

УДК 519.683, 519-7

DOI: [10.26102/2310-6018/2020.30.3.025](https://doi.org/10.26102/2310-6018/2020.30.3.025)

Сравнение эффективности различных методов отбора признаков для решения задачи бинарной классификации предсказания наступления беременности при проведении экстракорпорального оплодотворения

С.Л. Синотова¹, О.В. Лимановская², А.Н. Плаксина³, В.А. Макутина⁴

^{1,2}ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина», Екатеринбург, Российская Федерация,

³ФГБОУ ВО «Уральский государственный медицинский университет Министерства здравоохранения Российской Федерации», Екатеринбург, Российская Федерация,

⁴АО «Центр семейной медицины», Екатеринбург, Российская Федерация

Резюме: Определение круга факторов, влияющих на объект исследования, является важнейшей задачей медицинских исследований. Ее решение осложняется большим числом разнообразных данных, включающих в себя обширную анамнестическую информацию и данные клинических исследований часто сочетающимся с ограниченным количеством наблюдаемых пациентов. Данная работа посвящена сравнению результатов, полученных различными методами отбора признаков для поиска набора предикторов, на основе которого создана модель с лучшим качеством прогноза, для решения задачи бинарной классификации предсказания наступления беременности при проведении экстракорпорального оплодотворения (ЭКО). В качестве признаков использовались данные анамнеза женщин, представленные в бинарном виде. Выборка состояла из 68 признаков и 689 объектов. Признаки были исследованы на наличие взаимной корреляции, после чего применены методы и алгоритмы для поиска отбора значимых факторов: непараметрические критерии, интервальная оценка долей, Z-критерий для разности двух долей, взаимная информация, алгоритмы RFECV, ADD-DELL, Relief, алгоритмы, основанные на важности перестановок (Boruta, Permutation Importance, PIMP), алгоритмы отбора признаков при помощи модели (lasso, random forest). Для сравнения качества отобранных наборов признаков построены различные классификаторы, посчитана их метрика AUC и сложность модели. Все модели имеют высокое качество предсказания (AUC выше 95%). Лучшие из них построены на признаках, отобранных с помощью непараметрических критериев, отбора при помощи модели (lasso-регрессия), алгоритмов Boruta, Permutation Importance, RFECV, ReliefF. Оптимальным набором предикторов был выбран набор, состоящий из 30 бинарных признаков, полученный алгоритмом Boruta, из-за меньшей сложности модели при сравнительно высоком качестве (AUC модели 0,983). К значимым признакам отнесены: данные о наличии беременностей в анамнезе в целом, о внематочных и замерших беременностях, самостоятельных и срочных родах, абортax на ранних сроках в частности; гипертония, ишемия, инсульт, тромбозы, язвы, ожирение, сахарный диабет у ближайших родственников; проведение гормонального лечения в настоящее время, не связанного с процедурой ЭКО; аллергия; вредные профессиональные факторы; наличие нормальной продолжительности и стабильности менструального цикла без приема медицинских препаратов; гистероскопия, лапароскопия и лапаротомия в анамнезе; проведение резекций любого органа в мочеполовой системе; первая ли попытка ЭКО, наличие любых хирургических вмешательств, заболеваний мочеполовой системы; возраст и ИМТ пациентки; отсутствие хронических заболеваний; наличие диффузной фиброзно-кистозной мастопатии, гипотиреоза.

Ключевые слова: отбор признаков, задача бинарной классификации, анализ малых данных, машинное обучение, вспомогательные репродуктивные технологии

Для цитирования: Синотова С.Л., Лимановская О.В., Плаксина А.Н., Макутина В.А. Сравнение эффективности различных методов отбора признаков для решения задачи бинарной классификации предсказания наступления беременности при проведении экстракорпорального оплодотворения. *Моделирование, оптимизация и информационные технологии*. 2020;8(3). Доступно по: https://moit.vivt.ru/wp-content/uploads/2020/08/SinotovaSoavtors_3_20_1.pdf DOI: 10.26102/2310-6018/2020.30.3.025

Comparison of the efficiency of different selecting features methods for solving the binary classification problem of predicting in vitro fertilization pregnancy

S.L. Sinotova¹, O.V. Limanovskaya², A.N. Plaksina³, V.A. Makutina⁴
^{1,2} *FSAEI HE «UrFU named after the first President of Russia B.N. Yeltsin»,
Ekaterinburg, Russian Federation*
³ *FSBEI HE «USMU of the Ministry of Health of the Russian Federation»,
Ekaterinburg, Russian Federation*
⁴ *The Family Medicine Centre, Ekaterinburg, Russian Federation*

Abstract: Determination of the range of factors affecting the object of research is the most important task of medical research. Its solution is complicated by a large amount of diverse data, including extensive anamnestic information and data from clinical studies, often combined with a limited number of observed patients. This work is devoted to the comparison of the results obtained by various feature selection methods for the search for a set of predictors, on the basis of which a model with the best forecast quality was created, for solving the problem of binary classification of predicting the onset of pregnancy during in vitro fertilization (IVF). The data from the anamnesis of women, presented in binary form, were used as features. The sample consisted of 68 features and 689 objects. The signs were examined for the presence of cross-correlation, after which methods and algorithms were applied to search for a selection of significant factors: nonparametric criteria, interval estimate of the shares, Z-criterion for the difference of two shares, mutual information, RFECV, ADD-DELL, Relief algorithms, algorithms based on the permutation importance (Boruta, Permutation Importance, PIMP), feature selection algorithms using model feature importance (lasso, random forest). To compare the quality of the selected sets of features, various classifiers were built, their metric AUC and the complexity of the model were calculated. All models have high prediction quality (AUC above 95%). The best of them are based on features selected using nonparametric criteria, model selection (lasso regression), Boruta, Permutation Importance, RFECV and ReliefF algorithms. The optimal set of predictors is a set of 30 binary features obtained by the Boruta algorithm, due to the lower complexity of the model with a relatively high quality (AUC of the model 0.983). Significant signs includes: data about pregnancies in the anamnesis in general, ectopic and regressive pregnancies, independent and term childbirth, abortions up to 12 weeks; hypertension, ischemia, stroke, thrombosis, ulcers, obesity, diabetes mellitus in the immediate family; currently undergoing hormonal treatment not associated with the IVF procedure; allergies; harmful professional factors; normal duration and stability of the menstrual cycle without taking medication; hysteroscopy, laparoscopy and laparotomy; resection of any organ in the genitourinary system; is it the first IVF, the presence of any surgical interventions, diseases of the genitourinary system; the age and BMI of the patient; absence of chronic diseases; the presence of diffuse fibrocystic mastopathy, hypothyroidism.

Keywords: feature selection, binary classification problem, small data analysis, machine learning, assisted reproductive technologies

For citation: Sinotova S.L., Limanovskaya O.V., Plaksina A.N., Makutina V.A. Comparison of the efficiency of different selecting features methods for solving the binary classification problem of predicting in vitro fertilization pregnancy. *Modeling, Optimization and Information Technology*. 2020;8(3). Available from: <https://moit.vivt.ru/wp->

content/uploads/2020/08/SinotovaSoavtors_3_20_1.pdf DOI: 10.26102/2310-6018/2020.30.3.025 (In Russ).

Введение

Экстракорпоральное оплодотворение (ЭКО) – медицинская технология, помогающая забеременеть паре при диагнозе бесплодие. К сожалению, далеко не каждый протокол заканчивается успешной имплантацией и рождением здорового ребенка. Прогнозирование исхода ЭКО важно не только для пациентов, но и для врачей. Оценку взаимного влияния большого числа данных анамнеза можно получить, воспользовавшись машинным обучением, предполагающим создание прогностической модели для исхода ЭКО, в данном случае наступления беременности, на основе анализа репрезентативной выборки с информацией об анамнезе и исходах беременностей женщин, прошедших через протоколы.

После обработки, очистки и нормализации данных, любой объект выборки описан с помощью вектора признаков, каждый из которых определен на некотором множестве допустимых значений. Решающую роль для построения прогнозирующей экспертной системы с помощью машинного обучения играет отбор значимых предикторов из признакового описания объекта.

Тем не менее работы по методикам отбора факторов для ЭКО крайне редки. Так в обзорной работе [1] проведено исследование группы работ по прогнозированию исходов методик вспомогательных репродуктивных технологий (ВРТ). Авторы отобрали наиболее релевантные работы в этой области и провели сравнение значимости используемых факторов статистическими методами анализа. И хотя целью работы являлось проведение исследования по формированию рекомендации к набору факторов при построении модели по исходам ВРТ, авторы, к сожалению, сосредоточились только на 8 факторах, используемых в протоколе ВРТ: возраст женщины, базальный ФСГ, продолжительность репродуктивной функции, указание на недостаточность репродуктивной функции, количество извлеченных яйцеклеток, способ оплодотворения, количество эмбрионов и качество перенесенных эмбрионов. Поэтому никаких методик по отбору факторов авторами в работе не было использовано.

В работе по определению предикторов наступления беременности у бесплодных пар [2] был проведен отбор факторов методом обратного пошагового отбора. В исследовании рассматривались следующие факторы – уровень фолликулостимулирующего гормона, возраст женщины, возраст мужчины, продолжительность бесплодия, тип бесплодия (вторичный или первичный), мужской фактор, способ вызова овуляции, количество фолликулов, размер доминирующего фолликула, количество спермы, подвижность спермы и общее прогрессивное количество подвижных сперматозоидов. На первом этапе исследования авторы рассчитали отношения шансов, показывающее силу связи фактора с вероятностью наступления беременности, для каждого из используемых в исследовании факторов. Далее была построена логистическая регрессионная модель на всех факторах. Для отбора наиболее значимых признаков был использован метод пошагового обратного отбора, при котором из набора факторов удаляли по одному наименее значимому фактору. В конечной модели были выделены три фактора наиболее значимых фактора - возраст женщины, тип бесплодия (первичный или вторичный) и подвижность спермы.

В работе [3] по выявлению предикторов рождения живого ребенка при проведении процедуры ВРТ для отбора факторов использовалась оценка их корреляции с значением выходной переменной (фактом рождения живого ребенка). Для оценки корреляции факторов с исходом использовался метод обобщённых оценочных

уравнений, часто применимый при линейной регрессии. Для исключения переобучения модели на втором этапе исследования были удалены коллинеарные факторы, имеющие корреляцию между собой выше 0,8. В конечном итоге из 100 факторов было отобрано 36.

В большинстве работ указан только один из возможных способов отбора признаков без проведения сравнения результатов, полученных при помощи других алгоритмов. Например, для отбора признаков в модель предсказания живорождения в результате проведения протоколов ВРТ авторы статьи [4] использовали корреляционный анализ (t-критерий, критерий χ^2). В другом исследовании [5] объединен генетический алгоритм и методы обучения принятию решений для интеллектуального анализа медицинской базы данных ЭКО, в работе [6] – алгоритм на основе поиска с восхождением к вершине для отбора признаков в задаче прогнозирования результатов ЭКО. Авторы [7] используют отбор на основе моделей, проводят сравнение результатов, полученных с помощью SVM, RPART, RF, Adaboost и 1NN классификаторов, выбирая признаки из модели RPART для предсказания наступления беременности после протокола ЭКО / интрацитоплазматического введения сперматозоида в ооцит (ИКСИ). Подробный анализ различных моделей, предсказывающих результаты ВРТ, проведен в статье [8]. Авторы пишут, что отбор признаков был произведен для построения 19 моделей из 20 рассмотренных. При просмотре данных исследований не было обнаружено проведение сравнения результатов различных способов отбора признаков.

В практике машинного обучения существует большое количество алгоритмов для отбора признаков, которые условно можно разделить на несколько групп по отношению выбора признаков к построению модели: фильтры (не привязаны к модели), обертки (используют модель при отборе), встраиваемые методы (отбирают признаки в процессе построения модели) [9,10,11]. По используемым техникам для поиска предикторов алгоритмы могут быть разнообразными: на основе жадного выбора, эвристических методов, перемешивания данных и другие [9,10,12,13,14,15,16,17].

Данная работа посвящена сравнению результатов, полученных различными методами отбора признаков для поиска набора предикторов, на основе которого создана модель с лучшим качеством прогноза, для решения задачи бинарной классификации предсказания наступления беременности при проведении экстракорпорального оплодотворения. В качестве признаков использовались данные анамнеза женщин, представленные в бинарном виде. Признаки были исследованы на наличие взаимной корреляции, после чего применены методы и алгоритмы для поиска отбора значимых факторов: непараметрические критерии, интервальная оценка долей, Z-критерий для доли, взаимная информация, алгоритмы RFECV, ADD-DELL, Relief, алгоритмы, основанные на важности перестановок (Boruta, Permutation Importance, PIMP), алгоритмы отбора признаков при помощи модели (lasso, random forest). Для сравнения качества отобранных наборов признаков построены различные классификаторы и посчитаны их метрики качества AUC.

Материалы

Данные получены из медицинской организации АО «Центр семейной медицины» за 2014 – 2018 гг. Была проведена предварительная обработка информации, заключающаяся в выделении признаков из теста, бинарном кодировании признаков, заполнении пропусков либо удалении экземпляров с пропущенными значениями, в зависимости от смысла признака, балансировке данных методом ADASYN [18,19]. Итоговая выборка содержала 68 бинарных признаков и 689 записей, из которых 351 экземпляр нулевого класса («отсутствие беременности»), 338 экземпляров первого

класса («наступление беременности»). Для удобства работы названия всех признаков закодированы. Единица означает наличие данного признака у объекта, 0 – его отсутствие.

Большинство характеристик имеют малое число положительных значений, что объясняется природой данных: эти признаки отвечают за наличие какого-либо заболевания, хирургического вмешательства, предыдущих беременностей женщины, вредных производственных факторов на работе или наличие заболеваний у близких родственников. Признаки, имеющие больше единиц, чем нулей: «регулярные менструации без гормональных препаратов», «нормальная продолжительность цикла», «проводилась ли женщине гистероскопия», «отсутствие заболеваний мочеполовой системы» и «ИМТ норма». Признаки, имеющие значительное число единиц (более чем у 25% объектов выборки): «отсутствие предыдущих беременностей», «проводилась лапароскопия», «возраст 31-35 лет», «планируется первая попытка зачатия беременности при помощи вспомогательных репродуктивных технологий».

Методы

Удаление коллинеарных признаков. Коллинеарными называются признаки, коррелирующие друг с другом. Наличие таких признаков может отрицательно сказаться на скорости обучения модели (особенно для итерационных методов), снизить интерпретируемость и производительность, способствует переобучению модели. При наличии коррелированных признаков алгоритмы, основанные на оценке важности перестановок признака могут ошибаться, поскольку при удалении одного из них модель не снизит качество. Следовательно, важность перестановки для них может быть низкой и все они будут отброшены вне зависимости от полезности. Из совокупности взаимно коррелирующих признаков можно удалить все кроме одного без значимой потери информации.

Отбор признаков при помощи фильтров. Фильтры направлены на оценку информативности каждого признака для целевой переменной [12]. Наиболее часто встречаются в задачах классификации такие методы-фильтры, как корреляция и взаимная информация.

Один из самых популярных и простых методов – корреляционный анализ, оценивающий наличие связи с помощью коэффициента корреляции. Подход строится на обнаружении признаков, коррелирующих с целевой переменной, но не коррелирующих друг с другом. Вместо обычных критериев значимости можно использовать доверительные интервалы (ДИ). Существует мнение, что ДИ лучше справляются с задачами нахождения различий между группами [20, 21]. Для отбора бинарных признаков используют доверительные интервалы для доли. Принцип заключается в группировании данных по каждому признаку: первую подвыборку будут составлять образцы, у которых признак присутствует, вторую – отсутствует. Далее строятся доверительные интервалы для каждой выборки (чаще 95% ДИ). В случае их не пересечения, можно сделать вывод о том, что выборки различны, то есть признак значим. [20]. Для данных из распределения Бернулли можно применить Z-критерий для разности двух долей, проверяющий справедливость нулевой гипотезы о равенстве выборок. Это можно делать для любой двусторонней или односторонней альтернативы [20].

Еще одним примером одномерного оценивания качества признаков является подход, основанный на использовании взаимной информации [9]. Взаимная информация – метрика теории информации, показывающая, сколько информации содержится в одной величине относительно другой. Если она равна нулю, то величины независимы,

информативность признаков оценивают по модулю взаимной информации. Метрика используется для работы с дискретными данными и часто является составной частью других алгоритмов для отбора признаков [13,22,23].

Существуют и другие методы-фильтры для отбора признаков, например, Relief [14], который является прародителем прочих методов на его основе, разработанных для различных задач и входных данных [24, 25]. Relief рассчитывает значимость (вес) каждого признака, работает следующим образом: Первоначально все признаки имеют нулевые веса. Алгоритм выбирает произвольный элемент R (объект, строка в матрице признаков) и ищет к нему двух ближайших соседей: одного из того же класса («ближайшее попадание») - N и одного из противоположного («ближайший промах») - M. Обновляются веса всех признаков в зависимости от выбранного экземпляра и соседей: если R и N имеют разные значения признака X, то он разделяет объекты одного класса, что понижает его вес. Если R и M имеют разные значения признака X, то он разделяет объекты разных классов, что увеличивает его значимость, а значит и вес. Эти шаги повторяются заданное число раз, после чего делается вывод о значимости признаков на основе их веса.

Методы – фильтры хороши для данных высокой размерности, когда другие методы работают медленнее, устойчивы к переобучению [26], их целесообразно применять как предобработку для вычислительно более затратных методов. Не учитывают тип прогнозирующей модели. Одними из серьезных недостатков являются их неспособность оценивать влияние совокупности признаков на отклик.

Отбор признаков с помощью обертки. Принцип работы обертки – перебор подмножеств признаков и выбор из них такого набора, который оказывается наиболее эффективным, то есть дает модель лучшего качества. Поскольку полный перебор имеет большую вычислительную сложность (работа с 2^n подмножествами), применим он только для очень малого числа признаков. Большинство методов-обертки построено на основе жадной стратегии с помощью прямого или обратного перебора признаков [9].

Один традиционных методов этого класса – метод рекурсивного исключения признаков (recursive feature elimination, RFE [10]), исключающий признаки в порядке улучшения качества модели. В первую очередь тренируется и оценивается модель, построенная с использованием всех признаков. Далее признаки ранжируются по вкладу и удаляются из модели в порядке увеличения значимости, до тех пор, пока не останется требуемое число лучших признаков. Отбирает признаки под конкретную модель. Когда число отбираемых признаков заранее неизвестно, можно применить RFECV (recursive feature elimination and cross-validated selection), который предполагает выбор лучшего набора предикторов на основе кросс-валидации [27].

Другим примером можно назвать алгоритм ADD-DEL. Начинается алгоритм с выбора пары предикторов, которые максимизируют/минимизируют критерий отбора для выбранной модели, затем поочередно просматриваются остальные признаки: если добавление признака улучшает значение метрики качества, то признак добавляется в множество отобранных. Если за проход ни одного признака не было добавлено, начинается процесс удаления признаков из подмножества значимых. На каждом этапе удаления, убирается такой признак, при котором значение метрики качества является оптимальным на оставшемся наборе. Если за проход нет ни одного удаления, то прерываем удаление, возвращаемся к добавлению. Процесс завершается, когда за проход алгоритма не будет ни одного добавления и ни одного удаления [12].

В отличие от фильтров обертки позволяют обнаружить влияние признаков в совокупности и нелинейные зависимости, но обладают большей вычислительной сложностью, чем фильтры: на каждом шаге оценки модель обучается заново.

Отбор признаков на основе значимости при построении модели. Данный метод заключается в отборе признаков, которые имели наибольшую значимость в процессе построения конкретной модели [9]. Для получения оценки значимости в библиотеке Scikit-learn [28], можно посмотреть на параметр `feature_importance` построенной модели. Такой подход нельзя в полной мере назвать осознанным отбором признаков, поскольку оценки важности зачастую носят случайный характер. Например, при использовании моделей на основе деревьев есть риск, что после того как в отдельном дереве один из признаков сильно продвинет модель в решении задачи и получит большую значимость, другой, схожий по полезности, не будет оценен высоко и не попадет в отбор, поэтому `feature_importance` может и не отражать реальную значимость признаков для отклика, а иметь смысл только для одной конкретной построенной модели [29]. Тем не менее такой подход встречается на практике и часто составляет часть других алгоритмов для отбора признаков. Любая модель, которая ранжирует признаки может являться моделью для подобного отбора, кроме упомянутых моделей на основе дерева, на практике часто встречается линейная.

В линейной модели при масштабированных признаках их веса можно оценивать, как значимость для модели: чем больше вес, тем большее значение. Для обнуления веса при признаке применяется L1-регуляризатор (lasso-регрессия). Метод заключается в введении дополнительного параметра регуляризации в функционал оптимизации модели, величина которого равна минимальной ошибке прогноза на наблюдениях, не участвующих в построении модели [30].

Отбор признаков на основе оценки важности перестановок. Хорошие результаты при отборе признаков можно получить с помощью алгоритмов, которые используют для решения задачи перестановки данных. О важности признака можно судить по тому, как сильно у модели снижается качество без его использования. Алгоритмы-обертки способны оценить подобную важность использования признака для набора данных, но имеют недостаток в виде построения модели для каждого нового набора. Библиотека `eli5` [31] представляет алгоритм – обертку `PermutationImportance`, в котором признак удаляется только из тестовой части набора данных, что не требует перестроения модели, но и показывает важность только для нее. Поскольку модель ожидает определенное число параметров, то удаление заменяется заменой столбца на шум – перетасовку значений признака. – так вычисляется важность перестановки. Из-за значительной ресурсоемкости метод лучше применять для данных с небольшим числом признаков.

Еще один алгоритм – обертка, предполагающая перемешивание данных, – `boruta` [15]. В основе алгоритма лежит модель случайного леса. `Boruta` сравнивает важность функции (`feature importance`), полученную после построения модели, с важностью функции, полученной с помощью перемешанной копии набора данных (так называемые теневые признаки). Каждый раз, когда признак имеет более высокий Z-показатель, чем максимальный Z-показатель его теневых объектов, чем у лучших теневых объектов, `Boruta` записывает «попадания» для этого объекта. Если функция не записывает попадание в определенном количестве итераций, признак считается не важным.

Другим примером алгоритма, основанного на изменении важности перестановок и оценке значимости при построении модели, является эвристический алгоритм отбора признаков `PIMP` [13]. Метод может быть использован вместе с любым методом отбора, который рассчитывает значимость каждого признака (в данной работе со случайным лесом). Алгоритм измеряет увеличение ошибки предсказания модели после того, как переставляются значения вектора ответов, что нарушает связь между признаком и откликом. Признак значим, если перестановка увеличивает ошибку модели. Преимущество перестановки ответов заключается в сохранении зависимости между

признаками и возможности сделать перестановок меньше, чем число признаков. В данном методе изначально оценивается важность признаков (feature_importance) с помощью основного метода отбора. Далее для каждой перестановки вектора ответов оценивается релевантность всех признаков. Так для каждого признака формируется вектор null_importance, значениями которого являются важности признака в каждой перестановке. В заключении рассчитывается фактическая важность признака (actual_importance) по null_importance, с помощью параметров распределения, которому принадлежат значения null_importance [13].

Инструменты. Обработка данных, отбор признаков и построение классификаторов выполнялись с использованием Anaconda — дистрибутива Python 3.6 [32]. Для применения всех описанных выше методов использовались их реализации в библиотеках python: SciPy [33] (для корреляционного анализа), Scikit-learn [17] (для вычисления взаимной информации, алгоритма RFECV, моделей LogisticRegression, RandomForestClassifier), ReliefF [34], LightGBM[35] (для использования LGBMClassifier с параметром RF при реализации алгоритма изменения важности перестановок (PIMP), взятой на платформе Kaggle [36]), boruta_py [37] и eli5 [31] для алгоритма PermutationImportance. Для представления данных, вычислений, построения графиков использовались библиотеки NumPy [38], pandas [39], matplotlib [40], seaborn [41].

Оценка качества отбора признаков. Оценку того, насколько хорошо был произведен отбор признаков, можно получить, построив и оценив модель на данных предикторах. На каждой группе выбранных признаков построены классификаторы из библиотеки Scikit-learn [28]: LogisticRegression, RandomForestClassifier, GradientBoostingClassifier. Подбор параметров моделей осуществлялся при помощи библиотеки Hyperport [42], в качестве критериев значимости использованы AUC (площадь под ROC-кривой) и сложность модели - число используемых ей признаков. Значения AUC посчитаны как среднее арифметическое значение для ста разбиений выборки на обучающую и тестовую. Разбиения выполнялись в отношении 70% - обучающая выборка, 30% - тестовая.

Результаты

Удаление коллинеарных признаков. Корреляция признаков с друг другом была отражена с помощью матрицы корреляций. Диапазон корреляций: [-0,61;1]. Корреляция у большинства признаков очень слабая (< 0.2) [43] отрицательная. Обратная связь средней степени (-0.61 и -0.54), максимальная для данной матрицы, установлена между признаком 37 (аденомиоз) и 41 (нет заболеваний мочеполовой системы), признаками 41 и 42 (нет и одно заболевание мочеполовой системы), что объясняется распределением значений в признаках. Также есть две положительные связи средней степени, значения которых 0.52 для корреляции признаков 18 (гормональное лечение в настоящее время) и 66 (гипотиреоз) и 0.51 для признаков 1 (внематочная беременность) и 30 (лапаротомия). В целом связи не настолько сильные, чтобы исключать какие-либо признаки из-за дублирования.

Отбор признаков. После проведения отбора признаков при помощи непараметрических критериев, Z-критерия для доли, взаимной информации, алгоритмов ReliefF, PIMP, Boruta, PermutationImportance, RFECV, Add-Del, моделей Lasso и Random Forest было получено различное число признаков.

Меньше всего значимых признаков оставили алгоритмы PIMP (27 признаков), Add-Dell (26 признаков) для модели градиентного бустинга и отбор с помощью модели случайного леса (24 признака).

На каждом полученном наборе признаков были построены модели с помощью библиотеки scikit-learn [28]: логистическая регрессия (Logistic Regression), случайный лес (RandomForestClassifier), градиентный бустинг (GradientBoostingClassifier). Качество моделей оценено с помощью метрики AUC, дающей количественную оценку площади, ограниченной ROC-кривой и осью доли ложных положительных классификаций. Сравнение AUC моделей, построенных на выбранных признаках и число признаков, отобранных каждым методом представлены в таблице 1. Результатами применения обёрток (алгоритмы PermutationImportance, RFECV и Add-Del) представлены отдельным набором признаков для каждой модели.

Таблица 1 – число признаков, отобранных каждым методом отбора и значения AUC моделей, построенных на этих признаках

Table 1 - the number of features selected by each selection method and the AUC values of the models built on these features

№	Название метода отбора	число признаков	AUC (Logistic Regression)	AUC (Random Forest Classifier)	AUC (Gradient Boosting Classifier)
1	Непараметрические критерии	49	0,971	0,983	0,976
2	Интервальная оценка долей	46	0,971	0,981	0,975
3	Z-критерий для разности двух долей	50	0,945	0,971	0,951
4	Взаимная информация	53	0,967	0,980	0,968
5	ReliefF	36	0,978	0,980	0,981
6	Lasso	30	0,976	0,978	0,979
7	Random Forest	24	0,954	0,977	0,963
8	PIMP	27	0,961	0,979	0,969
9	Boruta	30	0,967	0,983	0,976
10	PermutationImportance	48 / 45 / 41	0,968	0,985	0,980
11	RFECV	58 / 58 / 38	0,973	0,987	0,981
12	Add-Del	49 / 37 / 26	0,962	0,968	0,963

Обсуждение

Коллинеарных признаков в данных обнаружено не было, поэтому все признаки были использованы при поиске лучшего набора предикторов. Модели, полученные на всех наборах, имеют высокое качество классификации (AUC выше 95%). Лучшими для задачи оказались методы отбора с помощью непараметрических критериев, интервальной оценки долей, алгоритмов Boruta, PermutationImportance и RFECV.

Поскольку при отборе признаков необходимо соблюсти баланс между количеством признаков и качеством модели, то для выбора лучшего набора признаков был использован поиск удовлетворяющего варианта с оптимизацией параметров (Satisficing and optimizing) [44]. Осуществлялось введение порогового значения для одного критерия – среднего качества классификации по метрике AUC (98%) и выбора минимума для второго критерия - сложности модели, измеряемой числом признаков.

Таким образом, оптимальным набором предикторов выбран набор, полученный алгоритмом Boruta из-за меньшего числа элементов при высоком качестве прогнозирующей модели (AUC = 0,983). Этот набор включает в себя 30 признаков бинарных признаков, отражающих следующую информацию из анамнеза пациентки:

данные о наличии беременностей в анамнезе в целом, о внематочных и замерших беременностях, самостоятельных и срочных родах, абортах на ранних сроках в частности; гипертония, ишемия, инсульт, тромбозы, язвы, ожирение, сахарный диабет у ближайших родственников; проведение гормонального лечения в настоящее время, не связанного с процедурой ЭКО; аллергия; вредные профессиональные факторы; наличие нормальной продолжительности и стабильности цикла без приема медицинских препаратов; гистероскопия, лапароскопия и лапаротомия в анамнезе; проведение резекций любого органа мочеполовой системы; первая ли попытка ЭКО, наличие любых хирургических вмешательств, заболеваний мочеполовой системы; возраст и ИМТ пациентки; отсутствие хронических заболеваний; наличие диффузной фиброзно-кистозной мастопатии, гипотиреоза.

Заключение

В работе было проведено исследование 12 методов (непараметрические критерии, интервальная оценка долей, Z-критерий для доли, взаимная информация, алгоритмы RFECV, ADD-DELL, Relief, алгоритмы, основанные на важности перестановок (Boruta, Permutation Importance, PIMP), алгоритмы отбора признаков при помощи модели (lasso, random forest)) для поиска набора предикторов, на основе которого создана модель с лучшим качеством прогноза, для решения задачи бинарной классификации предсказания наступления беременности при проведении экстракорпорального оплодотворения. Руководствуясь критериями числа признаков и качества модели, был выбран оптимальный набор, состоящий из 30 бинарных признаков. Построенная на нем модель показала высокое качество ($AUC > 98\%$).

Применение методов отбора признаков, используемых ранее в сфере машинного обучения, показало хорошие результаты и может быть успешно использовано в задачах построения медицинских прогностических моделей.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-37-90121.

Acknowledgments: The reported study was funded by RFBR, project number 19-37-90121.

ЛИТЕРАТУРА

1. van Loendersloot L.L., van Wely M., Limpens J., Bossuyt P.M., Repping S., van der Veen F. Predictive factors in in vitro fertilization (IVF): a systematic review and meta-analysis. *Hum Reprod Update*. 2010;16(6):577–589. DOI: 10.1093/humupd/dmq015
2. Atasever M., Namlı Kalem M., Hatırnaz Ş., Hatırnaz E., Kalem Z., Kalaylıoğlu Z. Factors affecting clinical pregnancy rates after IUI for the treatment of unexplained infertility and mild male subfertility. *J Turk Ger Gynecol Assoc*. 2016;17:134–138. DOI: 10.5152/jtgga.2016.16056
3. Vaegter K.K., Lakic T.G., Olovsson M., Berglund L., Brodin T., Holte J. Which factors are most predictive for live birth after in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments? Analysis of 100 prospectively recorded variables in 8,400 IVF/ICSI single-embryo transfers. *Fertil Steril*. 2017;107(3):641–648.e2. DOI:10.1016/j.fertnstert.2016.12.005

4. Vogiatzi, P., Pouliakis, A., Siristatidis, C. An artificial neural network for the prediction of assisted reproduction outcome. *J Assist Reprod Genet.* 2019;36:1441–1448. DOI: 10.1007/s10815-019-01498-7
5. Ruey-Shiang Guh, Tsung-Chieh Jackson Wu, Shao-Ping Weng. Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes. *Expert Systems with Applications.* 2011;38(4):4437–4449. DOI: 10.1016/j.eswa.2010.09.112
6. Hassan M.R., Al-Insaif S., Hossain M.I., Kamruzzaman J. A machine learning approach for prediction of pregnancy outcome following IVF treatment. *Neural Comput & Applic.* 2020;32:2283–2297. DOI: 10.1007/s00521-018-3693-9
7. Hafiz P., Nematollahi M., Boostani R., Namavar Jahromi B. Predicting Implantation Outcome of In Vitro Fertilization and Intracytoplasmic Sperm Injection Using Data Mining Techniques. *Int J Fertil Steril.* 2017;11(3):184–190. DOI: 10.22074/ijfs.2017.4882
8. Raef B, Ferdousi R. A Review of Machine Learning Approaches in Assisted Reproductive Technologies. *Acta Inform Med.* 2019;27(3):205–211. DOI:10.5455/aim.2019.27.205-211
9. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 2003;3:1157–1182.
10. Guyon, I., Weston, J., Barnhill, S., Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning.* 2002;46:389–422. DOI: 10.1023/A:1012487302797
11. Saeys Y., Inza I., Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics.* 2007;23(19):2507–2517. DOI: 10.1093/bioinformatics/btm344
12. Воронцов К. В. Лекции по методам оценивания и выбора моделей. Доступно по: <http://www.ccas.ru/voron/download/Modeling.pdf> (дата обращения 18.08.2020)
13. Altmann A., Toloşi L., Sander O., Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics.* 2010;26(10): 1340–1347. DOI: 10.1093/bioinformatics/btq134
14. Kenji K., Rendell A. L. The feature selection problem: traditional methods and a new algorithm. *AAAI.* 1992;129–134
15. Kursu, M., Rudnicki. Feature Selection with the Boruta Package. *Journal of Statistical Software.* 2010;36(11):1–13. DOI: 10.18637/jss.v036.i11
16. Mazaheri V., Khodadadi H. Heart arrhythmia diagnosis based on the combination of morphological, frequency and nonlinear features of ECG signals and metaheuristic feature selection algorithm. *Expert Systems with Applications.* 2020;161:113697. DOI: 10.1016/j.eswa.2020.113697
17. Faris H., Mafarja M.M., Heidari A.A., Aljarah I., Al-Zoubi A.M., Mirjalili S., Fujita H. An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems. *Knowledge-Based Systems.* 2018;154:43–67. DOI: 10.1016/j.knosys.2018.05.009
18. He H., Bai Y., Garcia E.A., Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).* 2008;1322–1328. DOI: 10.1109/IJCNN.2008.4633969
19. Lemaître G., Nogueira F., Aridas C.K. Imbalanced-learn: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *JMLR.* 2017;18(17):1–5.
20. Гланц С. Медико-биологическая статистика. Пер. с англ. М.: Практика;1998.

21. Rothman K.J. A Show of Confidence. *N Engl J Med.* 1978;299(24):1362–1363. DOI: 10.1056/NEJM197812142992410
22. Das A.K., Kumar S., Jain S., Goswami S., Chakrabarti A., Chakraborty B. An information-theoretic graph-based approach for feature selection. *Sādhanā.* 2020;45:11. DOI: 10.1007/s12046-019-1238-2
23. Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks.* 1994;5(4):537–550. DOI: 10.1109/72.298224
24. Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. *Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence).* 1994;784:171–182.
25. Robnik-Sikonja M., Kononenko I. An adaptation of Relief for attribute estimation in regression. *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning.* 1997;296–304.
26. Hamon J. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale. Applications [stat.AP]. Université des Sciences et Technologie de Lille - Lille I, 2013. Français. fftel-00920205
27. Реализация алгоритма RFECV в библиотеке Scikit-learn. Доступно по: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html#sklearn.feature_selection.RFECV (дата обращения 18.08.2020)
28. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: Machine Learning in Python. *JMLR.* 2011;12(85):2825–2830.
29. Натекин А. Градиентный бустинг: возможности, особенности и фишки за пределами стандартных kaggle-style задач. *Moscow Data Science Meetup.* 2017. Доступно по: https://www.youtube.com/watch?time_continue=746&v=cM2c-47Xlqk&feature=emb_logo (дата обращения 18.08.2020)
30. Шитиков В. К., Мاستицкий С. Э. Классификация, регрессия, алгоритмы Data Mining с использованием R. 2017. Доступно по: <https://github.com/ranalytics/data-mining>
31. Библиотека ELI5. Доступно по: <https://eli5.readthedocs.io/en/latest/index.html#> (дата обращения 18.08.2020)
32. Платформа для обработки данных и машинного обучения Anaconda. Доступно по: <https://www.anaconda.com> (дата обращения 18.08.2020)
33. Библиотека SciPy. Доступно по: <https://www.scipy.org/index.html> (дата обращения 18.08.2020)
34. Библиотека ReliefF. Доступно по: <https://pypi.org/project/ReliefF/#description> (дата обращения 18.08.2020)
35. Библиотека LightGBM. Доступно по: <https://lightgbm.readthedocs.io/en/latest/index.html#> (дата обращения 18.08.2020)
36. Grellier O. Feature Selection with Null Importances. Статья на Kaggle. Доступно по: <https://www.kaggle.com/ogrellier/feature-selection-with-null-importances> (дата обращения 18.08.2020)
37. Python-реализация метода Boruta. Доступно по: https://github.com/scikit-learn-contrib/boruta_py (дата обращения 18.08.2020)
38. Библиотека NumPy. Доступно по: <https://numpy.org/> (дата обращения 18.08.2020)
39. Библиотека pandas. Доступно по: <https://pandas.pydata.org/> (дата обращения 18.08.2020)
40. Библиотека Matplotlib. Доступно по: <https://matplotlib.org/index.html> (дата обращения 18.08.2020)

41. Библиотека seaborn. Доступно по: <https://seaborn.pydata.org/#> (дата обращения 18.08.2020)
42. Bergstra, J., Yamins D., Cox D.D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *JMLR Workshop and Conference Proceedings*. 2013;28(1):115–123.
43. Гржибовский А. М. Анализ номинальных данных (независимые наблюдения). *Экология человека*. 2008;6:58-68.
44. Ng A. Machine Learning Yearning. Доступно по: <https://www.mlyearning.org/> (дата обращения 18.08.2020)

REFERENCES

1. van Loendersloot L.L., van Wely M., Limpens J., Bossuyt P.M., Repping S., van der Veen F. Predictive factors in in vitro fertilization (IVF): a systematic review and meta-analysis. *Hum Reprod Update*. 2010;16(6):577–589. DOI: 10.1093/humupd/dmq015
2. Atasever M., Namlı Kalem M., Hatırnaz Ş., Hatırnaz E., Kalem Z., Kalaylıoğlu Z. Factors affecting clinical pregnancy rates after IUI for the treatment of unexplained infertility and mild male subfertility. *J Turk Ger Gynecol Assoc*. 2016;17:134–138. DOI: 10.5152/jtgga.2016.16056
3. Vaegter K.K., Lakic T.G., Olovsson M., Berglund L., Brodin T., Holte J. Which factors are most predictive for live birth after in vitro fertilization and intracytoplasmic sperm injection (IVF/ICSI) treatments? Analysis of 100 prospectively recorded variables in 8,400 IVF/ICSI single-embryo transfers. *Fertil Steril*. 2017;107(3):641–648.e2. DOI:10.1016/j.fertnstert.2016.12.005
4. Vogiatzi, P., Pouliakis, A., Siristatidis, C. An artificial neural network for the prediction of assisted reproduction outcome. *J Assist Reprod Genet*. 2019;36:1441–1448. DOI: 10.1007/s10815-019-01498-7
5. Ruey-Shiang Guh, Tsung-Chieh Jackson Wu, Shao-Ping Weng. Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes. *Expert Systems with Applications*. 2011;38(4):4437–4449. DOI: 10.1016/j.eswa.2010.09.112
6. Hassan M.R., Al-Insaif S., Hossain M.I., Kamruzzaman J. A machine learning approach for prediction of pregnancy outcome following IVF treatment. *Neural Comput & Applic*. 2020;32:2283–2297. DOI: 10.1007/s00521-018-3693-9
7. Hafiz P., Nematollahi M., Boostani R., Namavar Jahromi B. Predicting Implantation Outcome of In Vitro Fertilization and Intracytoplasmic Sperm Injection Using Data Mining Techniques. *Int J Fertil Steril*. 2017;11(3):184–190. DOI: 10.22074/ijfs.2017.4882
8. Raef B, Ferdousi R. A Review of Machine Learning Approaches in Assisted Reproductive Technologies. *Acta Inform Med*. 2019;27(3):205–211. DOI:10.5455/aim.2019.27.205-211
9. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J. Mach. Learn. Res*. 2003;3:1157–1182.
10. Guyon, I., Weston, J., Barnhill, S., Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*. 2002;46:389–422. DOI: 10.1023/A:1012487302797
11. Saeys Y., Inza I., Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507–2517. DOI: 10.1093/bioinformatics/btm344
12. Voroncov K. V. Lekcii po metodam ocenivaniya i vybora modelej. Available at: <http://www.ccas.ru/voron/download/Modeling.pdf> (accessed 18.08.2020) (In Russ)

13. Altmann A., Tološi L., Sander O., Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics*. 2010;26(10):1340–1347. DOI: 10.1093/bioinformatics/btq134
14. Kenji K., Rendell A. L. The feature selection problem: traditional methods and a new algorithm. *AAAI*. 1992;129–134.
15. Kursu, M., Rudnicki. Feature Selection with the Boruta Package. *Journal of Statistical Software*. 2010;36(11):1–13. DOI: 10.18637/jss.v036.i11
16. Mazaheri V., Khodadadi H. Heart arrhythmia diagnosis based on the combination of morphological, frequency and nonlinear features of ECG signals and metaheuristic feature selection algorithm. *Expert Systems with Applications*. 2020;161:113697. DOI: 10.1016/j.eswa.2020.113697
17. Faris H., Mafarja M.M., Heidari A.A., Aljarah I., Al-Zoubi A.M., Mirjalili S., Fujita H. An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems. *Knowledge-Based Systems*. 2018;154:43–67. DOI: 10.1016/j.knsys.2018.05.009
18. He H., Bai Y., Garcia E.A., Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008;1322–1328. DOI: 10.1109/IJCNN.2008.4633969
19. Lemaître G., Nogueira F., Aridas C.K. Imbalanced-learn: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *JMLR*. 2017;18(17):1–5.
20. Glantz S. Primer of biostatistics. M.: Practica;1998. (In Russ)
21. Rothman K.J. A Show of Confidence. *N Engl J Med*. 1978;299(24):1362–1363. DOI: 10.1056/NEJM197812142992410
22. Das A.K., Kumar S., Jain S., Goswami S., Chakrabarti A., Chakraborty B. An information-theoretic graph-based approach for feature selection. *Sādhanā*. 2020;45:11. DOI: 10.1007/s12046-019-1238-2
23. Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*. 1994;5(4):537–550. DOI: 10.1109/72.298224
24. Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. *Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*. 1994;784:171–182.
25. Robnik-Sikonja M., Kononenko I. An adaptation of Relief for attribute estimation in regression. *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*. 1997;296–304.
26. Hamon J. Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale. Applications [stat.AP]. Université des Sciences et Technologie de Lille - Lille I, 2013. Français. fftel-00920205
27. Implementation of the algorithm RFECV in Scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html#sklearn.feature_selection.RFECV (accessed 18.08.2020)
28. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: Machine Learning in Python. *JMLR*. 2011;12(85):2825–2830.
29. Natekin A. Gradientnyj busting: vozmozhnosti, osobennosti i fishki za predelami standartnyh kaggle-style zadach. *Moscow Data Science Meetup*. 2017. Available at: https://www.youtube.com/watch?time_continue=746&v=cM2c-47Xlqk&feature=emb_logo (accessed 18.08.2020) (In Russ)

30. Shitikov V. K., Mastickij S. Je. Klassifikacija, regressija, algoritmy Data Mining s ispol'zovaniem R. 2017. Available at: <https://github.com/ranalytics/data-mining> (accessed 18.08.2020) (In Russ)
31. ELI5 library. Available at: <https://eli5.readthedocs.io/en/latest/index.html#> (accessed 18.08.2020)
32. Anaconda - solutions for Data Science Practitioners and Enterprise Machine Learning. Available at: <https://www.anaconda.com> (accessed 18.08.2020)
33. SciPy library. Available at: <https://www.scipy.org/index.html> (accessed 18.08.2020)
34. ReliefF library. Available at: <https://pypi.org/project/ReliefF/#description> (accessed 18.08.2020)
35. LightGBM library. Available at: <https://lightgbm.readthedocs.io/en/latest/index.html#> (accessed 18.08.2020)
36. Grellier O. Feature Selection with Null Importances. Article on the Kaggle. Available at: <https://www.kaggle.com/ogrellier/feature-selection-with-null-importances> (accessed 18.08.2020)
37. Boruta implementation in Python. Available at: https://github.com/scikit-learn-contrib/boruta_py (accessed 18.08.2020)
38. NumPy library. Available at: <https://numpy.org/> (accessed 18.08.2020)
39. Pandas library. Available at: <https://pandas.pydata.org/> (accessed 18.08.2020)
40. Matplotlib library. Available at: <https://matplotlib.org/index.html> (accessed 18.08.2020)
41. Seaborn library. Available at: <https://seaborn.pydata.org/#> (accessed 18.08.2020)
42. Bergstra, J., Yamins D., Cox D.D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *JMLR Workshop and Conference Proceedings*. 2013;28(1):115–123.
43. Grijbovski A.M. Analysis of nominal data (independent observations). *Human Ecology*. 2008;6:58-68. (In Russ)
44. Ng A. Machine Learning Yearning. Available at: <https://www.mlyearning.org/> (accessed 18.08.2020)

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Синотова Светлана Леонидовна, аспирант, младший научный сотрудник кафедры интеллектуальных информационных технологий института фундаментального образования ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина», Екатеринбург, Российская Федерация
e-mail: sveta.volkova92@mail.ru
ORCID: [0000-0002-4318-5223](https://orcid.org/0000-0002-4318-5223)

Svetlana L. Sinotova, P.G., Junior Researcher Of The Department Of Intellectual Information Technologies, Institute Of Fundamental Education, Fsaei He «Urfu Named After The First President Of Russia B.N.Yeltsin», Ekaterinburg, Russian Federation

Лимановская Оксана Викторовна, к.х.н., доцент кафедры интеллектуальных информационных технологий института фундаментального образования ФГАОУ ВО «УрФУ имени первого Президента России Б.Н. Ельцина», Екатеринбург, Российская Федерация
e-mail: o.v.limanovskaia@urfu.ru
ORCID: [0000-0002-2084-3916](https://orcid.org/0000-0002-2084-3916)

Oksana V. Limanovskaya, Candidate of Chemical Sciences, Associate Professor of The Department of Intellectual Information Technologies, Institute of Fundamental Education, FSAEI HE «UrFU named after the first President of Russia B.N. Yeltsin», Ekaterinburg, Russian Federation

Плаксина Анна Николаевна, к.м.н., ассистент кафедры физической и реабилитационной медицины ФГБОУ ВО «Уральский государственный медицинский университет Минздрава РФ», Екатеринбург, Российская Федерация

e-mail: burberry20@yandex.ru

ORCID: [0000-0002-3119-478X](https://orcid.org/0000-0002-3119-478X)

Anna N. Plaksina, Candidate of Medical Sciences, Assistant of The Department of Physical and Rehabilitation Medicine, FSBEI HE «USMU of the Ministry of Health of the Russian Federation», Ekaterinburg, Russian Federation

Макутина Валерия Андреевна, к.б.н., эмбриолог АО «Центр семейной медицины», Екатеринбург, Российская Федерация

e-mail: makutina_v@rambler.ru

ORCID: [0000-0003-1127-2792](https://orcid.org/0000-0003-1127-2792)

Valerija A. Makutina, Candidate of Biological Sciences, Embryologist of The Family Medicine Centre, Ekaterinburg, Russian Federation