

ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОИСКА НОВЫХ ЛАБОРАТОРНЫХ МАРКЕРОВ ВОЗРАСТ-АССОЦИИРОВАННЫХ ЗАБОЛЕВАНИЙ

DOI: 10.37586/2949-4745-4-2024-208-213

УДК: 616-079.4

Захарчук С.А.^{}*, Миронов Н.А.^{}, Плисюк А.Г.^{}, Орлова Я.А.^{}

Факультет фундаментальной медицины МГУ им. М.В. Ломоносова, Москва, Россия

* Автор, ответственный за переписку, Захарчук Софья Александровна.

E-mail: szakharchuk5@gmail.com

Резюме

ВВЕДЕНИЕ. Глобальное старение населения ведет к росту возраст-ассоциированных заболеваний, таких как артериальная гипертензия. Существующие диагностические методы ограничены недостаточной чувствительностью клинико-лабораторных маркеров и ограничениями традиционных статистических методов, что подчеркивает необходимость новых подходов к анализу медицинских данных. Цель исследования — выявить новые лабораторные маркеры артериальной гипертензии с помощью моделей машинного обучения и сравнить их эффективность с традиционными методами. Проанализированы клинико-лабораторные данные 2228 пациентов старше 65 лет, обратившихся за медицинской помощью в клинику МНОИ МГУ. Были использованы логистическая регрессия с регуляризацией, случайный лес и градиентный бустинг. Эффективность оценивалась по AUC ROC, а вклад признаков — через SHAP-анализ. Модели машинного обучения превзошли традиционную логистическую регрессию. Градиентный бустинг достиг AUC ROC 0,85 против 0,78 у регрессионных методов. Ключевыми признаками АГ оказались возраст, показатель RDW-SD и уровень креатинина, тогда как традиционные методы фокусировались преимущественно на возрасте. Выводы подтверждают высокий потенциал методов машинного обучения в диагностике возраст-ассоциированных заболеваний. Модели обеспечивают более высокую точность и выявляют сложные взаимосвязи между показателями, что может улучшить раннюю диагностику и оптимизировать диагностические алгоритмы.

Старение населения является глобальной демографической тенденцией, приводящей к росту и распространенности возраст-ассоциированных заболеваний [1]. Ранняя и точная диагностика данных заболеваний приобретает особую актуальность для эффективного лечения и снижения рисков осложнений у пожилых пациентов. Однако современные диагностические алгоритмы часто ограничены недостаточной чувствительностью и специфичностью клинико-лабораторных маркеров, а стандартные методы анализа данных могут не всегда выявлять скрытые или нелинейные зависимости, существенные для диагностики и оценки прогноза [2]. В связи с этим возникает необходимость в разработке новых подходов к интерпретации медицинских данных и выявлению более точных диагностических крите-

риев. Развитие методов машинного обучения (МО) и искусственного интеллекта открывает новые возможности для анализа больших и сложных массивов медицинской информации. Эти методы способны обнаруживать сложные и нелинейные взаимосвязи между различными параметрами, что может значительно расширить понимание патогенеза, повысить точность диагностики и эффективность лечения возраст-ассоциированных заболеваний в условиях глобального старения населения [3, 4].

ЦЕЛЬ ИССЛЕДОВАНИЯ. Поиск новых лабораторных маркеров, ассоциированных с артериальной гипертензией (АГ) с использованием моделей МО, а также оценка их эффективности в сравнении с традиционными статистическими методами.

МАТЕРИАЛЫ И МЕТОДЫ. В исследовании проведен ретроспективный анализ базы данных, включающей клиническо-лабораторные показатели 2228 пациентов в возрасте старше 65 лет, прикрепленных к МНОИ МГУ им. М.В. Ломоносова и обращавшихся за медицинской помощью или обследованных в рамках профилактических осмотров в период с 2020 по 2021 год. Предварительная обработка данных включала удаление пропусков, обработку выбросов и нормализацию признаков. В представленном пилотном исследовании для отработки инструментов анализа в качестве целевой переменной рассматривался диагноз АГ. Использовалась бинарная логистическая регрессия, как стандартный и общепринятый метод поиска маркеров заболеваний, и модели МО, которые включали логистическую регрессию с регуляризацией, модель случайного леса и градиентный бустинг. Производительность методов оценивали с помощью AUC ROC. Вклад каждого признака интерпретировали путем SHAP-анализа.

РЕЗУЛЬТАТЫ. Модели МО показали превосходство по сравнению с традиционной логистической регрессией. Градиентный бустинг достиг AUC ROC 0,85, в то время как логистическая регрессия продемонстрировала AUC ROC 0,78 (рис. 1). Важнейшими признаками, ассоциированными с наличием АГ, стали возраст, показатель ширины распределения эритроцитов (RDW-SD) и уровень креатинина, в то время как стандартный регрессионный анализ опирался преимущественно на возраст (рисунки 2, 3). Возраст и уровень креатинина ожидаемо стали значимыми факторами, в то время как показатель RDW-SD, который обычно оценивается при анемиях и ассоциируется с состоянием эритропоэза, был также определен значимым маркером АГ. Полученные данные могут свидетельствовать о более широкой связи воспалительных процессов и состоянии эритропоэза с развитием АГ.

ВЫВОДЫ. Результаты исследования демонстрируют высокий потенциал применения моделей МО в медицинской диагностике для поиска новых лабораторных маркеров возраст-ассоциированных заболеваний. Методы МО не только обеспечивают более высокую точность предсказаний, но и позволяют выявлять сложные, ранее неизвестные взаимосвязи между клиническо-лабораторными показателями и развитием и прогрессированием заболеваний. В отличие от традиционных статистических методов они учитывают нелинейные и многомерные взаимодействия между признаками. Полученные результаты могут служить основой для проведения дальнейших клинических исследований, направленных на подтверждение выявленных ассоциаций и разработку новых мишеней для вмешательств, а в перспективе — на изме-

нение диагностических и терапевтических подходов. Применение моделей в клинической практике может способствовать улучшению процедуры ранней диагностики возраст-ассоциированных заболеваний, снижению заболеваемости и смертности, а также улучшению качества жизни пациентов.

Ключевые слова: возраст-ассоциированные заболевания; модели машинного обучения; лабораторные маркеры; анализ данных; артериальная гипертензия.

Для цитирования: Захарчук С.А., Миронов Н.А., Плисюк А.Г., Орлова Я.А. Использование моделей машинного обучения для поиска новых лабораторных маркеров возраст-ассоциированных заболеваний. *Проблемы геронауки*. 2024; 4(8):208-213. DOI: 10.37586/2949-4745-4-2024-208-213

USING MACHINE LEARNING MODELS TO DISCOVER NEW LABORATORY MARKERS OF AGE-RELATED DISEASES

Zakharchuk S.A.^{}*, Mironov N.A.^{}, Plisyuk A.G.^{}, Orlova Ya.A.^{}

Faculty of Fundamental Medicine, Lomonosov Moscow State University, Moscow, Russia

* Corresponding author: Sofia A. Zakharchuk. E-mail: szakharchuk5@gmail.com

Abstract

BACKGROUND. Global population aging leads to an increased rate of age-related diseases such as arterial hypertension. Existing diagnostic methods are limited by the insufficient sensitivity of clinical and laboratory markers and the restrictions of conventional statistical methods. This highlights the need for new approaches to medical data analysis. This study is aimed to identify new laboratory markers of arterial hypertension using machine learning models and to compare their effectiveness compared with conventional methods. It involved the analysis of clinical and laboratory data of 2,228 patients over 65 years of age who sought medical advice at the clinic of the Medical Research and Educational Center of Lomonosov Moscow State University. Logistic regression with regularization, random forest, and gradient boosting were employed. Effectiveness was assessed using the AUC ROC, while attribute contributions were assessed by means of SHAP analysis. Machine learning models outperformed conventional logit regression. Gradient boosting achieved an AUC ROC of 0.85 compared to 0.78 for regression methods. The key attributes associated with arterial hypertension were age, RDW-SD value and creatinine levels, whereas conventional methods focused mainly on age. The conclusions confirm the high potential of machine learning methods in diagnosing age-related diseases. These models ensure higher accuracy and reveal complex interrelationships between indicators, which can improve early diagnosis and optimize diagnostic algorithms. Population aging is a global demographic trend leading to an increase in the incidence of age-related diseases [1]. Early and accurate diagnosis of these conditions is becoming particularly important for effective treatment and minimization of the risk for complications in elderly patients. However, current diagnostic algorithms are often limited by the insufficient sensitivity and specificity of clinical and laboratory markers, while standard data analysis methods may not always detect hidden or

nonlinear dependencies that are critical for diagnosis and prognosis assessment [2]. This necessitates the development of new approaches to interpretation of medical data and identification of more precise diagnostic criteria. Advances in machine learning (ML) and artificial intelligence (AI) offer new possibilities for the analysis of large and complex medical datasets. These methods allow uncovering complex and nonlinear relationships between various parameters, which can significantly enhance our understanding of pathogenesis, improve diagnostic accuracy, and increase the treatment effectiveness of age-related diseases in conditions of global population aging [3, 4].

AIM. To identify new laboratory markers associated with arterial hypertension using ML models and to evaluate their effectiveness in comparison with conventional statistical methods.

MATERIALS AND METHODS. A retrospective analysis was performed on a database containing clinical and laboratory indicators of 2,228 patients over 65 years of age who were affiliated with the Medical Research and Educational Center of Lomonosov Moscow State University and sought medical advice or underwent preventive examinations from 2020 to 2021. Data preprocessing included handling of missing values, dealing with outliers, and normalization of attributes. In this pilot study conducted to refine the analytical tools, the presence of arterial hypertension was considered the target variable. Binary logistic regression, a standard and widely accepted method for identification of disease markers, as well as ML models such as logistic regression with regularization, a random forest model, and gradient boosting were employed. The performance of these methods was assessed using the area under the ROC curve (AUC ROC). The contribution of each attribute was interpreted using the SHAP analysis.

RESULTS. The ML models have been proven to be superior compared with conventional logistic regression. Gradient boosting achieved an AUC ROC of 0.85, whereas logistic regression showed an AUC ROC of 0.78 (Fig. 1). The most important attributes associated with the presence of arterial hypertension were age, red blood cell distribution width (RDW-SD), and creatinine levels. Meanwhile, the standard regression analysis relied primarily on age (Fig. 2, 3). Age and creatinine levels were expectedly significant factors, while RDW-SD, typically evaluated in the context of anemia and associated with erythropoiesis, was also identified as a significant marker of arterial hypertension. These findings may indicate a broader relationship between inflammatory processes and the state of erythropoiesis with the development of arterial hypertension.

CONCLUSIONS. Our findings reveal the high potential of ML models in medical diagnostics for identification of new laboratory markers of age-related diseases. ML methods not only provide higher predictive accuracy but also uncover complex, previously unknown relationships between clinical and laboratory indicators and the onset and progression of diseases. Unlike conventional statistical methods, they account for nonlinear and multidimensional interactions between features. These findings can serve as a basis for further clinical research aimed at confirming the associations found and developing new targets for intervention, potentially leading to changes in diagnostic and treatment approaches. Implementation of these models in clinical practice may improve early diagnosis of age-related diseases, reduce morbidity and mortality, and ultimately enhance patient quality of life.

Keywords: age-related diseases; machine learning models; laboratory markers; data analysis; arterial hypertension.

For citation: Zakharchuk S.A., Mironov N.A., Plisyuk A.G., Orlova Ya.A. Using machine learning models to discover new laboratory markers of age-related diseases *Problems of Geroscience*. 2024; 4(8):208-213. DOI: 10.37586/2949-4745-4-2024-208-213

ПРИЛОЖЕНИЯ / APPENDICES

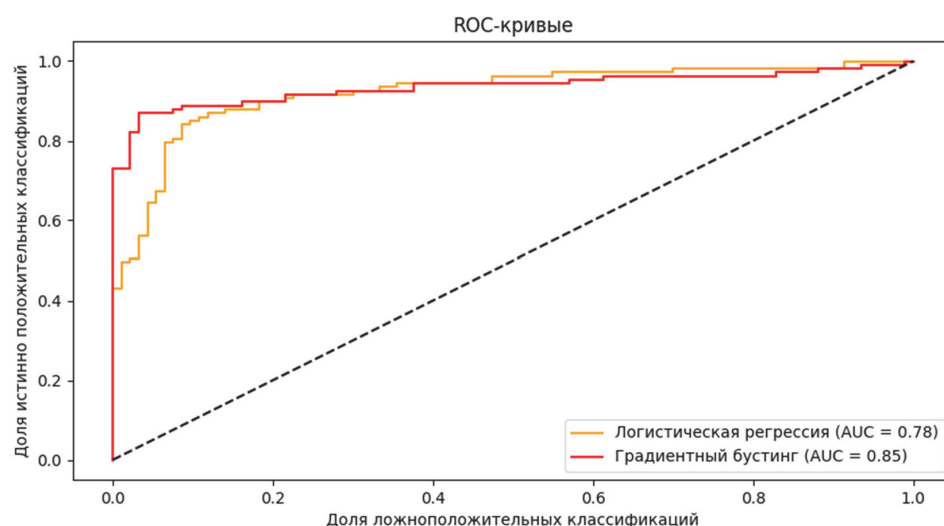


Рисунок 1. ROC-кривые логистической регрессии и градиентного бустинга
Figure 1. ROC curves of logistic regression and gradient boosting

Рисунок подготовлен авторами по собственным данным / The figure was prepared by the authors based on their own data



Рисунок 2. SHAP анализ признаков модели логистической регрессии
Figure 2. SHAP analysis of the logistic regression model attributes

Рисунок подготовлен авторами по собственным данным / The figure was prepared by the authors based on their own data

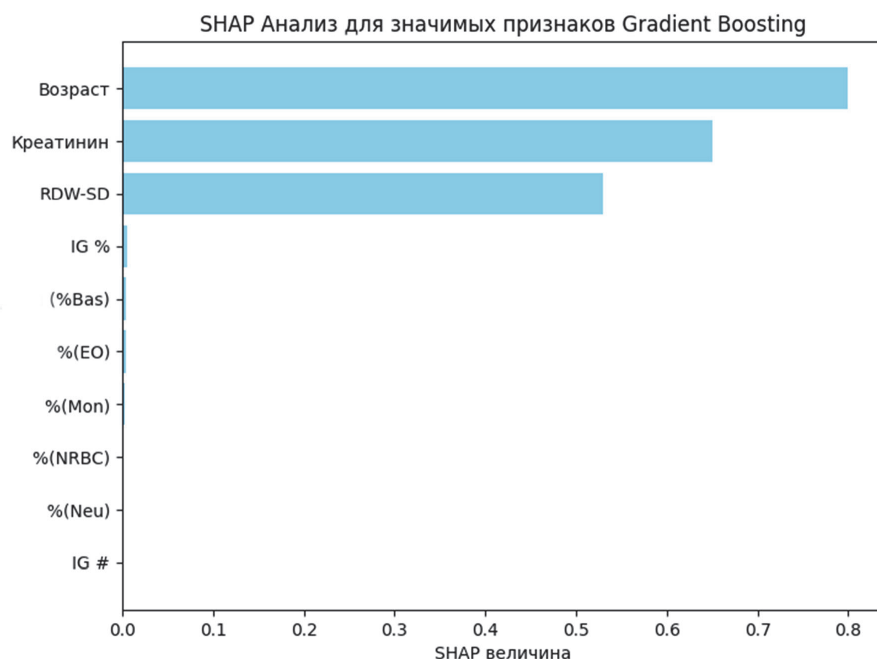


Рисунок 3. SHAP анализ признаков модели градиентного бустинга
Figure 3. SHAP analysis of the gradient boosting model attributes

Рисунок подготовлен авторами по собственным данным / The figure was prepared by the authors based on their own data

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ / ADDITIONAL INFORMATION

Финансирование. Научная работа выполнена без привлечения внешнего финансирования.

Funding sources. The study was conducted without external funding.

Конфликт интересов. Авторы декларируют отсутствие конфликта интересов.

Conflict of Interests. The authors declare no conflicts of interest.

ORCID АВТОРОВ:

Захарчук С.А. — 0009-0008-8220-351X
Миронов Н.А. — 0000-0001-6729-4371
Плисюк А.Г. — 0000-0003-2015-4712
Орлова Я.А. — 0000-0002-8160-5612

ORCID:

Zakharchuk S.A.: 0009-0008-8220-351X
Mironov N.A.: 0000-0001-6729-4371
Plisyuk A.G.: 0000-0003-2015-4712
Orlova Ya.A.: 0000-0002-8160-5612

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. **World Health Organization.** *Global Report on Ageism.* Geneva: WHO; 2021.
2. **Farrag M.A., Al-Harthi R., Saeed M., et al.** Machine Learning Applications in Geriatrics: A Systematic Review. *Journal of the American Medical Directors Association.* 2021;22(8):1621–1627.e1.
3. **Wang F., Preininger A.** AI in Health: State-of-the-Art, Challenges, and Future Directions. *Yearbook of Medical Informatics.* 2020;29(1):16–26.
4. **Mesko B., Gorog M.** A Short Guide for Medical Professionals in the Era of Artificial Intelligence. *NPJ Digital Medicine.* 2020;3:126.