/