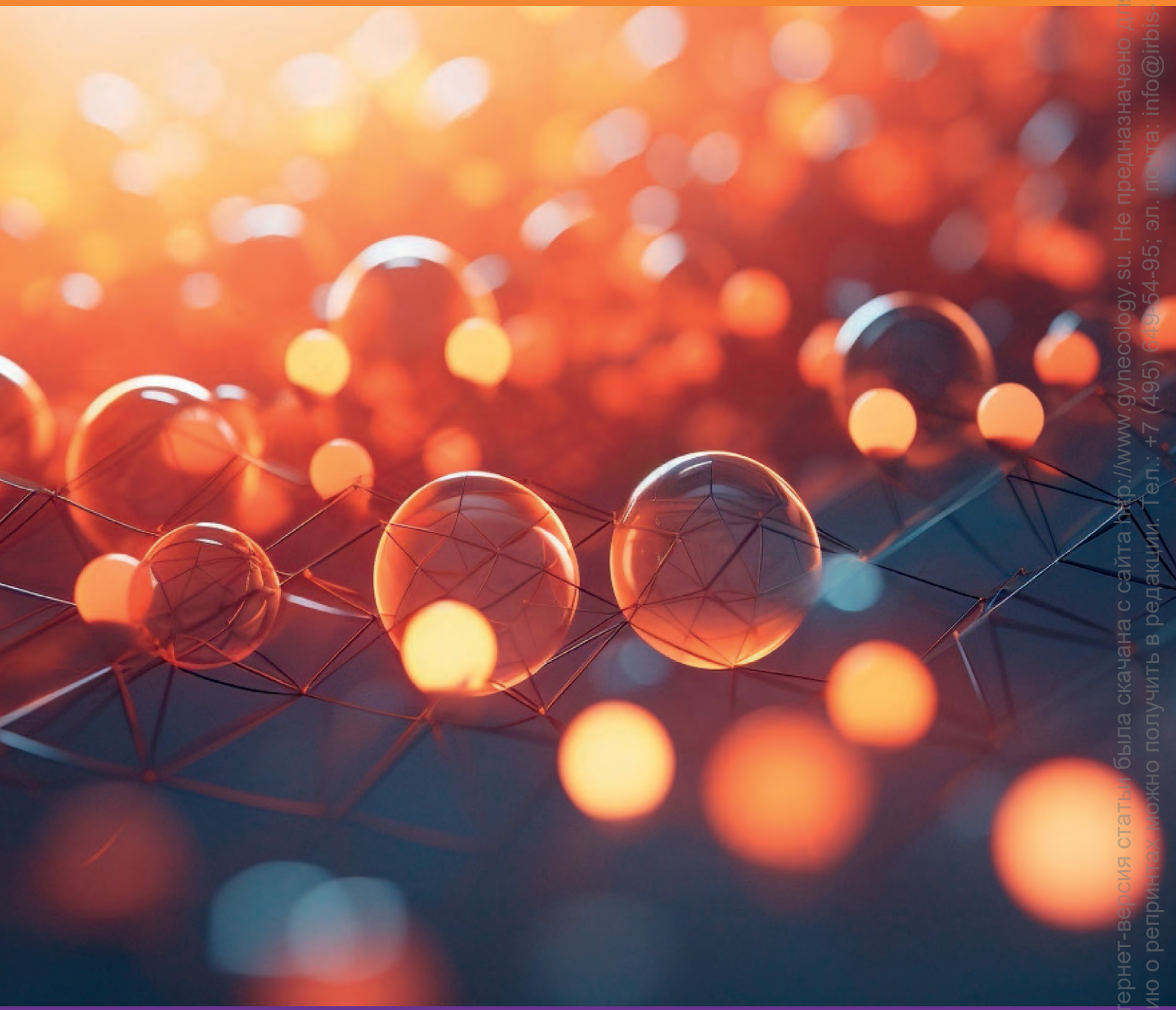
ISSN 2313-7347 (print)
ISSN 2500-3194 (online)

АКУШЕРСТВО ГИНЕКОЛОГИЯ РЕПРОДУКЦИЯ

Включен в перечень ведущих рецензируемых журналов и изданий ВАК

2026 • том 20 • № 1



OBSTETRICS, GYNECOLOGY AND REPRODUCTION

2026 Vol. 20 No 1

https://gynecology.su

ISSN 2313-7347 (print)
ISSN 2500-3194 (online)

2026 • Vol. 20 • № 1
Obstetrics, Gynecology and Reproduction

https://doi.org/10.17749/2313-7347/ob.gyn.rep.2025.701

От данных к прогнозу: разработка и клиническая апробация инструмента оценки риска преждевременных родов на основе технологий машинного обучения

Ю.С. Болдина, А.А. Ившин, К.С. Светова

ФГБОУ ВО «Петрозаводский государственный университет»; Россия, 185910 Петрозаводск, проспект Ленина, д. 33

Для контактов: *Александр Анатольевич Ившин, e-mail: scipeople@mail.ru*

Резюме

Введение. Преждевременные роды (ПР) остаются одним из наиболее серьезных осложнений беременности, выступают основной причиной неонатальной смертности и влекут за собой такие тяжелые последствия, как инвалидизация и развитие хронических заболеваний у новорожденных, а также приводят к значительным социально-экономическим издержкам. Глобальная частота ПР остается практически неизменной и составляет 5–18 %, несмотря на применяемые профилактические меры, что подчеркивает необходимость создания более эффективных инструментов прогнозирования для своевременной профилактики.

Цель: разработка и валидация на независимой выборке инструмента оценки риска ПР, основанного на технологиях машинного обучения (англ. Machine Learning, ML) и реальных клинических данных, полученных из электронных медицинских карт (ЭМК) беременных.

Материалы и методы. В работе использовался массив из 10000 анонимизированных записей ЭМК, содержащих 54 признака, включая анамнестические, клинические, лабораторные и инструментальные данные. Прогностическая система состояла из двух взаимосвязанных моделей ML: NLP-модели (обработка естественного языка; англ. Natural Language Processing, NLP) с использованием модели RuBERT (англ. Russian Bidirectional Encoder Representations from Transformers; предварительно обученная языковая модель для обработки русскоязычных текстов) для извлечения признаков ПР из неструктурированных русскоязычных текстов и предиктивной модели ML, для создания которой было протестировано 14 различных алгоритмов.

Результаты. NLP-модель показала высокое качество обработки данных с медианной чувствительностью = 0,998, F-мерой (гармоническое среднее между точностью и полнотой) = 0,976 и AUC-ROC = 0,974. Среди алгоритмов ML наилучшие результаты оценки риска продемонстрировал алгоритм на основе градиентного бустинга – CatBoost Classifier (англ. Categorical Boosting Classifier) с точностью (accuracy) = 0,81, чувствительностью = 0,87, точностью (precision) = 0,76, F-мерой = 0,81 и AUC-ROC = 0,82.

Заключение. Разработанная модель показала производительность, сопоставимую с зарубежными аналогами, а валидация подтвердила ее устойчивость к новым данным, что свидетельствует о перспективности системы для использования в реальной клинической практике. Данное исследование представляет собой первый этап создания комплексного решения для оценки риска ПР, объединяющего NLP и ML. Дальнейшее совершенствование разработанного алгоритма оценки может включать использование дополнительных признаков (например, биохимических маркеров) и проведение многоцентровых валидационных исследований.

Ключевые слова: преждевременные роды, ПР, оценка риска, машинное обучение, ML, обработка естественного языка, NLP, электронные медицинские карты, ЭМК

Для цитирования: Болдина Ю.С., Ившин А.А., Светова К.С. От данных к прогнозу: разработка и клиническая апробация инструмента оценки риска преждевременных родов на основе технологий машинного обучения. *Акушерство, Гинекология и Репродукция.* 2026;20(1):15–33. https://doi.org/10.17749/2313-7347/ob.gyn.rep.2025.701.

2026 • Том 20 • № 1
Акушерство, Гинекология и Репродукция

From data to prediction: development and clinical validation of a preterm birth risk assessment tool based on machine learning technologies

Yulia S. Boldina, Aleksandr A. Ivshin, Kristina S. Svetova

Petrozavodsk State University; 33 Lenin Avenue, Petrozavodsk 185910, Russia

Corresponding author: *Aleksandr A. Ivshin, e-mail: scipeople@mail.ru*

Abstract

Introduction. Preterm birth (PTB) remains one of the most serious complications of pregnancy, being the leading cause of neonatal mortality and contributing to long-term disability along with chronic morbidity in newborns, as well as imposing substantial socioeconomic costs. Despite preventive efforts, the global PTB rate has remained largely unchanged comprising 5–18 %, underscoring a need for developing more effective prediction tools to enable timely prevention.

Aim: using an independent sample to develop and validate a PTB risk-assessment tool based on machine learning (ML) and routinely collected clinical data retrieved from electronic health records (EHRs) of pregnant patients.

Materials and Methods. We analyzed a dataset of 10,000 de-identified EHRs entries containing 54 variables, including historical, clinical, laboratory, and instrumental (diagnostic/imaging) data. The predictive system comprised two interconnected ML components: (1) an NLP model based on RuBERT (a pre-trained ML model for processing Russian texts) for extracting PTB-relevant features from unstructured Russian-language clinical text, and (2) a downstream predictive ML model, for which 14 algorithms were benchmarked.

Results. The NLP model demonstrated high performance with a median sensitivity = 0.998, F1-score = 0.976, and AUC-ROC = 0.974. Among the ML algorithms, the algorithm based on gradient boosting – CatBoost Classifier (Categorical Boosting Classifier) achieved the best risk-prediction results: accuracy = 0.81, sensitivity (recall) = 0.87, precision = 0.76, F1-score = 0.81, and AUC-ROC = 0.82.

Conclusion. The developed model showed performance comparable to that of international counterparts, and validation confirmed its robustness to previously unseen data, indicating strong potential for use in routine clinical practice. This study represents the first step toward an integrated PTB risk-assessment solution combining NLP and ML. Future work will include incorporation of additional predictors (e.g., biochemical markers) and multicenter validation studies.

Keywords: preterm birth, PTB, risk assessment, prediction, machine learning, ML, natural language processing, NLP, electronic health records, EHRs

For citation: Boldina Yu.S., Ivshin A.A., Svetova K.S. From data to prediction: development and clinical validation of a preterm birth risk assessment tool based on machine learning technologies. *Akusherstvo, Ginekologia i Reprodukcia = Obstetrics, Gynecology and Reproduction*. 2026;20(1):15–33. (In Russ.). https://doi.org/10.17749/2313-7347/ob.gyn.rep.2025.701.

Введение / Introduction

Несмотря на существенные достижения в акушерской и неонатальной помощи, преждевременные роды (ПР) продолжают оставаться глобальной медико-социальной проблемой. Сохраняющаяся высокая распространенность ПР в мире (5–18 % по данным Всемирной организации здравоохранения) с ежегодным числом случаев, превышающим 13 млн, указывает на недостаточную эффективность существующих профилактических парадигм. Даже в странах с развитой системой здравоохранения прослеживаются значительные трудности в снижении частоты ПР, что в полной мере отражается и в отечественной статистике, где показатели варьируют от 4 до 6 %, достигая в специализированных перинатальных центрах более 9 % [1]. ПР влекут серьезные последствия, включая демографические потери, экономическую нагрузку на здравоохранение и психосоциальные последствия на уровне семей, столкнувшихся с данным осложнением. На долю ПР приходится около 1 млн предотвратимых смертей в год, что выводит их на первое место среди причин неонатальной смертности и на второе – в структуре смертности детей до 5 лет [2].

Тяжесть последствий для жизни и здоровья новорожденного варьирует в зависимости от срока гестации. Роды на 22–28-й неделе сопряжены с 98 % летальностью, а у сохранивших жизнь младенцев нередко развивается патология, сопровождающаяся глубокой инвалидизацией. Однако, даже при родах в более поздние сроки (34–37 недель) сохраняется повышенный риск бронхолегочной дисплазии, нарушений развития нерв-

2026 • Vol. 20 • № 1
Obstetrics, Gynecology and Reproduction

Основные моменты

Что уже известно об этой теме?

- Преждевременные роды (ПР) остаются глобальной нерешенной проблемой акушерства, являясь ведущей причиной неонатальной смертности и детской инвалидности. Несмотря на известные факторы риска (инфекции, истмико-цервикальная недостаточность, многоплодие), частота ПР стабильно сохраняется на высоком уровне (5–18 %) по всему миру. Существующие методы профилактики имеют ограниченную эффективность, а мультифакторная природа ПР затрудняет прогнозирование с помощью традиционных статистических методов.
- В последние годы машинное обучение (ML) активно исследуется для прогнозирования акушерских осложнений, демонстрируя высокий потенциал в международных исследованиях. Однако существующие модели часто разработаны для стандартизированных англоязычных наборов данных и не адаптированы к реалиям отечественного здравоохранения.

Что нового дает статья?

- Данное исследование представляет собой первый шаг по разработке комплексного инструмента прогнозирования ПР, адаптированного для работы с русскоязычными электронными медицинскими картами (ЭМК).
- Новизна метода заключается в создании и интеграции специализированной NLP-модели, которая автоматически извлекает 54 клинических признака ПР из неструктурированных текстов врачебных записей. Это устраняет главное препятствие для использования реальных клинических данных в отечественном здравоохранении.
- Сравнительный анализ 14 алгоритмов ML выявил, что алгоритм на основе градиентного бустинга CatBoost Classifier обеспечивает наилучшую производительность, подтвержденную внешней валидацией. Таким образом, в статье описывается технологический конвейер для автоматизированного анализа рутинной медицинской документации с целью оценки риска ПР.

Как это может повлиять на клиническую практику в обозримом будущем?

- Внедрение разработанного инструмента в клиническую практику позволит осуществлять автоматизированный скрининг риска ПР для каждой пациентки на основе данных, уже содержащихся в ее ЭМК. Это позволит врачам использовать объективный инструмент поддержки принятия решений для раннего выявления беременных высокого риска и своевременного назначения персонализированных профилактических мероприятий (например, прогестероновой поддержки).
- Интеграция системы в медицинские информационные системы позволит реализовать автоматическое оповещение о беременности высокого риска, что способствует оптимизации маршрутизации и тактики ведения. В перспективе это может привести к снижению частоты ПР и улучшению перинатальных исходов в регионах, применяющих данную технологию.

Highlights

What is already known about this subject?

- Preterm birth (PTB) remains an unresolved global challenge in obstetrics, being the leading cause of neonatal mortality and childhood disability. Despite well-established risk factors (infections, cervical insufficiency, multiple gestation), the global PTB rate has persistently remained high (5–18 %). Existing preventive strategies have limited effectiveness, and the multifactorial nature of PTB complicates prediction using conventional statistical methods.
- In recent years, machine learning (ML) has been actively investigated for predicting obstetric complications, demonstrating strong potential in international studies. However, many existing models were developed on standardized English-language datasets and are not adapted to the realities of the Russia-wide healthcare system.

What are the new findings?

- This study represents an initial step toward a comprehensive PTB prediction tool tailored to Russian-language electronic health records (EHRs).
- The methodological novelty lies in the design and integration of a specialized NLP (natural language processing) model that automatically extracts 54 PTB-related clinical features from unstructured physician notes. This directly addresses a key barrier to leveraging real-world clinical data within domestic healthcare.
- A comparative evaluation of 14 ML algorithms showed that the algorithm based on gradient boosting CatBoost Classifier delivers the best performance, as confirmed by external validation. The article thus describes an end-to-end pipeline for automated analysis of routine clinical documentation to estimate PTB risk.

How might it impact on clinical practice in the foreseeable future?

- Implementing the proposed tool in routine medical care would enable automated PTB risk screening for every female patient based on data already contained in her EHR. Clinicians would gain an objective decision-support instrument for early identification of high-risk pregnancies and timely initiation of personalized preventive measures (e.g., progesterone therapy).
- Integrating the system into medical information systems would allow automated high-risk alerts, helping optimize care pathways and management strategies. Over time, broader adoption could contribute to reducing PTB rates and improving perinatal outcomes in geographic regions deploying this technology.

ной системы и длительных хронических заболеваний [3, 4]. Социально-экономические издержки включают не только прямые затраты на интенсивную терапию и реабилитацию, но и отдаленные медицинские последствия, существенно влияющие на качество жизни. Помимо прочего, ПР прямо коррелируют с хроническими заболеваниями, такими как сердечно-сосудистые заболевания, сахарный диабет и патология почек [3, 5, 6].

До 70 % ПР развиваются спонтанно, пусковые механизмы во многих случаях остаются неясными, несмо-

тря на изученные предикторы, включая инфекционный фактор [5], истмико-цервикальную недостаточность и социально-демографические детерминанты [6–8]. Кроме того, определена роль маркеров развития преждевременного разрыва плодных оболочек (ПРПО) и ПР, например, плацентарного альфа-микроглобулина-1 (ПАМГ-1) [9, 10].

Современные подходы к ведению угрожающих ПР включают 2 ключевых элемента: временный токолиз (48 часов) для обеспечения маршрутизации пациентки и антенатальная профилактика респираторного дистресс-синдрома (РДС) глюкокортикостероидами. Тем не менее эффективность указанных мер для значимого пролонгирования беременности остается спорной [11]. Эффективность профилактики ПР также остается дискуссионной [12], в то же время ряд исследований подчеркивают целесообразность применения микронизированного прогестерона [13], акушерского разгружающего пессария и серкляжа [1, 14] у пациенток из группы риска по досрочным родам.

Преждевременные роды представляют собой следствие сложного взаимодействия множества факторов, уникального для каждой пациентки. Эта гетерогенность диктует необходимость перехода от универсальных подходов к персонализированным, основанным на интеграции широкого спектра клинических данных. В русле данной парадигмы современная медицина все чаще прибегает к использованию методов искусственного интеллекта (англ. Artificial Intelligence, AI), среди которых машинное обучение (англ. Machine Learning, ML) занимает ключевые позиции.

Машинное обучение открывает новые возможности для оценки риска ПР, преодолевая ограничения традиционной статистики. Его достоинство заключается в работе с большими данными, что позволяет выявлять скрытые предикторы. ML-модели способны к раннему и комплексному прогнозированию, одновременно анализируя медицинский анамнез, клиническую картину и факторы образа жизни. Особо следует отметить их способность к интеграции совершенно разнородной информации – от структурированных показателей до медицинских изображений и текстовых записей. Эта комплексность в совокупности с возможностью постоянного самообучения и совершенствования делает ML надежным инструментом прогнозирования, точность и клиническая значимость которого со временем только возрастают.

В последние годы ML-алгоритмы показали высокую эффективность в прогнозировании таких акушерских осложнений, как задержка роста плода [15], послеродовое кровотечение [16], преэклампсия (ПЭ) [17]. Отдельного внимания заслуживают отечественные разработки в сфере акушерства и гинекологии с применением ML-технологий. В исследовании А.Е. Андрейченко с соавт. (2023) разработаны и валидированы модели прогнозирования ПЭ и ее ранних форм на основе данных, полученных в I триместре беременности [18].

Демонстрируют потенциал результаты зарубежных исследований по разработке мультипараметрических моделей для оценки риска ПР на основе алгоритмов ML. Например, в недавнем исследовании Y. Chen с соавт. (2024) алгоритм XGBoost (англ. Extreme Gradient Boosting; библиотека машинного обучения, реализующая алгоритм градиентного бустинга) показал высокую точность прогнозирования спонтанных ПР (AUC = 0,89; 95 % доверительный интервал (ДИ) = 0,88–0,90), выявив 10 ключевых предикторов ПР, включая биохимические маркеры [19]. Y. Zhang с соавт. (2023) подтвердили перспективность использования алгоритма AdaBoost (точность 95,4 %, AUC = 0,93), выявив в качестве основных факторов риска ПР многоплодие, ПРПО, предлежание плаценты и дородовое кровотечение [20]. Алгоритм RF (англ. Random Forest; случайный лес) в исследовании Q. Sun с соавт. (2022), показал AUC = 0,885 (95 % ДИ = 0,873–0,897), используя для прогнозирования ПР клинико-биохимические параметры и данные 9550 беременных [21].

Весьма вариабельное качество существующих прогностических моделей и ограниченность внедрения их в клиническую практику подчеркивают необходимость дальнейших поисков надежных практико-ориентированных инструментов оценки риска ПР.

Цель: разработка и валидация на независимой выборке инструмента оценки риска ПР, основанного на технологиях ML и реальных клинических данных, полученных из электронных медицинских карт (ЭМК) беременных.

Материалы и методы / Materials and Methods

Источник данных / Data source

Проведено ретроспективное когортное исследование с анализом 10000 обезличенных неструктурированных записей ЭМК о пренатальном наблюдении женщин. Персональные данные были предварительно анонимизированы до начала анализа, что исключило необходимость получения информированного согласия и обеспечило соответствие правовым и этическим нормам защиты персональных данных.

Для обработки текстовой информации и извлечения признаков, связанных с ПР, использовалась специально разработанная авторами модель обработки естественного языка (англ. Natural Language Processing, NLP). Модель позволила идентифицировать и структурировать 54 фактора риска, представленных количественными (лабораторные, инструментальные) и категориальными (включая бинарные типа «да»/«нет») переменными. Детали создания NLP-модели представлены в соответствующем разделе.

2026 • Vol. 20 • № 1
Obstetrics, Gynecology and Reproduction

Дизайн исследования / Study design

В ретроспективный анализ исходно были включены 90046 обезличенных медицинских записей (по 10 атрибутов каждая) в формате JSON (англ. JavaScript Object Notation), сформированных из ЭМК пациенток за период с марта 2011 г. по июль 2020 г. JSON – это открытый стандартный файловый формат и формат обмена данными, который использует удобочитаемый текст для передачи объектов данных, состоящих из пар «атрибут-значение» и массивов (или других серий значений). Исходная документация включала карты планового дородового наблюдения, истории родов и завершенных случаев госпитализации, индивидуальные карты ведения беременности, а также записи о медицинских осмотрах и карты прерывания беременности.

Единицей анализа служила медицинская запись, отражающая документированный случай оказания медицинской помощи беременной с указанием срока гестации и актуальных клинико-лабораторных показателей на момент визита. Отбор данных осуществлялся по следующим принципам: 1) подтвержденный факт беременности согласно кодам МКБ-10; 2) сведения об исходе беременности (по соответствующим кодам МКБ-10 в ЭМК).

В результате отбора был сформирован датасет, включавший 10000 записей (n = 10000) ЭМК, разделенный в последующем на 2 равные части для обучения NLP-модели и автоматической обработки NLP-моделью с целью создания обучающей выборки для модели оценки риска ПР.

Валидационная выборка включала 500 уникальных случаев (m = 500) беременности с ПР, полученная на основе анкетирования медицинских карт пациенток ГБУЗ «Республиканский перинатальный центр имени Гуткина К.А.» (ГБУЗ РПЦ им. Гуткина К.А.) за 2016–2022 гг. Схематическое представление дизайна исследования приведено на **рисунке 1**.

Критерии включения и исключения / Inclusion and exclusion criteria

В группу с целевым событием включались все медицинские записи пациенток, у которых в течение текущей беременности был диагностирован либо статус «угрожающие преждевременные роды», либо непосредственно «преждевременные роды» по соответствующим кодам МКБ-10 (O47.0, 47.9, O60). Критериями исключения стали остальные записи без соответствующих кодов МКБ-10, они формировали контрольную группу.

Перечень кодов МКБ-10, использованных для идентификации участников, определения исходов беременности и оценки основного исхода, представлен в **таблице 1**.

Предикторы / Predictors

В анализ был включен широкий спектр признаков (54 параметра), доступных в ЭМК в рамках рутинного клинического наблюдения. Такой подход позволил оценить потенциальную информативность как общепризнанных клинических факторов риска, так и параметров, связь которых с ПР менее очевидна и требует дальнейшего изучения (например, социально-демографические характеристики).

Анамнестические факторы учитывали сопутствующую патологию, особенности репродуктивного анамнеза и факторы образа жизни, способные влиять на исход текущей беременности. В их число вошли: курение, употребление алкоголя и наркотических веществ, семейное и социально-трудовое положение, возраст менархе, нарушения фертильности, включая применение вспомогательных репродуктивных технологий (экстракорпоральное оплодотворение, ЭКО), паритет, отягощенный акушерский анамнез (неразвивающаяся беременность, самопроизвольные и искусственные аборты и ПР в анамнезе), привычное невынашивание беременности, подтвержденные наследственные тромбофилии, антенатальная гибель плода в анамнезе, гинекологические заболевания (доброкачественная и злокачественная патология шейки матки и ее лечение, в том числе конизация; миома матки, эндометриоз, объемные образования яичников, полипы эндометрия), а также перенесенные внутриматочные вмешательства (раздельное диагностическое выскабливание). Из соматической патологии анализировались наличие анемии, тромбоцитопении, сахарного диабета (СД), инфекций мочевыводящих путей (ИМВП) и генитального тракта.

Таблица 1. Перечень кодов МКБ-10, включенных в анализ электронных медицинских карт пациенток на различных этапах формирования базы данных.

Table 1. ICD-10 codes used for the extraction of patient electronic health records at different stages of the database creation.

Группа / Group	Коды МКБ-10 / ICD-10 codes
Случай обращения за медицинской помощью по беременности Seeking medical care for pregnancy	O10-O16*, O20-26*, O28-36*, O40-42*, O43-48*, O88*, O98*, O99*, Z32-36*
Исход беременности / Pregnancy outcome	O60*, O61-75*, O80-87*, O89-92*, O95*, O36.5, P95, Z37-39*
Наличие целевого исхода / Presence of a target outcome	O47.0, 47.9, O60*

Примечание: *коды МКБ-10 приведены с учетом подрубрик.*

Note: *ICD-10 codes include the required level of subclassification.*

2026 • Том 20 • № 1
Акушерство, Гинекология и Репродукция

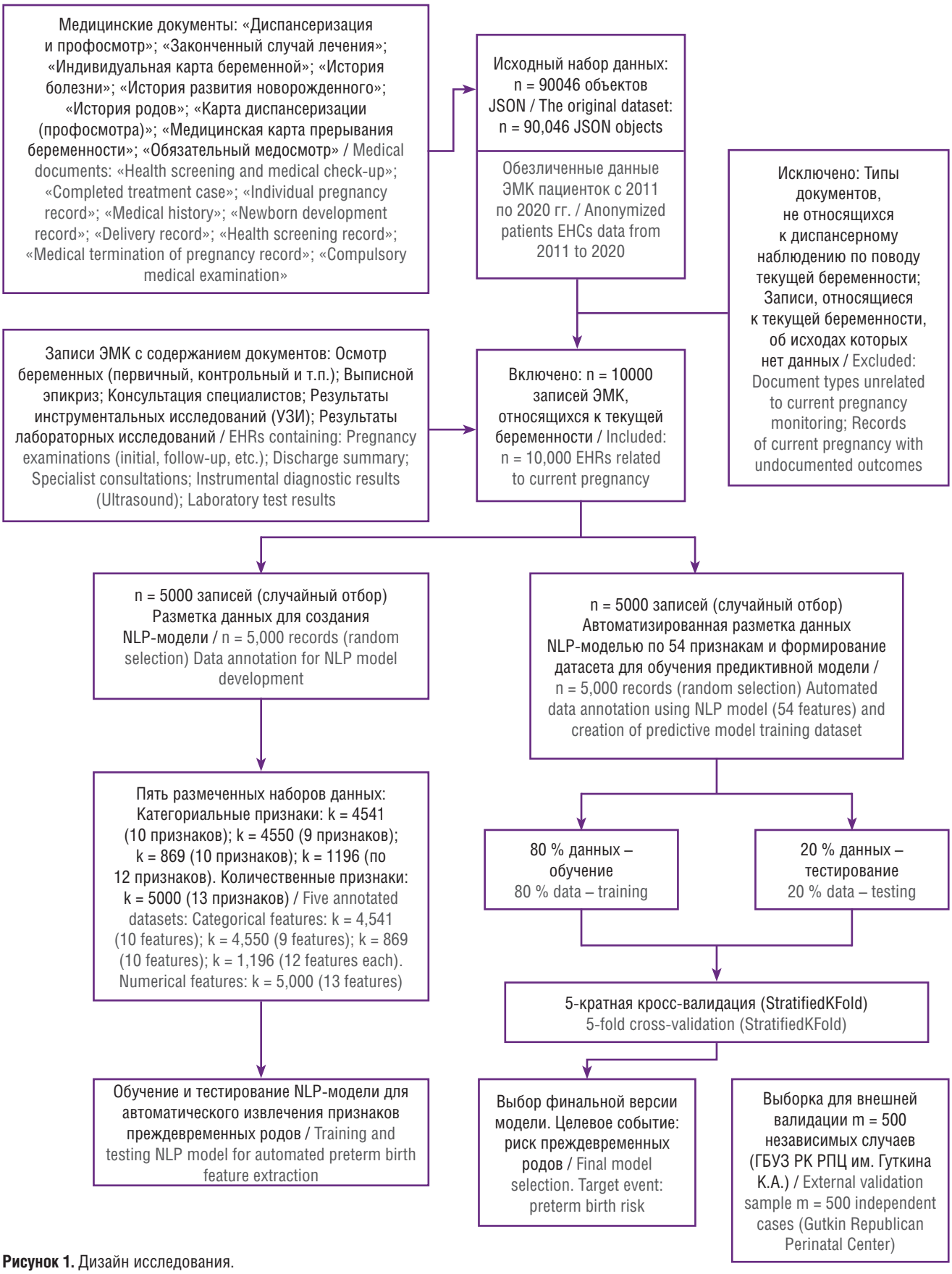


Рисунок 1. Дизайн исследования.

Примечание: *JSON – текстовый формат обмена данными, основанный на JavaScript; ЭМК – электронная медицинская карта; УЗИ – ультразвуковое исследование; NLP – обработка естественного языка.*

Figure 1. Study design.

Note: *JSON – JavaScript Object Notation; EHRs – electronic health records; Ultrasound – ultrasound examination; NLP – Natural Language Processing.*

2026 • Vol. 20 • № 1 Obstetrics, Gynecology and Reproduction

Конституциональные параметры включали возраст матери на момент зачатия, рост и массу тела до беременности, индекс массы тела (ИМТ) до беременности, наличие дефицита или избытка массы тела, ожирение, а также прибавку массы тела при беременности.

Клинические предикторы, характеризующие текущую беременность, учитывали многоплодие, наличие истмико-цервикальной недостаточности (ИЦН), угрозы прерывания беременности или угрозы ПР, применение гестагенов (микронизированный прогестерон или дидрогестерон), коррекцию ИЦН с помощью акушерского пессария или наложения серкляжа, признаки гипоксии плода или задержки его роста, а также нарушения количества околоплодных вод (многоводие или маловодие). Для пациенток с ПР также учитывался гестационный срок на момент родоразрешения.

Инструментальным предиктором, включенным в исследование, являлась длина сомкнутой части шейки матки, измеренная с помощью трансвагинальной цервикометрии.

Лабораторные параметры включали уровни гемоглобина, тромбоцитов и лейкоцитов, а также концентрацию С-реактивного белка (СРБ) в качестве неспецифического маркера воспаления. Также анализировались показатели коагулограммы: протромбиновый индекс (ПТИ), активированное частичное тромбопластиновое время (АЧТВ) и уровень фибриногена.

Предобработка и аннотация данных для разработки модели обработки естественного языка (NLP) / Data preprocessing and annotation for NLP model development

Исходный набор данных, содержащий 10000 записей, был подвергнут предварительной обработке. На этом этапе были удалены нерелевантные поля, и для дальнейшего анализа отобраны только те записи, которые включали описания физикальных осмотров, результаты инструментальной и лабораторной диагностики.

Аннотирование различных типов данных выполнялось с применением специализированных методов и охватывало числовые значения, категориальные признаки и коды МКБ-10. Для работы с категориальными признаками были созданы пользовательские словари, учитывающие ключевые термины, их синонимы и аббревиатуры. Например, термин «курение» мог быть заменен на «употребление табака» или «никотиновая зависимость». Извлечение терминов из текста осуществлялось с помощью регулярных выражений, которые учитывали контекст отрицания, как в случаях «не курит» или «отрицает употребление табака».

Для решения проблемы дисбаланса классов в данных был применен метод SMOTE (англ. Synthetic Minority Oversampling Technique; метод синтетической передискретизации меньшинства класса для несбалансированных данных). Этот метод генерирует синтетические примеры для наименьшего класса, создавая новые случаи между существующими примерами и их ближайшими соседями.

Числовые значения аннотировались путем извлечения паттернов, включающих название параметра, численное значение и соответствующие единицы измерения, например: «гемоглобин» со значением «120 г/л». Коды МКБ-10 извлекались с использованием стандартных правил именования с учетом возможных вариаций в форматировании: например, диагноз «истмико-цервикальная недостаточность» может быть представлен как «O34.3» или «O343». В результате процесса аннотации было создано 5 отдельных наборов данных: 4 набора с категориальными признаками (объемом 4541, 4550, 869 и 1196 записей, с 10, 9, 10 и 12 признаками соответственно; k = 4541, k = 4550, k = 869, k = 1196) и один набор с числовыми признаками, содержащий 5000 записей (k = 5000).

Описание принципа работы NLP-модели / Operating principles of the NLP model

Основой модели обработки естественного языка (NLP) стала предобученная языковая модель RuBERT (англ. Russian Bidirectional Encoder Representations from Transformers; предварительно обученная языковая модель для обработки русскоязычных текстов). Этот алгоритм был выбран благодаря своей специализированной оптимизации для работы с русскоязычными медицинскими текстами, включая надежное распознавание клинических контекстов и специальных аббревиатур, используемых в клинической практике.

Для сохранения целостности сложной медицинской терминологии выполнялась токенизация текста, подразумевающая под собой процесс разбиения на минимальные смысловые единицы (токены). Эти токены затем преобразовывались в 768-мерные векторные представления. В последующем для каждого клинического документа применялось усреднение векторных представлений токенов (mean pooling) с целью создания единого вектора фиксированной длины.

Классификация категориальных признаков реализована с использованием ансамблевого подхода, при котором для каждого признака обучался отдельный бинарный классификатор на основе градиентного бустинга – CatBoost Classifier (англ. Categorical Boosting Classifier; категориальный бустинг-классификатор). Числовые предикторы обрабатывались с помощью гибридной методики, сочетающей:

1. шаблонный анализ (англ. Pattern-based Extraction). Извлечение по шаблонам на основе регулярных выражений (Regex) для точной идентификации числовых значений и их единиц измерения в стандартных формулировках (например, «фибриноген – 3,5 г/л»);

2. контекстуальный анализ (англ. Contextual Analysis) с привлечением модели RuBERT для интерпретации случаев, когда параметры указаны неявно

(например, «фибриноген повышен»), а также для исключения ложных срабатываний путем выявления отрицаний (например, «фибриноген не обнаружен»).

Для обеспечения точности данных каждый извлеченный числовой параметр дополнительно проверялся на соответствие установленным физиологическим диапазонам и общему клиническому контексту. Коды заболеваний МКБ-10 извлекались методом точного строкового сопоставления со стандартной номенклатурой ввиду их строго регламентированного формата. Общая схема процесса обработки текста представлена на **рисунке 2**.

На заключительном этапе проекта NLP-модель была использована для автоматического аннотирования 5000 неструктурированных ЭМК. Результатом работы стал структурированный набор данных, содержащий информацию о каждом пациенте, включая распределение по классам для категориальных признаков, точные числовые значения для количественных предикторов и соответствующие коды МКБ-10 для диагнозов. Полученный в результате аннотирования набор данных был использован для обучения ML-модели, предназначенной для прогнозирования ПР.

Предобработка данных для обучения модели прогнозирования ПР / Data preprocessing for predictive PTB model training

Для повышения качества входных данных и улучшения производительности ML-моделей был реализован комплекс мероприятий по предварительной обработке данных. Этот комплекс включал анализ корреляций, обработку выбросов, импутацию пропущенных значений, нормализацию данных и балансировку классов в обучающем наборе.

В ходе предварительного анализа данных была проведена оценка взаимосвязей между признаками. Анализ выявил слабую линейную зависимость между большинством признаков и целевой переменной, что подтвердило целесообразность использования нелинейных классифицирующих алгоритмов при последующем моделировании.

Пропущенные и аномальные значения обрабатывались следующим образом: для числовых параметров, таких как лабораторные показатели (уровень гемоглобина, количество тромбоцитов и лейкоцитов), пропущенные значения заполнялись медианой соответствующего признака; для бинарных признаков (например, наличие сопутствующих заболеваний) пропуски заполнялись нулевыми значениями, что отражает клинически обоснованное предположение об отсутствии патологии при отсутствии записи данных.

Выбросы идентифицировались с использованием метода межквартильного размаха (IQR). Значения, выходящие за диапазон [Q_1 – 1,5×IQR; Q_3 + 1,5×IQR], считались аномальными. Для клинически значимых параметров это статистическое правило дополнялось применением физиологически обоснованных границ для исключения невозможных значений (например,

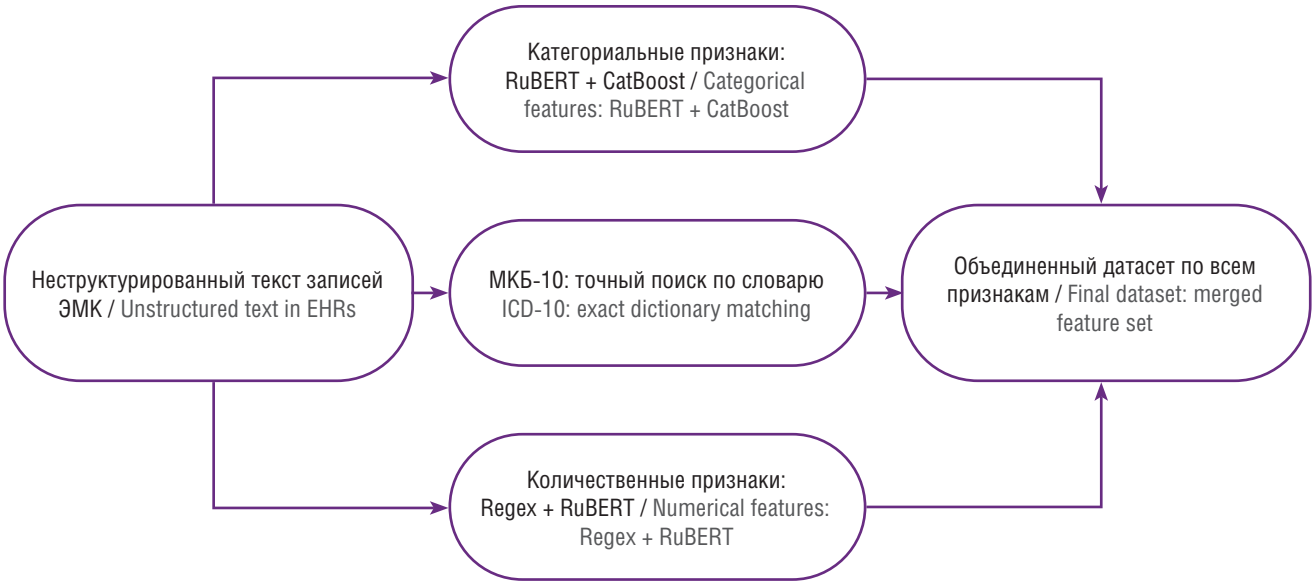


Рисунок 2. Схематическое представление принципа работы NLP-модели для извлечения признаков преждевременных родов.

Примечание: *RuBERT – Russian Bidirectional Encoder Representations from Transformers, предварительно обученная языковая модель для обработки русскоязычных текстов; CatBoost – Categorical Boosting Classifier, бинарный классификатор на основе градиентного бустинга CatBoost; ЭМК – электронные медицинские карты; МКБ-10 – Международная классификация болезни 10-го пересмотра; Regex – шаблоны на основе регулярных выражений.*

Figure 2. Schematic representation of the operating principle for the NLP model for extracting signs of premature birth.

Note: *RuBERT – Russian Bidirectional Encoder Representations from Transformers, a pre-trained language model for processing Russian-language texts; CatBoost – Categorical Boosting Classifier, a binary classifier based on the CatBoost gradient boosting framework; EHRs – electronic health records; ICD-10 – International Classification of Diseases, 10th Revision; Regex – Patterns based on regular expressions.*

2026 • Vol. 20 • № 1 Obstetrics, Gynecology and Reproduction

отрицательных концентраций). Обнаруженные выбросы систематизировались путем замены их на соответствующее граничное значение допустимого диапазона.

Для обеспечения сопоставимости признаков и улучшения сходимости алгоритмов был выполнен этап нормализации данных. Все числовые параметры подвергались Z-нормализации, при которой каждый признак преобразовывался к нулевому среднему значению и единичному стандартному отклонению. Этот подход особенно важен для алгоритмов, чувствительных к масштабу входных данных, таких как логистическая регрессия и нейронные сети.

Существенной проблемой был дисбаланс классов, для решения которой был применен алгоритм SMOTE, описанный выше. После обработки распределение классов стало сбалансированным (достигнуто примерное соотношение 50/50).

Все этапы предобработки данных были реализованы с использованием современных библиотек Python: *pandas* и *scipy.stats* для обработки пропущенных значений и выбросов, *sklearn.preprocessing* для нормализации и *imblearn* для балансировки классов. Такой комплексный подход к предобработке данных позволяет существенно повысить качество входных данных и, как следствие, улучшил производительность последующего прогнозного моделирования.

Методы статистического анализа и машинного обучения / Statistical analysis and machine learning methods

Для сравнения групп по количественным и категориальным признакам использовались стандартные статистические методы с учетом характера распределения данных. Статистически значимыми считались различия при $p < 0,05$. В ходе разработки предиктивной модели проводилась комплексная оценка эффективности 14 алгоритмов ML, включая сравнение традиционных статистических методов (логистическая регрессия) с современными ML-алгоритмами (ансамблевые методы и нейронные сети). В исследование были включены следующие классификаторы: логистическая регрессия (англ. Logistic Regression, LR), метод опорных векторов (англ. Support Vector Machine, SVM), линейный классификатор опорных векторов (англ. Linear Support Vector Classifier, Linear SVC), стохастический градиентный спуск (англ. Stochastic Gradient Descent, SGD), персептрон (англ. Perceptron), наивный байесовский классификатор (англ. Naive Bayes, NB), алгоритм k ближайших соседей (k-Nearest Neighbors, k-NN), ансамблевые методы: случайный лес (англ. Random Forest), деревья решений (англ. Decision Trees), бэггинг (англ. Bagging Classifier), методы градиентного бустинга (XGBoost, LightGBM, CatBoost) и искусственная нейронная сеть (англ. Artificial Neural Network, ANN).

Все модели оценивались с использованием 5-кратной стратифицированной кросс-валидации на сбалансированных данных (StratifiedKFold). Для комплексной оценки применялись следующие параметры: AUC-ROC (способность дифференцировать классы), accuracy (точность классификации), recall (чувствительность), precision (точность положительного прогноза) и F1-score (сбалансированная мера) [22, 23]. Доверительные интервалы (95 %) рассчитывались по t-распределению, что обеспечило статистическую надежность результатов. Порог классификации установлен на уровне 0,5 в соответствии с общепринятой практикой для бинарных медицинских моделей [24]. Ранжирование предикторов выполнено с использованием встроенных методов градиентного бустинга.

Финальная валидация проведена на независимой выборке – 500 клинических наблюдений из базы данных ГБУЗ РПЦ им. Гуткина К.А. с анализом рабочих характеристик, матрицы ошибок и ключевых метрик. Критериями выбора оптимальной модели стали: максимальное значение AUC-ROC, устойчивость метрик при кросс-валидации и клиническая интерпретируемость при стандартном пороге 0,5. Схема отбора итоговой модели отражена на **рисунке 3**.

Результаты / Results

Описательная статистика / Descriptive statistics

На этапе формирования выборки выделено 2 класса наблюдений: класс 1 с целевым событием (ПР) – 317 случаев (6,3 % от общего объема выборки) и класс 0 без целевого события – 4683 наблюдения (93,7 %). Такое распределение отражает характерный для данной клинической ситуации дисбаланс классов, соответствующий реальной эпидемиологической картине распространенности ПР, при этом обеспечивая достаточное количество положительных случаев для проведения статистического анализа.

Частотный анализ категориальных признаков выявил значительную вариабельность – от 0,1 до 37,1 %. Наиболее распространенными предикторами (встречаемость > 15 %) оказались анемии и тромбоцитопении, эндокринные заболевания, ИЦН и коррекция акушерским пессарием, привычное невынашивание, ЭКО и инфекции мочевыводящих путей. К среднечастотным признакам (5–15 %) относились такие признаки, как плацентарная недостаточность, инфекции половых путей, самопроизвольный выкидыш в анамнезе, прием гестагенов, многоплодная беременность. Состояния с низкой частотой встречаемости (1–5 %) включали в себя 17 переменных, включая мертворождение (4,8 %), угрожающий выкидыш (4,7 %) и первые роды (4,4 %). К редким состояниям (< 1 %) были отнесены табакокурение, дефицит массы тела и чрезмерная прибавка массы тела. Подробное представление частотных характеристик анализируемых категориальных признаков, позволяющих оценить репрезентативность выборки, представлено в **таблице 2**.

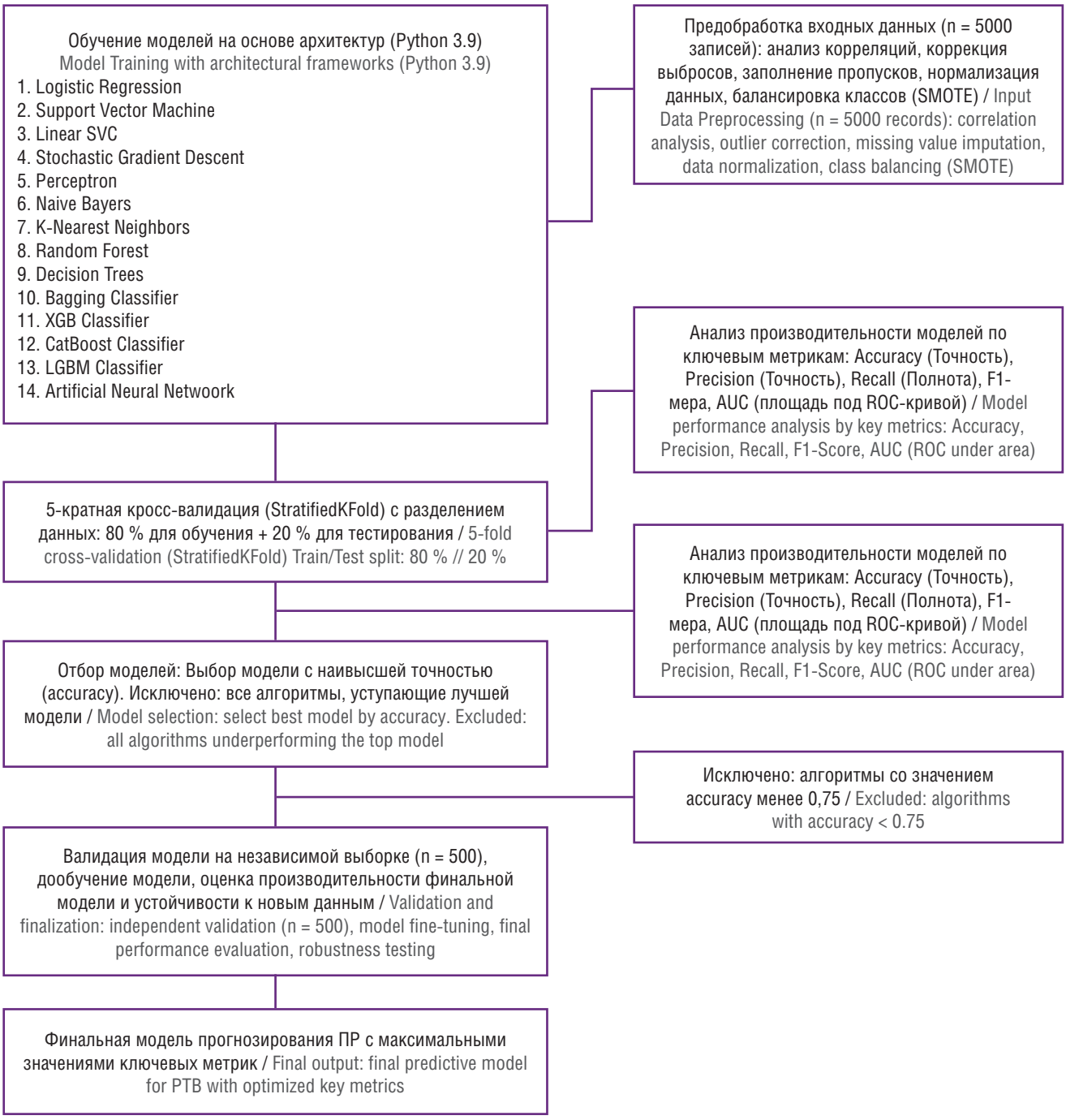


Рисунок 3. Алгоритм отбора финальной модели прогнозирования преждевременных родов.

Примечание: *Linear SVC – Linear Support Vector Classifier, линейная версия метода опорных векторов; SMOTE – Synthetic Minority Oversampling Technique, метод синтетической передискретизации меньшинства класса для несбалансированных данных; Accuracy – доля правильных предсказаний модели среди всех случаев; Recall – чувствительность; Precision – точность положительных предсказаний (доля истинно положительных среди всех положительных прогнозов); F-1 мера – гармоническое среднее между Precision и Recall; ПР – преждевременные роды.*

Figure 3. The algorithm for selecting the final model for predicting premature birth.

Note: *Linear SVC – Linea Support Vector Classifier, linear version of Support Vector Machines method; SMOTE – Synthetic Minority Oversampling Technique, synthetic minority over-sampling technique for imbalanced data; Accuracy – proportion of correct model predictions among all cases; Recall – sensitivity (true positive rate); Precision – accuracy of positive predictions (proportion of true positives among all positive predictions); F1-score – harmonic mean between Precision and Recall; PTB – Preterm Birth.*

Анализ количественных показателей выявил следующие ключевые характеристики: средний срок беременности на момент анализа составил 21 неделю. Наибольшую полноту данных продемонстрировали возраст (24 % записей), паритет (11,2 %) и срок беременности (28,8 %). Среди лабораторных и инструментальных параметров лидировали показатели УЗ-цервикометрии (16,7 %), уровень гемоглобина (5,4 %) и количество тромбоцитов (4,7 %). Подробная характеристика количественных признаков представлена в **таблице 3**.

2026 • Vol. 20 • № 1
Obstetrics, Gynecology and Reproduction

Таблица 2. Распределение частоты встречаемости категориальных признаков преждевременных родов.

Table 2. Distribution of categorical signs frequencies in preterm birth.

Признак Predictor	Частота, % Frequency, %	Количество случаев Number of cases	95 % ДИ 95 % CI	Погрешность MoE (± %) Margin of error (± %)
Анемии и тромбоцитопении / Anemias and thrombocytopenias	37,1	1853	[35,7–38,4]	1,34
Эндокринные заболевания (сахарный диабет) Endocrine disorders (diabetes mellitus)	24,3	1214	[23,1–25,5]	1,19
Истмико-цервикальная недостаточность (ИЦН) / Cervical insufficiency (CI)	20,8	1040	[19,7–21,9]	1,13
Экстракорпоральное оплодотворение / In vitro fertilization	19,1	954	[18,0–20,2]	1,1
Инфекции мочевыводящих путей / Urinary tract infections	18,4	922	[17,36–19,5]	1,08
Медицинское прерывание беременности Medical termination of pregnancy	18,1	904	[17,01–19,15]	1,07
Привычное невынашивание / Recurrent pregnancy loss	16,1	803	[15,04–17,08]	1,02
Коррекция ИЦН акушерским пессарием CI correction with obstetric pessary	15,0	751	[14,03–16,01]	0,99
Плацентарная недостаточность / Placental insufficiency	11,8	592	[10,94–12,74]	0,9
Инфекции половых путей / Genital tract infections	8,8	439	[8,00–9,56]	0,78
Самопроизвольный выкидыш / Spontaneous abortion (miscarriage)	8,0	398	[7,21–8,71]	0,75
Прием гестагенов / Progestogen therapy	7,1	357	[6,43–7,85]	0,71
Многоплодная беременность / Multiple pregnancy	6,9	345	[6,20–7,60]	0,7
Внутриматочные вмешательства / Intrauterine interventions	6,7	335	[6,01–7,39]	0,69
Угрожающие преждевременные роды / Threatened preterm birth	6,3	317	[5,66–7,02]	0,68
Миома матки / Uterine fibroids	6,0	301	[5,36–6,68]	0,66
Мертворождение / Stillbirth	4,76	238	[4,17–5,35]	0,59
Угрожающий выкидыш / Threatened abortion	4,66	233	[4,08–5,24]	0,58
Первородящая / Nulliparous	4,4	220	[3,83–4,97]	0,57
Полип полости матки / Endometrial polyp	3,46	173	[2,95–3,97]	0,51
Избыточная масса тела / Overweight	2,74	137	[2,29–3,19]	0,45
Повторнородящая / Multiparous	2,48	124	[2,05–2,91]	0,43
Наследственная тромбофилия / Inherited thrombophilia	1,66	83	[1,31–2,01]	0,35
Бесплодие / Infertility	1,76	88	[1,40–2,12]	0,36
Вредная зависимость (алкоголизм, наркотики) Substance use disorder (alcohol, drugs)	1,66	86	[1,36–2,08]	0,36
Маловодие / Oligohydramnios	1,36	68	[1,04–1,68]	0,32
Ожирение / Obesity	1,28	64	[0,97–1,59]	0,31
Замершая (неразвивающаяся) беременность / Missed abortion	1,28	64	[0,97–1,59]	0,31
Многоводие / Polyhydramnios	1,04	54	[0,76–1,32]	0,28
Брак зарегистрирован / Registered marriage	0,88	44	[0,62–1,14]	0,26
Коррекция ИЦН швом (серкляж) / CI correction with cerclage	0,84	42	[0,59–1,09]	0,25
Доброкачественные заболевания шейки матки / Benign cervical diseases	0,78	39	[0,54–1,02]	0,24
Эндометриоз / Endometriosis	0,76	38	[0,52–1,00]	0,24
Дистресс (гипоксия) плода / Fetal distress (hypoxia)	0,76	38	[0,52–1,00]	0,24
Брак не зарегистрирован / Unregistered marriage	0,62	31	[0,4–0,84]	0,22
Лечение заболеваний шейки матки / Treatment of cervical diseases	0,56	28	[0,35–0,77]	0,21
Злокачественные заболевания шейки матки / Malignant cervical diseases	0,56	28	[0,35–0,77]	0,21
Киста яичника / Ovarian cyst	0,42	21	[0,24–0,6]	0,18
Табакокурение / Tobacco smoking	0,14	7	[0,04–0,24]	0,10
Дефицит массы тела / Underweight	0,10	5	[0,01–0,19]	0,09
Чрезмерная прибавка массы тела / Excessive weight gain	0,10	5	[0,01–0,19]	0,09

Таблица 3. Характеристики количественных признаков в выборке для обучения предиктивной модели.

Table 3. Characteristics of quantitative signs in the sample for predictive model training.

Показатель Predictor	Количество карт (%) Number of records (%)	Среднее значение (диапазон) Mean value (range)	Медиана Median	SD
Анамнестические и конституциональные признаки / Anamnestic and constitutional signs				
Возраст, лет / Age, years	1200 (24,0)	30,0 (18,0–45,0)	29,0	4,8
Менархе, лет / Menarche, years	175 (3,5)	12,9 (11,0–17,0)	13	1,4
Индекс массы тела, кг/м^2 / Body mass index, kg/m^2	775 (15,5)	24,6 (17,3–35,0)	24	3,8
Паритет / Parity	560 (11,2)	6,1 (1,0–15,0)	5,0	2,0
Срок беременности, недель / Gestational age, weeks	1440 (28,8)	21,7 (10,0–36,0)	21	12,9
Лабораторные и инструментальные признаки / Laboratory and instrumental signs				
Гемоглобин, г/л / Hemoglobin, g/L	272 (5,4)	114,53 (63,0–185,0)	113	16,51
Тромбоциты, ×10^9/л / Platelets, ×10^9/L	235 (4,7)	244,1 (2,0–466)	248	89,91
Лейкоциты, ×10^9/л / Leukocytes, ×10^9/L	110 (2,2)	11,6 (1,0–37,0)	9,0	8,2
Фибриноген, г/л / Fibrinogen, g/L	105 (2,1)	3,4 (1,0–7,0)	3,0	1,1
АЧТВ, с / APTT, s	70 (1,4)	27,5 (15,0–39,0)	28	4,5
Протромбиновый индекс, % / Prothrombin index, %	75 (1,5)	94,8 (45,0–125,0)	95	17,5
С-реактивный белок, мг/л / C-reactive protein, mg/L	68 (1,4)	28,9 (2,0–78,0)	22,5	25,2
Ультразвуковая цервикометрия, мм Ultrasound cervicometry, mm	835 (16,7)	20,6 (4,0–38,0)	20	4,8

Примечание: *АЧТВ – активированное частичное тромбопластиновое время.*

Note: *APTT – activated partial thromboplastin time.*

Анализ показал, что в группе с ПР по сравнению с контрольной группой (без целевого события) достоверно чаще ($p < 0,05$) встречались такие факторы, как плацентарная недостаточность (встречалась в 2,5 раза чаще при ПР – 26,5 % против 10,8 %), применение ЭКО (25,9 % против 18,6 %) и ИЦН (25,6 % против 20,5 %). Другие значимые факторы включали: многоводие (3,2 % при ПР против 0,9 %), доброкачественные заболевания шейки матки (2,5 % против 0,7 %), первородящие (8,2 % против 4,1 %). Перечень факторов, внесших наибольший вклад в развитие ПР, представлен в **таблице 4**.

Соотношение классов валидационной выборки (m = 500 случаев из ГБУЗ РПЦ им. Гуткина К.А.) соответствовало структуре обучающей выборки и обеспечивало репрезентативность клинических и демо-

Таблица 4. Статистически значимые предикторы преждевременных родов по результатам анализа данных.

Table 4. Statistically significant predictors of preterm birth based on data analysis.

Фактор / Factor	Класс 1 Class 1 n = 317 %	Класс 2 Class 2 n = 4683 %	Повышение риска, % Risk increase, %	χ^2	p
Плацентарная недостаточность / Placental insufficiency	26,5	10,8	+15,7	68,18	< 0,001
Многоводие / Polyhydramnios	3,2	0,9	+2,3	12,59	0,0004
Доброкачественные заболевания шейки матки / Benign cervical diseases	2,5	0,7	+1,9	11,00	0,0009
Первородящая / Primiparous	8,2	4,1	**+4,1**	10,69	0,0011
Повторнородящая / Multiparous	5,7	2,3	+3,4	12,94	0,0003
Экстракорпоральное оплодотворение / In vitro fertilization	25,9	18,6	+7,2	9,64	0,0019
Лечение заболеваний шейки матки / Treatment of cervix diseases	1,9	0,5	+1,4	8,39	0,0038
Истмико-цервикальная недостаточность / Cervical insufficiency	25,6	20,5	**+5,1**	4,34	0,0373
Дефицит массы тела / Underweight	0,6	0,1	+0,6	4,72	0,0298
Чрезмерная прибавка массы тела / Excessive weight gain	0,6	0,1	+0,6	4,72	0,0298

Примечание: *выделены факторы, внесшие наибольший вклад в риск развития преждевременных родов.*

Note: *the factors that contributed most to the risk of preterm birth are highlighted in bold.*

2026 • Vol. 20 • № 1
Obstetrics, Gynecology and Reproduction

графических параметров (6,3 % случаев ПР против 93,7 % нормальных родов). Выборка для валидации модели показала более полный охват ключевых параметров, таких как возраст (100 %), ИМТ (98 %) и срок беременности (91,2 %) против 24, 15,5 и 28,8 % в обучающей выборке, а также повышенную частоту критически важных предикторов, таких как плацентарная недостаточность (24,2 % против 11,8 %) и многоводие (12,0 % против 1,04 %). Такая структура данных позволила провести более строгую оценку модели в условиях, максимально приближенных к реальной клинической практике, что особенно важно для подтверждения ее прогностической ценности.

Разработка предиктивной модели / Predictive model development

Для создания предиктивной модели были протестированы 14 алгоритмов машинного обучения (подробно представлены в разделе «Статистические методы»). Модель анализировала 54 клинико-анамнестических параметра, автоматически извлеченных из ЭМК с помощью разработанной NLP-системы. Полный список учитываемых потенциальных предикторов ПР приведен в приложении к исследованию.

Исходные сбалансированные данные были разделены на обучающую (80 %) и тестовую (20 %) выборки. Для каждого типа модели проводился подбор оптимальных параметров. Моделирование проводилось с учетом особенностей медицинских данных, включая дисбаланс классов и клиническую значимость предикторов. Для компенсации исходного дисбаланса классов была применена технология SMOTE, обеспечившая сбалансированное соотношение классов. Однако, чтобы избежать переобучения на синтетических данных, был использован комплексный подход валидации, включающий: 1) пятикратную стратифицированную кросс-валидацию, 2) тестирование на выборке с естественным распределением классов, 3) оценку по метрикам, устойчивым к дисбалансу (F1-score и AUC-ROC). Такой многоуровневый подход позволил создать модель, сохраняющую высокую прогностическую способность как на сбалансированных, так и на исходных несбалансированных данных, что особенно важно для клинической практики, где частота ПР традиционно колеблется в диапазоне 4–6 %.

Оптимизация алгоритма включала комплексную оценку доли правильных классификаций (accuracy) с 5-кратной кросс-валидацией и расчетом 95 % ДИ. Наиболее эффективный алгоритм дополнительно тестировался на независимой выборке по стандартному протоколу валидации.

Производительность моделей / Model performance

Разработанная NLP-модель продемонстрировала следующие медианные показатели: F1-мера = 0,976, полнота = 0,998 и AUC-ROC = 0,974. Полный перечень полученных метрик для категориальных и количественных предикторов отражен в **таблице 5**. Высокие показатели производительности NLP-модели обеспечили создание качественного датасета для обучения предиктивной модели.

Показатели ключевой метрики эффективности (accuracy) для алгоритмов ML, решающих задачу прогнозирования ПР, представлены в **таблице 6**. Наибольшую эффективность продемонстрировал ансамблевый алгоритм CatBoost Classifier, основанный на методе градиентного бустинга. Его показатели на внутренней валидации составили: точность = 0,8064 (95 % ДИ = 0,784–0,816), чувствительность = 0,76 (95 % ДИ = 0,748–0,772), F1-мера = 0,79 (95 % ДИ = 0,782–0,798) и AUC-ROC = 0,79 (95 % ДИ = 0,774–0,806).

При валидации на независимой выборке модель подтвердила высокую дискриминативную способность и устойчивость к новым данным, повысив AUC-ROC до 0,82 (95 % ДИ = 0,809–0,831), чувствительность – до 0,87 (95 % ДИ = 0,857–0,883), accuracy – до 0,81 (95 % ДИ = 0,799–0,821) и F1-score – до 0,81 (95 % ДИ = 0,805–0,815).

Итоговый перечень предикторов, используемых для разработки модели машинного обучения для оценки риска ПР, приведен в **приложении 1**.

Разработанная система оценки риска ПР демонстрирует высокую точность и способна автоматически анализировать факторы риска ПР из ЭМК, выявляя взаимосвязи между клиническими, анамнестическими и социально-демографическими данными.

Обсуждение / Discussion

Преждевременные роды являются грозным акушерским осложнением с высоким риском для жизни и здоровья матери и ребенка. Несмотря на существующие меры профилактики, частота ПР остается стабильно высокой, что требует поиска новых подходов к прогнозированию. Сложный многокомпонентный патогенез ПР, объединяющий инфекционные, эндокринные и коагуляционные нарушения [25, 26], обладает высокой степенью вариабельности, что ограничивает возможности применения стандартных профилактических подходов и снижает их клиническую результативность. В связи с чем особую актуальность приобретают методы искусственного интеллекта (ИИ), обеспечивающие анализ комплексных взаимосвязей между предикторами и индивидуальную оценку риска ПР. Их внедрение в клиническую практику может способствовать снижению частоты ПР и улучшению перинатальных исходов за счет повышения качества оценки риска.

В ходе исследования разработаны 2 взаимосвязанные ИИ-модели: NLP-модель для экстракции предикторов ПР из медицинских записей ЭМК и предиктивная модель, показавшая высокую точность при

Таблица 5 (начало). Метрики качества разработанной модели обработки естественного языка.

Table 5 (beginning). Performance metrics of the developed Natural Language Processing model.

Название предиктора / Predictor	Значение метрики / Metrics			
	Точность Precision	Полнота Recall	F-1 мера F1-score	Площадь под кривой ошибок AUC-ROC
Категориальные предикторы / Qualitative predictors				
Анемии и тромбоцитопении / Anemias and thrombocytopenias	0,825	0,857	0,841	0,828
Бесплодие / Infertility	0,978	1,0	0,989	0,989
Брак зарегистрирован / Registered marriage	0,977	1,0	0,988	0,989
Брак не зарегистрирован / Unregistered marriage	0,987	1,0	0,993	0,993
Внутриматочные вмешательства / Intrauterine interventions	0,971	0,979	0,975	0,969
Вредная зависимость (алкоголизм, наркотики) Substance use disorder (alcohol, drugs)	0,996	1,0	0,998	0,998
Дефицит массы тела / Underweight	0,997	1,0	0,998	0,998
Дистресс (гипоксия) плода / Fetal distress (hypoxia)	0,909	0,964	0,935	0,936
Доброкачественные заболевания шейки матки / Benign cervical diseases	0,965	0,995	0,980	0,978
Замершая беременность в анамнезе / Missed abortion	0,997	1,0	0,998	0,998
Злокачественные заболевания шейки матки / Malignant cervical diseases	0,991	1,0	0,995	0,995
Избыточная масса тела / Overweight	0,974	1,0	0,987	0,986
Инфекция мочевыводящих путей / Urinary tract infections	0,837	0,944	0,887	0,887
Инфекция половых путей / Genital tract infections	0,927	0,969	0,948	0,948
Истмико-цервикальная недостаточность (ИЦН) / Cervical insufficiency (CI)	0,915	0,953	0,933	0,933
Киста яичника / Ovarian cyst	1,0	1,0	1,0	1,0
Коррекция ИЦН акушерским разгружающим пессарием CI correction with obstetric pessary	0,958	0,988	0,973	0,973
Коррекция ИЦН швом на шейке матки (серкляж) / CI correction with cervical cerclage	0,981	0,998	0,989	0,989
Лечение заболеваний шейки матки / Cervical disease treatment	1,0	1,0	1,0	1,0
Маловодие / Oligohydramnios	0,917	0,986	0,950	0,947
Медицинское прерывание беременности / Medical termination of pregnancy	0,913	0,982	0,946	0,944
Мертворождение в анамнезе / Stillbirth	0,925	1,0	0,961	0,960
Миома матки / Uterine fibroids	0,904	0,947	0,925	0,931
Многоводие / Polyhydramnios	0,951	0,990	0,971	0,969
Многоплодная беременность / Multiple pregnancy	0,932	0,970	0,950	0,948
Наследственная тромбофилия / Inherited thrombophilia	0,988	1,0	0,994	0,993
Ожирение / Obesity	0,977	1,0	0,988	0,988
Первородящая / Primiparous	0,919	1,0	0,957	0,959
Плацентарная недостаточность / Placental insufficiency	0,863	0,968	0,912	0,900
Повторнородящая / Multiparous	0,911	0,989	0,949	0,943
Полип полости матки / Endometrial polyp	0,988	1,0	0,994	0,993
Привычное невынашивание / Recurrent pregnancy loss	0,964	0,997	0,980	0,981
Прием гестагенов / Progestogen therapy	0,948	0,993	0,970	0,970
Самопроизвольный выкидыш / Spontaneous miscarriage	0,989	1,0	0,994	0,994
Табакокурение / Tobacco smoking	0,903	0,986	0,943	0,941
Угрожающие преждевременные роды / Threatened preterm birth	0,885	0,992	0,936	0,941
Угрожающий выкидыш / Threatened abortion	0,986	1,0	0,993	0,993
Чрезмерная прибавка веса / Excessive weight gain	1,0	1,0	1,0	1,0
Экстракорпоральное оплодотворение / In vitro fertilization	0,960	0,988	0,974	0,974
Эндокринные заболевания (сахарный диабет) / Endocrine disorders (diabetes mellitus)	0,928	0,947	0,937	0,939
Эндометриоз / Endometriosis	1,0	1,0	1,0	1,0

2026 • Vol. 20 • № 1 Obstetrics, Gynecology and Reproduction

Таблица 5 (окончание). Метрики качества разработанной модели обработки естественного языка.

Table 5 (ending). Performance metrics of the developed Natural Language Processing model.

Название предиктора / Predictor	Значение метрики / Metrics			
	Точность Precision	Полнота Recall	F-1 мера F1-score	Площадь под кривой ошибок AUC-ROC
Количественные предикторы / Quantitative predictors				
Активированное частичное тромбопластиновое время Activated partial thromboplastin time	1,0	–	–	–
Возраст / Age	0,998020	0,988235	0,993103	0,993097
Гемоглобин / Hemoglobin	1,0	–	–	–
Индекс массы тела / Body mass index	0,997930	1,0	0,998964	0,999035
Лейкоциты / Leukocytes	0,995495	–	–	–
Менархе / Menarche	1,0	0,995960	0,997976	0,997980
Паритет / Parity	0,967871	0,969819	0,968844	0,969005
Протромбиновый индекс / Prothrombin index	1,0	–	–	–
С-реактивный белок / C-reactive protein	1,0	–	–	–
Ультразвуковая цервикометрия / Ultrasound cervicometry	0,967871	0,969819	0,968844	0,969005
Срок беременности / Gestational age	0,969325	1,0	0,984424	0,985741
Тромбоциты / Platelets	0,99593	–	–	–
Фибриноген / Fibrinogen	1,0	–	–	–

Таблица 6. Значения ключевой метрики (accuracy) в ходе обучения и выбора наиболее эффективной модели оценки риска преждевременных родов.

Table 6. Key metric (accuracy) magnitude during training and selection of the most effective preterm birth risk assessment model.

Модель машинного обучения Machine learning model	Доля правильных ответов Accuracy	95 % ДИ 95 % CI
ANN	0,758	[0,746–0,770]
Bagging Classifier	0,767	[0,755–0,779]
CatBoost Classifier	**0,806**	[0,795–0,817]
Decision Tree Classifier	0,767	[0,755–0,779]
k-NN	0,740	[0,727–0,753]
LightGBM	0,770	[0,758–0,782]
Linear SVC	0,682	[0,669–0,695]
LR	0,683	[0,670–0,696]
NB	0,652	[0,639–0,665]
Perceptron	0,667	[0,654–0,680]
Random Forest Classifier	0,776	[0,764–0,788]
SGD Classifier	0,683	[0,670–0,696]
SVM	0,766	[0,754–0,778]
XGB Classifier	0,677	[0,664–0,690]

Примечание: *ANN – Artificial neural network, искусственная нейронная сеть; k-NN – алгоритм k-ближайших соседей; LightGBM – Light Gradient Boosting Machine, алгоритм на основе градиентного бустинга; Linear SVC – Linear Support Vector Classifier, линейный алгоритм на основе метода опорных векторов; LR – Logistic Regression, логистическая регрессия; NB – Naive Bayes, Наивный Байесовский классификатор; SGD – Stochastic Gradient Descent, алгоритм на основе стохастического градиентного спуска; SVM – Support Vector Machine, метод опорных векторов; XGB – Extreme Gradient Boosting, экстремальный градиентный бустинг; 95 % ДИ – 95 % доверительный интервал; выделено максимальное значение ключевой метрики.*

Note: *ANN – Artificial Neural Network; k-NN – k-Nearest Neighbors algorithm; LightGBM – Light Gradient Boosting Machine, gradient boosting framework; Linear SVC – Linear Support Vector Classifier, linear support vector machine algorithm; LR – Logistic Regression; NB – Naive Bayes classifier; SGD – Stochastic Gradient Descent; SVM – Support Vector Machine; XGB – Extreme Gradient Boosting; 95 % CI – 95 % confidence interval; maximum value of the key metric is highlighted.*

оценке риска ПР (AUC = 0,82; Recall = 0,87). Производительность системы сопоставима с результатами зарубежных авторов [19–21]. Однако в отличие от исследований, использующих узкий набор биомаркеров (например, щелочную фосфатазу, альфа-фетопротеин или плацентарные факторы роста), данная модель включает широкий спектр предикторов, доступных в клинической практике, что позволяет учитывать комплексное влияние социально-демографических, акушерских и соматических факторов на риск ПР. Важным преимуществом разработанной системы является ее способность анализировать большие массивы неструктурированных медицинских данных и выявлять сложные, нелинейные взаимосвязи между предикторами ПР, что недоступно при традиционной статистической обработке данных.

Вместе с тем разработанная система адаптирована к особенностям русскоязычной электронной медицинской документации, включая обработку неструктурированных текстовых данных с помощью NLP-алгоритмов, в то время как зарубежные аналоги работают с англоязычными стандартизированными наборами данных. Применение для анализа данных специальных NLP-алгоритмов обеспечило формирование качественного обучающего датасета. Внешняя валидация подтвердила высокую точность оценки риска и стабильность при работе с новыми данными, что предполагает возможность клинического применения разработанного инструмента [20]. Возможности дальнейшего усовершенствования системы оценки риска ПР включают расширение выборки за счет мультицентровых данных и включения дополнительных лабораторных и генетических маркеров ПР.

ИИ – перспективный инструмент для оценки риска ПР, обладающий рядом ключевых преимуществ. Основное достоинство метода заключается в возможности анализа уже собранных в рамках стандартного диспансерного наблюдения клинических данных из ЭМК без необходимости проведения дополнительных дорогостоящих исследований. Алгоритмы ML эффективно работают со стандартными клинико-анамнестическими показателями, результатами лабораторных и инструментальных исследований, выявляя сложные взаимосвязи между многочисленными факторами риска. Несмотря на то что разработка точных прогностических моделей требует тщательной обработки больших массивов качественных данных, применение методов ML открывает новые перспективы для развития предиктивной медицины в акушерской практике, позволяя своевременно выявлять пациенток с высоким риском развития осложнений.

Ограничения исследования / Study limitations

Настоящее исследование имеет ряд ограничений, в основном связанных с его ретроспективным дизайном и природой исходных данных.

Ретроспективный дизайн и временной фактор

Отсутствие фиксации строго определенного срока беременности для измерения каждого предиктора не позволяет оценить временной интервал между оценкой признака и исходом, что является стандартом для проспективных прогностических исследований. Таким образом, модель следует рассматривать в первую очередь как инструмент для оценки риска ПР, а не для точного прогнозирования исхода во времени.

Характер используемых данных

Исследование было целенаправленно ограничено данными, рутинно используемыми в отечественной акушерской практике. В модель не включались специализированные лабораторные биомаркеры, что повышает ее потенциальную применимость в широкой клинической практике, но ограничивает прямое сравнение с зарубежными аналогами. Кроме того, для выявления сложных взаимосвязей был применен комплексный подход, включавший анализ широкого спектра параметров, в том числе социально-демографических, без предварительного строгого клинического отбора.

Неполное заполнение данных

Проблема пропущенных значений является общепризнанным вызовом при работе с ЭМК, которая ориентирована на клинические, а не исследовательские нужды. Для минимизации этого влияния был применен комплексный подход к предобработке данных, включая импутацию пропущенных значений. Перспективным направлением для повышения качества данных является разработка и внедрение стандартизированных шаблонов документации.

Заключение / Conclusion

Разработанная модель продемонстрировала высокую способность оценивать риск ПР на основе ретроспективных данных, полученных из ЭМК. Успешный результат был достигнут благодаря созданию специализированной NLP-модели для обработки русскоязычных медицинских текстов и создания качественного обучающего датасета. Устойчивость модели на основе алгоритма CatBoost Classifier подтверждена тестированием, что открывает перспективы ее применения в клинической практике. Данное исследование закладывает основы создания комплексного решения на основе технологий ИИ для анализа реальных клинических данных, с перспективой улучшения качества инструмента путем расширения спектра предикторов ПР, валидации на данных из других регионов и тщательной оптимизации предобработки данных в будущих исследованиях.

ИНФОРМАЦИЯ О СТАТЬЕ	ARTICLE INFORMATION
Поступила: 12.11.2025. **В доработанном виде:** 25.11.2025. **Принята к печати:** 05.12.2025. **Опубликована онлайн:** 06.12.2025.	**Received:** 12.11.2025. **Revision received:** 25.11.2025. **Accepted:** 05.12.2025. **Published online:** 06.12.2025.
Вклад авторов	**Author's contribution**
Все авторы внесли равный вклад в написание и подготовку рукописи.	All authors contributed equally to the article.
Все авторы прочитали и утвердили окончательный вариант рукописи.	All authors have read and approved the final version of the manuscript.
Конфликт интересов	**Conflict of interests**
Авторы заявляют об отсутствии конфликта интересов.	Authors declare no conflict of interest.
Финансирование	**Funding**
Исследование выполнено за счет гранта Российского научного фонда № 24-25-00429, https://rscf.ru/project/24-25-00429/.	This research was financially supported by the Russian Science Foundation, Grant No. 24-25-00429, https://rscf.ru/project/24-25-00429/.
Раскрытие данных	**Data sharing**
Данные, лежащие в основе результатов, протокол исследования, план статистического анализа, принципы анализа, представленные в этой статье, доступны по запросу автору, ответственному за корреспонденцию, после одобрения ведущим исследователем.	The underlying data, study protocol, statistical analysis plan, and principles of analysis presented in the article are available upon request to the corresponding author after approval by principle investigator.
Онлайн-контент	**Online content**
Онлайн-версия содержит дополнительные материалы, доступные на сайте журнала https://gynecology.su на странице публикации https://doi.org/10.17749/2313-7347/ob.gyn.rep.2025.701. **Приложение 1.** Итоговый список предикторов, используемых для разработки модели машинного обучения для оценки риска преждевременных родов (в алфавитном порядке).	The online version contains supplementary material available at the journal website https://gynecology.su at the paper webpage https://doi.org/10.17749/2313-7347/ob.gyn.rep.2025.701. **Appendix 1.** Final list of predictors used for the development of machine learning model for preterm birth risk assessment (in alphabetical order).
Комментарий издателя	**Publisher's note**
Содержащиеся в этой публикации утверждения, мнения и данные были созданы ее авторами, а не издательством ИРБИС (ООО «ИРБИС»). Издательство ИРБИС снимает с себя ответственность за любой ущерб, нанесенный людям или имуществу в результате использования любых идей, методов, инструкций или препаратов, упомянутых в публикации.	The statements, opinions, and data contained in this publication were generated by the authors and not by IRBIS Publishing (IRBIS LLC). IRBIS Publishing disclaims any responsibility for any injury to peoples or property resulting from any ideas, methods, instructions, or products referred in the content.
Права и полномочия	**Rights and permissions**
ООО «ИРБИС» обладает исключительными правами на эту статью по Договору с автором (авторами) или другим правообладателем (правообладателями). Использование этой статьи регулируется исключительно условиями этого Договора и действующим законодательством.	IRBIS LLC holds exclusive rights to this paper under a publishing agreement with the author(s) or other rightsholder(s). Usage of this paper is solely governed by the terms of such publishing agreement and applicable law.

Литература:

1. Ившин А.А., Погодин О.О., Шакурова Е.Ю. и др. Лапароскопический трансабдоминальный серкляж для лечения истмико-цервикальной недостаточности при беременности: клинический случай и обзор литературы. *Акушерство, Гинекология и Репродукция*. 2025;19(1):116–26. https://doi.org/10.17749/2313-7347/ob.gyn.rep.2025.578.
2. Серов В.Н., Сухорукова О.И. Эффективность профилактики преждевременных родов. *Акушерство и гинекология*. 2013;(3):48–53.
3. Risnes K., Bilsteen J.F., Brown P. et al. Mortality among young adults born preterm and early term in 4 Nordic nations. *JAMA Netw Open*. 2021;4(1):e2032779. https://doi.org/10.1001/jamanetworkopen.2020.32779.
4. Jeon G.W., Lee J.H., Oh M., Chang Y.S. Serial long-term growth and neurodevelopment of very-low-birth-weight infants: 2022 update on the Korean Neonatal Network. *J Korean Med Sci*. 2022;37(34):e263. https://doi.org/10.3346/jkms.2022.37.e263.
5. Горина К.А., Ходжаева З.С., Белоусов Д.М. и др. Преждевременные роды: прошлые ограничения и новые возможности. *Акушерство и гинекология*. 2020;(1):12–9. https://doi.org/10.18565/aig.2020.1.12-19.
6. Stedall P.M., Spencer-Smith M.M., Lah S. et al. Episodic and prospective memory difficulties in 13-year-old children born very preterm. *J Int Neuropsychol Soc*. 2023;29(3):257265. https://doi.org/10.1017/S1355617722000170.
7. Белоусова В.С., Стрижаков А.Н., Свитич О.А. и др. Преждевременные роды: причины, патогенез, тактика. *Акушерство и гинекология*. 2020;(2):82–7. https://doi.org/10.18565/aig.2020.2.82-87.
8. Thain S., Yeo G.S.H., Kwek K., Chern B., Tan K.H. Spontaneous preterm birth and cervical length in a pregnant Asian population. *PLoS One*. 2020;15(4):e0230125. https://doi.org/10.1371/journal.pone.0230125.
9. Друккер Н.А., Дурницына О.А., Никашина А.А., Селютина С.Н. Диагностическая значимость α-1-микроглобулина в развитии преждевременных родов. *Акушерство и гинекология*. 2019;(1):81–5. https://doi.org/10.18565/aig.2019.1.81-85.
10. Баев О.Р., Дикке Г.Б. Диагностика преждевременного разрыва плодных оболочек на основании биохимических тестов. *Акушерство и гинекология*. 2018;(9):132–6. https://doi.org/10.18565/aig.2018.9.132-136.
11. Клинические рекомендации – Преждевременные роды – 2020 (01.12.2020). *М.: Министерство здравоохранения Российской Федерации*, 2020. 66 с. Режим доступа: https://cr.minzdrav.gov.ru/schema/331_1. [Дата обращения: 28.08.2025].
12. Манухин И.Б., Фириченко С.В., Микаилова Л.У. и др. Прогнозирование и профилактика преждевременных родов – современное состояние проблемы. *Российский вестник акушера-гинеколога*. 2016;(3):9–15. https://doi.org/10.17116/rosakush20161639-15.
13. Ходжаева З.С., Дембовская С.В., Доброхотова Ю.Э. и др. Медикаментозная профилактика преждевременных родов

(результаты международного многоцентрового открытого исследования МИСТЕРИ). *Акушерство и гинекология.* 2016;(8):37–43. https://doi.org/10.18565/aig.2016.8.37-43.
14. Баринов С.В., Артымук Н.В., Новикова О.Н. и др. Опыт ведения беременных группы высокого риска по преждевременным родам с применением акушерского куполообразного пессария и серкляжа. *Акушерство и гинекология.* 2019;(1):140–8. https://doi.org/10.18565/aig.2019.1.140-148.
15. Crockart I.C., Brink L.T., du Plessis C., Odendaal H.J. Classification of intrauterine growthrestriction at 34-38 weeks gestation with machine learning models. *Inform Med Unlocked.* 2021;23:100533. https://doi.org/10.1016/j.imu.2021.100533.
16. Liu J., Wang C., Yan R. et al. Machine learning-based prediction of postpartum hemorrhage after vaginal delivery: combining bleeding high risk factors and uterine contraction curve. *Arch Gynecol Obstet.* 2022;306(4):1015–25. https://doi.org/10.1007/s00404-021-06377-0.
17. Melinte-Popescu A.S., Vasilache I.A., Socolov D., Melinte-Popescu M. Predictive performance of machine learning-based methods for the prediction of preeclampsia – a prospective study. *J Clin Med.* 2023;12(2):418. https://doi.org/10.3390/jcm12020418.
18. Андрейченко А.Е., Лучинин А.С., Ившин А.А. и др. Разработка и валидация моделей прогнозирования общего риска преэклампсии и риска ранней преэклампсии с использованием алгоритмов машинного обучения в первом триместре беременности. *Акушерство и гинекология.* 2023;(10):94–107. https://doi.org/10.18565/aig.2023.101.
19. Chen Y., Shi X., Wang Z., Zhang L. Development and validation of a spontaneous preterm birth risk prediction algorithm based on maternal bioinformatics: A single-center retrospective study. *BMC Pregnancy Childbirth.* 2024;24(1):763. https://doi.org/10.1186/s12884-024-06933-x.
20. Zhang Y., Du S., Hu T. et al. Establishment of a model for predicting preterm birth based on the machine learning algorithm. *BMC Pregnancy Childbirth.* 2023;23(1):779. https://doi.org/10.1186/s12884-023-06058-7.
21. Sun Q., Zou X., Yan Y. et al. Machine learning-based prediction model of preterm birth using electronic health record. *J Healthc Eng.* 2022;2022:9635526. https://doi.org/10.1155/2022/9635526.
22. Mavrogiorgou A., Kiourtis A., Kleftakis S. et al. A catalogue of machine learning algorithms for healthcare risk predictions. *Sensors (Basel).* 2022;22(22):8615. https://doi.org/10.3390/s22228615.
23. Hicks S.A., Strümke I., Thambawita V. et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep.* 2022;12(1):5979. https://doi.org/10.1038/s41598-022-09954-8.
24. Liu T., Krentz A., Lu L., Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *Eur Heart J Digit Health.* 2024;6(1):7–22. https://doi.org/10.1093/ehjdh/ztae080.
25. Khandre V., Potdar J., Keerti A. Preterm birth: an overview. *Cureus.* 2022;14(12):e33006. https://doi.org/10.7759/cureus.33006.
26. Фомина А.С. Преждевременные роды, современные реалии. *Научные результаты биомедицинских исследований.* 2020;6(3):434–46. https://doi.org/10.18413/2658-6533-2020-6-3-0-12.

References:

1. Ivshin A.A., Pogodin O.O., Shakurova E.Yu. et al. Experience of laparoscopic transabdominal cerclage for the correction of cervical insufficiency during pregnancy: a clinical case and literature review. [Laparoskopicheskij transabdominal'nyj serklyazh dlya lecheniya istmiko-cervikal'noj nedostatochnosti pri beremennosti: klinicheskij sluchaj i obzor literatury]. *Obstetrics, Gynecology and Reproduction.* 2025;19(1):116–26. (In Russ.). https://doi.org/10.17749/2313-7347/ob.gyn.rep.2025.
2. Serov V.N., Sukhorukova O.I. Effectiveness of preterm birth prevention. [Effektivnost' profilaktiki prezhdevremennykh rodov]. *Akusherstvo i ginekologiya.* 2013;(3):48–53. (In Russ.).
3. Risnes K., Bilsteen J.F., Brown P. et al. Mortality among young adults born preterm and early term in 4 Nordic nations. *JAMA Netw Open.* 2021;4(1):e2032779. https://doi.org/10.1001/jamanetworkopen.2020.32779.
4. Jeon G.W., Lee J.H., Oh M., Chang Y.S. Serial long-term growth and neurodevelopment of very-low-birth-weight infants: 2022 update on the Korean Neonatal Network. *J Korean Med Sci.* 2022;37(34):e263. https://doi.org/10.3346/jkms.2022.37.e263.
5. Gorina K.A., Khodzhaeva Z.S., Belousov D.M. et al. Preterm birth: past limitations and new opportunities. [Prezhdevremennye rody: proshlye ogranicheniya i novye vozmozhnosti]. *Akusherstvo i ginekologiya.* 2020;(1):12–9. (In Russ.). https://doi.org/10.18565/aig.2020.1.12-19.
6. Stedall P.M., Spencer-Smith M.M., Lah S. et al. Episodic and prospective memory difficulties in 13-year-old children born very preterm. *J Int Neuropsychol Soc.* 2023;29(3):257265. https://doi.org/10.1017/S1355617722000170.
7. Belousova V.S., Strizhakov A.N., Svitich O.A. et al. Preterm birth: causes, pathogenesis, management tactics. [Prezhdevremennye rody: prichiny, patogenez, taktika]. *Akusherstvo i ginekologiya.* 2020;(2):82–7. (In Russ.). https://doi.org/10.18565/aig.2020.2.82-87.
8. Thain S., Yeo G.S.H., Kwek K., Chern B., Tan K.H. Spontaneous preterm birth and cervical length in a pregnant Asian population. *PLoS One.* 2020;15(4):e0230125. https://doi.org/10.1371/journal.pone.0230125.
9. Drukker N.A., Durnitsyna O.A., Nikashina A.A., Selyutina S.N. Diagnostic value of α-1-microglobulin in the development of preterm birth. [Diagnosticheskaya znachimost' α-1-mikroglobulina v razvitii prezhdevremennykh rodov]. *Akusherstvo i ginekologiya.* 2019;(1):81–5. (In Russ.). https://doi.org/10.18565/aig.2019.1.81-85.
10. Baev O.R., Dicke G.B. Diagnosis of premature rupture of fetal membranes based on biochemical tests. [Diagnostika prezhdevremennogo razryva plodnykh obolochek na osnove biokhimicheskikh testov]. *Akusherstvo i ginekologiya.* 2018;(9):132–6. (In Russ.). https://doi.org/10.18565/aig.2018.9.132-136.
11. Clinical guidelines – Preterm birth – 2020 (01.12.2020). [Klinicheskie rekomendacii – Prezhdevremennye rody – 2020 (01.12.2020)]. *Moscow: Ministerstvo zdravoohraneniya Rossijskoj Federacii,* 2020. 66 p. (In Russ.). Available at: https://cr.minzdrav.gov.ru/schema/331_1. [Accessed: 28.08.2025].
12. Manukhin I.B., Firichenko S.V., Mikailova L.U. et al. Prediction and prevention of preterm birth – current state of the problem. [Prognozirovanie i profilaktika prezhdevremennykh rodov – sovremennoe sostoyanie problemy]. *Rossiiskii vestnik akushera-ginekologa.* 2016;(3):9–15. (In Russ.). https://doi.org/10.17116/rosakush20161639-15.
13. Khodzhaeva Z.S., Dembovskaya S.V., Dobrokhotova Yu.E. et al. Pharmacological prevention of preterm birth (results of the international multicenter open MISTERY study). [Medikamentoznaya profilaktika prezhdevremennykh rodov (rezul'taty mezhdunarodnogo mnogotsentrovogo otkrytogo issledovaniya MISTERY)]. *Akusherstvo i ginekologiya.* 2016;(8):37–43. (In Russ.). https://doi.org/10.18565/aig.2016.8.37-43.
14. Barinov S.V., Artymuk N.V., Novikova O.N. et al. Management experience in pregnant women at high risk for preterm birth using an obstetric dome-shaped pessary and cerclage. [Opyt vedeniya beremennykh gruppy vysokogo riska po prezhdevremennym rodam s primeneniem akusherskogo kupoloobraznogo pessariya i serklyazha]. *Akusherstvo i ginekologiya.* 2019;(1):140–8. (In Russ.). https://doi.org/10.18565/aig.2019.1.140-148.
15. Crockart I.C., Brink L.T., du Plessis C., Odendaal H.J. Classification of intrauterine growthrestriction at 34-38 weeks gestation with machine learning models. *Inform Med Unlocked.* 2021;23:100533. https://doi.org/10.1016/j.imu.2021.100533.
16. Liu J., Wang C., Yan R. et al. Machine learning-based prediction of postpartum hemorrhage after vaginal delivery: combining bleeding high risk factors and uterine contraction curve. *Arch Gynecol Obstet.* 2022;306(4):1015–25. https://doi.org/10.1007/s00404-021-06377-0.
17. Melinte-Popescu A.S., Vasilache I.A., Socolov D., Melinte-Popescu M.

Predictive performance of machine learning-based methods for the prediction of preeclampsia – a prospective study. *J Clin Med*. 2023;12(2):418. https://doi.org/10.3390/jcm12020418.
18. Andreychenko A.E., Luchinin A.S., Ivshin A.A. et al. Development and validation of models to predict total and early-onset preeclampsia in the first trimester of pregnancy using machine learning algorithms. [Razrabotka i validatsiya modelei prognozirovaniya obshchego riska preeklampsii i riska rannei preeklampsii s ispol'zovaniem algoritmov mashinnogo obucheniya v pervom trimestre beremennosti]. *Akusherstvo i ginekologiya*. 2023;(10):94–107. (In Russ.). https://doi.org/10.18565/aig.2023.101.
19. Chen Y., Shi X., Wang Z., Zhang L. Development and validation of a spontaneous preterm birth risk prediction algorithm based on maternal bioinformatics: A single-center retrospective study. *BMC Pregnancy Childbirth*. 2024;24(1):763. https://doi.org/10.1186/s12884-024-06933-x.
20. Zhang Y., Du S., Hu T. et al. Establishment of a model for predicting preterm birth based on the machine learning algorithm. *BMC Pregnancy Childbirth*. 2023;23(1):779. https://doi.org/10.1186/s12884-023-06058-7.
21. Sun Q., Zou X., Yan Y. et al. Machine learning-based prediction model of preterm birth using electronic health record. *J Healthc Eng*. 2022;2022:9635526. https://doi.org/10.1155/2022/9635526.
22. Mavrogiorgou A., Kiourtis A., Kleftakis S. et al. A catalogue of machine learning algorithms for healthcare risk predictions. *Sensors (Basel)*. 2022;22(22):8615. https://doi.org/10.3390/s22228615.
23. Hicks S.A., Strümke I., Thambawita V. et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep*. 2022;12(1):5979. https://doi.org/10.1038/s41598-022-09954-8.
24. Liu T., Krentz A., Lu L., Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *Eur Heart J Digit Health*. 2024;6(1):7–22. https://doi.org/10.1093/ehjdh/ztae080.
25. Khandre V., Potdar J., Keerti A. Preterm birth: an overview. *Cureus*. 2022;14(12):e33006. https://doi.org/10.7759/cureus.33006.
26. Fomina A.S. Premature birth, modern realities. [Prezhdevremennye rody, sovremennye realii]. *Research Results in Biomedicine*. 2020;6(3):434–46. (In Russ.). https://doi.org/10.18413/2658-6533-2020-6-3-0-12.

Сведения об авторах / About the authors:

Болдина Юлия Сергеевна / **Yuliya S. Boldina**, MD. ORCID: https://orcid.org/0000-0002-1450-650X. Scopus Author ID: 57356202000. WoS ResearcherID: PII-8685-2026. eLibrary SPIN-code: 2944-0409.

Ившин Александр Анатольевич, к.м.н. / **Aleksandr A. Ivshin**, MD, PhD. E-mail: scipeople@mail.ru. ORCID: https://orcid.org/0000-0001-7834-096X. Scopus Author ID: 57222275843. WoS ResearcherID: AAG-1507-2020. eLibrary SPIN-code: 8196-6605.

Светова Кристина Сергеевна / **Kristina S. Svetova**. ORCID: https://orcid.org/0009-0001-5552-638X.