УДК 618.146-006.6-07:616.15-076.3

ДИАГНОСТИКА РАКА ШЕЙКИ МАТКИ ПО РУТИННЫМ АНАЛИЗАМ КРОВИ: МАШИННОЕ ОБУЧЕНИЕ

А. И. Кузнецов[1], Е. В. Щепкина[2,3], Т. В. Сушинская[4], С. В. Епифанова[5,2], Н. И. Стуклов[7], Д. М. Фаур[6], А. Д. Каприн[4,8]

[1] ФГБУ ВО «Московский авиационный институт (национальный исследовательский университет)», Москва
[2] Российская академия народного хозяйства и государственной службы при Президенте РФ, Москва
[3] ГБУЗ «Научно-практический клинический центр диагностики и телемедицинских технологий» ДЗМ, Москва
[4] МНИОИ им. П. А. Герцена, филиал ФГБУ «НМИЦ радиологии» Минздрава России, Москва
[5] ФГБУ «Центральная клиническая больница с поликлиникой» Управления делами Президента Российской Федерации, Москва
[6] Российский экономический университет им. Г. В. Плеханова, Москва
[7] Медицинский институт ФГАОУ ВО «Российский университет дружбы народов», Москва
[8] ФГБУ «НМИЦ радиологии» Минздрава России, г. Обнинск

Обоснование. *Рак шейки матки (РШМ) продолжает оставаться серьезной проблемой здоровья, занимая четвертое место по распространенности и третье место по смертности среди женщин во всем мире. Риск РШМ увеличивается с возрастом, но основной контингент заболевших приходится на репродуктивный период. Эксперты прогнозируют рост заболеваемости и смертности от РШМ в будущем, особенно среди женщин старше 65 лет, а также в районах с ограниченным доступом к медицинским ресурсам. Это подчеркивает необходимость улучшения методов персонифицированной диагностики и прогнозирования риска развития РШМ.*

Искусственный интеллект и, в частности, машинное обучение (МО), предлагает многообещающий подход к разработке прогностических моделей на основе рутинных методов обследования, в том числе лабораторных показателей крови, что может помочь ранней диагностике РШМ.

Цель. *Разработка системы поддержки принятия врачебных решений — СППВР, для выявления пациенток с высоким риском развития РШМ.*

Методы. *Проведено одноцентровое когортное ретроспективное исследование женщин старше 18 лет на базе МНИОИ им. П. А. Герцена в 2000–2024 гг. В исследовании участвовали 452 женщины в возрасте 42,0 (медиана 33,75; 50,0) лет. Для построения прогностической модели предсказания наличия РШМ были использованы следующие алгоритмы машинного обучения: MLR — Lasso, MLR — Ridge, Extra Tree (ET), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Gradient Boosting (XGB), Catboost (CB) и LightGBM (LGMB).*

Результаты. *Лучшие результаты получены при построении прогностической модели на основе алгоритма МО XGB, который на тестовой выборке показал следующие рузультаты: ROC-AUC = 98,2 % (95 % ДИ 96,0; 99,7), точность = 94,5 % (95 % ДИ 91,0; 97,2), чувствительность = 95,0 % (95 % ДИ 91,2; 98,3), специфичность = 91,7 % (95 % ДИ 81,0; 100,0). В качестве основных предикторов использвались: агрегация тромбоцитов с АДФ, СОЭ, Д-димер, трансферин, возраст и ряд других.*

Заключение. *Это исследование представило новый способ выявления РШМ на основе рутинных лабораторных гематологических показателей. Построенная модель может быть использована в системе поддержки принятия врачебных решений (СППВР) для своевременного выявления женщин с подозрением на РШМ.*

Ключевые слова: *Рак шейки матки, диагностика, машинное обучение, рутинные гематологические показатели, прогностическая модель*

CERVICAL CANCER SCREENING USING ROUTINE LABORATORY BLOOD TESTS AS RISK FACTORS: MACHINE LEARNING

A. I. Kuznetsov[1], E. V. Schepkina[2,3], T. V. Sushinskaya[4], S. V. Epifanova[5,2], N. I. Stuklov[7], D. M. Faur[6], A. D. Kaprin[8]

[1] Moscow Aviation Institute (National Research University), Moscow, Russia
[2] Russian Presidential Academy of National Economy and Public Administration - RANEPA, Moscow, Russia
[3] Research and Practical Clinical Center for Diagnostic and Telemedical Technologies, Moscow, Russia
[4] P. Hertsen Moscow Oncology Research Institute — Branch of the National Medical Research Radiological Centre
[5] Central Clinical Hospital, Office of the President of the Russian Federation, Moscow, Russia
[6] Plekhanov Russian Economic University, Moscow, Russia
[7] Medical Institute of the Russian University of Peoples' Friendship, Moscow, Russia
[8] National Medical Research Radiological Centre, Obninsk, Russia

Background. *Cervical cancer (CC) continues to be a major health problem, ranking as the fourth most common cancer among women worldwide and the third leading cause of death. The risk of CC increases with age, but the majority of cases occur during the reproductive period. Experts predict an increase in incidence and mortality from CC in the future, especially among women over 65 years of age, and in areas with limited access to medical care. This emphasizes the need to improve methods for personalized diagnosis and prediction of the risk for developing cervical cancer.*
Artificial intelligence, and in particular machine learning (ML), offers a promising approach for developing predictive models based on routine examination methods of laboratory blood parameters, which can help in the early diagnosis of cervical cancer.

Aim. *Development of a medical decision support system – MDSS, to identify patients at high risk of developing cervical cancer.*

Methods. *A single-center cohort retrospective study of women over 18 years of age was conducted on data collected in 2000-2024 at the Moscow Research Institute named after P. A. Herzen. The study included 452 women aged 42.0 [median 33.75; 50.0] years. The following machine learning algorithms were used to build a model for predicting the presence of CC: MLR – Lasso, MLR – Ridge, Extra Tree (ET), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Gradient Boosting (XGB), Catboost (CB) and LightGBM (LGMB).*

Results. *The best predictive results were obtained when building a model based on the XGB MO algorithm, which on the test sample showed the following results: ROC-AUC = 98.2 % [95 % CI 96.0; 99.7], accuracy = 94.5 % [95 % CI 91.0; 97.2], sensitivity = 95.0 % [95 % CI 91.2; 98.3], specificity = 91.7 % [95 % CI 81.0; 100.0]. The major predictive factors were platelet aggregation with ADP, ESR, D-dimer, transferrin and age.*

Conclusions. *This study introduces a new method for predicting the risk of CC based on routine laboratory hematological parameters. The constructed model can be used in a medical decision support system (MDSS) for the timely identification of women with high risk of cervical cancer*

Keywords: *cervical cancer, diagnosis, machine learning, routine hematological parameters, prognostic model*

Обоснование

В 2022 году рак шейки матки (РШМ) занял четвертое место в мире по распространенности и третье место по смертности от злокачественных новообразований среди женщин, при этом стал причиной 662 301 новых случаев заболевания и 348 874 случаев смерти [1, 2].

В России в 2022 году распространенность РШМ составила 18 369, умерли в том же году 7 903 женщины [3].

Риск РШМ связан с инфицированием вирусом папилломы человека (ВПЧ), но может повышаться с возрастом, курением [4, 5], ранним началом половой жизни и наличием нескольких сексуальных партнеров, длительным использованием оральных контрацептивов [4, 6].

Заболевание часто развивается у женщин с хроническими и фоновыми заболеваниями шейки матки и на ранних стадиях проявляет себя неспецифическими симптомами или вообще протекает бессимптомно, что усложняет задачу своевременного выявления РШМ.

Симптомы могут появиться при развитии осложнений, связанных с большим местным распространением опухоли, когда процесс лечения становится более сложным, длительным

и дорогостоящим. Поэтому остро стоит задача ранней диагностики РШМ.

В настоящее время диагностика РШМ представлена цитологическим исследованием мазков с шейки матки и из цевикального канала. Метод простой, доступный, широко применяемый в скрининге и диагностике преинвазивного РШМ. Точность метода, по данным разных авторов, составляет от 50 до 97,5 % [7, 8].

Ошибки цитологической диагностики могут быть связаны с множеством причин: с качеством предоставленного для исследования биологического материала, с типом применяемого клиницистами инструмента для получения материала на цитологическое исследование, с особенностями нанесения биологического материала на предметные стекла [8].

В последние годы метод традиционной цитологии в скрининге и диагностике РШМ дополнен или заменен методом жидкостной цитологии, точность которого аналогична традиционному [9].

Оба метода требуют наличия цитологической лаборатории и квалифицированного цитолога. Однако, при наличии всех условий, одного цитологического исследования для верификации диагноза бывает недостаточно, поэтому в гинекологической практике широко распространился метод кольпоскопически направленной биопсии с последующим гистологическим исследованием. При проведении ограниченной биопсии шейки матки трудности возникают при выявлении наиболее информативного для гистологического исследования патологически-измененного участка, а также с мультифокальностью поражений эпителия. Таким образом, в этом случае, результативность диагностики РШМ зависит от выбора оптимальной локализации. Однако ограниченная биопсия дает большой процент расхождения результатов гистологического исследования биоптатов и последующей эксцизионной биопсии (конизации) по данным разных авторов от 27,9 % до 40 %, при этом гиподиагностика превалировала над гипердиагностикой (0,92–50,0 % и 5,3–21,7 % соответственно), что делает, по нашему мнению, данную верификационную методику практически непригодной для использования [7, 10–13].

При условии большого процента ложноотрицательных результатов становится понятным, что у большого процента даже регулярно обследуемых пациенток РШМ с помощью традиционных методов исследования не будет диагностирован своевременно.

Машинное обучение (МО) — это область искусственного интеллекта (ИИ), которая используя ретроспективные данные для построения прогностических моделей, которые прогнозируют будущие события [14].

Использование МО значительно продвинуло скрининг, диагностику, лечение и прогнозирование многих заболеваний, в том числе и злокачественных [15, 16].

Проведенные исследования показывают, что подходы, основанные на МО, могут обеспечить, например, скрининг рака молочной железы посредством высокоэффективного прогнозирования риска его выявления [17, 18].

Существующие алгоритмы машинного обучения демонстрируют хорошую прогностическую способность в различных областях медицины [19–24], в том числе и генетике [25–28].

Важным для медицинских исследований является то, что МО потенциально может работать со структурированными базами данных небольшого и среднего объема [29, 30].

В 2020 году нами была построена диагностическая модель прогнозирования метастазов РШМ с использованием алгоритмов машинного обучения на основе рутинных гематологических показателей [31]. Аналогичный результат представлен китайскими исследователями в 2023 году [32].

Оценка показателей периферической крови давно принята за основу определения состояния здоровья человека. Однако нами не найдено исследований по построению прогностических моделей для ранней диагностики РШМ с использованием алгоритмов МО на основе рутинных гематологических показателей, входящих в стандартное и обязательное обследование при любом заболевании, в том числе и злокачественном. Нам показалось логичным построить математическую модель, основанную на гематологических предикторах с целью своевременной и точной диагностики РШМ, особенно у пациенток с хроничекими фоновыми

и предраковыми заболеваниями шейки матки. Для достижения поставленной цели были собраны и подготовлены данные для математического анализа. В процессе подготовки данных мы использовали три стратегии: устранение избыточности данных, заполнение пропущенных значений и выбор значимых факторов для прогнозирования. Затем использовали алгоритмы МО на основе ранее обработанных данных для построения модели прогнозирования риска развития РШМ.

Цель

Разработка системы поддержки принятия врачебных решений — СППВР для выявления пациенток с высоким риском наличия РШМ.

Материалы и методы

Дизайн исследования

Одноцентровое когорное ретроспективное исследование.

Критерии соответствия

Критерии включения:

1. возраст старше 18 лет;
2. пациенки с фоновыми, предраковыми заболеваниями и раком шейки матки.

Критерии невключения:

1. синхронный опухолевый процесс;
2. отказ от продолжения обследования по предложенному алгоритму.

Условия проведения и продолжительность исследования

МНИОИ им. П. А. Герцена — филиал ФГБУ НМИЦ радиологии Министерства здравоохранения России. Участники исследования являлись жителями как Москвы, так и всех других регионов России и ближнего зарубежья. В исследование были включены пациентки, обратившиеся для обследования в период с 2000 по 2024 гг.

Основной исход исследования: выявление РШМ.

Методы регистрации исходов

В качестве метода регистрации исходов была выбрано гистологическое исследование опухоли, полученной в ходе операции или отдельной процедуры в рамках диагностических мероприятий (биопсии).

Анализ в подгруппах

В ходе исследования использовались две выборки: выборка для разработки прогностической модели и выборка для тестирования модели. Также в ходе исследования пациентки обеих выборок были разделены на 2 группы: с подтвержденным РШМ (злокачественное образование (ЗНО)) и с неподтвержденным РШМ (доброкачественное образование (ДНО)).

Статистический анализ

Размер выборки: Минимальный объем выборки при уровне значимости 5 % для сохранения статистической мощности в 80 % составляет 385 участников. Выборка в 452 пациенток является достаточной для того, чтобы выявить статистически значимые различия в оценке РШМ.

Статистическая обработка результатов проводилась средствами языка Питон (Python 3.11.).

Построение прогностической модели было выполнено в 10 этапов.

На первом этапе была проведена предобработка данных для повышения качества имеющихся данных: проверка на выбросы и пропуски значений. Для данных, которые имели менее, чем 5 % пропущенных значений, был применен метод вменения пропущенных значений, основанный на алгоритме машинного обучения k-Nearest Neighbors (kNN).

На втором этапе было проведено сравнение групп с наличием РШМ и без РШМ (исход) с использованием критерия Манна-Уитни. Также на этом этапе была проведена однофакторная логистическая регрессия для выявления факторов, оказывающих статистически значимое влияние на исход. В качестве количественной меры эффекта влияния нами использовался показатель отношения шансов (ОШ), определяемый как e^{β}, где e = 2,72 (число Эйлера).

На третьем этапе исходная выборка была разделена на обучающую и тестовую в соотношении 70/30, а также проведено сравнение выделенных групп на статистически значимое различие.

На четвертом этапе обучающая выборка была проверена на сбалансированность в зависимости от количества пациентов с исходом — 1 — есть РШМ (ЗНО) и 0 — нет РШМ (ДНО).

Для балансировки обучающей выборки в целях получения более стабильной модели прогнозирования, был использован метод SMOTE (Synthetic Minority Over-sampling Technique) [33, 34].

На пятом этапе был применен корреляционный анализ, при котором был рассчитан коэффициент ранговой корреляции Спирмена (r). Если пара переменных значительно (r>0,7) коррелировала между собой, то для дальнейшего анализа оставлялась одна переменная (наиболее значимая), вторая переменная из дальнейшего анализа исключалась.

На шестом этапе для отбора наиболее значимых переменных был применен метод RFE (Recursive Feature Elimination — рекурсивное исключение признаков) [35].

На седьмом этапе для построения модели предсказания наличия/отсутствия исхода были выбраны следующие алгоритмы машинного обучения: MLR — Lasso (MLR_L), MLR — Ridge (MLR_R), Extra Tree (ET), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), Gradient Boosting (XGB), Catboost (CB) и LightGBM (LGMB) [36].

На восьмом этапе была произведена оценка построенных моделей: они были протестированы с использованием независимого тестового набора. Для оценки качества построенных моделей на обучающей и тестовой выборках были использованы следующие метрики: точность, чувствительность, специфичность и ROC-AUC (ROC — Receiver Operating Characteristic — рабочая характеристика приёмника; AUC — Area Under the Curve — площадь под кривой). Метрики были рассчитаны с 95 % доверительными интервалами (ДИ) [37]. 95 % доверительный интервал рассчитывался методом Бутстреп (Bootstrap) при выборке 1000 экземпляров [38].

Также для максимизации чувствительности и специфичности полученной модели было рассчитано пороговое значение для определения наличия/отсутствия изучаемого исхода. Для оценки клинической полезности построенных моделей, был проведен анализ кривой принятия решений (DCA — Decision curve analysis) путем расчета чистых преимуществ модели прогнозирования по сравнению с решением «лечить всех» и «не лечить никого» [40].

На девятом этапе все построенные модели были сравнены между собой по метрике ROC-AUC с помощью критерия ДеЛонга (DeLong) [41].

На десятом этапе, для анализа «закрытых» моделей был применен метод SHAP, который основан на оценке важности признаков с использованием средних аддитивных объяснений Шепли (SHAP). Этот метод выдает значения SHAP, которые позволяют оценить степень влияния различных факторов риска на исход [42].

Таким образом, был выбран наиболее эффективный алгоритм машинного обучения, который строит наиболее точную прогностическую модель для достижения цели текущего исследования.

Результаты

Предварительная обработка данных

На этапе разработки исходная база данных состояла из 466 пациенток. При проверке базы на предмет пропущенных значений, было выявлено, что у 14 пациенток, включая 6 и 8 пациенток, относящихся к положительным и отрицательным случаям наличия РШМ, имеется более 5 % пропущенных значений. Они были исключены из исследования. Далее, с помощью алгоритма KNN были рассчитаны значения у 37 пациенток, которые имели менее 5 % пропущенных значений. Применение методов вменения пропущенных значений, использующие алгоритмы машинного обучения, позволяет уменьшить погрешность, чем при использовании других методов, например, при использовании значений, имеющих самую высокую частоту или среднее, и т. д.; следовательно, эффективность прогностической модели в отношении обобщаемости будет в значительной степени сохранена. Итого, в данном исследовании, как показано на рис. 1, осталось 452 пациентки, в том числе 381 с РШМ (ЗНО) и 71 без РШМ (ДНО).

Выбор признаков

Изначально было рассмотрено 33 признака, описывающих состояние пациенток: возраст и показатели общего и биохимического анализов

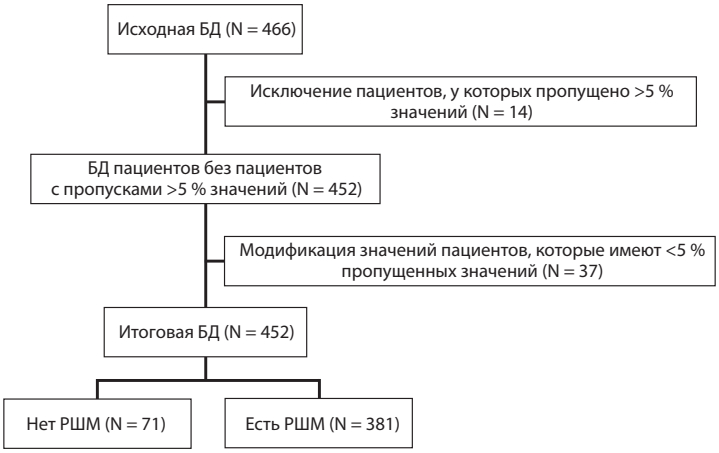

Рис. 1. *Этапы предварительной обработки БД на этапе разработки прогностической модели*

крови и коагулограммы, включающие следующие показатели: Показатели крови, использованные в исследовании: скорость оседания эритроцитов (СОЭ) (норма до 30 мм \час), гемоглобин (HGB) (норма 120–165 г\л), количество эритроцитов (RBC) (норма 4,0–5,5 10^{12}\л), гематокрит (HCT) (норма 36,0–48,0 %), средний объем эритроцита (MCV) (норма 80–100 фл), среднее содержание гемоглобина в эритроците (MCH) (норма 26,0–34,0 пг), количество тромбоцитов (PLT) (норма 180–350 10^{9}\л), количество лейкоцитов (WBC) (норма 4,0–9,0 10^{9}/л), активированное частичное тромбиновое время (АЧТВ) (норма 26,0–37,0 сек), протромбиновое время (ПТВ), тромбиновое время (ТВ) (норма 14,6–22,0 сек), фибриноген (норма 2,0–4,0 г\л), растворимые фибрин-мономерные комплексы (РФМК) (норма 0–4 мг/дл), Д-димер (норма 0–0,55 мг\л ФЭЕ), агрегация тромбоцитов с АДФ (норма 80–120 %), активность антитромбина III (норма 80–120 %), МНО (норма 0,90–1,30 отн. ед.), плазминоген (норма 75,0–140,0 %), Хагеманзависимый фибринолиз (ХЗФ) (норма 4,0–12,0 мин), общий белок (ОБ) (норма 64–83г\л), мочевина (норма 2,5–8,3 ммоль/л), креатинин (норма 53–97 ммоль/л), аспартатаминотрансфераза (АСТ) (норма до 41 ед/л), аланинаминотрансфераза (АЛТ) (норма до 40 ед/л), гамма-глутамилтранспептидаза (ГГТ) (норма до 30 ед/л), щелочная фосфатаза (ЩФ) (норма до 270 ед\л), С-реактивный белок (СРБ) (норма 0–5,0 мг\л), лактатдегидрокиназа (ЛДГ) (норма 0 — 480 ед\л), сывороточное железо (СЖ) (норма 10,7–32,2 мкмоль\л), трансферрин (ТФ) (норма 2–3,6 г\л), ферритин (норма 10–120 мкг\мл), насыщение трансферрина железом (НТЖ) (норма 15–50 %).

Характеристики пациенток представлены в таблице 1. При сравнении групп ДНО и ЗНО было выявлено, что статистически значимо ($p < 0,001$) группы различаются по следующим переменным: возраст, СОЭ, RBC, HGB, HCT, Д-димер, агрегация тромбоцитов с АДФ, МНО, РФМК, ЩФ, СРБ, ЛДГ и статистически значимо ($p < 0,05$) по АЧТВ, фибриноген, ОБ, АСТ, ТФ, НТЖ, креатинин. Именно эти переменные учитывались как факторы риска развития РШМ.

Согласно проведенному в дальнейшем анализу (однофакторной логистической регрессии) было установлено, что статистически значимо ($p<0,001$) группы ДНО и ЗНО различаются по переменным: возраст, СОЭ, RBC, HGB, Д-димер, агрегация тромбоцитов с АДФ, МНО, РФМК, АСТ, ЩФ, СРБ, ЛДГ, ТФ НТЖ и статистически значимо ($p < 0,05$) по переменным HCT, АЧТВ, фибриноген, ОБ.

Из таблицы 1 следует, что важными факторами рисками ЗНО являются: возраст СОЭ RBC, HGB, HCT, АЧТВ, фибриноген, Д-димер, агрегация тромбоцитов с АДФ, МНО, РФМК, ОБ, АСТ, ЩФ, СРБ, НТЖ.

Оставшиеся факторы риска ЗНО, включая WBC, MCV, MCH, PLT, ПТВ, ТВ, активность антитромбина III, плазминоген, ХЗФ, АЛТ, ГГТ, СЖ, ферритин, мочевина и креатинин, не имели статистической значимой разницы ($p > 0,05$), поэтому они были исключены из дальнейшего анализа.

Таблица 1

Характеристики пациенток: сравнение в разрезе групп и однофакторная логистическая регрессия

Факторы риска	Сравнение групп				Однофакторная логистическая регрессия			
	Все (n=452)	ДНО (n=71)	ЗНО (РШМ) (n=381)	p	β	ОШ [95 % ДИ]	p	R^2
Возраст (лет)	42,0 [33,75; 50,0],	34,0 [29,5; 39,0],	43,0 [35,0; 51,0],	<0,001*	0,08	1,083 [1,052, 1,116]	p<0,001*	0,092
СОЭ (мм \час	11,0 6,0; 21,0]	4,0 [2,0; 7,5]	13,04 [8,0; 22,65]	<0,001*	0,19	1,209 [1,14, 1,283]	p<0,001*	0,206
WBC, 10^9/л	6,4 [5,21; 7,76]	6,4 [5,35; 7,28]	6,4 [5,2; 7,8]	=0,963	0,04	1,041 [0,922, 1,176]	p=0,514	0,001
RBC 10^{12}\л	4,33 [4,1; 4,6]	4,6 [4,32; 4,83]	4,3 [4,05; 4,57]	<0,001*	-1,327	0,265 [0,149, 0,474]	p<0,001*	0,057
HGB, г/л	130,11 [118,0; 138,0]	136,0 [127,0; 140,0]	129,0 [115,0; 137,0]	<0,001*	-0,037	0,964 [0,945, 0,982]	p<0,001*	0,045
HCT, %	38,45 [35,8; 40,5]	40,6 [38,95; 42,25]	38,0 [35,19; 40,0]	p<0,001*	-0,054	0,947 [0,909, 0,988]	p=0,012*	0,020
MCV, фл	88,1 [83,6; 91,0]	87,9 [84,95; 91,0]	88,21 [83,2; 90,9]	p=0,570	-0,02	0,98 [0,946, 1,015]	p=0,265	0,004
MCH, пг	29,8 [28,29; 31,4]	29,2 [28,3; 30,8]	29,9 [28,2; 31,5]	p=0,220	0,025	1,025 [0,958, 1,098]	p=0,472	0,004
PLT, 10^9\л	255,0 [214,75; 306,25]	244,0 [209,0; 282,0]	257,0 [215,85; 310,0]	p=0,145	0,003	1,003 [0,999, 1,007]	p=0,132	0,006
АЧТВ, сек	30,7 [28,7; 33,37]	29,8 [27,65; 32,35]	30,8 [28,8; 33,57]	p=0,042*	0,077	1,08 [1,011, 1,154]	p=0,022*	0,015
ПТВ, сек	11,83 [11,3; 12,48]	11,9 [11,2; 12,95]	11,81 [11,32; 12,4]	p=0,716	-0,015	0,985 [0,955, 1,016]	p=0,348	0,002
ТВ, сек	18,0 16,7; 19,06]	18,5 [17,3; 18,98]	17,92 [16,6; 19,06]	p=0,198	-0,009	0,991 [0,967, 1,014]	p=0,437	0,001
Фибриноген, г/л	2,79 [2,4; 3,32]	2,6 [2,38; 2,9]	2,84 [2,41; 3,38]	p=0,002*	0,67	1,954 [1,289, 2,965]	p=0,002*	0,031
Д-димер, мг\л ФЭЕ	0,33 [0,21; 0,51]	0,15 [0,09; 0,24]	0,37 [0,25; 0,53]	p<0,001*	7,179	1311,596 [147, 11659]	p<0,001*	0,193
Агрегация тромбоцитов с АДФ, %	78,0 [68,38; 89,02]	94,49 [89,12; 100,55]	75,57 [65,9; 83,0]	p<0,001*	-0,108	0,898 [0,875, 0,92]	p<0,001*	0,283
Активность антитромбина III, %	105,66 [100,0; 108,96]	104,35 [100,09; 110,54]	105,7 [100,0; 108,81]	p=0,694	0,008	1,008 [0,982, 1,035]	p=0,560	0,001
МНО	1,04 [1,0; 1,06]	0,99 [0,96; 1,03]	1,04 [1,01; 1,07]	p<0,001*	14,197	1464464 [12553, 170753093]	p<0,001*	0,106
Плазминоген, %	95,6 [71,73; 121,09]	102,2 [79,51; 109,26]	94,12 [71,73; 122,25]	p=0,783	0,002	1,002 [0,994, 1,011]	p=0,647	0,001
РФМК мг/мкл	7,96 [4,99; 12,0]	4,34 [3,5; 7,0]	8,28 [5,5; 12,0]	p<0,001*	0,14	1,15 [1,075, 1,231]	p<0,001*	0,054
ХЗФ, мин	7,85 [6,76; 10,0]	8,63 [7,0; 9,13]	7,71 [6,73; 10,07]	p=0,978	0,053	1,054 [0,952, 1,168]	p=0,306	0,003
ОБ, г\л	73,1 [69,31; 76,3]	73,78 [71,5; 76,45]	72,79 [69,0; 76,3]	p=0,036*	-0,045	0,956 [0,917, 0,997]	p=0,034*	0,017
АЛТ, ед/л	17,98 [14,0; 22,29]	19,07 [14,01; 22,9]	17,7 [14,0; 22,0]	p=0,232	0,002	1,002 [0,984, 1,02]	p=0,818	0,000
АСТ, ед/л	17,56 [14,0; 21,35]	19,4 [15,87; 21,6]	17,0 [14,0; 21,0]	p=0,031*	-0,033	0,968 [0,949, 0,986]	p=0,001*	0,029
ГГТ, ед/л	20,77 [14,0; 34,96]	20,94 [13,77; 24,88]	20,6 [14,0; 38,87]	p=0,064	0	1,0 [0,998, 1,002]	p=0,696	0,000
ЩФ, ед\л	132,21 [77,0; 166,12]	50,5 [48,41; 64,33]	142,0 [110,0; 172,62]	p<0,001*	0,062	1,064 [1,048, 1,08]	p<0,001*	0,518

Окончание таблицы 1

Факторы риска	Сравнение групп				Однофакторная логистическая регрессия			
	Все (n=452)	ДНО (n=71)	ЗНО (РШМ) (n=381)	p	β	ОШ [95 % ДИ]	p	R^2
СРБ, мг\л	4,53 [1,0; 9,44]	0,71 [0,42; 2,8]	4,85 [1,3; 11,5]	p<0,001*	0,17	1,185 [1,098, 1,281]	p<0,001*	0,101
ЛДГ, ед\л	242,99 [185,13; 299,28]	211,58 [123,48; 279,41]	254,18 [192,16; 312,99]	p<0,001*	0,009	1,009 [1,006, 1,013]	p<0,001*	0,076
СЖ, мкмоль\л	13,96 [7,72; 17,79]	14,8 [11,59; 18,52]	13,6 [7,32; 17,78]	p=0,092	0,005	1,005 [0,991, 1,019]	p=0,500	0,002
ТФ, г\л	2,82 [2,45; 3,03]	2,9 [2,63; 3,05]	2,79 [2,38; 3,03]	p=0,005*	-0,052	0,949 [0,922, 0,977]	p<0,001*	0,038
Ферритин, мкг\мл	48,44 [19,08; 69,89]	32,92 [23,16; 62,7]	53,19 [17,0; 70,27]	p=0,070	0,005	1,005 [0,999, 1,01]	p=0,081	0,021
НТЖ, %	19,05 [9,83; 27,38]	11,86 [9,69; 24,3]	20,4 [9,89; 28,53]	p=0,004*	0,04	1,041 [1,016, 1,066]	p=0,001*	0,033
Мочевина, ммоль/л	4,42 [3,83; 5,22]	4,5 [3,8; 5,07]	4,41 [3,89; 5,29]	p=0,422	0,013	1,013 [0,973, 1,054]	p=0,540	0,001
Креатинин, ммоль/л	74,0 [67,0; 81,85]	71,12 [66,31; 76,31]	75,0 [67,0; 82,91]	p=0,005*	0,002	1,002 [0,991, 1,013]	p=0,729	0,000

ОШ — отношение шансов

Разделение базы на обучающую и тестовую выборки

Далее база данных с 452 пациентками была разделена на обучающую и тестовую выборки в соотношении 70/30. Таким образом, в обучающую базу попали 307 пациенток, в тестовую — 145. Выделенные выборки статистически значимо не различались между собой (табл. 2).

Балансировка базы данных

В ходе проверки сбалансированности обучающей выборки выявлено, что количество пациенток со злокачественными образованиями (ЗНО) было значительно больше, чем пациенток с доброкачественными образованиями (ДНО) (49 (16,0 %) и 258 (84,0 %), соответственно) (рис. 2).

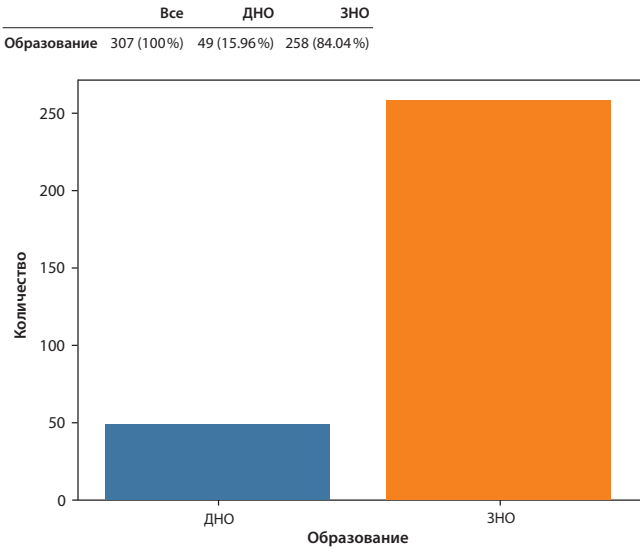

Рис. 2. *Распределение пациенток с ДНО и ЗНО в обучающей выборке — проверка сбалансированности базы*

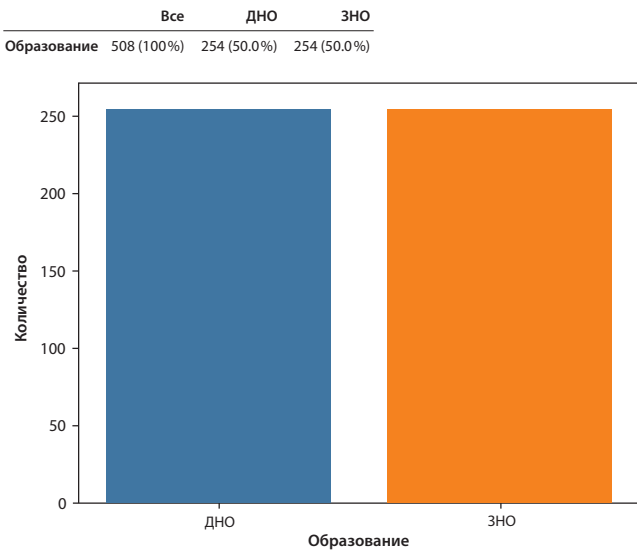

Рис. 3. *Распределение пациенток с ДНО и ЗНО в обучающей выборке после балансировки*

Таблица 2

Сравнение предикторов и исхода в обучающей и тестовой выборках

	Все (n=452)	Обучающая выборка (n=307)	Тестовая выборка (n=145)	p
Предикторы				
Возраст	42,0 [33,75; 50,0],	41,0 [33,0; 51,0],	44,0 [34,0; 50,0],	p=0,320
СОЭ (мм \час	11,0 [6,0; 21,0]	11,0 [6,0; 20,0],	12,0 [6,0; 22,0]	p=0,599
WBC, 10^9/л	6,4 [5,21; 7,76]	6,4 [5,3; 7,8]	6,3 [5,2; 7,6]	p=0,696
RBC 10^{12}\л	4,33 [4,1; 4,6]	4,33 [4,06; 4,66]	4,34 [4,13; 4,6]	p=0,814
ТФ, г\л	2,82 [2,45; 3,03]	2,78 [2,4; 3,01]	2,85 [2,49; 3,09]	p=0,090
СЖ, мкмоль\л	13,96 [7,72; 17,79]	13,75 [7,59; 17,25]	14,28 [8,68; 20,2]	p=0,111
СРБ, мг\л	4,53 [1,0; 9,44],	4,43 [1,0; 8,36]	4,63 [1,0; 11,42]	p=0,473
АЧТВ, сек	30,7 [28,7; 33,37]	30,64 [28,8; 33,4]	30,7 [27,8; 33,2]	p=0,510
АСТ ед/л	17,56 [14,0; 21,35]	17,61 [14,0; 21,86]	17,09 [13,82; 20,3]	p=0,365
ТВ, сек	18,0 [16,7; 19,06]	17,98 [16,7; 19,0]	18,0 [16,8; 19,2]	p=0,701
Фибриноген, г\л	2,79 [2,4; 3,32]	2,8 [2,41; 3,31]	2,72 [2,38; 3,33]	p=0,264
Д-димер, мг\л ФЭЕ	0,33 [0,21; 0,51]	0,33 [0,22; 0,49]	0,34 [0,2; 0,53]	p=0,440
Агрегация тромбоцитов с АДФ,	78,0 [68,38; 89,02]	77,29 [68,09; 87,45]	79,04 [71,0; 90,0]	p=0,136
Мочевина	4,42 [3,83; 5,22]	4,4 [3,8; 5,19]	4,55 [3,9; 5,4]	p=0,228
МНО	1,04 [1,0; 1,06]	1,03 [1,0; 1,06]	1,04 [1,0; 1,07]	p=0,965
Исход				
0 — ДНО	75 (16,6 %)	49 (16,0 %)	26 (17,9 %)	p=0,599
1 — ЗНО	377 (83,4 %)	258 (84,0 %)	119 (82,1 %)	

Для балансировки базы данных в целях получения более стабильной модели прогнозирования был использован метод SMOTE. В результате мы получили базу данных, в которой пациенток с ЗНО и с ДНО было равное количество — по 254 (50 %).

Корреляционный анализ

Для проверки отсутствия высокозначимой корреляционной связи ($|R| > 0,7$) был выполнен корреляционный анализ, который показал, что сильно коррелируют между собой HGB и HCT ($Rs = 0,791$, $p < 0,001$) и MCV с MCH ($Rs = 0,710$, $p < 0,001$) (табл. 3).

Исходя из полученных результатов следует удалить по одной переменной из каждой пары, поэтому было принято решение, что из этих четырех переменных в дальнейшем анализе останутся две переменные: гемоглобин и средний объем эритроцита.

Также из набора были удалены следующие показатели: плазминоген, ХЗФ и насыщение трансферрина железом, как редко встречающиеся в результатах.

Итого в исследуемом наборе осталось 27 показателей крови.

Выбор значимых признаков

Далее был проведен выбор значимых признаков с помощью метода RFE. Метод RFE на основе логистической регрессии отобрал следующие 15 признаков: Возраст, СОЭ, WBC,

Таблица 3

Корреляционный анализ

Переменная 1	Переменная 2	Rs	p
Гемоглобин (HGB)	Гематокрит (HCT)	0,791	p<0,001*
Средний объем эритроцита (MCV)	Среднее содержание гемоглобина в эритроците (MCH)	0,710	p<0,001*

RBC, ТФ, СЖ, СРБ, АЧТВ, АСТ, ТВ, фибриноген, Д-димер, агрегация тромбоцитов с АДФ, мочевина, МНО.

Разработка и тестирование модели

Результаты оценки качества моделей, обученных с помощью выбранных алгоритмов машинного обучения представлены в таблице 4.

Из данных, представленных в таблице 4, следует, что модель, обученная алгоритмом MLR — Lasso, с solver = 'saga' и C = 0,976, на тестовой выборке имеет ROC-AUC = 97,2 % (95,1; 99,0), точность = 90,3 % (86,2; 94,5), чувствительность = 88,4 % (83,5; 93,0), специфичность = 100,0 % (100,0; 100,0).

Модель, обученная алгоритмом MLR — Ridge, с C = 0,266, на тестовой выборке имеет ROC-AUC = 97,0 % (94,8; 98,9), точность = 91,7 % (87,6; 95,9), чувствительность = 90,1 % (85,2; 94,6), специфичность = 100,0 % (100,0; 100,0).

Модель, обученная алгоритмом Extra Tree, с количеством деревьев в лесу n_estimators=17 и с максимальным количеством пациентов конечных узлах max_leaf_nodes = 74, на тестовой выборке имеет ROC-AUC = 97,9 % (95,9; 99,3), точность = 93,1 % (89,0; 96,6) %, чувствительность = 95,0 % (91,5; 98,3), специфичность = 83,3 % (70,0; 95,5).

Модель, обученная алгоритмом kNN, с количеством соседей n_neighbors = 9, евклидовым расстоянием в качестве шкалы расстояний (weights='distance'), на тестовой выборке имеет ROC-AUC = 93,9 % (88,5; 98,1), точность = 86,9 % (82,1; 91,7), чувствительность = 86,8 % (81,5; 91,7), специфичность = 87,5 % (76,2; 97,2).

Модель, обученная алгоритмом SVM, при использовании ядра kernel=«linear», на тестовой выборке имеет ROC-AUC = 96,9 % (94,6; 98,9), точность = 92,4 % (89,0; 95,9), чувствительность = 93,4 % (89,5; 96,8), специфичность = 87,5 % (76,0; 100,0).

Модель, обученная алгоритмом Random Forest, с функцией измерения качества разделения criterion='entropy', с максимальным количеством пациентов конечных узлах max_leaf_nodes = 44 и с количеством деревьев в лесу n_estimators = 83, на тестовой выборке имеет ROC-AUC = 97,8 % (95,8; 99,2), точность = 93,1 % (89,7; 96,6), чувствительность = 94,2 % (90,5; 97,5), специфичность = 87,5 % (76,0; 96,7).

Модель, обученная алгоритмом Catboost, с максимальным количеством итераций early_stopping_rounds=15, с коэффициентом скорости обучения learning_rate=0,0997, n_estimators=61, на тестовой выборке имеет ROC-AUC = 97,3 % (95,0; 99,0), точность = 92,4 % (89,0; 95,9), чувствительность = 94,2 % (90,6; 97,5), специфичность = 83,3 % (70,0; 95,0).

Модель, обученная алгоритмом LGBM, с коэффициентом скорости обучения learning_rate=0,248, с максимальным количеством интервалов, в которых значения объектов делятся или «группируются» max_bin = 15, на тестовой выборке имеет ROC-AUC = 94,3 % (91,1; 97,1), точность = 86,2 % (81,4; 91,0), чувствительность = 85,1 % (79,7; 90,4), специфичность = 91,7 % (81,0; 100,0).

Модель, обученная алгоритмом Extreme Gradient Boosting (XGB), с коэффициентом скорости обучения learning_rate=0,425, с количеством деревьев в лесу n_estimators=51, на тестовой выборке имеет ROC-AUC = 98,2 % (96,0; 99,7), точность = 94,5 % (91,0; 97,2), чувствительность = 95,0 % (91,2; 98,3), специфичность = 91,7 % (81,0; 100,0).

Изучая результаты метрик качества выбранных алгоритмов машинного обучения, мы пришли к выводу, что модель XGB обладает самой высокой точностью, чувствительностью и специфичностью и, соответственно, лучшими возможностями прогнозирования РШМ, чем другие модели, обученные с помощью алгоритмов МО, рассмотренных в данной статье.

Напротив, Naive Bayes (NB) показал наименьшие метрики качества, чем другие модели МО, Модель, обученная алгоритмом Naive Bayes на значениях гиперпараметрах по умолчанию, на тестовой выборке имеет ROC-AUC = 90,2 % (84,3; 95,2), точность = 89,0 % (84,1; 93,1), чувствительность = 93,4 % (89,6; 97,2), специфичность = 66,7 % (50,0; 82,4).

Как видно из рис. 4, ROC-кривая, принадлежащая алгоритму XGB, имеет максимальную площадь под кривой: ROC-AUC = 98,2 % (96,0; 99,7). Поэтому модель XGB имеет больше возможностей для прогнозирования, чем другие алгоритмы в отношении диагностики РШМ.

Таблица 4

Сравнение метрик качества моделей на обучающей и тестовой выборках

Модель	MLR — Lasso	MLR — Ridge	Extra Tree	kNN	SVM
Cut-off	> 0,45	> 0,46	> 0,55	> 0,50	> 0,43
Обучение					
ROC-AUC	95,2 % [93,5; 96,8]	95,5 % [93,9; 97,0]	100,0 % [100,0; 100,0]	100,0 % [100,0; 100,0]	95,9 % [94,3; 97,4]
Точность	91,5 % [89,4; 93,7]	93,1 % [91,1; 94,9]	99,4 % [98,8; 100,0]	100,0 % [100,0; 100,0]	93,5 % [91,5; 95,1]
Чувствительность	91,3 % [88,2; 94,2]	92,9 % [90,0; 95,4]	99,2 % [98,1; 100,0]	100,0 % [100,0; 100,0]	93,7 % [91,0; 96,1]
Специфичность	91,7 % [89,0; 94,5]	93,3 % [90,8; 95,7]	99,6 % [98,8; 100,0]	100,0 % [100,0; 100,0]	93,3 % [90,9; 95,7]
R^2	66,1 % [57,5; 74,7]	72,4 % [64,5; 79,5]	97,6 % [95,3; 100,0]	100,0 % [100,0; 100,0]	74,0 % [66,1; 80,3]
Тестирование					
ROC-AUC	97,2 % [95,1; 99,0]	97,0 % [94,8; 98,9]	97,9 % [95,9; 99,3]	93,9 % [88,5; 98,1]	96,9 % [94,6; 98,9]
Точность	90,3 % [86,2; 94,5]	91,7 % [87,6; 95,9]	93,1 % [89,0; 96,6]	86,9 % [82,1; 91,7]	92,4 % [89,0; 95,9]
Чувствительность	88,4 % [83,5; 93,0]	90,1 % [85,2; 94,6]	95,0 % [91,5; 98,3]	86,8 % [81,5; 91,7]	93,4 % [89,5; 96,8]
Специфичность	100,0 % [100,0; 100,0]	100,0 % [100,0; 100,0]	83,3 % [70,0; 95,5]	87,5 % [76,2; 97,2]	87,5 % [76,0; 100,0]

Модель	Naive Bayes	Random Forest	XGB	Catboost	LightGBM
Cut-off	> 0,05	> 0,59	> 0,82	> 0,48	> 0,55
Обучение					
ROC-AUC	93,0 % [91,1; 94,8]	100,0 % [100,0; 100,0]	100,0 % [100,0; 100,0]	99,8 % [99,6; 100,0]	93,9 % [92,1; 95,6]
Точность	86,2 % [83,5; 88,6]	100,0 % [100,0; 100,0]	100,0 % [100,0; 100,0]	98,2 % [97,2; 99,2]	90,0 % [87,8; 92,1]
Чувствительность	86,2 % [82,2; 89,5]	100,0 % [100,0; 100,0]	100,0 % [100,0; 100,0]	98,0 % [96,5; 99,2]	89,8 % [86,7; 92,9]
Специфичность	86,2 % [82,7; 89,8]	100,0 % [100,0; 100,0]	100,0 % [100,0; 100,0]	98,4 % [97,1; 99,6]	90,2 % [86,9; 93,2]
R^2	44,9 % [33,9; 54,3]	100,0 % [100,0; 100,0]	100,0 % [100,0; 100,0]	92,9 % [89,0; 96,8]	59,8 % [51,2; 68,3]
Тестирование					
ROC-AUC	90,2 % [84,3; 95,2]	97,8 % [95,8; 99,2]	98,2 % [96,0; 99,7]	97,3 % [95,0; 99,0]	94,3 % [91,1; 97,1]
Точность	89,0 % [84,1; 93,1]	93,1 % [89,7; 96,6]	94,5 % [91,0; 97,2]	92,4 % [89,0; 95,9]	86,2 % [81,4; 91,0]
Чувствительность	93,4 % [89,6; 97,2]	94,2 % [90,5; 97,5]	**95,0 % [91,2; 98,3]**	94,2 % [90,6; 97,5]	85,1 % [79,7; 90,4]
Специфичность	66,7 % [50,0; 82,4]	87,5 % [76,0; 96,7]	**91,7 % [81,0; 100,0]**	83,3 % [70,0; 95,0]	91,7 % [81,0; 100,0]

Напротив, модель NB имеет минимальную площадь под кривой и поэтому является самой слабой моделью для прогнозирования РШМ.

Еще один важный результат, полученный в результате сравнения моделей, заключается в том, что качество модели XGB статистически значимо не отличалось от моделей, построенных с помощью алгоритмов Random Forest, Extra Tree и kNN. Иными словами, ансамблевые алгоритмы дают модели с более высокими метриками качества прогнозирования РШМ, чем другие алгоритмы МО.

Рассмотрим метрики модели, построенной на основе алгоритма XGB, как самой эффективной, более подробно. Результаты рассчитанных метрик показаны на рис. 5–7.

Результаты классификации данных (матрица путаницы) обучающей и тестовой выборок по TN, FP, FN и TN с использованием модели XGB показаны на рис. 6.

Чтобы оценить клиническую полезность модели, был использован Анализ кривой принятия решения (рис. 7). Анализ кривой принятия решения определил, что диапазон пороговых вероятностей составляет от 0 до 1, в которых модель имеет большую ценность, чем «лечить всех» или «не лечить никого».

На рисунке 8 указана важность факторов риска РШМ в модели XGB. Факторы Агрегация тромбоцитов с АДФ (0,228), СОЭ (0,088), Д-димер (0,086), ТФ (0,083) и возраст (0,078) имели наибольшую важность, чем другие.

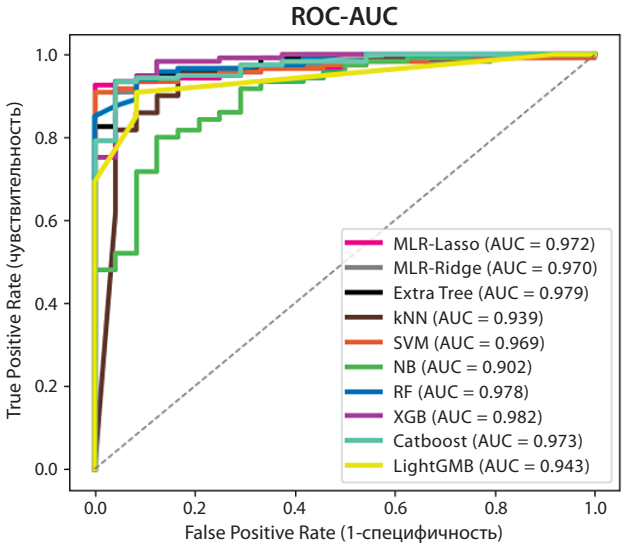

Рис. 4. *ROC кривые для всех алгоритмов на тестовой выборке*

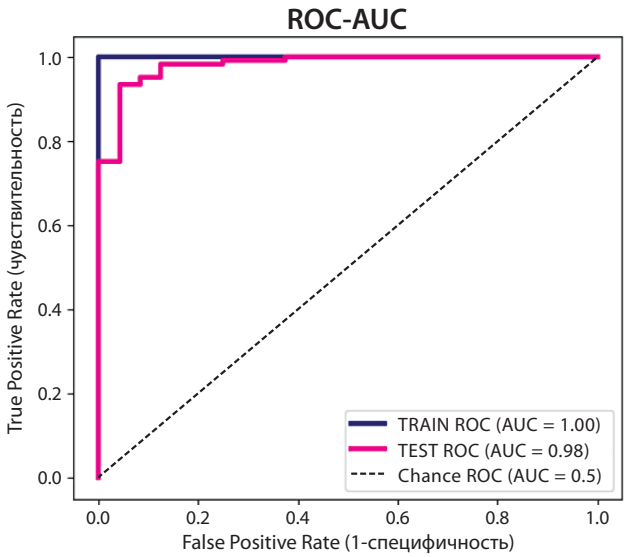

Рис. 5. *ROC-AUC модели, построенной на основе алгоритма XGB на обучающей и тестовой выборках*

Мы также отобразили важность факторов при диагностике РШМ на основе оценки признаков перестановки, средних аддитивных объяснений Шепли (SHAP) и значений SHAP на рис. 9. В этом случае наиболее важными факторами для прогнозирования РШМ оказались: агрегация тромбоцитов с АДФ, Д-димер, СОЭ, АСТ и СЖ.

Обсуждение

Учитывая растущую распространенность РШМ и не до конца изученную природу прогрессирования, использование эффективных скрининговых технологий может сыграть важную роль в ранней диагностике РШМ, снижении частоты его развития, неблагоприятных исходов.

Таблица 5

Сравнение качества построенных моделей между собой (критерий DeLong)

	MLR — Lasso	MLR — Ridge	Extra Tree	kNN	SVM	Naive Bayes	Random Forest	XGB	Catboost	LightGBM
MLR — Lasso		p=0,293	p<0,001*	p<0,001*	p=0,094	p=0,014*	p<0,001*	p<0,001*	p<0,001*	p=0,058
MLR — Ridge	p=0,293		p<0,001*	p<0,001*	p=0,192	p=0,008*	p<0,001*	p<0,001*	p<0,001*	p=0,039*
Extra Tree	p<0,001*	p<0,001*		p=0,135	p<0,001*	p<0,001*	p=0,259	p=0,183	p=0,013*	p<0,001*
kNN	p<0,001*	p<0,001*	p=0,135		p<0,001*	p<0,001*	p=0,096	p=0,087	p=0,826	p<0,001*
SVM	p=0,094	p=0,192	p<0,001*	p<0,001*		p=0,005*	p<0,001*	p<0,001*	p<0,001*	p=0,022*
Naive Bayes	p=0,014*	p=0,008*	p<0,001*	p<0,001*	p=0,005*		p<0,001*	p<0,001*	p<0,001*	p=0,694
Random Forest	p<0,001*	p<0,001*	p=0,259	p=0,096	p<0,001*	p<0,001*		p=0,452	p=0,006*	p<0,001*
Gradient Boosting	p<0,001*	p<0,001*	p=0,183	p=0,087	p<0,001*	p<0,001*	p=0,452		p=0,006*	p<0,001*
Catboost	p<0,001*	p<0,001*	p=0,013*	p=0,826	p<0,001*	p<0,001*	p=0,006*	p=0,006*		p<0,001*
LightGBM	p=0,058	p=0,039*	p<0,001*	p<0,001*	p=0,022*	p=0,694	p<0,001*	p<0,001*	p<0,001*	

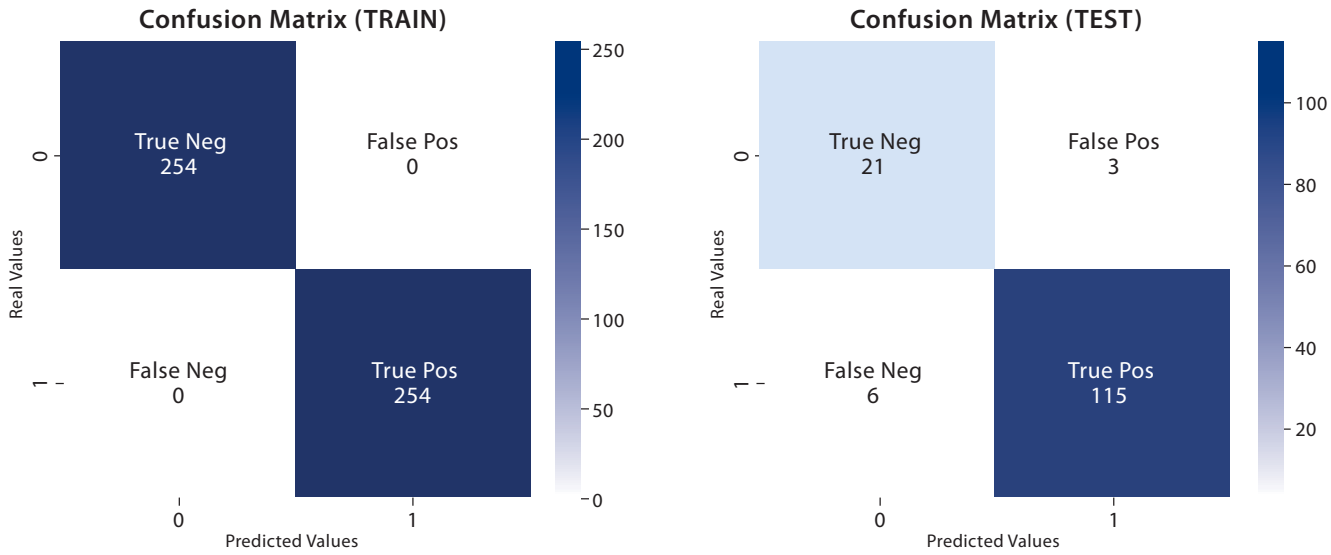

Рис. 6. *Матрица путаницы для обучающей (TRAIN) и тестовой (TEST) выборках*

Таким образом, это исследование было направлено на использование искусственного интеллекта для разработки потенциального прогностического решения в помощь врачу первичного звена здравоохранения для ранней диагностики РШМ на основе факторов риска. С этой целью, используя базу данных медицинского центра, занимающегося диагностикой РШМ, в который обращаются женщины с подозрением на РШМ со всей России и из ближнего зарубежья, мы разработали подход, основанный на данных МО.

После предварительной обработки и подготовки базы данных мы использовали выбранные

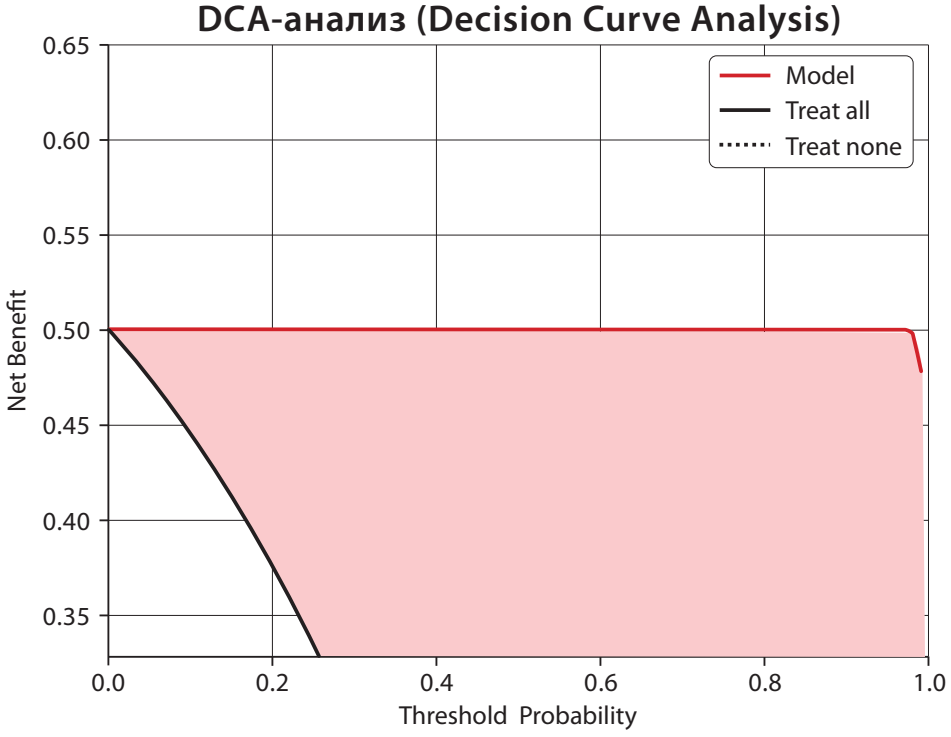

Рис. 7. *Метрики XGB модели: DCA-анализ*

Рис. 8. *Важность факторов риска, связанных с прогнозированием РШМ в модели XGB*

алгоритмы МО для анализа данных лабораторных исследований крови женщин с подтвержденным диагнозом РШМ, так и с отсутствием РШМ, для построения моделей прогнозирования. Наконец, для целей прогнозирования был выбран лучший алгоритм с наивысшими метриками качества диагностики, обученный МО.

Кроме того, наиболее значимые факторы риска, связанные с прогнозированием РШМ, были извлечены из наиболее эффективного алгоритма, обученного МО. Факторы, включающие агрегацию тромбоцитов с АДФ (0,228), СОЭ (0,088), Д-димер (0,086), Трансферин (0,083) и возраст (0,078), имели наибольшую важность, чем другие. После получения лучшего прогнозирования модели для РШМ мы протестировали ее с помощью метода Бутстреп и тестовой выборки. Текущее исследование показало, что на тестовой выборке модель XGB показала ROC-AUC = 98,2 % (96,0; 99,7), точность = 94,5 % (91,0; 97,2), чувствительность = 95,0 % (91,2; 98,3), специфичность = 91,7 % (81,0; 100,0).

Исследований по использованию МО для РШМ на основе факторов риска по рутинным показателям крови нами не найдено.

Однако было проведено несколько исследований по теме, касающейся рака яичника (РЯ). Например, Lu и др. использовали алгоритмы МО (МЛР и ДР) для прогнозирования РЯ с использованием данных, включающего 49 предикторов: демографических показателей, общей химии, опухолевых маркеров и рутинных показателях крови, относящихся к злокачественным и доброкачественным случаям РЯ. Выборки 235 и 114 использовались для обучения и тестирования алгоритма дерева решений (DT). Построенный алгоритм сравнивался с алгоритмом MLR. Результаты показали, что DT с ROC-AUC = 0,888 имеет лучшее качество, чем MLR (ROC-AUC = 0,877). Для диагностики РЯ авторы использовали биомаркеры: HE4 (human epididymis protein 4) и РЭА (carcinoembryonic antigen) [43].

Ahamad и др. использовали подход МО, основанный на клинических данных 349 пациентов с доброкачественными и злокачественными образованиями яичника, для создания модели выявления РЯ на ранних стадиях. Для построения моделей использовали следующие алгоритмы МО: RF (Random Forest), SVM (Support Vector Machine), DT (Decision Tree), XGB (Extreme Gradient Boosting), MLR (Logistic Regression), GBM (Gradient Boosting Machine) и LGBM (Light Gradient Boosting Machine). Для диагностики РЯ авторы использовали следующие биомаркеры: CA 125 (serumsamples

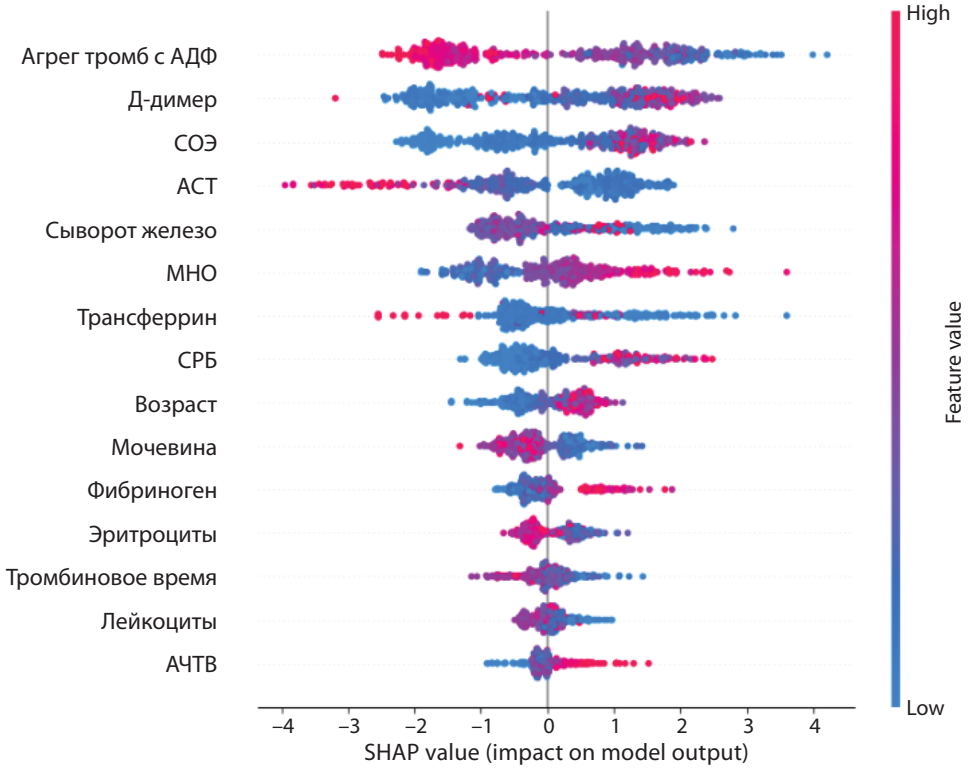

Рис. 9. *Значения SHAP, связанные с прогнозом РШМ, по обучающей выборке в модели XGB*

carbohydrate antigen 125), CA 19,9 (carbohydrate antigen 19–9), РЭА (carcinoembryonic antigen) и HE 4 (human epididymis protein), а также соотношение нейтрофилов и лимфоцитов (NLR), тромбокрит, гематокрит, аланинаминотрансфераза, кальций, непрямой билирубин, мочевая кислота, натрий. Алгоритм XGB в этом случае построил прогностическую модель наилучшего качества (ROC-AUC=0,86) [44].

Ограничения исследования

Ограничением данного исследования является его одноцентровой характер проведения. Некоторые потерянные данные, связанные со случаями РШМ, были восстановлены с использованием метода вменения (KNN), что повлияло на возможность обобщения. Для будущих исследований мы рекомендуем использовать большее количество данных. Мы также предлагаем в будущем проверить построенные модели на внешних данных, например, данных другого медицинского центра, чтобы обеспечить большую обобщаемость моделей.

Заключение

Это исследование было направлено на создание нового метода диагностики РШМ с использованием рутинных лабораторных показателей крови.

Разработанная модель, использующая алгоритм XGB, продемонстрировала высокую производительность с ROC-AUC 98,2 %, показав превосходную точность в стратификации ЗНО и ДНО. Кроме того, модель достигла высокой точности (94,5 %), чувствительности (95,0 %) и специфичности (91,7 %), укрепив свое положение как оптимального алгоритма для диагностики РШМ.

Факторы риска, извлеченные из модели XGB, могут быть интегрированы в интеллектуальные системы, такие как системы поддержки принятия врачебных решений (СППВР), что может помочь врачам в выявлении пациенток с высоким риском развития или наличия РШМ. Это, в свою очередь, может способствовать разработке и реализации целевых профилактических стратегий для групп высокого риска, в конечном итоге улучшив общие результаты в отношении здоровья.

Ранняя идентификация женщин с высоким риском развития или наличия РШМ в процессе скрининга позволит снизить количество расширенных инвазивных подходов к лечению из-за своевременного начала лечения. Это не только оптимизирует тактику лечения и профилактических мер, но и потенциально приведет к снижению затрат в здравоохранении.

СПИСОК ЛИТЕРАТУРЫ

1. https://gco.iarc.fr/today/en/dataviz/bars?mode=population&types=0&cancers=23&populations=900&key=total

2. https://gco.iarc.fr/today/en/dataviz/bars?mode=population&types=1&cancers=23&populations=900&key=totalhttps://gco.iarc.fr/today/en/dataviz/bars-compare-populations

3. https://gco.iarc.fr/today/en/dataviz/bars?mode=population&types=0&cancers=23&populations=643&key=total

4. *Zhang S, Xu H, Zhang L, Qiao Y.* Cervical cancer: Epidemiology, risk factors and screening. Chin J Cancer Res. 2020 Dec 31;32(6):720–728. doi: 10.21147/j.issn.1000–9604.2020.06.05. PMID: 33446995; PMCID: PMC7797226

5. *Louie KS, Castellsague X, de Sanjose S,* et al. International Agency for Research on Cancer Multicenter Cervical Cancer Study Group. Smoking and passive smoking in cervical cancer risk: pooled analysis of couples from the IARC multicentric case-control studies. Cancer Epidemiol Biomarkers Prev. 2011 Jul;20(7):1379–90. doi: 10.1158/1055–9965.EPI-11–0284. Epub 2011 May 24. PMID: 21610224

6. *Li XY, Li G, Gong TT,* et al. Non-Genetic Factors and Risk of Cervical Cancer: An Umbrella Review of Systematic Reviews and Meta-Analyses of Observational Studies. Int J Public Health. 2023 Mar 31;68:1605198. doi: 10.3389/ijph.2023.1605198. PMID: 37065642; PMCID: PMC10103589.

7. *Keskin N, Biyik I, Ince O.* et al. Evaluation of the consistency ratios of cervical smear, cervical biopsy and conization results. Ginekol Pol. 2021;92(11):778–783. doi: 10.5603/GP.a2021.0051. Epub 2021 Apr 29. PMID: 33914320.

8. Эффективность цитологической диагностики цервикальной интраэпителиальной неоплазии и рака шейки матки в зависимости от способа взятия материала / Т. В. Сушинская, Н. Н. Волченко, Ю. Э. Доброхотова [и др.] // Онкогинекология. — 2017. — № 3(23). — С. 51–59. — EDN ZHGTGB.]

9. *Савостикова М. В., Короленкова Л. И., Федосеева Е. С., Пименова В. В.* Опыт применения жидкостной технологии BD SUREPATH™ для ранней диагностики и скрининга предопухолевой и опухолевой патологии шейки матки в Ростовской области // Онкогинекология . 2018. — № 4(28). — С. 50–60.

10. Сопоставление результатов точечной и эксцизионной биопсий в диагностике цервикальных интраэпителиальных неоплазий и рака шейки матки / Е. К. Губская, Д. В. Бурцев, Т. А. Димитриади [и др.] // Вопросы онкологии. — 2023. — Т. 69, № 3S. — С. 118–120. — EDN NPVRBY.

11. *Petousis S, Christidis P, Margioula-Siarkou C,* et al. Discrepancy between colposcopy, punch biopsy and final histology of cone specimen: a prospective study. Arch Gynecol Obstet. 2018 May;297(5):1271–1275. doi: 10.1007/s00404–018–4714–8. Epub 2018 Feb 13. PMID: 29442140.

12. *Ye J, Cheng XD, Cheng B,* et al. MiRNA detection in cervical exfoliated cells for missed high-grade lesions in women with LSIL/CIN1 diagnosis after colposcopy-guided biopsy. BMC Cancer. 2019 Jan 30;19(1):112. doi: 10.1186/s12885–019–5311–3. PMID: 30700264; PMCID: PMC6354336.

13. *Ren, H., Jia, M., Zhao, S.,* et al. (2020). Factors Correlated with the Accuracy of Colposcopy-Directed Biopsy: A Systematic Review and Meta-Analysis. Journal of Investigative Surgery, 1–9. doi:10.1080/08941939.2020.1850944.

14. *Ongsulee P, Chotchaung V, Bamrungsi E, Rodcheewit T.* Big data, predictive analytics and machine learning. In: Ongsulee P, Chotchaung V, Bamrungsi E, Rodcheewit T, editors. 2018 16th international conference on ICT and knowledge engineering (ICT&KE); 2018 21–23 Nov. Bangkok: IEEE; 2018. p. 21–3.

15. *Lalmuanawma S, Hussain J, Chhakchhuak L.* Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. Chaos Solitons Fractals. 2020;139: 110059.

16. *Bertsimas D, Wiberg H*. Machine learning in oncology: methods, applications, and challenges. JCO Clin Cancer Inform. 2020;4:885–94.

17. *Stark GF, Hart GR, Nartowt BJ, Deng J.* Predicting breast cancer risk using personal health data and machine learning models. PLoS ONE. 2019;14(12): e0226765.

18. *Ming C, Viassolo V, Probst-Hensch N,* et al. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res. 2019;21(1):1–11.

19. *Akazawa M, Hashimoto K.* Artificial Intelligence in Ovarian Cancer Diagnosis. Anticancer Res. 2022 Aug;40(8):4795–4800. doi: 10.21873/anticanres.14482. PMID: 32727807.

20. *Kawakami E, Tabata J, Yanaihara N,* et al. Application of Artificial Intelligence for Preoperative Diagnostic and Prognostic Prediction in Epithelial Ovarian Cancer Based on Blood Biomarkers. Clin Cancer Res. 2019 May 15;25(10):3006–3015.

21. *Li H, Lin J, Xiao Y,* et al. Colorectal Cancer Detected by Machine Learning Models Using Conventional Laboratory Test Data. Technol Cancer Res Treat. 2021 Jan-Dec;20:15330338211058352. doi: 10.1177/15330338211058352. PMID: 34806496; PMCID: PMC8606732.

22. *Fan G, Cui R, Zhang R,* et al. Routine blood biomarkers for the detection of multiple myeloma using machine learning. Int J Lab Hematol. 2022 Jun;44(3):558–566. doi: 10.1111/ijlh.13806. Epub 2022 Feb 23. PMID: 35199461.

23. *Saito S, Sakamoto S, Higuchi K,* et al. Machine-learning predicts time-series prognosis factors in metastatic prostate cancer patients treated with androgen deprivation therapy. Sci Rep. 2023 Apr 18;13(1):6325. doi: 10.1038/s41598–023–32987–6. PMID: 37072487; PMCID: PMC10113215.

24. *Bentick K, Runevic J, Akula S,* et al. Machine learning models based on routinely sampled blood tests can predict the presence of malignancy amongst patients with suspected musculoskeletal malignancy. Methods. 2023 Dec;220:55–60. doi: 10.1016/j.ymeth.2023.10.012. Epub 2023 Nov 10. PMID: 37951558.

25. *Akbar S, Hayat M.* iMethyl-STTNC: identification of N6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences. J Theor Biol. 2018;455:205–11.

26. *Ali F, Ahmed S, Swati ZNK, Akbar S.* DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. J Comput Aided Mol Des. 2019;33(7):645–58.

27. *Akbar S, Hayat M, Tahir M,* et al. cACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model. Artif Intell Med. 2022;131: 102349.

28. *Akbar S, Ahmad A, Hayat M,* et al. iAtbP-Hyb-EnC: prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. Comput Biol Med.2021;137: 104778

29. *Atitallah SB, Driss M, Boulila W, Ghézala HB.* Leveraging deep learning and IoT big data analytics to support the smart cities development: review and future directions. Comput Sci Rev. 2020;38: 100303.

30. *Gong X, Zheng B, Xu G,* et al. Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cancer. J Thorac Dis. 2021;13(11):6240.

31. Малоинвазивная Прогностическая Модель предсказания наличия регионарных метастазов рака шейки матки по рутинным лабораторным показателям крови / Т. В. Сушинская, Н. И. Стуклов, Е. В. Щепкина [и др.] // Проблемы стандартизации в здравоохранении. — 2023. — № 5–6. — С. 12–18. — DOI 10.26347/1607–2502202305–06012–018. — EDN IOQYGC.

32. *Huan Zhao, Yuling Wang, Yilin Sun* et al. Hematological Indicator-Based Machine Learning Models for Preoperative Prediction of Lymph Node Metastasis in Cervical Cancer, 06 February 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2519676/v1]

33. *Blagus R, Lusa L.* SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013 Mar 22;14:106. doi: 10.1186/1471–2105–14–106. PMID: 23522326; PMCID: PMC3648438.

34. *Swana EF, Doorsamy W, Bokoro P.* Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. Sensors (Basel). 2022 Apr 23;22(9):3246. doi: 10.3390/s22093246. PMID: 35590937; PMCID: PMC9099503.

35. *Senan EM, Al-Adhaileh MH, Alsaade FW,* et al. Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques. J Healthc Eng. 2021 Jun 9;2021:1004767. doi: 10.1155/2021/1004767. PMID: 34211680; PMCID: PMC8208843.]

36. *Воронцов К. В.* Курс по машинному обучению. Course on machine learning from Vorontsov K. URL: http://www.machinelearning.ru (26.05.2024).

37. *Fawcett, Tom* (2006). «An Introduction to ROC Analysis» (PDF). Pattern Recognition Letters. 27 (8): 861–874. doi:10.1016/j.patrec.2005.10.010. S2CID 2027090

38. *Miwakeichi F, Galka A.* Comparison of Bootstrap Methods for Estimating Causality in Linear Dynamic Systems: A Review. Entropy (Basel). 2023 Jul 17;25(7):1070. doi: 10.3390/e25071070. PMID: 37510017; PMCID: PMC10378223.

39. *Van Calster B, Wynants L, Verbeek JFM,* et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. Eur Urol. 2018 Dec;74(6):796–804. doi: 10.1016/j.eururo.2018.08.038. Epub 2018 Sep 19. PMID: 30241973; PMCID: PMC6261531.

40. *Zhang L, Tang L, Chen S,* et al. A nomogram for predicting the 4-year risk of chronic kidney disease among Chinese elderly adults. Int Urol Nephrol. 2023 Jun;55(6):1609–1617. doi: 10.1007/s11255–023–03470-y. Epub 2023 Jan 31. PMID: 36720744

41. *DeLong E. R., DeLong D. M., Clarke-Pearson D. L.* Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach // Biometrics. 1988. № 3 (44). C. 837

42. *Lundberg, S., & Lee, S. I.* (2017). A Unified Approach to Interpreting Model Predictions.

43. *Lu M, Fan Z, Xu B,* et al. Using machine learning to predict ovarian cancer. Int J Med Inform. 2020 Sep;141:104195. doi: 10.1016/j.ijmedinf.2020.104195. Epub 2020 May 23. PMID: 32485554.

44. *Ahamad MM, Aktar S, Uddin MJ,* et al. Early-stage detection of ovarian cancer based on clinical data using machine learning approaches. J Pers Med. 2022;12(8):1211.

СВЕДЕНИЯ ОБ АВТОРАХ

Кузнецов Антон Игоревич, программист, студент кафедры «Прикладные программные средства и математические методы института № 3 "Системы управления, информатика и электроэнергетика"» Московского авиационного института (Национальный исследовательский университет), 125080, Москва, Волоколамское шоссе, д. 4, drednout5786@yandex.ru; ORCID ID 0000–0003–2182–5792, eLibrary SPIN: 8824–9080.

Kuznetsov Anton I., programmer, student of the Department of Applied Software and Mathematical Methods of the Faculty of Control Systems, Informatics & Electricity of Moscow Aviation Institute (National Research University), Russia, 125080, Moscow, Vernadsky Avenue, 4, e-mail: drednout5786@yandex.ru; ORCID ID 0000–0003–2182–5792, eLibrary SPIN: 8824–9080.

Щепкина Елена Викторовна, кандидат социологических наук, заместитель начальника Отдела сводного контингента и статистики Учебно-методического управления Российской академии народного хозяйства и государственной службы при Президенте Российской Федерации (РАНХиГС), исследователь данных в ГБУЗ «Научно-практический клинический центр диагностики и телемедицинских технологий» ДЗМ, 119571, Россия, Москва, проспект Вернадского, д. 82, стр.1; 127051, Россия, Москва, ул. Петровка, д. 24, стр. 1, elenaschepkina@gmail.com; ORCID ID 0000–0002–2079–1482, SPIN: 2347–9436, AuthorID: 959277, Scopus Author ID: 57211515165

Schepkina Elena V., PhD, Deputy Head of the Department of Statistics and the Consolidated Contingent of the Educational and Methodological Department at the Presidential Academy — RANEPA, Data scientist of the State Healthcare Institution Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Department of Healthcare, Russia, 119571, Moscow, Vernadsky Avenue, 82, building 1; 127051, Moscow, st. Petrovka, 24, building 1, 4, elenaschepkina@gmail.com; ORCID ID 0000–0002–2079–1482, SPIN: 2347–9436, AuthorID: 959277, Scopus Author ID: 57211515165

Сушинская Татьяна Валентиновна, кандидат медицинских наук, старший научный сотрудник отдела опухолей репродуктивных и мочевыводящих органов МНИОИ им. П. А. Герцена, филиала ФГБУ «НМИЦ радиологии» Минздрава России, 125284, Россия, Москва, 2-й Боткинский проезд, д. 3, ORCID ID 0000–0001–8812–9105. SPIN-код: 7283–0014, AuthorID: 446611

Sushinskaya Tatyana V., C. Sc. (Med.), senior researcher, P. Hertsen Moscow Oncology Research Institute — Branch of the National Medical Research Radiological Centre, 3, 2nd Botkinskiy lane, Moscow, 125284, Russian Federation; ORCID ID 0000–0001–8812–9105. SPIN-код: 7283–0014, AuthorID: 446611

Епифанова Светлана Викторовна, кандидат медицинских наук, врач-рентгенолог отделения рентгеновской диагностики и томографии Федерального Государственного Бюджетного Учреждения «Центральная клиническая больница с поликлиникой Управления делами Президента Российской Федерации», врач-рентгенолог в ГБУЗ «Научно-практический клинический центр диагностики и телемедицинских технологий» ДЗМ, Россия, 121359, Москва, ул. Маршала Тимошенко, д.15, 127051, Россия, Москва, ул. Петровка, д. 24, стр. 1, svepifanova@yandex.ru; ORCID ID 0000–0002–7591–5120, SPIN-код 9067–5033

Epifanova Svetlana V., PhD, Radiologist of Radiology and Tomography Department Federal State Institution Central Clinical Hospital with Out-Patient Clinic of the Presidential Administration of the Russian Federation, Radiologist of the Department of Expertise and Quality of the State Healthcare Institution Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Department of Healthcare, Russia, 121359, Moscow, st. Marshal Timoshenko, 15. Russia, 127051, Moscow, st. Petrovka, 24, building 1, svepifanova@yandex.ru; ORCID ID 0000–0002–7591–5120, SPIN-код 9067–5033

Стуклов Николай Игоревич, доктор медицинских наук, руководитель курса гематологии, профессор кафедры госпитальной терапии с курсами эндокринологии, гематологии и клинической лабораторной диагностики Медицинского института РУДН, 117198, Москва, ул. Миклухо-Маклая, д. 8; главный научный сотрудник отделения высокодозной химиотерапии с блоком трансплантации костного мозга МНИОИ им. П. А. Герцена — филиала ФГБУ «НМИЦ радиологии» Минздрава России, 125284, Россия, Москва, 2-й Боткинский проезд, д. 3, stuklovn@gmail.com; ORCID ID 0000–0002–4546–1578.

Stuklov Nikolay I., Dr. Sc. (Med.), Head of the Course of Hematology professor of the Department of Hospital Therapy with the Courses of Endocrinology, Hematology, & Clinical Laboratory Diagnostics, Medical Institute of the Russian University of Peoples' Friendship; 8, Miklukho-Maklaya str., Moscow, 117198, Russian Federation; principal researcher, P. Hertsen Moscow Oncology Research Institute — Branch of the National Medical Research Radiological Centre; 3, 2nd Botkinskiy pass., Moscow, 125284, Russian Federation; stuklovn@gmail.com; ORCID ID 0000–0002–4546–1578.

Фаур Дарий Мохамадович, студент факультета Высшей школы экономики и бизнеса РЭУ им. Г. В. Плеханова, 117997, Россия, Москва, Стремянный пер 36; dafa5801@gmail.com; ORCID ID 0009–0006–4756–4681.

Faur Dariy M., student of the faculty of the Higher School of Economics and Business of the Plekhanov Russian University of Economics, 117997, Russia, Moscow, Stremyanny lane 36; dafa5801@gmail.com; ORCID ID 0009–0006–4756–4681.

Каприн Андрей Дмитриевич, академик РАН, д.м.н., профессор, заслуженный врач РФ, член-корреспондент РАО, генеральный директор, ФГБУ «НМИЦ радиологии» Минздрава России, г. Обнинск, Российская Федерация; директор, МНИОИ им. П. А. Герцена — филиал ФГБУ «НМИЦ радиологии» Минздрава России, Москва, Российская Федерация; заведующий кафедрой урологии и оперативной нефрологии с курсом онкоурологии медицинского факультета, ФГАОУ ВО «Российский университет дружбы народов», Москва, Российская Федерация; mnioi@mail.ru; ORCID ID 0000–0001–8784–8415, SPIN: 1759–8101, AuthorID: 96775, ResearcherID: K-1445–2014, Scopus Author ID: 6602709853

Kaprin Andrey D., academician of Russian Academy of Sciences, Dr. Sci. (Med.), professor, honored doctor of the Russian Federation, corr. member of the RAE, general director National Medical Research Radiological Centre, Obninsk, Russian Federation; director at the P. A. Hertsen Moscow Oncology Research Institute — Branch of the National Medical Research Radiological Centre, Moscow, Russian Federation; head of the department of urology and operative Nephrology with the course of oncourology of the Medical Institute, Peoples' Friendship University of Russia, Moscow, Russian Federation. mnioi@mail.ru; ORCID ID 0000–0001–8784–8415, SPIN: 1759–8101, AuthorID: 96775, ResearcherID: K-1445–2014, Scopus Author ID: 6602709853