

Разработка модели машинного обучения для интерпретации результатов лабораторной диагностики с целью выявления подозрений на заболевания

© А.В. ГУСЕВ^{1, 2, 3}, Г.Г. ГАВРИЛЕНКО¹, Д.В. ГАВРИЛОВ¹

¹ООО «К-Скай», Петрозаводск, Республика Карелия, Россия;

²ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Минздрава России, Москва, Россия;

³ГБУЗ Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения Москвы», Москва, Россия

РЕЗЮМЕ

Введение. Интерпретация результатов количественных лабораторных исследований имеет ряд особенностей и ограничений. Для устранения этих ограничений обсуждается использование систем поддержки принятия врачебных решений. А использование в этих системах технологий искусственного интеллекта, в частности, NLP-технологий для автоматического извлечения симптомов и другой важной информации из электронных медицинских карт (ЭМК) с последующей их интерпретацией моделями машинного обучения позволяет оценивать вероятность у пациента того или иного заболевания.

Цель исследования. Изучение подходов к формированию наборов данных с использованием лабораторных показателей и связанных с ними заболеваний на примере разработки модели машинного обучения на данных лабораторного обследования, возраста и пола.

Материал и методы. Использована база данных ЭМК платформы Webiomed. Был сформирован набор данных пациентов, содержащий входную информацию о демографических (пол, возраст) и лабораторных данных, даты проведения анализов. Выходные данные были представлены сведениями о заключительном клиническом диагнозе, типе обращения (амбулаторный или стационарный случай) и исходе лечения. Для создания модели выявления подозрений на заболевания были применены следующие алгоритмы классификации: LogisticRegression, GaussianNB, DecisionTree, RandomForest, xgboost, AdaBoost, LGBM, MLP. В качестве метрики качества работы модели была выбрана точность (Accuracy). Данные исходного набора для обучения были обработаны различными способами с целью нормализации. Итоговое число записей набора данных для обучения составило 201 613.

Результаты. Ансамблевые алгоритмы, деревья решений и искусственные нейронные сети показали самые высокие результаты классификации: LGBM — 58%, xgboost — 59%, DecisionTree — 59%, MLP (многослойный перцептрон, число скрытых слоев — 3,147 нейронов в каждом) — 61%, RandomForest — 69%. В целях избегания переобучения моделей использовались методы кросс-валидации и регуляризации.

Выводы. Проведенное исследование показало, что использование наборов данных на извлеченных из ЭМК признаков и машинного обучения позволяет создавать модели для выявления подозрений на заболевания, причем поэтапная работа по анализу и подготовке наборов данных, а также применение различных алгоритмов машинного обучения и их настройка позволяют последовательно увеличивать точность работы моделей.

Ключевые слова: искусственный интеллект, машинное обучение, большие данные, системы поддержки принятия решений, электронные медицинские карты, лабораторные данные.

ИНФОРМАЦИЯ ОБ АВТОРАХ:

Гусев А.В. — <https://orcid.org/0000-0002-7380-8460>

Гавриленко Г.Г. — <https://orcid.org/0000-0001-9024-2756>

Гаврилов Д.В. <https://orcid.org/0000-0002-8745-857X>

Автор, ответственный за переписку: Гусев Александр Владимирович — e-mail: agusev@webiomed.ai

КАК ЦИТИРОВАТЬ:

Гусев А.В., Гавриленко Г.Г., Гаврилов Д.В. Разработка модели машинного обучения для интерпретации результатов лабораторной диагностики с целью выявления подозрений на заболевания. *Лабораторная служба*. 2022;11(2):9–17. <https://doi.org/10.17116/labs2022110219>

Development of a machine learning model to interpret the results of laboratory diagnostics in order to identify suspected diseases

© A.V. GUSEV^{1, 2, 3}, G.G. GAVRILENKO¹, D.V. GAVRILOV¹

¹K-Sky LLC, 17 Varkaus Embankment, Petrozavodsk, Republic of Karelia, Russia;

²FSBI «TSNIOIZ» of the Ministry of Health of the Russian Federation, Moscow, Russia;

³GBUZ «NPCC DITT DZ of Moscow», Moscow, Russia

ABSTRACT

Introduction. The interpretation of the results of quantitative laboratory studies has a number of features and limitations. To eliminate these limitations, the use of medical decision support systems is discussed. And the use of artificial intelligence technologies, in particular NLP technologies for automatically extracting symptoms and other important information from electronic medical records, followed by their interpretation by machine learning models designed to assess the likelihood of a patient having a particular disease.

Purpose of the study. Studying approaches to the formation of data sets using laboratory parameters and related diseases by the example of developing a machine learning model based on laboratory examination data, age and gender.

Material and methods. The database of electronic health records (EHR) of the Webiomed platform was used. A data set was formed containing input information about patients who underwent laboratory diagnostics, including demographic data (gender, age), laboratory data, date of analysis. The output was presented with information about the final clinical diagnosis, type of treatment (out-patient or inpatient) and treatment outcome. To create a model for identifying suspicions of diseases, the following classification algorithms were used: LogisticRegression, GaussianNB, DecisionTree, RandomForest, xgboost, AdaBoost, LGBM, MLP. Accuracy was chosen as the performance metric of the model. The original training set data was processed in various ways for the purpose of normalization. The total number of records in the training dataset was 201.613.

Results. Ensemble algorithms, decision trees and artificial neural networks showed the highest classification results: LGBM — 58%, xgboost — 59%, DecisionTree — 59%, MLP (multilayer perceptron, the number of hidden layers — 3.147 neurons in each) — 61%, Random Forest — 69%. In order to avoid retraining of models, cross-validation and regularization methods were used.

Conclusions. The study showed that the use of data sets based on features extracted from EHR and machine learning allows you to create models to identify suspected diseases, and step-by-step work on the analysis and preparation of data sets, as well as the use of various machine learning algorithms and their tuning, can consistently increase the accuracy of the models.

Keywords: artificial intelligence, machine learning, big data, decision support systems, electronic health records, laboratory data.

INFORMATION ABOUT THE AUTHORS:

Gusev A.V. — <https://orcid.org/0000-0002-7380-8460>

Gavrilenko G.G. — <https://orcid.org/0000-0001-9024-2756>

Gavrilov D.V. — <https://orcid.org/0000-0002-8745-857X>

Corresponding author: Gusev A.V. — e-mail: agusev@webiomed.ai

TO CITE THIS ARTICLE:

Gusev AV, Gavrilenko GG, Gavrilov DV. Development of a machine learning model to interpret the results of laboratory diagnostics in order to identify suspected diseases. *Laboratory Service = Laboratornaya sluzhba*. 2022;11(2):9–17. (In Russ.). <https://doi.org/10.17116/labs2022110219>

Введение

Интерпретация результатов количественных лабораторных исследований традиционно заключается в сравнении полученного результата с 1) референтным интервалом (РИ) — популяционной «нормой», отражающей разброс межиндивидуальной биологической вариации; 2) с использованием порогов принятия клинического решения (clinical decision limits) из клинических рекомендаций; 3) с предыдущими результатами пациента с использованием критической разницы значений RCV (Reference Change Value). Однако каждый из этих подходов сравнения имеет свои ограничения и недостатки. Низкий индекс индивидуальности, особенно меньше 0,6, что выражается в значительном преобладании внутри-

индивидуальной вариации над межиндивидуальной биологической вариацией, накладывает ограничение на использование РИ для трактовки результата исследования [1]. В таких случаях клинически значимые изменения уровня аналита у пациента могут не выходить за пределы РИ и оставаться незамеченными. Для большого количества аналитов отсутствует стандартизация методики определения и полная цепь метрологической прослеживаемости, что приводит к несопоставимости результатов, получаемых различными аналитическими системами. К сожалению, как правило, пороговые значения устанавливаются в исследовании с использованием какой-то одной аналитической системы, без учета отсутствия взаимозаменяемости аналитических систем и влияния данного факта на интерпретацию результата.

Многие маркеры обладают недостаточной специфичностью и чувствительностью для диагностики, в связи с чем для повышения точности результатов включается в анализ дополнительная информация о пациенте. В частности, ранее для комплексной оценки данных использовались различные номограммы и калькуляторы. Однако в реальной врачебной практике в условиях дефицита времени и необходимости интерпретации широкого спектра исследований применение разрозненных номограмм и калькуляторов затруднено [2].

Для устранения этих ограничений обсуждается использование систем поддержки принятия врачебных решений (СППВР). Как правило, при создании СППВР алгоритмы выявления подозрений на заболевания разрабатываются на основе подхода, называемого «экспертными системами». Такая технология подразумевает создание специальной «базы знаний», в которой в формализованном виде хранятся справочники, описывающие взаимосвязи между различными признаками и сопоставленными с ними заболеваниями, которые принято называть «решающими правилами» [2]. Этот подход позволяет получить высокую точность по выявлению подозрений на заболевания, прозрачность созданных алгоритмов и, как следствие, приемлемое с позиции практического использования доверие к таким системам.

Альтернативным подходом к созданию алгоритмов выявления подозрений на заболевания является использование технологий искусственного интеллекта (ИИ), в частности, NLP-технологий (Natural Language Processing) для автоматического извлечения признаков болезни и другой важной информации из электронных медицинских карт (ЭМК), с последующей их интерпретацией моделями машинного обучения (МО), предназначенными для оценки вероятности наличия у пациента того или иного заболевания [3]. ЭМК генерируют в себе большой объем информации о пациенте: демографические, лабораторные, инструментальные данные, диагнозы и назначаемые лекарства, что приводит к накоплению миллиардов различных признаков у миллионов человек. Несмотря на то что изначально данные собираются для персонального применения у пациента, существует возможность их обезличить и использовать для улучшения оказания помощи пациенту посредством использования передовых вычислительных технологий [4]. Эти технологии, в том числе связанные с МО, реализованы СППВР.

Развитие информатизации здравоохранения в России привело к тому, что в большинстве медицинских организаций внедрены различные медицинские информационные системы, в том числе и лабораторные информационные системы (ЛИС), обеспечивающие автоматизацию и информационную поддержку процессов работы клинической лаборатории, а также об-

мен данными с лабораторным оборудованием. Интеграция ЛИС с МИС в части автоматизированной передачи результатов лабораторного обследования в ЭМК является рутинной практикой. Таким образом, использование наборов данных из ЭМК для задач МО позволяет автоматически создавать алгоритмы выявления подозрений на заболевания.

Практическое применение методов МО в медицине на данный момент носит преимущественно узкоспециализированный характер, например, для анализа медицинских изображений и сигналов, полученных в результате инструментальных исследований. Однако возможности МО имеют широкие перспективы применения в области диагностических задач, таких как детекция подозрений на заболевание при помощи анализа биомаркеров.

Ф. Meherwar и соавт. на примере пяти заболеваний показали преимущества применения методов МО для диагностики и перспективы этого подхода. Отмечено, что результаты лабораторных исследований являются наиболее информативными и точными источниками медицинских данных [5]. В исследовании Gresco A. и др. на данных тестирования крови разработаны модели МО для ранней диагностики заболеваний двигательных нейронов [6]. Kumar и соавт. [7] разработали две модели, выявляющие при помощи параметров анализа крови гематологические заболевания и продемонстрировали возможность их клинического применения. Был сделан вывод, что данные анализа крови являются исчерпывающей информацией для диагностики исследуемых заболеваний. В исследовании Park и др. [8] продемонстрированы возможности моделей МО для решения задачи классификации заболеваний на основе лабораторных тестов. Также имеются публикации [9, 10], в которых описывается анализ данных лабораторных исследований с помощью моделей МО для выявления подозрений на коронавирусную инфекцию.

Разработка модели выявления подозрения на заболевания с использованием лабораторных данных при помощи методов МО — это трудоемкая и сложная задача, предполагающая предварительное решение ряда задач. Это исследование стало первым шагом на этом пути. Цель работы заключалась в изучении подходов к формированию наборов данных с использованием лабораторных показателей и связанных с ними заболеваний на примере разработки модели МО на данных лабораторного обследования, возраста и пола.

Материал и методы

СППВР Webiomed.DHRA является частью платформы прогнозной аналитики Webiomed (регистрационное удостоверение Росздравнадзора 2020/9958). Стратегией развития Webiomed.DHRA является

совершенствование возможностей по определению подозрений на заболевания.

Источником информации для создания модели стала база данных (БД) платформы прогнозной аналитики Webiomed, содержащая деперсонифицированные формализованные данные ЭМК 2,9 миллионов пациентов, проходивших обследование и лечение в медицинских организациях различных регионов Российской Федерации. Для сбора данных компанией-разработчиком платформы Webiomed были подписаны соглашения с соответствующими операторами персональных медицинских данных на их обезличивание и передачу результатов такой обработки данных для анализа в платформу Webiomed, в том числе для научно-исследовательских целей. Поскольку анализировались обезличенные медицинские данные, информированное добровольное согласие не оформлялось. Медицинские данные и выявляемые признаки получены рефрактивно из ЭМК с помощью различных технологий извлечения информации из «сырых» данных, в том числе NLP. БД содержит более 2600 категориальных и количественных признаков. Общий вид и детализация платформы Webiomed представлены на рис. 1 на цв. вклейке, рис. 2.

Критериями отбора информации о пациентах для последующего анализа стали возраст ≥ 18 лет и нали-

чие хотя бы одного протокола биохимического или клинического анализа крови.

Характеристика базы данных Webiomed представлена в табл. 1.

Дополнительные характеристики базы данных Webiomed в разрезе показателей лабораторной диагностики представлены в табл. 2 и 3.

Формирование набора данных

Для создания модели МО из базы данных Webiomed был сформирован набор данных, содержащий входную информацию о пациентах, прошедших лабораторную диагностику, включая демографические данные (пол, возраст), лабораторные данные, дату проведения анализов. Выходные данные были представлены сведениями о заключительном клиническом диагнозе, типе обращения (амбулаторный или стационарный случай) и исходе лечения.

Изначально для создания модели выявления подозрений на заболевания было принято решение о рассмотрении всех видов заключительного основного диагноза. Сопутствующие заболевания и осложнения основного диагноза в ходе исследования было решено не рассматривать. Параметр «Пол пациента» был преобразован из факторного формата (Мужской, Женский) в числовой — 0, 1 соответственно.

The screenshot displays the Webiomed database interface. At the top, there is a table titled 'Лист окончательных диагнозов' (List of final diagnoses) with columns: Код МКБ-10 (ICD-10 code), Формулировка (Formulation), Дата регистрации (Registration date), Дата обращения (Referral date), and Тип (Type). The table lists three diagnoses: 004.9 (Medication-induced abortion at an early stage of pregnancy, 3 weeks on the uterus), N80.8 (Endometriosis), and I80.0 (Varicose expansion of veins of the lower extremities without ulcers). Below the table is a button 'Показать ещё' (Show more).

Below the table, there is a section 'Заболевания пациента' (Patient's diseases) with the text 'Заболевания не указаны' (Diseases are not specified).

Below that, there is a section 'Извлечённые признаки пациента' (Extracted patient features) with a table showing various features and their sources. The features include: Вес (Weight), Гастрит (Гастрит (источник: анкета диспансеризации) (Gastritis (source: questionnaire)), Глюкоза крови (любое измерение) (Blood glucose (any measurement)), Диастолическое АД (Diastolic blood pressure), Нарушение сознания (Consciousness disturbance), Общий холестерин (Total cholesterol), Окружность талии (Waist circumference), Отягощенная наследственность по инфаркту миокарда в возрасте до 60 лет у ближайших родственников (Heavy hereditary predisposition to myocardial infarction at the age of 60 years in the closest relatives), Пребывания в домах престарелых (Residence in nursing homes), Рациональное питание (источник: анкета диспансеризации) (Rational nutrition (source: questionnaire)), Рост (Height), Систолическое АД (Systolic blood pressure), Табакокурение (Smoking), Уровень употребления алкоголя (источник: анкета диспансеризации) (Alcohol consumption level (source: questionnaire)), and Язвенная болезнь (источник: анкета диспансеризации) (Ulcer disease (source: questionnaire)). To the right of the table, there are buttons 'Показать значения' (Show values) for each feature. At the bottom of the table is a button 'Свернуть' (Collapse).

Рис. 2. Обезличенные социальные, категориальные и количественные признаки пациента в БД Webiomed.

Fig. 2. Impersonal social, categorical and quantitative characteristics of the patient in the Webiomed database.

Таблица 1. Характеристика базы данных обезличенных ЭМК, накопленная в платформе Webiomed
Table 1. Characteristics of the database of depersonalized EHRs accumulated in the Webiomed platform

Показатель	Значение
Количество медицинских организаций, поставивших данные в платформу	86
Количество уникальных пациентов, зарегистрированных в платформе	2 915 994
Распределение пациентов по полу, <i>n</i> (%)	Мужчин — 1 374 410 (47,13), Женщин — 1 541 584 (52,87)
Распределение пациентов по возрасту, <i>n</i> (%)	0—17 лет — 452 458 (15,52) 18—29 лет — 324 357 (11,12) 30—39 лет — 497 383 (17,06) 40—49 лет — 410 637 (14,08) 50—59 лет — 356 293 (12,22) 60—69 лет — 409 296 (14,04) 70 и старше лет — 465 570 (15,97)
Общее число пациентов с любым протоколом лабораторной диагностики в ЭМК (%)	1 104 002 (37,86)
Доля таких пациентов среди взрослого (от 18 лет) населения, %	30,72
Число пациентов с двумя и более протоколами лабораторной диагностики в ЭМК	1 026 795 (35,21%)
Доля таких пациентов среди взрослого (от 18 лет) населения, %	28,53
Число загруженных в платформу документов «Протокол лабораторной диагностики» (% от общего числа загруженных документов)	27 186 420 (24,71)
Число типов признаков, которые были извлечены из протоколов лабораторной диагностики	213
Общее число признаков, извлеченных в базе данных Webiomed из протоколов лабораторной диагностики	85 292 379

Таблица 2. Распределение числа протоколов лабораторной диагностики по видам исследований
Table 2. Distribution of laboratory diagnostic protocols by types of studies

Вид лабораторного исследования	Количество	Доля
Гематологические исследования	13 275 900	48,83
Биохимические исследования	5 571 903	20,50
Химико-микроскопические исследования	3 693 391	13,59
Коагулологические исследования	352 746	1,30
Иммунологические исследования	63 828	0,23
Аутоиммунная диагностика	63 414	0,23
Терапевтический лекарственный мониторинг	4026	0,01
Молекулярная диагностика инфекционных заболеваний	2002	0,01
Вид исследования не был распознан Webiomed	4 159 210	15,30
Итого	27 186 420	100,00

Таблица 3. Распределение числа извлеченных из протоколов лабораторной диагностики признаков по видам
Table 3. Distribution of the quantity of features extracted from the protocols of laboratory diagnostics by types

Вид лабораторного исследования	Количество	Доля
Гематологические исследования	47 077 784	55,20
Биохимические исследования	28 669 632	33,61
Химико-микроскопические исследования	7 961 748	9,33
Коагулологические исследования	1 533 465	1,80
Аутоиммунная диагностика	48 487	0,06
Иммунологические исследования	1143	0,00
Терапевтический лекарственный мониторинг	120	0,00
Итого	85 292 379	100,00

Во время исследования было принято решение о балансировке классов в наборе данных по причине их дисбаланса. Для этого был применен ресэмплинг, суть которого заключается либо в удалении элементов из слишком большого набора (андерсэмплинг)

и/или добавлении большего количества элементов в недостаточно большой набор (оверсэмплинг) [11] (рис. 3). В нашем исследовании был применен метод оверсэмплинга, т.е. число записей в каждом классе было увеличено до максимального из всех классов.

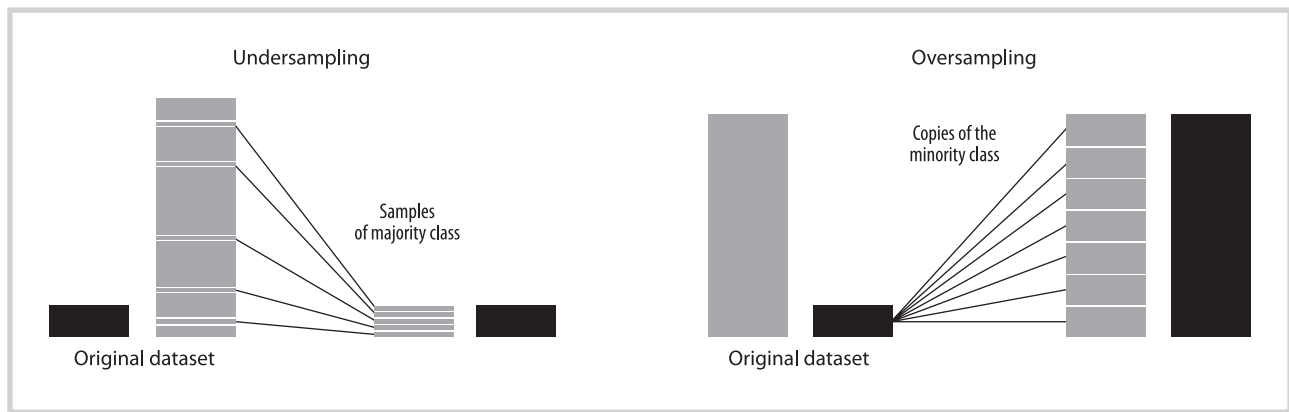


Рис. 3. Балансировка классов методами андерсэмплинга (undersampling) и оверсэмплинга (oversampling).

Fig. 3. Class balancing with the use of undersampling and oversampling methods.

В полученном наборе данных ряд результатов лабораторной диагностики содержал два, три и более значений ввиду повторного выполнения лабораторного обследования во время случая лечения (табл. 4).

По итогам обсуждения этой особенности с медицинскими советниками было решено выбрать самое первое значение параметра, т.е. самое раннее для каждой записи, так как именно оно характеризовало начальные проявления заболевания, имеющегося у пациента. Строки набора данных, в которых содержатся несколько значений признака, были преобразованы при помощи Python-функции *pivot_table* в отдельные столбцы, хранящие свои значения для каждой записи, так как эти значения являются основными входными признаками модели. Фрагмент этого преобразования показан в табл. 5.

Стоит отметить, что после преобразования набора в нем появились пропущенные значения (NaN — Not a Number), так как ряд лабораторных данных, имеющих у одного пациента, может отсутство-

вать у другого. В дальнейшем пропущенные значения были обнулены.

Заключительным шагом формирования набора данных стала агрегация обработанных записей в единый файл, который затем был использован для создания модели.

Отбор информативных признаков

Среди демографических (пол, возраст) и лабораторных данных пациентов был проведен отбор информативных признаков по методу χ^2 . Были выделены наименее значимые признаки, исключенные впоследствии из обучающего набора.

Результаты

В итоговом наборе данных значения демографических и лабораторных данных пациентов были использованы в качестве входных параметров алгоритмов МО. Итоговый набор данных был случайным

Таблица 4. Пример записей лабораторных данных с несколькими значениями в одной ячейке

Table 4. Example of laboratory data records with multiple values in one cell

ID пациента	Дата получения признака	Название признака	Код признака	Значение признака
532418	2020-1-20 00:00:00+03:03	Гемоглобин крови	260	[100,00; 0]
613056	2018-12-1 00:00:00+03:03	Гемоглобин крови	260	[118,00; 359,00]
597316	2019-12-17 00:00:00+03:03	Уровень глюкозы (любое измерение)	14	[13,00; 15,10; 14,30]
640791	2019-4-11 00:00:00+03:03	Уровень глюкозы (любое измерение)	14	[10,40; 5,70; 7,70]
522228	2019-6-12 00:00:00+03:03	Уровень глюкозы (любое измерение)	14	[6,19; 9,40; 13,10]

Таблица 5. Фрагмент преобразованной таблицы с извлеченными лабораторными данными

Table 5. Fragment of the transformed table with extracted laboratory data

ID пациента	Дата получения признака	Альфа-амилаза общая крови	Аланинамино- трансфераза	Анизоцитоз	Активность альфа-2 антиплазмина
500009	2016-2-10 00:00:00+03:03	NaN	20,30	NaN	NaN
500009	2017-4-25 00:00:00+03:03	NaN	20,90	13,60	NaN
500009	2017-5-31 00:00:00+03:03	NaN	NaN	14,00	NaN

выявления подозрений на заболевания, причем поэтапная работа по анализу и подготовке наборов данных, а также применение различных алгоритмов МО и их настройка позволяют последовательно увеличивать точность работы моделей.

Вместе с этим полученные метрики модели были признаны как недостаточные для ее включения в СППВР Webiomed.DHRA, так как их работа в условиях реальной клинической практики вероятнее всего привела бы к дальнейшей деградации точности, что является неприемлемо высоким риском дополнительных ошибок.

По итогам данного исследования команда разработчиков сформулировала несколько гипотез, нуждающихся в дальнейшей проверке:

— Интерпретация только протоколов лабораторного обследования без учета дополнительной информации из ЭМК, включая данные об уже зарегистрированных ранее заболеваниях, объективные данные и другие симптомы пациентов не является точной для постановки диагноза. Целесообразно повторить исследования на более полных данных ЭМК, характеризующих пациента и его заболевания.

— Требуется дополнительная исследовательская работа по изучению различных методов работы с пропущенными значениями, что может позволить дополнительно увеличить точность работы модели.

— Целесообразно разрабатывать отдельные модели для каждого заболевания или группы заболеваний, объединенных по общим патогенетическим критериям. Вероятно, что путь разработки одной универсальной модели, которая могла бы формировать подозрения на любые заболевания, является тупиковым.

— Целесообразно проведение внешней валидации моделей с целью формирования необходимого доверия и последующего принятия решения о возможности их встраивания в СППВР.

Заключение

Результатом проведенного исследования стала модель выявления подозрений на заболевания на основе результатов лабораторной диагностики с использованием данных ЭМК. Точность модели не позволяет применять ее в практике. В ходе обсуждения возможных причин невысокой точности был сформулирован ряд гипотез, требующих дополнительных исследований и разработок. Анализ данных литературы относительно аналогичного опыта разработки диагностических моделей позволяет утверждать, что данные лабораторных анализов являются значимыми для многих заболеваний. Однако применение алгоритмов МО для детекции подозрения на заболевание на данном этапе является экспериментальной областью, в которой обмен данными между медициной и наукой направлен в обе стороны. Актуальной задачей остается поиск биомаркеров и осуществление корреляционных связей между ними, на основе которых может проводиться достоверная и точная оценка вероятности наличия у пациента заболевания. В целом, наше исследование подтверждает необходимость развития этого направления, так как его результаты могут быть использованы в практической медицине.

**Авторы заявляют об отсутствии конфликта интересов.
The authors declare no conflicts of interest.**

ЛИТЕРАТУРА/REFERENCES

1. Фрейзер К.Г. *Биологическая вариация: от теории к практике*. М.: Медиздат; 2010.
Fraser CG. *Biological Variations From principles to Practice*. М.: ААСС; 2001. (In Russ.).
2. Евгина С.А., Савельев Л.И. Современные теория и практика референтных интервалов. *Лабораторная служба*. 2019;8(2):36-44.
Evagina SA, Savelyev LI. Current theory and practice of reference intervals. *Laboratory Service*. 2019;8(2):36-44. (In Russ.).
<https://doi.org/10.17116/labs2019802136>
3. Гаврилов Д.В., Кирилкина А.В., Серова Л.М. Алгоритм формирования подозрения на новую коронавирусную инфекцию на основе анализа симптомов для использования в системах поддержки принятия врачебных решений. *Врач и информационные технологии*. 2020;4:51-58.
4. Гусев А.В., Новицкий Р.Э., Ившин А.А., Алексеев А.А. Машинное обучение на лабораторных данных для прогнозирования заболеваний. *ФАРМАКОЭКОНОМИКА. Современная фармакоэкономика и фармакоэпидемиология*. 2021;14(4):581-592.
<https://doi.org/10.17749/2070-4909/farmakoekonomika.2021.115>
5. Tang A, Oskotsky T, Sirota M. Personalizing routine lab tests with machine learning. *Nat Med*. 2021;27:1514-1515.
<https://doi.org/10.1038/s41591-021-01486-4>
6. Fatima Meherwar, Maruf Pasha. «Survey of Machine Learning Algorithms for Disease Diagnostic». *Journal of Intelligent Learning Systems and Applications*. 2017;9:1:1-16.
<https://doi.org/10.4236/jilsa.2017.91001>
7. Greco Alberto, Maria Rosa Chiesa, Ilaria Da Prato, Anna Maria Romanelli, Cristina Dolciotti, Gabriella Cavallini, Silvia Maria Masciandaro, Enzo Pasquale Scilingo, Renata Del Carratore, Paolo Bongioanni. «Using Blood Data for the Differential Diagnosis and Prognosis of Motor Neuron Diseases: A New Dataset for Machine Learning Applications». *Scientific Reports*. 2021;11:1:3371.
<https://doi.org/10.1038/s41598-021-82940-8>
8. Kumar Naresh, Nripendra Narayan Das, Deepali Gupta, Kamali Gupta, Jatin Bindra. «Efficient Automated Disease Diagnosis Using Machine Learning Models». *Journal of Healthcare Engineering*. 2021.
<https://doi.org/10.1155/2021/9983652>
9. Park Dong Jin, Min Woo Park, Homin Lee, Young-Jin Kim, Yeongsic Kim, Young Hoon Park. «Development of Machine Learning

- Model for Diagnostic Disease Prediction Based on Laboratory Tests». *Scientific Reports*. 2021;11:1.
<https://doi.org/10.1038/s41598-021-87171-5>
10. Wu Jiangpeng, Pengyi Zhang, Liting Zhang, Wenbo Meng, Junfeng Li, Chongxiang Tong, Yonghong Li, et al. «Rapid and Accurate Identification of COVID-19 Infection through Machine Learning Based on Clinical Available Blood Test Results». *medRxiv*. 2020.
<https://doi.org/10.1101/2020.04.02.20051136>.
 11. Yao Haochen, Nan Zhang, Ruochi Zhang, Meiyu Duan, Tianqi Xie, Jiahui Pan, Ejun Peng, et al. «Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests». *Frontiers in Cell and Developmental Biology*. 2020;8.
<https://www.frontiersin.org/article/10.3389/fcell.2020.00683>.
 12. Estabrooks Andrew, Taeho Jo, Nathalie Japkowicz. «A Multiple Resampling Method for Learning from Imbalanced Data Sets». *Computational Intelligence*. 2004;20:1:18-36.
 13. Schaffer Cullen. Selecting a Classification Method by Cross-Validation. *Machine Learning*. 1993;13:1:135-143.
<https://doi.org/10.1007/BF00993106>.
 14. Tian Yingjie, Yuqi Zhang. A Comprehensive Survey on Regularization Strategies in Machine Learning. *Information Fusion*. 2022;80:146-166.
<https://doi.org/10.1016/j.inffus.2021.11.005>.
 15. Belete Daniel Mesafint, Manjaiah D. Huchaiah. «Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results». *International Journal of Computers and Applications*. 2021;0:0:1-12.
<https://doi.org/10.1080/1206212X.2021.1974663>.
 16. Bergstra James, Yoshua Bengio. «Random Search for Hyper-Parameter Optimization». *Journal of Machine Learning Research*. 2012;13:281-305.

Поступила 16.03.2022

Received 16.03.2022

Принята к печати 29.04.2022

Accepted 29.04.2022