УДК 004.89
DOI 10.17513/snt.40419

ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОПРЕДЕЛЕНИЯ ГРУПП РИСКА ХРОНИЧЕСКИХ ЗАБОЛЕВАНИЙ СРЕДИ ПАЦИЕНТОВ

[1]Королева Я.А., [2]Родионов А.В.

[1]ОГБУЗ «Иркутская городская клиническая больница № 3», Иркутск;
[2]ФГБОУ ВО «Байкальский государственный университет», Иркутск,
e-mail: avr-v@yandex.ru

Хронические неинфекционные заболевания являются важной медико-социальной проблемой, оказывающей существенное влияние на структуру заболеваемости и смертности населения. Актуальность исследования обусловлена необходимостью повышения точности диагностики, персонализации подходов к профилактике и снижению нагрузки на систему здравоохранения. Цель работы – оценка возможности применения современных методов машинного обучения для прогнозирования вероятности развития повышенного артериального давления у взрослых пациентов, наблюдающихся в условиях поликлинического звена. В работе проанализированы обезличенные карты 1843 пациентов. После предварительной обработки, включающей очистку и нормализацию данных, были исследованы следующие алгоритмы: Random Forest, Gradient Boosting, XGBoost, метод K-ближайших соседей и рекуррентная нейронная сеть LSTM. Для верификации качества построенных моделей применялись метрики точности, полноты, F1-мера и ROC-AUC. Результаты апробации показали, что Gradient Boosting и рекуррентная нейронная сеть LSTM наиболее успешно справились с задачей стратификации выборки: пациенты были корректно распределены на группы с отсутствием заболевания, наличием артериальной гипертензии и повышенным риском ее развития. Были показаны ключевые факторы риска – гиперхолестеринемия, неправильное питание и избыток массы тела. Полученные результаты подтверждают целесообразность и перспективность внедрения инструментов машинного обучения, в частности градиентного бустинга и нейросетевых моделей, в клинические информационные системы с целью автоматизированного скрининга артериальной гипертензии и последующего планирования профилактических мероприятий.

Ключевые слова: машинное обучение, нейронные сети, медицинская аналитика, хронические заболевания, классификация данных, интеллектуальный анализ данных

MACHINE LEARNING ALGORITHMS FOR IDENTIFYING CHRONIC DISEASE RISK GROUPS AMONG PATIENTS

[1]Koroleva Ya.A., [2]Rodionov A.V.

[1]Irkutsk City Clinical Hospital № 3, Irkutsk;
[2]Baikal State University, Irkutsk, e-mail: avr-v@yandex.ru

Chronic non-communicable diseases represent a major medical and social challenge, exerting a substantial influence on morbidity and mortality patterns in the population. The relevance of this study lies in the need to improve diagnostic accuracy, personalize preventive strategies, and reduce the burden on the healthcare system. Objective: to assess the feasibility of using modern machine-learning methods to predict the probability of developing elevated blood pressure in adult patients followed up in an outpatient (polyclinic) setting. We analyzed anonymized records of 1,843 patients. After preliminary data processing, including cleansing and normalization, we investigated the performance of the following algorithms: Random Forest, Gradient Boosting, XGBoost, k-Nearest Neighbours, and a recurrent LSTM neural network. Model quality was validated with Accuracy, Recall, F1-score, and ROC-AUC metrics. The results showed that Gradient Boosting and the LSTM network performed best at stratifying the cohort, correctly assigning patients to groups with no disease, existing arterial hypertension, or an elevated risk of its development. Key risk factors identified were hypercholesterolemia, poor diet, and excess body weight. These findings confirm the advisability and promise of integrating machine-learning tools–particularly gradient boosting and neural-network models–into clinical information systems for automated arterial-hypertension screening and subsequent planning of preventive interventions.

Keywords: machine learning, neural networks, medical analytics, chronic diseases, data classification, data mining

Введение

В настоящее время в современной медицине отмечается рост распространенности хронических неинфекционных заболеваний (ХНИЗ), таких как сахарный диабет, сердечно-сосудистая патология, хроническая бронхолегочная патология и др. Эти заболевания оказывают существенное негативное влияние на продолжительность и качество жизни пациентов.

В свете повсеместной цифровизации и постоянного увеличения объемов медицинских данных методы машинного обучения (ML) представляют собой мощный инструмент для выявления лиц, входящих в группу риска развития ХНИЗ [1, 2]. Точное

и своевременное определение группы риска позволяет улучшить результативность проводимых профилактических мероприятий и повысить эффективность лечебно-диагностического процесса. При этом для обеспечения точности и надежности прогнозов важно подобрать и адаптировать алгоритмы машинного обучения с учетом специфики медицинских данных, разнообразия клинических показателей и наличия пропусков и иных искажений в исходных данных.

Цель исследования – построение классификатора и оценки качества классификации пяти алгоритмов – случайного леса, градиентного бустинга, XGBoost, метода k-ближайших соседей и рекуррентной нейронной сети с ячейками долгой краткосрочной памяти (LSTM).

Материалы и методы исследования

Для решения задачи классификации в данной работе рассмотрены следующие алгоритмы: Random Forest (RF) – ансамблевый метод, способный обрабатывать нелинейные зависимости, что в свою очередь обеспечивает вывод интерпретируемых результатов через значения важности признаков [3, 4]; Gradient Boosting (GB) – продвинутый алгоритм машинного обучения для решения задач классификации и регрессии. Он строит предсказание в виде ансамбля слабых предсказывающих моделей, которыми в основном являются деревья решений [5, 6]; XGBoost (XGB) представляет собой усовершенствованный градиентный бустинг [7, 8], он достаточно эффективен на малых и несбалансированных выборках; метод K-ближайших соседей (kNN) – простой алгоритм, анализирующий расстояния между точками [9, 10], который подходит для данных с небольшим числом признаков; LSTM – это рекуррентная нейронная сеть, способная учитывать временные зависимости и сложные зависимости между признаками [11, 12] с использованием оптимизатора Adam. Для обучения моделей использован массив эмпирических данных: 1843 записи о взрослых пациентах, наблюдавшихся в ОГБУЗ «Иркутская городская клиническая больница № 3» г. Иркутска в период с 2021 по 2023 г. Собранный датасет содержит различные факторы риска (ФР), представленные в табл. 1, и целевую переменную – диагноз «артериальная гипертензия». Целевая переменная делит пациентов на три класса: Класс 0 (Здоров): пациент не имеет хронического заболевания; класс 1 (Заболевание): установлен диагноз «артериальная гипертензия»; класс 2 (Риск): высокая вероятность постановки диагноза в будущем.

Так как в датасете имеется дисбаланс классов, то в процессе обработки и анализа были использованы следующие операции:

1. Стратифицированное разбиение выборки – сохранение пропорций классов при разделении данных, при этом 70% записей использовалось для обучения, 30% – для тестирования.

2. Метрики, устойчивые к дисбалансу, такие как ROC-AUC и F1-мера.

3. Балансировка весов классов – автоматическое увеличение значимости меньших классов в градиентных методах.

Для проведения исследования было применено следующее программное обеспечение:

1) один из самых популярных языков программирования для машинного обучения и анализа данных – Python;

2) Google Colab – облачная среда для разработки на Python [13], которая особенно удобна для решения ML-задач.

Для нормализации данных использовался *StandardScaler* с целью приведения значений переменных к единому масштабу. Для модели LSTM данные преобразовывались в трехмерный формат (*samples, timesteps, features*), чтобы учесть сложные зависимости.

Таблица 1

Факторы риска

Переменная	Описание
Избыточная масса тела	Индекс массы тела выше 25
Курение	Наличие вредной привычки
Алкоголь	Наличие вредной привычки
Гиподинамия	Низкий уровень физической активности
Нерациональное питание	Нарушение сбалансированности рациона
Гиперхолистеринемия	Повышенный уровень холестерина

Источник: составлено авторами.

Для моделей Random Forest, Gradient Boosting, kNN была использована библиотека *Scikit-learn*, улучшенный градиентный бустинг реализован с использованием одноименной библиотеки *XGBoost*, а LSTM (Long Short-Term Memory) реализована с помощью библиотек *TensorFlow и Keras* [13].

Для оценки качества классификации моделей применялись следующие метрики [14, 15]:

1. Accuracy – отражает долю правильно прогнозируемых событий относительно их общего числа:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

где TP (True Positive) – количество истинно положительных предсказаний (модель правильно определила положительный класс), TN (True Negative) – количество истинно отрицательных предсказаний (модель правильно определила отрицательный класс), FP (False Positive) – количество ложно положительных предсказаний (модель ошибочно определила отрицательный класс как положительный), FN (False Negative) – количество ложно отрицательных предсказаний (модель ошибочно определила положительный класс как отрицательный).

2. Recall (Полнота) – доля корректно распознанных положительных событий для каждого класса:

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}, i \in \{0,1,2\}.$$

3. Precision (Точность) – это метрика, которая измеряет, насколько точно модель классифицирует положительные случаи. Она показывает, сколько объектов, предсказанных моделью как положительные, действительно являются положительными.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

4. F1-мера: взвешенное среднее точности и полноты:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

5. Метрика ROC-AUC отражает среднюю площадь под ROC-кривой [13] для классов (One-vs-Rest):

$$\text{ROC-AUCсред} = \frac{\text{ROC} - \text{AUC0} + \text{ROC} - \text{AUC1} + \text{ROC} - \text{AUC2}}{3}.$$

Таблица 2

Сравнительный анализ метрик оценки.

Алгоритм	Accuracy	Recall (сред.)	F1-мера (сред.)	ROC-AUC (сред.)
Random Forest	0,93	0,87	0,89	0,978
Gradient Boosting	0,93	0,88	0,90	0,975
XGBoost	0,92	0,86	0,88	0,973
kNN	0,88	0,82	0,84	0,928
LSTM	0,94	0,89	0,91	0,977

Источник: составлено авторами.

Результаты исследования и их обсуждение

В табл. 2 представлены результаты расчета основных метрик по всем моделям. Согласно им, наилучшая модель по пяти метрикам – LSTM, она хорошо справляется с задачами, где важно минимизировать пропуски. В качестве альтернативы высокую эффективность и стабильные результаты показывают Random Forest, Gradient Boosting, XGBoost. kNN показывает худшие результаты по всем метрикам.

На рис. 1–5 представлены графики кривой ROC-AUC, которые позволяют оценить качество срабатывания классификационной модели, где ось X отражает FPR (False Positive Rate) – ложно положительное срабатывание, а ось Y – TPR (True Positive Rate) – истинно положительное срабатывание.

На основе матриц ошибок и значений ROC-AUC для всех моделей можно провести сравнительный анализ ошибок классификации с акцентом на FP (ложно положительные) и FN (ложно отрицательные) ошибки.

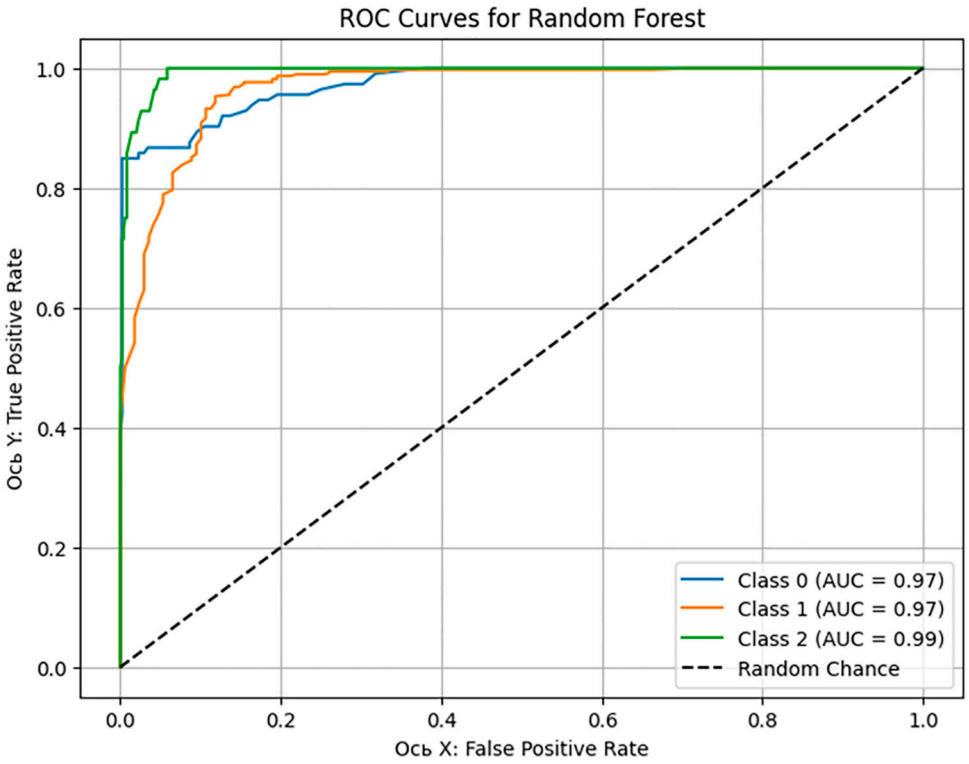

Рис. 1. ROC-AUC кривая модели Random Forest
Источник: составлено авторами

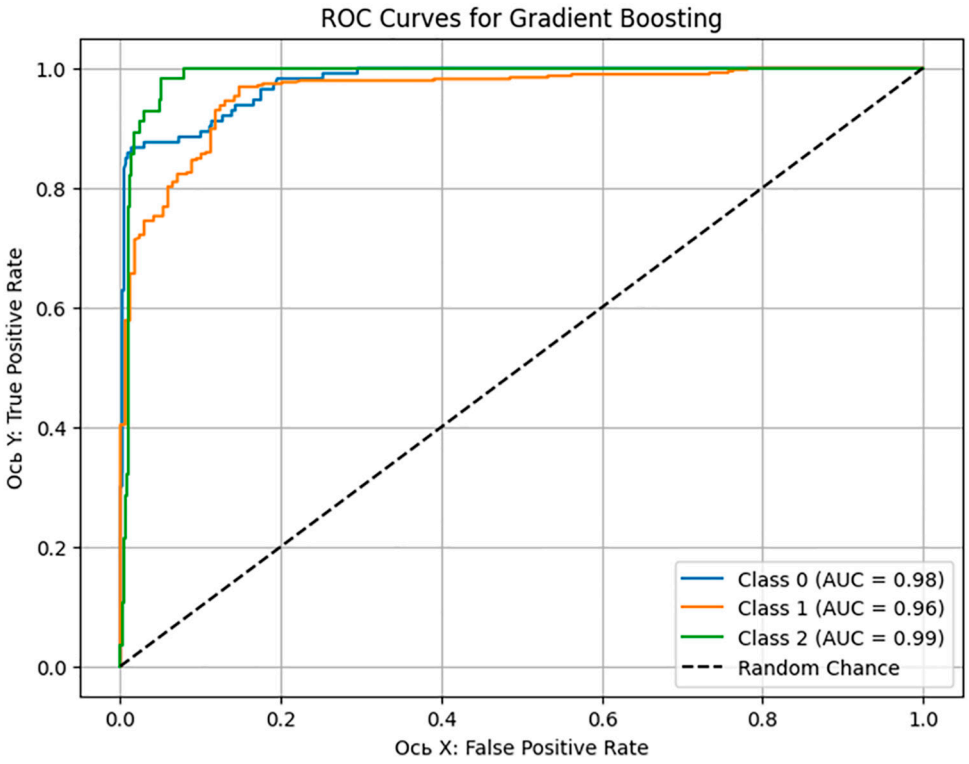

Рис. 2. ROC-AUC кривая модели Gradient Boosting
Источник: составлено авторами

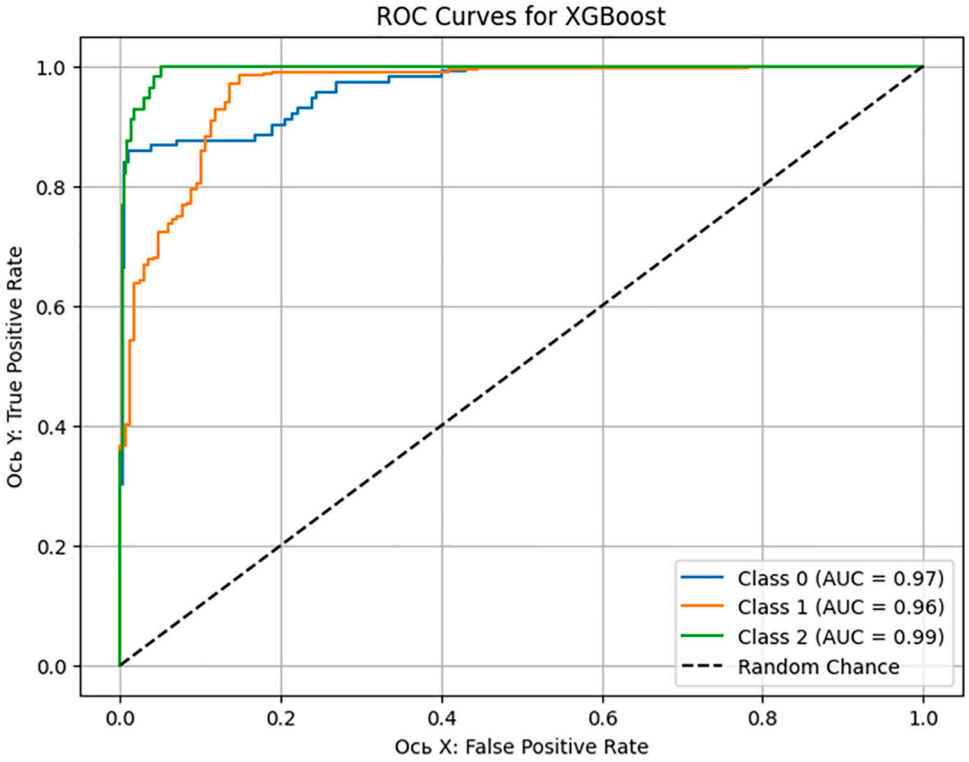

Рис. 3. ROC-AUC кривая модели XGBoost
Источник: составлено авторами

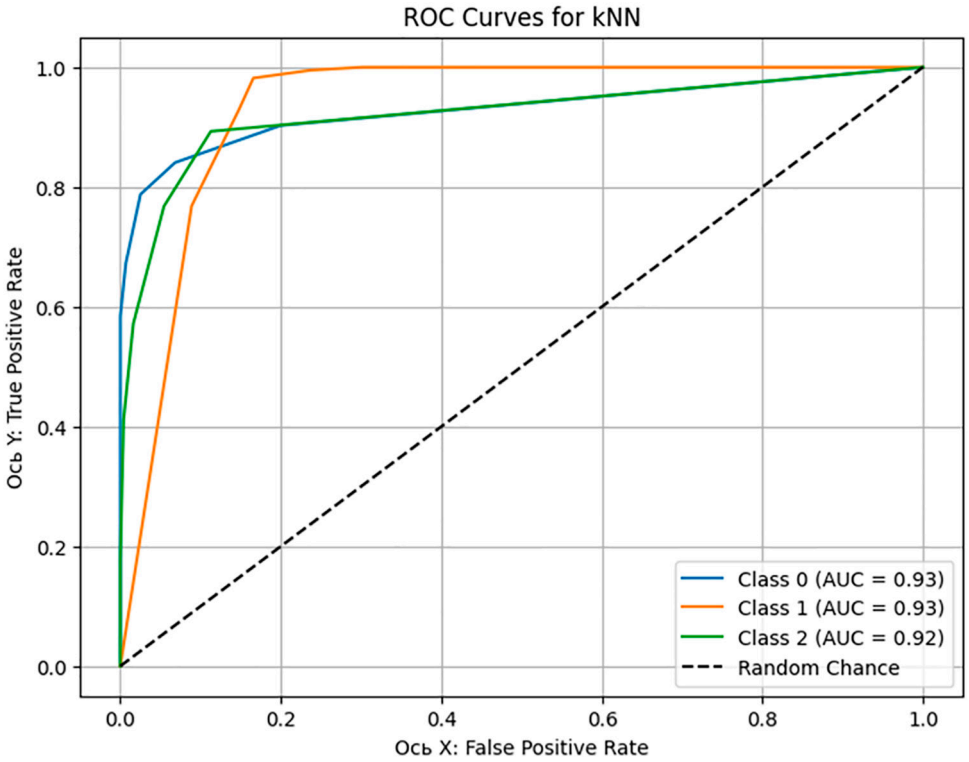

Рис. 4. ROC-AUC кривая модели kNN
Источник: составлено авторами

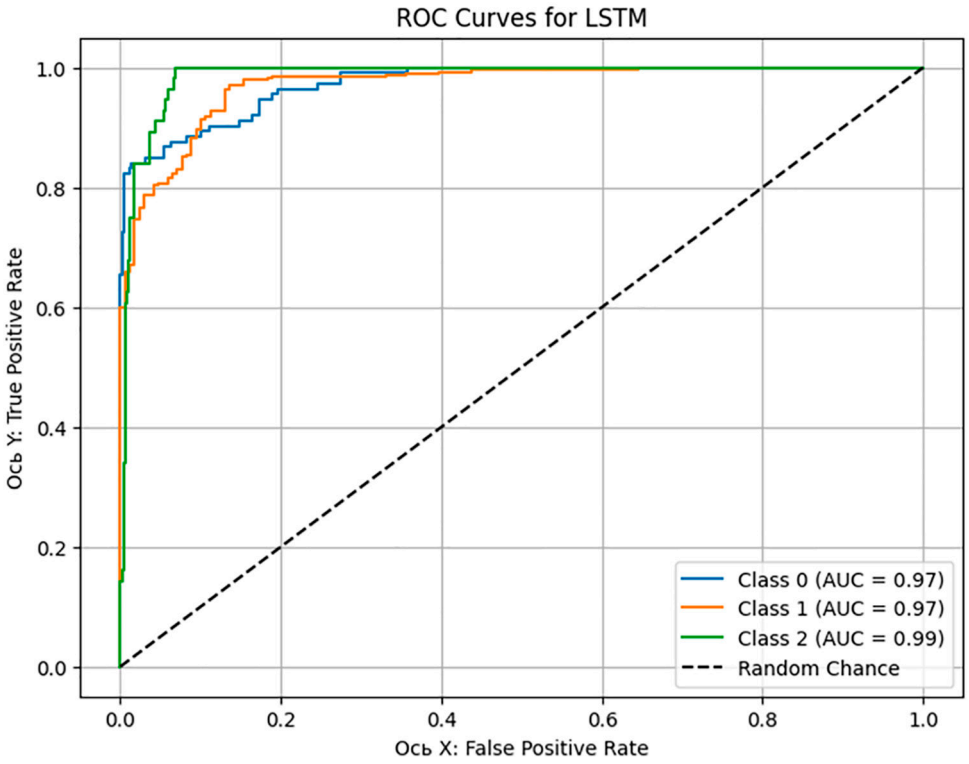

Рис. 5. ROC-AUC кривая модели LSTM
Источник: составлено авторами

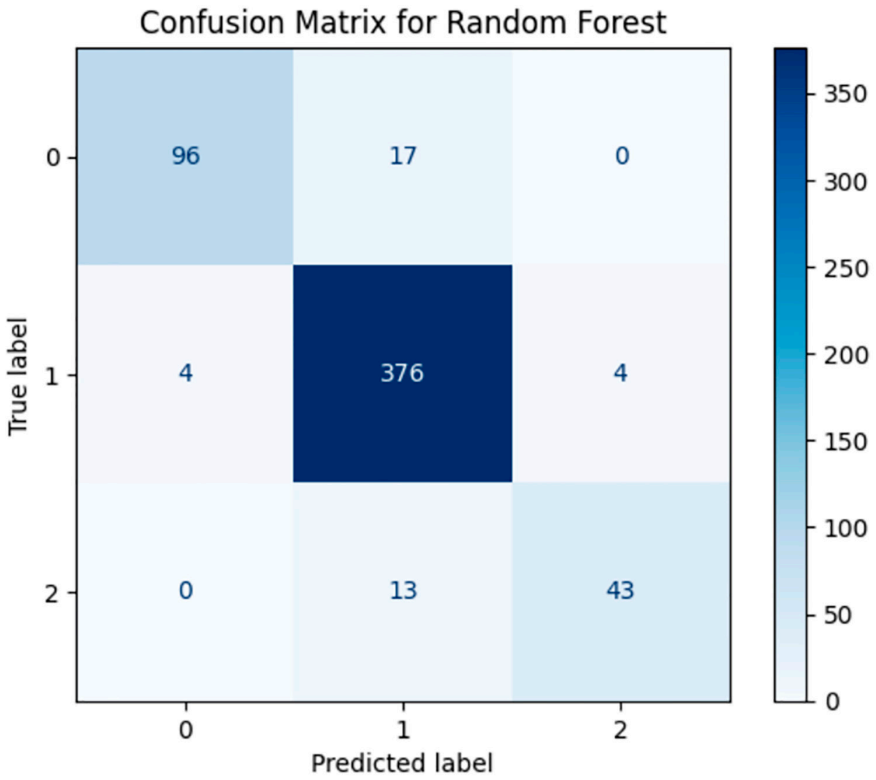

Рис. 6. Матрица ошибок модели Random Forest
Источник: составлено авторами

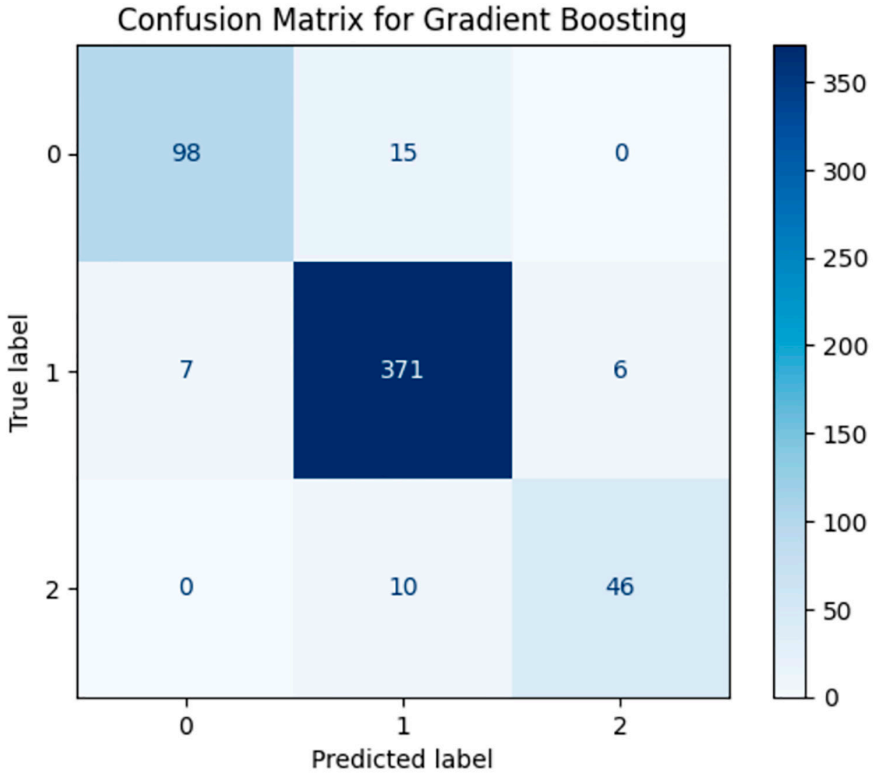

Рис. 7. Матрица ошибок модели Gradient Boosting
Источник: составлено авторами

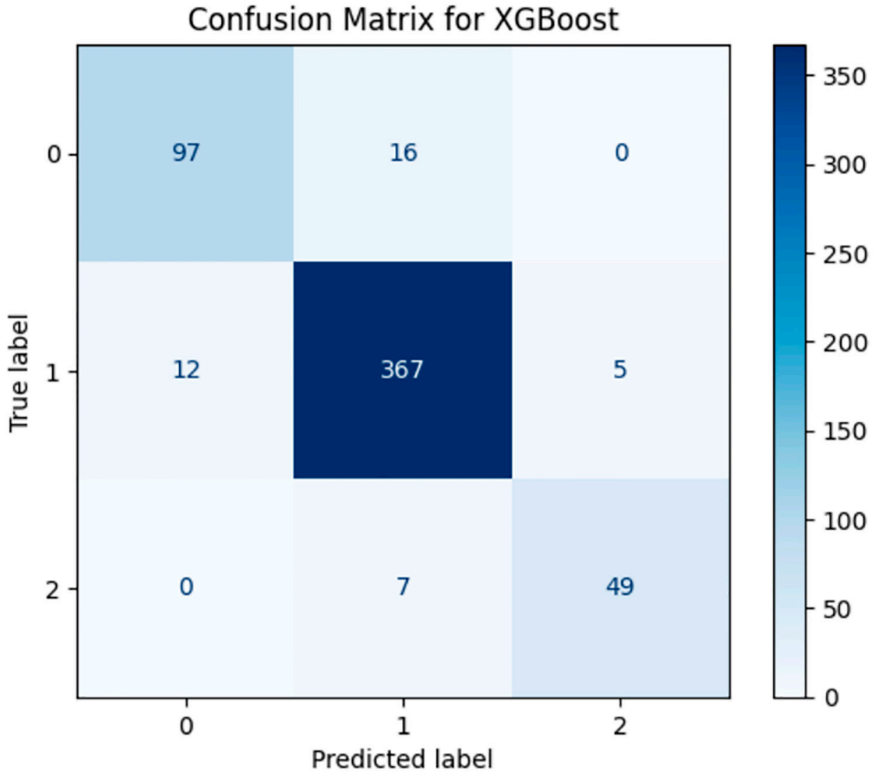

Рис. 8. Матрица ошибок модели XGBoost
Источник: составлено авторами

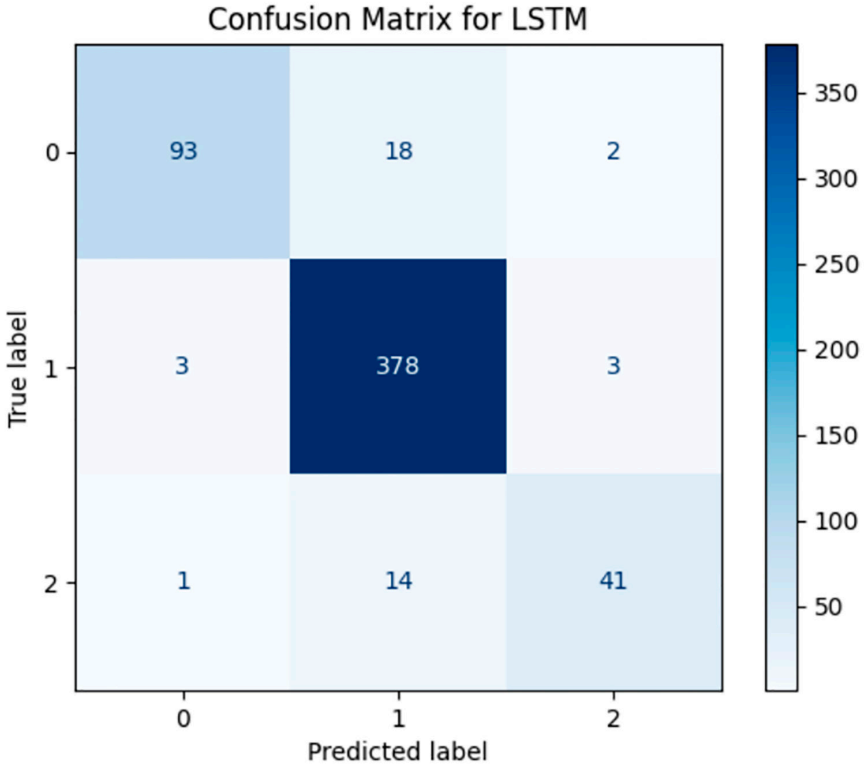

Рис. 9. Матрица ошибок модели kNN
Источник: составлено авторами

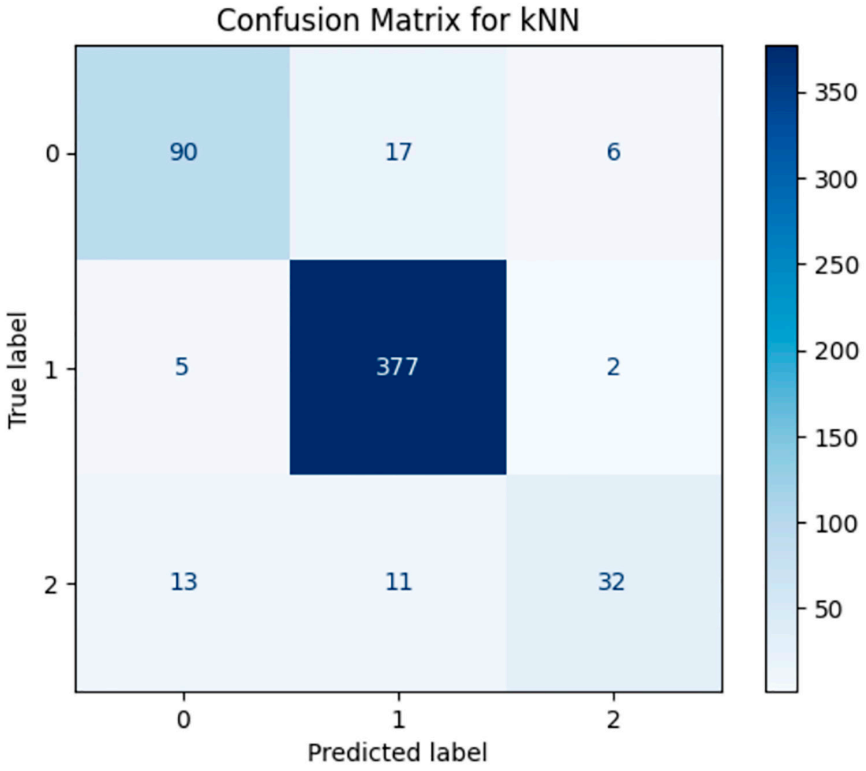

Рис. 10. Матрица ошибок модели LSTM
Источник: составлено авторами

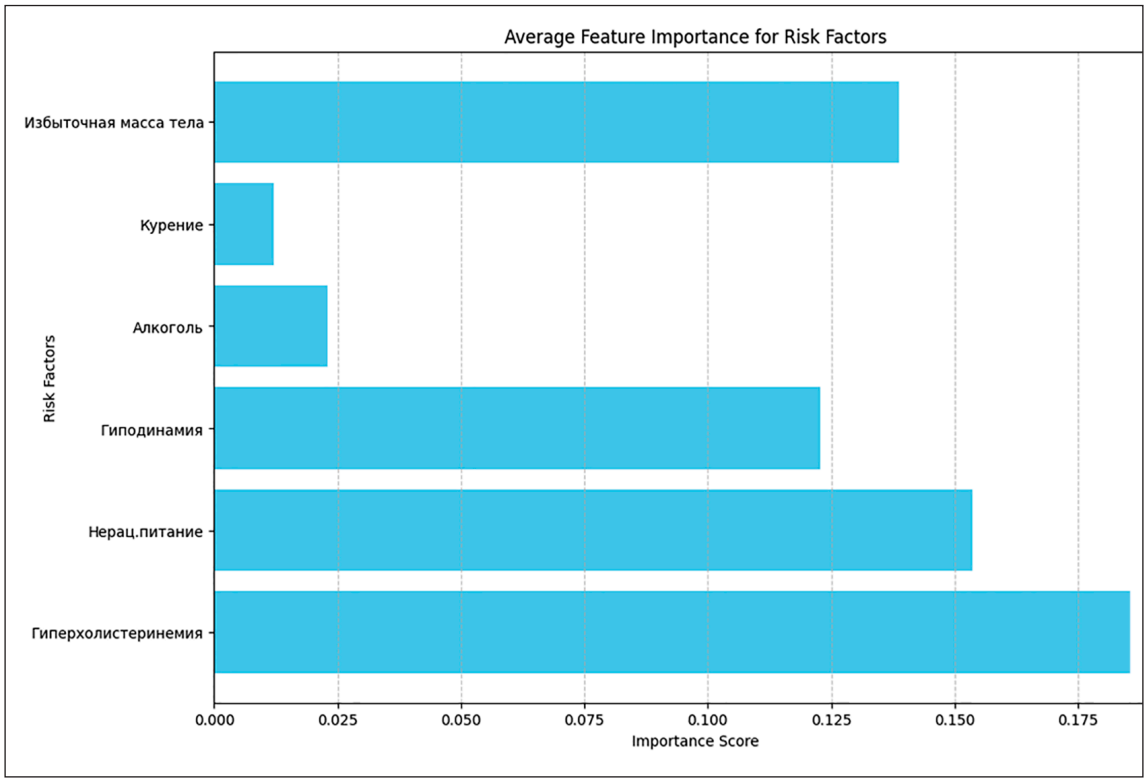

Рис. 11. Средняя важность факторов риска
Источник: составлено авторами

Для решения поставленной задачи, с использованием алгоритмов машинного обучения, анализ данных метрик является важным этапом, при этом высокое значение FN более критично, чем FP, так как это ведет к пропуску диагностики пациентов, требующих внимания (высокое значение FP приводит к лишним и часто ненужным медицинским вмешательствам). На рис. 6–10 приведены матрицы ошибок, рассчитанные на основе тестовой выборки (30% от исходного датасета).

Анализ моделей показал, что универсального идеального алгоритма не существует, выбор метода зависит от целей применения. Для диагностики больных пациентов (класс 1) Gradient Boosting и LSTM является хорошим выбором благодаря высокой точности и минимальным FN-ошибкам. С прогнозированием группы риска (класс 2) лучше прочих справились Gradient Boosting и XGBoost, что делает их применимыми в задачах профилактической медицины. Для общей классификации Random Forest остается неплохим вариантом, однако возможная путаница между классами 1 и 2 затрудняет четкое разделение больных и предрасположенных к заболеванию пациентов. Метод kNN показал низкую эффективность и не рекомендуется для решения данной задачи.

Для моделей можно оценить значимость факторов в классификации. Важность рассчитывается различными способами для моделей: через разбиения деревьев, частоту использования признаков или путем перестановочных оценок. На рис. 11 приведена усредненная гистограмма оценки важности факторов риска.

Наиболее значимыми факторами риска оказались гиперхолистеринемия, нерациональное питание и избыточная масса тела.

Заключение

Результаты проведенного исследования продемонстрировали хорошую эффективность применения алгоритмов машинного обучения для прогнозирования групп риска ХНИЗ. В условиях цифровизации медицины и накопления большого количества медицинских данных такие методы позволяют автоматизировать процесс предварительной диагностики и прогнозирования заболеваний, а также повышать качество и скорость принятия врачебных решений.

В ходе исследования были протестированы пять алгоритмов машинного обучения: Random Forest, Gradient Boosting, XGBoost,

kNN и LSTM. На основе анализа базовых метрик лидирующие позиции занимают следующие модели: Gradient Boosting и LSTM. Они демонстрируют достаточную точность для решения задач классификации. Random Forest, Gradient Boosting и XGBoost обеспечили хорошую интерпретируемость факторов риска. Метод kNN показал наихудшие результаты, особенно в отношении прогнозирования группы риска, что делает его наименее подходящим для данной задачи.

Таким образом, использование ансамблевых методов (Gradient Boosting, XGBoost) и нейросетевых подходов (LSTM) является наиболее оправданным для задач медицинского прогнозирования, а использование более объемных и разнообразных медицинских данных (например, данных лабораторной диагностики и дополнительных исследований – МРТ и КТ) позволит решать более сложные задачи скоринговой диагностики заболеваний. Внедрение технологий машинного обучения в сферу здравоохранения является перспективным направлением, способствующим повышению эффективности медицинской диагностики и улучшению качества жизни пациентов.

Список литературы

1. Жукова Т.В., Шатов А.Ю., Ушакова А.А., Зверева О.В., Бондаренко М.А. Методы машинного обучения компьютерных программ для целей профилактической медицины // Актуальные вопросы гигиены и диетологии на современном этапе: материалы 5-й Всероссийской научно-практической конференции (Ростов-на-Дону, 15 февраля 2024 г.). Ростов-на-Дону: Ростовский государственный медицинский университет, 2024. С. 117–121. EDN: CXEMMO.

2. Степанян И.В., Алимбаев Ч.А., Савкин М.О., Лю Д., Зидун М. Сравнительный анализ методов машинного обучения для прогнозирования болезней сердца // Проблемы машиностроения и автоматизации. 2022. № 2. С. 84–95. DOI: 10.52261/02346206_2022_2_84.

3. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45, Is. 1. P. 5–32. DOI: 10.1023/A:1010933404324.

4. Носова Г.С., Абдуллин А.Х. Машинное обучение на основе непараметрического и нелинейного алгоритма Random Forest (RF) // Инновации. Наука. Образование. 2021. № 35. С. 33–39. URL: https://innovjourn.ru/nomer/35-nomer/ (дата обращения: 01.04.2025). EDN: HXWMCZ.

5. Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine // Annals of Statistics. 2001. Vol. 29, Is. 5. P. 1189–1232. DOI: 10.1214/aos/1013203451.

6. Семашкин Н.М. Разработка алгоритма градиентного бустинга со случайными поворотами признакового пространства для решения задачи классификации // Молодежный вестник ИрГТУ. 2018. Т. 8. № 2. С. 17–22. URL: http://мвестник.рф/journals/2018/02/articles/03 (дата обращения: 01.04.2025).

7. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16). 2016. P. 785–794. DOI: 10.1145/2939672.2939785.

8. Ананенков В.А. Применение XGBoost для решения задач классификации // Инновации. Наука. Образование. 2020. № 24. С. 1663–1669. URL: https://innovjourn.ru/nomer/24-nomer (дата обращения: 01.04.2025).

9. Dudani S.A. The Distance-Weighted k-Nearest-Neighbor Rule. IEEE Transactions on Systems, Man, and Cybernetics. 1976. Vol. SMC-6, Is. 4. P. 325–327. DOI: 10.1109/TSMC.1976.5408784.

10. Родионов А.В., Ищенко К.Л. Исследование влияния параметров алгоритма k-ближайших соседей на метрики качества моделей // System Analysis and Mathematical Modeling. 2024. Т. 6. № 2. С. 251–262. DOI: 10.17150/2713-1734.2024.6(2).251-262.

11. Свекольникова Е.А., Пановский В.Н. Обзор open-source библиотек для решения задач прогнозирования временных рядов // Моделирование и анализ данных. 2024. Т. 14. № 2. С. 45–61. DOI: 10.17759/mda.2024140203.

12. Павленко Д.В., Татарис Ш.Э., Овчаренко В.В. Применение глубокого обучения в интерфейсах мозг – компьютер для распознавания движений // Программные продукты и системы. 2024. Т. 37. № 2. С. 164–169. DOI: 10.15827/0236-235X.142.164-169.

13. Bisong E. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Berkeley, CA: Apress, 2019. 619 p. DOI: 10.1007/978-1-4842-4470-8.

14. Bishop C.M. Pattern Recognition and Machine Learning. New York: Springer, 2006. 738 p. DOI: 10.1007/978-0-387-45528-0.

15. Powers D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation // Journal of Machine Learning Technologies. 2011. Vol. 2, Is. 1. P. 37–63. URL: https://arxiv.org/abs/2010.16061 (дата обращения: 01.04.2025). DOI: 10.9735/2229-3981.