Универсальная Лабораторная Модель: прогнозирование патологий по результатам анализов крови

Карпов Павел Владимирович, к.х.н.[*]
МГУ имени М. В. Ломоносова
г. Москва, Россия
carpovpv@qsar.chem.msu.ru

Райман Руслан Адамович[†]
ООО «РУЛИС», ЛИС RosLIS
г. Москва, Россия
ruslan.raiman@roslis.ru

Аннотация

В документе изложены методологические и практические разработки применения Универсальной Лабораторной Модели в работе диагностической потоковой лаборатории. Описаны алгоритмы построения моделей, их статистические характеристики, а также параметры исходных данных. Представлена актуальная библиография по современным подходам применения методов машинного обучения в лабораторной диагностике.

1 Введение

Методы машинного обучения (искусственного интеллекта — ИИ) активно применяются в медицине для автоматического выявления патологий путём анализа изображений и текстов [1]. Тем не менее, в лабораторной диагностике эти методы еще не используются в виду объективной сложности моделирования табличных данных[2], и, что более важно, из-за отсутствия методологии такого применения. Мы предлагаем использовать методы искусственных нейронных сетей для **прогнозирования вероятности патологических значений неназначенных врачом лабораторных тестов, по результатам выполнения назначенных тестов** и исторических данных пациента.

В качестве примера рассмотрим ситуацию, когда пациенту был назначен липидный профиль, общеклинический анализ крови и несколько биохимических тестов. После выполнения всех этих исследований осуществляется прогноз возможных патологических значений по другим лабораторным тестам (которые заложены в модели), и если этот прогноз положителен, то тесты автоматически назначаются и выполняются на анализаторах (при условии наличия необходимого биологического материала)[3]. В результате **пациент получит результат как по тем тестам, которые ему непосредственно назначил лечащий врач, так и по другим показателям**. Тем самым если назначенные показатели действительно оказались завышенными или заниженными, врач сможет скорректировать лечение или направить пациента на дополнительную консультацию к другому специалисту.

Для реализации такого подхода целесообразно ввести дополнительную услугу «Анализ результатов лабораторных исследований с использованием искусственного интеллекта», на подобие недавно введённой услуги с кодом 001601 «Описание и интерпретация данных маммографического исследования с использованием искусственного интеллекта» [4]. Так как дополнительные исследования будут назначаться не по всем поступающим направлениям, а только по тем, по которым получен положительный прогноз от сервиса ИИ, то у каждой лаборатории появляется широкий спектр вариантов оптимизации алгоритма, включая экономические и медицинские аспекты.

В рамках данного подхода нами развивается специальная архитектура искусственной нейронной сети — универсальная лабораторная модель[4] (УЛМ) [5, 6], которая спроектирована таким образом, чтобы учесть как можно больше имеющейся информации по результатам анализов конкретного пациента для одновременного прогнозирования спектра других тестов, по результатам которых вероятны отклонения от нормы.

---

[*] МГУ, Химический факультет, кафедра медицинской химии, 119991, Москва, Ленинские горы, д. 1, стр. 3, к. 523.

[†] ООО «РУЛИС», 115191, г. Москва, Холодильный пер., д. 3, к. 1, стр. 2, https://roslis.ru

[3] В рамках прогнозирования одного показателя — низкого уровня ферритина — такой подход в автоматическом добавлении незначенного теста был применён и встроен в Лабораторную Информационную Систему ЛИС в больнице Йерун Бош в г. Хертогенбос (Нидерланды) в октябре 2021 [3].

[4] https://ulm.roslis.ru

2 Универсальная лабораторная модель

В данном разделе представлено краткое теоретическое введение в архитектуру УЛМ[5], более подробно описанную в нашей работе [6], представленной на международной конференции ICECCME-2025 и опубликованной IEEE.

Множество результатов лабораторных тестов представляет собой неупорядоченный набор пар тест-значение, который варьируется от пациента к пациенту, или, с точки зрения машинного обучения, представляет собой таблицу с пропущенными значениями. В УЛМ **мы формулируем задачу табличного моделирования как задачу перевода множеств**, где исходный набор содержит пары GPT-эмбеддингов тестов [7] и соответствующих им значений, а целевой набор состоит только из эмбеддингов. Предлагаемый подход может эффективно справляться с пропущенными значениями без их неявной оценки и связывает большие языковые модели с традиционными табличными данными.

При построении модели на основе табличных данных модель искусственной нейронной сети «ничего не знает» о самой сути этих данных. Всё, что известно модели, это то, что есть несколько столбцов и что в этих столбцах есть некоторые значения. Она не имеет никакого представления о том, что именно означает каждый столбец и как эти столбцы потенциально связаны друг с другом — то есть отсутствует контекст, который мог бы помочь модели при обучении. Без этого контекста модели глубокого обучения обычно показывают результаты не лучше классических алгоритмов машинного обучения, таких как деревья решений или машины опорных векторов [8].

Чтобы добавить информацию об этих взаимосвязях между входными и выходными данными и обеспечить контекст, можно использовать механизм внимания. В работе [9] для этой цели использовались обучаемые эмбеддинги признаков. Вместо этого мы используем GPT- подобные эмбеддинги, которые остаются фиксированными в процессе обучения. Кроме того, мы также предлагаем использовать такие же эмбеддинги и для целевых выходов модели.

В нашем подходе используется схема энкодера-декодера, рис. 1. Блок внимания принимает три параметра, называемые традиционно запросами Q, ключами K, и значениями V, соответственно, [10]. Если за X обозначить входные данные модели, а за P — прогнозируемые признаки, то для энкодера $E(X)$, мы используем так называемый механизм самовнимания (Self-Attention), где все три параметра одинаковы и представляют входные данные для модели; для декодера $D(P, X)$ в качестве запросов поступают признаки, которые модель обучается прогнозировать ($Q = P$):

$$E(X) = I = Attention(Q = X, K = X, V = X)$$
$$D(P, I) = Attention(Q = P, K = I, V = I)$$

При формировании результатов лабораторных тестов на бланке отчёта, как правило, нет строго определённого порядка, за исключением традиционной формы выдачи результатов, поэтому предлагаемая модель УЛМ проектировалась инвариантной к перестановкам. Блок Attention (а также Multi-Head Attention) [10] является эквивариантным, что означает, что любая перестановка во входных данных приводит к такой же перестановке на выходе. Для достижения инвариантности необходимо использование фиксированных запросов Q, которые не зависят от порядка, [11]. Таким образом, благодаря своей конструкции **УЛМ может работать со множествами различных размеров, то есть учитывать сколь угодно большой перечень входных атрибутов, описывающих состояние пациента, и предоставлять согласованный прогноз, не зависящий от порядка входных данных**.

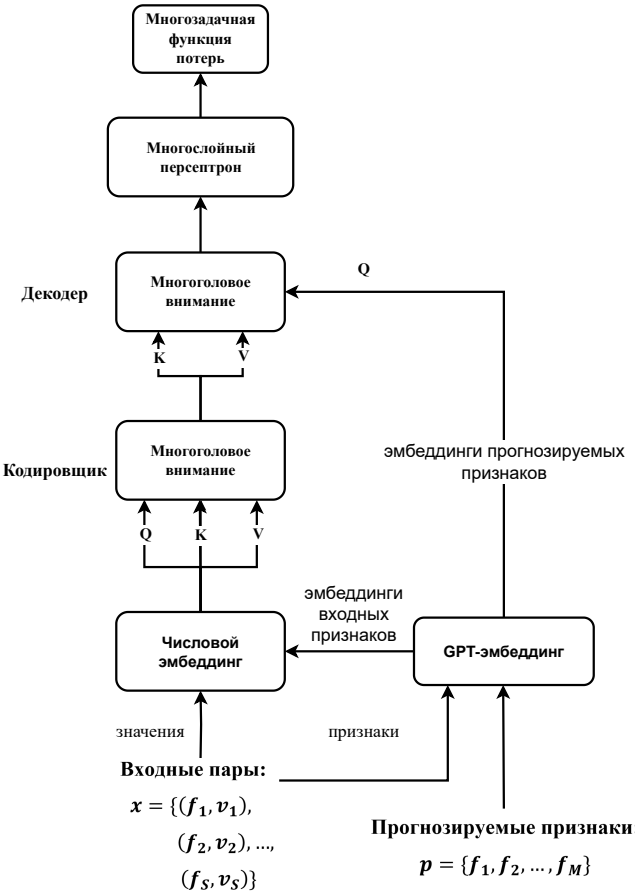

Рис. 1: Архитектура УЛМ: схема Энкодера-Декодера, Декодер использует эмбеддинги прогнозируемых признаков в качестве запросов (Q).

[5]Свидетельство о государственной регистрации программы для ЭВМ № 2024684524 от 17 октября 2024 г.

3 Применение моделей УЛМ

Универсальная лабораторная модель рассчитывает два прогноза по показателю:

1. Классификационный — **прогнозируется, что измеренное значение показателя будет патологическим**. Например, что значение гликированного гемоглобина будет выше 6,5%, что является диагностическим критерием сахарного диабета дополнительно к определению уровня глюкозы крови натощак и тесту на толерантность к глюкозе [12]. При этом какое именно будет значение не вычисляется. То есть оно может быть и 7,0%, и 10,0%, или какое-либо другое.
2. Регрессионный — прогнозируется непосредственно числовое значение этого показателя. Задача восстановления регрессии сложнее классификации и сопряжена с большими ошибками определения. В УЛМ регрессионный выход используется главным образом для подкрепления принятия решения модели в процессе обучения.

С практической точки зрения классификационный подход наиболее перспективен[6], так как в случае выполнения теста, пациент получит реальный, а не спрогнозированный результат. При оценке моделей использовались стандартные метрики для решения классификационной задачи, табл. 3.

Модели машинного обучения по прогнозированию лабораторных показателей имеют ограниченное применение, так как не позволяют полностью отказаться от выполнения подтверждающих лабораторных тестов [13]. Несмотря на то, что модели со значениями $AUC \geqslant 0,80$ допустимы в медицинском применении [4, с. 49], стандартные пороговые значения, определяемые для максимизации чувствительности и специфичности одновременно, приводят к ощутимым ошибкам первого и второго родов. В связи с этим мы предлагаем выбирать пороговые значения для максимизации специфичности, как описано в частности в [14] путём калибровки при внедрении в конкретную лабораторию.

Если в результате прогноза тест требуется выполнить, то доля таких дополнительных постановок определяется по формуле для W, при этом доля определяемой патологии составит P, а доля ошибочно пропущенных — E:

$$W = \frac{TP + FP}{TP + FP + FN + TN}$$
$$P = W * PPV = W * \frac{TP}{TP + FP} = \frac{TP}{TP + FP + FN + TN}$$
$$E = \frac{FN}{TP + FP + FN + TN}$$

Статистические параметры моделей приведены в табл. 1, данные рассчитаны на внешней выборке, не участвовавшей в процессе обучения. Все классификационные модели характеризуются высокими параметрами — $AUC \geqslant 0,80 (\geqslant 80,0\%)$.

Таблица 1: Статистические параметры модели УЛМ

Показатель	AUC, %	TP	TN	FP	FN	$\sum$	Чувствительность	Специфичность	W	P	E
↑ Гликированный гемоглобин, HbA1c	90,1	1839	4242	1030	474	7585	79,5	80,5	0,378	0,242	0,062
↑ Глюкоза	82,4	5070	96 243	32 842	1 872	136 027	73,0	74,6	0,278	0,037	0,013
↑ Липопротеины низкой плотности	87,1	11 277	9 391	2 715	3 195	26 578	77,9	77,6	0,526	0,424	0,120
↑ Мочевая кислота	80,2	1 994	11 883	4 714	847	19 438	70,2	71,6	0,345	0,103	0,044
↓ Ферритин	89,7	1 838	15 493	3 444	497	21 272	78,7	81,8	0,248	0,086	0,023
↑ Холестерин	83,4	108 851	91 462	30 422	36 612	267 347	74,8	75,0	0,521	0,407	0,137

[6]При построении и валидации моделей мы также активно использовали статистику по регрессионному выходу главным образом для определения, что в модели нет скрытых спутывающих факторов (кофаундеров), который приводят к хорошим статистическим показателям моделей. Однако такие модели абсолютно непригодны на практике, см. Спутывающие факторы в регрессионной модели по ЛПНП на с. 16.

В среднем по перечисленным в табл. 1 показателям загрузка лаборатории возрастёт на 38% (W), выявляемость патологий увеличится на 22% (P).

Гликированный гемоглобин (гликозилированный гемоглобин) отражает состояние углеводного обмена за последние 6–8 недель и в норме составляет 4–5,5% [15, с. 143]. В соответствии с рекомендациями Всемирной организации здравоохранения, этот тест признан необходимым для контроля и диагностики ($\geqslant 6,5$ % [16, 12]) сахарного диабета. Рост доли гликированного гемоглобина на 1% связывают с увеличением концентрации глюкозы на 2 ммоль/л, а результат зависит от факторов, влияющих на срок жизни эритроцитов в крови [16, с. 76].

Прогнозированию уровня HbA1c с использованием методов машинного обучения уделяется большое внимание в литературе [17, 18, 19], при этом авторы используют разные пороговые значения. В данной работе для отметки патологии использовалось условие $\geqslant 6,0\%$. Классификационная модель характеризуется $AUC = 0,90$; ROC-кривая приведена на рис. 36.

Глюкоза один из важнейших компонентов крови, количество которого отражает состояние углеводного обмена. Превышение концентрации глюкозы в крови натощак выше 7,0 ммоль/л (такой критерий использовался также в данной работе) служит веским аргументом в постановке диагноза сахарного диабета [15, с. 138]. В научной литературе уделяется большое внимание построению моделей машинного обучения для прогнозирования уровня глюкозы особенно у больных диабетом, например, [20, 21]. Классификационная модель УЛМ характеризуется $AUC = 0,82$; ROC-кривая приведена на рис. 37.

Липопротеины низкой плотности (ЛПНП) широко используются как маркер сердечно-сосудистого риска с целью профилактики осложнений атеросклероза. Многие лаборатории используют расчётный метод определения ЛПНП по формулам Фридвальда [22], Мартина[23], Сэмпсона [24], также разрабатываются новые подходы [25]. Пограничные значения риска ИБС и атеросклероза составляют $\geqslant 3,4$ ммоль/л [15, с. 152]. Построению прогностических моделей машинного для определения ЛПНП посвящены в частности работы [26, 27]. В настоящем исследовании использовался критерий $\geqslant 3,4$ ммоль/л для отметки патологического состояния. Классификационная модель характеризуется $AUC = 0,87$; ROC-кривая приведена на рис. 38.

Мочевая кислота имеет диагностическое значение при подагре при повышенных значениях (гиперурикемия) выше 0,48 ммоль/л для мужчин и 0,38 ммоль/л для женщин, соответственно, [15, с. 134]. Вопросам прогнозирования значений мочевой кислоты и определения групп риска посвящены в частности работы [28, 29]. Полученная классификационная модель характеризуется $AUC = 0,80$; ROC-кривая приведена на рис. 39.

Ферритин выявляет железодефицитную анемию (ЖДА) и расчетным методам определения ферритина уделяется большое внимание в литературе [30, 31, 32]. Кроме того, по-видимому, это первый тест, который был принят для автоматического добавления в заказ при условии положительного прогноза по модели машинного обучения [3]. В настоящей работе за патологический уровень используется значение ферритина в сыворотке $\leqslant 12,0$ нг/мл [15, с. 233]. Полученная классификационная модель характеризуется $AUC = 0,89$; ROC-кривая приведена на рис. 40.

Холестерин является важным показателем состояния липидного обмена; при значениях выше 6,5 ммоль/л является фактором риска развития атеросклероза и ишемической болезни сердца [15, с. 148]. Желательный уровень общего холестерина составляет $< 5,18$ ммоль/л [16, с. 103]. В данной работе использовалось пороговое значение 5,2 ммоль/л. В литературе холестерин как отдельный целевой параметр обычно не моделируется, только как часть в моделях по прогнозу ЛПНП, уровней риска атеросклероза и т.д. Полученная классификационная модель характеризуется $AUC = 0,83$; ROC-кривая приведена на рис. 41.

4 Выводы

Моделирование таких многофакторных данных как результаты лабораторных анализов сопряжено с риском появления спутывающих факторов и требует как большой и разнообразной базы данных для построения и валидации моделей, где представлен широкий спектр пациентов с разными проблемами со здоровьем, так и учёта как можно большей информации по результатам исследований. В связи с этим традиционные подходы машинного обучения к построению моделей на табличных данных для этой цели не подходят, так как не могут учитывать всю имеющуюся информацию, и к тому же не в состоянии выявлять скрытые ложные корреляции.

Предлагаемый подход Универсальной Лабораторной Модели решает эти проблемы органично благодаря специальной архитектуре искусственной нейронной сети, которая может принимать вариативное множество результатов

исследований и одновременно прогнозировать несколько показателей, при этом не ограничиваясь только анализами крови. Такой режим прогноза разрушает возможные скрытые ложные корреляции и опосредованно имитирует врача-диагноста: чем больше результатов анализов, тем точнее прогноз. Показанные в работе статистические характеристики моделей на уровне[7] современных исследований в этой области.

Нами разработан программный интерфейс для интеграции моделей УЛМ с Лабораторными и Медицинскими информационными системами. Описание доступно на сайте https://ulm.roslis.ru. В качестве апробации работы прогнозирование части показателей запущено через этот сервис для Лаборатории IX в рамках отдельного модуля ЛИС RosLIS.

Использование УЛМ увеличивает выявляемость заболеваний на 20–30%. Мы считаем, что **использование лабораторной предикативной медицины это будущее лабораторной диагностики, где результаты анализов не будут «лежать пустым грузом» в базах данных, а оперативно заработают на улучшение здоровья граждан**.

Список литературы

[1] Искусственный интеллект в медицине: обзор текущей ситуации и тенденции / Чихачева Я.Г., Мирук А.К., Ломоносова А.В. и Козлова А.А. // Cifra. Медико-биологические науки. — 2024. — № 2 (2). — Режим доступа: https://medbio.cifra.science/archive/2-2-2024-september/10.60797/BMED.2024.2.4.

[2] Revisiting Deep Learning Models for Tabular Data / Gorishniy Yury, Rubachev Ivan, Khrulkov Valentin, and Babenko Artem. — 2023. — 2106.11959.

[3] Automated prediction of low ferritin concentrations using a machine learning algorithm / Kurstjens Steef, de Bel Thomas, van der Horst Armando, Kusters Ron, Krabbe Johannes, and van Balveren Jasmijn // Clinical Chemistry and Laboratory Medicine (CCLM). — 2022. — Vol. 60, no. 12. — P. 1921–1928. — Access mode: https://doi.org/10.1515/cclm-2021-1194.

[4] Арзамасов К. М. Технологии искусственного интеллекта при массовых профилактических и диагностических лучевых исследованиях : дис. докт. наук ; ГБУЗ г. Москвы, “Научно-практический клинический центр диагностики и телемедицинских технологий департамента здравоохранения г. Москвы". — 443079, г. Самара, пр. К. Маркса, 165 Б, 2024. — Режим доступа: https://samsmu.ru/files/referats/2024/arzamasov/dissertation.pdf.

[5] Карпов П. В. Виртуальный скрининг библиотек органических структур на основе одноклассовой классификации : дис. канд. наук ; МГУ. — 119991, Москва, ГСП-1, Ленинские горы, д. 1, стр. 3, Химический факультет, 2011. — Нояб.

[6] Karpov Pavel, Petrenkov Ilya, Raiman Ruslan. Universal Laboratory Model: Prognosis of Abnormal Clinical Outcomes based on Routine Tests // 2025 5th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). — 2025. — P. 1–7.

[7] Режим доступа: https://yandex.cloud/en-ru/services/yandexgpt.

[8] Representation Learning of Lab Values via Masked AutoEncoders / Restrepo David, Wu Chenwei, Jia Yueran, Sun Jaden K., Gallifant Jack, Bielick Catherine G., Jia Yugang, and Celi Leo A. — 2025. — 2501.02648.

[9] TabTransformer: Tabular Data Modeling Using Contextual Embeddings. — 2020. — 2012.06678.

[10] Attention is All you Need / Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser-Lukasz, and Polosukhin Illia // Advances in Neural Information Processing Systems / ed. by Guyon I., Luxburg U. Von, Bengio S. et al. — Curran Associates, Inc. — 2017. — Vol. 30.

[11] Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. — 2019. — 1810.00825.

[12] Бирюкова Е. В. Роль гликированного гемоглобина в диагностике и улучшении прогноза сахарного диабета // Медицинский Совет. — 2017. — № 3. — С. 48–53.

[13] Labrador: Exploring the Limits of Masked Language Modeling for Laboratory Data / Bellamy David R., Kumar Bhawesh, Wang Cindy, and Beam Andrew. — 2024. — 2312.11502.

[14] Fan Jerome, Upadhye Suneel, Worster Andrew. Understanding receiver operating characteristic (ROC) curves // Canadian Journal of Emergency Medicine. — 2006. — Vol. 8, no. 1. — P. 19–20.

[7] В настоящее время отсутствует стандартный набор данных, на котором разные группы исследователей могли бы сравнить качество своих моделей.

[15] Назаренко Г. И., Кишкун А. А. Клиническая оценка результатов лабораторных исследований. — Москва "Медицина 2006. — С. 544. — ISBN: 5-225-04579-0.

[16] Лабораторная диагностика / под ред. Кондрашева Е.А., Островский А. Ю. — Москва. Медиздат, 2018. — С. 720. — ISBN: 978-5-902943-46-4.

[17] Use of Machine Learning and Routine Laboratory Tests for Diabetes Mellitus Screening / Cardozo Glauco, Pintarelli Guilherme Brasil, Andreis Guilherme Rettore, Lopes Annelise Correa Wengerkievicz, and Marques Jefferson Luiz Brum // BioMed Research International. — 2022. — Vol. 2022, no. 1. — P. 8114049. — https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/8114049.

[18] Cichosz Simon Lebech, Bender Clara, Hejlesen Ole. A Comparative Analysis of Machine Learning Models for the Detection of Undiagnosed Diabetes Patients // Diabetology. — 2024. — Vol. 5, no. 1. — P. 1–11. — Access mode: https://www.mdpi.com/2673-4540/5/1/1.

[19] Выявление сахарного диабета при диспансеризации у обследуемых лич с нормальным уровнем глюкозы плазмы натощак с помощью инструментов машинного обучения / Гимадиев Р. Р., Губина Е. В., Кокорин В. А., Долгих Т. И., Щеголев О. Б. и Радченко А. В. // (XXXI) Национальный диабетологический конгресс с международным участием. — 2025. — С. 200. — Режим доступа: https://pureportal.spbu.ru/files/140712137/NDC_TEZIS_Book_v23.05.25.pdf.

[20] Deep learning for blood glucose level prediction: How well do models generalize across different data sets? / Ghimire Sarala, Celik Turgay, Gerdes Martin, and Omlin Christian W. // PLOS ONE. — 2024. — 09. — Vol. 19, no. 9. — P. 1–26. — Access mode: https://doi.org/10.1371/journal.pone.0310801.

[21] Blood Glucose Diabetic Prediction using Machine Learning Algorithm / Devi A., Sasireka P., Kovardhani K., Premalatha G., Sivasakthi T., and Subash R. // 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS). — 2024. — P. 1519–1522.

[22] Friedewald W. T., Levy Robert I., Fredrickson Donald S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. // Clinical chemistry. — 1972. — Vol. 18 6. — P. 499–502. — Access mode: https://api.semanticscholar.org/CorpusID:45990531.

[23] Comparison of a novel method vs the Friedewald equation for estimating low-density lipoprotein cholesterol levels from the standard lipid profile. / Martin Seth Shay, Blaha Michael Joseph, Elshazly Mohamed Badreldin, Toth Peter P., Kwiterovich Peter O. Jr., Blumenthal Roger S., and Jones Steven R // JAMA. — 2013. — Vol. 310 19. — P. 2061–8. — Access mode: https://api.semanticscholar.org/CorpusID:12515149.

[24] A New Equation for Calculation of Low-Density Lipoprotein Cholesterol in Patients With Normolipidemia and/or Hypertriglyceridemia / Sampson Maureen, Ling Clarence, Sun Qian, Harb Roa, Ashmaig Mohmed, Warnick Russell, Sethi Amar, Fleming James K., Otvos James D., Meeusen Jeff W., Delaney Sarah R., Jaffe Allan S., Shamburek Robert, Amar Marcelo, and Remaley Alan T. // JAMA Cardiology. — 2020. — May. — Vol. 5, no. 5. — P. 540–548. — Access mode: https://doi.org/10.1001/jamacardio.2020.0013.

[25] Садовников П. С., Ольховик А. Ю., Гуревич В. С. Расчетный метод определения уровня холестерина липопротеинов низкой плотности на основании современной парадигмы метаболизма липидов // Атеросклероз и Дислипидемии. — 2022. — Окт. — № 3 (48). — С. 21–28. — Режим доступа: https://jad.noatero.ru/index.php/jad/article/view/347.

[26] Öter Ali. Deep learning-based LDL-C level prediction and explainable AI interpretation // Computers in Biology and Medicine. — 2025. — Vol. 188. — P. 109905. — Access mode: https://www.sciencedirect.com/science/article/pii/S0010482525002562.

[27] Прогнозирование концентрации липопротеинова низкой плотности с помощью инструментов машинного обучения / Гимадиев Р. Р., Варакина-Митрай К. А., Щеголев О. Б., Радченко А. В., Артемьева О. А., and Вареха Н. В // Вычислительная биология и искусственный интеллект для персонализированной медицины. — 2024. — Access mode: https://www.endocrincentr.ru/sites/default/files/all/EVENTS_2024/07-09.08_vychislitelnaya_biologiya_i_iskusstvennyj_intellekt_dlya_personalizirovannoj_mediciny/theses/023.pdf.

[28] Ensemble machine learning prediction of hyperuricemia based on a prospective health checkup population / Zhang Yongsheng, Zhang Li, Lv Haoyue, and Zhang Guang // Frontiers in Physiology. — 2024. — Vol. 15. — Access mode: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2024.1357404.

[29] Blood Uric Acid Prediction With Machine Learning: Model Development and Performance Comparison / Sampa Masuda Begum, Hossain Md Nazmul, Hoque Md Rakibul, Islam Rafiqul, Yokota Fumihiko, Nishikitani Mariko, and Ahmed Ashir // JMIR Med Inform. — 2020. — Vol. 8, no. 10. — P. e18331. — Access mode: https://medinform.jmir.org/2020/10/e18331.

[30] Pullakhandam Siddartha, McRoy Susan. Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning // BioMedInformatics. — 2024. — Vol. 4, no. 1. — P. 661–672. — Access mode: https://www.mdpi.com/2673-7426/4/1/36.

[31] Создание и оценка значимости прогностических моделей для определения уровня ферритина сыворотки с помощью машинного обучения в разных клинических группах. / Вареха Н.В., Стуклов Н.И., Гимадиев Р.Р., Гордиенко К.В., Щеголев О.Б., Макарчев А.И. и Гуркина А.А. // Клиническая лабораторная диагностика. — 2025. — Т. 70, № 3. — С. 172–181.

[32] Лучинин А. С. Как врач сделал себе ИИ помощника. — Режим доступа: https://habr.com/ru/articles/709050/.

[33] Корнеенков А.А., Рязанцев С.В., Вяземская Е.Э. Вычисление и интерпретация показателей информативности диагностических медицинских технологий. // Медицинский Совет. — 2019. — № 20. — С. 45–51.

[34] Khurshid Shakir, Loganathan Bharath Kumar, Duvinage Matthieu. Comparative Evaluation of Applicability Domain Definition Methods for Regression Models. — 2024. — 2411.00920.

Приложение

Содержание

A Метрики качества классификационных моделей

В табл. 3 представлен используемый в данной работе набор классификационных метрик. Более полный перечень — в [33]. Большинство метрик зависит от порогового значения, которое разделяет непрерывную величину, рассчитываемую моделью, на два класса: патология и норма. Так как пороговое значение калибруется на целевых данных, то наиболее важной метрикой считаем площадь под кривой AUC, так как она не зависит от порогового значения и отражает кумулятивную мощь метода. Все остальные метрики, используемые в данной работе, приводим для полноты.

Таблица 3: Метрика качества классификационных моделей

Переменная	Описание
TP	(true positive) — количество верно спрогнозированных случаев патологии. Модель прогнозирует патологию, и лабораторный тест её подтверждает.
TN	(true negative) — количество верно спрогнозированных случаев отсутствия патологии. Модель прогнозирует, что показатель будет в норме, что подтверждается выполнением лабораторного теста.
FP	(false positive) — количество неверно спрогнозированных случаев патологии (ошибки первого рода). Модель ошибочно прогнозирует патологию, которая не подтверждается лабораторным тестом.
FN	(false negative) — количество неверно спрогнозированных случаев отсутствия патологии (ошибка второго рода). Модель не прогнозирует патологию, хотя она есть в соответствии с результатами лабораторного теста.
AUC	(area under curve) — площадь под характеристической кривой приёмника. Меняется в пределах $[0-1]$ или $[0-100]\%$. Значение 0,5 соответствует случайному классификатору, 1 — идеальному. На основании этой метрики сравнивались разные модели и выбирались лучшие.
Sensitivity	Чувствительность — процент корректно спрогнозированных патологических случаев среди всех патологических примеров. Значение 100% соответствует идеальному классификатору, который не пропускает патологию.
Specificity	Специфичность — процент корректно спрогнозированных случаев отсутствия патологии среди непатологических примеров. Значение 100% соответствует идеальному классификатору, который правильно идентифицирует примеры без патологий.
PPV	(positive predictive value, precision) — предсказательная ценность положительного результата модели.

B Описание данных для построения прогностических моделей

Исходные обезличенные данные результатов анализов пациентов были собраны в период 2022–2025 гг. в коммерческих лаборатория и частных медицинских центрах (со своей лабораторией) в России через драйверы оборудования в рамках функционирования лабораторной информационной системы RosLIS[8]. Всего было задействовано для данной работы 8 лабораторий, в разных федеральных округах РФ, табл. 4. В базе содержится 2 800 856 записей пациентов с результатами по общеклиническому анализу крови и биохимическим тестам.

Для построения моделей использовались все данные по всем лабораториям, чтобы исключить влияние потенциальных спутывающих (конфаундер) факторов, при которых модели показывают хорошие статистические параметры, но ограничены в применении из-за невозможности учёта этого фактора.

Таблица 4: Распределение количества записей по пациентам по лабораториям

Федеральный округ	Источник данных	Количество пациентов	Из них женщин, %	Средний возраст, годы	Всего по округу
Центральный	Лаборатория I	970 881	60,7	42 ± 19	2 681 815
	Лаборатория IV	3 629	81,6	38 ± 16	
	Лаборатория VIII	172 664	53,5	42 ± 15	
	Лаборатория IX	1 534 641	50,9	37 ± 14	
Северо-Западный	Лаборатория III	31 441	58,4	41 ± 14	31 441
Южный	Лаборатория VI	61 380	63,7	48 ± 19	61 380
Дальневосточный	Лаборатория V	22 428	59,8	38 ± 22	26 220
	Лаборатория VII	3 792	31,5	41 ± 17	
Итого		2 800 856	54,8	39 ± 16	

B.1 Область применимости модели

Любая математическая модель может надежно рассчитывать прогнозы только в том случае, если входные данные попадают в область её применимости (Applicability Domain) [34], которая обычно включает диапазоны для всех непрерывных входных и выходных параметров. Допустимые диапазоны значений могут быть определены для каждого параметра индивидуально путем построения функции плотности вероятности, которая отражает его изменение и учёта всех тех значений, которые входят в некоторую долю площади под этой кривой.

Наша реализация [6] этого алгоритма включает в себя построение гистограммы всех значений показателя, сортировку по ним, нахождение максимального значения и перемещение от этой точки в обоих направлениях до тех пор, пока доля совокупного числа просмотренных значений не достигнет порогового значения в 99%. Рассчитанные таким образом значения представлены на сайте УЛМ — `https://ulm.roslis.ru`.

B.2 Статистические характеристики исходных данных

Статистические характеристики всех исходных данных приведены в табл. 5. Для каждого параметра указывается[9] количество записей, среднее значение, медиана, максимум и минимум, а также изображена гистограмма с графиком ядерной оценки плотности вероятности распределения значений. Перед расчётом статистики из данных были исключены выбросы, определённые как описано в разделе Область применимости модели.

В приведенных диаграммах мы сознательно не «укорачивали длинные хвосты» распределений, а также использовали логарифмическое преобразование $ln(x)$ по всем параметрам кроме возраста, поэтому в некоторых случаях средние значения и медианы величин могут существенно различаться, а значения стандартных отклонений иметь ограниченное применение. Среди гематологических анализаторов, с которых были получены данные для данного исследования, есть как 5-DIFF, так и 3-DIFF, поэтому в таблице также присутствуют средние клетки и гранулоциты, при этом пересчет из 5-DIFF в 3-DIFF не производился.

[8] Лабораторная Информационная Система ЛИС RosLIS — `https://roslis.ru`

[9] Параметры определены на момент написания данного отчёта. Так как ежедневно поступают новые данные, то статистика может незначительно измениться.

Таблица 5: Статистические параметры исходных данных

Параметр	Ед. изм.	Кол-во	Среднее	Медиана	Стандартное отклонение	Минимум	Максимум	Гистограмма
Возраст	годы	2 800 856	39,5	39,0	16,3	1	120	Рис. 2
АЛТ	Ед/л	412 692	26,2	18,6	33,57	0,10	1 694,4	Рис. 3
АСТ	Ед/л	382 397	27,7	21,6	33,67	0,10	1 514,0	Рис. 4
Альбумин	г/л	24 459	42,8	43,4	4,88	0,04	62,3	Рис. 5
Базофилы	%	1 057 871	0,7	0,6	0,57	0,02	27,2	Рис. 6
Белок общий	г/л	238 484	71,9	72,1	5,41	19,20	136,6	Рис. 7
Билирубин непрямой	мкмоль/л	31 247	10,7	8,7	9,18	0,05	221,7	Рис. 8
Билирубин общий	мкмоль/л	331 643	12,7	10,6	10,73	0,03	428,3	Рис. 9
Билирубин прямой	мкмоль/л	116 862	2,7	2,0	6,32	0,04	242,9	Рис. 10
Витамин B12	пг/мл	28 195	511,7	427,6	777,83	1,00	39 833,0	Рис. 11
Гемоглобин (HGB)	г/л	2 800 524	138,6	139,0	16,64	30,00	213,0	Рис. 12
Гликированный гемоглобин	%	39 234	6,0	5,6	1,35	2,20	21,5	Рис. 13
Глюкоза	ммоль/л	2 254 856	5,1	5,0	1,20	0,01	26,3	Рис. 14
Гранулоциты	%	1 706 592	61,0	61,3	7,94	11,30	95,0	Рис. 15
Креатинин	мкмоль/л	333 506	98,0	82,0	102,87	0,30	1 618,8	Рис. 16
ЛДГ	Ед/л	44 786	236,0	198,0	132,53	2,00	4 983,0	Рис. 17
Лейкоциты	10^9/л	2 800 465	6,7	6,5	2,00	0,10	44,5	Рис. 18
Лимфоциты	%	2 799 488	31,7	31,3	8,55	0,10	87,8	Рис. 19
Моноциты	%	1 065 103	8,0	7,6	2,77	0,10	55,4	Рис. 20
Мочевая кислота	ммоль/л	111 753	1,7	0,3	20,44	0,00	491,0	Рис. 21
Мочевина	ммоль/л	265 633	5,7	5,1	3,50	0,50	67,0	Рис. 22
Нейтрофилы	%	1 093 929	56,4	56,8	11,27	3,00	100,0	Рис. 23
ПСА общий	нг/мл	21 237	9,0	1,1	101,35	0,01	4212,0	Рис. 24
С-реактивный белок	мг/л	24 071	9,4	2,2	22,55	0,01	332,4	Рис. 25
Средние клетки	%	1 735 923	7,8	7,1	3,56	0,10	37,4	Рис. 26
Средний объем эритроцитов	фл	2 709 100	88,7	89,0	6,66	12,00	130,3	Рис. 27
Триглицериды	ммоль/л	133 030	1,4	1,2	0,96	0,02	17,3	Рис. 28
Тромбоциты	10^9/л	2 800 422	258,1	252,0	66,02	1,00	917,0	Рис. 29
Ферритин	нг/мл мкг/л	117 244	82,0	42,3	113,36	0,01	1 813,0	Рис. 30
Фолиевая кислота	нг/мл	17 985	9,6	8,1	7,96	0,56	330,1	Рис. 31
Холестерин	ммоль/л	1 853 628	5,1	4,9	1,23	0,01	13,7	Рис. 32
Холестерин ЛПНП	ммоль/л	133 316	3,6	3,5	1,06	0,01	10,2	Рис. 33
Эозинофилы	%	1 056 838	2,6	2,1	2,14	0,09	43,9	Рис. 34
Эритроциты	10^{12}/л	2 800 521	4,7	4,7	0,53	0,78	7,9	Рис. 35

B.3 Гистограммы распределения параметров

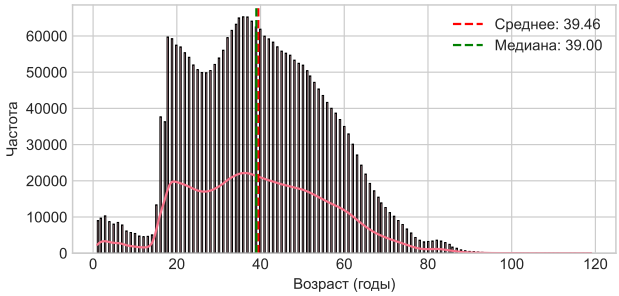

Рис. 2: Распределение по возрастам

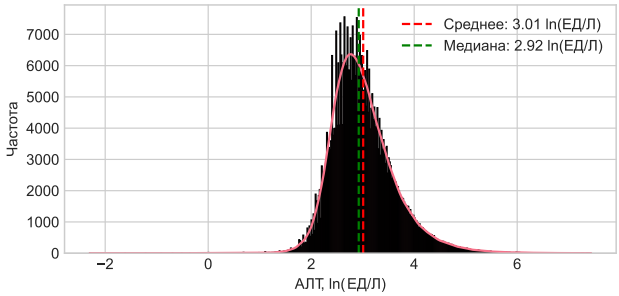

Рис. 3: Распределение по АЛТ

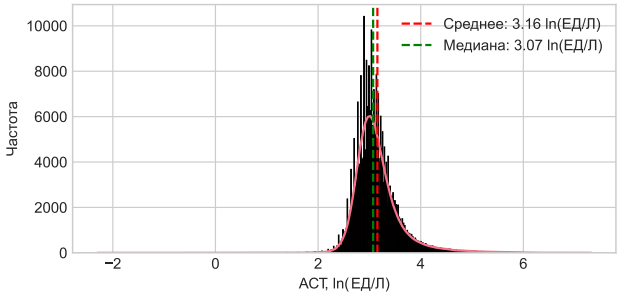

Рис. 4: Распределение по АСТ

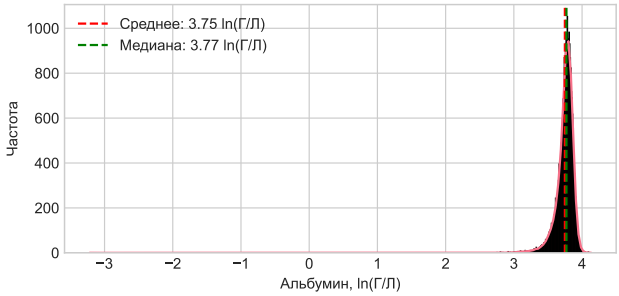

Рис. 5: Распределение по альбумину

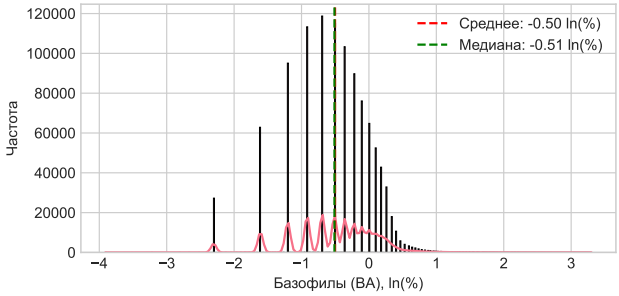

Рис. 6: Распределение по базофилам

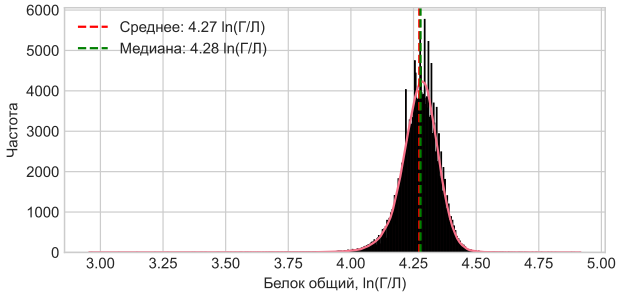

Рис. 7: Распределение по общему белку

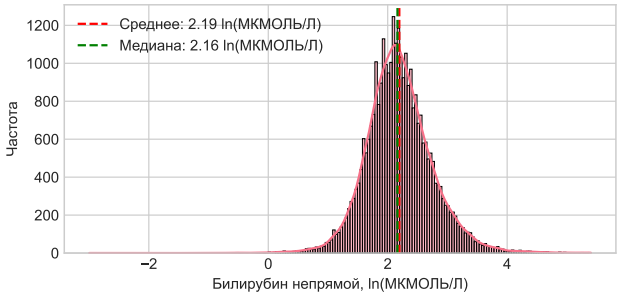

Рис. 8: Распределение по непрямому билирубину

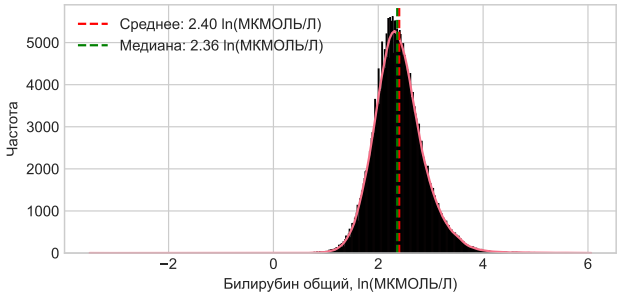

Рис. 9: Распределение по общему билирубину

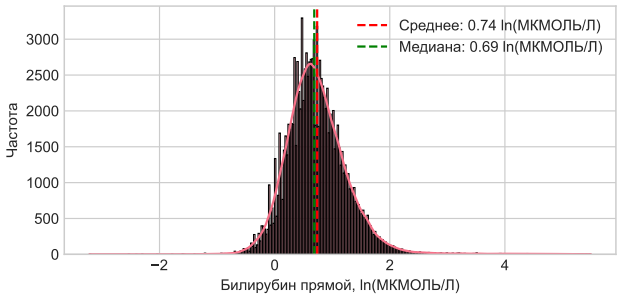

Рис. 10: Распределение по прямому билирубину

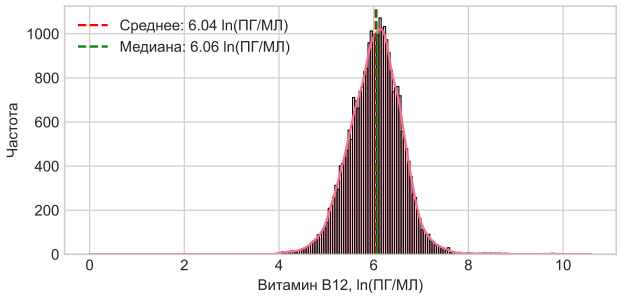

Рис. 11: Распределение по витамину B12

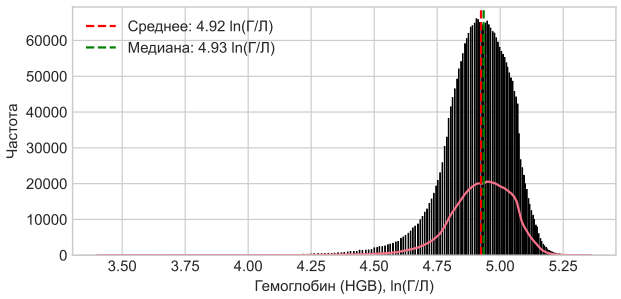

Рис. 12: Распределение по гемоглобину

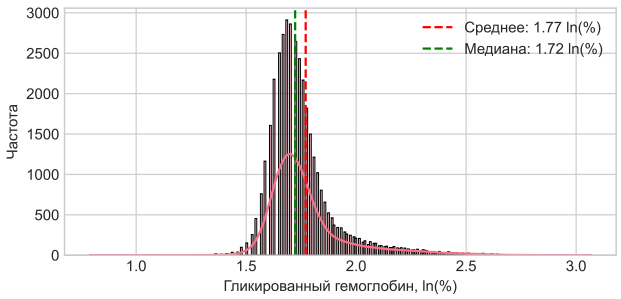

Рис. 13: Распределение по гликированному гемоглобину

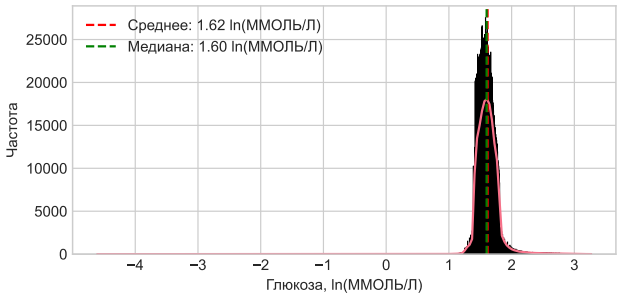

Рис. 14: Распределение по глюкозе

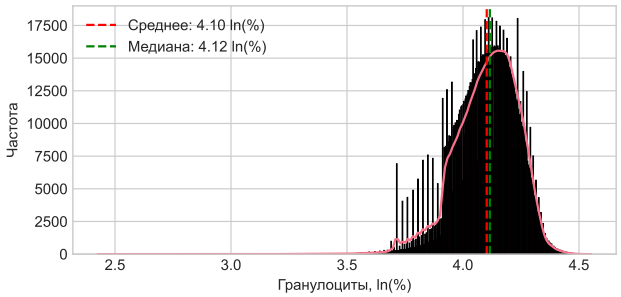

Рис. 15: Распределение по гранулоцитам

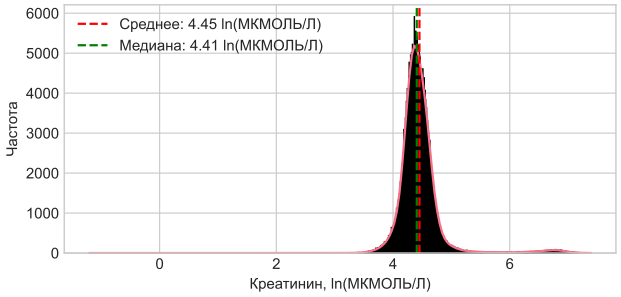

Рис. 16: Распределение по креатинину

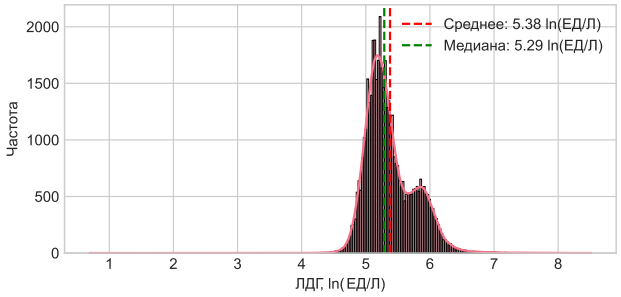

Рис. 17: Распределение по ЛДГ

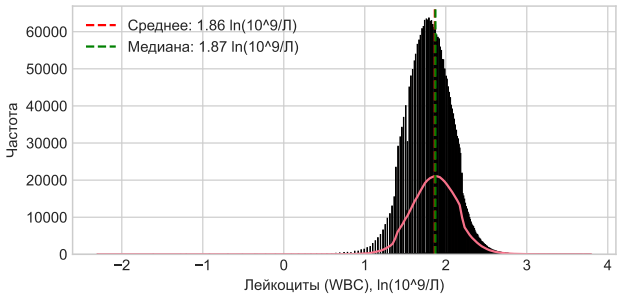

Рис. 18: Распределение по лейкоцитам

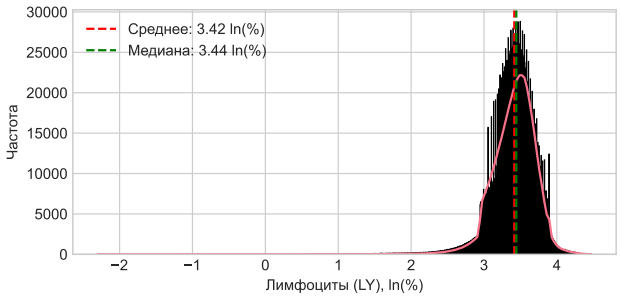

Рис. 19: Распределение по лимфоцитам

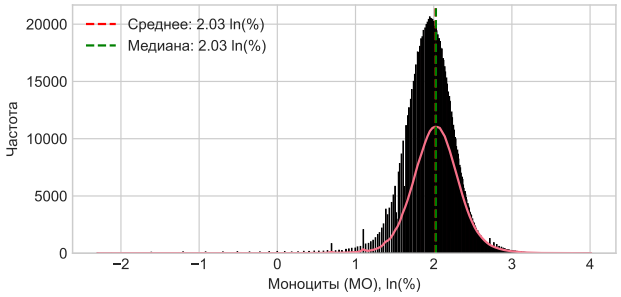

Рис. 20: Распределение по моноцитам

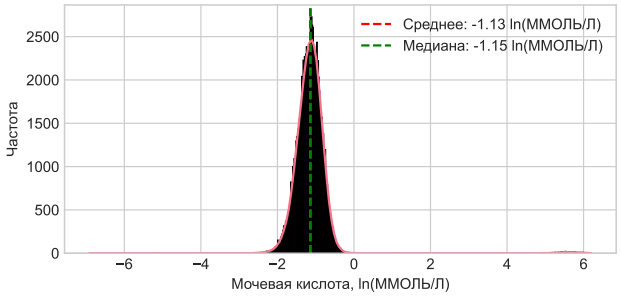

Рис. 21: Распределение по мочевой кислоте

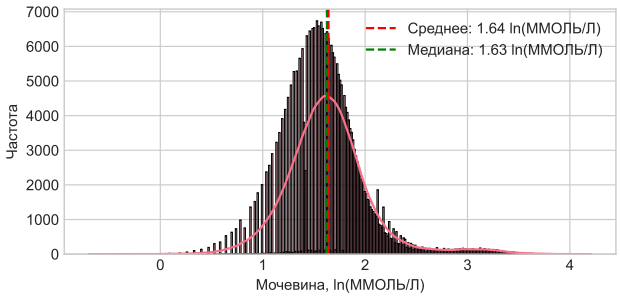

Рис. 22: Распределение по мочевине

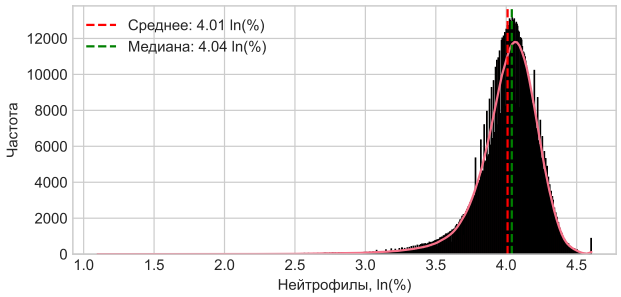

Рис. 23: Распределение по нейтрофилам

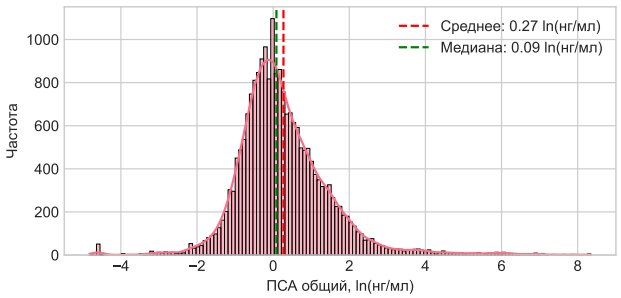

Рис. 24: Распределение по общему ПСА

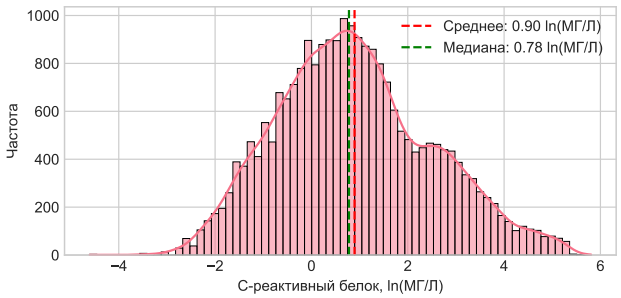

Рис. 25: Распределение по С-реактивному белку

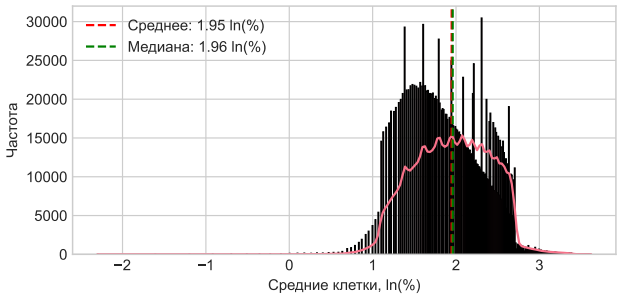

Рис. 26: Распределение по средним клеткам

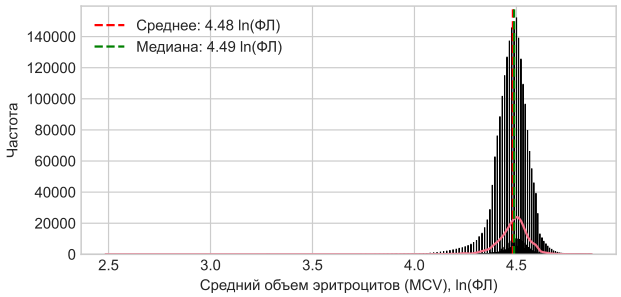

Рис. 27: Распределение по среднему объему эритроцитов

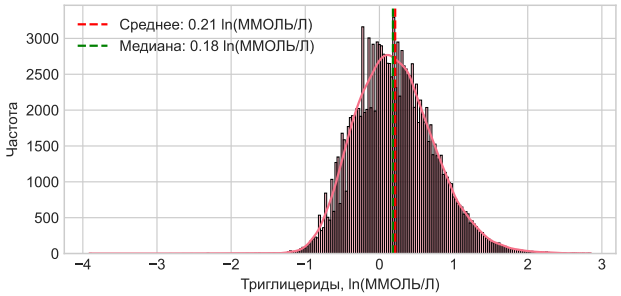

Рис. 28: Распределение по триглицеридам

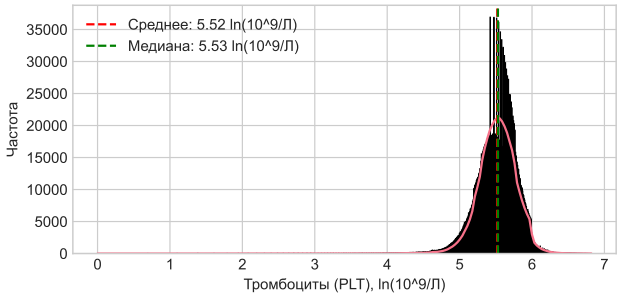

Рис. 29: Распределение по тромбоцитам

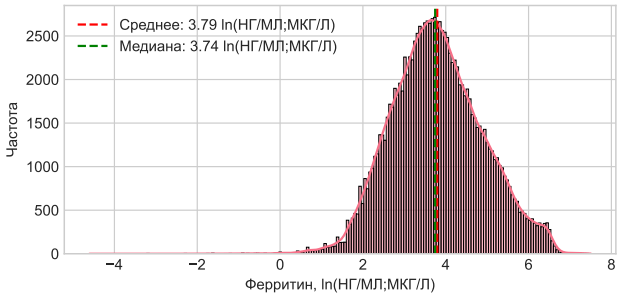

Рис. 30: Распределение по ферритину

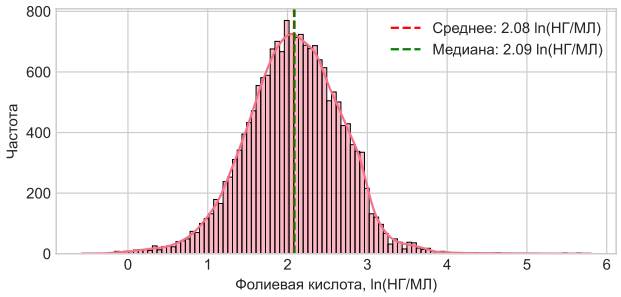

Рис. 31: Распределение по фолиевой кислоте

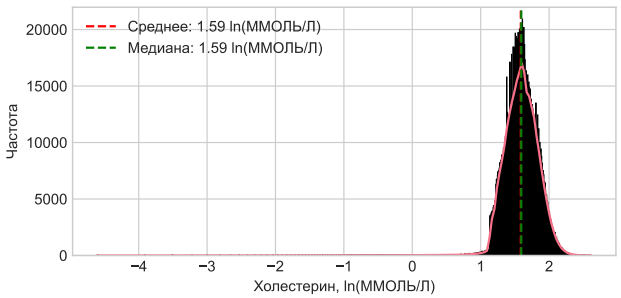

Рис. 32: Распределение по общему холестерину

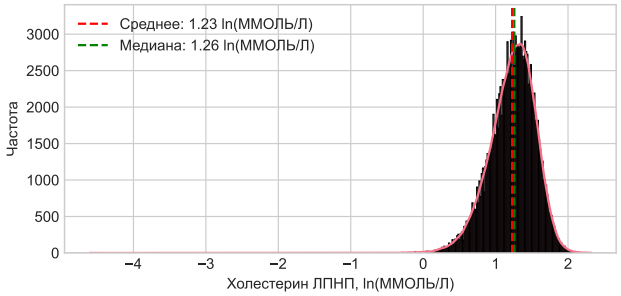

Рис. 33: Распределение по липопротеинам низкой плотности

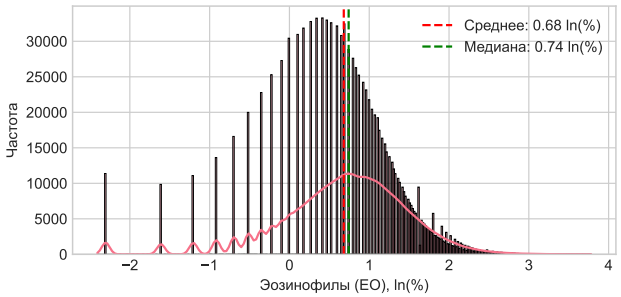

Рис. 34: Распределение по эозинофилам

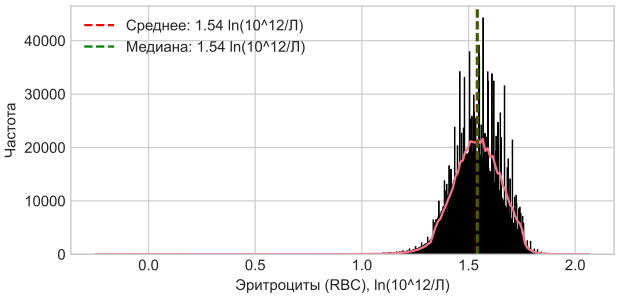

Рис. 35: Распределение по эритроцитам

C Параметры классификационных моделей

C.1 ROC кривые для моделируемых показателей

На графиках изображены характеристические кривые (ROC — Receiver Operating Characteristics, операционная характеристика приёмника) для всех моделируемых показателей из табл. 1. Жирной цветной кривой отражена характеристика классификационной модели, серая прямая $y = x$ соответствует случайному классификатору. Чем больше площадь под кривой, тем лучше классификатор отличает образцы обоих классов.

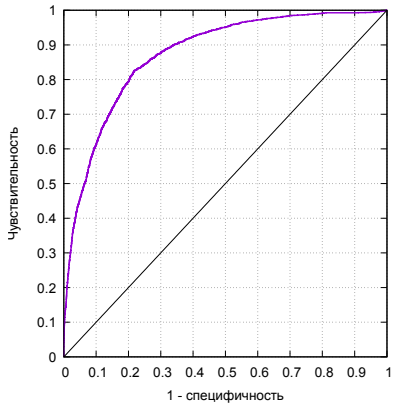

Рис. 36: Гликированный гемоглобин ($\geqslant 6,0\%$)

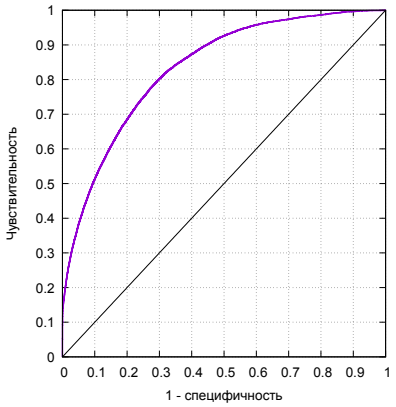

Рис. 37: Глюкоза ($\geqslant 7,0$ ммоль/л)

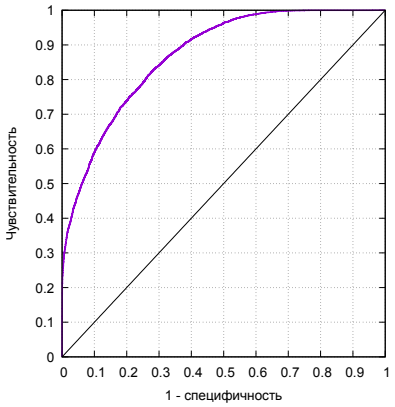

Рис. 38: Липопротеины НП ($\geqslant 3,4$, ммоль/л)

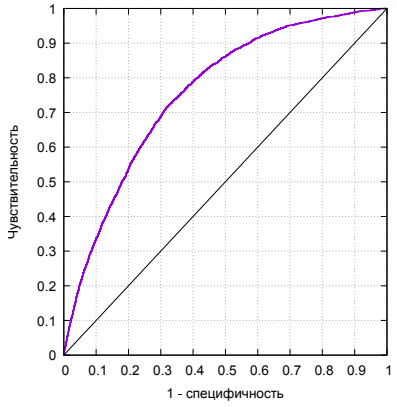

Рис. 39: Мочевая кислота ($\geqslant 0,48/0,38$, ммоль/л (М/Ж)

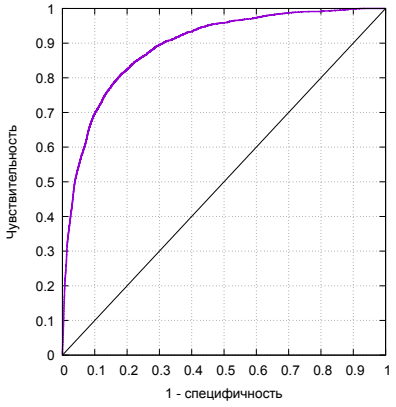

Рис. 40: Ферритин ($\leqslant 12$, нг/мл)

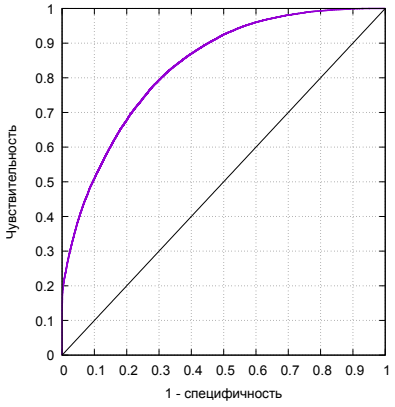

Рис. 41: Холестерин ($\geqslant 5,2$, ммоль/л)

C.2 Интерпретация моделей: анализ чувствительности

Для интерпретации моделей использовался метод анализа чувствительности, в рамках которого каждому входному параметру модели поочередно произвольным образом добавлялся произвольный шум и оценивалось изменение выходных параметров, инициированное этим шумом. Тем самым можно определить наиболее значимые параметры, на основе которых строится прогноз. Нормированные на максимальное изменение целевых значений коэффициенты чувствительности представлены в табл. 6.

Таблица 6: Коэффициенты чувствительности (безразмерная величина): чем больше значение тем больший вклад вносит параметр (в строках) в моделируемое свойство (в столбцах)

Наименование параметра	Гликированный гемоглобин	Глюкоза	Липопротеины низкой плотности	Мочевая кислота	Ферритин	Холестерин
Возраст	0,3149	0,0752	0,9831	0,3271	0,0594	0,8794
АЛТ	0,6151	0,1941	0,5651	0,7163	0,0939	0,4843
АСТ	0,1732	0,1388	0,1982	0,3526	0,1119	0,1851
Альбумин	0,0726	0,1450	0,2371	0,2593	0,1231	0,1304
Базофилы	0,0440	0,0241	0,0730	0,0924	0,0338	0,0816
Белок общий	0,1170	0,0881	0,4287	0,3044	0,0630	0,4925
Билирубин непрямой	0,0200	0,0513	0,2076	0,1032	0,0575	0,1551
Билирубин общий	0,0848	0,0515	0,3322	0,4652	0,2862	0,2219
Билирубин прямой	0,1431	0,0728	0,3239	0,5319	0,0567	0,3248
Витамин B12	0,0231	0,0209	0,1848	0,0685	0,0292	0,0916
Гемоглобин (HGB)	0,2352	0,5841	0,5861	0,7745	1,0000	0,4917
Гранулоциты	0,0620	0,0225	0,1310	0,1380	0,0899	0,1195
Креатинин	0,2417	0,0442	0,4601	0,7001	0,1802	0,4098
ЛДГ	0,1605	0,0919	0,0919	0,1613	0,0511	0,0899
Лейкоциты	0,8741	0,5347	0,3363	0,6228	0,2331	0,1960
Лимфоциты	0,0250	0,0204	0,2795	0,0775	0,0539	0,3335
Моноциты	0,0610	0,0687	0,1434	0,1144	0,0908	0,0787
Мочевина	0,4578	0,2517	0,4740	0,6982	0,1453	0,3866
Нейтрофилы	0,0599	0,0481	0,0952	0,2332	0,0535	0,1362
ПСА общий	0,1393	0,0249	0,1650	0,2844	0,0641	0,0978
С-реактивный белок	0,2001	0,0551	0,4186	0,4310	0,3404	0,3441
Средние клетки	0,0766	0,0400	0,0925	0,1411	0,0718	0,1198
Средний объем эритроцитов	0,0996	0,0400	0,4688	0,2406	0,0725	0,3320
Триглицериды	1,0000	1,0000	1,0000	1,0000	0,3331	0,8163
Тромбоциты	0,1243	0,1046	0,2533	0,1172	0,2337	0,3484
Фолиевая кислота	0,2493	0,1548	0,2735	0,3602	0,2302	0,1950
Эозинофилы	0,0416	0,0337	0,0767	0,1110	0,0613	0,0999
Эритроциты	0,3742	0,2577	0,3787	0,2843	0,1290	0,3312

Вклады всех параметров в окончательное значение моделируемого свойства существенны ($\geqslant 10^{-3}$). Это косвенно подтверждает важность каждого из этих параметров, а также отсутствие одного доминирующего свойства. Данная работа не ставила перед собой цель подробную интерпретацию модели — это задача будущих исследований.

D Спутывающие факторы в регрессионной модели по ЛПНП

При добавлении новых тестов в УЛМ была обнаружена хорошая регрессионная модель по прогнозированию значений липопротеинов низкой плотности с коэффициентом детерминации $R^2 = 0,93$ и абсолютной средней

ошибкой всего 0,23 ммоль/л на данных лаборатории I. Похожие характеристики моделей сообщает Гимадиев Р. Р. в своей презентации, а также в тезисах конференции [27].

По нашему опыту высокие коэффициенты корреляции часто свидетельствуют об ошибке. В данном случае, модель построенная на данных лаборатории I неудовлетворительно показала себя при прогнозе данных других лабораторий, при этом ошибка MAE увеличивалась более, чем в два раза.

Анализ чувствительности данной модели показал, что практически весь сигнал приходит от одного параметра — общего холестерина, а простая линейная регрессия вида

$$LDLC = 0,72486751 * CHOL - 0,44804567$$

имеет характеристики $R^2 = 0,91, RMSE = 0,31$ ммоль/л, $MAE = 0,23$ ммоль/л или в процентах 0,23 / 3,6 * 100% = 6,3%, что выглядит неправдоподобно, учитывая общую допустимую ошибку определения общего холестерина с учетом биологической вариации в 11,9% [25, с. 25].

Если же строить данные только по лаборатории V, то $R^2 = 0,75$; по лаборатории III: $R^2 = 0,70$. Полагаем, что в данном случае наблюдается эффект «спутывающего» фактора, поэтому в его отсутствии модели не работают. Возможно, также наблюдается эффект линеаризации, учитывая что зависимость ЛПНП от общего холестерина линейна по используемым в лабораторной практике расчётным формулам Фридмана, Мартина и др.