DOI 10.37882/2223-2966.2025.06-2.33

ИННОВАЦИОННЫЕ ПОДХОДЫ К НЕИНВАЗИВНОЙ ДИАГНОСТИКЕ ЭНДОМЕТРИОЗА: КОНСТРУИРОВАНИЕ И ОТБОР ФАКТОРОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В КЛИНИЧЕСКОЙ ПРАКТИКЕ

INNOVATIVE APPROACHES TO NON-INVASIVE ENDOMETRIOSIS DIAGNOSIS: CONSTRUCTION AND SELECTION OF FACTORS USING MACHINE LEARNING METHODS IN CLINICAL PRACTICE

A. Rusinova

Summary. Endometriosis is a widely prevalent pathological condition; however, its clinical manifestations and underlying mechanisms remain insufficiently studied. The time gap between the onset of the first symptoms and the establishment of a diagnosis can sometimes exceed ten years, significantly hindering early diagnosis and adequate treatment. Currently, there is no universal treatment capable of completely eradicating endometriosis, underscoring the necessity for developing new diagnostic approaches. This study examines the construction and selection of factors associated with the risk of developing endometriosis using modern machine learning techniques to formulate an optimal mathematical model. An analysis of the significance of selected features was conducted, allowing for the reduction of the factor set to those that do not degrade the dynamic characteristics of the model, including accuracy, responsiveness, and stability. As a result, a risk prediction algorithm for endometriosis based on logistic regression was developed, incorporating 30 significant features. The effectiveness of the developed model was evaluated using standard metrics such as accuracy, sensitivity, specificity, F1-score, and the area under the ROC curve. The best results were achieved with an AUC value of 0.950, indicating a high predictive ability of the model.

Keywords: endometriosis, non-invasive diagnostics, machine learning, forecasting, logistic regression.

Русинова Анастасия Константиновна
ФГБОУ ВО Воронежский государственный медицинский университет им. Н.Н. Бурденко
rusiknastya@mail.ru

Аннотация. Эндометриоз представляет собой широко распространенное патологическое состояние, однако его клиническое проявление и механизмы остаются недостаточно изученными. Временной разрыв между появлением первых симптомов и установлением диагноза иногда составляет более десяти лет, что значительно затрудняет раннюю диагностику и адекватное лечение. На сегодняшний день отсутствует универсальное лечение, способное полностью устранить эндометриоз, что подчеркивает необходимость разработки новых подходов к диагностике. В данном исследовании рассматривается конструирование и отбор факторов, ассоциированных с риском развития эндометриоза, с использованием современных методов машинного обучения для формирования оптимальной математической модели. Проведен анализ значимости выбранных признаков, что позволило сократить набор факторов до тех, которые не ухудшают динамические характеристики модели, включая точность, быстродействие и стабильность. В результате был создан алгоритм прогнозирования риска эндометриоза на основе логистической регрессии, который включает 30 значимых признаков. Эффективность разработанной модели была оценена с помощью стандартных метрик, таких как точность, чувствительность, специфичность, F1-score и площадь под ROC-кривой. Наилучшие результаты достигнуты с показателем AUC, равным 0,950, что свидетельствует о высокой прогностической способности модели.

Ключевые слова: эндометриоз, неинвазивная диагностика, машинное обучение, прогнозирование, логистическая регрессия.

Введение

Одной из значительных проблем, связанных с диагностикой эндометриоза, является разнообразие его клинических проявлений. Это разнообразие приводит к тому, что от появления первых симптомов до окончательного подтверждения диагноза, которое происходит только после гистологического анализа, может пройти около десяти лет. Важно отметить, что в современных клинических рекомендациях отсутствует эффективный неинвазивный метод для подтверждения диагноза эндометриоза. Данная патология по-прежнему представляет собой серьезную и недостаточно изученную проблему общественного здравоохранения [1]. По статистическим данным, эндометриоз поражает примерно 10 % женщин репродуктивного возраста и затрагивает около 20–25 % женщин, перенесших операции по поводу бесплодия или тазовой боли [2, 3]. Симптомы данного заболевания существенно ухудшают качество жизни многих женщин, что делает проблему особенно актуальной [4–6, 8].

Симптоматическое лечение эндометриоза варьируется от консервативных методов, таких как занятия йогой и прием комбинированных оральных контрацептивов, до более радикальных хирургических вмешательств, таких как резекция яичника и цистэктомия [7]. Однако, эффективных препаратов для лечения эндометриоза на сегодняшний день не существует. Таким образом, перед медицинским сообществом стоит задача совершенствования методов диагностики эндометриоза с целью сокращения времени, необходимого для установления точного диагноза [9]. Это позволит своевременно начать симптоматическое лечение и предотвратить дальнейшее прогрессирование заболевания, что, в свою очередь, может значительно улучшить качество жизни пациенток и повысить шанс на зачатие.

Современные информационные технологии и методы машинного обучения предоставляют новые возможности для решения данной задачи. Их интеграция с клиническими и экспериментальными методами позволяет разработать математическую модель, которая будет прогнозировать риск развития эндометриоза на основе анкетного опроса пациенток. Важно отметить, что создание такой модели включает в себя ключевой этап — построение и отбор признаков, определяющих качественные характеристики диагностики и предсказательной способности модели. Это может стать основой для более ранней и точной диагностики эндометриоза, что несомненно окажет положительное влияние на их качество жизни.

Материалы и методы

В целях оценки риска возникновения и раннего прогнозирования эндометриоза среди женщин разрабатывается прогностическая модель, основанная на анализе клинических признаков, полученных из стандартизированного вопросника. Данный вопросник был адаптирован в соответствии с актуальными руководящими принципами в области акушерства и гинекологии. Использование современных алгоритмов машинного обучения для обработки многомерных данных позволяет эффективно идентифицировать наборы признаков, характеризующие различные классы заболеваний, что усиливает теоретическую и практическую значимость проекта.

Прогностические модели разрабатываются в рамках обучения с учителем, где выявляются закономерности, служащие прогностическими маркерами для оценки риска эндометриоза, его прогрессирования и рецидивов. Классические метки для обучения модели включают данные, полученные из гистологических исследований, а также ответов респондентов, обозначающих свое состояние как практически здоровое.

Процесс машинного обучения, описанный в исследовании, охватывает этапы построения и отбора признаков, которые играют ключевую роль в диагностике эндометриоза. Как показывает ряд исследований, количество таких признаков может достигать 200, что подчеркивает важность правильного их формирования. Построение признаков влияет на производительность прогностических моделей, и без тщательной проработки этого этапа невозможно осуществить последующий отбор признаков, что также критически важно для достижения высоких показателей точности и эффективности машинного обучения.

В практическом плане процесс построения признаков является многоступенчатым и включает реше-

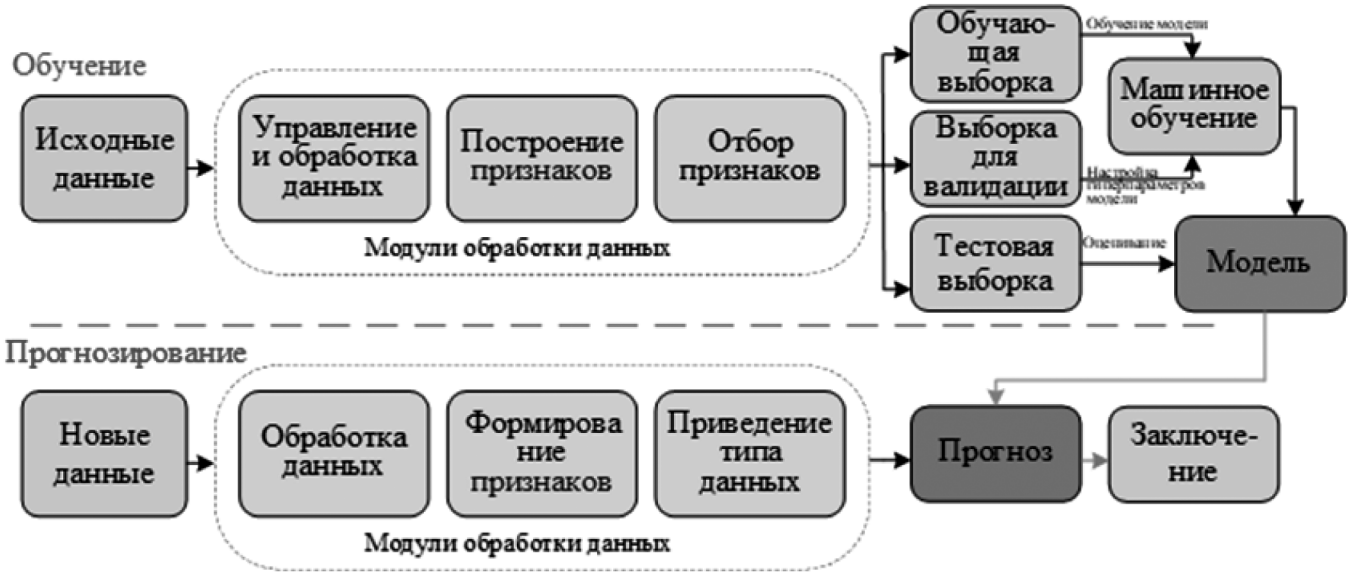

Рис.1. Процесс машинного обучения, используемый при разработке и обучении модели оценки риска при неинвазивной диагностике эндометриоза

ние задач, таких как обработка пропущенных данных, обнаружение выбросов, а также применение методов кодирования и масштабирования данных. Это может включать, например, использование метода One-hot encoding для кодирования категориальных переменных и нормализацию для унификации диапазонов числовых данных. Также на данный процесс оказывают влияние параметры, такие как структура и объем данных, доступные вычислительные ресурсы, а также специфика предметной области.

Методы отбора признаков формируют подмножество исходных данных, которые используются для обучения модели. При этом могут применяться методы фильтрации, оберточные и встроенные методы. Методы фильтрации, такие как критерий хи-квадрат и критерий Фишера, позволяют оценивать значимость и взаимосвязь между переменными. Оберточные методы, например, позволяют оптимизировать отбор, анализируя важность признаков, а встроенные методы, такие как Lasso-регуляризация, помогают одновременно выполнять выбор признаков и обучение модели.

В рамках исследования также была разработана анкета для медико-социологического опроса, включающая около шестидесяти признаков, ассоциированных с эндометриозом, многие из которых были подробно обсуждены в научной литературе [9–11]. Ключевое внимание уделено формализации болевых симптомов, которые являются субъективными и зависят от множества факторов, включая характер патологии, психологическое состояние и жизненный опыт пациентки.

Для оценки болевых симптомов в анкете применяются различные шкалы, такие как Лидсская шкала оценки нейропатических симптомов (LANSS) и опросник DN4. Визуальная аналоговая шкала применяется для определения интенсивности боли, позволяя пациенткам обозначать уровень своего discomfort на 10-сантиметровой линии. Многомерные оценки болевых симптомов могут быть также проведены с использованием опросника Мак-Гилла [12].

Данная анкета предполагает, что участницы медико-социологического опроса будут отвечать на вопросы о наличии тех или иных симптомов в течение последнего месяца. В исследовании задействованы современные методы машинного обучения, включая линейную, множественную и логистическую регрессию, а также деревья решений, что призвано повысить точность и надежность прогнозирующих моделей в контексте диагностики эндометриоза.

Результаты и их обсуждение

В целях прогнозирования вероятности развития эндометриоза была создана обучающая выборка данных, включающая две группы пациентов. Первая группа составлена из пациентов, у которых установлен диагноз эндометриоз, основанный на результатах предыдущего лечения или клинических обследований, подтверждающих наличие глубокого эндометриоза. Вторая, контрольная группа, включает пациентов с одним симптомом, указывающим на возможность эндометриоза, однако без предварительного лечения или клинических исследований, подтверждающих диагноз глубокого эндометриоза. Обучающая выборка данных включает три типа данных: числовые, категориальные и текстовые.

Формирование выборки позволило оценить генеральную совокупность по таким параметрам, как пол, возраст и предрасположенность к эндометриозу. Выборка состоит из женщин в возрасте от 18 до 45 лет (со средним возрастом 28 ± 9 лет), при этом объем выборочной совокупности в 393 пациента обеспечивает репрезентативность с точностью 95 % и погрешностью ± 5 %. Все участники исследования подписали информированное согласие.

Медико-социологический опрос завершили 393 респондента, из которых 202 имели диагностированный эндометриоз, а 191 считали себя здоровыми, не проходя диагностические исследования. Учитывается вероятность, что часть недиагностированных женщин может страдать от эндометриоза, что потенциально влияет на разработанную модель, в результате чего они могут попасть в группу ложноотрицательных результатов. Однако, учитывая, что распространенность эндометриоза в популяции оценивается в пределах 5–10 %, предполагается, что такое смещение будет относительно незначительным.

На основании описательной статистики и значимости признаков, полученных с применением методов машинного обучения, был проанализирован вклад каждого признака в способность модели корректно классифицировать ответы респондентов, а также исследована корреляция между признаками. Высокая корреляция может сигнализировать о избыточности признака, что было визуализировано на тепловой карте корреляции между каждой парой значений признаков, где отсутствие одноцветных строк и столбцов указывает на коррелирование отобранных признаков между собой.

В процессе формирования обучающей выборки отбор признаков был осуществлен с применением селектора голосования, в который интегрированы три метода: метод фильтрации, основанный на корреляции Пирсона; метод обучения без учителя, использующий мультиколлинеарность; и метод рекурсивного исключения признаков. Селекция признаков проводилась следующим образом: результаты каждого из трех методов фиксировались в формате бинарных значений, где 1 указывало на необходимость сохранения признака, а 0 — на его ис-

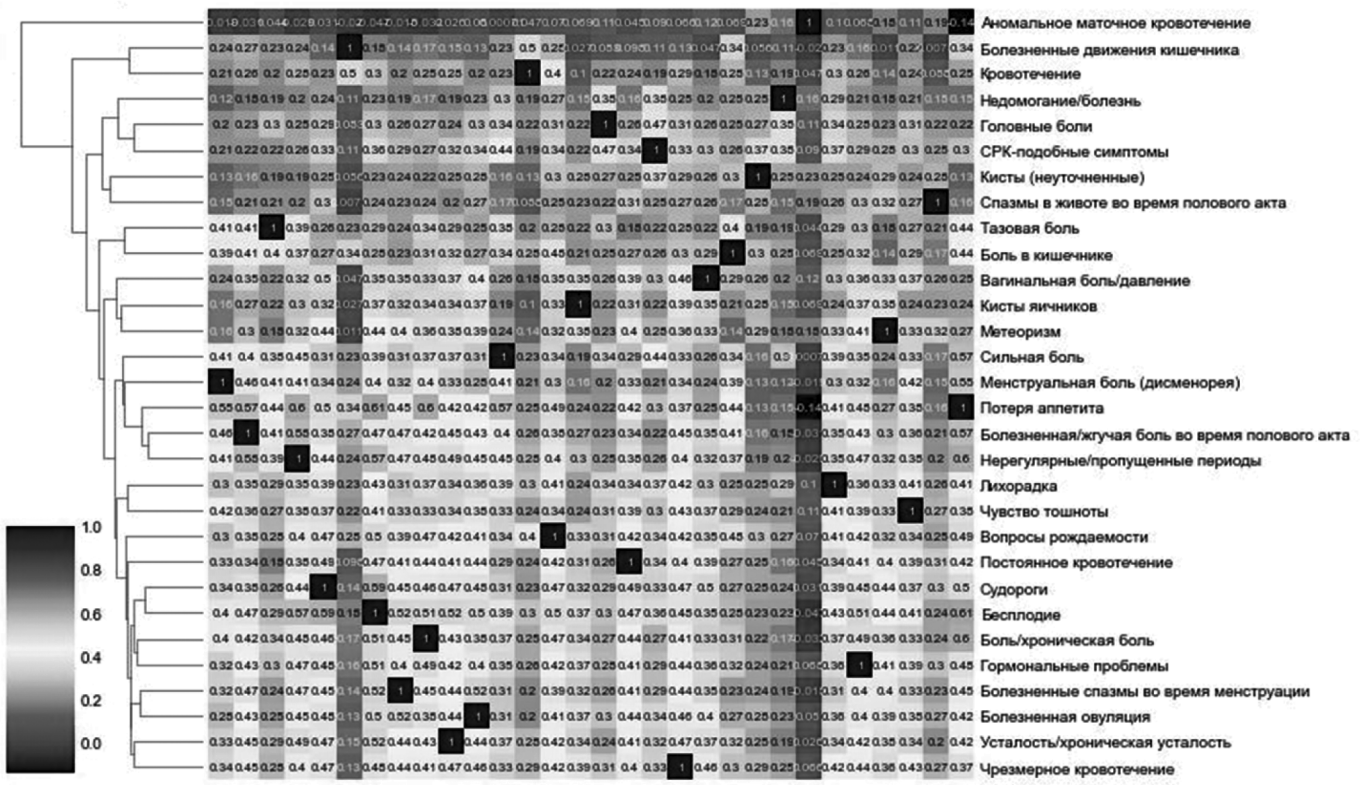

Рис. 2. Тепловая карта корреляции признаков эндометриоза, сгруппированных с помощью метода кластеризации

ключение. Затем вычислялось среднее значение голосов по каждому признаку. Если среднее значение превышало установленные 0,5, что означало, что два из трех методов поддержали сохранение признака, он оставлялся в наборе. В результате был сформирован подмножество из 30 признаков, обеспечивающих оптимальные характеристики производительности модели машинного обучения. Для подтверждения необходимости исключения каких-либо дополнительных признаков проводился итеративный процесс удаления каждого признака с последующим переобучением и тестированием модели, где во всех случаях показатели производительности демонстрировали ухудшение.

Среди отобранных признаков, входящих в наиболее производительную модель, представленных в порядке убывания важности, были такие, как тяжелое или сильное менструальное кровотечение, нерегулярные или пропущенные менструации, аномальное маточное кровотечение, менструальная боль (дисменорея), болезненные ощущения при движении кишечника, боль в кишечнике, тазовая боль, симптомы, напоминающие синдром раздраженного кишечника, болезненные спазмы во время менструации и другие. На основе выбранных признаков была разработана модель логистической регрессии с использованием методов машинного обучения.

Endometriosis
0 191
1 202
dtype: int64
Менструальная боль (дисменорея) ... Endometriosis
1 1 ... 1
1 1 ... 1
0 1 ... 1
1 0 ... 1
0 1 ... 1
[5 rows x 31 columns]
Точность DecisionTreeClassifier: 0.8764044943820225
Матрица ошибок
[[72 13]
[9 84]]
precision recall f1-score support
0 0.89 0.85 0.87 85
1 0.87 0.90 0.88 93
accuracy 0.88 178
macro avg 0.88 0.88 0.88 178
weighted avg 0.88 0.88 0.88 178

Среднее абсолютное отклонение: 0.12359550561797752
Среднеквадратичная ошибка: 0.12359550561797752
Корень из среднеквадратичной ошибки: 0.3515615246553262

Коэффициенты регрессии: [-3.49543333] [[1.45885081 0.76242693 0.92791622 1.23194836 0.70039648 1.460676
1.15791625 1.13007877 0.60017117 1.71682529 -0.91069957 0.79625285

1.60343103 -0.85365601 1.66803964 -0.35314205 -1.65131403 1.4674266
0.07601714 -0.30986979 -1.32778715 1.03996341 -1.07582961 -1.54173294
-1.53315974 0.75692754 0.70849066 -0.47969551 0.54825219 -1.10146085]]
Точность LogisticRegression: 0.9504504504504504
Важность: 0, Score: 1.45885
Важность: 1, Score: 0.76243
Важность: 2, Score: 0.92792
Важность: 3, Score: 1.23195
Важность: 4, Score: 0.70040
Важность: 5, Score: 1.46068
Важность: 6, Score: 1.15792
Важность: 7, Score: 1.13008
Важность: 8, Score: 0.60017
Важность: 9, Score: 1.71683
Важность: 10, Score: -0.91070
Важность: 11, Score: 0.79625
Важность: 12, Score: 1.60343
Важность: 13, Score: -0.85366
Важность: 14, Score: 1.66804
Важность: 15, Score: -0.35314
Важность: 16, Score: -1.65131
Важность: 17, Score: 1.46743
Важность: 18, Score: 0.07602
Важность: 19, Score: -0.30987
Важность: 20, Score: -1.32779
Важность: 21, Score: 1.03996
Важность: 22, Score: -1.07583
Важность: 23, Score: -1.54173
Важность: 24, Score: -1.53316
Важность: 25, Score: 0.75693
Важность: 26, Score: 0.70849
Важность: 27, Score: -0.47970
Важность: 28, Score: 0.54825
Важность: 29, Score: -1.10146

Эффективность разработанной модели была оценена с применением общепринятых метрик в области машинного обучения, таких как точность, чувствительность, специфичность, точность, F1-score и площадь под ROC-кривой [13].

Для повышения значимости полученных результатов была проведена процедура перекрестной проверки. Хотя множество моделей продемонстрировало высокие показатели производительности, модель логистической регрессии оказалась наиболее эффективной, достигнув значения AUC, равного 0,95.

Данные, полученные от респондентов с подтвержденным диагнозом «эндометриоз», стали основой для разработки прогностической модели раннего выявления эндометриоза у женщин, основанной на признаках, с использованием различных методов машинного об-

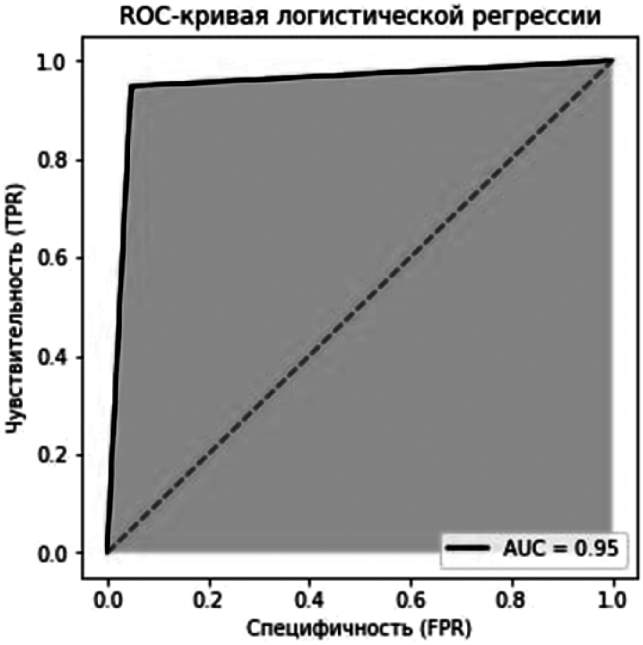

Рис. 3. ROC-кривая модели логистической регрессии для прогнозирования риска развития эндометриоза

учения. На основе многомерных данных, подвергнутых анализу с помощью алгоритмов машинного обучения, был выделен оптимальный набор признаков, эффективно характеризующий разнообразные состояния и позволяющий идентифицировать классы заболеваний. В ходе применения методов машинного обучения были определены закономерности, обладающие прогностическим значением, что дало возможность оценить риск заболевания, его прогрессирования и рецидива. Достоверность созданной прогностической модели была проверена путем сопоставления с гистопатологическими данными, использовавшимися в качестве меток класса в процессе машинного обучения.

Кроме формирования прогнозов и генерации заключений, указанные модели также демонстрируют значимость каждого признака, что позволяет выявлять и исключать неинформативные признаки из будущих медико-социологических опросов.

Следует подчеркнуть, что в обучающую выборку не были включены данные о социально-демографическом статусе респондентов, такие как возраст, образование, статус в браке, место проживания, индекс массы тела, результаты физикального обследования и наличие сопутствующих заболеваний, что означает, что в разрабатываемых моделях эти переменные не учитывались.

Заключение

В рамках исследования была решена задача построения и отбора признаков для оценки состояния пациентов с подтвержденным диагнозом эндометриоз. В про-

цессе работы был применен селектор, использующий несколько методов анализа важности признаков, что позволило получить наиболее эффективную модель на основе подмножества из тридцати признаков.

Разработанная модель прогнозирования эндометриоза основана на данных, полученных от респондентов посредством заполнения анкеты, отражающей самооценки их состояния здоровья. Модель логистической регрессии продемонстрировала наилучшие результаты, достигнув значений AUC=0.95, точности и F1-score=0.94, а также чувствительности=0.93 и специфичности=0.95.

На основании полученных результатов показана целесообразность применения методов машинного обучения для формирования прогностических моделей, оценивающих риск развития эндометриоза у женщин, и обозначена возможность создания рекомендательного блока, основывающегося на этих моделях, в рамках подсистемы неинвазивной диагностики. Разработанная модель может быть использована женщинами, испытывающими симптомы, на начальных этапах их диагностического обследования для оценки вероятности того, что их симптомы могут быть вызваны эндометриозом.

В результате создания и отбора признаков была сформирована модель машинного обучения, способная прогнозировать эндометриоз с высокой точностью, достигающей 95 %, основываясь на подмножестве из тридцати самооценочных симптомов. Ожидается, что данная модель значительно сократит время, необходимое для постановки диагноза, которое в настоящее время составляет от 6 до 10 лет с момента появления симптомов [14]. Более того, модель реализована в виде веб-приложения, позволяющего женщинам провести самодиагностику и получить оценки вероятности наличия эндометриоза. Одной из рекомендаций, предоставляемых данным приложением, является направление женщин на диагностическое обследование на наличие эндометриоза.

ЛИТЕРАТУРА

1. Benagiano G., Brosens I., Lippi D. The history of endometriosis. Gynecol Obstet Invest. 2014, 78: 1–9.
2. Wheeler J.M. Epidemiology of endometriosis-associated infertility. J Reprod Med. 1989, 34: 41–6.
3. Eskenazi B., Warner M.L. Epidemiology of endometriosis. Obstet Gynecol Clin North Am. 1997, 24: 235–58.
4. Nnoaham K.E., Hummelshoj L., Webster P., et al. Impact of endometriosis on quality of life and work productivity: a multicenter study across ten countries. Fertility Sterility. 2011, 96(2): 366–73. e8.
5. Dunselman G.A.J., Vermeulen N., Becker C., et al. ESHRE guideline: management of women with endometriosis. Human Reproduction. 2014, 29: 400–12.
6. Andres M.P., Borrelli, G.M., Abrão, M.S. Endometriosis classification according to pain symptoms: can the ASRM classification be improved? Best Practice & Research Clinical Obstetrics & Gynaecology. 2018, 51: 111–118.
7. Koga K., Takamura M., Fujii T, Osuga Y. Prevention of the recurrence of symptom and lesions after conservative surgery for endometriosis. Fertility Sterility. 2015, 104: 793–801.
8. Johnson N.P., Hummelshoj L., Adamson G.D., et al. World endometriosis society consensus on the classification of endometriosis. Human Reproduction. 2017, 32: 315–24.
9. Fauconnier A. et al. Early identification of women with endometriosis by means of a simple patient-completed questionnaire screening tool: A diagnostic study. Fertility Sterility. 2021, 116: 1580–1589.
10. Eskenazi B. et al. Validation study of nonsurgical diagnosis of endometriosis. Fertility Sterility. 2001, 76: 929–935.
11. Chapron C. et al. A new validated screening method for endometriosis diagnosis based on patient questionnaires. eClinicalMedicine. 2022, 44: 101263.
12. Абрамович С.Г. Физиотерапия боли: учеб. пособие / С.Г. Абрамович Иркутск: РИО ИГМАПО. 2020, 72.
13. Лимановская О.В. Основы машинного обучения: учебное пособие / О.В. Лимановская, Т.И. Алферьева; Мин-во науки и высш. образования РФ. Екатеринбург: Изд-во Урал. ун-та. 2020, 88.
14. Линде В.А. Эндометриозы. Патогенез, клиническая картина, диагностика и лечение / Линде В.А., Татарова Н.А. — М.: ГЭОТАР-Медиа. 2010, 192.

Журнал «Современная наука: актуальные проблемы теории и практики»