DOI: 10.25205/978-5-4437-1691-6-24

ИСПОЛЬЗОВАНИЕ ИНВАЗИВНЫХ И НЕИНВАЗИВНЫХ ДАННЫХ ПРИ ДИАГНОСТИКЕ САХАРНОГО ДИАБЕТА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

INVASIVE AND NON-INVASIVE DATA FOR MACHINE LEARNING BASED DIAGNOSIS OF DIABETES MELLITUS

И. А. Корзун[1], К. С. Егоров[2]

[1]*Московский физико-технический институт*
[2]*Лаборатория искусственного интеллекта Sber AI Lab, Москва*

I. A. Korzun[1], K. S. Egorov[2]

[1]*Moscow Institute of Physics and Technology*
[2]*Sber AI Lab, Moscow*

✉ ivan.korzun@gmail.com

Аннотация

В настоящей работе проведено сравнение предиктивной значимости данных анамнеза, результатов лабораторных анализов и записей ЭКГ, а также их сочетанного применения при диагностике сахарного диабета методами машинного обучения.

Abstract

The present work compares the predictive value of medical history data, laboratory test results and ECG recordings, as well as their combined use in diagnosis of diabetes mellitus by machine learning methods.

Введение. Сахарный диабет (СД) является одним из наиболее распространенных и клинически значимых заболеваний. По прогнозам, к 2035 г. более чем у 590 млн пациентов будет данное заболевание [1]. Серьезной проблемой является несвоевременная диагностика СД. В 2021 г. почти каждый второй взрослый с СД не был осведомлен о своем диабетическом статусе [2]. Существующие сейчас модели машинного обучения для диагностики СД 2-го типа в большинстве случаев унимодальные [3] и/или используют результаты инвазивных исследований [4], что ограничивает их применение для скрининга. Перспективна разработка мультимодальной модели на основе данных анамнеза и результатов ЭКГ как простого и доступного для скрининга метода.

Цель. Настоящая работа призвана сравнить предиктивную значимость отдельных групп признаков при диагностике СД при использовании различных методов машинного обучения.

Материалы и методы. Была использована общедоступная база данных MIMIC-IV с данными электронных медицинских карт о более чем 160 000 пациентов, получавших терапию в Beth Israel Deaconess Medical Center с 2008 по 2019 г. [5]. Были выбраны данные анамнеза (пол, возраст, рост, вес, индекс массы тела, расовая принадлежность, систолическое и диастолическое артериальное давление), инвазивных исследований (гликированный гемоглобин, глюкоза, С-реактивный белок, липопротеины высокой плотности (ЛПВП), липопротеины низкой плотности, общий холестерин (ОХС), отношение ОХС / ЛПВП, триглицериды), исходные записи ЭКГ (12 отведений, длительность 10 с, частота 500 Гц) и автоматически полученные на их основе признаки. При наличии нескольких измерений одного показателя использовалось среднее значение. После удаления выбросов в выборку попало 6633 пациента (из них 2621 с СД 1-го или 2-го типа) в возрасте от 18 лет до 91 года.

Для предупреждения избыточного влияния пациентов с большим количество записей ЭКГ на всю выборку у каждого пациента было выбрано не более 5 случайным образом отобранных ЭКГ. Всего было получено 23 513 ЭКГ. ЭКГ были обработаны фильтром Баттерворта 3-го порядка с частотой фильтрации от 1 до 47 Гц, приведены к частоте 1280 Гц.

Для обучения использовалось 3 модели: 1) градиентный бустинг с помощью алгоритма LightGBM [6]; 2) полносвязная нейронная сеть (MLP); 3) гибридная сверточно-трансформерная нейронная сеть (CNN + Transformer). Модель 3 получала на вход исходный сигнал ЭКГ, а модели 1 и 2 — автоматически сформированные на основе ЭКГ признаки.

Результаты приведены в таблице. Методы глубокого обучения оказались предпочтительнее при работе с ЭКГ. Даже простейшая модель MLP была эффективнее градиентного бустинга на табличных данных с автома-

тически сформированными ЭКГ признаками (AUC 0,60 и 0,53 соответственно). Наилучшие результаты были достигнуты при использовании лабораторных данных в сочетании с анамнезом и ЭКГ (AUC 0,91), при этом данные ЭКГ имели небольшое влияние (без ЭКГ AUC снизился на 0,002). При использовании данных анамнеза и ЭКГ наиболее эффективной оказалась модель (3) (AUC 0,76).

Заключение. Хотя результаты инвазивных диагностических исследований и имеют наибольшую предиктивную значимость при диагностике СД, сочетанное использование ЭКГ, данных анамнеза и неинвазивных исследований представляется перспективным для создания скринингового метода.

Результаты обучения моделей на различных группах признаков

Данные	Модель								
	LightGBM			MLP			CNN+Transformer		
	AUC	Sn	Sp	AUC	Sn	Sp	AUC	Sn	Sp
Анамнез	0,651	0,474	0,829	**0,720**	0,641	0,711			
Анализы	0,825	0,749	0,902	**0,896**	0,778	0,886			
Анамнез + анализы	0,835	0,758	0,911	**0,904**	0,795	0,897			
ЭКГ	0,530	0,327	0,734	0,601	0,378	0,738	**0,676**	0,199	0,840
Анамнез + ЭКГ	0,625	0,477	0,773	0,744	0,631	0,716	**0,756**	0,588	0,754
Анамнез + анализы + ЭКГ	0,812	0,751	0,873	**0,906**	0,810	0,882	0,905	0,835	0,832

Примечание. AUC — площадь под кривой; Sn — чувствительность; Sp — специфичность.

Литература

1. Reed J., Bain S., Kanamarlapudi V. A Review of Current Trends with Type 2 Diabetes Epidemiology, Aetiology, Pathogenesis, Treatments and Future Perspectives // Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy. 2021. Vol. 14. P. 3567–3602.

2. Ogurtsova K., Guariguata L., Barengo N. C. et al. IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021 // Diabetes Research and Clinical Practice. 2022. Vol. 183. P. 109118.

3. Mohsen F., Al-Absi H. R. H., Yousri N. A. et al. A scoping review of artificial intelligence-based methods for diabetes risk prediction // Npj Digital Medicine. 2023. Vol. 6 (1). P. 197.

4. Wee B. F., Sivakumar S., Lim K. H. et al. Diabetes detection based on machine learning and deep learning approaches // Multimedia Tools and Applications. 2023. Vol. 83 (8). P. 24153–24185.

5. Johnson, A. E. W., Bulgarelli L., Shen L. et al. MIMIC-IV, a freely accessible electronic health record dataset // Scientific Data. 2023. Vol. 10 (1). P. 1.

6. Ke G., Meng Q., Finley T. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, December 2017. P. 3149–3157.